Incorporating Covariates into Multipoint Association Mapping in the Case-Parent Design

Yen-Feng Chiu^a Kung-Yee Liang^{a, b} Wen-Harn Pan^c

^aDivision of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan, ROC; ^bDepartment of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Md., USA; ^cInstitute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, ROC

Key Words

Case-parent designs • Gene-gene interactions • Gene-environment interactions • Gene-covariate interactions • Relative efficiency • Parametric • Non-parametric

Abstract

Background/Aims: To improve the efficiency of disease locus localization in association mapping using case-parent designs and to assess or account for the main covariate effects and gene-covariate interaction effects, while localizing the disease locus. **Methods:** The present study extends a multipoint fine-mapping approach to incorporate covariates into the association mapping of case-parent designs through parametric and non-parametric modeling. This approach is based on the expected preferential-allele-transmission statistics for transmission from either parent to an affected child. **Results:** Simulation studies indicate that the efficiency in estimating the disease locus increases considerably when incorporating a covariate associated with the disease. This is especially true when the genetic effect of the disease locus is small. The proposed approach was applied to a young-onset hypertension data sample. The relative efficiency of estimating the locus of young-onset hypertension increases 110-fold after incorporating triglyceride into the association mapping while localizing the disease variant in the lipoprotein lipase

gene in the non-parametric model. By incorporating the information of SNP variants into the fine-mapping, the proposed method further assesses the gene-gene interactions between the SNP and the disease locus. **Conclusion:** With the incorporation of covariates, the proposed method cannot only improve efficiency in estimating disease loci, but can also elucidate the etiology of a complex disease.

Copyright © 2010 S. Karger AG, Basel

Introduction

The etiologies of complex diseases with an early onset in life, such as asthma, schizophrenia, birth defects, etc., often involve both genetic and environmental components. The genetic component might itself be complex, involving the affected individual's inherited genes as well as maternal-mediated mechanisms or environmental exposures during gestation [1, 2]. Hence, researchers often adopt the case-parent design to assess the gene-environment interaction. This design includes affected offspring and both of their parents, allowing for testing and estimation of offspring- and maternally mediated genetic effects [3, 4].

Schaid [5] illustrated that the case-parent design is more powerful than the case-control design in detecting gene-environment interactions (GxE) for rare disease

susceptibility alleles. Specifically, he used a logistic regression model to assess GxE by estimating the genotype relative risks for one locus at a time. This approach permits testing for GxE, subject to the constraints of an assumed genetic relative risk model. Umbach and Weinberg [6] proposed a log-linear modeling approach to estimate the genotype-exposure interaction effect, but it was infeasible to include continuous environmental factors. Lim et al. [7] used a multinomial logistic regression method to overcome this constraint, but their approach included many parameters and often encountered small or zero frequencies in some categories. Baksh et al. [8] used a two-stage modeling approach to model the covariate and gene-covariate interaction terms over conditional score tests that include covariate information. They estimated nuisance parameter(s) in the penetrance function instead of using external estimates [9, 10], which can be unreliable [11]. By relaxing the conditioning on parental genotypes to allow exchangeability of parental genotypes, Cordell et al. [11] extended the conditional logistic approach to fitting models for multiallelic markers, multiple linked loci, multiple linked loci in multiple unlinked regions, and gene-gene and gene-environment interactions without the requirement of estimating the large number of nuisance parameters in these situations. Although these likelihood-based approaches can assess gene-environment interactions, the number of parameters may increase substantially when the genetic model becomes too complex. The parameters in a genetic model, such as disease allele frequency, disease locus, and penetrances, are mostly unknown, and must be estimated with likelihood-based approaches.

On the other hand, the family-based approach proposed by Rabinowitz and Laird [12] is based on computing p values by comparing test statistics for association with their conditional distributions given the minimal sufficient statistic under the null hypothesis for the genetic model, sampling plan, and population admixture. This approach can be applied to any kind of phenotype, and multi-allelic markers may be examined without estimating or specifying many parameters, such as parental mating type frequencies. This approach can also accommodate gene-environment effects, but focuses on testing associations with or without the presence of linkage rather than estimating the disease locus or gene-environment effects.

The semi-parametric linkage mapping for the case-parent trios design proposed by Liang et al. [13] utilizes all the markers simultaneously to localize the disease locus while making no assumption about genetic mechanism,

except that only one disease gene lies in the region under study. The advantage of this approach is that it does not require the specification of an underlying genetic model, so estimating the position of a disease locus and its standard error is robust to a wide variety of genetic mechanisms. Hsu et al. [14] extended this approach to incorporate the genotypic information of an unlinked locus into the linkage disequilibrium (LD) mapping to assess gene-gene interaction. Through stratification on a categorical variable, their approach [14] is applicable to incorporate qualitative covariates only. Incorporating covariates associated with the disease locus in fine-mapping, be they qualitative or quantitative, can improve the efficiency of estimating disease loci and elucidates the covariate's role in the genetic mechanism underlying a complex trait. Hence, it is important to extend this approach of Liang et al. [13] to incorporate quantitative and qualitative covariates. We consider quantitative covariates in the present study subsequently as one can always stratify the sample by qualitative covariates [14]; the proposed method can also handle qualitative covariates, however.

Method

Notation and Preferential-Transmission Statistic

Consider n case-parent trios sampled for an association study, and let R be a chromosomal region of length T cM which contains no more than one susceptibility locus at τ in the region R . Suppose M markers were genotyped in the region R at locations of $0 \leq t_1 < t_2 < \dots < t_M \leq T$. Following the notation in Liang et al. [13], we assume there are two alleles per marker with $H(t)$ being the target allele at marker t , and $h(t)$ the non-target allele, and define $Y(t)$ as the paternal preferential-transmission statistic

$$Y(t) = Y_{d1}(t) - Y_{d2}(t), \quad (1)$$

where t is an arbitrary location in this region R , and

$$Y_{d1}(t) = \begin{cases} 1, & \text{if the transmitted paternal allele at } t \text{ is the target allele } H(t) \\ 0, & \text{if the transmitted paternal allele at } t \text{ is the non-target allele } h(t), \end{cases} \quad (2)$$

$$Y_{d2}(t) = \begin{cases} 1, & \text{if the non-transmitted paternal allele at } t \text{ is the target allele } H(t) \\ 0, & \text{if the non-transmitted paternal allele at } t \text{ is the non-target allele } h(t). \end{cases}$$

Similarly, the maternal preferential-transmission statistic can be defined by

$$X(t) = X_{d1}(t) - X_{d2}(t). \quad (3)$$

Let Φ denote the affected child; consider a situation where n independent case-parent trios are sampled and the paternal and maternal preferential-transmission statistics $Y_i(t_j)$ and $X_i(t_j)$, respectively, are computed at each of M markers located at $0 \leq t_1 < t_2 < \dots < t_M \leq T$, $i = 1, \dots, n$. Liang et al. [13] showed that the expected number of a target allele at t_j being transmitted to the affected child Φ is

$$\mu(t_j) = E[Y_i(t_j)|\Phi] = E[X_i(t_j)|\Phi] = (1 - 2\theta_{t_j,\tau})C(1 - \theta_{t_j,\tau})^N \pi_j, \quad (4)$$

where $C = E[Y(\tau)|\Phi] = E[X(\tau)|\Phi]$, $\theta_{t_j,\tau}$ is the recombination fraction between t_j and τ , N is the number of previous generations to the one in which a disease-causing mutation at τ was introduced, $i = 1, \dots, n$, and $\pi_j = \Pr[h(t_j)|h(\tau)]$, $j = 1, \dots, M$.

The genetic effect of τ , characterized by C , is the transmission probability that the affected offspring will carry the disease allele, $H(\tau)$, at τ . This genetic effect is postulated to be associated with a vector of covariates $\mathbf{Z}$ (assuming that recombination and linkage disequilibrium do not depend on $\mathbf{Z}$). Therefore, the expectation of the preferential-transmission statistic can be re-written as

$$\mu(t_j) = E[Y(t_j)|\mathbf{Z} = \mathbf{z}, \Phi] = (1 - 2\theta_{t_j,\tau})C(\mathbf{z})(1 - \theta_{t_j,\tau})^N \pi_j, \quad (5)$$

where $C(\mathbf{z})$ is $E[Y(\tau)|\mathbf{Z} = \mathbf{z}, \Phi]$. The parameter C , which measures the genetic effect of τ , is a function of the covariate $\mathbf{Z}$, where $\mathbf{Z} = (Z_1, Z_2, \dots, Z_p)^T$ is a vector of p covariates. The relationship between C and $\mathbf{Z}$ can be modeled parametrically or non-parametrically.

The Parametric Model

Multiple parametric methods can be utilized to model C as a function of the covariates. The proposed approach employed logistic type models to establish the relationship between a vector of covariates $\mathbf{Z}$ and $C(\mathbf{z})$, where $C(\mathbf{z})$ is the dependent variable [15].

$$\text{logit}[\{C(\mathbf{z}) + 1\}/2] = \text{log} \left[\frac{1 + C(\mathbf{z})}{1 - C(\mathbf{z})} \right] = \alpha + \beta^T \mathbf{z}. \quad (6)$$

$(C(\mathbf{z}) + 1)/2$ represents the probability that an affected child will receive a target allele at τ from his or her heterozygous parent. Thus,

$$C(\mathbf{z}) = \frac{\exp(\alpha + \beta^T \mathbf{z}) - 1}{\exp(\alpha + \beta^T \mathbf{z}) + 1}. \quad (7)$$

Replacing $C(\mathbf{z})$ in (5) by Eq. (7) leads to

$$E[Y(t_j)|\mathbf{Z} = \mathbf{z}, \Phi] = (1 - 2\theta_{t_j,\tau}) \left(1 - \frac{2}{\exp(\alpha + \beta^T \mathbf{z}) + 1} \right) (1 - \theta_{t_j,\tau})^N \pi_j. \quad (8)$$

Following Liang et al. [13], this method assumes the Haldane [16] map function

$$\theta_{t_j,\tau} = (1 - \exp(-0.02|t_j - \tau|))/2, \quad (9)$$

and estimates π_j by $\hat{\pi}_j$, namely,

$$\hat{\pi}_j = \frac{\sum_{i=1}^n [1 - Y_{id2}(t_j) + 1 - X_{id2}(t_j)]}{2n}, \quad j = 1, \dots, M. \quad (10)$$

Through the Generalized Estimating Equation (GEE) [17] approach, one can solve equation (11) to estimate the vector of parameters $\delta = (\tau, N, \alpha, \beta_1, \dots, \beta_p)$, where p is the dimension of covariates.

$$S(\delta) = \sum_{i=1}^n \left[\frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \text{Cov}^{-1}(Y_i) \{Y_i - \mu(\delta, \hat{\pi})\} + \frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \text{Cov}^{-1}(X_i) \{X_i - \mu(\delta, \hat{\pi})\} \right] = 0, \quad (11)$$

The Non-parametric Model

A criticism of multiple parametric modeling is that it may not accurately reflect the underlying mechanism. This method also employs the non-parametric model proposed by Chiou et al. [18] for linkage analysis to estimate the function C by spline and kernel smoothing methods as local polynomial regression [19]. The expected number of disease alleles being transmitted at $\tilde{\tau}$ was estimated by

$$Y_i^*(\tilde{\tau}) = \sum_{j=1}^M w_j(\tilde{\tau}) Y_i(t_j), \quad (12)$$

where $w_j(\tilde{\tau})$ is the weight function of nearby markers centered at $\tilde{\tau}$. The weight function here comprises the statistic at the two nearest markers, $Y_i(t_l)$ and $Y_i(t_k)$ with $t_k < \tilde{\tau} < t_l$ such that

$$Y_i^*(\tilde{\tau}) = w Y_i(t_k) + (1 - w) Y_i(t_l), \quad (13)$$

where $w = (t_l - \tilde{\tau})/(t_l - t_k)$. Similarly, $X_i^*(\tilde{\tau})$ can be estimated. Let

$$Y_i^{**}(\tilde{\tau}) = \frac{(X_i^*(\tilde{\tau}) + Y_i^*(\tilde{\tau}))}{2},$$

then $\hat{C}(\mathbf{z})$ can be obtained by minimizing the following kernel weighted least squares function

$$\sum_{i=1}^n [(Y_i^{**}(\tilde{\tau}) - 1) - \alpha - \beta^T (\mathbf{Z}_i - \mathbf{z})]^2 K_H(\mathbf{Z}_i - \mathbf{z}); \quad (14)$$

where $\beta^T = (\beta_1, \dots, \beta_p)$, $\mathbf{H}$ is a $p \times p$ symmetric positive definite matrix depending on sample size n ; K is a p variate kernel such that $\int K(\mathbf{u}) d\mathbf{u} = 1$; and $K_H(\mathbf{u}) = |\mathbf{H}^{-1/2}| K(\mathbf{H}^{-1/2} \mathbf{u})$, $\mathbf{H}^{1/2}$ is the bandwidth matrix, where bandwidths describe the width of the convolution kernel used; and $\hat{\alpha}$ is the estimate of C [20]. Here, the kernel function K and the bandwidth matrix $\mathbf{H}^{1/2}$ are

$$K(\mathbf{u}) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{u}\right), \quad (15)$$

and

$$\mathbf{H}^{1/2} = \text{diag} \left\{ \frac{\max(Z_1) - \min(Z_1)}{3}, \frac{\max(Z_2) - \min(Z_2)}{3}, \dots, \frac{\max(Z_p) - \min(Z_p)}{3} \right\}. \quad (16)$$

Letting

$$\mathbf{Z}_z = \begin{bmatrix} 1 & (\mathbf{Z}_1 - \mathbf{z})^T \\ \vdots & \vdots \\ 1 & (\mathbf{Z}_n - \mathbf{z})^T \end{bmatrix}, Y^{**} = [Y_1^{**}, \dots, Y_n^{**}]^T \text{ and} \quad (17)$$

$$\mathbf{W}_z = \text{diag} \{K_H(\mathbf{Z}_1 - \mathbf{z}), \dots, K_H(\mathbf{Z}_n - \mathbf{z})\},$$

and assuming that $\mathbf{Z}_z^T \mathbf{W}_z \mathbf{Z}_z$ is nonsingular, Eq. (14) produces the following solution:

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = (\mathbf{Z}_z^T \mathbf{W}_z \mathbf{Z}_z)^{-1} \mathbf{Z}_z^T \mathbf{W}_z \mathbf{Y}^{**}(\bar{\tau}). \quad (18)$$

After obtaining $\hat{C}(z)$, plug it into the following GEE equation to update the estimate of $\hat{\delta}$:

$$S(\hat{\delta}) = \sum_{i=1}^n \left[\frac{\partial \mu(\hat{\delta}, \hat{\pi})}{\partial \hat{\delta}} \text{Cov}^{-1}(Y_i) \{Y_i - \mu(\hat{\delta}, \hat{\pi})\} + \frac{\partial \mu(\hat{\delta}, \hat{\pi})}{\partial \hat{\delta}} \text{Cov}^{-1}(X_i) \{X_i - \mu(\hat{\delta}, \hat{\pi})\} \right] = 0, \quad (19)$$

where $\hat{\delta} = (\tau, N)$. The iterative process of updating $\hat{C}(z)$ in the non-parametric model and $\hat{\delta}$ in the GEE approach produces the estimates of $C(z)$ and $\hat{\delta}$ upon reaching convergence. The parameter estimates obtained from the GEE approach exhibit asymptotic normality, making it possible to make inference about the parameters (see Appendix for the details).

Simulation Study

To assess the performance of the parametric and nonparametric methods under different genetic mechanisms, three genetic models were assumed to generate the affection status in the simulation study.

1. Fixed Penetrance Models

The first genetic model is a fixed penetrance model, which determines the affection status by the penetrance rates f_0 , f_1 , and f_2 for genotypes hh , Hh , and HH at the disease locus, respectively.

2. Threshold Models

The second genetic model is a threshold model, which determines the affection status by a threshold (see fig. 1) for a quantitative trait following a normal distribution [21]. An individual is classified as a case if his or her trait exceeds the threshold. This model sets the threshold at 1.02 to produce a prevalence rate of 0.05.

3. GxE Models

The third genetic model postulates a prospective logistic regression model [22] as the penetrance function, and generates binary disease outcomes for case-parent trio data

$$\log \left(\frac{P(D=1 | g_1, g_2, x)}{1 - P(D=1 | g_1, g_2, x)} \right) = \beta_0 + \beta_1 g_1 + \beta_2 g_2 + \beta_3 x + \beta_4 (xg_1) + \beta_5 (xg_2), \quad (20)$$

where $\beta_0 = \ln(0.001)$, $\beta_1 = \ln(14)$, $\beta_2 = \ln(7)$, $\beta_3 = \ln(15)$, $\beta_4 = \ln(7)$, $\beta_5 = \ln(7)$, and

$$g_1 = \begin{cases} 1, & \text{if } HH \\ 0, & \text{if } Hh \text{ or } hh \end{cases}, g_2 = \begin{cases} 1, & \text{if } Hh \\ 0, & \text{if } HH \text{ or } hh \end{cases}, X \sim N(1, 1); \quad (21)$$

where $D = 1$ for affected individuals and $D = 0$ for unaffected individuals, and the parameter vector $\beta = (\beta_1, \dots, \beta_5)$ contains the natural logarithms of the odds ratios (ORs) ($\beta_1, \beta_2, \beta_3$) and the natural logarithms of the ratios of odds ratios (β_4, β_5). The environmental factor appears as a covariate in the fine-mapping.

Fig. 1. Graphical illustration of disease models. D = Disease; E = environmental factor; G = gene (susceptibility locus).

On the other hand, a quantitative trait [21] and age at onset were also generated as a covariate based on the genotypes at the disease locus for affected individuals. The quantitative trait is

$$z_i = \mu + g_i + e_i, \quad e_i \sim N(0, 1), \quad i = 1, \dots, n, \quad (21)$$

where n is the total number of trios, $g_i = a, d, -a$ for the genotypes HH, Hh , or hh respectively; $a = 1$; $d = 1$ for dominant models, $d = -1$ for recessive models, and $d = 0$ for additive models; and e_i are independent random variables that follow a standard normal distribution. Age at onset is also a covariate (z_i) that follows an extreme value distribution, that is, the logarithm of age at onset is a Weibull distribution [23],

$$\text{Log}(\text{age at onset } (z_i)) = -\log \lambda - \beta z_i + v_i / \gamma, \quad (22)$$

where $\lambda = 0.02$, $\beta = 0.1$, $\gamma = 40$, z_i is the number of H at τ , and v_i follows the standard extreme value distribution, $i = 1, \dots, n$. Using the same algorithm as Liang and Chiu [24], where random mating and the Hardy-Weinberg equilibrium were assumed, genotype data at the disease locus and marker loci were generated for parents, then for their child. Ten or 20 bi-allelic markers spanning 0.9

cM were simulated for each configuration. The disease locus was located in the middle of the region, either at 0.45 or at 0.475 cM. In addition, 91 SNPs with a SNP every 0.01 cM where a marker locus at 0.45 cM is also the disease locus were generated, a covariate (trait) controlled by different quantitative trait loci (QTLs) was simulated to study the QTL's impact on estimating the disease locus.

A child's affection status was determined by the genotypes at the disease locus using the genetic models of fixed penetrance, threshold, or logistic regression with gene-environment (GxE) interactions. Each scenario was simulated for 1,000 replicates with 200 case-parent trios in each replicate. The relative efficiency of estimating the disease locus using the parametric and non-parametric approaches with an incorporated covariate was compared to the original approach without a covariate. The performance of these approaches under different genetic models was also examined.

Results

The performance of approaches that incorporate a quantitative covariate through parametric and non-parametric modeling was assessed in this study. The estimate of the disease locus, its standard error, bias, and 95% coverage probability, and the relative efficiency of comparing estimates for the disease locus with and without covariates are presented. The parametric model also assesses the relationship between the genetic effect at the trait locus and the covariate and its significance.

Since the covariate is either a quantitative trait controlled by the disease locus or a quantitative environmental factor associated with the disease, incorporating the covariate provides additional information about the disease locus. This, in turn, increases the efficiency of estimating τ (table 1). The increase in the disease allele frequency leads to an increase in C . Hence, the standard error for the estimate of τ is smaller at larger disease allele frequencies. Regardless of the underlying genetic disease model, the efficiency gain of incorporating a covariate is higher for small C since a larger C refers to a relatively more homogeneous genetic effect. As a result, the incorporation of covariates improves efficiency more when C is small. The relative efficiency of estimating the disease locus also depends on the covariates incorporated. Under the GxE genetic model, incorporating age at onset and the quantitative trait at the disease locus improves efficiency more than incorporating the environmental variable. Incorporating age at onset is most effective as it has a greater genetic effect on the disease locus than the other two covariates (table 1c). Denser markers also improve the efficiency in estimating τ for all the es-

timates, regardless of whether or not a covariate was incorporated. As a result, the magnitude of the relative efficiency decreases (table 2). Table 3 shows that under the GxE genetic model, relative efficiency increases as the environmental effect increases, even though the standard errors of $\hat{\tau}$ also increase. On the other hand, when the covariate is a trait, its genetic component has some impact on the relative efficiency – the higher the additive genetic effect is, the higher efficiency becomes when incorporating covariates under the GxE model (table 4). Table 5 shows that the relative efficiency and the significance of the covariate decrease when the QTL's position for the covariate is farther away from the disease locus, regardless of the genetic model of the covariate (quantitative trait). These features can help investigators select a covariate with the strongest genetic association with the disease locus.

Three different LD patterns among the non-disease SNPs were simulated with 50, 100, and 150 generations N to examine the impact of LD patterns (table 6). The results suggest that the efficiency, as well as the relative efficiency in estimating the disease locus τ , tends to decrease when the LD between markers increases; however, the effect is not substantial. In addition, incorporating a covariate not associated with the disease locus does not reduce the efficiency of estimating the disease locus in either parametric or non-parametric models compared to the estimate without the covariate. With three additional irrelevant traits as covariates, the efficiency in estimating τ remains similar for both parametric and non-parametric approaches (table 7); however, the convergence rate dropped to 74.5% for the non-parametric approach (when there is only one covariate incorporated in the non-parametric model, the convergence rate is around 99.2–100%, depending on the specific covariate). The convergence rate and efficiency are fairly stable in the simulation studies for the parametric approach when incorporating additional covariates not associated with the disease locus. As a result, selecting relevant covariates into the models remains crucial, as, in general, with many irrelevant covariates there can be problems of non-convergence or loss of efficiency, particularly in the non-parametric approach. The relative efficiency of the estimate of the disease locus also depends on whether or not a marker is genotyped at the disease locus. The relative efficiency is higher when a marker is genotyped at the disease locus than when it is not.

Table 1. Relative efficiency of estimating the disease locus using different genetic models

f_2, f_1, f_0	P(H τ)		C	τ	Bias	R.E. ^d	β	p value	95% coverage probability
a Fixed penetrance genetic model: Covariate: the trait with a QTL at τ									
0.67, 0.05, 0.007	0.05	parametric ^a	0.27	0.4504 $\pm$ 0.0237	0.0004	1.26	0.37 $\pm$ 0.050	1.46E-13	0.98
		non-parametric ^b		0.4505 $\pm$ 0.023	0.0005	1.34			0.97
		original ^c		0.4502 $\pm$ 0.0266	0.0002				0.98
0.67, 0.05, 0.007	0.1	parametric	0.42	0.4495 $\pm$ 0.0147	-0.0005	1.12	0.52 $\pm$ 0.0047	<0.00E-15	0.99
		non-parametric		0.4495 $\pm$ 0.0146	-0.0005	1.15			0.99
		original		0.4498 $\pm$ 0.0156	-0.0002				0.98
0.95, 0.9, 0.01	0.05	parametric	0.42	0.4493 $\pm$ 0.0161	-0.0007	1.015	0.14 $\pm$ 0.05	1.02E-02	0.98
		non-parametric		0.4491 $\pm$ 0.0168	-0.0009	0.94			0.98
		original		0.4492 $\pm$ 0.0162	-0.0008				0.98
0.95, 0.9, 0.01	0.1	parametric	0.40	0.4498 $\pm$ 0.0165	-0.0002	1.015	0.12 $\pm$ 0.06	5.06E-02	0.98
		non-parametric		0.4498 $\pm$ 0.0168	-0.0002	0.98			0.98
		original		0.4498 $\pm$ 0.0166	-0.0002				0.98
b Threshold genetic model: Covariate: the trait with a QTL at τ									
	0.05	parametric	0.17	0.4504 $\pm$ 0.0381	0.0004	1.68	0.25 $\pm$ 0.04	2.26E-08	0.96
		non-parametric		0.4509 $\pm$ 0.0388	0.0009	1.62			0.95
		original		0.4519 $\pm$ 0.0494	0.0019				0.95
	0.1	parametric	0.27	0.4508 $\pm$ 0.0231	0.0008	1.23	0.34 $\pm$ 0.057	3.35E-09	0.96
		non-parametric		0.4507 $\pm$ 0.0233	0.0007	1.21			0.96
		original		0.4509 $\pm$ 0.0256	0.0009				0.97
	0.2	parametric	0.39	0.4505 $\pm$ 0.0152	0.0005	1.078	0.45 $\pm$ 0.09	6.56E-07	0.98
		non-parametric		0.4505 $\pm$ 0.0153	0.0005	1.060			0.98
		original		0.4502 $\pm$ 0.0157	0.0002				0.98
c Gene-environment genetic model: Covariate: the trait with a QTL at τ									
	0.05	parametric	0.21	0.4474 $\pm$ 0.0504	-0.0026	1.83	0.22 $\pm$ 0.0447	1.39E-06	0.94
		non-parametric		0.4473 $\pm$ 0.0507	-0.0027	1.81			0.92
		original		0.4505 $\pm$ 0.0682	0.0005				0.92
	0.1	parametric	0.26	0.4497 $\pm$ 0.0342	-0.0003	1.72	0.28 $\pm$ 0.0547	4.64E-07	0.95
		non-parametric		0.4498 $\pm$ 0.0342	-0.0002	1.72			0.96
		original		0.4507 $\pm$ 0.0449	0.0007				0.95
	0.2	parametric	0.26	0.4508 $\pm$ 0.0287	0.0008	1.34	0.33 $\pm$ 0.0745	9.16E-06	0.96
		non-parametric		0.4513 $\pm$ 0.0295	0.0013	1.27			0.97
		original		0.4511 $\pm$ 0.0332	0.0011				0.96
Covariate: age at onset									
	0.05	parametric	0.21	0.4487 $\pm$ 0.034	-0.0013	4.014	-0.16 $\pm$ 0.0193	4.44E-16	0.96
		non-parametric		0.4485 $\pm$ 0.0326	-0.0015	4.36			0.95
		original		0.4505 $\pm$ 0.0682	0.0005				0.92
	0.1	parametric	0.26	0.4505 $\pm$ 0.0268	0.0005	2.80	-0.18 $\pm$ 0.0219	<1.00E-16	0.96
		non-parametric		0.4506 $\pm$ 0.0263	0.0006	2.92			0.96
		original		0.4507 $\pm$ 0.0449	0.0007				0.95
	0.2	parametric	0.26	0.4498 $\pm$ 0.0229	-0.0002	2.095	-0.21 $\pm$ 0.0304	1.55E-12	0.98
		non-parametric		0.4497 $\pm$ 0.0223	-0.0003	2.21			0.97
		original		0.4511 $\pm$ 0.0332	0.0011				0.96

^a Parametric: incorporating a covariate based on the parametric model; ^b non-parametric: incorporating a covariate based on the non-parametric model; ^c original: without incorporating a covariate; ^d R.E. = relative efficiency, the ratio of variances of disease locus estimates from approaches with a covariate vs. without. R.E. >1 indicates that the proposed method is more efficient than the original method.

Table 2. Impact of marker density on estimating a disease locus using the gene-environment genetic model

No. marker	C = 0.1958	τ	Bias	R.E.	β	p value	95% coverage probability
<i>Covariate: the environmental factor in the gene-environment genetic model</i>							
10	parametric	0.4501 $\pm$ 0.0356	0.0001	1.59	-0.38 $\pm$ 0.084	4.98E-06	0.95
	non-parametric	0.4505 $\pm$ 0.0369	0.0005	1.48			0.95
	original	0.4507 $\pm$ 0.0449	0.0007				0.95
20	parametric	0.4511 $\pm$ 0.0233	0.0011	1.43	-0.38 $\pm$ 0.078	1.02E-06	0.95
	non-parametric	0.4511 $\pm$ 0.0233	0.0011	1.43			0.96
	original	0.4517 $\pm$ 0.0279	0.0017				0.97
<i>Covariate: the trait with a QTL at τ</i>							
10	parametric	0.4497 $\pm$ 0.0342	-0.0003	1.72	0.28 $\pm$ 0.055	4.64E-07	0.95
	non-parametric	0.4498 $\pm$ 0.0342	-0.0002	1.72			0.96
	original	0.4507 $\pm$ 0.0449	0.0007				0.95
20	parametric	0.451 $\pm$ 0.0226	0.0010	1.52	0.28 $\pm$ 0.051	4.71E-08	0.96
	non-parametric	0.4509 $\pm$ 0.023	0.0009	1.46			0.96
	original	0.4517 $\pm$ 0.0279	0.0017				0.97

Table 3. Impact of the environmental factor (β_3) on estimating the disease locus using the gene-environment genetic model

β_3		C	τ	Bias	R.E.	β	p value	95% coverage probability
<i>Covariate: the environmental factor in the gene-environment genetic model</i>								
ln(2)	parametric	0.3448	0.4497 ± 0.0161	-0.0003	0.98	0.031 ± 0.092	7.32E-01	0.98
	non-parametric		0.4497 ± 0.0164	-0.0003	0.94			0.99
	original		0.4497 ± 0.0159	-0.0003				0.98
ln(10)	parametric	0.2817	0.449 ± 0.0277	-0.0010	1.24	-0.36 ± 0.088	3.51E-05	0.96
	non-parametric		0.4492 ± 0.0301	-0.0008	1.050			0.96
	original		0.449 ± 0.0308	-0.0010				0.95
ln(15)	parametric	0.2558	0.4501 ± 0.0356	0.0001	1.59	-0.38 ± 0.084	4.98E-06	0.95
	non-parametric		0.4505 ± 0.0369	0.0005	1.48			0.95
	original		0.4507 ± 0.0449	0.0007				0.95
<i>Covariate: the trait with a QTL at τ</i>								
ln(2)	parametric	0.3448	0.4496 ± 0.0159	-0.0004	1.0079	0.14 ± 0.062	2.51E-02	0.99
	non-parametric		0.4496 ± 0.0161	-0.0004	0.98			0.99
	original		0.4497 ± 0.0159	-0.0003				0.98
ln(10)	parametric	0.2817	0.4486 ± 0.0278	-0.0014	1.23	0.28 ± 0.059	2.40E-06	0.95
	non-parametric		0.4487 ± 0.0294	-0.0013	1.10			0.96
	original		0.449 ± 0.0308	-0.0010				0.95
ln(15)	parametric	0.2558	0.4497 ± 0.0342	-0.0003	1.72	0.28 ± 0.055	4.64E-07	0.95
	non-parametric		0.4498 ± 0.0342	-0.0002	1.72			0.96
	original		0.4507 ± 0.0449	0.0007				0.95

Table 4. Impact of the additive genetic effect from the covariate on estimating the disease locus using the gene-environment model

Additive genetic component		τ	Bias	R.E.	β	p value	95% coverage probability
1	parametric	0.4497 ± 0.0342	-0.0003	1.72	0.28 ± 0.055	4.64E-07	0.95
	non-parametric	0.4498 ± 0.0342	-0.0002	1.72			0.96
	original	0.4507 ± 0.0449	0.0007				0.95
5	parametric	0.4505 ± 0.0251	0.0005	3.21	0.21 ± 0.021	<1.00E-16	0.97
	non-parametric	0.4506 ± 0.0255	0.0006	3.11			0.97
10	parametric	0.4506 ± 0.0246	0.0006	3.33	0.11 ± 0.0104	<1.00E-16	0.97
	non-parametric	0.4506 ± 0.0254	0.0006	3.12			0.97

A Data Example

To illustrate the proposed method, the original, parametric, and non-parametric approaches were applied to a family-based hypertension study conducted at four community hospitals in northwestern Taiwan [25]. The sample of the study included a total of 88 young-onset hypertension trios from 66 families. Forty-four SNPs of the lipoprotein lipase (LPL) gene with a minor allele frequency (MAF) of at least 0.05 in the 43 trios having the covariate plasma triglyceride (TG) were included in these analyses. ‘Young-onset’ hypertension indicates hypertensive patients first diagnosed at less than 40 years old. The unit of base pair (bp) was roughly converted into the unit of centi-Morgan (cM) by dividing the bp by 10^6 . The TG level was incorporated into the association mapping to compare the efficiency of estimating the disease locus between this approach and the original approach without a covariate. The estimated disease locus was at 19.85367 cM, near SNP rs343 (fig. 2), with a standard error of 0.002418 (95% CI = [19.84893, 19.85841]) in the original approach. The genetic effect C was 0.134 with a p value of 1.48×10^{-4} , strongly suggesting an association between young-onset hypertension and the estimated disease locus at 19.85367 cM (table 8). Incorporating the TG level into the parametric approach produced an estimate of 18.85321 (cM), near SNP rs343, with a standard error of 0.002463 (95% CI = [19.84838, 19.85804]). TG was also positively correlated with the genetic effect at the estimated disease locus in young-onset hypertensive subjects (p value = 4.6×10^{-3}). However, the efficiency of estimating τ remained similar after incorporating TG, which may be because one more regression parameter for the covariate must be estimated based on the limited sample size. In the non-parametric approach, the disease locus estimate, its standard error, and 95% CI were estimated to be 19.85682 cM (near SNP

rs258), 0.00023 and [19.85636, 19.85727], respectively (fig. 2). The relative efficiencies in estimating τ were 0.96 and 110.52, compared with the original estimate for the parametric approach and non-parametric approach, respectively. These estimated locations are consistent with the results of Chen et al. [25]. Since the non-parametric model is more robust against the one-locus assumption, the non-parametric approach obtained three additional estimates for τ with different initials: (i) 19.86076 cM, near SNP rs295, with a 95% CI = [19.86064, 19.86088]; (ii) 19.8634 cM, between SNPs rs320 and rs322, with a 95% CI = [19.86328, 19.86352], and (iii) 19.86522 cM, near SNP rs331 with a 95% CI = [19.86472, 19.86572]. Compared with the estimate that did not incorporate a covariate, the relative efficiencies in the non-parametric approach were 1530.86, 1445.43, and 88.52 for these three additional estimates, respectively. The estimated disease locus 19.8634 (between rs320 and rs322) with the smallest standard error (6.36×10^{-5}) is located at exon 8, which is known to be associated with hypertension [26]. To further study the interactions between a SNP and the disease locus, and the SNP’s relation with TG, the genotypes of rs295 at 19.860518 cM were incorporated into the parametric model. Two dummy variables indicating genotype 2/2 (wild type) and 2/1 were created. The relative efficiency of estimating the disease locus improved 1.30-fold after adding rs295 to the fine-mapping. With the incorporation of SNP rs295, TG remained significant (p = 0.041), and rs295 also remained significantly associated with the estimated disease locus (p = 7.6×10^{-13} for genotype 2/2 and p = 1.07×10^{-6} for genotype 2/1, compared to the genotype of 1/1). In addition, the interaction between the estimated disease locus and rs295 is significant (p = 2.44×10^{-6}). These results suggest that TG and rs295 are independently associated with the disease locus, and there is a gene-gene interaction between rs295 and the estimated disease locus near rs343.

Table 5. Impact of the QTL's position on estimating the disease locus using the gene-environment genetic model ($C = 0.21$)

QTL/genetic model of the covariate		τ	Bias	R.E.	β	p value	95% coverage probability
Additive							
0.45 (τ)	parametric	0.4506 ± 0.0104	0.0006	1.44	0.28 ± 0.047	5.23E-09	0.96
	non-parametric	0.4505 ± 0.0105	0.0005	1.42			0.95
	original	0.4507 ± 0.0125	0.0007				0.95
0.5 cM	parametric	0.4523 ± 0.0109	0.0023	1.32	0.23 ± 0.048	1.99E-06	0.94
	non-parametric	0.4522 ± 0.0109	0.0022	1.32			0.94
0.7 cM	parametric	0.4531 ± 0.0116	0.0031	1.16	0.17 ± 0.049	6.53E-04	0.96
	non-parametric	0.4527 ± 0.0113	0.0027	1.23			0.94
0.9 cM	parametric	0.4522 ± 0.0118	0.0022	1.11	0.12 ± 0.048	9.71E-03	0.95
	non-parametric	0.4521 ± 0.0116	0.0021	1.16			0.94
1.1 cM	parametric	0.4508 ± 0.0122	0.0008	1.04	0.095 ± 0.051	5.88E-02	0.95
	non-parametric	0.4508 ± 0.0118	0.0008	1.11			0.95
1.3 cM	parametric	0.4508 ± 0.0119	0.0008	1.099	0.068 ± 0.051	1.84E-01	0.96
	non-parametric	0.4507 ± 0.0114	0.0007	1.20			0.95
unlinked	parametric	0.4507 ± 0.0124	0.0007	1.013	-0.0003 ± 0.049	9.95E-01	0.96
	non-parametric	0.4509 ± 0.0122	0.0009	1.047			0.95
	non-parametric	0.4503 ± 0.0104	0.0003	1.43			0.96
Dominant							
0.5 cM	parametric	0.4528 ± 0.0094	0.0028	1.75	0.26 ± 0.040	2.00E-10	0.94
	non-parametric	0.4526 ± 0.0096	0.0026	1.68			0.97
0.7 cM	parametric	0.4543 ± 0.0105	0.0043	1.42	0.19 ± 0.040	1.90E-06	0.94
	non-parametric	0.4541 ± 0.0104	0.0041	1.44			0.95
Recessive							
0.5 cM	parametric	0.4516 ± 0.012	0.0016	1.076	0.098 ± 0.052	5.93E-02	0.96
	non-parametric	0.4517 ± 0.0112	0.0017	1.24			0.93
0.7 cM	parametric	0.4512 ± 0.0121	0.0012	1.061	0.067 ± 0.053	2.08E-01	0.96
	non-parametric	0.4514 ± 0.0117	0.0014	1.13			0.93
Using the environmental factor as a covariate parametric		0.4504 ± 0.0103	0.0004	1.46	-0.38 ± 0.077	1.04E-06	0.96

Table 6. Comparisons of relative efficiency of estimating the disease locus among three LD patterns (by varying $N = 50, 100, 150$)

N	C	τ	Bias	R.E.	β	p value	95% coverage probability
<i>Covariate: the trait with a QTL at τ</i>							
150	parametric	0.26	0.4497 ± 0.0342	-0.0003	0.28 ± 0.0547	4.64E-07	0.95
	non-parametric		0.4498 ± 0.0342	-0.0002			0.96
	original		0.4507 ± 0.0449	0.0007			0.95
100	parametric	0.26	0.4488 ± 0.0475	-0.0012	0.28 ± 0.053	1.74E-07	0.95
	non-parametric		0.4489 ± 0.0456	-0.0011			0.94
	original		0.4509 ± 0.0589	0.0009			0.95
50	parametric	0.26	0.4511 ± 0.0744	0.0011	0.28 ± 0.0517	6.17E-08	0.95
	non-parametric		0.4519 ± 0.0697	0.0019			0.94
	original		0.4522 ± 0.0859	0.0022			

Table 7. Performance of the proposed methods with incorporation of four covariates: one quantitative trait with a QTL at τ and three unlinked quantitative traits

	τ	Bias	R.E.	β_1 (for the quantitative trait with a QTL at τ)	p value	β_2 (for the 1st unlinked quantitative trait)	p value	β_3 (for the 2nd unlinked quantitative trait)	p value	β_4 (for the 3rd unlinked quantitative trait)	p value	95% coverage probability	Convergence rate (out of 1,000 replicates)
Parametric	0.4494 $\pm$ 0.034	-0.0006	1.75	0.276 $\pm$ 0.0554	6.3E-07	-0.0009 $\pm$ 0.0488	0.98	0.0036 $\pm$ 0.035	0.92	-0.0001 $\pm$ 0.0306	0.997	0.96	1
Non-parametric	0.4489 $\pm$ 0.0293	-0.0011	2.35									0.95	0.745
Original	0.4507 $\pm$ 0.0449	0.0007										0.95	1

Table 8. Association mapping for young-onset hypertension using the SNPs of the LPL gene

	τ	N	C	β_1 (TG)	β_2 (rs295) (2/2*)	β_3 (rs295) (2/1)	Testing $H_0 : \beta_2 = \beta_3$
Without incorporating covariates							
Est.	19.85367	5,939.331	0.133768				
S.E.	0.002418	2,927.626	0.03526				
95% CI	[19.84893, 19.85841]						
Z			3.7938				
p value			0.000148				
Incorporating TG based on the parametric model							
Est.	19.85321	7,093.473	0.146494	0.00175			
S.E.	0.002463	3,124.493		0.000618			
95% CI	[19.84838, 19.85804]						
Z				2.832174			
p value				0.004623			
Incorporating TG and rs295 based on the parametric model							
Est.	19.85325	5,736.296	0.135005	0.000731	-1.13849	-0.59154	
S.E.	0.002156	2,024.27		0.000358	0.158826	0.121277	
95% CI	[19.84902, 19.85748]						
Z					-7.16816	-4.87759	-4.7131
p value					7.6E-13	1.07E-06	2.44E-06
Incorporating TG based on the non-parametric model							
Est.	19.85682	37,411.91	0.390251				
S.E.	0.00023	8,218.808					
95% CI	[19.85636, 19.85727]						
Est.	19.86076	312,199	0.608378				
S.E.	6.18E-05	77,263.94					
95% CI	[19.86064, 19.86088]						
Est.	19.8634	163,886.4	0.435007				
S.E.	6.36E-05	42,583.58					
95% CI	[19.86328, 19.86352]						
Est.	19.86522	51,429.97	0.504052				
S.E.	0.000257	9,713.551					
95% CI	[19.86472, 19.86572]						
* Allele 2 for rs295 is the major allele.							

Fig. 2. Empirical values for the expected preferential-transmission statistics for the 44 SNPs of the LPL gene from the family study of hypertension. The solid line (—) denotes the empirical $E[Y(t)|\Phi]$; the dotted line (·····) denotes the empirical $E[Y(t)|\Phi]/Pr[h(t)h(\tau)]$. The estimates of the disease locus and their 95% CIs are denoted by 'x' and brackets, respectively, on the x-axis.

Discussion

The case-parent design is robust against genetic population structures. However, spurious gene-environment interactions may still result when trios are sampled from subpopulations with different alleles and exposure frequencies [6]. The proposed approaches allow investigators to account for population stratification or heterogeneity by incorporating covariates. Researchers can also assess gene-environment interactions while identifying the disease locus in the case-parent design.

Through the key representation stated in equation (5), the GEE approach allows investigators to use multiple markers from the same region to estimate the location of the trait locus, τ , and the genetic effect in C. Thus, one could test the null hypothesis of no linkage and association in the region by testing whether $C = 0$ without the need to test the same hypothesis one marker at a time and hence avoids the multiple testing issue, at least within the

region considered. Multiple markers were incorporated into the fine-mapping through the GEE approach, effectively eliminating the concern of multiple testing in the chromosomal region of interest. Another advantage of the GEE approach is that it provides valid standard error estimates when using multiple trios from the same family [13]. In addition, this method is robust to the genetic models as the only assumption it makes is that there is only a susceptible locus located in this region. The data example in the present study shows that the non-parametric approach is robust to this assumption, which may be due to local peaks and obtaining the estimate for the disease locus locally in the non-parametric approach. Simulation study and data example both demonstrate that incorporating covariates into fine-mapping improves the efficiency of estimating the disease locus and assessing gene-gene and gene-environment interactions. By comparing the results of incorporating different covariates, this approach can help investigators select a

proper environmental factor or biomarker to mediate or moderate the function of the disease gene. In addition, comparisons of the disease locus estimates from the original and proposed parametric and non-parametric approaches help investigators confirm the estimated disease locus and elucidate the association between the covariates and the disease.

In the simulation study, incorporating a covariate unlinked to the disease locus does not reduce the efficiency of estimating the disease locus in either the parametric or non-parametric models, compared to the estimate without the covariate. The efficiency remained similar when incorporating three additional irrelevant covariates (four covariates in total) into the fine-mapping. This underlines the utility and robustness of both the parametric and non-parametric approaches. However, although the number of covariates is arbitrary, one needs to be cautious – if the sample size is not large enough, non-convergence problems may appear when the total number of covariates is too big, in the non-parametric approach, in particular. The efficiency in estimating τ is comparable between parametric and non-parametric approaches, although, when C is small, the non-parametric approach tends to have slightly higher efficiency than the parametric approach.

This simulation study and its examples also illustrate that the proposed method is helpful for identifying a covariate, which improves the efficiency of estimating the disease locus. In the young-onset hypertension example, in addition to confirming previous results [25], the proposed method assesses gene-gene interactions between the estimated disease locus and SNP rs295. These findings can help investigators understand the underlying genetic mechanisms of a disease.

The approach described in the present study requires the availability of a dataset consisting of parent-offspring trios with no missing genotype data. A natural extension of the proposed method would be to apply it to nuclear families [27], extended pedigrees, or missing genotype data. For missing genotype data, single and/or multiple imputation approaches or averaging all the possible genotype configurations consistent with the observed genotype data could be solutions [28].

In the situation when a disease is controlled by more than one unlinked loci, the information from some regions can be incorporated to estimate the marginal effect of a disease locus in other regions. The example of young-onset hypertension above shows that the proposed non-parametric model is more robust to the one-locus assumption, which is a useful feature for the analysis of ge-

nome-wide association (GWA) data. Further studies should examine the robustness of these two models against this one-locus assumption. The potential extension of this approach to GWA data will allow researchers to incorporate risk factors and gene-gene and gene-environment interactions into GWA analysis through a gene-based or sliding window approach. This may help a great deal in identifying disease loci and dissecting the etiology of complex diseases.

Acknowledgments

The authors would like to express their gratitude to Ms. Hui-Yi Kao and Mr. Yu-Wei Lee for their computing assistance and to Dr. Chia-Min Chung for his assistance with preparing the young-onset hypertension data set. This project is supported by a grant from the National Science Council, Taiwan (NSC97-2311-B-003), a grant from the National Health Research Institutes, Taiwan (PH-098-pp-04) and a grant from the National Institutes of Health, USA (HL090577).

Appendix

The estimates for the vector of parameters δ in this method were obtained by solving the estimating equation [13], where $\delta = (\tau, N, \alpha, \beta_1, \dots, \beta_p)$ in the parametric modeling, and $\delta = (\tau, N)$ in the non-parametric modeling [17, 20]:

$$S(\delta) = \sum_{i=1}^n \left[\frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \text{Cov}^{-1}(Y_i) \{Y_i - \mu(\delta, \hat{\pi})\} + \frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \text{Cov}^{-1}(X_i) \{X_i - \mu(\delta, \hat{\pi})\} \right] = 0, \quad (23)$$

where

$$\begin{aligned} Y_i &= [Y_{i1}(t_1) - Y_{i2}(t_1), \dots, Y_{i1}(t_M) - Y_{i2}(t_M)]', \\ X_i &= [X_{i1}(t_1) - X_{i2}(t_1), \dots, X_{i1}(t_M) - X_{i2}(t_M)]', \end{aligned} \quad (24)$$

and

$$\mu(\delta, \hat{\pi}) = [\mu(t_1; \delta, \hat{\pi}_1), \dots, \mu(t_M; \delta, \hat{\pi}_M)]'. \quad (25)$$

The vector of parameter estimates $\hat{\delta}$ are consistent estimates of δ with the normality asymptotic property. Based on the asymptotic property, the variance estimates of $\hat{\delta}$ can be calculated as

$$\widehat{\text{Var}}(\hat{\delta}) = A^{-1} B A^{-1} \quad (26)$$

where

$$\begin{aligned} A &= \sum_{i=1}^n \left[\left(\frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \right)' \text{Cov}^{-1}(Y_i) \left(\frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \right) \right. \\ &\quad \left. + \left(\frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \right)' \text{Cov}^{-1}(X_i) \left(\frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \right) \right] \Bigg|_{\delta = \hat{\delta}} \end{aligned} \quad (27)$$

and

$$B = \sum_{i=1}^n \left[\left(\frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \right)' \text{Cov}^{-1}(Y_i) \{Y_i - \mu(\delta, \hat{\pi})\} \{Y_i - \mu(\delta, \hat{\pi})\}' \text{Cov}^{-1}(Y_i) \left(\frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \right) + \left(\frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \right)' \text{Cov}^{-1}(X_i) \{X_i - \mu(\delta, \hat{\pi})\} \{X_i - \mu(\delta, \hat{\pi})\}' \text{Cov}^{-1}(X_i) \left(\frac{\partial \mu(\delta, \hat{\pi})}{\partial \delta} \right) \right] \Bigg|_{\delta = \hat{\delta}} \quad (28)$$

This approach makes it possible to make inferences for the parameters of interest. In addition, it is possible to test whether the covariates z_i of allele transmission are significant by testing the null hypothesis: $\beta_i = 0$, $i = 1, \dots, p$.

Following the approach in Liang et al. [13], one minor modification is necessary when applying the GEE method, since the $\mu(\delta, \hat{\pi})$ is not differentiable with respect to τ (strictly speaking) through $|t - \tau|$ in the Haldane [16] mapping function. This concern can be addressed by replacing $|t - \tau|$ with

$$\begin{cases} |t - \tau| & \text{if } |t - \tau| \geq \varepsilon \\ \frac{1}{2\varepsilon}(t - \tau)^2 + \frac{1}{2}\varepsilon & \text{if } |t - \tau| < \varepsilon \end{cases}, \quad (29)$$

where ε is a small positive number. In the simulation study reported in the present study, the selection of ε has little impact on the estimates.

References

- Lee PJ, Ridout D, Walter JH, Cockburn F: Maternal phenylketonuria: Report from the United Kingdom registry 1978–97. *Arch Dis Child* 2005;90:143–146.
- Zeegers M, Rijdsdijk F, Sham P: Adjusting for covariates in variance components qtl linkage analysis. *Behav Genet* 2004;34:127–133.
- Weinberg CR, Wilcox AJ, Lie RT: A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet* 1998;62:969–978.
- Wilcox AJ, Weinberg CR, Lie RT: Distinguishing the effects of maternal and offspring genes through studies of ‘case-parent triads’. *Am J Epidemiol* 1998;148:893–901.
- Schaid DJ: Case-parents design for gene-environment interaction. *Genet Epidemiol* 1999;16:261–273.
- Umbach DM, Weinberg CR: The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Genet* 2000;66:251–261.
- Lim S, Beyene J, Greenwood CM: Continuous covariates in genetic association studies of case-parent triads: gene and gene-environment interaction effects, population stratification, and power analysis. *Stat Appl Genet Mol Biol* 2005;4:Article20.
- Baksh MF, Balding DJ, Vyse TJ, Whittaker JC: A likelihood ratio approach to family-based association studies with covariates. *Ann Hum Genet* 2006;70:131–139.
- Lunetta KL, Faraone SV, Biederman J, Laird NM: Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am J Hum Genet* 2000;66:605–614.
- Liu Y, Tritchler D, Bull SB: A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits. *Genet Epidemiol* 2002;22:26–40.
- Cordell HJ, Barratt BJ, Clayton DG: Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet Epidemiol* 2004;26:167–185.
- Rabinowitz D, Laird N: A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered* 2000;50:211–223.
- Liang KY, Hsu FC, Beaty TH, Barnes KC: Multipoint linkage-disequilibrium-mapping approach based on the case-parent trio design. *Am J Hum Genet* 2001;68:937–950.
- Hsu FC, Liang KY, Beaty TH: Multipoint linkage disequilibrium mapping approach: incorporating evidence of linkage and linkage disequilibrium from unlinked region. *Genet Epidemiol* 2003;25:1–13.
- Chiu YF, Liang KY, Chuang LM, Beaty TH: Incorporation of covariates into multipoint linkage disequilibrium mapping in case-control studies. *Genet Epidemiol* 2008;32:143–151.
- Haldane JBS: The combination of linkage values and the calculation of distances between the loci of linked factors. *J genetics* 1919;8:299–309.
- Liang KY, Zeger SL: Longitudinal data analysis using generalized linear models *Biometrika* 1986;73:13–22.
- Chiou JM, Liang KY, Chiu YF: Multipoint linkage mapping using sibpairs: non-parametric estimation of trait effects with quantitative covariates. *Genet Epidemiol* 2005;28:58–69.
- Fan J, Gijbels I: Local polynomial modelling and its applications. London: Chapman and Hall, 1996.
- Ruppert D, Wand MP: Multivariate locally weighted least squares regression. *The Annals of Statistics* 1994;22:1346–1370.
- Haseman JK, Elston RC: The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 1972;2:3–19.
- Schmidt S, Qin X, Schmidt MA, Martin ER, Hauser ER: Interpreting analyses of continuous covariates in affected sibling pair linkage studies. *Genet Epidemiol* 2007;31:541–552.
- Li H, Hsu L: Effects of age at onset on the power of the affected sib pair and transmission/disequilibrium tests. *Ann Hum Genet* 2000;64:239–254.
- Liang KY, Chiu YF: Multipoint linkage disequilibrium mapping using case-control designs. *Genet Epidemiol* 2005;29:365–376.
- Chen P, Jou YS, Fann CS, Chen JW, Chung CM, Lin CY, Wu SY, Kang MJ, Chen YC, Jong YS, Lo HM, Kang CS, Chen CC, Chang HC, Huang NK, Wu YL, Pan WH: Lipoprotein lipase variants associated with an endophenotype of hypertension: hypertension combined with elevated triglycerides. *Hum Mutat* 2009;30:49–55.
- Chen P, Jou YS, Fann CS, Chen JW, Wu SY, Pan WH: Lipoprotein lipase gene is linked and associated with hypertension in Taiwan young-onset hypertension genetic study. *J Biomed Sci* 2005;12:651–658.
- Chen YH, Lin HW: Simple association analysis combining data from trios/sibships and unrelated controls. *Genet Epidemiol* 2008;32:520–527.
- Cordell HJ: Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. *Genet Epidemiol* 2006;30:259–275.