國立交通大學

統計學研究所

碩士論文

加入共變數於三元體資料分析

以估計疾病基因位置

Incorporating Covariates into Linkage-Disequilibrium Mapping Using the Case-Parent Trio Design

研 究 生：李昱緯

指導教授：邱燕楓　博士

中華民國九十七年六月

加入共變數於三元體資料分析
以估計疾病基因位置

Incorporating Covariates into Linkage-Disequilibrium Mapping Using the Case-Parent Trio Design

研究生：李昱緯　　Student: Yu-Wei Lee

指導教授：邱燕楓　　Advisor: Dr. Yen-Feng Chiu

國立交通大學

統計學研究所

碩士論文

A Thesis
Submitted to institute of Statistics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics
June 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年六月

加入共變數於三元體資料分析
以估計疾病基因位置

研究生：李昱緯　　　　指導教授：邱燕楓　博士

國立交通大學統計學研究所

摘要

Case-parent trio design 常被用在遺傳流行病學研究中，相較於其他的傳統方法，例如：affected-sib-pair (ASP) sign，case-parent trio design 更適合應用在罕見疾病。Liang 等人在 2001 年時根據 case-parent trio design 提出一種疾病基因相關定位的方法，他們利用偏好傳遞統計量(expected preferential-transmission statistic)估計疾病基因的位置。相較於傳統的 TDT 方法，他們指出，這個方法不但較有效力，且可以應用於更廣泛的資料。此外，除了利用假設檢定去尋找疾病基因的位置，這個方法還能對疾病基因位點，提供準確的估計值及其相對應的標準差，以對這疾病位置作推論。因為許多複雜的疾病是由基因和環境因素的交互作用所造成，因此，加入這些基因或環境因素於三元體的資料分析，應能對疾病基因位點，做更精準的定位。在本研究中，我們用 case-parent trios 的資料，分別利用有母數和無母數的方法，將相關的共變數併入模型中，以幫助我們估計疾病基因的位置。模擬結果和兔唇資料分析均顯示，估計疾病基因位置時，加入共變數，會使得估計值更有效率。

關鍵詞：多點檢定；連鎖不平衡；三元體資料；連續型變數；有母數方法；無母數方法。

Incorporating Covariates into Linkage-Disequilibrium Mapping Using the Case-Parent Trio Design

Student: Yu-Wei Lee   Advisor: Dr. Yen-Feng Chiu
Institute of Statistics
National Chiao Tung University

ABSTRACT

Case-parant trio design is commonly used in genetic epidemiological family studies. It is more suitable for rare disorders than other conventional designs for family studies, such as affected-sib-pair (ASP) designs. Liang et al. (2001b) proposed a multipoint linkage disequilibrium (LD) mapping approach to localize disease genes based on a preferential-transmission statistic in the case-parent trio design. They found that their approach was more powerful and could accommodate a wider variety of data than the conventional TDT approach. In addition, instead of conducting hypothesis testing to search for a disease locus, it provided a precise estimate for a postulated disease locus along with its standard error, so that one can make inference for the disease locus. Most complex diseases involve both genetic and environmental components, incorporating genetic or environmental factors into the LD mapping may be helpful in localizing the disease locus. We therefore incorporated trait-related covariates into the LD mapping to estimate the disease locus through parametric and nonparametric models in the case-parent trio design in the present study. Simulation studies and the example of oral cleft study both suggested that incorporating covariates into the LD mapping approach helps a great deal to improve efficiency in localizing the disease locus.

Key words: Multipoint; Linkage disequilibrium; Case-parent trio design; Covariates; Parametric approaches; Nonparametric approaches.

誌謝

從口試委員的口中聽到了「恭喜」這兩個字時，內心的激動是不可言喻的。回想起開始做論文時，對自己的能力總是存在著一些不確定感，所幸這一年來，謝謝我的指導教授邱燕楓老師不僅一步步地引導我進入基因統計這門陌生的領域，在每週一次的報告中也不厭其煩地解決我的疑惑、指正錯誤的觀念。雖然我不能算是一個很認真的學生，但是您從不吝嗇在別人面前稱讚我，而您給我的鼓勵也總是多過於責難，因為您認真嚴謹的要求才能讓這篇論文順利地誕生。

感謝所上的教授對我們的指導與照顧，讓我們在交大統研的這兩年內學到了許多有用的知識；感謝最美麗的所辦小姐郭姐，常常在忙碌的工作之餘，還得傾聽我們的苦水，認真的工作態度讓我們省去許多時間去處理學校生活上的瑣碎事。當然也要感謝陪伴我這兩年生涯的好夥伴們，謝謝小爸不時給予我適時的安慰與鼓勵，謝謝小胖廷常常傾聽我的抱怨給予我適當的建議，謝謝嗡嗡在我忙碌之餘提醒我該吃飯了，謝謝小老婆提供許多在研究室的娛樂活動，謝謝重耕和彥銘在我需要幫忙的時候毫不猶豫地伸出援手，謝謝小螞蟻經常幫我解決程式方面的問題，謝謝小樹林帶給研究室許多歡樂，謝謝班代這兩年的辛苦，同時也謝謝邰嵐、香菱、佩芳、姿蒨和鈺婷妳們的陪伴和照顧，特別感謝每天陪伴我超過十二小時的許小雞總是不辭辛勞、不計代價地擔任我的司機和顧問。另外我想感謝的是國衛院的徽宜學姊、君儀學姊和素梅學姊，妳們不只是幫助我解決論文和程式方面的問題，對我的關心也讓我感到溫馨，能認識妳們是我進國衛院除了論文以外最大的收穫。謝謝各階段和各方面的親朋好友，你們對我的鼓勵和祝福是我完成論文最大的動力。此外，謝謝黃冠華老師、陳君厚老師和楊欣洲老師，感謝您們不僅抽空擔任我的口試委員，並且在口試過程中給予我寶貴的意見和鼓勵，讓我能更深入地去了解這領域的內涵和發展。

最後謹以此篇論文獻給我最愛的家人們，雖然奶奶沒辦法親眼看到我穿上碩士服的那一刻，但我相信在天國的您還是會像以前一樣以我這孫子為榮。親愛的爸媽、大姐佩芳、二姐衣樺和輝賓大哥，謝謝你們幫我解決許多生活上的雜事，讓我可以毫無顧慮地專注在我的課業上，而且在我煩悶之餘帶我去郊外吹風散心。也謝謝小舅和小舅媽對我的照顧與關心，因為你們的鼓勵我才能更有信心地往自己的未來邁進。

李昱緯
于交通大學統計所
民國九十七年七月二日

Content

The List of Tables

The List of Figures

1. Introduction

Case-parent trio design is commonly used in present genetic epidemiology. It is more suitable for rare disorders than other conventional designs, such as affected-sib-pair (ASP) designs. In addition, the trio design does not require multiplex siblings needed in ASP designs. For trio data, the method named Transmission/disequilibrium test (TDT) (Spielman et al. 1993) was proposed to detect linkage when a disease-susceptibility locus is found to be associated with a marker in family triads, including two parents and one affected child. Risch and Merikangas (1996) proved that TDT is more powerful statistically to test genes of modest effect than ASP designs, even in the presence of population stratifications.

Many extended methods of TDT were proposed in recent year to deal with more complex situations. For example, (i) TDT without parents marker data—Sib-ship disequilibrium test (SDT) (Horvath and Laird, 1998) and Sib transmission/disequilibrium (S-TDT) (Spielman et al. 1998). These two methods exploited one or more unaffected siblings' marker data instead of parents' marker data that may be absent. The defect is that these methods are not as powerful as TDT, so they are only adaptable when lacking parents data; (ii) TDT with pedigree data—pedigree disequilibrium test (PDT) (Martin et al. 2000) can catch extra information from general pedigrees out of original trio data regardless of their size and obtains a valid TDT even when there is misclassification of unaffected individuals, especially with a high-prevalence model; (iii) TDT with multi-allele markers. (Bickeböller and Clerget-Darpoux 1995, Sham and Curtis 1995; Terwilliger 1995; Schaid 1996; Spielman and Ewens 1996; Cleves et al. 1997; Kaplan et al. 1997; Lazzeroni and Landge 1998), Sham and Curtis (1995) proposed an extension of transmission/disequilibrium test for dealing with multi-allele problem, but the approach has good power only when linkage disequilibrium is strong and the disease is recessive. On the other hand, Spielman and Ewens (1996) also revised their biallelic TDT to muiltiallelic TDT; and (iv) TDT with multiple markers (Terwilliger 1995; Lazzeroni and Landge 1998; Clayton and Jones 1999; Clayton 1999; Dudbridge et al. 2000). Zhao et al. (2000) also proposed a new approach about multiple markers and corrected the disadvantage of prior approaches. (E.g. Lazzeroni and Landge's approach ignores the dependence of marker, Clayton's approach is not robust to population stratification, and for Dudbridge's approach, ambiguous haplotypes have to be discarded.) In solving the problem of unknown haplotype frequency, it is important and bounden to know the information of parents' genotype. Besides, although haplotype with multimarker is more informative than single marker, it also results in a larger number of degrees of freedom and reduces the power of these tests simultaneously. The new approach-- Haplotype-sharing TDT (HS-TDT) (Zhang et al. 2003), not only remains informative as traditional haplotype-based tests, but decreases the degrees of freedom. HS-TDT is applicable to both qualitative and quantitative traits, arbitrary size of nuclear family with or without ambiguous phase information, and whatever number of alleles at each marker. However, Knapp et al. (2004) declared that if the genotyping error exists, even the probability of genotyping errors is low, HS-TDT cannot have a precise type I error.

Although the original TDT was powerful and robust, it could not include the informative trait or covariate. In earlier research, Haseman and Elston (1972) used sib pairs' data, not trios' data we required in TDT, to estimate linkage between a known marker with $\boldsymbol{m}$ alleles and a susceptibility disease locus which governs a quantitative trait with biallelic genotype. Many other researchers developed a lot of extended methods for dealing with quantitative-covariate with IBD (e.g. Sham et al. 2002). Recently, some researches devoted on connecting TDT and a quantitative or qualitative covariate and then proposed some effective tests (Allison 1997; Abecasis et al. 2000, 2002; Liang et al. 2001; Wheeler and Cordell, 2007), such as QTDT (Rabinowitz 1997; Lunetta et al. 2000). QTDT makes use of quantitative phenotype as a dependent variable, which improved and redefined quantified genotype as an independent variable to generate linear regression. In addition, Hierarchical QTDT (HQTDT) (Fulker et al., 1999) separates genotype (independent variable) by different mating-type-- QTDT $_M$ (Gauderman, 2003) utilized the information of mating-type instead of the intercept of original regression model, and in Retrospective QTDT (RQTDT), the genotype is modeled as a function according to their phenotype and the parental genotypes (Liu et al. 2002). Gauderman (2003) employed above tests to detect three effect, genetic main effect, gene-environment interaction effect, and gene-gene interaction effect. After that, he found QTDT $_M$ is more efficient (i.e. required less sample size) than other tests under the necessary condition that the all genotypes of markers of trios data should be known, but it is not realistic.

In multipoint linkage analysis using affected sib pairs, Liang et al. (2001a) capitalized upon IBD information of multiple markers around a susceptibility gene and then obtained a simple formula between the expected numbers of allele-sharing of these markers and the susceptibility gene by careful assumption and complicated calculation. According to the formula, they applied generalized estimating equation (GEE) method (Liang and Zeger 1986) to estimate all parameters (including the disease location τ) in the model and variances of the estimates at the same time. The parameter C represents estimated expected number of allele-sharing of τ , and the range of value is from -1 to 1. The magnitude (absolute value) of parameter C in their method indicates the ability of estimating the true location of susceptibility gene. The advantage of this approach is that it did not require specification of penetrance or a mode of inheritance.

On the side, based on the conception of TDT, Liang et al. (2001b) also used allele-transmitted information of trio data instead of allele-sharing information of sibling data, and rewrote the formula between information of markers and the parameter C for case-parent trios data. In the traditional TDT method, only heterozygous parent data are informative and could be included, but in Liang et al.'s model, homozygous parent data could be recruited simultaneously. Furthermore, they could test if there is linkage or linkage disequilibrium between a disease gene and multiple genetic markers over the region at the same time, which is not like the conventional TDT where each marker－ is tested separately resulting in a multiple testing problem. Specially, the method is not restricted to trio designs only, it can also be extended for other types of data. On the other hand, the approach of Liang et al. is usually more powerful than the traditional TDT approach (Liang et al. 2001b).

Glidden et al. (2003) quoted Liang's formula for ASP designs (2001a) and added age-at-onset information as a covariate to support the estimation of parameter C. The information of covariates can yield substantial efficiency gains on finding the location of susceptibility gene. Chiou et al. (2005) also adopted Liang's formula in ASP designs (2001a), they utilized nonparametric approach to model and estimate C as a function of covariates at first, and then applied the GEE method to estimate the location of τ. By an iterative process, the estimation of C and τ could be obtained until convergence was reached. According to Chiou et al. (2005), the nonparametric method is better than the quadratic and linear models, because the nonparametric method avoids the flaw of using misspecified parametric regression models. Under case-parent trio designs, we propose a new multipoint approach for estimating the location of a susceptibility gene, τ. The proposed approach is based on transmission information of markers near an unobserved disease gene and a quantitative or a qualitative covariate associated with the disease gene. We model C as a function of covariates through parametric and nonparametric approaches, so as to incorporating covariates into the association mapping in estimating the location of susceptibility locus τ .

2. Literature Review

2.1 Transmission/Disequilibrium Test (TDT)

The transmission/disequilibrium test (TDT) can be utilized if a heterozygous parent transmits his or her target allele and alternative allele to affected child with equal frequency. It only requires affected children of trio data rather than multiple affected or unaffected family members. Besides, it detects the linkage between susceptibility gene and marker locus when association is present.

Consider two bi-allelic (a target allele D_1 and a normal allele D_2) markers M_1 and M_2, and suppose there have n trio families which have two parents and an affected child. After collecting this type of data, researchers arrange $2n$ parents of trio data into a 2×2 table shown in the Table 2 in Spielman et al. 1993.

	Nontransmitted Allele		
Transmitted Allele	M_1	M_2	Total
M_1	a	b	$a+b$
M_2	c	d	$c+d$
Total	$a+c$	$b+d$	$2n$

The above table shows every parent's genotype and the alleles which he or she transmits and does not transmit to affected child. Then, they assume a coefficient δ represents linkage disequilibrium (- freq(M_1D_1)− mp , m and p are the population frequency of allele M_1 and D_1), and θ represents the recombination fraction between marker M and locus D. With these coefficients, the Table can be rewritten as the Table 3 in Spielman et al. 1993:

	Nontransmitted Allele		
Transmitted Allele	M_1	M_2	Total
M_1	$m^2+(m\delta/p)$	$m(1-m)+[(1-\theta-m)\delta/p]$	$m+[(1-\theta)\delta/p]$
M_2	$m(1-m)+[(\theta-m)\delta/p]$	$(1-m)^2+[(1-m)\delta/p]$	$1-m-[(1-\theta)\delta/p]$
Total	$m+(\theta\delta/p)$	$1-m-(\theta\delta/p)$	1

The null hypothesis is that there is no linkage ($\theta = 1/2$), it also represents $E(b) = E(c)$ whatever the value of m and p, but the necessary condition is that the value, δ, should not be zero. On the other word, a heterozygous parent transmits target allele and normal allele with equal frequency. Under the null hypothesis, we suppose $\boldsymbol{b}$ is distributed in binomial distribution with b+c sample sizes, which are the total numbers of heterozygous parents, and the probability is 1/2.

$$\boldsymbol{b \sim Binomial(b+c,\frac{1}{2}) \Rightarrow E(b)=\frac{b+c}{2}, Var(b)=\frac{b+c}{4}} \text{ .}$$

Under this hypothesis, the χ^2 statistic has the form (McNemar's test, Sokal and Rohlf, 1969)

$$\chi^2=\left((b-\frac{b+c}{2})\Big/\sqrt{\frac{b+c}{4}}\right)^2=\frac{(b-c)^2}{b+c}.$$

The TDT is often more powerful than other conventional linkage tests and it is not affected by population structure which can lead association in the absence of linkage, since it exploits within-family comparisons only. Although TDT is much more sensitive than traditional haplotype sharing test (Risch and Merikangas 1996), and only requires a single affected child, it should be utilized under the existence of population association, even the linkage is strong.

2.2 Extension of TDT from one marker to multiple markers

Since the Human Genome Project is progressing rapidly, the genetic marker can be identified and genotyped easily and that can help us to acquire more information. After the information of multiple markers is obtained easily, many researchers proposed relative tests. We will introduce some existing and known methods below.

Lazzeroni and Lange (1998) analyzed each marker separately and obtained the adjusted $\boldsymbol{P}$-value which is the minimum of $\boldsymbol{P}$-values under the null hypothesis that there is no linkage between the region over each markers, but it ignored the dependence which may result in linkage between markers.

Some researchers use the haplotype instead of the information of multiple markers, and assume the haplotype of parents and affected child are known. Clayton

(1999) estimated the frequency of haplotype and calculated the likelihood after considering all possible solutions, but it is not robust when population stratification is present. Dudbridge et al. (2000) proposed an unbiased TDT for individual haplotype, they calculated the correct variance of the transmission count within family, and used extra information from multiple siblings if they are available. Similar to Clayton's work, they utilized missing data techniques to estimate the uncertain haplotype, so this method is also not robust when population stratification is present. To avoid this kind of problem, some family data with equivocal haplotype should be discarded, but it discards a part of information simultaneously.

Under knowing all haplotype information of each parent, Zhao et al. (2000) displayed a $\boldsymbol{h \times h}$ transmission/nontransmission table T as

$$\begin{array}{c|cccc} & \mathbf{1} & \mathbf{2} & \cdots & \boldsymbol{h} \\ \hline \mathbf{1} & \boldsymbol{t}_{11} & \boldsymbol{t}_{12} & \cdots & \boldsymbol{t}_{1h} \\ \mathbf{2} & \boldsymbol{t}_{21} & \boldsymbol{t}_{22} & \cdots & \boldsymbol{t}_{2h} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{h} & \boldsymbol{t}_{h1} & \boldsymbol{t}_{h2} & \cdots & \boldsymbol{t}_{hh} \end{array},$$

where t_{ij} is the number of parents with haplotypes $H_i H_j$ and they transmit H_γ to the affected child but not transmit H_δ, where $\boldsymbol{h}$ is the total number of possible haplotype. After completing this table, they can calculate a statistic:

$$\boldsymbol{T} = \frac{\boldsymbol{h}-\mathbf{1}}{\boldsymbol{h}} \sum_{\gamma=1}^{h} \frac{\left(\boldsymbol{t}_{\gamma.} - \boldsymbol{t}_{.\gamma}\right)^2}{\boldsymbol{t}_{\gamma.} + \boldsymbol{t}_{.\gamma} - 2\boldsymbol{t}_{\gamma\gamma}}.$$

The statistic is a marginal homogeneity test, since it may not approximate a χ^2 distribution with $\boldsymbol{h}-\mathbf{1}$ degree of freedom, we can use simulation methods to assess the $\boldsymbol{P}$ -value. With ambiguous parents' haplotype, they detected $\boldsymbol{T}_{ml}$ (estimating haplotype frequency by assuming that parents are random samples of individuals from population under Hardy-Weinberg equilibrium), which has the highest power than $\boldsymbol{T}_u$ (estimating haplotype frequency by making use of unambiguous families) and $\boldsymbol{T}_c$

(estimating haplotype frequencies by making use of both unambiguous families and ambiguous families, and assigning each compatible haplotype group equal probability for each ambiguous family). Furthermore, testing each marker separately and discarding the ambiguous families have lower power.

Although the approach using multiple markers is more informative than using a single marker, there exists some difficulties. For example, if we consider each haplotype as an allele in TDT, the degree of freedom will increase rapidly according to the number of markers and then result in lower power. On the other hand, the haplotype of parents are not always unequivocal. Zhang et al. (2003) proposed a haplotype-sharing TDT (HS-TDT) which utilized the similarity of haplotype as the information. Let $S_{H_i,H_j}(l)$ be the distance between the leftmost and the rightmost markers with identical alleles l (See figure 1 of Zhang et al.). For any haplotype H, the score of l th marker is defined as

$$X_H(l)=\frac{1}{4n}\sum_{i=1}^{n}\sum_{j=1}^{4}S_{H,H_{ij}}(l),$$

where H_{ij} is four kinds of parental haplotypes in the i^{th} family. Then let $x_{ik}=\sum_{j=1}^{4}\xi_{ijk}X_{ij}(l)$ be the difference of the haplotype-sharing score between the transmitted parental haplotypes and non-transmitted parental haplotypes, ξ_{ijk} = 1 means the haplotype $X_{ij}(l)$ transmitted to k^{th} child and ξ_{ijk} = -1 means the haplotype $X_{ij}(l)$ is not transmitted to k^{th} child. They estimated the covariance between the value of trait y_{ik} (for the qualitative case, y_{ik} = 1 means the child is affected, and y_{ik} = 0 means the child is not affected, for the quantitative case, y_{ik} can represent the quantitative value directly.) and the transmitted score x_{ik},

$$U_i(l)=\sum_{k=1}^{t_i}(y_{ik}-c)x_{ik}(l),$$

where c can be arbitrary constant, Zhang et al. (2003) set it as the average of trait value over all children. Under the null hypothesis of no linkage and association, $E\left[U_i(l)\right]$ is equal to zero for any value of c. We can find that if the disease mutation causes high trait value, the value of $U_i(l)$ should be positive. Similarly, if the disease mutation causes low trait value, the value of $U_i(l)$ should be negative. Let $U(l)=\sum_{i=1}^{n} w_i U_i(l)$, where $w_i > 0$ is a weight function over each family and the statistic of HS-TDT is defined by

$$U = \max_{1 \le l \le L} \left| U(l) \right|,$$

where L is the total number of markers. It is noticeable that the choice of c and w_i will influence the power of test. Finally, they utilized the permutation procedure to evaluate the P-value of test.

HS-TDT is applicable to both qualitative and quantitative traits, it decreases the degree of freedom with traditional haplotypes method, it has correct false-positive error rate, and it is more powerful than single-marker TDTs and haplotype-based TDTs.

2.3 Extension of TDT from bi-alleles marker to multiallele marker

A biallelic marker is assumed under traditional TDT method, but sometimes many markers over chromosome of human have more than two alleles, such as blood type which has A, B, and O, three alleles basically. So when TDT is introduced and popular over the world, some researchers devoted to extending TDT to multiallele marker. The original approach, generalized TDT (Bickeböller and Clerget-Darpoux, 1995), is to combine HHRR (haplotype-based haplotype relative risk) statistic (Terwilliger and Ott, 1992) and TDT:

$$T_c = \sum_{i<j} \frac{\left(t_{ij} - t_{ji}\right)^2}{t_{ij} + t_{ji}},$$

where t_{ij} is the number of parents who transmitted allele i and not transmitted allele j. The statistic has asymptotically a χ^2 distribution with $m(m-1)/2$ D.F. under the absence of linkage. But under the null hypothesis ($\theta = 1/2$) and the presence of linkage disequilibrium, the statistic is invalid and has lower power, since the transmitted and non-transmitted allele are not independent. In addition to the test T_c described above, they also proposed another statistic,

$$T_m = \sum_{i=1}^{t} \frac{\left(t_{i\cdot} - t_{\cdot i}\right)^2}{\left(t_{i\cdot} + t_{\cdot i}\right)},$$

where $t_{i\cdot}$ and $t_{\cdot i}$ are the row and column marginal totals. The statistic is an extension of the discussion of Ewens and Spielman (1995) for biallelic markers.

Sham and Curtis (1995) proposed an extended method of TDT. First, they calculated the probability (P_{ij}) of each type of transmitted and non-transmitted alleles conditional on parental genotype. Under $\theta = 0$, $\ln(P_{ij}/P_{ji}) = b_i - b_j$, so there are $m-1$ independent parameter b_i which related to the marker alleles M_i. For convenience, b_m is set to zero, and then they define a likelihood ratio statistic by

$$T_l = -2\ln{}^*(\frac{L_0}{L_1}) \sim \chi^2(m-1),$$

where L_0 is the likelihood under null hypothesis that $b_i = 0$ for all i, and L_1 is the maximized likelihood with respect to b_i. Then, they utilize the statistic to test if there is linkage in the presence of linkage disequilibrium. They pointed out this approach has good power when linkage disequilibrium is strong if the disease is recessive.

Spielman and Ewens (1996) also proposed a new statistic for multi-allelic marker:

$$T_{mhet} = \frac{m-1}{m}\sum_{i=1}^{m}\frac{\left(t_{i\cdot}-t_{\cdot i}\right)^2}{t_{i\cdot}+t_{\cdot i}-2t_{ii}} \sim \chi^2(m-1).$$

Kaplan et al. (1997) compared these tests mentioned above and applied Monte Carlo test to guarantee valid tests and then concluded that T_c has the lowest power than other three tests (T_m ,T_{mhet} , and T_l), and the three tests almost have similar power over all situations (the variation of recombination fraction θ, and the different disease model) and population they classified.

2.4 Extension of TDT from trio data to affected sib pair data

When considering the case of families with two affected children, Spielman et al. (1993) provided three categories to define the information from heterozygous parents by

$$\begin{aligned} i &= \textit{number of parents who transmit } M_1 \textit{ to both children} \\ j &= \textit{number of parents who transmit } M_2 \textit{ to both children} \\ h-i-j &= \textit{number of parents who transmit } M_2 \textit{ to one child and } M_2 \textit{ to the other}, \end{aligned}$$

where h is the number of heterozygous parents, and then they rewrote the parameters b and c of TDT as

$$\begin{cases} b = 2i + (h-i-j) \\ c = 2j + (h-i-j) \end{cases}.$$

By this definition, the TDT statistic could be written as

$$TDT = \frac{2(i-j)^2}{h}.$$

They also proposed other statistics for families with more than two affected offspring.

Martin et al. (1997) devised a statistic with ASP data, and called the statistic T_{sp}. Among children of heterozygous parents, let n_{11} be the number of ASPs who all accepted target allele M_1, let n_{22} be the number of ASPs who all accepted referent allele M_2, and let n_{12} be the number of ASPs who one accepted M_1 and the other accepted M_2. Then, the statistic would be

$$T_{sp} = \frac{\left(n_{11} - n_{22}\right)^2}{n_{11} + n_{22}}.$$

Wicks (2000) simulated two tests (TDT and T_{sp}) and pointed out that T_{sp} is valid when testing for both linkage and linkage disequilibrium, while TDT is only valid when testing for linkage, but not linkage disequilibrium. However, TDT is more powerful than T_{sp} since TDT utilizes excess sharing—that is the tendency for $n_{11} + n_{22}$ exceeding n_{12} as linkage is present. Wicks also defined a general TDT-like statistics for ASPs as

$$T(\alpha) = \frac{\left(n_{11} - n_{22}\right)^2}{(1-\alpha)\left(n_{11} + n_{22}\right) + \alpha n_{12}}, 0 \le \alpha \le 1.$$

We can observe that T_{sp} and TDT are the special case for $\alpha = 0$ and $\alpha = 1/2$, respectively. He found $T(1)$ is most powerful test for detecting linkage and it has the correct asymptotic false-positive error rate under the null hypothesis, since the statistic $T(1)$ exploits excess sharing to the fullest extent possible.

2.5 Extension of TDT without parents' data

Traditional TDT method required marker information of trio data, included an affected child and his or her parent, but in some late onset, such as cardiovascular, non-insulin-dependent diabetes, and other age related diseases, it's difficult to know that. To handle this form of problem, some researchers tried to reform TDT method, for example, sib transmission/disequilibrium (S-TDT) (Spielman et al. 1998) and sibship disequilibrium test (SDT) (Horvath and Laird, 1998). They all utilized the marker data of unaffected sibs instead of parents.

The S-TDT determines if the marker allele frequency is different between affected offspring and their unaffected sibs significantly. It has two procedures, one is the permutation procedure, it can calculate the $\boldsymbol{P}$ -value that tests if the number of interested allele $\boldsymbol{M_1}$ is randomly arrange in affected and unaffected groups, but it needs sufficiently large number of replicates to keep a precise $\boldsymbol{P}$ -value . The other one is a Z-score procedure; it utilizes the hypergeometric distribution to estimate the expected mean $\boldsymbol{U}$ and variance $\boldsymbol{V}$ of interested allele $\boldsymbol{M_1}$, and calculates the Z score,

$$\boldsymbol{Z = \frac{(Y-U)}{\sqrt{V}}},$$

where $\boldsymbol{Y}$ is the observed number of $\boldsymbol{M_1}$, or the Z score with a continuity correction as

$$\boldsymbol{z' = \frac{(|Y-U| - \frac{1}{2})}{\sqrt{V}}},$$

and then the $\boldsymbol{P}$ -value can be calculated by normal distribution approximation. They also combine the TDT and the S-TDT by assuming the expected mean and variance of TDT, $\boldsymbol{n/2}$ and $\boldsymbol{n/4}$, respectively, and adding them with expected mean and variance of S-TDT. Lastly, we can calculate the combined Z score and corresponding $\boldsymbol{P}$ -value.

The formula of S-TDT is similar with the Mantel-Haenszel test (Laird et al. 1998). It is noticeable that if we have the information of parent, we should choose TDT rather than S-TDT, because under such circumstance, TDT is more powerful than S-TDT. Although S-TDT is useful when the parent data are missed, it has some restriction: (1) the sibship must have at least one affected and one unaffected member; and (2) in the sibship, all members should not have the same genotype. Another method, SDT, is a nonparametric sign test. First, it denoted the mean number of target allele among affected (m_A^1) and unaffected (m_U^1) siblings as

$$\boldsymbol{m_A^T} = \textbf{(total number of target alleles among the affecteds)/n}_{\mathrm{A}}$$
$$\boldsymbol{m_U^T} = \textbf{(total number of target alleles among the unaffecteds)/n}_{\mathrm{U}}\,,$$

where n_A and n_U are the total number of affected and unaffected members in the sibship. They denoted the difference of m_A^T and m_U^T by d^T, let b be the number of $d^T > 0$, and let c be the number of $d^T < 0$, so the statistic of SDT can be defined by the form of TDT. The two tests have similar power in most situations, but SDT is better than S-TDT, because it avoids accounting for correlation between the siblings, and it's relatively simple. Similar to S-TDT, SDT can also combine with TDT by $b_{SDT} = b + b_{TDT}$, and $c_{SDT} = c + c_{TDT}$.

2.6 Extension of TDT from qualitative traits to quantitative traits

Due to the increasing availability of genetic data, many quantitative traits are noticed and related with susceptibility gene. At the start of research, one might related the phenotype and genotype with linear regression model:

$$\Upsilon_i = \alpha + \beta G_i + e_i,$$

where Υ_i is the quantitative phenotype and G_i is marker's genotypes, and then we can test if the value of β equals to zero.

QTDT is proposed by Rabinowitz (1997), the linear regression model was revised as

$$\Upsilon_i = \alpha + \beta Z_i + e_i$$

where $Z_i = H_{im}(T_{im} - 1/2) + H_{if}(T_{if} - 1/2)$, $H_{im}(H_{if})$ is an indicator of heterozygosity in the mother (father), and $T_{im}(T_{if})$ is an indicator of that if the mother (father) transmits a target allele to affected child. Furthermore, Lunetta et al. (2000) rewrote the QTDT,

$$\Upsilon_i = \alpha + \beta\left(G_i - E\left[G_i \mid g_{im}, g_{if}\right]\right) + e_i.$$

Fulker et al. expanded the linear model to partition the covariate into between- and within- mating type information, two variables. They called the approach hierarchical QTDT (HQTDT) which has the form

$$\Upsilon_i = \alpha + \beta_B G_M + \beta_W (G_i - G_M) + e_M + e_i,$$

where e_M is a mating-type specific residual and it is assumed to be $N(0, \tau^2)$, G_M is some average genotype for mating type M. The test of association is based on an LRT of β_W.

On the other hand, the value of Υ in the original QTDT model is restricted to α, regardless of the mating type, so Gauderman (2003) proposed a reformatory method, $QTDT_M$,

$$\Upsilon_i = \alpha_M + \beta G_i + e_i.$$

The difference from other models is that the extra term α_M considers the different effect of 6 mating type.

The models described above are all prospective, but there are some other models that are retrospective, such as retrospective QTDT (RQTDT) (Liu et al. 2002), which lets the genotype of affected children be modeled as a function conditional on their phenotype and their parental genotypes. Then, by Bayes rule, the likelihood becomes

$$L(\alpha,\beta,\sigma) = \prod_{i=1}^{N} \frac{f\left(\Upsilon_i \mid G_i, \alpha, \beta, \sigma\right) \Pr(g_i \mid g_{im}, g_{if})}{\sum_{g^* \mid g_{im}, g_{if}} f\left(\Upsilon_i \mid G^*, \alpha, \beta, \sigma\right) \Pr(g^* \mid g_{im}, g_{if})},$$

where σ^2 is the residual variance. The summation in the denominator includes four genotype (g^*) and it could be transmitted to a child conditional on parental genotypes.

Gauderman (2003) compared these models under genetic main effects, gene-environment interaction, and gene-gene interaction, and then pointed out

$QTDT_M$ needs less sample size than other models for testing these effects, i.e. $QTDT_M$ is the most efficient approach.

2.7 Localization of disease locus in case-parent trio designs

Liang et al. (2001b) applied the conception of TDT and developed a new statistic $Y(t)$, called the preferential- transmission statistic (It would be described in more detailed in Chapter 3). Through complicated calculation, they showed the relationship between $Y(t)$ and $Y(\tau)$, the preferential-transmission statistic with arbitrary marker and susceptibility gene's locus, respectively is:

$$E\left[Y(t)\,|\,\Phi\right]=(1-2\theta_{t,\tau})E\left[Y(\tau)\,|\,\Phi\right](1-\theta_{t,\tau})^{N}\left\{\Pr\left[h(t)\,|\,h(\tau)\right]\right\},$$

where Φ represents the event that the offspring of trio is affected, $\theta_{t,\tau}$ is the recombination fraction between marker locus t and the postulated disease gene location τ, and N is the number of generations since the introduction, into the population, of a disease-causing mutation at location τ.

Finally, they applied the generalized-estimating-equation (GEE) (Liang and Zeger, 1986) to estimate the parameter $\delta = (\tau, C, N)$.

The approach can test the null hypothesis that there is no linkage or linkage disequilibrium (LD) to the region R by testing if $C \equiv 0$. In contrast to TDT, Liang et al.'s approach simultaneously uses all the markers' information, so it is more powerful than TDT. The approach uses the data of every marker over the specific region regardless of whether the parent's genotype is heterozygous or homozygous, and also provides valid standard-error estimates of parameter through GEE. Most of all, there is no need to assume the genetic model of the disease in this approach.

2.8 Multipoint approach with covariate data

Liang et al. (2001a) also proposed a multipoint approach with affected sib pair (ASP) data by the model as follows.

$$E\left\{S(t)\,|\,\Phi\right\}=1+(1-2\theta_{t,\tau})^{2}C\ ,\ C=E\left\{S(\tau)\,|\,\Phi\right\}-1,$$

where $S(t)$ and $S(\tau)$ represent the number of alleles shared identical-by-descent (IBD) at a marker locus t and a susceptibility locus τ, respectively. Glidden et al. (2003) incorporated age as a covariate X into the model of Liang et al. and assume C is a function of covariate X. Their model has the form

$$\begin{aligned}\mu(t\,|\,x)=E\left\{S(t)\,|\,X=x,\Phi\right\}&=1+(1-2\theta_{t,\tau})^{2}C(x)\\&=1+\exp\left(-0.04\left|t-\tau\right|\right)C(x).\end{aligned}$$

Furthermore, since the value of C is -1 to 1, it could be transformed and postulated as a logistic formula:

$$\text{logit}\left[\left\{C(x)+1\right\}/2\right]=\alpha+\beta^{T}x$$

Then, we could utilize GEE method to estimate the parameters, $\delta=(\tau,\alpha,\beta_{1},\ldots,\beta_{p})$.

Conclusively, they find incorporating covariate data could provide more information, increase precision in localizing susceptibility gene and other parameters, and minimize the effect of the unknown heterogeneity process, even when it is mismodelled.

2.9 Multipoint approach with covariate data and non-parametric approaches

Although multipoint linkage analysis using sibpair designs is a popular approach to test the location of interested trait, some issues, such as genetic heterogeneity, gene-gene, and gene-environment interaction, should be addressed properly. Chiou et al. (2005) proposed an approach which assumes trait locus’ genetic effect is a function of covariate, and the function represents the probability of a sibpair sharing the same allele at the trait locus. Then, they estimated the susceptibility gene locus with GEE method and the genetic effect with a nonparametric approach iteratively.

For the j^{th} marker and i^{th} sibpair, they applied Liang et al.'s (2001a) model and rewrote it as

$$E\left\{S_i(t_j)\,|\,\Phi\right\}=1+(1-2\theta_{t_j,\tau})^2 C(x_{i1},x_{i2})$$

Let $g=(g_1(x_1,x_2),g_2(x_1,x_2))$ be some transformed predictor of covariate pair (x_1,x_2) which is in relation to C, and estimate C and τ iteratively between equation (1) and (2).

$$\sum_{i=1}^{n}\left[\left(S_i^*(\tilde{\tau})-1\right)-\beta_0-\beta_1(g_{i1}-g_1)-\beta_2(g_{i2}-g_2)\right]^2 K_2\left(H^{-1}\left(g-G_i\right)\right) \qquad \textbf{(1)}$$

where $S_i^*(\tilde{\tau})$ is the imputed IBD sharing at $\tilde{\tau}$, $G_i=(g_{i1},g_{i2})$ with $g_{ik}=g_k\left(x_{i1},x_{i2}\right)$, K_2 is a bivariate kernel function, and H is a nonsingular square bandwidth matrix; and

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

where $w_j(\tilde{\tau})$ is the weight function centering at $\tilde{\tau}$. It may depend on the distance of t_j and $\tilde{\tau}$ or the average of the two nearest IBD sharing at two markers.

When we obtain the estimates $\tilde{\tau}$, we can calculate $S_i^*(\tilde{\tau})$ and the covariate data G_i, and then get the estimate $\hat{C}(g)=\hat{\beta}_0$ for the function C, then plug the estimate $\hat{C}(g)$ into the estimating equation to estimate the parameter of interest $\delta=(\tau)$ again. This procedure is repeated until convergence is met.

$$\sum_{i=1}^{n}\left(\frac{\partial\mu_i(\tau)}{\partial\tau}\right)' Cov^{-1}(S_i)(S_i-\mu_i(\tau))=0 \qquad \textbf{(2)}$$

This approach not only keeps the preciseness when using Liang et al.'s model (2001), it does not need to assume the relation between C and covariates and avoids mis-specifying the function C by an invalid model.

2.10 Interpreting analyses of continuous covariates in ASP

Schmidt et al. (2007) discussed three plausible models for the relationship between continuous covariate and disease risk or linkage heterogeneity. First, the covariate distribution is determined by a quantitative trait locus (QTL). Second, the covariate affects the disease risk through statistical interaction with a disease susceptibility locus. Third, the covariate distribution is different in families linked or unlinked to a particular disease susceptibility locus. Then, they utilize three approaches, a regression-based QTL analysis, a nonparametric analysis of the binary affection status, and the ordered subset analysis (OSA), to analysis above three relations.

They used a prospective logistic regression model as the penetrance function to generate binary disease outcomes in their simulation studies as follows.

$$\ln\left(\frac{P\left(D=1\,|\,x_1,x_2\right)}{1-P\left(D=1\,|\,x_1,x_2\right)}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$$

$$\beta_1=\ln(OR(G)),\beta_2=\ln(OR(E)),\beta_3=\ln(OR(G\times E)),$$

where D=1 for affected, D=0 for unaffected individuals, x_1=1 for the susceptibility genotype(s), x_1=0 for the referent genotype(s), and x_2 is the value of a normally distributed continuous covariate represents environmental factor.

Among the three approaches, QTL analysis is useful to detect $G\times E$ interaction between the covariate and a disease susceptibility locus when the data included unaffected sib pair that can provide information only in the QTL analysis, but not other two approaches. But the data analyzed by the QTL analysis should be dealt with by a standardized process. OSA has a significant result when a gene influences variability in the population distribution of a continuous disease risk factor, rather than a disease susceptibility locus influencing the disease risk directly. Finally, the NPL is more powerful then other two analyses when the $\boldsymbol{OR(G \times E)}$ is high, whether the data included unaffected sib pair or not.

3. The Proposed Method

3.1 Notation and Preferential-Transmission Statistic

Apply the approach of Liang et al. (2001b), consider n case-parent trios are sampled for an association study, and let $\boldsymbol{R}$ be a chromosomal region of length $\boldsymbol{T}$ cM (centimorgan) which contains no more than one susceptibility gene at unknown location $\boldsymbol{\tau}$ over region $\boldsymbol{R}$. Denote $\boldsymbol{M}$ markers framed region $\boldsymbol{R}$ with locations of $\boldsymbol{0 \le t_1 < t_2 < ... < t_M \le T}$. For simplicity, we suppose there are two alleles per marker and define $\boldsymbol{Y(t)}$ as the paternal preferential-transmission statistic

$$\boldsymbol{Y(t) = Y_1(t) - Y_2(t),}$$

where $\boldsymbol{t}$ is one of $\boldsymbol{M}$ markers and

$$Y_1(t) = \begin{cases} \textbf{1, if the transmitted paternal allele} \\ \quad \textbf{at t is target allele } \boldsymbol{H(t)} \\ \textbf{0, if the transmitted paternal allele} \\ \quad \textbf{at t is nontarget allele } \boldsymbol{h(t)} \end{cases} ,$$

$$Y_2(t) = \begin{cases} \textbf{1, if the nontransmitted paternal} \\ \quad \textbf{allele at t is target allele } \boldsymbol{H(t)} \\ \textbf{0, if the nontransmitted paternal} \\ \quad \textbf{allele at t is nontarget allele } \boldsymbol{h(t)} \end{cases} .$$

Similarly, maternal preferential-transmission statistic also can be defined as $\boldsymbol{X(t) = X_1(t) - X_2(t)}$, accordingly. From now on, we only discuss the property and extension of $\boldsymbol{Y(t)}$, since it applies to $\boldsymbol{X(t)}$ completely as well.

The expected number of preferential-transmission statistic of Liang et al.'s model has the form

$$\begin{aligned} \boldsymbol{\mu(t_j)} &= \boldsymbol{E\left[Y_i(t_j) \mid \Phi\right] = E\left[X_i(t_j) \mid \Phi\right]} \\ &= \boldsymbol{(1-2\theta_{t,\tau})C(1-\theta_{t,\tau})^N \pi_j,} \end{aligned}$$

where $\boldsymbol{C = E\left[Y(\tau) \mid \Phi\right] = E\left[X(\tau) \mid \Phi\right]}$, $\boldsymbol{\theta_{t,\tau}}$ is the recombination fraction between $\boldsymbol{t}$ and $\boldsymbol{\tau}$, $\boldsymbol{N}$ is the number of generations when a disease-causing mutation at $\boldsymbol{\tau}$ was introduced into the generation, $\boldsymbol{i = 1,...,n}$, $\boldsymbol{n}$ is the number of trios, and

$\pi_j = \Pr\left[h(t_j) \mid h(\tau)\right],\ j = 1, \ldots M$.

Since some diseases are associated with covariates like hypertension, BMI, fat in the blood, age, or the level of disease, and some notable recent researches showed that incorporating covariates information can amplify the signals of linkage (Glidden et al. 2003; Chiou et al. 2005), we rewrote the formula and added a covariate Z associated with an affected child into C (assuming the recombination does not depend on Z) as

$$\mu(t) = E\left[Y(t) \mid Z = z, \Phi\right] = (1 - 2\theta_{t,\tau})C(z)(1 - \theta_{t,\tau})^{N} \pi_j, \tag{3}$$

where $C(z)$ is $E\left[Y(\tau) \mid Z = z, \Phi\right]$. We expect the covariate Z will be helpful to estimate the location of the susceptibility gene more accurately. Equation (3) represents the transmitted number at t as a function of recombination $\theta_{t,\tau}$, the number of generations N , and the expected transmitted number at susceptibility locus τ and covariates Z . Assuming the Haldane (1919) map function,

$$\theta_{t,\tau} = (1 - \exp(-0.02\left|t - \tau\right|)) / 2. \tag{4}$$

On the other hand, π_j represents the probability of the non-target allele is carried at marker t_j upon the normal allele at susceptibility locus τ , as it is difficult to be observed among collected data, we replace it with $\hat{\pi}_j$ by

$$\hat{\pi}_j = \frac{\sum_{i=1}^{n}\left[1 - Y_{i2}(t_j) + 1 - X_{i2}(t_j)\right]}{2n}. \tag{5}$$

The parameter $C(z)$ plays an important role in our approaches, it measures the degree of overall linkage to R, and decides how precise the estimation of the disease locus τ is. If the absolute value of $C(z)$ is close to 1, the magnitude of linkage is more strong, and the estimation of τ is more precise, in other words, the variance is smaller. We will illustrate it in the next Chapter. By the same token, if the absolute value of $\boldsymbol{C(z)}$ is close to 0, there is little linkage over the region and has minimal information about the estimation of $\boldsymbol{\tau}$. Some complex diseases may involve interactions of gene and environment factors, or different patients may have different genetic effects from the same disease-locus, or the phenocopies may result from environment factors…etc. The complexities of the underlying genetic mechanism of a disease may weaken the signal of linkage, if a covariate $\boldsymbol{Z}$ is associated with the underlying mechanism of a disease, by incorporating the covariate into the linkage mapping, one may obtain more precise estimation of $\boldsymbol{\tau}$ (Glidden et al. 2003).

Now, we introduce two approaches to estimate $\boldsymbol{\tau}$ by incorporating a covariate $\boldsymbol{Z}$ through parametric and nonparametric methods.

3.2 The Parametric Approach with Covariates

There are multiple parametric methods that could be utilized to model $\boldsymbol{C}$ as a function of the covariates, we employed the logistic type models to establish the relation of a covariate $\boldsymbol{Z}$ and $\boldsymbol{C(z)}$ as a dependent variable Glidden et al. (2003). First, since the range of $\boldsymbol{C(z)}$ is $\boldsymbol{[-1,1]}$, we must transform its range into $\boldsymbol{[0,1]}$, hence, the model takes the form

$$\begin{aligned} logit\left[E\left\{S(\tau)\mid Z=z\right\}\right] &= logit\left[\left\{C(z)+1\right\}/2\right] \\ &= log\left[\frac{1+C(z)}{1-C(z)}\right] \\ &= \alpha+\beta^{T}Z \qquad . \end{aligned}$$

$\boldsymbol{\left(C(z)+1\right)/2}$ characterizes the probability that an affected child received a target allele at $\boldsymbol{\tau}$ from his or her heterozygous parent. Thus,

$$C(z)=\frac{\exp\left(\alpha+\beta^{T}z\right)-1}{\exp\left(\alpha+\beta^{T}z\right)+1}. \qquad \textbf{(6)}$$

The vector of parameters $\delta=(\tau,N,\alpha,\beta_1,...,\beta_p)$, p is the dimension of covariates. By replacing $C(z)$ with a logistic regression model,

$$E\left[Y(t)\,|\,Z=z,\Phi\right]=(1-2\theta_{t,\tau})\left(1-\frac{2}{\exp\left(\alpha+\beta^T z\right)+1}\right)(1-\theta_{t,\tau})^N\pi_j. \qquad (7)$$

We then apply the Generalized Estimating Equation (GEE) (Liang and Zeger, 1986) approach to solve the parameters. That is, estimating $\delta=(\tau,N,\alpha,\beta_1,...,\beta_p)$ by solving

$$\begin{aligned}S(\delta)=\sum_{i=1}^{n}\Bigg[&\frac{\partial\mu(\delta,\hat{\pi})}{\partial\delta}Cov^{-1}(Y_i)\left\{Y_i-\mu(\delta,\hat{\pi})\right\}+\\ &\frac{\partial\mu(\delta,\hat{\pi})}{\partial\delta}Cov^{-1}(X_i)\left\{X_i-\mu(\delta,\hat{\pi})\right\}\Bigg]=0\,,\end{aligned}$$

where

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

and

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

The parameter estimates $\hat{\delta}$ are consistent estimates, hence, have the asymptotic property. Based on the asymptotic property, we could calculate the variance estimates of $\hat{\delta}$ by

$$\widehat{Var}(\hat{\delta})=A^{-1}BA^{-1}$$

where

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

and

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

This approach allows one to make inferences for the parameters of interest. In addition, we could test if the covariates z_p on allele transmitting is significant by testing the null hypothesis: $\beta_p = 0$.

One minor modification is necessary when applying the GEE method, since the variable $\mu(\delta, \hat{\pi})$ is not differentiable with respect to τ (strictly speaking) through $|t - \tau|$ in the Haldane mapping function (1919). This concern could be fixed by replacing $|t - \tau|$ by

$$
\begin{cases} |t - \tau| & \text{if } |t - \tau| \geq \varepsilon \\ \dfrac{1}{2\varepsilon}(t - \tau)^2 + \dfrac{1}{2}\varepsilon & \text{if } |t - \tau| > \varepsilon \end{cases}, \qquad \textbf{(8)}
$$

where ε is a positive number. We will discuss the effect of the value of ε in the next Chapter.

3.3 The Nonparametric Approach with Covariates

A criticism of multiple parametric modeling is that the approaches imposed may not reflect the underlying mechanism properly. Here, we refer to a nonparametric method proposed by Chiou et al. (2005) who estimated the function C by spline and kernel smoothing methods as local polynomial regression (Fan and Gijbels, 1996). Before estimating $C(z)$, we need the information about imputed allele transmitting at $\hat{\tau}$, $Y_i^*(\tilde{\tau})$, we utilize the allele transmitting information at markers, $Y_i(t_j)$, near $\tilde{\tau}$, to impute $Y_i^*(\tilde{\tau})$ with an weighted average, i.e.,

$$Y_i^*(\tilde{\tau}) = \sum_{j=1}^{M} w_j(\tilde{\tau}) Y_i(t_j), \tag{9}$$

where $w_j(\tilde{\tau})$ is the weight function of nearby markers centering at $\tilde{\tau}$. The weight function we employ here is to take allele transmitting at two nearest markers, $Y_i(t_k)$ and $Y_i(t_k)$ with $t_k < \tilde{\tau} < t_\ell$ such that

$$Y_i^*(\tilde{\tau}) = wY_i(t_k) + (1-w)Y_i(t_\ell),$$

where $w = (t_\ell - \tilde{\tau}) / t_\ell - t_k$. When the location of gene falls between two reasonably close marker loci, this weight function could work well.

Next, we could obtain $\hat{C}(z)$ by minimizing the following kernel weighted least squares function

$$\sum_{i=1}^{n} \left[\left(Y_i^*(\tilde{\tau}) - 1 \right) - \beta_0 - \beta_j (G - g(z)) \right]^2 K_H \left(G - g(z) \right),\ j = 1, \cdots, p \tag{10}$$

where G could be covariates z, or other transformation of z, like $\exp(z)$, $\log(z)$ and so on; H is a $p \times p$ symmetric positive definite matrix depending on sample size n; K is a p-variate and $\int K(u)du = 1$; $K_H(u) = \left| H^{-1/2} \right| K(H^{-1/2}u)$, and we called $H^{1/2}$ the bandwidth matrix; and $\hat{\beta}_0$ is the estimate of C (Ruppert and Wand, 1994). Here, we choose the kernel function K and the bandwidth matrix $H^{1/2}$ to be

$$K(u) = (2\pi)^{-1/2} \exp(-\frac{u^2}{2}), -\infty < u < \infty,$$

and

$$H^{1/2} = \frac{\max(Z_i) - \min(Z_i)}{3}, i = 1, \cdots, n,$$

and then we could solve equation (10) by

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \left(Z_z^T W_z Z_z \right)^{-1} Z_z^T W_z Y^*(\tilde{\tau}), \quad \text{where } Z_z = \begin{bmatrix} 1 & (Z_1 - z)^T \\ \vdots & \vdots \\ 1 & (Z_n - z)^T \end{bmatrix},$$

where

$$Y^* = \left[Y_1^*, \ldots, Y_n^* \right]^T ,$$

and

$$W_z = diag\left\{ K_H(Z_1 - z), \ldots, K_H(Z_n - z) \right\} .$$

Then we employ $\hat{C}(z)$ and put it in the equation below to update the estimate of τ by solving this equation (GEE),

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

where $\delta = (\tau, N)$.

Through the iterative process between updating $\hat{C}(z)$ in the nonparametric model and $\hat{\tau}$ in the GEE method, the estimate $C(z)$ and τ could be obtained when convergence is reached.

4. Simulation Studies

4.1 Disease models

4.1.1 Logistic regression models

In the simulation study, we carry out three different disease models to assess the performance of the two proposed methods for quantitative trait-related covariates. First, we assume a prospective logistic regression model (See Figure 1) as the penetrance function to generate binary disease outcomes for a case-parent trio data:

$$\mathbf{ln}\left(\frac{P(D=1\,|\,g,e)}{1-P(D=1\,|\,g,e)}\right)=\beta_0+\beta_1 g+\beta_2 e\,,$$

where $D=1$ for affected and $D=0$ for unaffected individuals, $g=1$ for the susceptible genotype(s), $g=0$ for the referent genotype(s) (For dominant model, $g=1$ when genotype is HH or Hh, $g=0$ when genotype is hh, for recessive model, $g=1$ when genotype is HH, $g=0$ when genotype is Hh or hh, and for additive model, we separate g into g_1 and g_2, β_1 into β_{11} and β_{12}, and then $g_1=1$ when genotype is HH, $g_1=0$ when genotype is Hh or hh, $g_2=1$ when genotype is Hh, and $g_2=0$ when genotype is HH or hh.), e is a value of environmental effect, E, which follows a standard normally distribution, and the parameter vector β are the natural logarithm of the odds ratio (ORs). By the logistic regression model, we set up the relative risk β according to the inheritance mode, and then we can calculate the penetrance f_0, f_1, and f_2 for genotype HH, Hh, and hh, respectively. On the other hand, we generate the trait depending on the genotype of affected individual:

$$z_i=\mu+g_i+e_i,\;\; e_i\sim N(0,1)\quad \text{(Haseman and Elston, 1972)}$$

($g_i=a$ when genotype is homozygous HH, $g_i=d$ when genotype is heterozygous Hh, $g_i=-a$ when genotype is homozygous hh. For dominant models, $d>0$; for recessive models, $d<0$; and for additive models, $d=0$).

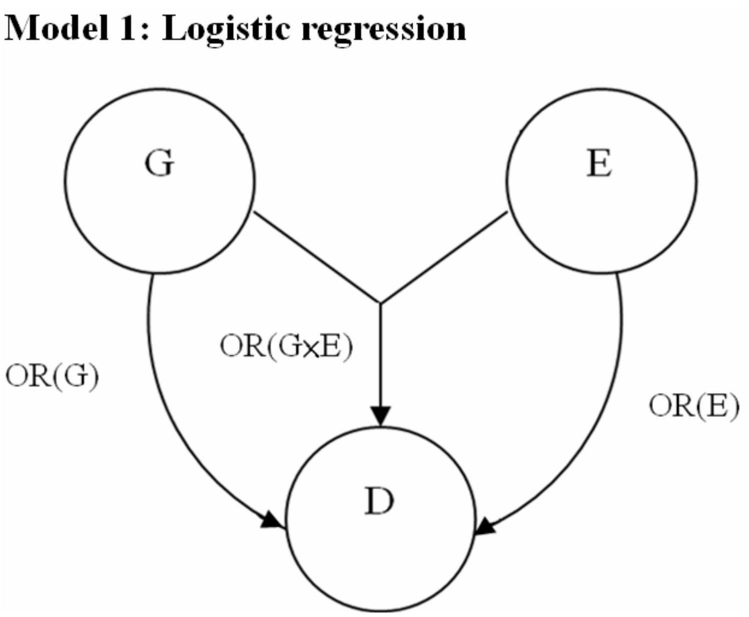

Figure 1. Graphical illustration of logistic regression disease models

4.1.2 Threshold models

Second, we decided whether an individual is affected or not by a threshold model (See Figure 2). For a start, we generate trait z_i for each individual directly,

$$z_i = \mu + g_i + e_i, i = 1, \ldots, n,$$

where μ is the mean of quantitative traits for all individual, g_i is genetic effect and the value is determined by personal genotype ($g_i = a$ when genotype is HH, $g_i = d$ when genotype is Hh, $g_i = -a$ when genotype is hh. For dominant model $d > 0$, for recessive model $d < 0$, and for additive model $d = 0$), and e_i is the environmental effect with a standard normal distribution. After knowing the value of traits, we take a threshold (T) depends on the prevalence of population (See Equation (11) and Figure 3) under the simulation, and if a trait of person exceeds the threshold, he or she will be diagnosed to be affected.

$$\begin{aligned} prevalence = {} & p^2 \Pr\left[Z > T \mid Z \sim N(a,1), HH\right] \\ & + 2p(1-p)\Pr\left[Z > T \mid Z \sim N(d,1), Hh\right] \\ & + (1-p)^2 \Pr\left[Z > T \mid Z \sim N(-a,1), hh\right] \end{aligned} \qquad \textbf{(11)}$$

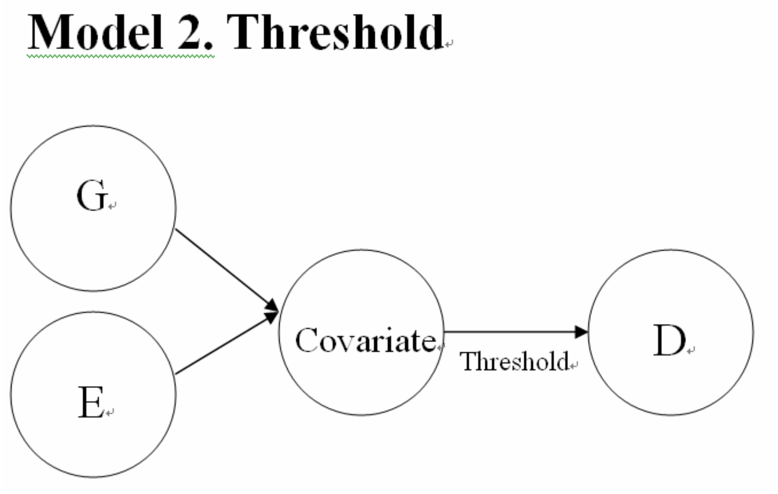

Figure 2. Graphical illustration of threshold disease models

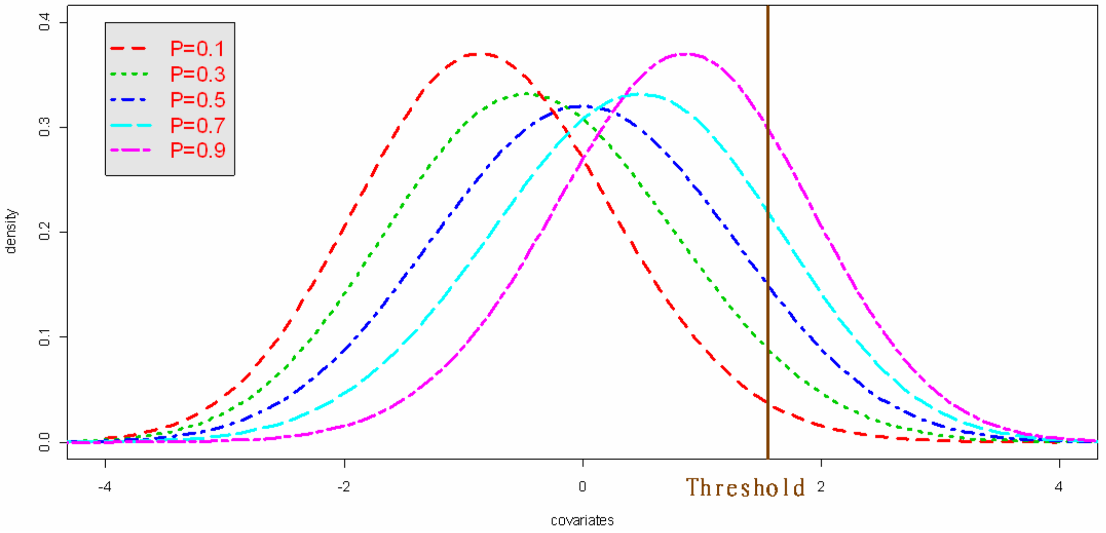

Figure 3. Disease allele frequencies and the probability density function for a covariate in the threshold models (P=disease allele frequency, a=1, d=0)

4.1.3 Fixed penetrance models

The last and simplest one is fixed penetrance models (See Figure 4), the probability for an individual being affected depends on predetermined penetrance $\boldsymbol{f_0}$, $\boldsymbol{f_1}$, and $\boldsymbol{f_2}$ for genotype $\boldsymbol{HH}$, $\boldsymbol{Hh}$, and $\boldsymbol{hh}$, respectively, and we generate the traits (covariates) in the same way as that in the logistic regression model.

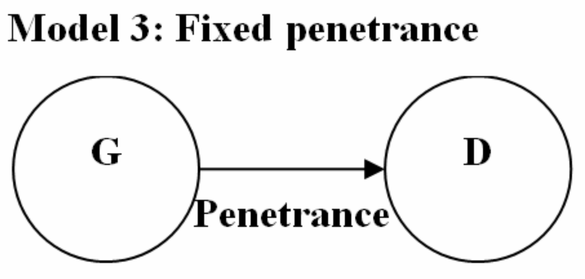

Figure 4. Graphical illustration of fixed penetrance disease models

4.2 Genotype Data

We assume the joint probability of target alleles $H(t)$ and $H(\tau)$ at marker t and disease locus τ at present generation N as

$$\Pr(H(t), H(\tau)) = \Pr(H(t))\Pr(H(\tau)) + (1-\theta_{t,\tau})^N \Delta_t^{(0)},$$

where $\Delta_t^{(0)}$ represent the degree of LD between marker t and disease locus τ at $N = 0$. Here, we apply the equation (3) of Liang et al. (2001b),

$$d(t) = P\left[H(t) \mid H(\tau)\right] - P\left[H(t) \mid h(\tau)\right],$$

and the related formula of $d(t)$ proposed by Devlin and Risch (1995),

$$d(t) = (1-\theta_{t,\tau})^N P\left[h(t) \mid h(\tau)\right].$$

We set the value of $\Delta_t^{(0)}$ to be 0.009 for all markers, and then we can calculate two important probabilities $P(H(t) \mid H(\tau))$ and $P(h(t) \mid h(\tau))$. In the simulation work, we have set the number of trios at 200 and a region with 10 fully polymorphic markers that are equally spaced between 0 cM and 0.9 cM (0.1 cM between adjacent markers) and the disease locus τ at 0.45 cM. Then, we provide the genotype at τ of parents with a disease allele probability p, then utilize these genotypes, and the conditional probabilities $P(H(t) \mid H(\tau))$ and $P(h(t) \mid h(\tau))$ to generate genotypes for markers.

After completing the genotypes of parents, we use the information of recombination $\theta_{t,\tau}$ to generate their child's genotypes at τ and markers, and then determine if the child is affected through the logistic regression, threshold, or fixed penetrance disease models.

4.3 Simulation Results

In the following simulation results, we simulate 1,000 replicates including 200 case-parent trios, and compare the relative efficiency (R.E.= $\left\{\boldsymbol{SE(\hat{\tau}_2)/(SE(\hat{\tau}_1)}\right\}^2$) in between our parametric and nonparametric procedures where covariates were incorporated with the original approach where no covariates were incorporated (Liang et al., 2001b). In addition, we examined the performance of the approaches when the disease models were logistic regression models, threshold models, and fixed penetrance models in our simulation.

For the logistic regression disease model, we assume the inheritance mode is additive, that is,

$$\mathbf{ln}\left(\frac{\boldsymbol{P(D=1\,|\,g_1,g_2,e)}}{\boldsymbol{1-P(D=1\,|\,g_1,g_2,e)}}\right)=\boldsymbol{\beta_0+\beta_1 g_1+\beta_2 g_2+\beta_3 e}\,,$$

where $\beta_0=\ln(0.01)$, $\beta_1=\ln(9)$, $\beta_2=\ln(5)$, $\beta_3=\ln(2)$, and

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

We set $a=1, d=0$ (See the definition in section 4.1). Figure 7 illustrates the true, observed, and three fitted curves in one of these models.

For the threshold model, we generated a trait based on the logistic model, and used the threshold with prevalence of 0.05, the threshold was 1.022636. Those with a trait greater than 1.022636 were affected. We will show the estimating results of different prevalences in the following simulation studies.

In Tables 1-5, we display scenarios with different numbers of generation, different disease allele frequencies at $\boldsymbol{\tau}$, sample sizes, frequencies of a targeted allele of markers, and numbers of markers over the same region. From the five Tables, we summarize results as follows: In Table 1, smaller standard errors for estimates of τ were found in larger generations. In Table 2, it shows that the higher the value of C , the transmitted probability of the targeted allele, the more precise and efficient of the estimate for τ . The magnitude of C depends on many factors including the disease allele frequency, we plot two simple diagrams (Figure 5 & 6) to display the association of them.

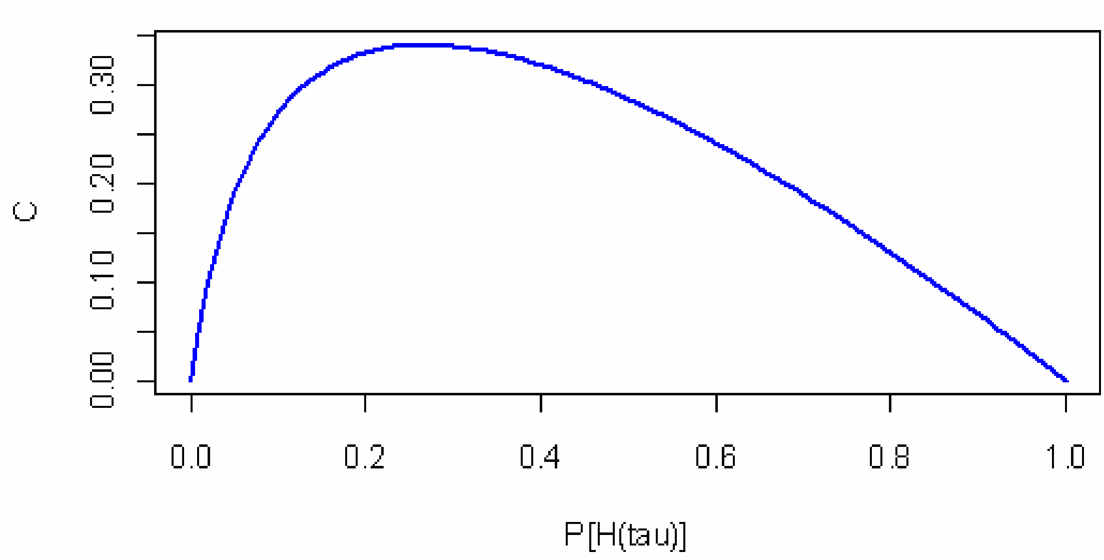

Figure 5. The curves of the transmitted probability C at τ depend on the disease allele frequency $P[H(\tau)]$ with penetrance rates $f_0 = 0.491$, $f_1 = 0.153$, and $f_2 = 0.022$ in the logistic regression disease model.

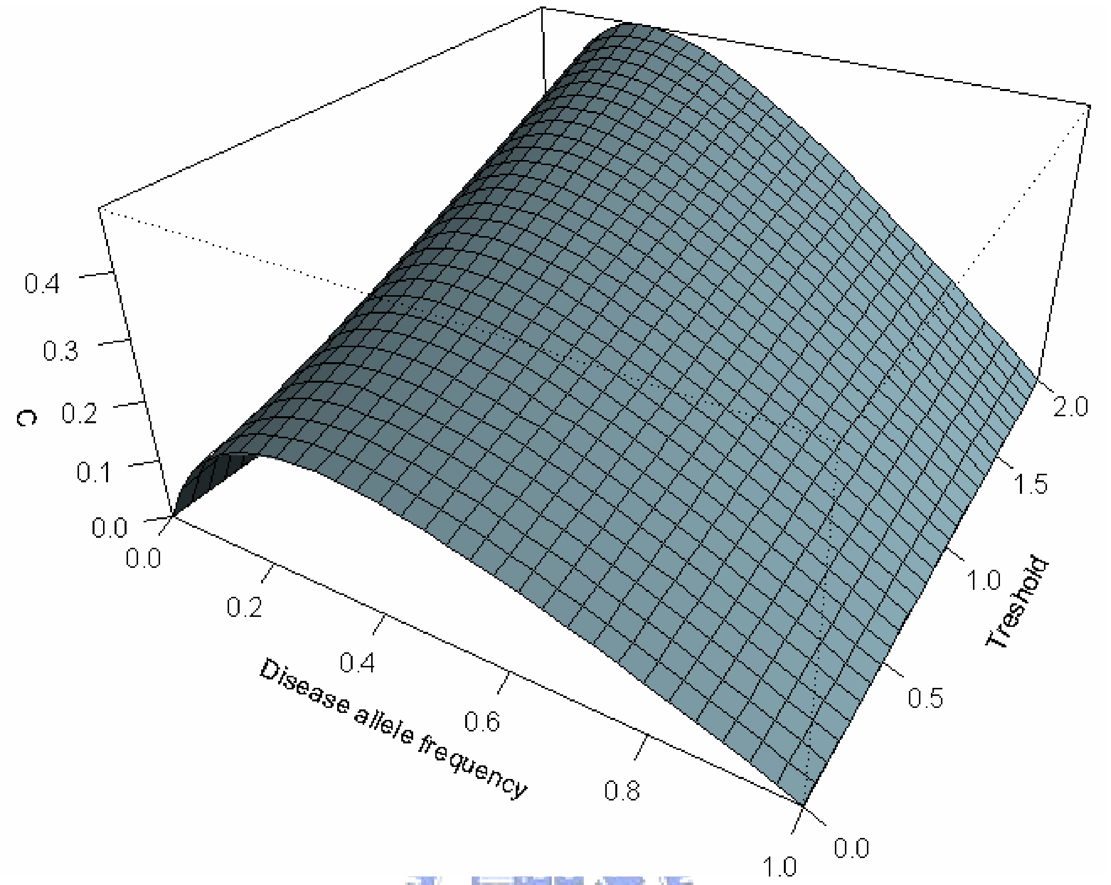

Figure 6. The curves of the transmitted probability at τ depend on the disease allele frequency $P[H(\tau)]$ and the chosen threshold in the threshold model.

Additionally, we compare the relative efficiency among different underlying disease models including logistic regression and threshold models in Table 2. The estimates for τ from the phenotype data generated by the logistic regression model are more efficient than those from those generated by the threshold model regardless of parametric or nonparametric approaches for a specific C. Apparently, the larger the C value, the more efficient of the estimate for τ and β in both parametric and non-parametric approaches. As demonstrated in Table 3, the precision of estimates will be improved by enlarging the sample sizes. It is notable that our proposed approaches were more efficient with a smaller sample size compared to the estimate without incorporating a trait (covariate) when the disease model is the threshold model, and it maintained accurate estimation of susceptibility locus τ

even if there were only 50 trios when the underlying genetic model was the logistic regression model. From Table 4, we found that if the frequency of the targeted allele at marker is more deviant from the frequency of the disease allele at τ, the variance estimate will be much larger for $\hat{\tau}$, and the ***p-value*** of testing $\boldsymbol{\beta} = \mathbf{0}$ became not as significant, our proposed approach was quite robust in terms of efficiency compared to the original approach (without a covariate). Table 5 illustrates the results from the scenarios with 10 markers and 20 markers on the same region of length 0.9 cM. Apparently, denser marker could make the estimates more precise (less bias), but the difference between the results from approaches with and without a covariate remained similar.

In Table 6, we have tried three different values (smaller than the distance of two adjacent markers) of ε in equation (6) to find the most optimum one and to study the robustness of various ε. The results were similar except for the convergence rate of the 1,000 replicates. It is quite obvious that when the value of ε equal to 0.05 cM, the convergence rate is the highest, i.e. half of distance between two adjacent markers (0.1 cM) was an optimum choice.

Tables 7 ~ 9 reveal the influence from different relative risks β_0, β_1, β_2, and β_3 in the underlying disease models. The value of $\boldsymbol{C}$ varied according to these risks, and it's again showed that a larger value of $\boldsymbol{C}$ made the estimates more precise. Moreover, estimates from our proposed approaches were more efficient than the original approach when $\boldsymbol{C}$ was small. In Table 10, we changed the value of a, the genotypic effect at τ when simulating the covariate. We found that the results were different in the two models. For the logistic regression disease model, apparently, increasing value of a can keep the estimate more accurate and more efficient until— $\boldsymbol{a}$ exceeding 5, but for the threshold model, since $\boldsymbol{C}$ changed corresponding to a, the comparison was hard to make. It is expected that the estimate of β decreased with an increased value of a. The result from changes of prevalence rates are displayed in Table 11, which depended on the magnitude of $\boldsymbol{C}$. After checking the results from a variety of disease models including logistic regression and threshold models, we simulated the fixed penetrance models as shown in Table 12, we examine if it has the same performance as other two models. Basically, it is mostly affected by $\boldsymbol{C}$, but we observed that recessive model ($\boldsymbol{f_0 = 0.67}$, $\boldsymbol{f_1 = 0.05}$**, and** $\boldsymbol{f_2 = 0.007}$**)** is more efficient and more significant (referring to testing for $\boldsymbol{\beta}$) compared to the dominant model ($\boldsymbol{f_0 = 0.95}$, $\boldsymbol{f_1 = 0.9}$**, and** $\boldsymbol{f_2 = 0.01}$**)** with the similar average $\boldsymbol{C}$ values**.**

From Table 13 to Table 18, we compare the difference in bias and relative efficiency of estimating $\boldsymbol{\tau}$ when the covariate is controlled by a locus near the disease locus rather than the locus τ itself. With an exception for the nearest marker at 0.5 cM in some cases, the farther distance between a locus controlling the covariate and τ was, the smaller the estimates of β was. When the covariate was controlled by an unlinked locus, in spite of the estimate was better than that without a covariate, the corresponding ***p-value*** of testing $\boldsymbol{\beta = 0}$ is almost near 1, which was as expected. We found the bias of the estimate for τ was a useful index to distinguish whether the covariate's locus is actually τ itself or it is near but not τ, since the loci close to τ induced more serious bias (See Table 13 ~ Table18), if we want to make sure whether the covariate is controlled by τ, in addition to evaluate the estimate for β, we could also check if the estimate for τ is similar from that obtained from the mapping without incorporating the covariate. Although sometimes, we won't be able to obtain the estimate without incorporating a covariate due to the lack of statistical power.

In addition, a covariate (quantitative trait) with a dominant genetic model mostly provided a more efficient estimate for τ than that under an additive or recessive model regardless of the parametric or nonparametric approaches, and regardless of the disease models of logistic regression, threshold models, or inheritance modes. Finally, we added the genotype at τ, two qualitative variable $\boldsymbol{Z}_{fa}$ and $\boldsymbol{Z}_{mo}$, into the equation (6) of the proposed parametric method, and let $\boldsymbol{\beta}$, $\boldsymbol{\beta}_{fa}$, and $\boldsymbol{\beta}_{mo}$ be the regression coefficients for the covariate $\boldsymbol{Z}$, genotypes $\boldsymbol{Z}_{fa}$ and $\boldsymbol{Z}_{mo}$, respectively. We compared the results with the original results from the parametric method, we found that $\boldsymbol{\beta}$ becomes non-significant because the estimate of $\boldsymbol{\beta}$ was close to 0, on the contrary, $\boldsymbol{\beta}_{fa}$ and $\boldsymbol{\beta}_{mo}$ were all significant when the two covariates were added (see column 1 and column 2 of Table 19). The reason is that the covariate $\boldsymbol{Z}$ no longer carries any additional information on $\boldsymbol{\tau}$, when the genotype of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ were incorporated.

5. A Data Example

We applied our proposed approaches to a case-parent trios study of oral cleft from four population (Korea, Maryland, Singapore, and Taiwan) reported in Sull et al. (2008). In this international study, they recruited 383 case-parent trios of oral cleft (see Table 20) and gathered their genotypes at 635 SNPs spanning about 175 cM on chromosome 4p16.

Figure 8 shows the plot of the empirical transmitted statistic over the region ranging from 2.7 cM to 175 cM. We found the leftmost region was most informative, so we focused on this region as displayed in Figure 9. It is clear that the most informative region is from 4 cM to 6 cM as shown in Figure 10 so we plotted the narrower region from 4.5 cM to 5cM in Figure 11, this smaller region includes only one highest peak which meets our model assumption.

Further, since the SNPs markers are in LD, we selected some of the tag SNPs to conduct the linkage mapping. The SNPs around 4.7cM (see Figure 12) include: (1) rs9995063, rs4689885, rs11728302, rs10027615, rs10012509, rs10428352, rs6446666, rs11733672, rs11725796,and rs10937875--the ten markers located from 4.674158 cM to 4.731674 cM for the Taiwanese and all populations; (2) SNPs rs9995063, rs4689885, rs11728302, rs10027615, rs10012509, rs10428352, rs6446666, rs11725796, and rs10937875--the nine markers located from 4.674158 cM to 4.731674 cM for the Korean population; (3) SNPs rs7682040, rs9654059, rs12504020, rs7681821, rs3910659, rs7437213, rs9995063, rs4689885, rs11728302, rs10027615, and rs10012509--the eleven markers located from 4.634028 cM to 4.700255 cM for the Marylander population; (4) SNPs rs11728302, rs10027615, rs10012509, rs10428352, rs6446666, rs11733672, rs11725796, rs10937875, rs2165431, rs4689907, rs838958, rs6840368, rs6826063, and rs6824609--the fourteen markers from 4.683682 cM to 4.771068 cM for Singaporean population. We estimated the disease locus τ for oral cleft, the corresponding standard errors for the estimates, the p-value of β and the 95% coverage probability for τ by incorporating different covariates through the proposed parametric and nonparametric models.

We applied three methods in estimating τ. One is the original model without a covariate proposed by Liang et al. (2001b), the other two are our proposed parametric and nonparametric approaches with covariates incorporated. The estimated results were listed in Table 21 ~ Table 25 for the four combined population, Korea, Maryland, Singapore, and Taiwan, respectively. The data and the fitted curves were also demonstrated in Figure 13 ~ Figure 26.

Since the data did not include any quantitative covariates, we employed the following 5 qualitative covariates in localizing of the disease locus: GENDER (gender of proband, male=1, female=2,), CLF_(father) (condition of father, affected=1, unaffected=0) , CLF_(mother) (condition of mother, affected=1, unaffected=0), SMOKE and DRINK (yes=1, no=0) (The data from Singapore also include information of having taken vitamin or not). Some of the drinking and smoking data were missing in Taiwanese population, hence, there were only 104 out of 172 Taiwanese were included in the analyses when the incorporated covariate was drinking or smoking status.

The results showed that our proposed approaches were mostly more efficient than the original approach where no covariate was incorporated. In addition, the estimates from our approaches were more precise (bias was smaller) unless the covariate is not associated with the oral cleft syndromes in a specific population. Besides, the nonparametric approach seemed to be more efficient than the parametric approach. Sometimes an irrelevant covariate not only makes the estimate less efficient, but also induces higher bias for the estimate of τ. For example, the factor, SMOKE, in populations of Taiwan and Singapore was not helpful in estimating the disease locus.

It is worth noting that the covariate POPULATION helped improved the disease localization greatly in the combined (four) populations (see Table 21 and Figure 17). The result suggested that the genetic effects in the four separate populations were different. The estimate for τ was at around 4.7 cM, and Korea has the substantial linkage effect than other populations. The order of significance magnitudes of β (the population difference on the effect from the estimated disease locus) was Korea, Taiwan, Maryland, and Singapore as illustrated in Figure 12. Moreover, Table 21 revealed that adding POPULATION as a covariate increased the efficiency of estimating the disease locus (also see Figure 17).

Finally, we find the patterns of the transmitted targeted alleles were similar in Korea and Taiwan, but different from the other two populations, so we tried to combine the data of Korea and Taiwan to see if the efficiency gets improved (Table 26). The real data and the fitted transmitted frequencies of the targeted alleles from the original and the proposed approaches were illustrated in Figure 27 ~ 29. Comparing the results from Table 26 to Table 21, we found the relative efficiency of estimating τ did get improved when including the data of two populations--Korea and Taiwan only than including all populations.

6. Discussions

In the modern society, many families have only one child, so case-parent trios data are easier to collect than affected sibling pairs data except for some late onset diseases. Using case-parent trios data, one can estimate the disease susceptibility locus τ precisely and robustly by the preferential-transmission statistic $\boldsymbol{Y(t)}$ proposed by Liang et al. (2001b) through the generalized estimating equation approach (GEE, Liang and Zeger, 1986). But when the number of sample size is small (rare disorder) or when the preferential-transmission statistic at τ (empirical $\boldsymbol{C}$) is near 0, the estimation may not be accurate, sometimes it does not converge due to the heterogeneous genetic effects at τ even. Most complex diseases are induced by interactions between multiple genetic and environmental factors, incorporating those factors into the LD mapping can add more information into the mapping and therefore is very likely to increase the efficiency in estimating the disease locus. In the present study, we proposed two multipoint fine-mapping methods that incorporate covariates into the LD fine-mapping approach proposed by Liang et al. (2001b). The expected preferential-transmission statistic at τ (denoted by $\boldsymbol{C}$) is modeled as a parametric or nonparametric function of covariates, and all the parameters were estimated through the GEE approach. By testing whether the covariate is associated with the estimated disease locus, we can explore the underlying genetic mechanism and etiology of the disease. This information is very helpful on disease preventions and controls for public health.

Further, we illustrated this approach by applying the proposed methods to real data of a case-parent trios study of oral cleft and found significant covariate effects on the locus identified at 4.7 cM on chromosome 4p16 in populations of Korea and Taiwan. Hence, incorporating covariates associated with the disease did improve the efficiency in estimating the disease locus. These results showed that the proposed approaches can not only help researchers to estimate the disease locus more efficiently, but also to identify risk factors associated with diseases.

References

Abecasis GR, Cardon LR, Cookson WO. (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66: 279-292.

Abecasis GR, Cherney SS, Cookson WO, and Cardon LR. (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97-101.

Allison DB. (1997) Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet 60: 676-690.

Bickeböller H, Clerget-Darpoux F. (1995) Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. Genet Epidemiol 12: 865-870.

Chiou JM, Liang KY, and Chiu YF. (2005) Multipoint linkage mapping using sibpairs: Non-parametric estimation of trait effects with quantitative covariates. Genet Epidemiol 28: 58-69.

Clayton D, Jones H. (1999) Transmission/disequilibrium tests for extended marker haplotypes. Am J Hum Genet 65: 1161-1169.

Clayton D. (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65: 1170-1177.

Cleves MA, Olson JM, and Jacobs KB. (1997) Exact transmission-disequilibrium tests with multiallelic markers. Genet Epidemiol 14: 337-347.

Dudbridge F, Koeleman BP, Todd JA, Clayton DG. (2000) Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. Am J Hum Genet 66: 2009-2012.

Fan J, Gijbels I. (1996) Local polynomial modelling and its applications. London: Chapman and Hall.

Fulker DW, Cherny SS, Sham PC, and Hewitt JK. (1999) Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64: 259-267.

Gauderman WJ. (2003) Candidate gene association analysis for a quantitative trait, using parent-offspring trios. Genet Epidemiol 25: 327-338.

Glidden DV, Liang KY, Chiu YF, and Pulver AE. (2003) Multipoint affected sibpair linkage methods for localizing susceptibility genes of complex diseases. Genet Epidemiol 24: 107-117.

Haldane JBS. (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8: 299-309.

Haseman JK, Elston RC. (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3-19.

Horvath S, Laird NM. (1998) A discordant-sibship test for disequilibrium and linkage: No need for parental data. Am J Hum Genet 63: 1886-1897.

Kaplan NL, Martin ER, and Weir BS. (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 60: 691-702.

Knapp M. (1999) A note on power approximations for the transmission/disequilibrium test. Am J Hum Genet 64: 1177-1185.

Knapp M, Becker T. (2004) Letters to the editor. "Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT)" Am J Hum Genet 74: 589-591.

Laird NM, Blacker D, Wilcox M. (1998) The sib transmission/disequilibrium test is a Mantel-Haenszel test. Am J Hum Genet 63: 1915-1916.

Lazzeroni LC, Lange K. (1998) A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered 48: 67-81.

Liang KY, Zeger SL. (1986) Longitudinal data analysis using generalized linear models. Biometrika 73: 13-22.

Liang KY, Chiu YF, and Beaty TH. (2001a) A robust identity-by-descent procedure using affected sib pairs: Multipoint mapping for complex diseases. Hum Hered 51: 64-78.

Liang KY, Chiu YF, and Beaty TH. (2001b) Multipoint analysis using affected sib pairs: Incorporating linkage evidence from unlinked regions. Genet Epidemiol 21: 105-122.

Liu Y, Tritchler D, Bull SB. (2002) A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits. Genet Epidemiol 22: 26-40.

Lunetta KL, Faraone SV, Biederman J, and Laird NM. (2000) Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. Am J Hum Genet 66: 605-614.

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am. J Hum Genet 61: 439-448.

Martin ER, Monks SA, Warren LL, and Kaplan NL. (2000) A test for linkage and association in general pedigrees: The pedigree disequilibrium test. Am J Hum Genet 67: 146-154.

Rabinowitz D. (1997) A transmission disequilibrium test for quantitative trait loci. Hum Hered 47: 342-350.

Risch N, Merikangas K. (1996) The future of genetic studies of complex human disease. Science 273: 1516-1517.

Ruppert D, Wand MP. (1994) Multivariate locally weighted least squares regression.

The Annals of Statistics 22: 1346-1370.

Schaid DJ. (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13: 423-449.

Schmidt S, Qin X, Schmidt MA, Martin ER, and Hauser ER. (2007) Interpreting analyses of continuous covariates in affected sibling pair linkage studies. Genet Epidemiol 31: 541-552.

Sham PC, Curtis D. (1995) An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. Ann Hum Genet 59:323-336.

Sham PC, Purcell S, Cherny SS, and Abecasis GR. (2002) Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet 71: 238-253.

Sokal RR, Rohlf FJ. (1969) Biometry. WH Freeman, San Francisco.

Spielman RS, McGinnis RE, and Ewens WJ. (1993) Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506-516.

Spielman RS, Ewes WJ. (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59: 983-989.

Spielman RS, Ewens WJ. (1998) A sibship test for linkage in the presence of association: The sib Transmission/Disequilibrium test. Am J Hum Genet 62: 450-458.

Sull JW, Liang KY, Hetmanski JB, Fallin MD, Ingersoll RG, Park J, Wu-Chou YH, Chen PK, Chong SS, Cheah F, Yeow V, Park BY, Jee SH, Jabs EW, Redett R, Jung E, Ruczinski I, Scott AF, and Beaty TH. (2008) Differential parental transmission of markers in RUNX2 among cleft case-parent trios from four populations. Genet Epidemiol 32: 1-9.

Terwilliger JD, and Ott, J. (1992) A haplotype-based ‘haplotype relative risk’ approach to detecting allelic associations. Hum Hered. 42: 337-346.

Terwilliger JD. (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56: 777-787.

Wheeler E, Cordell HJ. (2007) Quantitative trait association in parent offspring trios: Extension of case/pseudocontrol method and comparison of prospective and retrospective approaches. Genet Epidemiol 31: 813-833.

Wicks J. (2000) Exploiting excess sharing: A more powerful test of linkage for affected sib pairs than the transmission/disequilibrium test. Am J Hum Genet 66: 2005-2008.

Zhang S, Sha Q, Chen HS, Dong J, and Jiang R. (2003) Transmission/Disequilibrium test based on haplotype sharing for tightly linked markers. Am J Hum Genet 73:

566-579.

Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, and Kidd KK (2000) Transmission/disequilibrium tests using multiple tightly linked markers. Am J Hum Genet 67: 936-946.

Table 1. Impact of number of generations on estimating a disease locus

Gene-Environment disease model: C=0.1879

Number of Generation (N)		τ	Bias	R.E.	β	P-value	95% Coverage Probability
100	**Parametric**	0.45±0.044	0.0008	1.56	0.2944±0.0533	<1.0e-6	0.96
	Nonparametric	0.45±0.044	0.0005	1.55			0.95
	Original	0.45±0.055	-0.0018				0.95
150	**Parametric**	0.45±0.034	0.0010	1.61	0.2929±0.0551	<1.0e-6	0.96
	Nonparametric	0.45±0.035	0.0008	1.50			0.96
	Original	0.45±0.043	-0.0021				0.95
200	**Parametric**	0.45±0.028	0.0012	1.67	0.2918±0.0568	<1.0e-6	0.96
	Nonparametric	0.45±0.029	0.0010	1.62			0.96
	Original	0.45±0.036	-0.0010				0.95

Threshold disease model: C=0.2744

Number of Generation (N)		τ	Bias	R.E.	β	P-value	95% Coverage Probability
100	**Parametric**	0.45±0.031	-0.0003	1.03	0.4873±0.1802	0.006835	0.96
	Nonparametric	0.45±0.032	-0.0003	0.98			0.97
	Original	0.45±0.032	-0.0007				0.97
150	**Parametric**	0.45±0.024	0.0008	1.05	0.4850±0.1831	0.008081	0.98
	Nonparametric	0.45±0.025	0.0007	0.99			0.97
	Original	0.45±0.025	0.0004				0.97
200	**Parametric**	0.45±0.021	-0.0002	1.05	0.4838±0.1858	0.009200	0.97
	Nonparametric	0.45±0.021	-0.0003	0.98			0.98
	Original	0.45±0.021	-0.0004				0.97

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without

Table 2. Impact of disease allele frequency on estimating a disease locus

Gene-Environment disease model:

Pr(H_τ)		C	τ	Bias	R.E.	β	P-value	95% Coverage Probability
0.05	**Parametric**	**0.1262**	0.45±0.044	0.0021	2.18	0.2214±0.0457	0.000001	0.95
	Nonparametric		0.45±0.048	0.0026	1.78		0.000091	0.91
	Original		0.45±0.064	-0.0008				0.94
0.1	**Parametric**	**0.1879**	0.45±0.034	0.0010	1.61	0.2929±0.0551	<1.0e-6	0.96
	Nonparametric		0.45±0.035	0.0008	1.50		0.000003	0.96
	Original		0.45±0.043	-0.0021				0.95
0.2	**Parametric**	**0.2331**	0.45±0.027	<1.0e-4	1.12	0.3579±0.0794	0.000002	0.96
	Nonparametric		0.45±0.028	-0.0002	1.09		0.000008	0.96
	Original		0.45±0.030	0.0006				0.97

Threshold disease model:

Pr(H_τ)		C	τ	Bias	R.E.	β	P-value	95% Coverage Probability
0.05	**Parametric**	**0.1683**	0.45±0.046	0.0004	1.09	0.3565±0.1412	0.011575	0.94
	Nonparametric		0.45±0.046	0.0002	1.10			0.94
	Original		0.45±0.048	0.0009				0.94
0.1	**Parametric**	**0.2744**	0.45±0.024	0.0008	1.05	0.4850±0.1831	0.008081	0.98
	Nonparametric		0.45±0.025	0.0007	0.99			0.97
	Original		0.45±0.025	0.0004				0.97
0.2	**Parametric**	**0.3851**	0.45±0.016	-0.0005	1.04	0.6527±0.2972	0.028058	0.99
	Nonparametric		0.45±0.016	-0.0005	1.00			0.99
	Original		0.45±0.016	-0.0004				0.99

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without

Table 3. Impact of sample sizes on estimating the disease locus

Gene-Environment disease model: C=0.1879

Sample Size		τ	Bias	R.E.	β	P-value	95% Coverag Probability
50	**Parametric**	0.45±0.071	-0.0025	1.37	0.3149±0.1222	0.009969	0.93
	Nonparametric	0.45±0.068	-0.0006	1.46			0.89
	Original	0.45±0.083	-0.0073				0.91
200	**Parametric**	0.45±0.034	0.0010	1.61	0.2929±0.0551	<1.0e-6	0.96
	Nonparametric	0.45±0.035	0.0008	1.50			0.96
	Original	0.45±0.043	-0.0021				0.95
400	**Parametric**	0.45±0.022	0.0003	1.63	0.2917±0.037	<1.0e-6	0.97
	Nonparametric	0.45±0.023	<1.0e-4	1.59			0.97
	Original	0.45±0.028	0.0004				0.97
1000	**Parametric**	0.45±0.013	-0.0002	1.48	0.2920±0.0251	<1.0e-6	0.98
	Nonparametric	0.45±0.013	-0.0002	1.53			0.98
	Original	0.45±0.016	0.0006				0.97

Threshold disease model: C=0.2744

Sample Size		τ	Bias	R.E.	β	P-value	95% Coverage Probability
50	**Parametric**	0.45±0.058	0.0008	1.15	0.5187±0.3964	0.190659	0.94
	Nonparametric	0.45±0.058	0.0007	1.16			0.92
	Original	0.45±0.062	-0.0004				0.94
200	**Parametric**	0.45±0.024	0.0008	1.05	0.4850±0.1831	0.008081	0.98
	Nonparametric	0.45±0.025	0.0007	0.99			0.97
	Original	0.45±0.025	0.0004				0.97
400	**Parametric**	0.45±0.017	0.0005	1.04	0.4753±0.1287	0.000222	0.98
	Nonparametric	0.45±0.017	0.0005	0.99			0.98
	Original	0.45±0.017	0.0007				0.98
1000	**Parametric**	0.45±0.010	0.0004	1.03	0.4719±0.0806	<1.0e-6	0.97
	Nonparametric	0.45±0.010	0.0004	1.01			0.97
	Original	0.45±0.010	0.0003				0.97

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

Table 4. Impact of markers' targeted allele frequencies on estimating a disease locus

Gene-Environment disease model:		**C=0.1879**			**Pr(H_τ)=0.1**		
Pr(H_t)		τ	**Bias**	**R.E.**	β	**P-value**	**95% Coverage Probability**
0.1	**Parametric**	0.45±0.034	0.0010	1.61	0.2929±0.0551	<1.0e-6	0.96
	Nonparametric	0.45±0.035	0.0008	1.50			0.96
	Original	0.45±0.043	-0.0021				0.95
0.2	**Parametric**	0.45±0.044	0.0011	1.80	0.2956±0.0627	0.000002	0.96
	Nonparametric	0.45±0.043	0.0015	1.88			0.95
	Original	0.45±0.060	-0.0015				0.95
0.3	**Parametric**	0.45±0.054	0.0002	1.90	0.3010±0.0718	0.000027	0.94
	Nonparametric	0.45±0.056	-0.0003	1.71			0.93
	Original	0.45±0.074	0.0002				0.94
random	**Parametric**	0.45±0.042	0.0012	1.64	0.2942±0.0614	0.000002	0.96
	Nonparametric	0.45±0.043	0.0014	1.60			0.94
	Original	0.45±0.054	-0.0021				0.93
Threshold disease model:		**C=0.2744**			**Pr(H_τ)=0.1**		
Pr(H_t)		τ	**Bias**	**R.E.**	β	**P-value**	**95% Coverage Probability**
0.1	**Parametric**	0.45±0.024	0.0008	1.05	0.4850±0.1831	0.008081	0.98
	Nonparametric	0.45±0.025	0.0007	0.99			0.97
	Original	0.45±0.025	0.0004				0.97
0.2	**Parametric**	0.45±0.034	-0.0007	1.09	0.4913±0.2018	0.014905	0.95
	Nonparametric	0.45±0.036	-0.0007	1.03			0.96
	Original	0.45±0.036	-0.0012				0.96
0.3	**Parametric**	0.45±0.045	0.0004	1.09	0.4958±0.2170	0.022336	0.94
	Nonparametric	0.45±0.047	-0.0003	1.01			0.94
	Original	0.45±0.047	0.0003				0.95
random	**Parametric**	0.45±0.032	0.0006	1.06	0.4865±0.1923	0.011391	0.97
	Nonparametric	0.45±0.033	0.0014	0.99			0.97
	Original	0.45±0.032	0.0004				0.97

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without

Table 5. Impact of markers' density on estimating a disease locus (τ=0.45 with 10 markers, τ=0.475 with 20 markers)

Gene-Environment disease model: C=0.1879

No. marker		τ	Bias	R.E.	β	P-value	95% Coverage Probability
10	**Parametric**	0.45±0.034	0.0010	1.61	0.2929±0.0551	<1.0e-6	0.96
	Nonparametric	0.45±0.035	0.0008	1.50			0.96
	Original	0.45±0.043	-0.0021				0.95
20	**Parametric**	0.475±0.022	-0.0003	1.66	0.2933±0.0495	<1.0e-6	0.96
	Nonparametric	0.475±0.022	-0.0004	1.64			0.96
	Original	0.476±0.029	0.0010				0.96

Threshold disease model: C=0.2744

No. marker		τ	Bias	R.E.	β	P-value	95% Coverage Probability
10	**Parametric**	0.45±0.024	0.0008	1.05	0.4850±0.1831	0.008081	0.98
	Nonparametric	0.45±0.025	0.0007	0.99			0.97
	Original	0.45±0.025	0.0004				0.97
20	**Parametric**	0.475±0.017	-0.0004	1.08	0.4747±0.1794	0.008132	0.96
	Nonparametric	0.475±0.017	-0.0004	1.05			0.97
	Original	0.475±0.018	-0.0004				0.97

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without

Table 6. Impact of ε on estimating a disease locus

Gene-Environment disease model: C=0.1879

ε		τ	Bias	R.E.	β	P-value	95% Coverage Probability	times
0.01	**Parametric**	0.45±0.034	0.0007	1.53	0.2934±0.0553	<1.0e-6	0.95	974
	Nonparametric	0.45±0.033	0.0011	1.65			0.95	987
	Original	0.45±0.042	-0.0020				0.95	959
0.05	**Parametric**	0.45±0.034	0.0010	1.61	0.2929±0.0551	<1.0e-6	0.96	998
	Nonparametric	0.45±0.035	0.0008	1.50			0.96	999
	Original	0.45±0.043	-0.0021				0.95	997
0.09	**Parametric**	0.45±0.041	0.0008	0.99	0.2927±0.0551	<1.0e-6	1.00	963
	Nonparametric	0.45±0.038	0.0008	1.56			1.00	827
	Original	0.45±0.050	-0.0020				1.00	945

Threshold disease model: C=0.2744

ε		τ	Bias	R.E.	β	P-value	95% Coverage Probability	times
0.01	**Parametric**	0.45±0.024	0.0009	1.07	0.4870±0.1829	0.007741	0.96	990
	Nonparametric	0.45±0.025	0.0009	0.98			0.97	993
	Original	0.45±0.024	0.0005				0.96	989
0.05	**Parametric**	0.45±0.024	0.0008	1.05	0.4850±0.1831	0.008081	0.98	1000
	Nonparametric	0.45±0.025	0.0007	0.99			0.97	1000
	Original	0.45±0.025	0.0004				0.97	1000
0.09	**Parametric**	0.45±0.025	0.0006	1.05	0.4846±0.1828	0.008035	1.00	990
	Nonparametric	0.45±0.025	0.0005	1.00			1.00	876
	Original	0.45±0.025	0.0004				1.00	983

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without

Table 7. Impact of β_0 (the risk of referent population) on estimating a disease locus

Gene-Environment disease model:

Beta0		C	τ	Bias	R.E.	β	P-value	95% Coverage Probability
ln(0.001)	Parametric	0.1986	0.45±0.032	0.0006	1.43	0.2974±0.0551	<1.0e-6	0.96
	Nonparametric		0.45±0.034	0.0005	1.31			0.95
	Original		0.45±0.038	0.0005				0.96
ln(0.01)	Parametric	0.1879	0.45±0.034	0.0010	1.61	0.2929±0.0551	<1.0e-6	0.96
	Nonparametric		0.45±0.035	0.0008	1.50			0.96
	Original		0.45±0.043	-0.0021				0.95
ln(0.1)	Parametric	0.1343	0.45±0.041	-0.0002	2.27	0.2743±0.0540	<1.0e-6	0.96
	Nonparametric		0.45±0.044	0.0014	1.96			0.93
	Original		0.45±0.062	0.0007				0.95

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

Table 8. Impact of β_1 and β_2 (genetic effect) on estimating a disease locus

Gene-Environment disease model:

Beta1,Beta2		C	τ	Bias	R.E.	β	P-value	95% Coverage Probability
ln(9),ln(5)	Parametric	0.1879	0.45±0.034	0.0010	1.61	0.2929±0.0551	<1.0e-6	0.96
	Nonparametric		0.45±0.035	0.0008	1.50			0.96
	Original		0.45±0.043	-0.0021				0.95
ln(29),ln(15)	Parametric	0.3031	0.45±0.022	0.0006	1.10	0.2775±0.0582	0.000002	0.97
	Nonparametric		0.45±0.023	0.0007	1.01			0.97
	Original		0.45±0.023	0.0004				0.98
ln(49),ln(25)	Parametric	0.3375	0.45±0.019	0.006	1.09	0.2542±0.0607	<1.0e-6	0.96
	Nonparametric		0.45±0.020	0.0008	1.03			0.93
	Original		0.45±0.020	0.0005				0.95

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

Table 9. Impact of β_3 (environment effect) on estimating a disease locus

Gene-Environment disease model:

Beta3		C	τ	Bias	R.E.	β	P-value	95% Coverage Probability
ln(2)	**Parametric**	0.1879	0.45±0.034	0.0010	1.61	0.2929±0.0551	<1.0e-6	0.96
	Nonparametric		0.45±0.035	0.0008	1.50			0.96
	Original		0.45±0.043	-0.0021				0.95
ln(5)	**Parametric**	0.1472	0.45±0.039	-0.0019	2.06	0.2808±0.0555	<1.0e-6	0.95
	Nonparametric		0.45±0.041	-0.0018	1.89			0.94
	Original		0.45±0.056	0.0014				0.94
ln(10)	**Parametric**	0.1059	0.45±0.045	-0.0001	2.70	0.2642±0.0528	0.000001	0.95
	Nonparametric		0.45±0.050	0.0006	2.16			0.92
	Original		0.45±0.074	0.0037				0.93

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without

Table 10. Impact of the additive genetic effect "a" on estimating a disease locus

Gene-Environment disease model: C=0.1879

a		τ	Bias	R.E.	β	P-value	95% Coverage Probability
1	**Parametric**	0.45±0.034	0.0010	1.61	0.2929±0.0551	<1.0e-6	0.96
	Nonparametric	0.45±0.035	0.0008	1.50			0.96
	Original	0.45±0.043	-0.0021				0.95
5	**Parametric**	0.45±0.024	0.0007	3.18	0.2132±0.0215	<1.0e-6	0.97
	Nonparametric	0.45±0.024	0.0006	3.16			0.97
10	**Parametric**	0.45±0.023	0.0004	3.53	0.1156±0.0105	0.000001	0.95
	Nonparametric	0.45±0.024	0.0005	3.21			0.92

Threshold disease model:

a		C	τ	Bias	R.E.	β	P-value	95% Coverage Probability
0.5	**Parametric**	0.1217	0.45±0.064	-0.0021	1.12	0.2143±0.1809	0.236275	0.93
	Nonparametric		0.45±0.063	-0.0029	1.14			0.90
	Original		0.45±0.067	-0.0023				0.94
1	**Parametric**	0.2744	0.45±0.024	0.0008	1.05	0.4850±0.1831	0.008081	0.98
	Nonparametric		0.45±0.025	0.0007	0.99			0.97
	Original		0.45±0.025	0.0004				0.97
1(d=0.5)	**Parametric**	0.3529	0.45±0.019	0.0003	1.02	0.3000±0.1622	0.064350	0.98
	Nonparametric		0.45±0.020	0.0002	0.94			0.99
	Original		0.45±0.019	0.0003				0.98
2	**Parametric**	0.4672	0.45±0.014	0.0003	0.97	0.5727±0.1418	0.000054	0.99
	Nonparametric		0.45±0.014	0.0003	0.98			0.99
	Original		0.45±0.014	0.0002				0.99

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

Table 11. Impact of prevalence rates on estimating a disease locus

Threshold disease model:

prevalence		C	τ	Bias	R.E.	β	P-value	95% Coverage Probability
0.01	**Parametric**	**0.4072**	0.45±0.017	0.0003	1.01	0.5532±0.2485	0.025994	0.98
	Nonparametric		0.45±0.017	0.0004	0.95			0.98
	Original		0.45±0.017	0.0004				0.98
0.05	**Parametric**	**0.2744**	0.45±0.024	0.0008	1.05	0.4850±0.1831	0.008081	0.98
	Nonparametric		0.45±0.025	0.0007	0.99			0.97
	Original		0.45±0.025	0.0004				0.97
0.1	**Parametric**	**0.2117**	0.45±0.035	0.0011	1.11	0.4339±0.1548	0.005060	0.96
	Nonparametric		0.45±0.036	0.0016	1.03			0.97
	Original		0.45±0.037	0.0005				0.95

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

Table 12. Fixed penetrance disease model

f0,f1,f2	$P(H_\tau)$		C	τ	Bias	R.E.	β	P-value	95% Coverage Probability
0.67,0.05,0.007	**0.05**	**Parametric**	**0.2678**	0.45±0.023	0.0005	1.28	0.3701±0.0501	<1.0e-6	0.99
		Nonparametric		0.45±0.023	0.0006	1.36			0.97
		Original		0.45±0.026	0.0006				0.98
0.67,0.05,0.007	**0.1**	**Parametric**	**0.4241**	0.45±0.015	-0.0005	1.12	0.5200±0.0690	<1.0e-6	0.98
		Nonparametric		0.45±0.015	-0.0005	1.14			0.98
		Original		0.45±0.016	-0.0002				0.98
0.95,0.9,0.01	**0.05**	**Parametric**	**0.4157**	0.45±0.016	-0.0007	1.01	0.1388±0.0537	0.009811	0.97
		Nonparametric		0.45±0.017	-0.0009	0.93			0.98
		Original		0.45±0.016	-0.0008				0.97
0.95,0.9,0.01	**0.1**	**Parametric**	**0.4039**	0.45±0.028	-0.0001	1.02	0.1235±0.0635	0.051736	0.98
		Nonparametric		0.45±0.029	<1.0e-4	0.97			0.98
		Original		0.45±0.036	-0.0001				0.98

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

Table 13. Impact of the QTL's position on estimating a disease locus

Gene-Environment disease model:		C=0.1879			Additive model		
Covariate		τ	**Bias**	**R.E.**	β	**P-value**	**95% Coverage Probability**
0.45(tau)	**Parametric**	0.45±0.034	0.0010	1.61	0.2929±0.0551	<1.0e-6	0.96
	Nonparametric	0.45±0.035	0.0008	1.50			0.96
	Original	0.45±0.043	-0.0021				0.95
0.5cM	**Parametric**	0.46±0.026	0.0129	2.62	0.2804±0.0578	0.000001	0.97
	Nonparametric	0.46±0.026	0.0137	2.63			0.92
0.7cM	**Parametric**	0.48±0.042	0.0297	1.04	0.2071±0.0517	0.000063	0.85
	Nonparametric	0.48±0.043	0.0250	0.97			0.91
0.9cM	**Parametric**	0.48±0.054	0.0279	0.61	0.1518±0.0489	0.001891	0.93
	Nonparametric	0.47±0.048	0.0195	0.80			0.92
1.1cM	**Parametric**	0.46±0.045	0.0076	0.88	0.1147±0.0503	0.022595	0.95
	Nonparametric	0.46±0.042	0.0070	1.03			0.93
1.3cM	**Parametric**	0.45±0.040	-0.0006	1.11	0.0840±0.0543	0.122233	0.96
	Nonparametric	0.45±0.042	-0.0002	1.05			0.95
unlinked	**Parametric**	0.45±0.041	-0.0026	1.10	-0.0014±0.0520	0.979139	0.95
	Nonparametric	0.45±0.039	-0.0035	1.17			0.96
Dominant:							
0.5cM	**Parametric**	0.47±0.020	0.0156	4.62	0.3161±0.0522	<1.0e-6	0.96
	Nonparametric	0.47±0.021	0.0164	3.95			0.95
0.7cM	**Parametric**	0.49±0.039	0.0425	1.17	0.2314±0.0436	0.000063	0.85
	Nonparametric	0.49±0.043	0.0391	1.00			0.91
Recessive:							
0.5cM	**Parametric**	0.46±0.036	0.0057	1.43	0.1408±0.0584	0.015895	0.96
	Nonparametric	0.46±0.031	0.0092	1.94			0.90
0.7cM	**Parametric**	0.46±0.043	0.0073	0.98	0.0954±0.0581	0.100280	0.94
	Nonparametric	0.46±0.042	0.0106	1.02			0.92

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without
QTL: quantitative trait locus, the quantitative trait refers to the covariate incorporated

Table 14. Impact of the QTL's position on estimating a disease locus

Threshold disease model:		C=0.2744			Additive model		
Covariate		τ	Bias	R.E.	β	P-value	95% Coverage Probability
0.45(τ)	**Parametric**	0.45±0.024	0.0008	1.05	0.4850±0.1831	0.008081	0.98
	Nonparametric	0.45±0.025	0.0007	0.99			0.97
	Original	0.45±0.025	0.0004	1.00			0.97
0.5cM	**Parametric**	0.46±0.020	0.0088	1.48	0.3194±0.620	<1.0e-6	0.97
	Nonparametric	0.46±0.020	0.0096	1.47			0.95
0.7cM	**Parametric**	0.47±0.028	0.0163	0.80	0.2407±0.0583	0.000036	0.91
	Nonparametric	0.46±0.028	0.0143	0.81			0.93
0.9cM	**Parametric**	0.46±0.029	0.0137	0.71	0.1782±0.0568	0.001705	0.95
	Nonparametric	0.46±0.028	0.0109	0.77			0.95
1.1cM	**Parametric**	0.45±0.026	0.0046	0.91	0.1393±0.0553	0.011720	0.98
	Nonparametric	0.45±0.025	0.0045	0.97			0.97
1.3cM	**Parametric**	0.45±0.025	0.0016	0.99	0.1042±0.0576	0.070589	0.98
	Nonparametric	0.45±0.026	0.0014	0.93			0.97
unlinked	**Parametric**	0.45±0.025	0.0001	0.97	0.0012±0.025	0.983060	0.97
	Nonparametric	0.45±0.026	-0.0001	0.94			0.98
Dominant:							
0.5cM	**Parametric**	0.46±0.018	0.0097	1.85	0.3358±0.0600	<1.0e-6	0.96
	Nonparametric	0.46±0.020	0.0106	1.51			0.97
0.7cM	**Parametric**	0.47±0.027	0.0226	0.82	0.2504±0.0514	0.000001	0.86
	Nonparametric	0.47±0.029	0.0205	0.75			0.92
Recessive:							
0.5cM	**Parametric**	0.46±0.024	0.0056	1.11	0.2051±0.0573	0.000340	0.98
	Nonparametric	0.46±0.027	0.0075	0.86			0.93
0.7cM	**Parametric**	0.46±0.027	0.0061	0.84	0.1430±0.0604	0.017877	0.96
	Nonparametric	0.46±0.026	0.0078	0.89			0.94

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without
QTL: quantitative trait locus, the quantitative trait refers to the covariate incorporated

Table 15. Impact of the QTL's position and genetic models of the quantitative trait on estimating a disease locus

Gene-Environment disease model:		C=0.1744			Dominant model		
Covariate		τ	Bias	R.E.	β	P-value	95% Coverage Probability
Dominant 0.45(τ)	Parametric	0.45±0.030	0.0014	2.42	0.3283±0.0468	<1.0e-6	0.94
	Nonparametric	0.45±0.034	0.0015	1.89			0.97
	Original	0.45±0.046	-0.0013				0.95
Dominant 0.5cM	Parametric	0.47±0.020	0.0167	5.17	0.3051±0.0502	<1.0e-6	0.97
	Nonparametric	0.47±0.022	0.0171	4.30			0.92
Additive 0.5cM	Parametric	0.46±0.027	0.0136	3.02	0.2610±0.0574	0.000005	0.96
	Nonparametric	0.46±0.027	0.0143	2.89			0.93
Recessive 0.5cM	Parametric	0.46±0.038	0.0056	1.47	0.1105±0.0597	0.064213	0.96
	Nonparametric	0.46±0.034	0.0094	1.87			0.91

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without
QTL: quantitative trait locus, the quantitative trait refers to the covariate incorporated

Table 16. Impact of the QTL's position and genetic models of the quantitative trait on estimating a disease locus (with an underlying recessive model)

Gene-Environment disease model:		C=0.1910			Recessive model		
Covariate		τ	Bias	R.E.	β	P-value	95% Coverage Probability
Recessive 0.45(τ)	Parametric	0.45±0.028	0.0022	2.08	0.3972±0.0533	<1.0e-6	0.98
	Nonparametric	0.45±0.024	0.0017	2.75			0.97
	without covariate	0.45±0.040	0.0012				0.96
Recessive 0.5cM	Parametric	0.46±0.026	0.0076	2.42	0.3727±0.0586	<1.0e-6	0.98
	Nonparametric	0.46±0.022	0.0079	3.37			0.94
Additive 0.5cM	Parametric	0.46±0.020	0.0127	3.83	0.4411±0..0627	<1.0e-6	0.97
	Nonparametric	0.46±0.019	0.0122	4.41			0.93
Dominant 0.5cM	Parametric	0.47±0.018	0.0156	5.19	0.4379±0.0670	<1.0e-6	0.95
	Nonparametric	0.47±0.018	0.0153	4.85			0.93

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without
QTL: quantitative trait locus, the quantitative trait refers to the covariate incorporated

Table 17. Impact of the QTL’s position and genetic models of the quantitative trait on estimating a disease locus (with an underlying dominant model)

Threshold disease model:		C=0.3951			Dominant model		
Covariate		τ	Bias	R.E.	β	P-value	95% Coverage Probability
Dominant	Parametric	0.45±0.017	0.0004	1.00	0.1057±0.1305	0.417662	0.98
0.45(τ)	Nonparametric	0.45±0.018	0.0005	0.94			0.98
	Original	0.45±0.017	0.0003				0.98
Dominant	Parametric	0.45±0.016	0.0022	1.12	0.1287±0.0620	0.037859	0.98
0.5cM	Nonparametric	0.45±0.017	0.0040	0.97			0.98
Additive	Parametric	0.45±0.016	0.0024	1.10	0.1359±0.0603	0.024182	0.98
0.5cM	Nonparametric	0.45±0.017	0.0035	1.04			0.97
Recessive	Parametric	0.45±0.016	0.0021	1.07	0.1092±0.0520	0.035712	0.98
0.5cM	Nonparametric	0.45±0.016	0.0042	1.13			0.95

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without
QTL: quantitative trait locus, the quantitative trait refers to the covariate incorporated

Table 18. Impact of the QTL’s position and genetic models of the quantitative trait on estimating a disease locus (with an underlying recessive threshold model)

Gene-Environment disease model:		C=0.1031			Recessive model		
Covariate		τ	Bias	R.E.	β	P-value	95% Coverage Probability
Recessive	Parametric	0.45±0.062	0.0003	1.81	0.5633±0.1777	0.001527	0.95
0.45(τ)	Nonparametric	0.45±0.056	0.0017	2.19			0.91
	Original	0.45±0.083	0.0044				0.92
Recessive	Parametric	0.46±0.043	0.0077	3.68	0.2809±0.0574	0.000001	0.97
0.5cM	Nonparametric	0.46±0.033	0.0090	6.23			0.91
Additive	Parametric	0.47±0.026	0.0181	10.12	0.3588±0.0589	<1.0e-6	0.97
0.5cM	Nonparametric	0.47±0.026	0.0170	10.27			0.87
Dominant	Parametric	0.47±0.020	0.0230	17.82	0.3620±0.0583	<1.0e-6	0.91
0.5cM	Nonparametric	0.47±0.021	0.0221	15.56			0.86

Original: without incorporating a covariate
R.E.: Relative efficiency from approaches with a covariate vs. without
QTL: quantitative trait locus, the quantitative trait refers to the covariate incorporated

Table 19. Impact of the genotype τ on estimating β

Gene-Environment disease model: C=0.1879 N=150

	τ(0.45cM)	parametric	0.5cM	parametric	0.7cM	parametric	0.9cM	parametric	unlinked	parametric
τ	0.4503	0.4510	0.4509	0.4629	0.4517	0.4797	0.4515	0.4779	0.4510	0.4474
S.E.(τ)	0.0233	0.0335	0.0221	0.0263	0.0226	0.0417	0.0228	0.0544	0.0230	0.0405
N	143.4489	158.6520	144.9560	187.0526	143.2264	154.3638	141.4642	130.4114	142.9235	163.4912
S.E.(N)	25.2158	33.8415	26.1238	37.3240	25.7249	33.0865	26.0691	38.8114	25.6668	31.5500
C	0.1969	0.2059	0.1967	0.2164	0.1960	0.2042	0.1953	0.1966	0.1965	0.2073
S.E.(C)	0.0284	0.0307	0.0281	0.0303	0.0281	0.0292	0.0281	0.0294	0.0291	0.0324
α	-0.2237	0.5523	-0.2034	0.5470	-0.1984	0.5157	-0.1994	0.4856	-0.2274	0.4216
S.E.(α)	0.0765	0.0708	0.0658	0.0682	0.0648	0.0669	0.0651	0.0663	0.0611	0.0682
β	0.0014	0.2929	0.0294	0.2804	0.0348	0.2071	0.0328	0.1518	0.0029	-0.0014
S.E.(β)	0.0457	0.0551	0.0413	0.0578	0.0392	0.0517	0.0374	0.0489	0.0379	0.0520
P-value(β)	0.976083	<1.0e-6	0.476265	0.000001	0.373807	0.000063	0.379500	0.001891	0.938493	0.979139
β1	1.1764		1.1596		1.1579		1.1598		1.1760	
S.E.(β1)	0.1388		0.1318		0.1313		0.1318		0.1236	
P-value(β1)	<1.0e-6		<1.0e-6		<1.0e-6		<1.0e-6		<1.0e-6	
β2	1.1685		1.1549		1.1532		1.1552		1.1792	
S.E.(β2)	0.1334		0.1308		0.1302		0.1309		0.1295	
P-value(β2)	<1.0e-6		<1.0e-6		<1.0e-6		<1.0e-6		<1.0e-6	
95% Coverage	0.98	0.96	0.98	0.97	0.98	0.85	0.98	0.93	0.97	0.95

Table 20. The proportions of each covariates' category for probands recruited from four populations from the oral cleft study

	Korea	Maryland	Singapore	Taiwan	Total
Gender (Male)	57%	55%	55%	59%	57%
Affected father (Y)	0%	2%	0%	1%	1%
Affected mother (Y)	0%	3%	0%	0%	1%
Mother Smoking (Y*)	0%	24%	8%	4%	10%
Mother Drinking (Y*)	2%	16%	5%	3%	7%
Vitamin	10%	81%			
Total	**42**	**103**	**66**	**172 (104*)**	**383 (315*)**

Table 21. Incorporating different covariates for four combined populations (Korea, Maryland, Singapore, and Taiwan) from the non-syndromic oral cleft study

Covariate	Method	τ	S.E.	R.E.	N	S.E.	C	β	P-value	
Sample Size: 383										
	Original	4.706	0.0022		7122.33	3466.80	0.1065			
POPULATION	**parametric**	4.707	0.0009	5.63	12966.68	3519.34	0.1633	1.1258	0.999993	**Korea**
								0.4228	0.999997	**Maryland**
								0.0558	1.000000	**Singapore**
								0.5153	0.999996	**Taiwan**
GENDER	**parametric**	4.707	0.0022	1.00	7119.21	3481.56	0.1066	0.0357	0.801413	
	nonparametric	4.707	0.0019	1.39	8702.04	2437.74	0.1270			
CLP(mother)	**parametric**	4.706	0.0022	0.97	6951.74	3438.09	0.1043	-0.1745	0.463335	
	nonparametric	4.707	0.0019	1.40	8782.35	2457.90	0.1276			
CLP(father)	**parametric**	4.707	0.002	1.21	7856.95	3509.16	0.1145	-0.1473	0.903556	
	nonparametric	4.708	0.0014	2.60	13277.77	4476.98	0.1558			
Sample Size: 315										
	Original	4.704	0.0038		4620.77	3912.21	0.0765			
SMOKE	**parametric**	4.704	0.0037	1.10	4610.41	3922.03	0.0763	-0.0565	0.744450	
	nonparametric	4.704	0.0039	0.98	4318.15	1911.41	0.0752			
DRINK	**parametric**	4.704	0.0039	0.99	4620.75	3910.38	0.0765	0.0030	0.985917	

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

Table 22. Incorporating different covariates for Korean population from the non-syndromic oral cleft study

Covariate	Method	τ	S.E.	R.E.	N	S.E.	C	β	P-value
Sample Size: 42									
	Original	4.708	0.0012		12458.35	4095	0.4523		
GENDER	parametric	4.708	0.0012	1.04	12608.6	4283.49	0.4563	0.2479	0.710034
	nonparametric	4.708	0.0011	1.15	12327.15	3248.3	0.4437		
DRINK	parametric	4.708	0.0012	1.01	12592.79	4149.29	0.456	-0.2123	0.505898
	nonparametric	4.708	0.0011	1.13	12931.6	3293.52	0.4537		

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

Table 23. Incorporating different covariates for population in Maryland from the non-syndromic oral cleft study

Covariate	Method	τ	S.E.	R.E.	N	S.E.	C	β	P-value
Sample Size: 103									
	Original	4.663	0.0025		8285.22	3493.80	0.2830		
GENDER	parametric	4.664	0.0026	0.89	8434.68	3390.84	0.2794	0.3610	0.373772
	nonparametric	4.663	0.0017	2.10	9463.42	3020.97	0.3254		
SMOKE	parametric	4.663	0.0028	0.79	6929.11	3390.84	0.2450	-0.4348	1.000000
	nonparametric	4.662	0.0018	1.94	8967.58	3020.97	0.3218		
DRINK	parametric	4.663	0.0025	0.97	8513.10	3612.09	0.2890	0.1531	0.800657
	nonparametric	4.663	0.0017	2.07	9770.79	3037.28	0.3355		
VATAMIN	parametric	4.663	0.0026	0.87	8094.21	3419.44	0.2778	-0.1636	1.000000
	nonparametric	4.662	0.0018	1.88	9594.13	2991.27	0.3328		

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

Table 24. Incorporating different covariates for Singaporean population from the non-syndromic oral cleft study

Covariate	Method	τ	S.E.	R.E.	N	S.E.	C	β	P-value
	Sample Size: 66								
GENDER	Original	4.727	0.0104		4487.01	1794.17	0.1295	0.5249	0.204076
	parametric	4.727	0.0114	0.84	3464.61	1868.28	0.2087		
	nonparametric	4.725	0.0065	2.61	4890.88	1211.29	0.2885		

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

Table 25. Incorporating different covariates for Taiwanese population from the non-syndromic oral cleft study

Covariate	Method	τ	S.E.	R.E.	N	S.E.	C	β	P-value
	Sample Size: 172								
GENDER	Original	4.709	0.0039		7505.51	3856.9	0.1351	-0.1017	1.000000
	parametric	4.708	0.0031	1.62	7787.11	3628.49	0.1391		
	nonparametric	4.708	0.0021	3.53	8299.81	2147.92	0.1527		
	Sample Size: 104								
DRINK	Original	4.712	0.0086		3813.78	2618.17	0.0955	-0.0599	1.000000
	parametric	4.712	0.0082	1.09	3853.75	2689.63	0.0961		
	nonparametric	4.711	0.0049	3.07	4545.06	1569.35	0.1104		

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

Table 26. Incorporating different covariates for Korean and Taiwanese population from the non-syndromic oral cleft study

Covariate	Method	τ	S.E.	R.E.	N	S.E.	C	β	P-value
Sample Size: 214									
	Original	4.708	0.0014		10822.54	3198.38	0.2160		
GENDER	parametric	4.708	0.0014	1.05	10750.36	3116.72	0.2147	-0.0728	1.000000
	nonparametric	4.708	0.0012	1.41	11091.76	2412.58	0.2229		
Sample Size: 146									
	Original	4.707	0.0022		7802.33	2936.51	0.1836		
SMOKE	parametric	4.708	0.0023	0.91	7708.71	2883.37	0.1821	0.1182	0.754960
	nonparametric	4.707	0.0016	1.87	8724.44	1812.26	0.2114		
DRINK	parametric	4.707	0.0019	1.36	8257.80	2908.89	0.1921	-0.2149	1.000000
	nonparametric	4.707	0.0015	2.17	9078.12	1824.40	0.2117		

Original: without incorporating a covariate

R.E.: Relative efficiency from approaches with a covariate vs. without

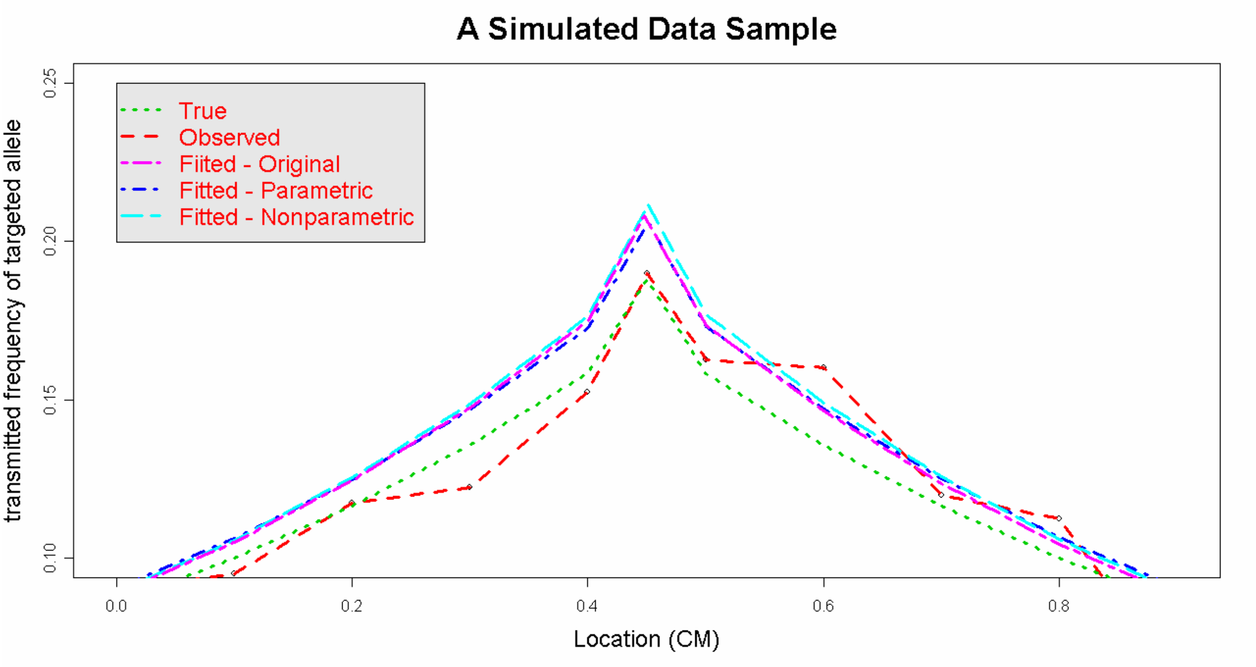

Figure 7. True, observed and fitted curves by the original approach, the proposed parametric approach and the proposed nonparametric approach

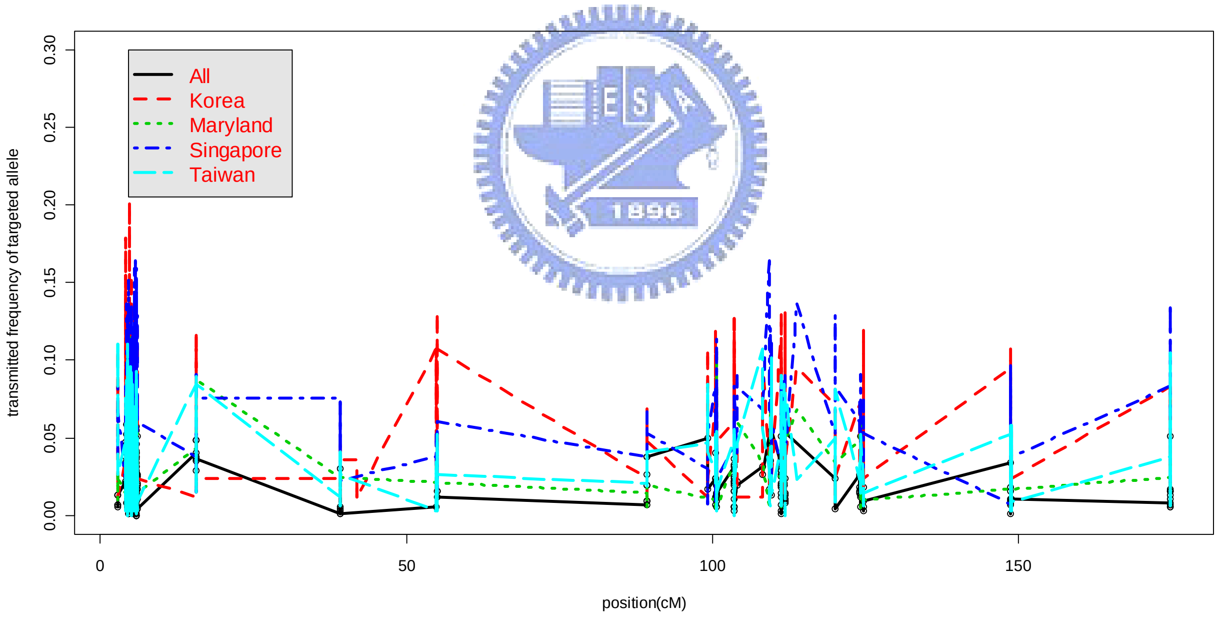

Figure 8. The transmitted statistic from 2.7 cM to 175 cM on chromosome 4p16 from oral clefts data (Sull et al. 2008)

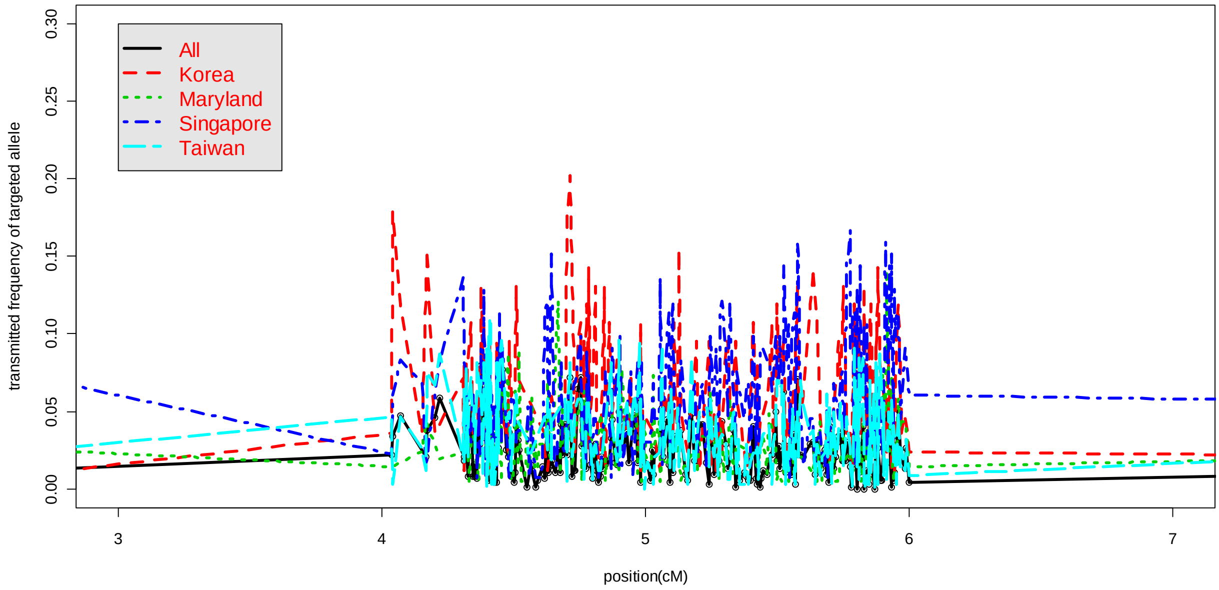

Figure 9. The transmitted statistic from 3 cM to 7 cM on chromosome 4p16 from oral clefts data (Sull et al. 2008)

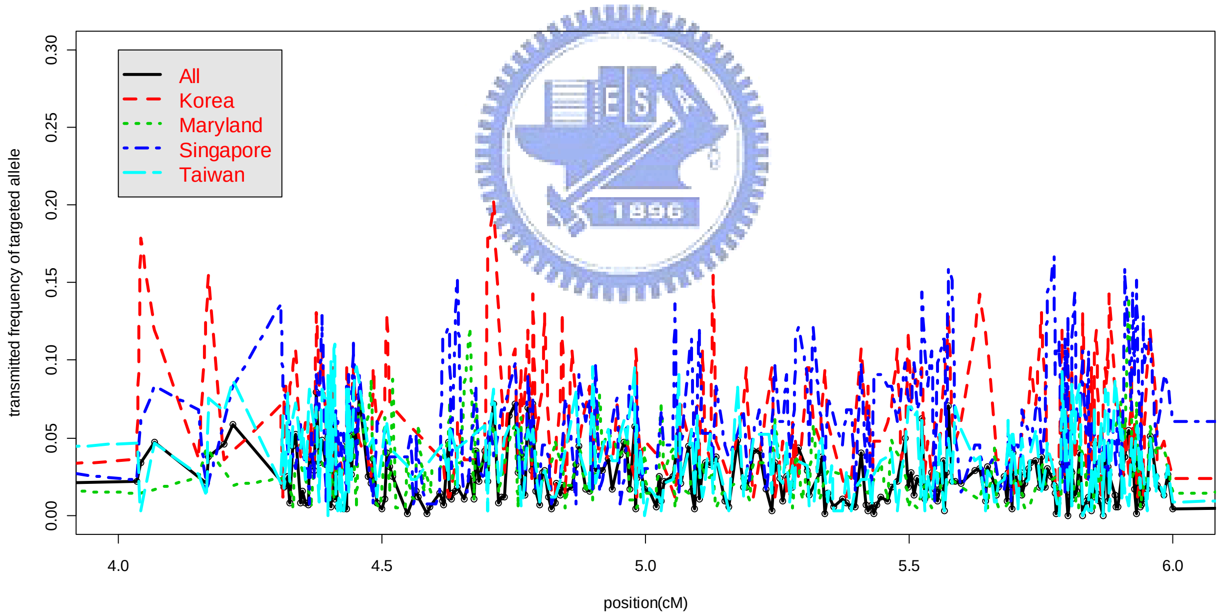

Figure 10. The transmitted statistic from 4 cM to 6 cM on chromosome 4p16 from oral clefts data (Sull et al. 2008)

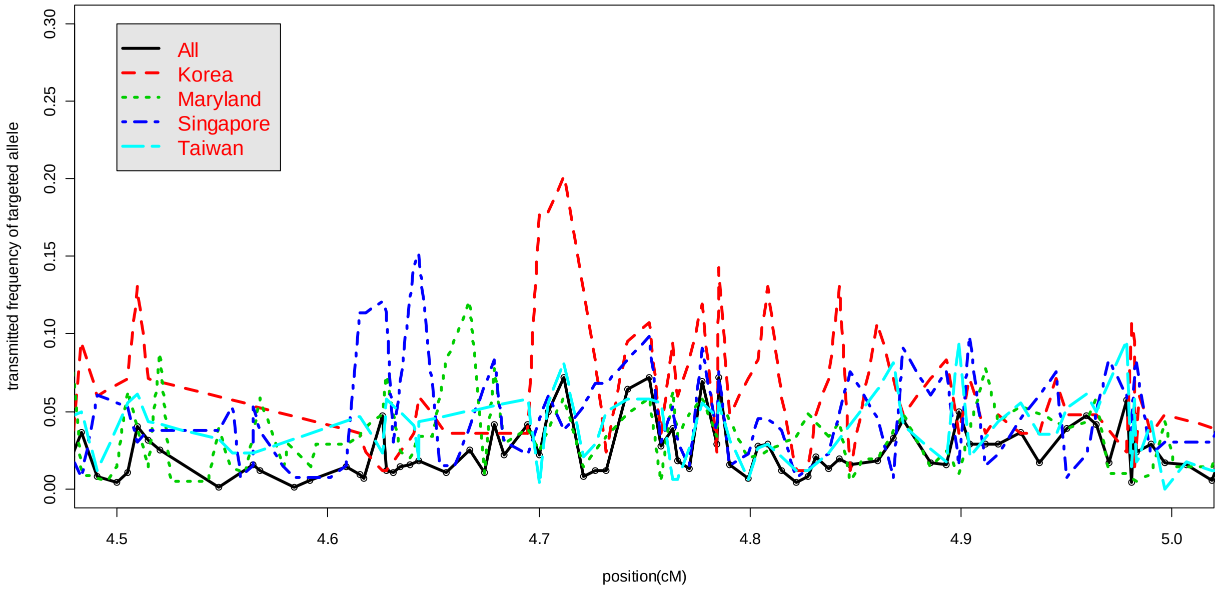

Figure 11. The transmitted statistic from 4.5 cM to 5 cM on chromosome 4p16 from oral clefts data (Sull et al. 2008)

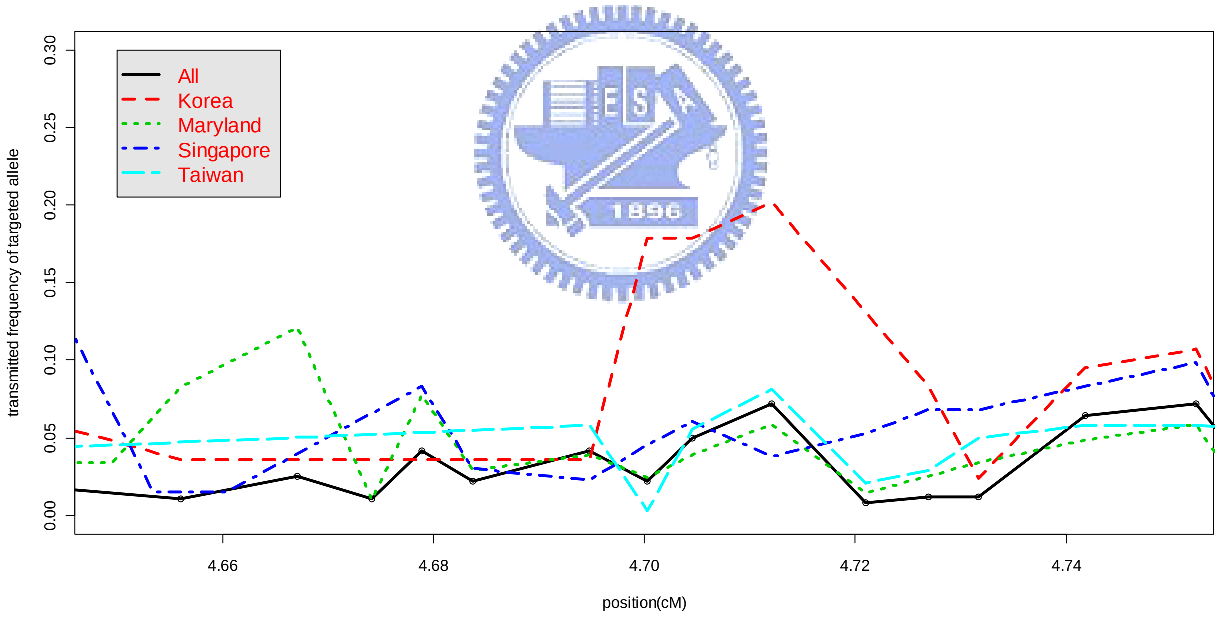

Figure 12. The transmitted statistic from 4.65 cM to 4.75 cM on chromosome 4p16 from oral clefts data (Sull et al. 2008)

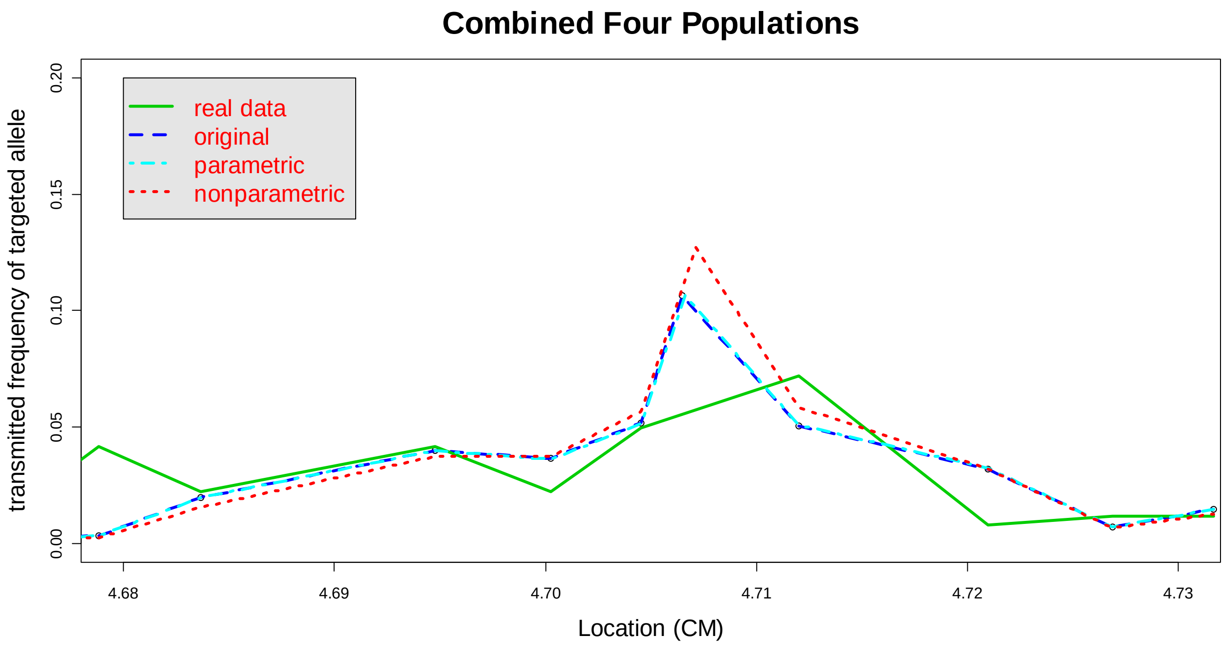

Figure 13. Comparisons of three approaches by incorporating gender into the LD mapping

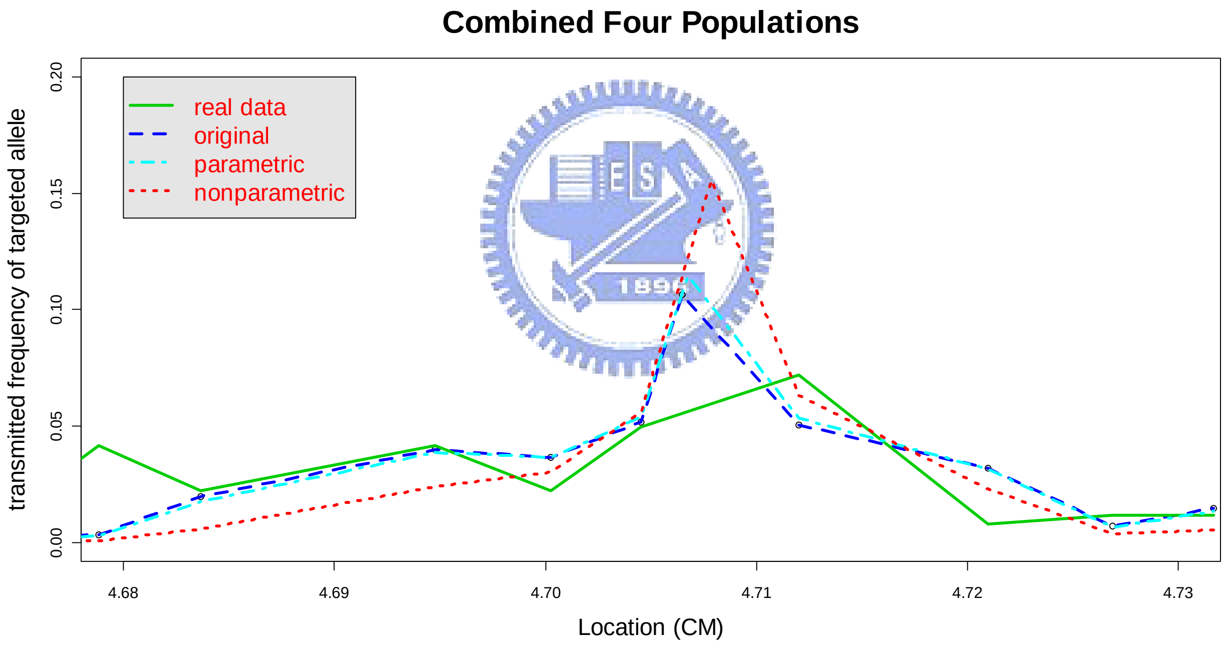

Figure 14. Comparisons of three approaches by incorporating affected father into the LD mapping

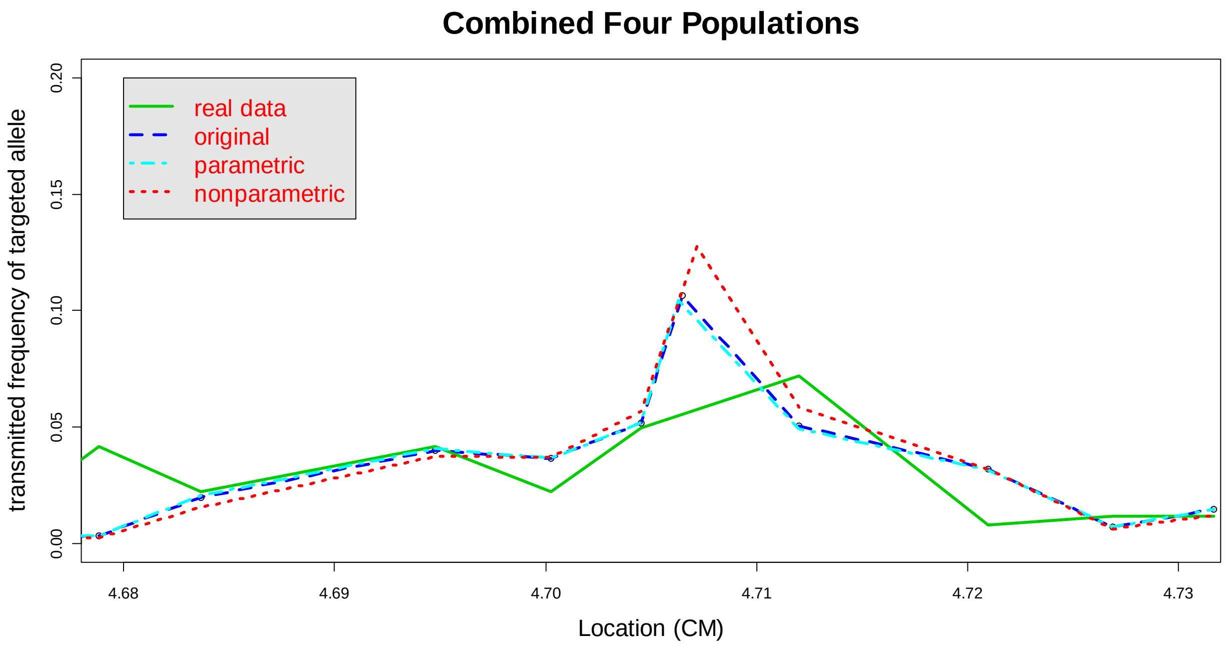

Figure 15. Comparisons of three approaches by incorporating affected mother into the LD mapping

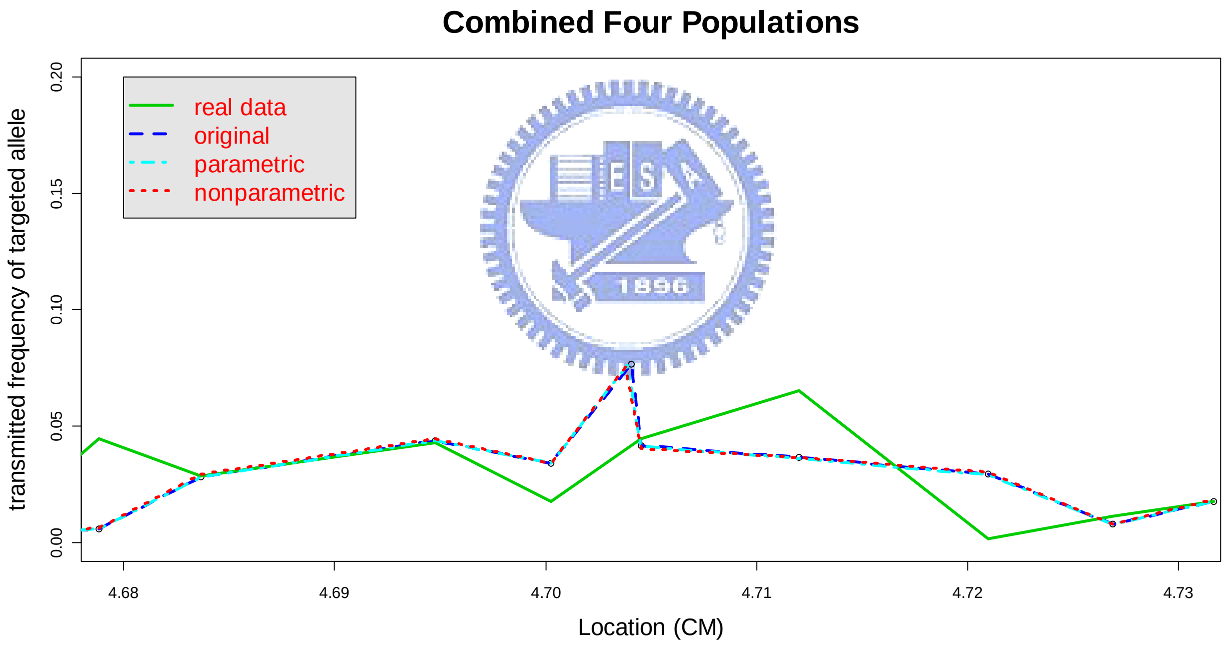

Figure 16. Comparisons of three approaches by incorporating smoking into the LD mapping

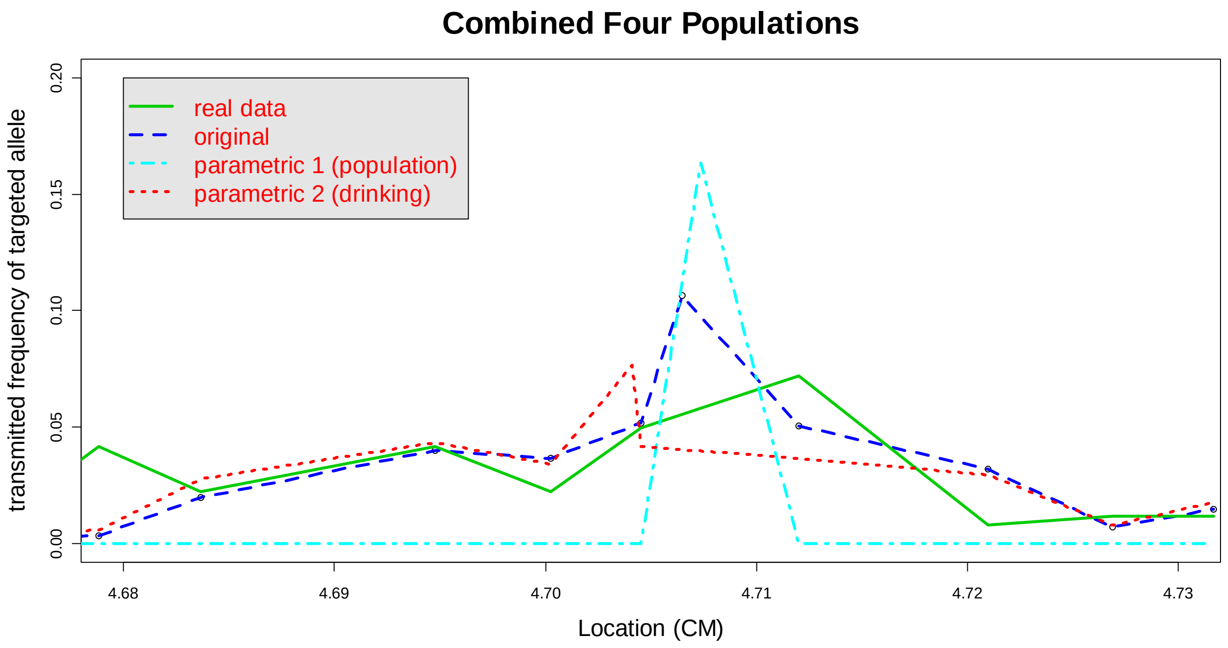

Figure 17. Comparisons of differences in incorporating population types or drinking in the parametric approach

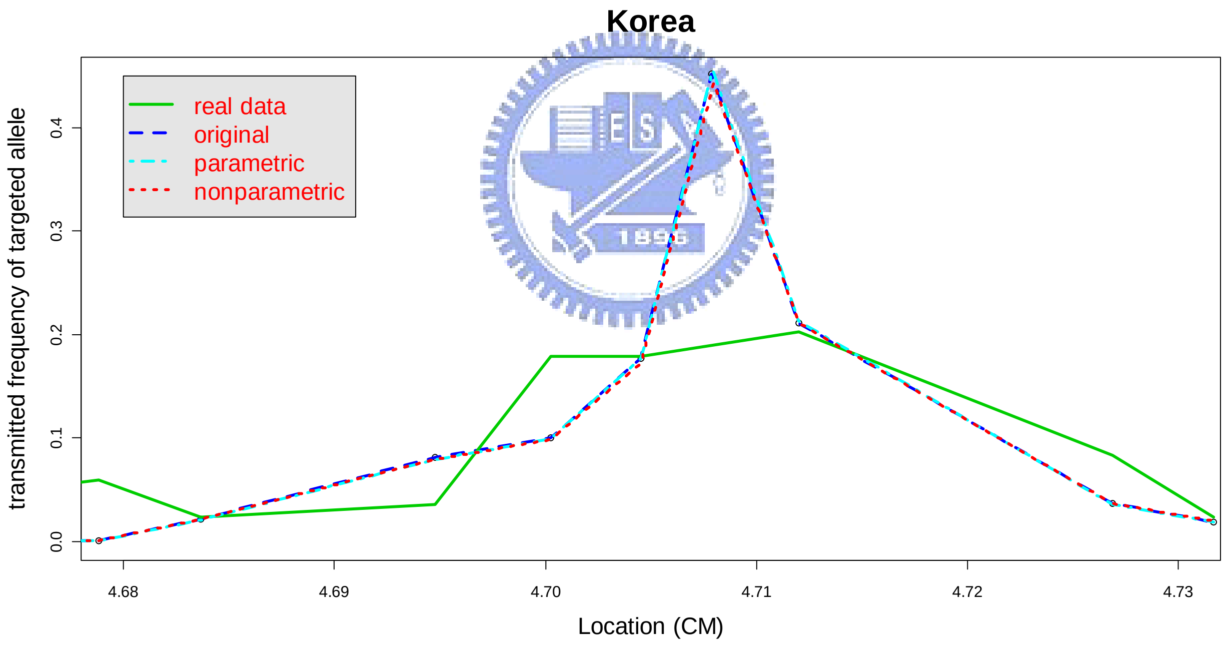

Figure 18. Comparisons of three approaches by incorporating gender into the LD mapping

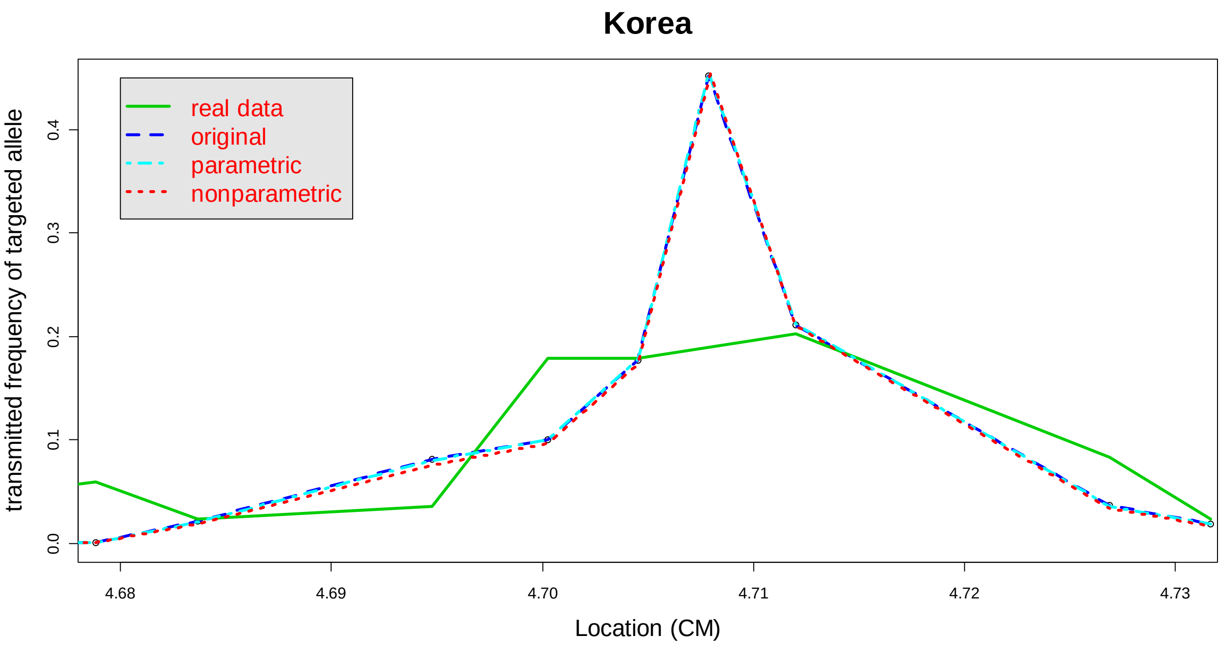

Figure 19. Comparisons of three approaches by incorporating drinking into the LD mapping

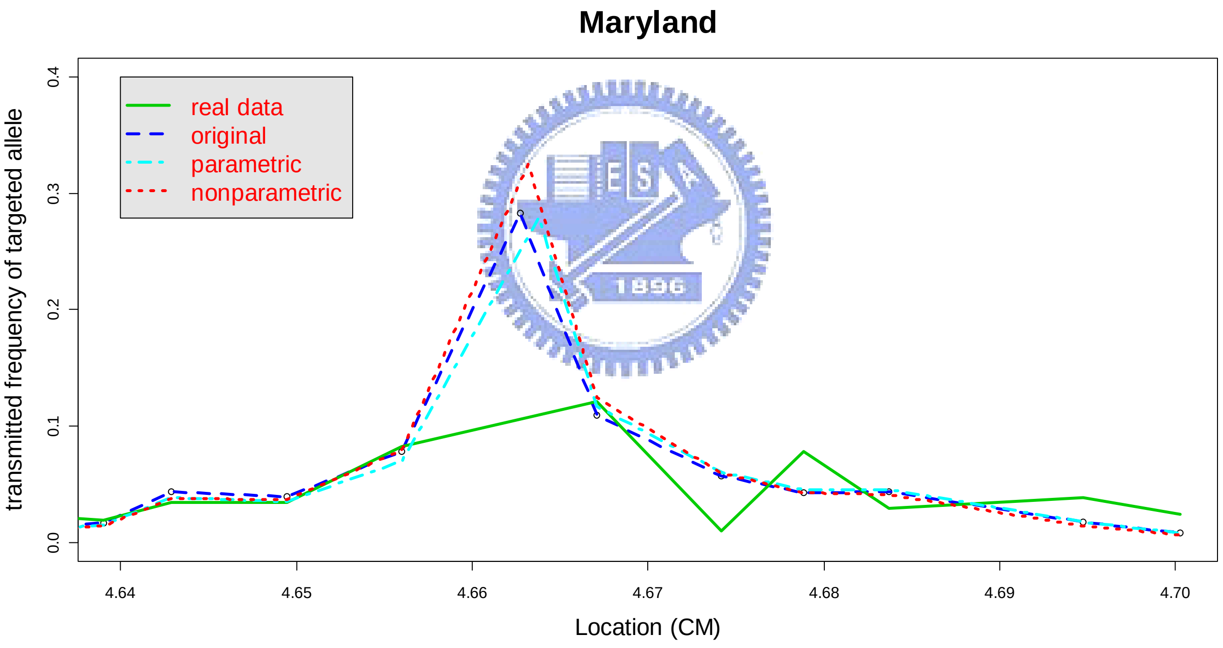

Figure 20. Comparisons of three approaches by incorporating gender into the LD mapping

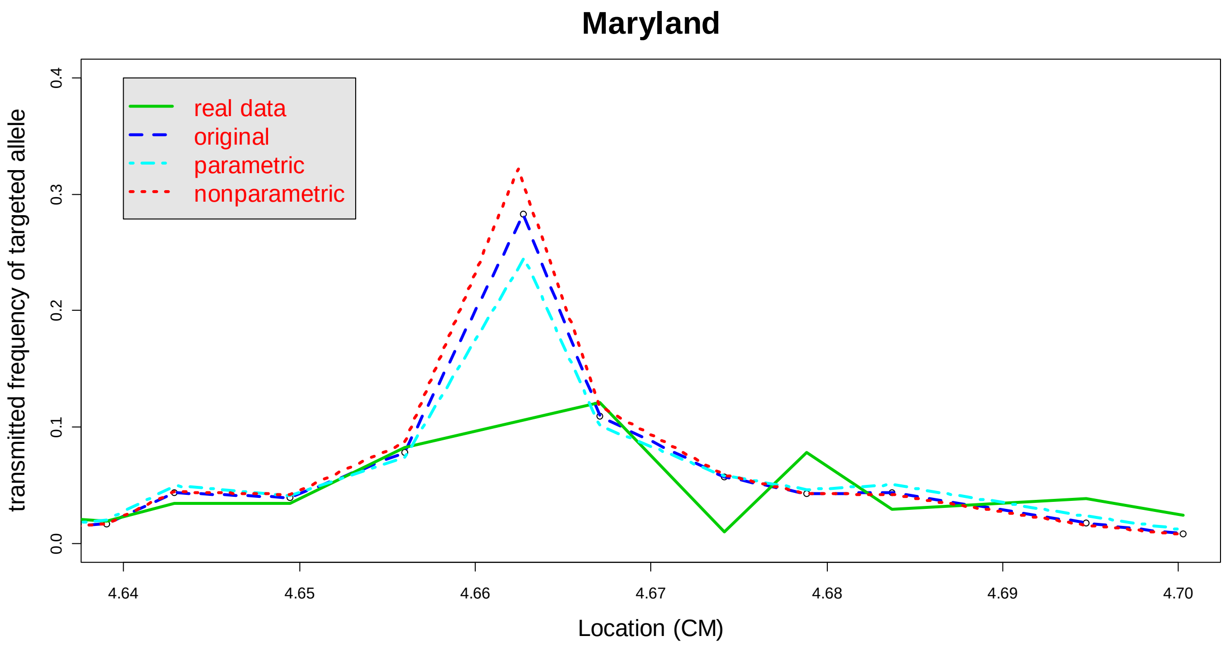

Figure 21. Comparisons of three approaches by incorporating smoking into the LD mapping

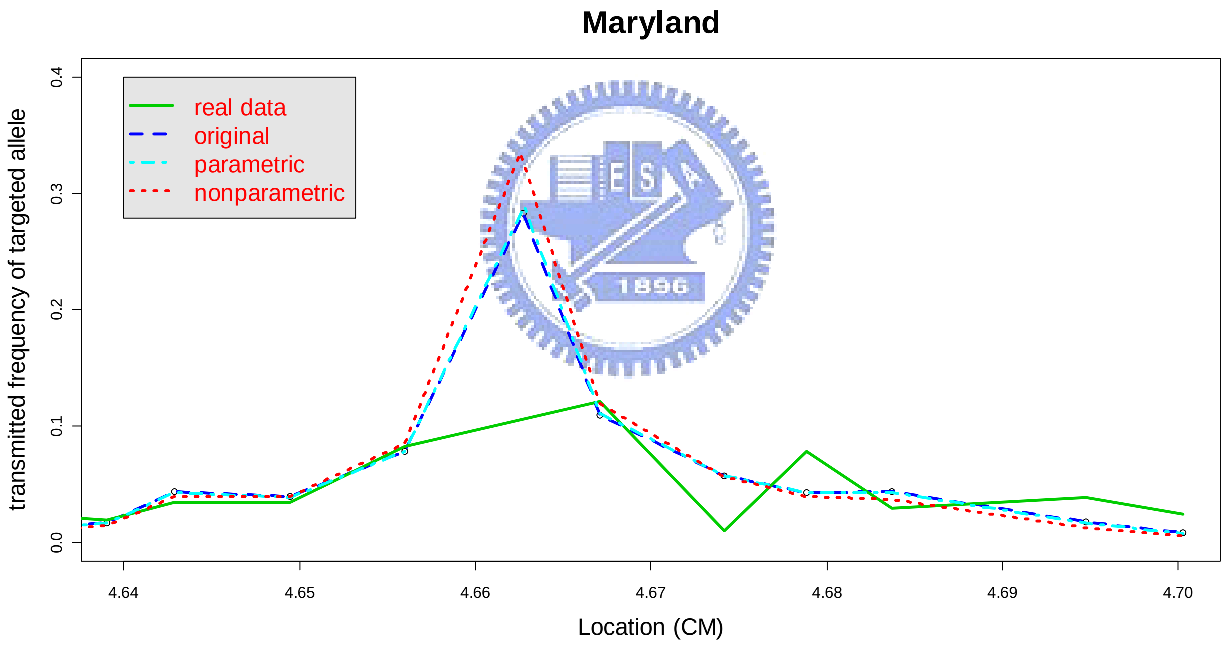

Figure 22. Comparisons of three approaches by incorporating drinking into the LD mapping

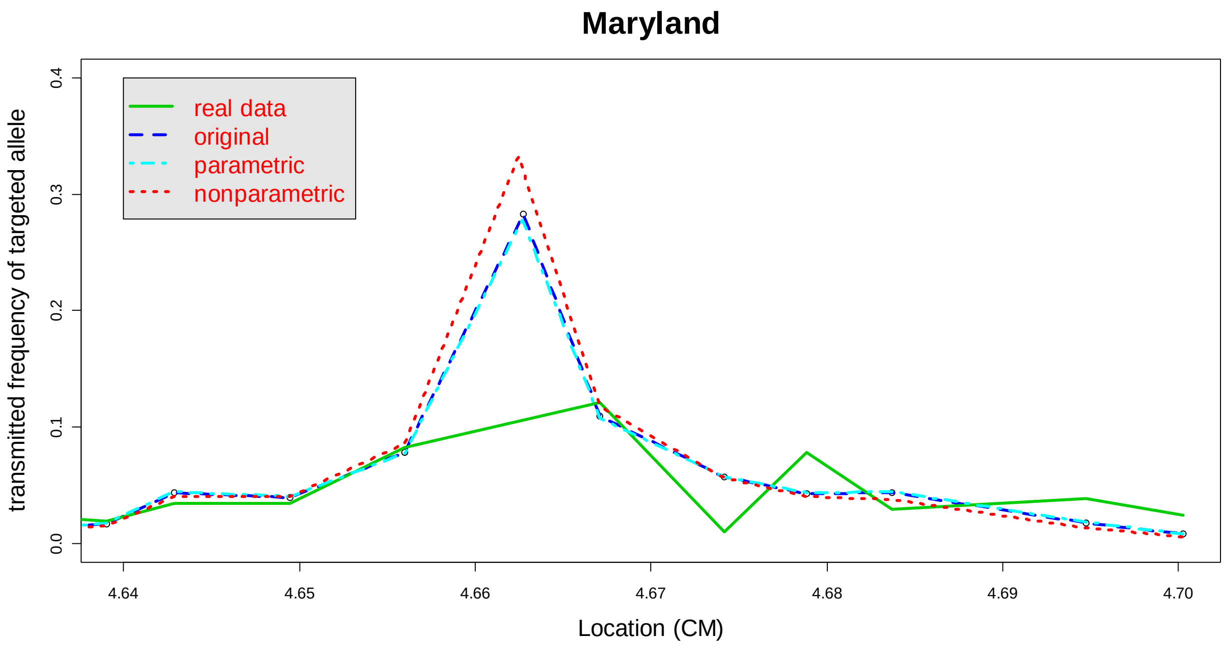

Figure 23. Comparisons of three approaches by incorporating vitamin into the LD mapping

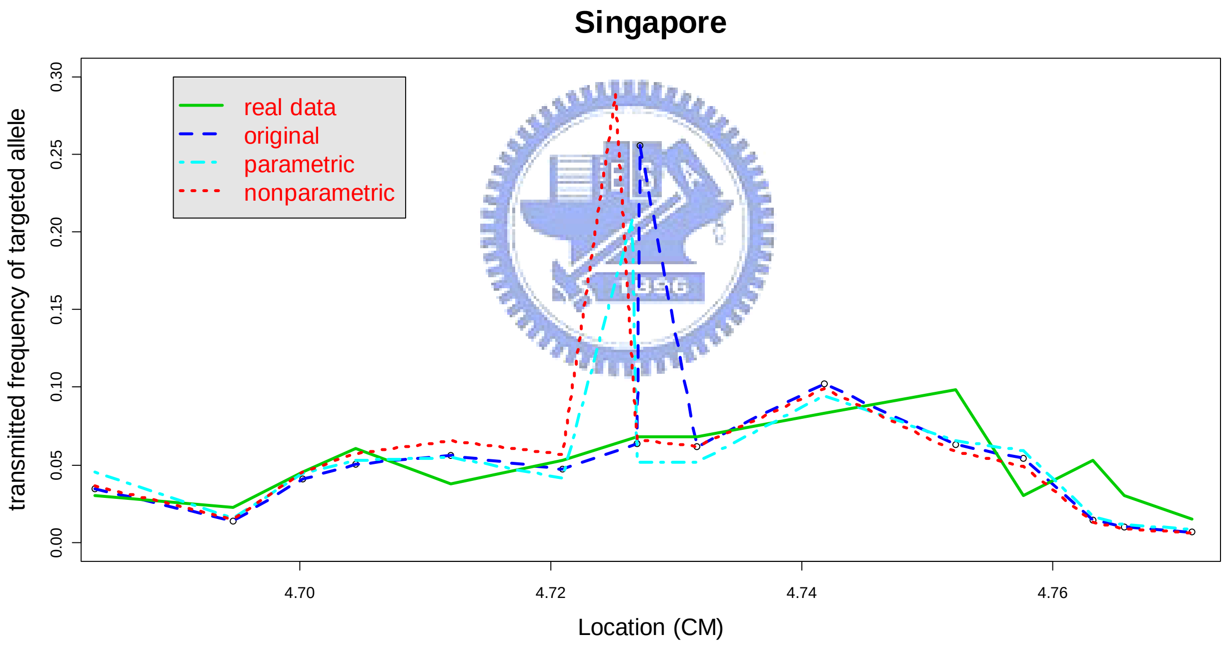

Figure 24. Comparisons of three approaches by incorporating gender into the LD mapping

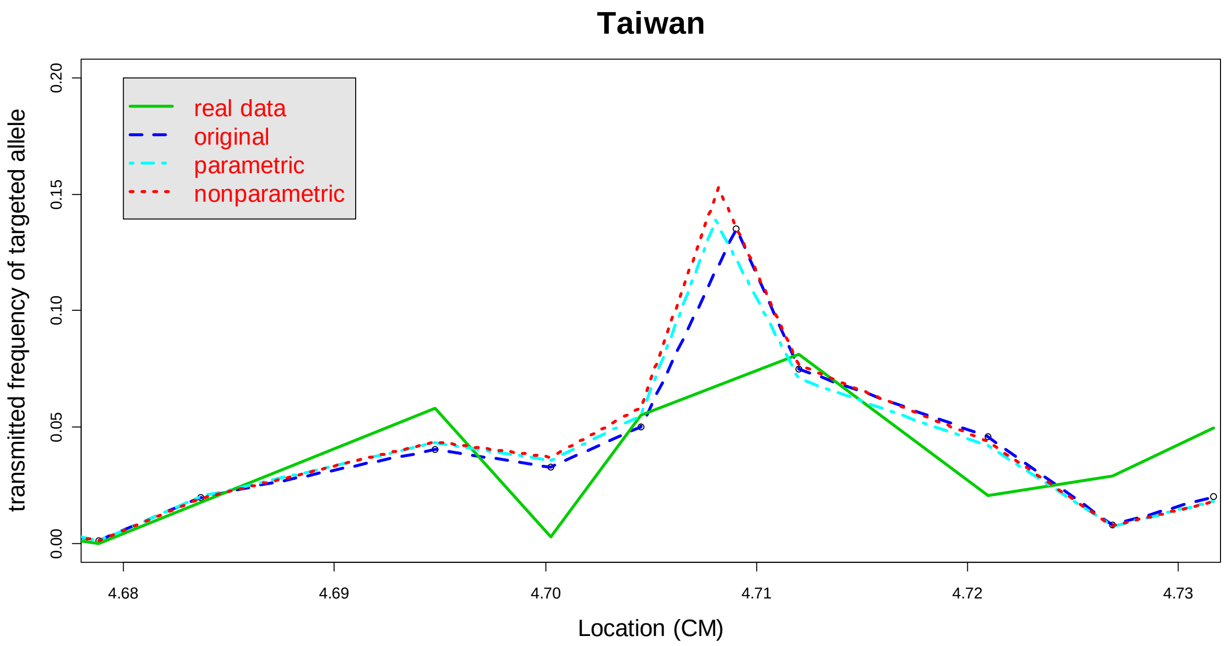

Figure 25. Comparisons of three approaches by incorporating gender into the LD mapping

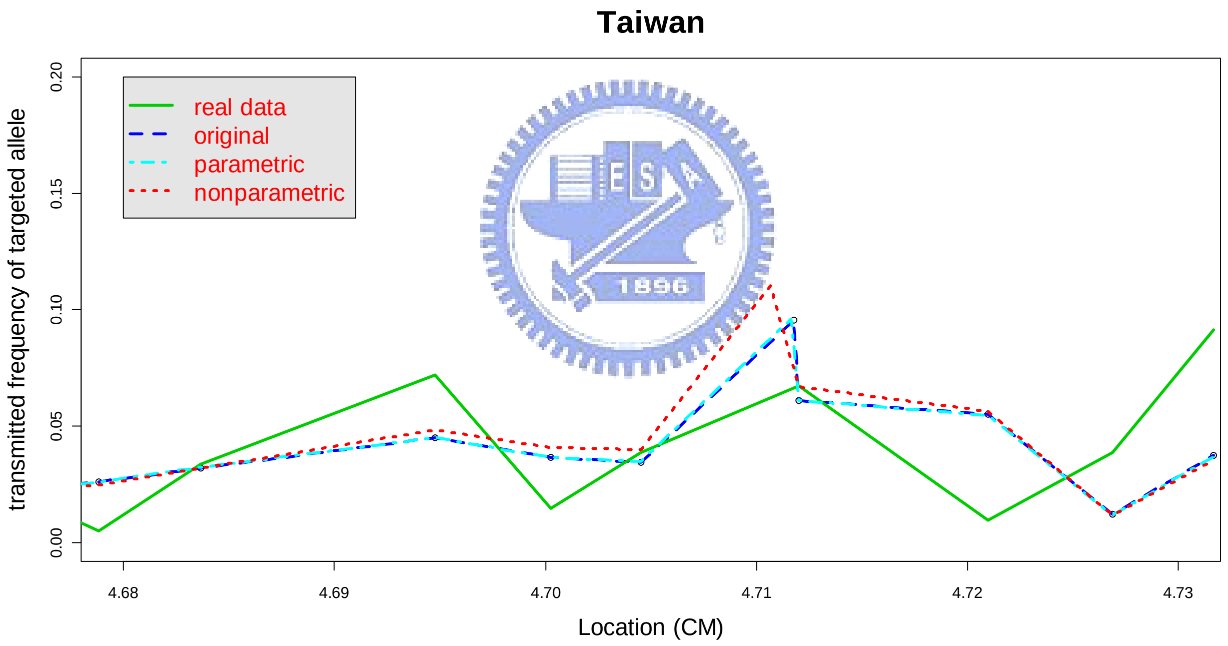

Figure 26. Comparisons of three approaches by incorporating drinking into the LD mapping

.

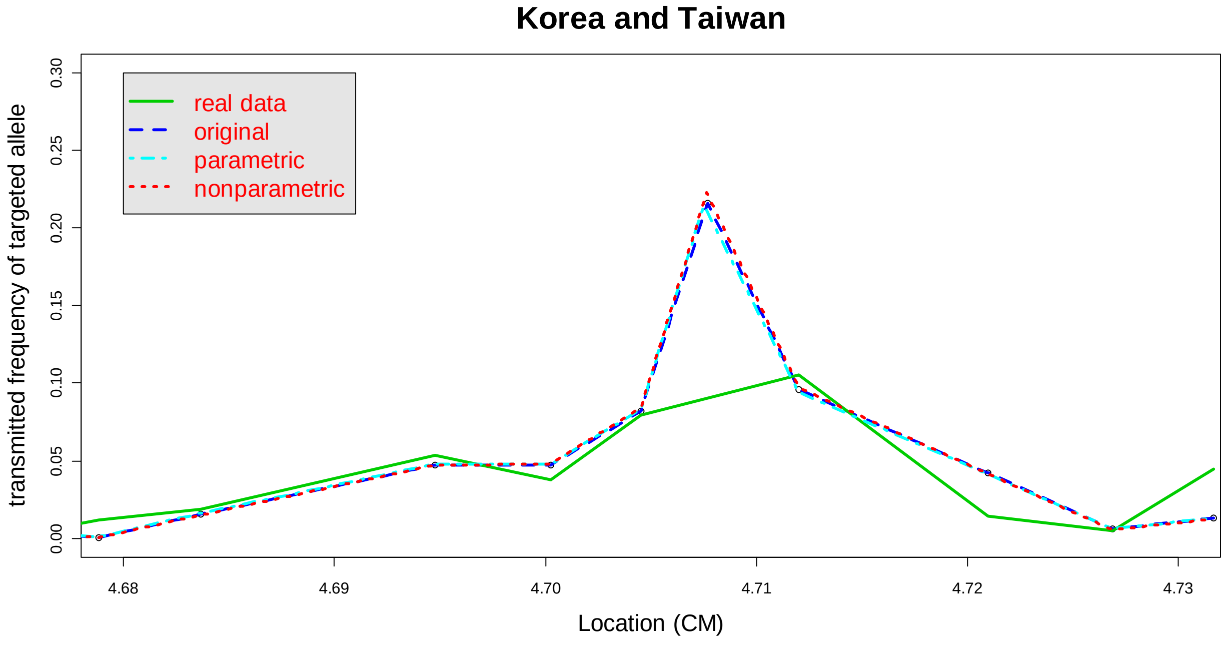

Figure 27. Comparisons of three approaches by incorporating gender into the LD mapping

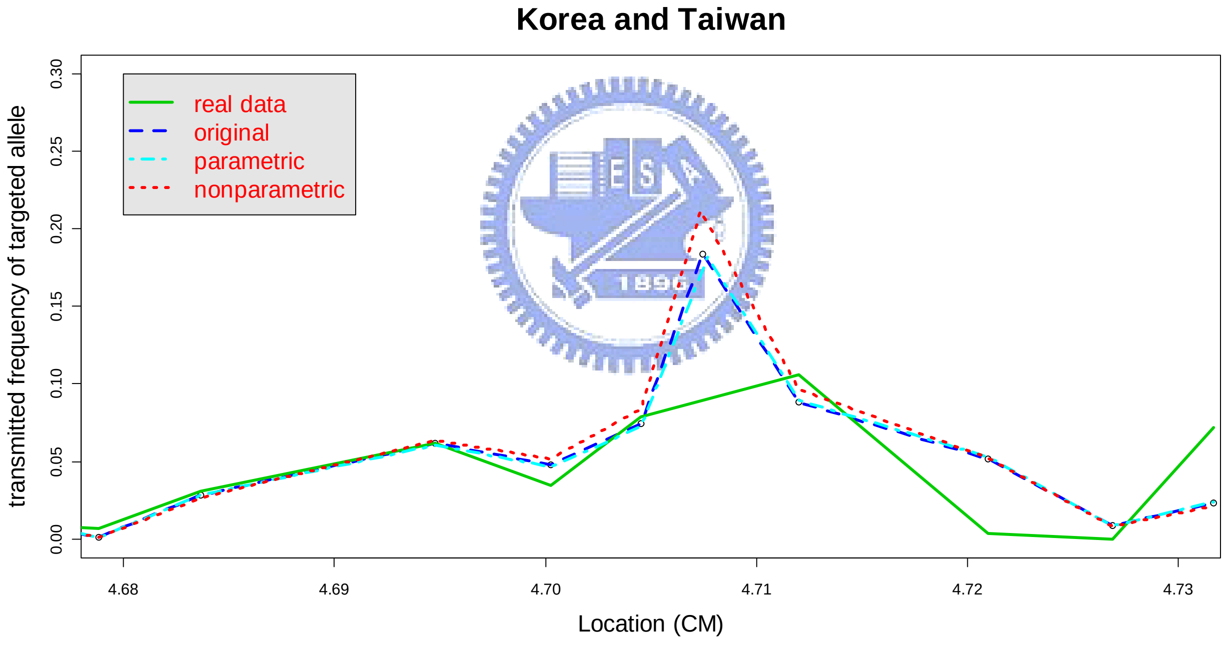

Figure 28. Comparisons of three approaches by incorporating smoking into the LD mapping

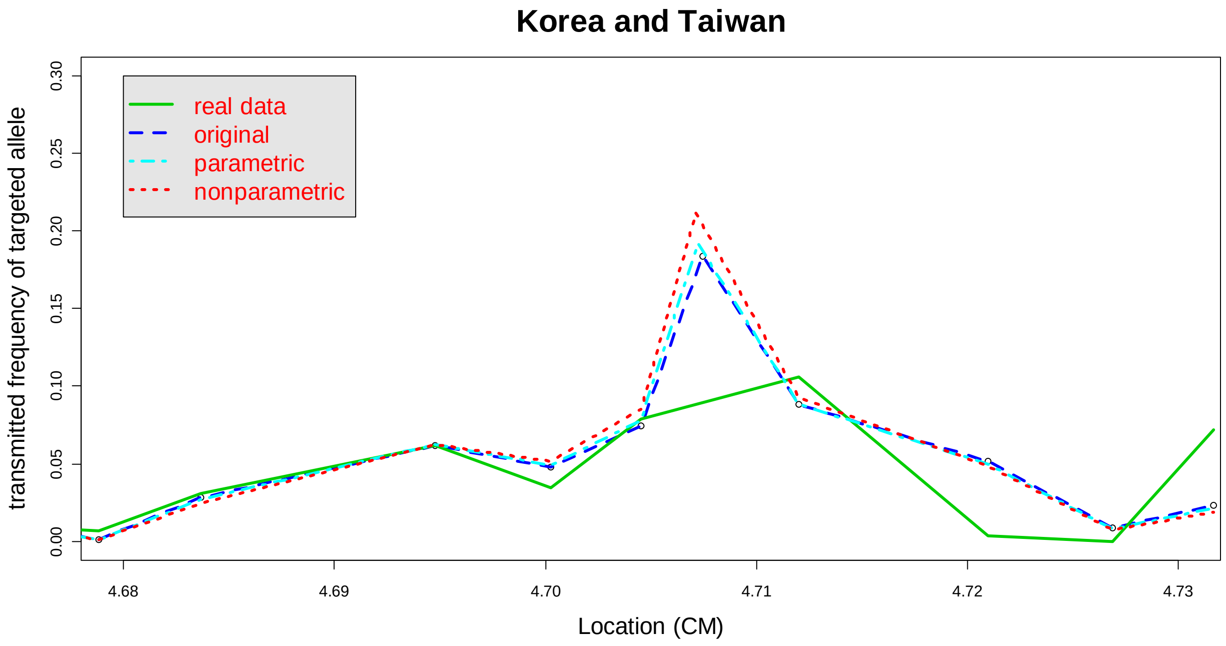

Figure 29. Comparisons of three approaches by incorporating drinking into the LD mapping