Published in final edited form as:

Genet Epidemiol. 2009 February ; 33(2): 151–163. doi:10.1002/gepi.20368.

A Joint Association Test for Multiple SNPs in Genetic Case-Control Studies

Tao Wang^{1,2,*}, Howard Jacob², Soumitra Ghosh³, Xujing Wang³, and Zhao-Bang Zeng⁴

¹Division of Biostatistics, Department of Population Health, Medical College of Wisconsin, Milwaukee, Wisconsin

²Human Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, Wisconsin

³Max McGee National Research Center for Juvenile Diabetes, Medical College of Wisconsin, Milwaukee, Wisconsin

⁴Bioinformatics Research Center, Department of Statistics, North Carolina State University, Raleigh, North Carolina

Abstract

For a dense set of genetic markers such as single nucleotide polymorphisms (SNPs) on high linkage disequilibrium within a small candidate region, a haplotype-based approach for testing association between a disease phenotype and the set of markers is attractive in reducing the data complexity and increasing the statistical power. However, due to unknown status of the underlying disease variant, a comprehensive association test may require consideration of various combinations of the SNPs, which often leads to severe multiple testing problems. In this paper, we propose a latent variable approach to test for association of multiple tightly linked SNPs in case-control studies. First, we introduce a latent variable into the penetrance model to characterize a putative disease susceptible locus (DSL) that may consist of a marker allele, a haplotype from a subset of the markers, or an allele at a putative locus between the markers. Next, through using of a retrospective likelihood to adjust for the case-control sampling ascertainment and appropriately handle the Hardy-Weinberg equilibrium constraint, we develop an expectation-maximization (EM)-based algorithm to fit the penetrance model and estimate the joint haplotype frequencies of the DSL and markers simultaneously. With the latent variable to describe a flexible role of the DSL, the likelihood ratio statistic can then provide a joint association test for the set of markers without requiring an adjustment for testing of multiple haplotypes. Our simulation results also reveal that the latent variable approach may have improved power under certain scenarios comparing with classical haplotype association methods.

Keywords

haplotype association; retrospective likelihood; latent variable; logistic mixture model; EM algorithm

INTRODUCTION

There is growing utilization of high-density genetic markers such as single nucleotide polymorphisms (SNPs) in fine-mapping of genes within candidate regions of interest. The traditional regression approach by treating each marker as an independent variable is limited by high correlation of redundant markers and does not attempt to infer underlying haplotypes. Haplotype-based association is attractive in reducing the data complexity especially for tightly linked SNPs within a small candidate region. As a combination of closely linked alleles on the same chromosome, the haplotypes may also provide better approximation to functional genes than single marker alleles or genotypes [Schaid, 2004; Clark, 2004]. As a result, it may have increased statistical power in detecting disease susceptible loci (DSLs) comparing with the single marker analysis. Haplotype analysis, however, is complicated by the fact that the observed genotype data are incomplete in the sense that the haplotype pairs that an individual inherited from his or her parents cannot be uniquely identified from his or her observed genotype data only as long as there are more than one marker loci having heterozygous genotypes, i.e., the so-called phase problem. Along with the unknown underlying genetic architectures, haplotype-based analysis often faces potential challenges such as uncertainty in haplotype assignment and frequency estimates and multiple testing problems.

There have been extensive studies on haplotype analysis. In most of the current haplotype association studies, people often partition the SNPs into haplotype blocks and then focus on common haplotypes from the whole set of contiguous markers within blocks. Since genetic markers are usually selected without knowing their underlying genetic functions, it is possible that a true functional disease variant may consist of a single marker allele or a haplotype from a subset of the markers. It could also be a mutant allele at a putative DSL located between two flanking marker loci. When the underlying disease variant is embedded into several common or rare haplotypes, considering only the whole-set marker haplotypes may lead to biased estimate of genetic effects, and probably reduced power because the effect of the disease variant is diluted across several haplotypes. We may also lose the chance to detect a disease variant if it is only embedded in a number of rare haplotypes even though itself may not be so rare. Therefore, in order to fully exploit association between a disease phenotype and a set of markers, it may require consideration of haplotypes from any subset of the markers. This, however, often leads to severe multiple testing problems [Schaid, 2004; Becker and Knapp, 2004].

Several studies have explored possible extensions of the classical haplotype association tests. Browning and Browning [2007] proposed a localized haplotype clustering method to test for single marker alleles and haplotypes but not sub-haplotypes. Gauderman et al. [2007] developed a principal-component clustering approach that could capture a disease variant consisting of either a single SNP allele or as a haplotype or sub-haplotype of the markers. But it is less likely to capture a disease variant at a DSL between the markers, and sometimes may have difficulty in interpretation of specific components. In Wang et al. [2006], we developed a likelihood-based latent variable approach to test for haplotype-based association between a quantitative trait and multiple SNPs in cohort or cross-sectional designed genetic studies. In our approach, we introduced a latent variable in the genetic model to describe a putative DSL within a candidate genome region. An expectation-maximization (EM) algorithm was then developed to impute the genotypes at the DSL. Due to a flexible role of the latent variable, which can characterize a marker allele, a haplotype from a subset of the markers, or even a disease variant at a putative locus between two flanking markers, the likelihood ratio (LR) statistic can then provide a comprehensive test between the disease phenotype and the set of markers without requiring adjustment for testing of multiple haplotypes.

In this article, we extend this latent variable approach to genetic case-control studies. Case-control studies have been widely applied in epidemiological studies due to their low cost and high efficiency. A common approach to test for haplotype association using case-control samples is to estimate haplotype frequencies in cases and controls either separately or combined and then assign the most likely haplotype pair to each individual, as if it were observed, given his or her observed genotypes. After that, one can compare the haplotype frequencies between the cases and controls by using standard χ^2 tests for contingency tables. The classical logistic regression method can also be applied to adjust for other potential confounding variables. However, this approach is flawed in several aspects. Firstly, by treating the most likely haplotype pair as observed, we ignore other possible haplotype pairs (i.e., haplotype uncertainty) which may lead to smaller variance estimation of parameter estimates and inflated type I errors. Secondly, as pointed out in Epstein and Satten [2003], the over-sampling of cases in most of the case-control studies may lead to violation of the Hardy-Weinberg equilibrium (HWE) near a DSL. HWE is an assumption based on which the algorithm for estimation of haplotype frequencies is usually constructed. Under the null hypothesis of no association, the haplotype distribution at the marker loci should be the same between the cases and controls. Therefore, the assumption of HWE does not lead to inflated type I error rate. However, under the alternative hypothesis of there existing association, the HWE assumption may lead to biased estimates of haplotype frequencies. So, whether we should assume HWE in cases becomes a difficult dilemma to handle.

To account for the haplotype uncertainty, Schaid et al. [2002], Stram et al. [2003], and Lake et al. [2003] developed prospective likelihood-based methods to allow for reconstruction of all possible haplotype pairs for each subject with weights determined by their conditional frequencies within the subject given the observed genotypes. Through an EM-procedure, haplotype frequencies and effects can then be estimated and tested simultaneously, and it was found that the EM-based iteration procedure could provide better estimation. In their prospective likelihood-based approach, the HWE assumption still poses a potential problem. Estimation of the haplotype frequencies and their effects could be biased even with correctly specified penetrance models. Another potential problem, which is also related to the HWE assumption, is the validity of the prospective likelihood-based approach itself. In epidemiological studies, it has been well known that case-control data can usually be analyzed through standard logistic regression, as if it comes from a prospective cohort study. The rationale behind this is due to the fact that a prospective likelihood of the case-control data could provide, except the intercept term, the equivalent maximum likelihood estimates (MLE) of the regression parameters as the retrospective likelihood gives provided that the regression parameters are unconstrained and the effects are fixed [Prentice and Pyke, 1979]. In haplotype association analysis, however, haplotypes of the sampled individuals are partially missing but restricted by their observed genotypes, and the HWE assumption also adds a constraint on distribution of the paternal and maternal haplotypes. As a result, the standard logistic regression may no longer provide the equivalent parameter estimates as the one from its retrospective likelihood [Cheng and Lin, 2005].

Case-control data naturally lead to a so-called retrospective likelihood, which is the conditional probability of exposures given the disease status. Using the retrospective likelihood allows us to automatically adjust for the sampling scheme involved in the ascertainment of cases and controls. Recently, a number of retrospective likelihood-based methods have also been proposed for haplotype association analysis [Spinka et al., 2005; Chatterjee and Carroll, 2005; Lin and Zeng, 2006]. The retrospective likelihood-based approach can automatically correct for potential bias due to over-sampling of cases in the case-control studies. In addition, it can handle the HWE dilemma more appropriately by assuming HWE in the sampled population instead of cases or controls. Meanwhile, it can still account for haplotype uncertainty by considering all possible haplotype pairs per subject in the likelihood analysis.

As a trade-off, the retrospective likelihood-based methods usually require some prior knowledge about the disease prevalence in the sampled population. Besides, they are computationally more intensive and often require construction of a profile likelihood to replace the nuisance parameters involved in the distributions of some environmental factors.

In this article, we first formulate a logistic mixture regression model to describe penetrance of risk haplotypes. A latent predictive variable is then introduced to model the influence of an unknown DSL. Next, we develop an EM algorithm to fit the logistic mixture latent variable model by using a retrospective profile likelihood derived by Spinka et al. [2005]. A joint association test is then constructed for haplotype-based association between the disease and multiple markers based on the LR statistics. Finally, we implement three different latent variable methods, i.e., a prospective likelihood-based, a retrospective likelihood-based, and a rare disease approximation, under one framework by treating the ratio of sampling rates in controls and cases as a tuning parameter. Performance of these three methods and their comparisons with some other haplotype association methods are assessed through extensive simulation studies.

METHODS

LOGISTIC MIXTURE MODEL

Suppose that we have a random sample of N unrelated individuals from a sampled population. Let y_i be an indicator for the disease status ($y_i = 1$ for affected cases; $y_i = 0$ for controls). Let g_i , and z_i , $i = 1, 2, \dots, N$, denote the observed marker genotypes, and p covariates of individual i , respectively. Then a general setting for traditional haplotype association analysis is given by the following logistic regression model:

$$\begin{aligned} \text{logit } P(y_i=1 | (h_{i0}, h_{i1}), z_i) \\ = \beta_0 + z_i \alpha + X(h_{i0}, h_{i1}) \beta \quad \text{for } i=1, \dots, N, \end{aligned} \quad (1)$$

where (h_{i0}, h_{i1}) is the phase-known haplotype configuration (or the so-called diplotype) of individual i with h_{i0} denoting the maternally transmitted haplotype and h_{i1} the paternally transmitted haplotype, α is a p -dimensional vector of parameters for fixed effects of the covariates, β is a q -dimensional vector of parameters for fixed effects of the haplotypes which may include additive, dominance, and probable genetic interactions between the haplotypes, and the $X(h_{i0}, h_{i1})$ is a $1 \times q$ incidence vector corresponding to β with its components coded according to diploypes (h_{i0}, h_{i1}) of the individual. For example, for a specific haplotype of interest h^* , we can let $X(h_{i0}, h_{i1})\beta$ take $a \cdot [I(h_{i0} = h^*) + I(h_{i1} = h^*)] + d \cdot I(h_{i0} = h_{i1} = h^*)$ with a, d corresponding to the additive, dominance effect of the haplotype h^* . In haplotype analysis, we are interested in testing genetic association for markers within a small candidate region. Genetic markers outside of the region could be included as co-factors to adjust for influence contributed by other genetic factors outside the region — a strategy proposed by Zeng [1994] for composite interval mapping. Interaction between haplotype and environmental covariates may also be included into the model.

A specialty in the genotype data is that the diplotype (h_{i0}, h_{i1}) of each individual i is not directly observable and usually needs to be inferred from its observed genotypes g_i . The haplotypes of markers are usually assumed to follow a multinomial distribution, i.e., $\text{Multinomial}(N, P_h)$, where N is the number of sampled individuals and h denotes the haplotype of the markers. Marker allele frequencies, pairwise linkage disequilibria (LDs) between, and higher-order disequilibria can all be derived from P_h . In addition, it is often assumed that there are HWEs at the marker loci (or, more precisely, the gametic phase equilibrium at the marker loci) in the sampled population — a common assumption that holds as long as there is a random mating

in the parental population. Then the maternal and paternal haplotypes are independent and the frequency of a phase-known diplotype (h_{i0}, h_{i1}) is a product of its two haplotype frequencies, i.e., $P\{(h_{i0}, h_{i1})\} = P(h_{i0})P(h_{i1})$. Note that here we also assume that the paternal gametes and maternal gametes have the same haplotype distribution.

Model (1) is different from classical logistic models in that the diplotypes (h_{i0}, h_{i1}) are not directly observable and need to be treated as random from a specific distribution even though they have fixed effects in the model — a situation similar to the measurement error model for predictors [Roeder et al., 1996]. In general, the observed genotypes g_i provide only partial information for the phased diplotypes (h_{i0}, h_{i1}) , $i = 1, 2, \dots, N$. Given the observed genotypes g_i at multiple markers for an individual i , its phased diplotypes (h_{i0}, h_{i1}) are constrained by its observed genotypes g_i through $(h_{i0}, h_{i1}) \in H(g_i)$ where $H(g_i)$ denotes a set of diplotypes that are compatible with g_i . As long as there are more than one heterozygous marker genotypes in g_i , there will be more than one possible diplotypes (h_{i0}, h_{i1}) that are compatible with g_i . Missing genotypes at a marker can also be treated as no constraint at this locus.

Given the observed covariates z_i and assuming haplotype-environment independency between (h_{i0}, h_{i1}) and z_i , the probability of observed data (y_i, g_i) can be constructed hierarchically in terms of the distribution of diplotype (h_{i0}, h_{i1}) as in the following equation

$$P(y_i, g_i | z_i) = \sum_{(h_{i0}, h_{i1}) \in H(g_i)} P(y_i | (h_{i0}, h_{i1}), z_i) P(h_{i0}) P(h_{i1})$$

which is a mixture of logistic regression models. This logistic mixture model has been utilized in many prospective and retrospective likelihood-based haplotype studies, although it was not presented in such a formal manner. One naive way of fitting this model is to estimate haplotype frequencies first by assuming HWE and applying the EM algorithm [Excoffier and Slatkin, 1995]. Then one can reconstruct all possible haplotype pairs for each subject and estimate and test for haplotype effects through weighted regression [Zaykin et al., 2002]. However, this method does not take into account variation in estimation of the haplotype frequencies. It has been shown that by estimating the haplotype frequencies P_h and fitting the penetrance model (1) simultaneously we could improve parameter estimation [Schaid et al., 2002; Lake et al., 2003]. A key procedure in fitting the model jointly was to reconstruct all possible haplotype pairs for each subject based on its observed genotypes, phenotypes, and current parameter setting through which the phenotypic values can also contribute to estimation of haplotype frequencies. The haplotype effects are evaluated through a weighted regression with weights determined by the current haplotype frequency estimates.

In most of current haplotype studies, people often treat h_{i0} and h_{i1} , $i = 1, 2, \dots, N$, as haplotypes from a whole set of contiguous markers within a small candidate region, which implicitly assume that the risk factors consist of haplotypes from the whole set of markers. However, as we mentioned before, a risk factor of the disease could consist of a single allele or a haplotype from a subset of the markers. It could also be a mutant allele at a putative DSL between the markers. To characterize a risk factor with such unknown status, we further incorporate a latent variable into model (1). Let (q_{i0}, q_{i1}) denote the i th individual's diplotypes at the DSL. A logistic mixture model with the latent variable (q_{i0}, q_{i1}) is then given by

$$\begin{aligned} \text{logit } P(y_i=1 | (q_{i0}h_{i0}, q_{i1}h_{i1}), z_i) \\ = \beta_0 + z_i\alpha + X(q_{i0}, q_{i1})\beta, \quad i=1, \dots, N \\ q_{i0}h_{i0}, \quad q_{i1}h_{i1} \sim \text{Multinomial}(N, P_{qh}), \quad \text{and} \\ q_{i0}h_{i0} \perp q_{i1}h_{i1}, \end{aligned}$$

where β is the (fixed) genetic effects of a DSL and P_{qh} denotes the joint haplotype frequencies of the DSL and markers. In this model, we implicitly assume that the disease phenotype does not depend on marker genotypes within the region of interest given the diplotype at the DSL. Unlike marker loci where observed genotypes are available, the diplotypes (q_{i0}, q_{i1}) , $i = 1, 2, \dots, N$, in the model are totally missing. So, the marker diplotypes $(h_{i0}, h_{i1}) \in H(g_i)$ are constrained by the observed genotypes g_i while there is no constraint on the DSL diplotypes (q_{i0}, q_{i1}) . Still, we assume the gametic equilibrium at the DSL and marker loci so that the probability of the joint diplotypes $(q_{i0}h_{i0}, q_{i1}h_{i1})$ is a product of its two haplotype frequencies. The probability of observed (y_i, g_i) given covariates z_i now becomes

$$P(y_i, g_i | z_i) = \sum_{(q_0, q_1) \in Z^2} \sum_{(h_{i0}, h_{i1}) \in H(g_i)} \times P(y_i | (q_0 h_{i0}, q_1 h_{i1}), z_i) P(q_0 h_{i0}) P(q_1 h_{i1})$$

where Z is the state space for alleles at the DSL. Assuming a biallelic DSL, we can take $Z = \{0, 1\}$ with 0, 1 indicating the common and rare alleles, respectively. As the summation on $(q_{i0}, q_{i1}) \in Z^2$ is the same for all individuals, we drop the subject index i to simplify the notation. We would like to see whether the latent DSL in the model could play a flexible role of a putative risk factor that consists of an allele, haplotypes, or sub-haplotypes of the markers. We first proposed this latent variable approach to prospective cohort or cross-sectional studies for quantitative traits [Wang et al., 2006]. In this article, we extend the method to retrospective case-control studies.

One major characteristic of the case-control design is its retrospective sampling scheme of cases based on their disease status. It has been shown in Prentice and Pyke [1979] that, under no restriction on missing covariate information, a regression based on a prospective likelihood $L' = \pi_i P(y_i | g_i, z_i)$ can actually provide the same MLE of regression coefficients except the intercept term as that based on a retrospective likelihood $L = \pi_i P(g_i, z_i | y_i)$. In general, the prospective approach is easy to implement through standard regression, while retrospective approach may become more efficient than the prospective likelihood-based MLE. In genetic studies, however, the case-control sampling scheme creates more complications that need not be addressed in conventional epidemiology studies. The retrospective sampling scheme may not only over-sample the cases but also distort the genotypic distribution in the cases. Besides, with Hardy-Weinberg constraint on diplotypes, the MLE from prospective likelihood is in general no longer equivalent to MLE based on the retrospective likelihood [Spinka et al., 2005].

In the following, using the same profile likelihood derived by Spinka et al. [2005], we develop a likelihood-based approach to estimate the haplotype distribution $\{\pi_k\}$ of $q_k h_k$ and the regression model parameters $\theta = (\beta_0, \alpha, \beta)$ jointly. Three different methods (i.e., a prospective likelihood-based, a retrospective likelihood-based, and a rare disease approximation) will be implemented under one framework. In addition, we construct the LR statistic to test for association between the disease phenotype and DSL. Due to the flexible role of the DSL in our latent variable model, this LR test provides a composite test for haplotype-based association between the disease phenotype and the set of marker loci regardless of whether the true risk factor consists of an allele, a sub-haplotype, or a haplotype of the markers.

ESTIMATION METHODS

We assume that (i) given DSL genotypes, each phenotype y_i is conditionally independent of marker genotypes; and (ii) the genotypic distribution of DSL and markers is independent of the non-genetic covariates. For a case-control design, the diseased cases are sampled from the conditional distribution of $g, z | y = 1$, while the disease-free controls are sampled separately

from $g, z|y = 0$. Therefore, as in ascertainment-corrected case-control studies, the retrospective likelihood is given by

$$\begin{aligned} L &= \prod_{i=1}^N P(g_i, z_i | y_i) = \prod_{i=1}^N \sum_{h_q \in Z^2} \sum_{h_m \in H(g_i)} P(h, z_i | y_i) \\ &= \prod_{i=1}^N \sum_{h_q} \sum_{h_m \in H(g_i)} \frac{P(z_i)P(h|z_i)P(y_i|h_q, z_i)}{P(y_i)} \\ &= \prod_{i=1}^N \frac{P(z_i) \sum_{h_q} \sum_{h_m \in H(g_i)} P(h|z_i)P(y_i|h_q, z_i)}{\int_z P(z) \sum_{h_q} \sum_{h_m \in H} P(h|z)P(y_i|h_q, z) dF(z)} \end{aligned}$$

where $h_q = (q_0, q_1)$ is a diplotype at the DSL, $h_m = (h_0, h_1)$ a diplotype at the markers, $h = (h_q, h_m)$ the joint diplotype of the DSL and markers, and $F(z)$ the marginal distribution function of z_i .

In practice, z_i often contains information about several environmental covariates. Thus $F(z)$ may potentially involve many nuisance parameters in the likelihood. Assuming that z_i is discrete and can take finite values $c_1, \dots, c_S$ with mass probabilities $P(z_i = c_k) = d_k$ for $k = 1, \dots, S$, Spinka et al. [2005] derived a profile log-likelihood $l^* = \sup_{F(z)} \{L\}$ by maximizing L on the probability mass function $F(z)$. Let N_1 be the number of cases and N_0 be the number of controls. Using an i.i.d. Bernoulli sequence $\{R_i\}$, $i = 1, 2, \dots, N^*$, where N^* is the total number of subjects in the underlying population, to denote the case-control sampling scheme: $R_i = 1$ if individual i is selected into the case-control sample and $R_i = 0$ otherwise. It turns out that the profile log-likelihood can be expressed in the form

$$\begin{aligned} l^* &= \sum_{i=1}^N \log P(y_i, g_i | z_i, R_i = 1) \\ &= \sum_i \log \left[\sum_{h_q} \sum_{h_m \in H(g_i)} P(h | z_i, R_i = 1) P(y_i | h_q, z_i, R_i = 1) \right] \end{aligned}$$

where

$$\begin{aligned} P(y_i | h_q, z_i, R_i = 1) &= \frac{\exp\{y_i(k + z_i\alpha + X(h_q)\beta)\}}{1 + \exp\{k + z_i\alpha + X(h_q)\beta\}}, \\ P(h | z_i, R_i = 1) &= \frac{P(h)\gamma(h_q, z_i)}{\sum_{h_q} P(h_q)\gamma(h_q, z_i)} \end{aligned}$$

and

$$\gamma(h_q, z_i) = \frac{1 + \exp\{k + z_i\alpha + X(h_q)\beta\}}{1 + R^* \cdot \exp\{k + z_i\alpha + X(h_q)\beta\}}$$

with $k = \beta_0 - \log R^*$ and $R^* = P(R_i = 1 | y_i = 0) / P(R_i = 1 | y_i = 1)$ being the ratio of sampling rates in controls vs. cases. When the cases and controls have the same sampling rates, i.e., $P(R_i = 1 | y_i = 1) = P(R_i = 1 | y_i = 0)$, we have $R^* = 1$ and $\gamma(h_q, z_i) \equiv 1$. Then l^* becomes the same as the classical prospective log-likelihood as we would have expected in cohort studies. For rare diseases, we usually have $R^* \approx 0$ and $\gamma(h_q, z_i) \approx 1 + \exp\{k + z_i\alpha + X(h_q)\beta\}$. In general, with $0 \leq R^* < 1$, we can see that the MLE from this profile likelihood may not always be the same as

the one from the prospective likelihood. This could lead to different estimates in both the intercept term and other regression coefficients. Note that the sampling rates $P(R_i = 1|y_i = 0)$ and $P(R_i = 1|y_i = 1)$ in the two disease strata can be approximated by $P(R_i = 1|y_i = 0) \approx N_0/P(y_i = 0)N^*$ and $P(R_i = 1|y_i = 1) \approx N_1/P(y_i = 1)N^*$. Therefore, when the disease prevalence $P(y = 1)$ in the sampled population is known, R^* can be estimated directly through

$$\hat{R}^* = N_0 P(y=1) / N_1 P(y=0).$$

The overall unknown parameters involved in our above setting are $\Theta = (k, R^*, \alpha, \beta, \{\pi_k\})$, where $\{\pi_k\}$ denote the joint haplotype frequencies of $h_k = (q_k, h_k)$ in the underlying population. In Spinka et al. [2005], another interception parameter $\beta_0 = k + \log R^*$ was used in the profile log-likelihood optimization procedure instead of using R^* . Here, we choose to replace the parameter β_0 by the sampling ratio R^* because the latter varies in a narrow region of $[0, 1]$ which makes a grid searching for it easier to implement. For the latent DSL in the prevalence model (2), we treat genotypes at the DSL as missing data for all sampled cases and controls. Note that the disease prevalence model of $P(y_i|h_q, z_i, R_i = 1)$ and the modifier function $\gamma(h_q, z_i)$ depend on genotypes at the DSL but not marker genotypes. The parameters can then be estimated through the following procedures.

EM algorithm:

- i. Set $t = 0$, and initialize the parameters $\Theta^{(0)}$.
- ii. Based on the current parameter setting of $\Theta^{(t)}$, update the haplotype frequencies $\{\pi_k\}$ using

$$\pi_k^{(t+1)} = \left(\frac{N_k^{(t)}}{2} \right) \left[\frac{\sum_{q=1}^N \frac{\sum P(q_k) P(q) \gamma((q_k, q), z_i)}{\sum_{h_q} P(h_q) \gamma(h_q, z_i)} \right]^{-1} \quad (2)$$

where $N_k^{(t)}$ is the expected total number of haplotype (q_k, h_k) in the sample under the current parameter setting which can be calculated by

$$N_k^{(t)} = \sum_{i=1}^N \frac{\sum_{h_q, h_m \in H(g_i)} n_k(h_q, h_m) P(h_q, h_m) \gamma(h_q, z_i) P(y_i|h_q, z_i, R_i=1)}{\sum_{h_q, h_m \in H(g_i)} P(h_q, h_m) \gamma(h_q, z_i) P(y_i|h_q, z_i, R_i=1)}$$

where $n_k(h_q, h_m)$ counts the number of haplotype (q_k, h_k) in diplotype (h_q, h_m) under the current parameter setting.

- iii. Based on the current parameter setting of $\Theta^{(t)}$, update the regression parameters $\theta = (k, \alpha, \beta)$ through a modified weighted logistic regression for maximizing the following function (see more details in the Appendix):

$$\begin{aligned} l^*(t) &= \sum_{i=1}^N \sum_{h_q \in Z^2} W_{i, h_q}^{(t)} \log P(y_i|h_q, z_i, R_i=1) \\ &= - \sum_{i=1}^N \sum_{h_q \in Z^2} V_{i, h_q}^{(t)} \log \gamma(h_q, z_i) \end{aligned}$$

where

$$W_{i,h_q}^{(t)} = \frac{\sum_{h_m \in H(g_i)} P_t(h_q, h_m) \gamma_t(h_q, z_i) P_t(y_i | h_q, z_i)}{\sum_{h_q} \sum_{h_m \in H(g_i)} P_t(h_q, h_m) \gamma_t(h_q, z_i) P_t(y_i | h_q, z_i)}$$

and

$$V_{i,h_q}^{(t)} = \frac{P_t(h_q) \gamma_t(h_q, z_i)}{\sum_{h_q} P_t(h_q) \gamma_t(h_q, z_i)}$$

for each individual $i = 1, 2, \dots, n$ and $h_q \in Z^2$.

- iv. We keep the parameter R^* as fixed in the updating procedures above. For the prospective likelihood-based method, take $R^* = 1$. For rare disease approximation, we set $R^* = 0$. For the retrospective likelihood-based method,

$\widehat{R}^* = N_0 P(y=1) / N_1 P(y=0)$ gives a direct estimate of R^* when the disease prevalence in the sampled population is known. If the disease prevalence is unknown, we can optimize $R^* \in [0, 1]$ separately based on a one-dimensional grid search.

It has been known that the EM algorithm cannot guarantee convergence to a global maximum of the likelihood function. In general, different starting points should be examined to initialize the parameters, and the optimum can then be selected as the one which achieves the largest likelihood value. To initialize the joint frequencies of the DSL and markers, we can first estimate the haplotype frequencies of markers only through the classical EM algorithm [Excoffier and Slatkin, 1995] using observed marker genotypes either in controls or the cases and controls combined. As the haplotype distributions in cases and controls are the same under the null hypothesis of no association, this estimation strategy would not change the asymptotic distribution of the LR test statistics under the null. Under the alternatives, however, the haplotype distributions in cases and controls would not be the same as the distribution of marker haplotypes in cases is likely to be distorted by disease phenotypes. So, using controls only to estimate the haplotype frequencies of markers is preferred as long as the control is a good representative of the sampled population and has a reasonable size of the control subjects. Given the frequency estimates of the marker haplotypes, the joint haplotype frequencies of the DSL and markers can be initialized by simply assigning the DSL on one of the markers. Through the updating procedures, the haplotype frequencies of the markers are iteratively updated along with the DSL. A single marker analysis can also be performed at the beginning to initialize the regression model parameters as well as to assign the DSL position for initialization of the joint haplotype frequencies of the DSL and markers.

The EM algorithm above allows missing data in marker genotypes, supposing that the missingness does not depend on observed marker genotypes and disease phenotypes, i.e., missing completely at random [Little and Rubin, 2002]. For an individual with missing genotypes at some marker loci, there are basically no constraints on possible phase-known genotypes at these loci. This leads to increased phase-known genotype configurations that are compatible with the observed marker genotypes. At the DSL, the genotypes of all individuals are treated as missing. So, we need to consider all possible configurations of diplotypes (q_{i0} , q_{i1}) for each individual i . We recognize that interchanging labels of the two alleles at the DSL does not change distribution of the disease phenotype. Besides, the baseline genotype of the DSL is not specified due to missing genotypes. So, in general, we might have more than one solution for MLE of the additive and dominance effects a , d that can give the same distribution

of disease phenotype. Similar to the identifiability problem in normal mixture distributions [McLachlan and Peel, 2000], here we define the parameter identifiability in terms of a class of mixture distributions and estimate the DSL effects by imposing constraints such as $a > 0$ or $d > 0$. Through the updating equation (2) above, we also see that the joint haplotype frequencies of the DSL and markers, which include the marker haplotype information, make use of the phenotype information through the posterior weights computed in the E-step. This approach could provide better estimation of haplotype frequencies since marker genotypes are supposedly to be correlated with disease phenotypes. On the other hand, the approach may be subject to possible mis-specification of the penetrance model.

Based on either the prospective or retrospective likelihood, we can construct the LR statistic $LR = 2(\ln L_F - \ln L_R)$ to test for association between the observed marker genotypes and the disease phenotype through a null hypothesis of $H_0 : a = d = 0$ vs. $H_a : a$ or $d \neq 0$. It has been well known that the finite mixture model is a non-regular parametric family, and classical results on the asymptotic properties of LR statistics may not be directly applied [Liang and Self, 1996]. In our case, under the null hypothesis of no DSL associated with the markers, the phenotypic values carry no information about the genotypic distribution of the DSL. So, the maximum likelihood of the reduced model L_R depends only on the mean of phenotypic values, the covariates, and the haplotype distribution of markers, and the joint haplotype distribution of DSL and markers contributes to the likelihood only through the DSL allele frequency. It

appears that the LR statistics under the null hypothesis asymptotically follow a χ^2_{df} with degrees of freedom $df = 3$ accounting for the parameters a , d , and the DSL allele frequency. Under the alternative hypothesis, computing the maximum likelihood of the full model L_F involves estimation of the joint haplotype distribution of DSL and markers as well as the model fitting of the phenotypic values. Note that the DSL-related parameters under our multinomial distributional model include allele frequency of the DSL, pairwise and various higher-order LDs between the DSL and marker alleles, which lead to a total number of $2^{m+1} - 2^m = 2^m$ independent parameters provided that all the 2^m marker haplotypes are present in the data. Therefore, a conjecture for the asymptotic distribution of the LR statistic $LR = 2(\ln L_F - \ln L_R)$ under the alternative H_a would be a noncentral χ^2 distribution χ^2_{df} with $df = 2^m + 2$, where 2 accounts for the parameters a and d . In practice, not all 2^m haplotypes of the markers may truly be involved in inducing the marker genotypes. Therefore, $df = df_m + 2 \leq 2^m + 2$ with df_m being the number of marker haplotypes that are truly present in the data.

SIMULATION RESULTS

Simulations are performed to evaluate the performance and properties of the three methods we described. To approximate the real setting, we adopt a simulation scheme used by Spinka et al. [2005] and Lin and Zeng [2006] from the Finland-United States NIDDM data set [Valle et al., 1998]. There are five biallelic SNPs within a candidate region on chromosome 22. The haplotypes and haplotype frequencies are listed as in Table I. For the five SNPs with eight possible haplotypes, we first randomly generate haplotypes based on this haplotype distribution. Then we randomly select haplotype pairs to form genotypes of individuals in the sampled population.

Assuming haplotype 5 (01100) as being the true disease-causing set of variants, we determine the disease status of each individual based on the logistic regression model: $\text{logit } P(y_i = 1 | h_i, z_i) = \beta_0 + a \cdot x(h_i) + b \cdot z_i + c \cdot x(h_i)z_i$, where $x(h_i)$ counts the number of risk haplotype (01100) in diplotype h_i of individual i , a is the (additive) haplotype effect, b is the effect of a pseudo-environmental covariate z , and c denotes the gene by environmental interaction. It has been known that population-based genetic association studies are complicated by the fact that a current population should be treated as just a single replicate sample generated through a

complex stochastic evolutionary process from its founder population. To account for variation from other coalescent populations, we set a random disease prevalence $\beta_0 \sim \text{Uniform}(-4, -3)$ which corresponds to varied disease prevalence between 1.6 and 7%. We also set the covariate z_i i.i.d. $\sim \text{Bernoulli}(0.2)$ being independent of the distribution of haplotype pairs with a fixed covariate effect $b = \log(1.5)$. For making inference on a and c , we consider odds ratios (OR) of 1, 1.25, 1.5, and 2 for the additive effect a , and OR of 1 (i.e., no interaction) and 1.5 for interaction effect c between DSL and the covariate. The cases and controls are retrospectively selected based on their disease status. In each sample, we select equal number of cases and controls. To compare the power, different number of cases of 200, 500, 800, and 1,000 are considered. We generate 500 replicate samples for each combination of these parameters.

For each simulated sample, we use three different methods to analyze the data: the prospective likelihood-based method (L1); the method based on simplified rare disease model (L2); and the retrospective likelihood-based method with known disease prevalence (L3). We use combined cases and controls to estimate the haplotype frequencies of markers and then initialize the joint haplotype frequencies of the DSL and markers by assigning the DSL as on marker 2. Since the five SNPs are tightly linked, we found that the algorithm was quite robust to different initialization points. To check the characteristics of the latent variable in our model, we also calculate the conditional probabilities $P(D|\text{hap5})$ of the risk allele given the true risk haplotype (i.e., haplotype 5) and the correlation coefficients r between the latent variable and the risk haplotype based on the estimated joint haplotype frequencies $q_j h_k$. Table II gives the means and the standard deviation of the parameter estimates from 500 replicates under various parameter settings.

On average, a higher OR level corresponds to stronger genetic effects in the penetrance model and therefore leads to better estimation of parameters. With increased sample size, we tend to get better estimates of parameters. The estimates of genetic effects for the putative DSL are reasonably close to the true parameters. Besides, as the genetic effects become stronger, the estimates of risk allele frequencies keep dropping toward the true frequency (i.e., 0.2514) of haplotype 5. Meanwhile, the conditional probability $P(D|\text{hap5})$ of the risk allele given the true risk haplotype becomes very close to 1, and the correlation coefficient r between the latent variable and the risk haplotype also increases toward 1. This indicates that the latent DSL in the model tends to play the same role as that of the true risk haplotype 5, even though the risk haplotype itself was not fitted into the penetrance model directly. Note that the correlation coefficient between haplotype 5 and the second marker on which we start our initialization is about 0.54, while the strongest correlation between haplotype 5 and the markers is achieved at marker 3 that has a value of 0.82.

Regarding the three estimation methods, we found that all the three methods provide quite compatible estimates for the OR parameters. But it appears that the prospective likelihood-based method generates greater bias in estimates of the risk allele frequencies, and the correlation coefficient between the risk allele and the true risk haplotype, perhaps due to the fact that this method does not account for the ascertainment of the case-control samples. The rare disease approximation method also has reduced bias in estimates of the risk allele frequencies but gives worse estimates in the conditional probabilities $P(D|\text{hap5})$ and the correlation coefficient r between the latent variable and the risk haplotype when there is haplotype by covariate interaction. We also noticed in our simulation studies that this method is not as stable as the other two methods and sometimes could not reach convergence, which might be caused by the boundlessness of the exponential function $\gamma(h_q, z_i)$ in this case. All the three methods, however, tend to have the DSL allele frequency and its additive effect over-estimated due to perhaps the large number of unknown parameters involved in modeling the joint haplotype distribution of the DSL and markers. Larger sample size might be required to overcome this problem.

A joint test is also performed for the composite association of no DSL additive effect and its interaction with the covariate (i.e., the null hypothesis of $H_0: a = c = 0$ vs. its alternative of $H_a: a \neq 0$ or $c \neq 0$) using the LR statistic. We first examined the asymptotic properties of the LR test statistic under the null hypothesis of no DSL and various alternative hypotheses. Figure 1 shows the histograms of the statistic values based on the retrospective likelihood-based method for the number of cases 200, 500, 800, and 1,000 and OR being 1.0, 1.2, 1.5, and 2.0 of the additive effect without gene by environmental interaction. Histograms of the statistic values from the other two methods have very similar patterns (not shown). Under the null H_0 , it seems that the LR statistics approximately follow a χ^2_3 distribution as sample size increases. With increased additive effects or sample sizes, the LR statistics tend to take greater values and therefore provide better power.

To calculate the power under various scenarios, we first computed the 95th percentiles of the statistic values under the null hypothesis of no DSL and interaction as the cut point for rejection of the null hypothesis according to different sample sizes and estimation methods under each combination of parameters. Then we calculated the proportion of the LR statistic values that are greater than or equal to the cut point as power for various parameter settings. To compare the power of our latent variable methods with the classical haplotype analysis, we also performed the standard haplotype analysis on association testing of haplotype 5 using the prospective and retrospective likelihood-based methods proposed by Spinka et al. [2005] (H1—prospective; H2—retrospective). Figure 2 shows power curves from the three latent variable methods and the two standard haplotype methods for the number of cases of 200, 500, 800, and 1,000 under four sets of OR parameters. Overall, the three latent variable methods are quite compatible in terms of power, although the retrospective likelihood-based method has about 9% gain in power when the OR of the additive and interaction effects are 1.5 and 1.5, respectively. It seems that the classical haplotype association methods have slightly better power than the latent variable methods in this case. This is no surprise because haplotype 5 is the true disease variant and we only test for association of haplotype 5 in the H1 and H2 methods.

As we have pointed out before, under certain circumstances, the latent variable approach may provide better power than the classical haplotype analysis. Here, we consider a scenario in which the true disease variant is given by a sub-haplotype $r_{35} = [1\ 0]$ with allele '1' at SNP 3 and '0' at SNP 5. Suppose the r_{35} has an additive effect on the disease phenotype. We performed both the prospective (L1) and the retrospective (L3) methods using the latent variable approach and methods H1 and H2 for the classical haplotype analysis, with H1 and H2 testing for association of haplotype 5 only. Note that haplotype 5 is the only common haplotype (over the 10% threshold) that contains the true risk variant r_{35} . Table III gives the powers for the four model fitting methods according to different sample sizes and OR parameters. With the same 5% significance level, the latent variable methods show clear power advantages over the classical haplotype association methods at least for number of cases 500 and 1,000 with the additive OR 1.2, and number of cases 200 and 500 with the additive OR 1.5, due to perhaps the fact that the latent variable methods tend to capture the effect of the sub-haplotype r_{35} across all the common and rare haplotypes 2, 3, 5, 10, 11, and 13 in which r_{35} is embedded. In practice, we do not know which haplotype really contains the true risk sub-haplotype. A standard haplotype analysis needs to be tested for all the four common haplotypes 4, 5, 9, and 12. After adjustment for multiple testing, the actual power of the classical haplotype analysis would be even lower than that of the H1 and H2 given in Table III because the latter used more information on haplotypes that contain the true disease variant.

DISCUSSION

In this article, we introduced a latent variable approach to test for haplotype-based association between a disease and multiple SNPs within a candidate region of interest. As a retrospective likelihood-based method, this approach can correct for the sampling ascertainment of the case-control samples when knowledge of the sampling rates in cases and controls or prevalence of disease is available. It can also properly handle the HWE constraint on the genotypic distribution of markers. Meanwhile, it can appropriately account for uncertainty in both haplotype frequency estimates and the assignment of haplotype pairs to individuals. Estimates of the joint haplotype frequency of the latent DSL and markers also provide useful information for further dissection of the status of the DSL by examining, for example, the correlation structure between the DSL and markers [Wang et al., 2006].

This latent variable approach has some special features as a complement to the existing haplotype association methods. Firstly, the latent variable approach may show a power gain compared to the simple haplotype analysis when the underlying disease variant is embedded in several haplotypes. In this case, effect of the disease variant from one haplotype may be attenuated due to ignoring of other haplotypes — rare or common. Its detection could also be completely missed when it is embedded in only rare haplotypes even though itself may not be so rare in the sampled population. For example, in our simulation setting, the sub-haplotype $r_{13} = [1 \ 1]$ with both alleles ‘1’ at SNP loci 1 and 3 is embedded only in several rare haplotypes but has a population frequency of 0.097. On the other hand, the latent variable approach tends to capture the characteristics of a risk sub-haplotype across all embedded common and rare haplotypes. Secondly, the standard haplotype analysis often tests for all common haplotypes within a block separately which requires an adjustment for testing of multiple haplotypes with a raised criterion for declaring significance. With the flexible role of the latent variable in the penetrance model to characterize a putative DSL with unknown status, an allele, or a haplotype from a subset of the markers, the LR statistics from the latent variable approach can provide a comprehensive association test between a disease and a set of markers regardless of the underlying composition of the true disease susceptible variant and therefore alleviate the multiple testing issue.

Recently, genome-wide association (GWA) studies on complex diseases using hundreds of thousands of SNPs are gaining momentum. As a strategy to simplify the complex structural variation of the genome, haplotype blocks have been applied to characterize variation of multiple tightly linked SNPs. However, identification of the haplotype blocks is often problematic and different methods may give inconsistent results. Testing of only haplotypes from the whole set of SNPs within blocks is also questionable due to the unknown status of the true functional disease variant. Alternatively, the latent variable approach can potentially be applied to GWA studies by incorporating it with the classical sliding window approach. Since the latent variable approach can provide a comprehensive association test for a moderate number of multiple linked SNPs within a candidate region of interest and these SNPs do not have to be in high LD within a block, we can choose a fixed or variable length of windows (e.g., with a moderate number of 5–15 tagged SNPs) sliding across the genome and test for composite association within each window using the approach. An LOD score profile can then be constructed on each chromosome for further identification of candidate gene regions after an appropriate adjustment for testing of the multiple windows.

Through our simulation studies, we found that both the prospective and the retrospective likelihood-based methods can provide quite compatible power in testing the genetic effects. The latter appears to be more efficient in the estimation of the DSL allele frequencies but it is computationally more intensive and requires some prior knowledge of the disease prevalence or the sampling rates in cases and controls. In general, the power of these methods depends

not only on sample sizes and genetic effects but also the number of SNPs and their underlying haplotype structures. It has to be pointed out that our simulation is limited to a small number of genetic markers and the particular set of markers we used are unlikely to be representative for other genomic regions with different haplotype structures. We also noticed that the computing time mainly depends on the number of SNPs involved at one time. For our simulation data, it usually takes less than 1 minute to run one sample using our Matlab program on a desktop PC with an Intel Dual Core 2.40 GHz processor.

The latent variable approach we proposed has some limitations. First, we assume that the case-control samples come from a underlying homogeneous population. An admixture of the sampled population may generate spurious association and inflate the false-positive rate. Fortunately, classical epidemiological strategies could help for adjustment of population admixture through population stratification. A stratification of the mixture logistic regression on population admixture is also plausible. Second, in estimation of haplotype frequencies of DSL and markers, a gametic phase equilibrium in the sampled population is assumed across the candidate region including the DSL and marker loci. It is known that there exists a low level of inbreeding in human population due to limited population size during the evolutionary history. With the inbreeding, the gametic phase equilibrium may be violated slightly. Departure from this equilibrium may lead to biasness in the parameter estimation. Robustness of the method to this departure needs further exploration. Third, the LR statistic can provide a composite test for association but it does not specify whether the association is due to a marker allele, a sub-haplotype, or a haplotype of the markers, even though the joint haplotype frequency of the latent DSL and markers may help for further dissection of the DSL status in certain circumstances. Finally, common diseases are most likely contributed by more than one disease genes. Current method assumes a single disease variant within a candidate region. An extension of the method to multiple disease variants is straightforward when they are located on unlinked candidate regions separately. If more than one disease variant is located within a candidate region, however, this method may have difficulty in modeling the correlation structure among these variants. Further exploration on these issues is required.

ACKNOWLEDGMENTS

The authors thank two anonymous reviewers and Dr. Raymond Hoffmann in the Division of Biostatistics at the Medical College of Wisconsin for their constructive comments. We acknowledge the technical assistance of John Blimke in the Department of Pediatrics at the Medical College of Wisconsin. This work was partially supported by the National Institute of Diabetes and Digestive and Kidney Diseases under grant R01DK080100 and the Biostatistical Genetics Program Development Fund from the Medical College of Wisconsin.

Contract grant sponsor: National Institute of Diabetes and Digestive and Kidney Diseases; Contract grant number: R01DK080100; Contract grant sponsor: Medical College of Wisconsin.

Appendix

APPENDIX

If we recode $y_i = 0$ as $y_i = -1$, then the logistic regression model

$$P(y_i|h, z_i, R_i=1) = \frac{\exp\{y_i(k + z_i\alpha + X(h_q)\beta)\}}{1 + \exp\{k + z_i\alpha + X(h_q)\beta\}}$$

can be re-written as

$$P(y_i = \pm 1 | h, z_i, R_i = 1) = \frac{1}{1 + \exp\{-y_i(k + z_i\alpha + X(h_q)\beta)\}}.$$

To simplify the notation, we write $k + z_i\alpha + X(h_q)\beta = \theta^T X_i$, where $\theta = (k, \alpha, \beta)'$. Then, through some simple derivation, we can show that the gradient of $l^*(t)$ is given by

$$\begin{aligned} \nabla_{\theta} l^*(t) &= \frac{\partial l^*(t)}{\partial \theta} \\ &= \sum_{i=1}^N \sum_{h_q} W_{i,h_q}^{(t)} \frac{\partial \log P(y_i | h, z_i, R_i = 1)}{\partial \theta} \\ &\quad + \sum_{i=1}^N \sum_{h_q} \left(W_{i,h_q}^{(t)} - V_{i,h_q}^{(t)} \right) \frac{\partial \log \gamma(h, z_i)}{\partial \theta}. \end{aligned}$$

Define $\sigma(y_i \theta^T X_i) = 1/(1 + \exp(-y_i \theta^T X_i))$. Note also that $\log \gamma(h_q, z_i) = \log[1 + \exp(\theta^T X_i)] - \log[1 + R^* \cdot \exp(\theta^T X_i)]$. Therefore,

$$\begin{aligned} \nabla_{\theta} l^*(t) &= \sum_{i=1}^N \sum_{h_q} W_{i,h_q}^{(t)} (1 - \sigma(y_i \theta^T X_i)) (y_i X_i) \\ &\quad + \sum_{i=1}^N \sum_{h_q} \left(W_{i,h_q}^{(t)} - V_{i,h_q}^{(t)} \right) \\ &\quad \times \left[\frac{1}{1 + R^* \cdot \exp(\theta^T X_i)} - \frac{1}{1 + \exp(\theta^T X_i)} \right] X_i. \end{aligned}$$

The Hessian matrix of $l^*(t)$ is given by

$$\begin{aligned} H &= \frac{\partial^2 l^*(t)}{\partial \theta \partial \theta^T} = - \sum_{i=1}^N \sum_{h_q} \left[V_{i,h_q}^{(t)} \sigma(\theta^T X_i) (1 - \sigma(\theta^T X_i)) \right. \\ &\quad \left. + \left(W_{i,h_q}^{(t)} - V_{i,h_q}^{(t)} \right) \sigma^*(\theta^T X_i) (1 - \sigma^*(\theta^T X_i)) \right] (X_i X_i^T) \end{aligned}$$

where $\sigma^*(\theta^T X_i) = 1/(1 + R^* \cdot \exp(\theta^T X_i))$. A Newton-Raphson-type optimization procedure can then be constructed to search for the maximum $\theta = (k, \alpha, \beta)'$ of function $l^*(t)$.

REFERENCES

- Becker T, Knapp M. A powerful strategy to account for multiple testing in the context of haplotype analysis. *Am J Hum Genet* 2004;75:561–570. [PubMed: 15290652]
- Browning BL, Browning SR. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet Epidemiol* 2007;31:365–375. [PubMed: 17326099]
- Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation in case-control studies of gene environment interactions. *Biometrika* 2005;92:399–418.
- Cheng KF, Lin WJ. Retrospective analysis of case-control studies when the population is in Hardy-Weinberg equilibrium. *Stat Med* 2005;24:3289–3310. [PubMed: 16206241]
- Clark AG. The role of haplotypes in candidate gene studies. *Genet Epidemiol* 2004;27:321–333. [PubMed: 15368617]
- Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 2003;73:1316–1329. [PubMed: 14631556]

- Excoffier L, Slatkin M. Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995;12:921–927. [PubMed: 7476138]
- Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol* 2007;31:383–395. [PubMed: 17410554]
- Lake SL, Lyon H, Tanisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered* 2003;55:56–65. [PubMed: 12890927]
- Liang KY, Self SG. On the asymptotic behavior of the pseudolikelihood ratio test statistic. *J R Stat Soc B* 1996;58:785–796.
- Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies. *J Am Stat Assoc* 2006;101:89–104.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. Vol. 2nd edition. Wiley; New Jersey: 2002.
- McLachlan, G.; Peel, D. *Finite Mixture Models*. Wiley; New York, USA: 2000.
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979;66:403–412.
- Roeder K, Carroll RJ, Lindsay BG. A semiparametric mixture approach to case-control studies with errors in covariates. *J Am Stat Assoc* 1996;91:722–732.
- Schaid JD. Evaluating association of haplotypes with trait. *Genet Epidemiol* 2004;27:348–364. [PubMed: 15543638]
- Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002;70:425–434. [PubMed: 11791212]
- Spinka C, Carroll RJ, Chatterjee N. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genet Epidemiol* 2005;29:108–127. [PubMed: 16080203]
- Stram B, Pearce L, Bretsky P, Freedman M, Hirschhorn J, Altshuler D, Kolonel L, Henderson B, Thomas D. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum Hered* 2003;55:179–190. [PubMed: 14566096]
- Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, Nylund SJ, Kohtamaki K, Toivanen L, Vidgren G, Tuomilehto-Wolf E, Ehnholm C, Blaschak J, Langefeld CD, Watanabe RM, Magnuson V, Ally DS, Hagopian WA, Ross E, Buchanan TA, Collins F, Boehnke M. Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study. *Diabetes Care* 1998;21:949–958. [PubMed: 9614613]
- Wang T, Weir B, Zeng ZB. A population-based latent variable approach for association mapping of quantitative trait loci. *Ann Hum Genet* 2006;70:506–523. [PubMed: 16759182]
- Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 2002;53:79–91. [PubMed: 12037407]
- Zeng ZB. Precision mapping of quantitative trait loci. *Genetics* 1994;136:1457–1468. [PubMed: 8013918]

Fig. 1. Histograms of the LR statistics for number of cases of 200, 500, 800, and 1,000 and odds ratios 1, 1.2, 1.5, and 2 of the additive effect (no interaction) using the retrospective likelihood method with 500 replicates.

Fig. 2.

Power curves for number of cases of 200, 500, 800, and 1,000 and odds ratios (1.2,1), (1.5,1), (2,1), and (1.5,1.5) of the additive and interaction effects.

TABLE I

Haplotype distribution and allele frequencies

Haplotypes	SNP loci					Hap. frequencies
	1	2	3	4	5	
1	0	0	0	1	1	0.0042
2	0	0	1	0	0	0.0035
3	0	0	1	1	0	0.0018
4	0	1	0	1	1	0.1292
5	0	1	1	0	0	0.2514
6	0	1	1	0	1	0.0012
7	0	1	1	1	1	0.0019
8	1	0	0	0	0	0.0136
9	1	0	0	1	1	0.3573
10	1	0	1	0	0	0.0521
11	1	0	1	1	0	0.0317
12	1	1	0	1	1	0.1392
13	1	1	1	0	0	0.0109
14	1	1	1	1	1	0.0020
Frequencies of allele "1"	0.6194	0.5453	0.3607	0.6720	0.6403	1

TABLE II
Means and standard deviations of parameter estimates: with number of cases = 200, 500 and, 800, 1,000

# Cases	e^a	e^c	e^d	e^e	$\hat{P}_D$	$\hat{P}(D hap\delta)$	$\hat{r}$
(a) 200 (Method L1)	1	1	1.29 (0.23)	1.02 (0.55)	0.44 (0.04)	0.94 (0.23)	0.66 (0.06)
	1.2	1	1.39 (0.31)	0.91 (0.52)	0.45 (0.04)	0.90 (0.31)	0.67 (0.06)
	1.5	1	1.62 (0.42)	1.05 (0.52)	0.44 (0.04)	0.94 (0.25)	0.72 (0.07)
	2	1	2.22 (0.49)	1.22 (0.73)	0.42 (0.04)	0.95 (0.22)	0.80 (0.07)
	1.5	1.5	1.90 (0.68)	1.72 (0.92)	0.44 (0.04)	0.75 (0.42)	0.71 (0.12)
	2	1.5	2.57 (0.76)	1.83 (0.88)	0.44 (0.04)	0.84 (0.36)	0.77 (0.09)
200 (Method L2)	1	1	1.27 (0.21)	1.00 (0.43)	0.45 (0.04)	0.87 (0.34)	0.64 (0.05)
	1.2	1	1.34 (0.27)	0.93 (0.32)	0.44 (0.05)	0.91 (0.29)	0.66 (0.06)
	1.5	1	1.58 (0.36)	1.01 (0.35)	0.39 (0.06)	0.94 (0.23)	0.72 (0.09)
	2	1	2.07 (0.36)	0.92 (0.25)	0.35 (0.07)	0.96 (0.20)	0.78 (0.12)
	1.5	1.5	1.76 (0.60)	1.57 (0.78)	0.40 (0.08)	0.82 (0.37)	0.67 (0.14)
	2	1.5	2.46 (1.04)	1.69 (0.96)	0.32 (0.16)	0.75 (0.42)	0.57 (0.28)
200 (Method L3)	1	1	1.28 (0.21)	0.98 (0.42)	0.45 (0.04)	0.87 (0.34)	0.64 (0.05)
	1.2	1	1.35 (0.27)	0.91 (0.34)	0.44 (0.05)	0.90 (0.30)	0.66 (0.06)
	1.5	1	1.59 (0.36)	1.06 (0.41)	0.39 (0.06)	0.95 (0.21)	0.73 (0.07)
	2	1	2.18 (0.40)	0.99 (0.33)	0.35 (0.05)	0.98 (0.13)	0.81 (0.07)
	1.5	1.5	1.85 (0.63)	1.78 (0.88)	0.39 (0.06)	0.89 (0.29)	0.71 (0.11)
	2	1.5	2.52 (0.72)	1.95 (0.94)	0.35 (0.06)	0.98 (0.14)	0.79 (0.10)
500 (Method L1)	1	1	1.12 (0.11)	0.95 (0.29)	0.45 (0.02)	0.98 (0.13)	0.64 (0.03)
	1.2	1	1.26 (0.17)	1.02 (0.34)	0.44 (0.03)	0.99 (0.08)	0.68 (0.04)
	1.5	1	1.60 (0.23)	1.04 (0.28)	0.42 (0.04)	0.99 (0.08)	0.75 (0.06)
	2	1	2.15 (0.27)	1.04 (0.31)	0.40 (0.03)	1.00 (0.00)	0.84 (0.05)
	1.5	1.5	1.76 (0.36)	1.75 (0.67)	0.43 (0.04)	0.84 (0.36)	0.73 (0.10)
	2	1.5	2.41 (0.54)	1.76 (0.68)	0.43 (0.04)	0.89 (0.31)	0.79 (0.09)
500 (Method L2)	1	1	1.12 (0.10)	0.96 (0.23)	0.46 (0.02)	0.98 (0.15)	0.64 (0.03)
	1.2	1	1.25 (0.17)	0.99 (0.22)	0.43 (0.04)	0.99 (0.08)	0.67 (0.04)
	1.5	1	1.57 (0.21)	1.02 (0.18)	0.37 (0.05)	1.00 (0.00)	0.76 (0.06)
	2	1	2.08 (0.27)	0.94 (0.15)	0.32 (0.06)	0.98 (0.15)	0.82 (0.13)
	1.5	1.5	1.75 (0.44)	1.77 (0.64)	0.40 (0.07)	0.88 (0.30)	0.68 (0.12)
	2	1.5	2.33 (0.83)	1.50 (0.54)	0.30 (0.17)	0.74 (0.43)	0.55 (0.30)

# Cases	e^a	e^c	e^d	$\hat{P}_D$	$\hat{P}(D \text{hap5})$	$\hat{r}$
500 (Method L3)	1	1	1.12 (0.10)	0.45 (0.02)	0.98 (0.13)	0.64 (0.03)
	1.2	1	1.25 (0.18)	0.43 (0.04)	0.99 (0.08)	0.67 (0.04)
	1.5	1	1.57 (0.21)	0.37 (0.05)	1.00 (0.00)	0.76 (0.06)
	2	1	2.14 (0.25)	0.32 (0.03)	1.00 (0.00)	0.85 (0.05)
	1.5	1.5	1.76 (0.35)	0.38 (0.05)	0.95 (0.20)	0.74 (0.09)
800 (Method L1)	2	1.5	2.50 (0.58)	0.35 (0.05)	0.99 (0.11)	0.80 (0.09)
	1	1	1.11 (0.08)	0.45 (0.02)	0.99 (0.08)	0.64 (0.03)
	1.2	1	1.23 (0.16)	0.44 (0.03)	0.99 (0.08)	0.68 (0.04)
	1.5	1	1.59 (0.16)	0.40 (0.03)	1.00 (0.06)	0.77 (0.05)
	2	1	2.10 (0.21)	0.39 (0.03)	1.00 (0.00)	0.86 (0.04)
800 (Method L2)	1.5	1.5	1.67 (0.26)	0.42 (0.04)	0.91 (0.28)	0.75 (0.08)
	2	1.5	2.32 (0.42)	0.42 (0.04)	0.92 (0.27)	0.81 (0.08)
	1	1	1.10 (0.07)	0.46 (0.02)	0.99 (0.08)	0.64 (0.02)
	1.2	1	1.22 (0.15)	0.43 (0.03)	1.00 (0.00)	0.67 (0.04)
	1.5	1	1.57 (0.14)	0.36 (0.04)	1.00 (0.00)	0.78 (0.05)
800 (Method L3)	2	1	2.06 (0.19)	0.31 (0.03)	0.99 (0.09)	0.85 (0.08)
	1.5	1.5	1.71 (0.30)	0.38 (0.07)	0.94 (0.22)	0.70 (0.12)
	2	1.5	2.16 (0.71)	0.29 (0.16)	0.76 (0.42)	0.57 (0.31)
	1	1	1.10 (0.08)	0.46 (0.02)	0.99 (0.08)	0.64 (0.02)
	1.2	1	1.22 (0.15)	0.43 (0.03)	1.00 (0.00)	0.67 (0.04)
1,000 (Method L1)	1.5	1	1.57 (0.14)	0.36 (0.04)	1.00 (0.00)	0.78 (0.05)
	2	1	2.09 (0.18)	0.31 (0.03)	1.00 (0.00)	0.87 (0.04)
	1.5	1.5	1.67 (0.25)	0.36 (0.05)	0.99 (0.09)	0.77 (0.08)
	2	1.5	2.39 (0.40)	0.34 (0.05)	1.00 (0.04)	0.82 (0.08)
	1	1	1.10 (0.07)	0.45 (0.01)	0.99 (0.11)	0.64 (0.02)
1,000 (Method L2)	1.2	1	1.21 (0.14)	0.44 (0.03)	0.98 (0.13)	0.68 (0.04)
	1.5	1	1.58 (0.14)	0.40 (0.03)	1.00 (0.00)	0.78 (0.05)
	2	1	2.08 (0.18)	0.38 (0.02)	1.00 (0.00)	0.87 (0.04)
	1.5	1.5	1.65 (0.22)	0.42 (0.04)	0.93 (0.24)	0.77 (0.07)
	2	1.5	2.28 (0.35)	0.42 (0.04)	0.95 (0.23)	0.82 (0.08)
1,000 (Method L3)	1	1	1.08 (0.07)	0.46 (0.02)	1.00 (0.00)	0.63 (0.01)

# Cases	e^a	e^c	e^a	e^c	$\hat{P}_D$	$\hat{P}(D \text{hapS})$	$\hat{r}$
1,000 (Method L3)	1.2	1	1.21 (0.13)	1.01 (0.15)	0.43 (0.03)	0.99 (0.08)	0.67 (0.04)
	1.5	1	1.57 (0.12)	1.00 (0.13)	0.35 (0.03)	1.00 (0.00)	0.79 (0.05)
	2	1	2.07 (0.15)	0.93 (0.09)	0.31 (0.02)	1.00 (0.00)	0.86 (0.03)
	1.5	1.5	1.66 (0.24)	1.66 (0.39)	0.37 (0.07)	0.96 (0.19)	0.71 (0.12)
	2	1.5	2.17 (0.64)	1.44 (0.36)	0.29 (0.14)	0.81 (0.39)	0.61 (0.30)
	1	1	1.09 (0.07)	0.97 (0.14)	0.46 (0.02)	1.00 (0.00)	0.64 (0.01)
	1.2	1	1.21 (0.13)	1.01 (0.16)	0.43 (0.03)	0.99 (0.08)	0.67 (0.04)
	1.5	1	1.57 (0.13)	1.01 (0.13)	0.35 (0.03)	1.00 (0.00)	0.79 (0.05)
	2	1	2.08 (0.18)	0.97 (0.13)	0.31 (0.02)	1.00 (0.00)	0.88 (0.03)
	1.5	1.5	1.66 (0.24)	1.70 (0.48)	0.36 (0.05)	0.99 (0.08)	0.78 (0.08)
	2	1.5	2.38 (0.37)	1.57 (0.61)	0.33 (0.04)	1.00 (0.01)	0.83 (0.08)

TABLE III
Power of the latent variable methods (L1 and L3) and classical haplotype association methods (H1 and H2) for number of cases 200, 500, 800, and 1,000 at a significance level of $\alpha = 0.05$

Methods	Odds ratio	200	500	800	1,000
L1	$e^d = 1$	0.049	0.049	0.051	0.051
L3		0.050	0.051	0.051	0.051
H1		0.050	0.050	0.050	0.050
H2		0.049	0.050	0.051	0.051
L1	$e^d = 1.2$	0.151	0.431	0.542	0.624
L3		0.152	0.428	0.533	0.658
H1		0.118	0.337	0.471	0.531
H2		0.123	0.344	0.500	0.537
L1	$e^d = 1.5$	0.620	0.982	0.998	1.000
L3		0.597	0.975	0.996	1.000
H1		0.457	0.895	0.991	0.998
H2		0.382	0.873	0.974	0.981
L1	$e^d = 2$	0.992	1.000	1.000	1.000
L3		0.994	1.000	1.000	1.000
H1		0.893	1.000	1.000	1.000
H2		0.684	0.987	0.997	1.000