Original Paper

Human Heredity

Hum Hered 2006;61:104–110
DOI: 10.1159/000093476

Received: January 12, 2006
Accepted after revision: March 13, 2006
Published online: May 24, 2006

Estimating Haplotype Effects on Dichotomous Outcome for Unphased Genotype Data Using a Weighted Penalized Log-Likelihood Approach

Olga W. Souverein Aeilko H. Zwinderman Michael W.T. Tanck

Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, The Netherlands

Key Words
Haplotype analysis · Association study · Penalty function · EM algorithm

Abstract
Objective: To develop a method to estimate haplotype effects on dichotomous outcomes when phase is unknown, that can also estimate reliable effects of rare haplotypes. ***Methods:*** In short, the method uses a logistic regression approach, with weights attached to all possible haplotype combinations of an individual. An EM-algorithm was used: in the E-step the weights are estimated, and the M-step consists of maximizing the joint log-likelihood. When rare haplotypes were present, a penalty function was introduced. We compared four different penalties. To investigate statistical properties of our method, we performed a simulation study for different scenarios. The evaluation criteria are the mean bias of the parameter estimates, the root of the mean squared error, the coverage probability, power, Type I error rate and the false discovery rate. ***Results:*** For the unpenalized approach, mean bias was small, coverage probabilities were approximately 95%, power ranged from 15.2 to 44.7% depending on haplotype frequency, and Type I error rate was around 5%. All penalty functions reduced the standard errors of the rare haplotypes, but introduced bias. This trade-off decreased power. ***Conclusion:*** The unpenalized weighted log-likelihood approach performs well. A penalty function can help to estimate an effect for rare haplotypes.

Introduction

Recent interest has been to associate haplotypes with common complex diseases as a way to identify causal genetic variants. However, in most genetic association studies haplotype information is not available since the study population consists of unrelated individuals of whom the genotypes are determined independently. Haplotypes of individuals who are homozygous or heterozygous at only one locus are unambiguous (i.e., phase is known), since there is only one possible haplotype pair for these individuals. However, haplotypes of multiple heterozygotes are ambiguous (unknown phase) and it is necessary to rely on statistical methods to deduce haplotypes of these individuals. Several methods have been developed to infer haplotypes and haplotype frequencies from unphased genotype data [reviewed by Niu, 1]. Besides algorithms to estimate haplotype frequencies, methods have also been developed to associate haplotypes with disease or phenotype [reviewed by Schaid et al., 2]. Several of these methods are based on cladistic analysis, genealogy or clustering of haplotypes [e.g. 3–6], while others are based on regression models [e.g. 7–14.]

One of these regression methods, described by Tanck et al. [12], is a method of weighted penalized log-likelihood to estimate haplotype effects on continuous outcome data incorporating the uncertainty about phase ambiguous individuals as weights in the model. The present study is a generalisation of this method to dichotomous outcome data.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2006 S. Karger AG, Basel
0001–5652/06/0612–0104$23.50/0

Accessible online at:
www.karger.com/hhe

O.W. Souverein
Academic Medical Center, Department of Clinical Epidemiology and Biostatistics
PO Box 22700
NL–1100 DE Amsterdam (The Netherlands)
Tel. +31 20 566 6945, Fax +31 20 691 2683, E-Mail o.w.souverein@amc.uva.nl

Since the parameter estimates of rare haplotypes often show large variances which could lead to model instability, a penalty function was introduced to shrink these effects. Shrinking the effects of rare haplotypes is, theoretically, an appealing approach [2]. Other approaches to deal with rare haplotypes, for example pooling all rare haplotypes in one category or pooling the rare haplotypes with similar common haplotypes, often lead to results that are hard to interpret. The penalty function used by Tanck et al. [12] is based on the assumption that similar haplotypes show similar effects. However, practice and simulations indicate that this might not always be the best choice. Therefore, this current study also compares four different penalty functions.

Methods

Data and Model

Consider a sample of N unrelated individuals, where G_i is the genotypic vector of the i-th individual.

First, haplotype frequencies were estimated using for instance the EM algorithm of Excoffier and Slatkin [15]. Second, haplotypes were assigned to all unambiguous individuals. For the remaining ambiguous individuals the number of haplotype pairs (k_i) compatible with their genotype (G_i) were determined and the posterior probabilities (w_{ij}) were calculated using Bayes' theorem given the estimated haplotype frequencies (p) under the assumption that the underlying population is in Hardy-Weinberg equilibrium (HWE)

$$w_{ij} = P(j|G_i) = \frac{p(h)p(r)d_{hri}}{\sum_{h=1}^{m}\sum_{r=1}^{m} p(h)p(r)d_{hri}} \tag{1}$$

where j is haplotype pair j ($j = 1, \dots, k_i$), h and r the haplotypes forming the haplotype pair j ($h, r = 1, \dots, m$), m is the number of haplotypes estimated to be present in the population, and d_{hri} is an indicator function, which is 1 when haplotype pair (h, r) is compatible with G_i and 0 otherwise. This is identical to the procedure described by Tanck et al. [12].

The logistic regression model had the following form:

$$\ln\left[\frac{\pi_{ij}}{1-\pi_{ij}}\right] = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_{m-1} X_{ijm-1} + \cdots, \tag{2}$$

where X_{ijr} ($r = 1, \dots, m - 1$) attains values 0, 1 or 2 denoting presence of 0, 1 or 2 copies of haplotype r in haplotype pair j of patient i, assuming that haplotypes show additive effects, β_r ($r = 1, \dots, m - 1$) are the parameters to be estimated, m denotes the number of haplotypes estimated to be present in the population, and $\pi_{ij} = e^{\beta X_{ij}}/(1 + e^{\beta X_{ij}})$. The most frequent haplotype was chosen to be the reference category. The haplotype effects (β) can be estimated by maximizing the log-likelihood

$$l(\beta) = \sum_{i=1}^{N} log\left(\sum_{j} \pi_{ij}^{\ y_i}\left(1-\pi_{ij}\right)^{1-y_i} w_{ij}\right), \tag{3}$$

where y_i denotes the dichotomous outcome variable for individual i. Estimating equation for β was derived by equating to zero the first order derivative of this log-likelihood

$$\begin{aligned}\frac{\partial l}{\partial \beta_r} &= \sum_i \sum_j X_{ijr}\left(y_i - \pi_{ij}\right)\frac{w_{ij}\left(e^{y_i\beta x_{ij}}/\left(1+e^{\beta x_{ij}}\right)\right)}{\sum_j w_{ij}\left(e^{y_i\beta x_{ij}}/\left(1+e^{\beta x_{ij}}\right)\right)}\\ &= \sum_i \sum_j X_{ijr}\left(y_i - \pi_{ij}\right) f_{ij} = 0.\end{aligned} \tag{4}$$

However, maximizing this log-likelihood is difficult since f_{ij} depends on β. Therefore, we chose to use an EM algorithm [16] in which we maximize the expectation of the joint log-likelihood of y_i, and $X_{ij1}, \dots, X_{ijm}$, given provisional estimates of the parameters (β^0, w^0)

$$\begin{aligned}E[l^*] &= \sum_i E\left[\ln\left(P\left(Y_i = y_i, X_i = x_i \middle| G_i = g_i\right)\right)\middle|\left(\beta^0, w^0\right)\right]\\ &= \sum_i \sum_j \left[y_i \ln\left(\pi_{ij}\right) + \left(1-y_i\right)\ln\left(1-\pi_{ij}\right) + \ln w_{ij}\right] f_{ij},\end{aligned} \tag{5}$$

where f_{ij} is the posterior probability that $X_i = x_{ij}$ given the data evaluated with (β^0, w^0)

$$\begin{aligned}f_{ij} = P\left(X_i = x_{ij} \mid Y_i, G_i\right) &= \frac{\pi_{ij}^{y_i}\left(1-\pi_{ij}\right)^{1-y_i} w_{ij}^0}{\sum_{j=1}^{k_i}\pi_{ij}^{y_i}\left(1-\pi_{ij}\right)^{1-y_i} w_{ij}^0}\\ &= \frac{w_{ij}^0\left(e^{y_i\beta^0 x_{ij}}/\left(1+e^{\beta^0 x_{ij}}\right)\right)}{\sum_{j=1}^{k_i} w_{ij}^0\left(e^{y_i\beta^0 x_{ij}}/\left(1+e^{\beta^0 x_{ij}}\right)\right)}.\end{aligned} \tag{6}$$

In the M-step of the EM algorithm f_{ij} was fixed and β and w_{ij} were estimated by maximizing (5) using Newton-Raphson algorithm. The estimating equation of β for the expectation of the joint log-likelihood (5) is identical to the estimating equation of β for the log-likelihood, which is shown in equation (4). In the E-step, f_{ij} was re-estimated using (6) with β and w_{ij} from the M-step. These two steps were iterated until the parameter estimates reached convergence.

The weighted (penalized) logistic regression maximization routine was programmed in MATLAB® 7.0 (The Mathworks, Natick, Mass., USA) and is freely available upon request from the corresponding author.

The Penalty Functions

The estimation of rare haplotypes often shows large variation. To circumvent this problem, Tanck et al. [12] proposed using a penalty function. A penalty function reduces the standard errors of the parameter estimates (β) at the cost of introducing a small bias in these estimates. To estimate the haplotype effects the EM algorithm was used as described above.

We considered four different penalty functions. The first one was the ridge penalty

$$l_{pen1} \propto l(\beta) - \frac{1}{2}\lambda\sum_r \beta_r^2, \tag{7}$$

Table 1. Frequency, mean bias, mean SE, coverage probability, power and Type I error rate of weighted unpenalized logistic regression[a]

Haplotype	Frequency	Mean bias ($\times 10^2$)	Mean SE	Coverage probability	Power	Type I error rate
00111	0.08	–0.49	0.37	0.92	26.5	4.1
01000	0.13	–2.5	0.30	0.95	26.0	5.3
01100	0.05	–3.0	0.45	0.96	18.7	5.4
01111	0.03	–31.6	689.59	0.97	15.2	5.3
11111	0.29	–0.27	0.23	0.97	44.7	4.3

[a] Five different (sub)scenarios in which the haplotype reported in the table has an OR of 1.5 and all other haplotypes show no effect. The mean bias, mean SE, coverage probability and power are reported for the haplotype with effect and the Type I error rate is reported for the other four haplotypes in a subscenario.

where $l(\beta)$ is the unpenalized log-likelihood, λ is penalty coefficient, and β_r is the parameter estimate of haplotype *r*. The second penalty function was the similarity penalty, which was used previously by Tanck et al. [12]. It is based on the assumption that similar haplotypes show similar effects

$$l_{pen2} \propto l(\beta) - \frac{1}{2}\lambda \sum_{h=1}^{m} \sum_{r=h+1}^{m} a_{hr}\left(\beta_h - \beta_r\right)^2, \tag{8}$$

where a_{hr} is the similarity between haplotypes *h* and *r (h, r* = 1, ... , *m)*, expressed as the number of alleles that these haplotypes share (*a* = 0, ... , number of polymorphisms-1), and $\beta_h - \beta_r$ is the difference in estimated effects of haplotypes *h* and *r*. Furthermore, we considered whether weighting the haplotype frequencies in the penalty would improve the results, since the penalty function is only included as a way to estimate effects of the rare haplotypes. Therefore, penalizing the rare haplotypes more than the common haplotypes might yield better results. The ridge penalty now becomes

$$l_{pen3} \propto l(\beta) - \frac{1}{2}\lambda \sum_{r} \frac{\beta_r^2}{p_r}, \tag{9}$$

where p_r is the frequency of haplotype *r*. This penalty function will be referred to as the ridge-frequency penalty. The similarity penalty is

$$l_{pen4} \propto l(\beta) - \frac{1}{2}\lambda \sum_{h} \sum_{r} \frac{a_{hr}}{p_h p_r}\left(\beta_h - \beta_r\right)^2, \tag{10}$$

where p_h and p_r are the frequencies of haplotype *h* and haplotype *r* respectively. This penalty will be called the similarity-frequency penalty in the remainder of this article.

Generalized cross-validation (GCV) [17] was used to determine the magnitude of λ. This was done by minimizing the mean-squared error (MSE) with respect to λ. For various values of λ the MSE was calculated as follows [18]:

$$\mathrm{MSE}_{\mathrm{GCV}} = n^{-1} \frac{\sum_{i=1}^{N} \sum_{j=1}^{k_i} \left(y_i - \pi_{ij}\right)^2}{\left(1 - n^{-1} \sum_{i=1}^{N} \sum_{j=1}^{k_i} h_{ij}\right)^2} \tag{11}$$

where $h_{ij} = v_{ij} X_{ij} (\Omega(\hat{\beta}^{\lambda}) + \lambda I)^{-1} X'_{ij}$, $v_{ij} = \pi_{ij}(1 - \pi_{ij})$, β^{λ} is the maximizer of the penalized log-likelihood as shown in equations (7)–(10), and $\Omega(\hat{\beta}^{\lambda})$ is the negative of the matrix of second derivatives.

Simulation Settings

To investigate properties of our method, we performed a simulation study. A total of six haplotypes consisting of five SNPs were chosen to be present in the population with haplotype frequencies similar to those we previously found in the CETP gene [12] (see table 1). Haplotypes were randomly assigned to 500 individuals and will be presented as a combination of zeros and ones with 1 representing the least common allele. Disease status was sampled from the binomial distribution with probability depending on the haplotypes using a logistic model assuming haplotypes to have an additive effect on the log-odds scale. Baseline disease prevalence was set to 10%. We evaluated performance of our method in different scenarios with one or more (comparable or different) haplotypes associated with disease. Three different scenarios were investigated. The first scenario considered one haplotype with an odds ratio (OR) of 1.5 and all other haplotypes showed no effect. The five different subscenarios involved the effect being placed on the five different haplotypes. For the second scenario two similar haplotypes, namely 01000 and 01100, were both given an OR of 1.5. Theoretically, this scenario favours the similarity and similarity-frequency penalty. In the third scenario, ORs of 1.5 were placed on dissimilar haplotypes, namely 00111 and 01100, which does not favour the similarity and similarity-frequency penalties. For each scenario, 500 replicates were carried out. The 00000 haplotype (frequency 0.42) was considered as the reference category in all analyses. The statistical properties were evaluated using three different measures, namely the mean bias of the parameter estimates, the root of the mean squared error and the coverage probability, which is defined as the probability that the 95% confidence interval of the parameter estimate contains the true theoretical value of the parameter estimate. Furthermore, for each haplotype the percentage of replicates which identified the haplotype as being significantly associated with the outcome (i.e., power or Type I error rate) was calculated. The significance level used to calculate the power and the Type I error rate was set to $\alpha = 0.05$.

Table 2. Mean bias, mean SE and power for three different scenarios comparing results of the different penalties with the unpenalized results

	Haplotype	Unpenalized	Ridge	Ridge-frequency	Similarity	Similarity-frequency
1[a]						
Mean bias	01111	–0.32	–0.36	–0.35	–0.35	–0.44
Mean SE	01111	689.59	0.35	0.13	0.48	0.33
Power	01111	15.2	3.4	3.4	6.9	9.1
Type I error rate		5.3	1.0	1.5	8.6	6.6
2[b]						
Mean bias	01000	–0.01	–0.23	–0.23	–0.22	–0.19
	01100	–0.05	–0.28	–0.30	–0.26	–0.30
Mean SE	01000	0.30	0.17	0.17	0.28	0.27
	01100	0.45	0.31	0.18	0.40	0.32
Power	01000	30.3	13.4	14.2	12.6	13.4
	01100	18.4	7.2	7.6	9.3	8.2
Coverage probability	01000	0.95	0.48	0.45	0.88	0.90
	01100	0.96	0.43	0.28	0.91	0.78
Type I error rate		5.8	1.5	2.5	7.4	4.3
3[c]						
Mean bias	00111	–0.07	–0.27	–0.29	–0.25	–0.26
	01100	–0.04	–0.25	–0.27	–0.24	–0.25
Mean SE	00111	0.38	0.19	0.18	0.35	0.31
	01100	0.45	0.22	0.18	0.40	0.32
Power	00111	19.8	8.7	7.6	9.7	9.3
	01100	21.3	8.3	7.9	13.0	10.3
Coverage probability	00111	0.95	0.42	0.33	0.90	0.82
	01100	0.95	0.40	0.28	0.89	0.79
Type I error rate		6.3	1.9	2.3	9.5	5.4

[a] Haplotype 01111 associated with OR = 1.5. All other haplotypes showed no effect in this scenario.

[b] In this scenario haplotypes 01000 and 01100 were associated with ORs of 1.5. These haplotypes are very similar. All other haplotypes were not associated with risk.

[c] In this scenario haplotypes 00111 and 01100 were associated with ORs of 1.5. These haplotypes are dissimilar. All other haplotypes showed no effect.

Results

Unpenalized Log-Likelihood Approach

Table 1 shows results for the weighted unpenalized logistic regression method for the five subscenarios in which an (small) adverse effect (OR = 1.5) was simulated for the five haplotypes. Therefore, in all subscenarios the haplotype in the table had an OR of 1.5 while the other four haplotypes showed no effect. Mean bias ranged from –0.32 to –0.0027, corresponding to mean ORs of 1.65 for haplotype 01111 and 1.53 for haplotype 11111. The power of detecting the haplotype with effect ranged from 15.2% for the 01111 haplotype to 44.7% for the 11111 haplotype. The Type I error rate was calculated for the other four haplotypes in each subscenario, and differed between 4.1 and 5.4%. In all subscenarios the coverage probability was approximately 95%, for haplotypes with effect as well as for haplotypes without effect. As can be seen in the fourth column of table 1, the mean standard error was quite large for the infrequent 01111 haplotype (i.e., approximately 690). The mean standard errors for the other haplotypes varied between 0.23 and 0.45.

The Penalty Functions

Table 2 shows results for the comparison of the different penalties for three different scenarios. In the first scenario the least frequent haplotype (01111) was associated with an effect of 1.5. As can be seen from the results in the table, the mean SE is reduced from 689.59 for the unpenalized analysis to a minimum of 0.13 for the ridge-frequency penalty and at most 0.48 for the similarity penalty. Bias varied between –0.32 and –0.44, showing a slight increase for the penalized analyses compared to the unpenalized analysis. However, power was decreased from 15.2% for the unpenalized model to 3.4% for the ridge and the ridge-frequency penalties, to 6.9% for the similarity penalty and 9.1% for the similarity-frequency penalty. The Type I error rate, which is approximately 5% in the unpenalized model, is decreased to 1.0 and 1.5% for the ridge and ridge-frequency penalties, respectively, and slightly increased to 8.6 and 6.6% for the similarity and similarity-frequency penalties, respectively.

In the second scenario haplotypes 01000 and 01100 were associated with ORs of 1.5. These haplotypes are very similar since only one allele is different. Therefore, this scenario should favour the similarity and similarity-frequency penalties. However, mean bias (around 0.22 and 0.28) and power (around 14 and 8%) are similar for all penalties. Mean SE is somewhat higher for the similarity and similarity-frequency penalties, resulting in markedly higher coverage probabilities (see table 2). The Type I error rate was 5.8% for the unpenalized model, 1.5% for the ridge penalty, 2.5% for the ridge-frequency penalty, 7.4% for the similarity penalty and 4.3% for the similarity-frequency penalty. The false discovery rate (FDR) of the penalties were 0.18 for the ridge penalty, 0.25 for the ridge-frequency penalty, 0.50 for the similarity penalty, and 0.37 for the similarity-frequency penalty, compared to 0.26 for the unpenalized method.

The third scenario compares results when a similar effect of 1.5 is associated with the dissimilar haplotypes 00111 and 01100, which is a scenario that does not favour the similarity and similarity-frequency penalties. Mean bias, mean SE and power are similar to the previous scenario. For both the 00111 and the 01100 haplotype, the coverage probabilities of the ridge and ridge-frequency penalty are lower than the coverage probabilities of the similarity and similarity-frequency penalty. The Type I error rate is decreased for the ridge, the ridge-frequency, and the similarity-frequency penalties (i.e., 1.9, 2.3, and 5.4% respectively) and increased for the similarity penalty (9.5%) compared to the unpenalized model (6.3%). The FDR were 0.32, 0.25, 0.31, 0.56, and 0.45 for the unpenalized method, the ridge penalty, the ridge-frequency penalty, the similarity penalty, and the similarity-frequency penalty, respectively.

Discussion

The present study shows a generalisation of the weighted log-likelihood method to estimate haplotype effects of Tanck et al. [12] to dichotomous outcome data. Some statistical properties of this model have been investigated with a simulation study. Furthermore, to deal with the problem that estimates of rare haplotypes show large variation, which can lead to model instability, statistical properties of four different penalty functions were investigated in a simulation study.

The coverage probabilities of the weighted (unpenalized) log-likelihood approach were good for all investigated scenarios. The mean bias of the parameter estimates was usually small, although it increased when the haplotype frequency decreased. The power decreased with decreasing haplotype frequencies, as was expected. The frequency of the haplotype was inversely related to the standard error of the parameter estimate, with rare haplotypes showing extremely large standard errors. This last issue is not specific for our method but is a general statistical property. The polymorphisms of the *CETP* haplotypes on which the haplotype frequencies in our simulation study were based, show high linkage disequilibrium (LD) [19]. Polymorphisms in other genes or genomic regions might show less LD. Therefore, we tested the performance of our method in the extreme scenario of no LD. Although a haplotype analysis would not be the method of first choice in this case, it turns out that the mean bias in this simulation was small (data not shown).

The main aim of introducing a penalty in the log-likelihood was to get a more accurate estimate of the effects of the rare haplotypes, and indeed, all penalty functions reduced the standard errors of the rare haplotypes markedly, but they also introduced bias. Unfortunately, this trade-off decreased power in all investigated scenarios. Therefore, using a penalty function might only be useful in a pilot study where the unpenalized approach cannot estimate the effect of rare haplotypes. In this situation, the penalized approach decreases the variance of the parameter estimates, thereby giving some indication of whether rare haplotype show association with disease, after which further research might be conducted. Overall, the similarity and similarity-frequency penalties showed higher power than the ridge and ridge-frequency penalties, even in scenarios that did not favour to the underlying assumption that similar haplotypes show similar effects. However, the similarity and similarity-frequency penalty also show a higher FDR than the ridge and ridge-frequency penalty. Consequently, using these penalties means that more false discoveries will be made. Therefore, considering both power and FDR, there is not one penalty function that performs markedly better than the others. However, the Type I error rate of the similarity penalty is somewhat higher than 5%, indicating that this penalty function might be anti-conservative. Other approaches to deal with rare haplotypes, like pooling them into one category, or pooling them with common haplotypes that are very similar, lead to pooled categories that are hard to interpret. These methods seem to increase power [20], but only in specific situations where pooled haplotypes have similar effects.

Besides the penalty functions investigated in the present study, many other penalty functions, for example the Lasso [21] or Garotte [22] penalty functions, are possible. These penalty functions shrink the regression coefficients to zero in a manner similar to, for example, the ridge penalty function. A completely different penalty function has been suggested by Warm [23] within the framework of the item response theory. He used a function of the variance of the regression parameter estimate to weight the log-likelihood, through which shrinkage of the coefficients as well as shrinkage of the variance was accomplished. So far, this method has not been used outside of the item response theory.

The weighted log-likelihood approach described in this paper is a flexible method allowing for adjustment for (environmental) covariates as well as haplotype-environment interactions. Furthermore, although not incorporated in the present software yet, our method can be easily extended to deal with missing genotype data. Missing genotype data would simply increase the number of possible haplotype pairs for a particular subject.

The logistic regression technique is valid for cohort as well as case-control sampling. However, most methods to infer haplotype frequencies from population-based data assume HWE. This assumption is also made in our method when calculating weights from haplotype frequencies (w_{ij}). This HWE assumption could be violated in the cases of a case-control study when an allele is associated with disease. It is possible to calculate haplotype frequencies only in the sample of controls, although this might be problematic when a certain haplotype is only present in the case sample. Among others, Fallin and Schork [24] and Niu et al. [25] have investigated the impact of deviations from HWE on the performance of the EM algorithm. They found that deviations in HWE did not dramatically increase the error, especially when deviation resulted in excess homozygosity. Moreover, our method re-estimates the weights based on the parameter estimates, which might lead to better estimates of the weights and haplotype frequencies when deviations from HWE are present.

The standard errors for the penalized analysis that are presented in the present paper do not account for the uncertainty related to λ since this parameter was not incorporated in the information matrix. In practice, the correct standard errors are obtained by bootstrapping.

The unpenalized weighted log-likelihood approach is a good method for estimating multilocus haplotype effects on dichotomous outcome. The penalty function can help estimate an effect for rare haplotypes with large standard errors in the unpenalized model. Although this estimate is biased, it is a more efficient estimate than the unpenalized estimate, which may help to indicate whether further studying this haplotype is useful.

Acknowledgement

M.W.T. Tanck was financially supported by the Netherlands Heart Foundation grant no 2000:125.

References

1 Niu T: Algorithms for inferring haplotypes. Genet Epidemiol 2004;27:334–347.

2 Schaid DJ: Evaluating associations of haplotypes with traits. Genet Epidemiol 2004;27:348–364.

3 Templeton AR, Boerwinkle E, Sing CF: A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. Genetics 1987;117:343–351.

4 Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP: Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 2004;75:35–43.

5 Seltman H, Roeder K, Devlin B: Evolutionary-based association analysis using haplotype data. Genet Epidemiol 2003;25:48–58.

6 Yu K, Gu CC, Province M, Xiong CJ, Rao DC: Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. Genet Epidemiol 2004;27:182–191.

7 Chiano MN, Clayton DG: Fine genetic mapping using haplotype analysis and the missing data problem. Ann Hum Genet 1998;62:55–60.

8 Epstein MP, Satten GA: Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 2003;73:1316–1329.

9 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.

10 Sham PC, Rijsdijk FV, Knight J, Makoff A, North B, Curtis D: Haplotype association analysis of discrete and continuous traits using mixture of regression models. Behav Genet 2004;34:207–214.

11 Stram DO, Leigh PC, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC: 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.

12 Tanck MW, Klerkx AH, Jukema JW, Knijff PD, Kastelein JJ, Zwinderman AH: Estimation of multilocus haplotype effects using weighted penalised log-likelihood: analysis of five sequence variations at the cholesteryl ester transfer protein gene locus. Ann Hum Genet 2003;67:175–184.

13 Tregouet DA, Escolano S, Tiret L, Mallet A, Golmard JL: A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm. Ann Hum Genet 2004;68:165–177.

14 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.

15 Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12:921–927.

16 Dempster AP, Laird NM, Rubin DB: Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc B 1977;39:1–38.

17 Golub G, Heath M,Wahba G: Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 1979;21:215–223.

18 le Cessie S, van Houwelingen JC: Ridge estimators in logistic regression. Applied Statistics 1992;41:191–201.

19 Klerkx AHEM, Tanck MWT, Kastelein JJP, Molhuizen HOF, Jukema JW, Zwinderman AH, Kuivenhoven JA: Haplotype analysis of the CETP gene: not TaqIB, but the closely linked –629c→a polymorphism and a novel promotor variant are independently associated with CETP concentration. Hum Mol Genet 2003;12:111–123.

20 Jannot AS, Essioux L, Clerget-Darpoux F: Association in multifactorial traits: how to deal with rare observations? Hum Hered 2004;58:73–81.

21 Tibshirani R: Regression shrinkage and selection via the Lasso. JRSS-B 1996;58:267–288.

22 Breiman L: Better subset selection using the nonnegative Garotte. Technometrics 1995;37:373–383.

23 Warm TA: Weighted likelihood estimation of ability in item response theory. Psychometrika 1989;54:427–450.

24 Fallin D, Schork NJ: Accuracy of haplotype frequency estimation for biallelic loci, via the Expectation-Maximization algorithm for unphased diploid genotype data. Am J Hum Genet 2000;67:947–959.

25 Niu T, Qin ZS, Xu X, Liu JS: Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 2002;70:157–169.