*Genet. Res., Camb.* (2005), **86**, *pp.* 223–231. *With* 1 *figure.*  
doi:10.1017/S0016672305007792 Printed in the United Kingdom

# Haplotype association analysis of human disease traits using genotype data of unrelated individuals

QIHUA TAN[1,2]*, LENE CHRISTIANSEN[2], KAARE CHRISTENSEN[2], LISE BATHUM[1,2], SHUXIA LI[2], JING HUA ZHAO[3] AND TORBEN A. KRUSE[1]
[1] *Department of Clinical Biochemistry and Genetics (KKA), Odense University Hospital, Sdr. Boulevard 29, 5000 Odense C, Denmark*
[2] *Institute of Public Health, University of Southern Denmark, Odense, Denmark*
[3] *Department of Epidemiology and Public Health, University College London, London, UK*

(*Received 7 March 2005 and in revised form 27 June and 16 August 2005*)

**Summary**

Haplotype inference has become an important part of human genetic data analysis due to its functional and statistical advantages over the single-locus approach in linkage disequilibrium mapping. Different statistical methods have been proposed for detecting haplotype – disease associations using unphased multi-locus genotype data, ranging from the early approach by the simple gene-counting method to the recent work using the generalized linear model. However, these methods are either confined to case – control design or unable to yield unbiased point and interval estimates of haplotype effects. Based on the popular logistic regression model, we present a new approach for haplotype association analysis of human disease traits. Using haplotype-based parameterization, our model infers the effects of specific haplotypes (point estimation) and constructs confidence interval for the risks of haplotypes (interval estimation). Based on the estimated parameters, the model calculates haplotype frequency conditional on the trait value for both discrete and continuous traits. Moreover, our model provides an overall significance level for the association between the disease trait and a group or all of the haplotypes. Featured by the direct maximization in haplotype estimation, our method also facilitates a computer simulation approach for correcting the significance level of individual haplotype to adjust for multiple testing. We show, by applying the model to an empirical data set, that our method based on the well-known logistic regression model is a useful tool for haplotype association analysis of human disease traits.

## 1. Introduction

Combating complex diseases is one of the significant challenges for twenty-first-century medicine. New advances in human genetics, especially the sequencing of the human genome and the recent development of high-throughput single nucleotide polymorphism (SNP) genotyping technology, facilitate the deciphering of the genetic nature in complex diseases. Haplotypes, the combination of alleles at closely linked multiple loci on the same chromosome, may play a key role in the study of complex diseases due to their functional and statistical advantages over the single-locus approach in linkage disequilibrium mapping (Akey *et al.*, 2001; Schaid, 2004; Clark, 2004). Although improving genotyping efficiency enables large-scale population-based studies, there is a critical need to develop efficient analytical methods for haplotype inference (risk and frequency) using unphased genotype data.

Haplotype estimation attracts attention from many researchers in statistical genetics. In the literature, different statistical approaches have been proposed for haplotype analysis. By assuming Hardy–Weinberg equilibrium (HWE), the expectation-maximization (EM) algorithm (Excoffier & Slatkin, 1995; Zhao & Sham, 2002) can be applied to estimate haplotype frequencies using unphased multi-locus genotype data. The likelihood ratio test is used to compare haplotype frequencies between affected cases and normal controls when HWE holds in both groups. Epstein & Satten (2003) proposed a retrospective

* Corresponding author. Tel: +45 65412822. Fax: +45 65411911. e-mail: qihua.tan@ouh.fyns-amt.dk

likelihood method for haplotype inferences. Under HWE in the controls, the method maximizes the likelihood of the data to estimate and test the parameters for all individual haplotypes. The likelihood-based procedure also produces interval estimates on the risks of haplotypes. However, the application of the method is restricted to case-control data. Using the generalized linear model (GLM), Schaid *et al.* (2002) proposed a score test for haplotype inference. The model can be generalized to a variety of different disease traits and performs efficient tests on individual haplotypes. It is necessary to point out that, in clinical studies, the sampling is usually conditional on the disease traits (such as the case – control studies). Such a situation can lead to an over-representation of the cases or the extremes of the traits compared with the general population. In this case, the GLM approach can produce biased estimates on the haplotype parameters because it does not account for how the sample was ascertained (Schaid, 2004).

This paper is aimed at introducing a new approach for haplotype inference using the popular logistic regression model (Hosmer & Lemeshow, 2000). Instead of prospectively modelling disease phenotype (dependent variable) as a function of haplotypes (independent variable), we retrospectively model haplotype frequency conditional on the disease phenotype which, as we show, can be accomplished by the logistic regression model. Modelling haplotype frequency conditional on trait is advantageous in that: (a) the logit of haplotype frequency is regressed on the disease trait, which can be binary, categorical or continuous; (b) it offers unbiased estimates. Our logistic regression model allows haplotype-based parameterization and maximization of the likelihood function built upon the multinomial distribution of the observed multi-locus genotypes. Point and interval estimates of haplotype relative risks can be obtained by exponentiating the corresponding point and interval estimates of the slope coefficients in the polytomous logistic regression model. In addition to haplotype relative risks, our model also provides haplotype frequency estimates for given disease status (case and control), or given trait values (continuous) which can be plotted against the continuous trait. Different statistics can be applied to assess the asymptotic significance levels for single or grouped haplotypes. Moreover, we propose a simulation-based approach for correcting the significance levels of individual haplotypes to adjust for multiple testing. As an example, we apply our model to an empirical data set to associate the haplotypes in the promoter region of the interleukin-6 gene with a cognitive impairment trait derived from the Mini Mental State Examination in a cohort of elderly subjects. Data analysing strategies are also illustrated in the example application.

## 2. Materials and methods

### (i) *The haplotype logistic regression (HLR) model*

We start by introducing the polytomous logistic regression model with haplotype-based parameterization by assuming that each of the haplotypes is observed explicitly as phase known. We denote the collection of all the haplotypes arising from the typed loci with $H$. Following the typical formulation of a logistic regression model, we define the logit function for the frequency of a haplotype pair $(h_i, h_j)$ as

$$g_{i,j}(x) = \ln[\pi_{i,j}(x)/\pi_{0,0}(x)] = \alpha_{i,j} + \beta_{i,j}x \quad (1)$$

where $\pi_{i,j}(x)$ is the frequency of haplotype pair $(h_i, h_j)$ and $\pi_{0,0}(x)$ is the frequency of the baseline haplotype pair at the given trait value $x$, $\alpha_{i,j}$ is the intercept coefficient as in an ordinary regression model, and $\beta_{i,j}$ is the slope coefficient that measures the association between haplotype pair $(h_i, h_j)$ and the trait which is our primary interest. Equation (1) is a typical multinomial logistic regression model with polytomous responses which are the haplotype pairs that occur over the multiple loci. For the baseline haplotype pair, we set $\alpha_{0,0}$ and $\beta_{0,0}$ to zero to make the model identifiable. The parameterization of (1) is based on haplotype pairs. To introduce the haplotype-based parameterization, let $\alpha_{i,j} = \alpha_i + \alpha_j$ and $\beta_{i,j} = \beta_i + \beta_j$. Substituting them into (1), we have

$$\begin{aligned}\pi_{i,j}(x) &= \exp[(\alpha_i+\alpha_j)+(\beta_i+\beta_j)x] \Big/ \sum_{i',j'\in H} \exp[g_{i',j'}(x)] \\ &= \left\{ \exp(\alpha_i+\beta_i x) \Big/ \sqrt{\sum_{i',j'\in H} \exp[g_{i',j'}(x)]} \right\} \\ &\quad \times \left\{ \exp(\alpha_j+\beta_j x) \Big/ \sqrt{\sum_{i',j'\in H} \exp[g_{i',j'}(x)]} \right\}. \quad (2)\end{aligned}$$

Note that, since frequencies of all possible haplotype pairs sum up to 1, we have $\sum_{i',j'\in H}\pi_{i',j'}(x) = \pi_{0,0}(x)\sum_{i',j'\in H}\exp(\alpha_{i',j'}+\beta_{i',j'}x) = \pi_{0,0}(x)\sum_{i',j'\in H}\exp[g_{i',j'}(x)] = 1$. Rearranging, we have $\pi_{0,0}(x) = 1/\sum_{i',j'\in H}\exp[g_{i',j'}(x)]$. When $i=j$ and HWE holds,

$$\begin{aligned}\pi_{i,i}(x) &= \pi_i(x)^2 \\ &= \left\{ \exp(\alpha_i+\beta_i x) \Big/ \sqrt{\sum_{i',j'\in H} \exp[g_{i',j'}(x)]} \right\}^2. \quad (3)\end{aligned}$$

From (3), it is easy to see that under HWE (2) means that the frequency of any haplotype pair can be calculated by multiplying the frequencies of the two corresponding haplotypes. Similar to the situation of heterozygous genotypes, because we can not distinguish haplotype pair $(h_i, h_j)$ from $(h_j, h_i)$, we have,

https://doi.org/10.1017/S0016672305007792 Published online by Cambridge University Press

for any $i, j \in H$,

$$\pi^*_{i,j}(x) = \begin{cases} \exp[(\alpha_i+\alpha_j)+(\beta_i+\beta_j)x] / \sum_{i',j' \in H} \exp[g_{i',j'}(x)] \\ i=j \\ 2\exp[(\alpha_i+\alpha_j)+(\beta_i+\beta_j)x] / \sum_{i',j' \in H} \exp[g_{i',j'}(x)] \\ i<j. \end{cases} \quad (4)$$

Equation (4) will be used in constructing the likelihood function later. For the baseline haplotype pair, since $\alpha_{0,0}=2\alpha_0=0$ and $\beta_{0,0}=2\beta_0=0$, we have $\alpha_0=0$ and $\beta_0=0$. Under HWE, once the intercept and the slope coefficients have been estimated, we can calculate the frequency for any haplotype (including the baseline) as

$$\pi_i(x) = \exp(\alpha_i+\beta_i x) / \sqrt{\sum_{i',j' \in H} \exp[g_{i',j'}(x)]},$$

$$\alpha_0 = 0,\ \beta_0 = 0. \quad (5)$$

(ii) *Risk of haplotype*

In our model, the relative risk (RR) of haplotype $h_i$ can be calculated as $RR_i(x)=\pi_i(x)/\pi_0(x)=\exp(\alpha_i+\beta_i x)$ at the given trait value of $x$. Based on RR, we define the relative risk ratio (RRR) for comparing the relative risks of haplotype $h_i$ at two given trait values $x_1$ and $x_2$ ($x_2-x_1=k$) as

$$\begin{aligned} RRR_i &= RR_i(x_2)/RR_i(x_1) \\ &= \exp(\alpha_i+\beta_i x_2)/\exp(\alpha_i+\beta_i x_1) \\ &= \exp[\beta_i(x_2-x_1)] = \exp(k\beta_i). \end{aligned} \quad (6)$$

When $k=1$, such as in a case – control study, we have $RRR_i=\exp(\beta_i)$. Similarly, we can estimate RRR for any haplotype pair $(h_i, h_j)$ as

$$\begin{aligned} RRR_{i,j} &= \frac{\pi_{i,j}(x_2)}{\pi_{0,0}(x_2)} \Big/ \frac{\pi_{i,j}(x_1)}{\pi_{0,0}(x_1)} \\ &= \exp(\alpha_{i,j}+\beta_{i,j}x_2)/\exp(\alpha_{i,j}+\beta_{i,j}x_1) \\ &= \exp[(x_2-x_1)\beta_{i,j}] = \exp(k\beta_{i,j}). \end{aligned} \quad (7)$$

In a multiplicative model,

$$\begin{aligned} &\beta_{i,j}=\beta_i+\beta_j \text{ and } RRR_{i,j} = \exp[k(\beta_i+\beta_j)] \\ &\quad = \exp(k\beta_i)\ \exp(k\beta_j) = RRR_i\ RRR_j. \end{aligned}$$

Based on the estimated variance of the slope coefficients (see below), we obtain the confidence interval (CI) of RRR for haplotype $h_i$ by exponentiating $k$ times the endpoints of the CI of $\beta_i$, i.e. $\exp[k\beta_i \pm z_{1-\alpha/2}kSE(\beta_i)]$ where $z_{1-\alpha/2}$ is the upper $100(1-\alpha/2)\%$ point from the standard normal distribution with a type 1 error rate of $\alpha$ and $SE(\beta_i)$ the standard error of $\beta_i$. Again using the variance information for the slopes, we can estimate the CI of RRR for any haplotype pair. Since

$$\begin{aligned} Var(\ln RRR_{i,j}) &= Var(\ln RRR_i + \ln RRR_j) \\ &= Var(k\beta_i + k\beta_j) = k^2[Var(\beta_i) \\ &\quad + Var(\beta_j) + 2Cov(\beta_i, \beta_j)], \end{aligned}$$

we construct CI of RRR for haplotype pair $(h_i, h_j)$ as

$$\begin{aligned} &\exp[\ln RRR_{i,j} \pm z_{1-\alpha/2}\sqrt{Var(\ln RRR_{i,j})}] \\ &\quad = \exp[k(\beta_i+\beta_j) \\ &\quad \pm z_{1-\alpha/2}k\sqrt{Var(\beta_i)+Var(\beta_j)+2Cov(\beta_i,\beta_j)}]. \end{aligned}$$

(iii) *Non-multiplicative effects*

Under HWE, (4) is a multiplicative effects model for all the haplotypes with $\beta_{i,j}=\beta_i+\beta_j$ for $(h_i, h_j)$ and $\beta_{i,i}=2\beta_i$ for $(h_i, h_i)$. The parameterization of the slope coefficients in our model also allows the fitting of non-multiplicative effects models (dominant, recessive). If haplotype $h_i$ is dominant over the other haplotypes, we have $\beta_{i,j}=\beta_{i,i}=\beta_i$. In this case, (7) becomes $RRR_{i,j}=\exp(k\beta_i)=RRR_i$ which means that the disease risk is only imposed by haplotype $h_i$. Likewise, when it is recessive, we simply have $\beta_{i,j}=0$ and $\beta_{i,i}=\beta_i$ so that only homozygous carriers of the haplotype are at risk of the disease. It is important to note, however, that when the disease trait value $x=0$, no matter how the slopes are parameterized, (4) becomes

$$\pi^*_{i,j}(0) = \begin{cases} \exp(\alpha_i+\alpha_j) / \sum_{i',j' \in H} \exp[g_{i',j'}(0)] & i=j \\ 2\exp(\alpha_i+\alpha_j) / \sum_{i',j' \in H} \exp[g_{i',j'}(0)] & i<j. \end{cases} \quad (8)$$

Equation (8) means that, in any model, HWE is a prerequisite in the population whose trait value is zero. Since it is sensible to assume HWE in a normal population, we suggest coding or transforming the trait value such that it is zero in those free from the disease. For example, in a case – control study, we can code the controls with 0 and the cases with 1.

(iv) *Sex-specific effects*

By specifying sex-specific slope parameters, our model can measure sex-dependent haplotype effects to infer haplotype – sex interactions. When the effect

https://doi.org/10.1017/S0016672305007792 Published online by Cambridge University Press

of haplotype $h_i$ is sex-dependent, we have, instead of (1),

$$g_{i,j}(x) = \alpha_{i,j} + [{}_m\beta_i U + {}_f\beta_i(1-U) + \beta_j]x \quad (9)$$

where ${}_m\beta_i$ and ${}_f\beta_i$ stand for the effects of haplotype $h_i$ in males and in females; $U$ is an indicator of sex (0 for females and 1 for males). The sex-dependent $RRR_i$ can be calculated by introducing formula (9) to (6). Given the haplotype and sex, for example male ($U=1$), we have ${}_mRRR_i = \exp(\alpha_{i,j} + {}_m\beta_i x_2)/\exp(\alpha_{i,j} + {}_m\beta_i x_1) = \exp(k\,{}_m\beta_i)$. In a case – control study ($k=1$), the risk of the haplotype for males is $\exp({}_m\beta_i)$ as compared with the baseline haplotype. In the same manner, we have ${}_fRRR_i = \exp(k\,{}_f\beta_i)$. Using the covariance information available from the maximum likelihood procedure (see next section), we can construct the Wald test for comparing ${}_m\beta_i$ and ${}_f\beta_i$ as $W = ({}_m\beta_i - {}_f\beta_i)/\sqrt{Var({}_m\beta_i - {}_f\beta_i)}$ with the null hypothesis of ${}_m\beta_i = {}_f\beta_i$. Here $Var({}_m\beta_i - {}_f\beta_i) = Var({}_m\beta_i) + Var({}_f\beta_i) - 2Cov({}_m\beta_i, {}_f\beta_i)$. When ${}_m\beta_i$ and ${}_f\beta_i$ are not significantly different, we set ${}_m\beta_i = {}_f\beta_i = \beta_i$ so that (9) reduces to (1). Note that, based on the law of segregation, we assign the same intercept parameter for both sexes to reduce the number of parameters in the model.

### (v) *Linking haplotype with genotype and the likelihood function*

Up to now, our model has been formulated on unambiguous haplotypes. However, in practice the phase information is missing in the genotype data of unrelated subjects. Since what we observe are multi-locus genotypes, it is necessary to set up a link between haplotype and multi-locus genotype in order to estimate the haplotype parameters using the observed genotype data. For a multi-locus genotype $g$, we can find the collection of haplotype pairs that are consistent with $g$ and which we denote as $S(g)$. Based on this, we can calculate the frequency of the multi-locus genotype as the summation over the frequencies of all the haplotype pairs in $S(g)$, i.e.

$$\pi_g(x) = \sum_{i,j \in s(g)} \pi^*_{i,j}(x) \quad (10)$$

where $\pi^*_{i,j}(x)$ is the frequency of haplotype pair $(h_i, h_j)$ in $S(g)$ as expressed in (4). Expression (10) links the observed multi-locus genotypes with the ambiguous haplotypes which we do not observe. With this relationship, we can construct the likelihood function to estimate the haplotype parameters in the model. To do this, we first create an indicator variable for grouping the membership of an individual. Let $y_{g,s}$ take the value of 1 if the multi-locus genotype of subject $s$ is $g$ and 0 otherwise. Then the likelihood function for the whole data set consisting of $n$ subjects is

$$L(\vec{\alpha}, \vec{\beta}) = \prod_s \prod_{g \in G} \pi_g(x_s)^{y_{g,s}}. \qquad \sum_{g \in G} y_{g,s} = 1,$$
$$s = 1, 2, 3, \ldots\ldots, n. \quad (11)$$

Here $\vec{\alpha}$ and $\vec{\beta}$ are vectors of the intercept and slope coefficients to be estimated, $G$ is a collection of all the multi-locus genotypes observed in the data. The covariance matrix obtained by inverting the observed information matrix for (11) can be used to calculate the univariate Wald statistic (asymptotically standard normal) for significance inferences on specific slope parameters (Jennings, 1986).

### (vi) *Multiple haplotypes and multiple testing*

Haplotypes as super-alleles are highly polymorphic. Such a situation weakens the statistical power of the current haplotype inference methods (Schaid, 2004). As the power for detecting associations with rare haplotypes is very low (Comeron *et al.*, 2003), we suggest grouping the rare haplotypes to form one combined haplotype (Lake *et al.*, 2003). The synthetic haplotype can serve as a baseline haplotype in our logistic regression model since the heterogeneous group is biologically meaningless. Grouping of rare haplotypes helps us to focus on more informative haplotypes. At the same time, the strategy also reduces the number of multiple test and thus false positive results.

Since in our polytomous logistic regression model, there are two parameters (one intercept and one slope) for each haplotype, the total number of parameters in the model is twice the number of non-baseline haplotypes. In order to increase the statistical power, one can first estimate the slope parameter for each haplotype separately by assuming no effect from the other haplotypes with their slopes set to zero. The estimation can be done by assuming that the haplotype effect is multiplicative, dominant or recessive. The Akaike information criterion (AIC) (Akaike, 1973) can be calculated and recorded for model selection. Haplotypes with low AICs can be picked up for fitting an extended model in which the slope parameters for the rest of haplotypes are set to zero. To find the best-performing model, different combinations of the number of the selected haplotypes can be included in the extended model and their recorded AICs compared. Final parameter estimates are obtained from the best-performing model that has the lowest AIC amongst all the combinations. When assuming that the effects of all haplotypes are multiplicative, a full haplotype model can be fitted. In this case, we suggest using the convenient log-likelihood ratio test to obtain an overall significance level for a group or all the haplotypes (equivalent to

https://doi.org/10.1017/S0016672305007792 Published online by Cambridge University Press

the testing of goodness-of-fit). That is, we calculate minus twice the change in the log-likelihood between the saturated full-haplotype model and the intercept-only model, $G = -2[\ln L(\vec{\alpha}) - \ln L(\vec{\alpha}, \vec{\beta})]$, which asymptotically follows a chi-squared distribution with the length of $\vec{\beta}$ as degrees of freedom.

Note that the $P$ values obtained by the univariate Wald test are haplotype-specific and are prone to false positive results due to multiple comparisons. In order to reduce the type 1 error rate inflated by multiple testing, we introduce a simulation-based empirical approach similar to that used in linkage analysis. To do that, we first shuffle the observed phenotype values to form random samples. For each random sample, we apply our model to estimate the haplotype parameters and record the absolute Wald statistics for the slopes of all the haplotypes. When this is done for a total number of $B$ random samples, we obtain the adjusted $P$ value for an individual haplotype with the observed univariate Wald statistic $W_{obs}$ as

$$p \equiv \sum_{i=1}^{B} I\left\langle \left\{ \sum_{j=1}^{L} I[abs(W_{i,j}) \geqslant abs(W_{obs})] \right\} \geqslant 1 \right\rangle / B \quad (12)$$

where $I(\cdot)$ is an indicator function, $L$ is the number of testing haplotypes and $W_{i,j}$ is the univariate Wald statistic calculated for haplotype $j$ from random sample $i$.

Table 1. *Parameters from the multiplicative and non-multiplicative effects models*

| Haplotype model | Intercept $\alpha$ | Slope $\beta$ | SE | $P$ value | AIC |
|---|---|---|---|---|---|
| *Multiplicative* | | | | | |
| 1-1-2 | 2·038 | 0·016 | 0·010 | 0·118 | 2681·911 |
| 1-2-1 | 1·621 | −0·016 | 0·013 | 0·206 | 2682·726 |
| 1-4-1 | 1·549 | −0·004 | 0·012 | 0·735 | 2684·243 |
| 1-3-1 | 0·509 | −0·026 | 0·022 | 0·244 | 2682·925 |
| 2-3-1 | −0·474 | 0·018 | 0·026 | 0·494 | 2683·904 |
| 1-2-2 | −0·990 | 0·019 | 0·034 | 0·578 | 2684·056 |
| *Dominant* | | | | | |
| 1-1-2 | 2·110 | 0·011 | 0·013 | 0·404 | 2683·654 |
| 1-2-1 | 1·590 | −0·015 | 0·014 | 0·269 | 2683·119 |
| 1-4-1 | 1·495 | 0·003 | 0·013 | 0·802 | 2684·295 |
| 1-3-1 | 0·446 | −0·018 | 0·022 | 0·400 | 2683·629 |
| 2-3-1 | −0·464 | 0·017 | 0·027 | 0·513 | 2683·945 |
| 1-2-2 | −0·990 | 0·019 | 0·034 | 0·578 | 2684·056 |
| *Recessive*[a] | | | | | |
| 1-1-2 | 2·119 | 0·017 | 0·014 | 0·221 | 2682·904 |
| 1-2-1 | 1·512 | −0·011 | 0·024 | 0·652 | 2684·147 |
| 1-4-1 | 1·555 | −0·027 | 0·026 | 0·296 | 2683·157 |
| 1-3-1 | 0·365 | −0·229 | 0·234 | 0·328 | 2682·005 |
| 2-3-1 | −0·332 | 0·019 | 0·109 | 0·861 | 2684·328 |

AIC, Akaike information criterion.
[a] No estimation on the 1-2-2 haplotype due to its very low frequency.

## 3. Results

As an example, we apply our model to a multi-locus genotype data set collected in a study on the *interleukin-6* gene (IL-6) and human ageing in the Danish population (Christiansen *et al*., 2004). In this study, multi-locus genotypes are available from two SNP loci (–572G/C and –174G/C) and one AT-stretch locus (–373AnTm, four alleles) in the promoter region of the IL-6 gene. In the sample, full genotype information is available on 555 participants aged 93 years who also underwent the Mini Mental State Examination (MMSE), a test that assesses mental status. The reported MMSE score varies from 2 to 30, with the lowest score indicating the most severe cognitive impairment. In general, a score under 24 is an indication of dementia. We chose to examine IL-6 and MMSE phenotype because a significant association between the –174C allele and Alzheimer's disease, the most common form of dementia, was reported in a recent study (Licastro *et al*., 2003). Using our example data, we show how our haplotype logistic regression model can be used to estimate haplotype frequencies in the cases (MMSE $\leqslant$24) and controls (MMSE >24) or at given (continuous) trait values, while at the same time measuring the haplotype association with MMSE trait. To ensure that our HWE assumption is for the normal population, we use (30 – MMSE) instead of the MMSE score in fitting the model.

We start by fitting our model for each haplotype while setting the slope parameters of the others to zero. As described before, this is done for different modes of haplotype function (multiplicative, dominant and recessive). Results from analyses of the six most common haplotypes (Christiansen *et al*., 2004) are presented in Table 1. The haplotypes in Table 1 are formed with the sequence of –572G/C (G = 1, C = 2), –373AnTm ($A_8T_{12} = 1$, $A_9T_{11} = 2$, $A_{10}T_{10} = 3$, $A_{10}T_{11} = 4$) and –174G/C (G = 1, C = 2). The rare haplotypes are combined to form the baseline haplotype in the logistic regression model. Wald tests on the estimated slope parameters in Table 1 showed that no haplotype displays a significant association with the MMSE trait. However, it is interesting to see that the lowest AIC is achieved by the multiplicative model of the 1-1-2 haplotype which contains the – 174C allele, a result that supports the finding by Licastro *et al*. (2003). The slope parameter of this haplotype suggests a frequency increase of the haplotype with (30 – MMSE), which means increased haplotype frequency in the poor-performance individuals. Christiansen *et al*. (2004) reported a modest but harmful influence of the same haplotype on human survival, with a lower frequency in aged

https://doi.org/10.1017/S0016672305007792 Published online by Cambridge University Press

Table 2. *Model outputs for the grouped case – control data*[a]

| Haplo-type | Inter-cept $\alpha$ | Slope | | | | Risk | | Frequency | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Case | | Control | |
| | | $\beta$ | SE | *P* value, unadjusted | *P* value, adjusted | RRR | 95% CI | HLR | EM | HLR | EM |
| 1-1-2 | 2·034 | 0·237 | 0·404 | 0·558 | 0·851 | 1·267 | 0·574–2·800 | 0·439 | 0·440 | 0·410 | 0·409 |
| 1-2-1 | 1·451 | 0·087 | 0·438 | 0·843 | 0·993 | 1·090 | 0·463–2·571 | 0·211 | 0·211 | 0·229 | 0·227 |
| 1-4-1 | 1·412 | 0·190 | 0·405 | 0·640 | 0·923 | 1·209 | 0·546–2·675 | 0·225 | 0·224 | 0·220 | 0·223 |
| 1-3-1 | 0·381 | −0·118 | 0·384 | 0·759 | 0·974 | 0·889 | 0·419–1·887 | 0·059 | 0·059 | 0·079 | 0·080 |
| 2-3-1 | −0·430 | 0·191 | 0·594 | 0·748 | 0·971 | 1·210 | 0·378–3·876 | 0·036 | 0·035 | 0·035 | 0·035 |
| 1-2-2 | −1·105 | 0·462 | 0·616 | 0·453 | 0·717 | 1·588 | 0·475–5·311 | 0·024 | 0·023 | 0·018 | 0·017 |

RRR, relative risk ratio; HLR, haplotype logistic regression; EM, expectation-maximization.
[a] By definition cases had a Mini Mental State Examination (MMSE) score of $\leqslant$24 and controls MMSE >24.

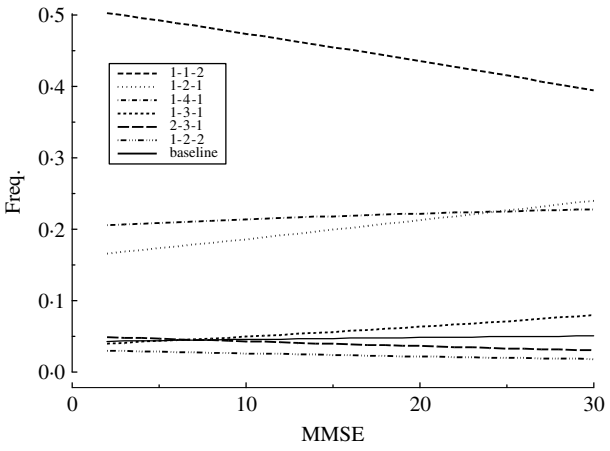


Fig. 1. The estimated haplotype frequencies plotted against the Mini Mental State Examination (MMSE) score for all haplotypes (including the baseline haplotype) by the multiplicative effects model. A flat frequency pattern is shown by the combined haplotype group which serves as the baseline haplotype in the model.

subjects. Our analysis could imply that carriers of the haplotype who manage to survive to the high age of 93 years are likely to be more mentally impaired than the others of the same age.

By assuming multiplicative effects for all the haplotypes, we fitted a full haplotype model for the six haplotypes in Table 1, again using the combined group as baseline, and displayed the haplotype frequency trajectory as a function of the MMSE trait value (Fig. 1). Although the most frequent haplotypes exhibit decreasing or increasing trends, none is statistically significant. Note that the frequency trajectory of the baseline haplotype (unbroken line) is nearly flat. To illustrate how our model works on discrete traits, we code the MMSE phenotype value to 0 (control) for MMSE >24 or 1 (case) for MMSE $\leqslant$24. This roughly converts our continuous phenotype into a binary trait (dementia or non-dementia). In Table 2, we present the results of a full haplotype model again assuming multiplicative haplotype effects. Though the same trend remains for the 1-1-2 haplotype, the *P* value has increased greatly, perhaps as a result of reduced power due to dichotomizing the trait values. Most importantly, with the outputs from our logistic regression model, we can estimate RRR and its confidence interval for each haplotype as shown in Table 2. Moreover, using the intercept and the slope coefficients, our model estimates the frequencies of each haplotype in the cases and controls. Although we know that no significant haplotype was found in our data, for illustrative purposes we conducted a likelihood ratio test to get an overall significance level of IL-6 haplotypes and MMSE association. We have $G = -2(-1335{\cdot}1872 - (-1333{\cdot}6039)) = 3{\cdot}1666$. From the chi-squared distribution with 6 degrees of freedom, we obtained a *P* value of 0·7877 which means no significant association. In Table 2, we also present the adjusted *P* values for specific haplotypes obtained by the simulation-based empirical approach ($B = 1000$, $L = 6$). Although no adjusted *P* value was expected to be significant, the *P* values were all increased as a result of correcting for multiple testing.

Finally, as shown at the right-hand side of Table 2, we compared the frequencies of the six haplotypes estimated by our model with the frequencies estimated by GENECOUNTING software (Zhao *et al.*, 2002), which implements the EM algorithm. One can see that the haplotype frequency estimates by these two different approaches are very close.

## 4. Discussion

We have introduced a haplotype logistic regression model for haplotype-based association analysis using unphased multi-locus genotype data. The formulation of our model allows us to infer the effect of a specific haplotype (point estimation), to construct a confidence

https://doi.org/10.1017/S0016672305007792 Published online by Cambridge University Press

interval for the risk of the haplotype (interval estimation), to estimate haplotype frequency conditional on the trait value for both discrete and continuous traits, and to obtain an overall significance level for the association between the disease trait and a group or all of the haplotypes. Although HWE is assumed for fitting a multiplicative haplotype effects model, it is only required in the control or the normal population in fitting the non-multiplicative models. Moreover, our direct maximization also supports the simulation-based empirical significance for correcting multiple testing.

Although the estimated haplotype frequency in the controls in a case – control study is only relevant to the intercept parameter in our model, the estimation of the intercept parameter together with the slope coefficients is done using all the data. This means that the genotype information for the cases also contributes to the estimation of haplotype frequencies in the controls. This is different from and more efficient than the traditional approach using the EM algorithm (Excoffier & Slatkin, 1995), which separately estimates haplotype frequency for a given disease status. Meanwhile, as the number of genotyped loci increases, there will be a large number of multi-locus genotypes in $S(g)$. Since our model uses haplotype-based parameterization, the parameter space is considerably reduced. This is the same as using allele- instead of genotype-based analysis in single-locus studies (Sasieni, 1997; Tan *et al*., 2004). On the other hand, the number of intercept parameters parallels the number of haplotypes included in the model. As already suggested, one can always group the rare haplotypes to reduce the number of parameters to be estimated. Nevertheless, since our interest in haplotype analysis concerns both risk and frequency, all parameters in our model are meaningful. Since our model uses haplotype-based parameterization, the same risk parameter can be assigned to haplotypes with a similar effect on the trait to reduce the number of parameters in the model. Schaid (2004) summarized the approaches for a cladistic analysis and clustering of haplotypes. Due to potential limitations in the existing methods, for example disregard of the association of haplotype with trait, important aspects remain to be solved. However, once the haplotypes have been clustered, parameterization based on the haplotype clusters will help to largely reduce the parameter space and thus increase the statistical power.

The formulation of our haplotype logistic regression model resembles the retrospective likelihood approach for case – control data proposed by Epstein & Satten (2003). In our model, parameter estimation is achieved by modelling the frequencies of the polytomous haplotype pairs which, with haplotype-based parameterization, link the unobserved haplotypes with the observed multi-locus genotypes. In our retrospective framework, the observed disease trait is assigned as an independent variable which can be easily modelled as in any ordinary logistic regression model whether the trait is discrete or continuous. Since in our haplotype logistic regression model the disease phenotype is an independent variable, the estimated haplotype frequency (dependent variable) is thus unaffected by the over-representation of cases in the samples. This is important in clinical applications where the sampling is highly dependent on the disease phenotypes. In addition, because our retrospective approach summarizes all possible haplotype pairs that are consistent with the observed multi-locus genotypes, the model produces unbiased estimates of the effects of specific haplotypes.

For a given trait value $x$, our model calculates the relative risk of a haplotype as the ratio between the frequency of the haplotype in question and that of the baseline haplotype. Since the relative risk ratio is defined as the ratio between the relative risks at two given trait values, it measures the risk of the haplotype on the transition over the disease status. Since our relative risk ratio is estimated from a retrospective model, it is necessary to study its connection with the relative risk parameter in a general prospective model. In the Appendix, we derive the relationship between the risk parameter in our retrospective model and that in the prospective model. It is shown that, when the disease is rare, the relative risk in a prospective model can be approximated by the relative risk ratio estimated from our retrospective model. This is important because, provided the disease incidence is low in the population, our model can estimate the haplotype risk parameters that can be interpreted in term of trait penetrance as in a prospective model. Likewise, in case of a sex-specific effect, our estimated sex-specific RRR for a haplotype can be seen as the relative risk of the haplotype in developing the disease for given sex. Finally, as shown by equation (7), testing the null hypothesis of $\beta = 0$ is equivalent to testing $H_o: RRR = 1$. This is also shown by the 95% confidence intervals for the estimated RRRs in Table 2. Since no slope parameter for the haplotypes is statistically different from zero, all the 95% confidence intervals of RRR cover the null risk of one.

Although the logit is linearly dependent on the trait, as shown in formula (5), the relationship between haplotype frequency and the trait is actually not linear. Because the logistic regression model does not make any assumptions on normality and linearity of the independent variable, it is less stringent than the ordinary least squares (OLS) regression model. Hosmer & Lemeshow (2000) illustrated different types of models for the relationship between the logit and a covariate or independent variable and concluded

https://doi.org/10.1017/S0016672305007792 Published online by Cambridge University Press

Table A1.

| Haplotype pair | Disease status | | |
|---|---|---|---|
| | Case ($x=1$) | Control ($x=0$) | Total |
| $(a, b)$ | $\pi_{a,b}(1)pN$ | $\pi_{a,b}(0)(1-p)N$ | $[\pi_{a,b}(1)p+\pi_{a,b}(0)(1-p)]N$ |
| $(c, d)$ | $\pi_{c,d}(1)pN$ | $\pi_{c,d}(0)(1-p)N$ | $[\pi_{c,d}(1)p+\pi_{c,d}(0)(1-p)]N$ |
| …… | …… | …… | …… |
| $(i, j)$ | $\pi_{i,j}(1)pN$ | $\pi_{i,j}(0)(1-p)N$ | $[\pi_{i,j}(1)p+\pi_{i,j}(0)(1-p)]N$ |
| …… | …… | …… | …… |
| $(o, o)$ | $\pi_{o,o}(1)pN$ | $\pi_{o,o}(0)(1-p)N$ | $[\pi_{o,o}(1)p+\pi_{o,o}(0)(1-p)]N$ |
| Total | $pN$ | $(1-p)N$ | $N$ |

that the logistic regression model can capture the main effect except for a U-shaped pattern. From formula (5), one can see that the fitted haplotype frequency is a monotonic function of the trait value. Efforts have been made in transforming the independent variable into modelling a non-monotonic frequency pattern (Hosmer & Lemeshow, 2000), but only when the dependent variable is dichotomous. In our case, if a U-shaped frequency exists, it would mean that the same haplotype is responsible for both the low and the high status of the disease phenotype, which is biologically contradictory. Although a non-linear relationship can be modelled using the fractional polynomials when the dependent variable is dichotomous, transformation of the independent variable in the multinomial logistic regression model has been rare. When the functional form of dependence is questioned in multinomial logistic regression, Hosmer & Lemeshow (2000) suggested approximating the fit of a multinomial logistic model by fitting separate binary models using fractional polynomials. In haplotype analysis, this is infeasible because individual haplotypes are not observable. Given the situation, we think that the linear model is a practical approach because the main effect of haplotype association is captured. In addition, because haplotype frequency is modelled conditional on the trait, non-genetic covariates can not be analysed alone in our model. However, as described in Section 2, our model does allow us to assess the interaction effects between haplotypes and other binary or categorical covariates, for example the sex-dependent effects. More work is needed in extending the present model to cover additional non-genetic covariates.

The study was jointly supported by the US National Institute on Aging (NIA) research grants NIA P01 AG08761, -AG13196, the microarray center project under the Biotechnological Research Program financed by the Danish Research Agency and the Danish Medical Research Council. The authors are grateful to Kirsten Pagh for her help in preparing the manuscript.

## Appendix

Suppose we have a case – control study with $x=1$ for cases and $x=0$ for controls. We designate the haplotypes arising from the observed multiple genotype data as $h_a$, $h_b$, $h_c$, ……, $h_o$ with haplotype $h_o$ as the baseline or reference haplotype. Let $N$ be the size of the population from which the samples were taken and $p$ be the frequency of the outcome (disease) in the population. With all the parameters, we can present the entire population in Table A1 in which the total population is divided according to their disease status and haplogenotype (haplotype pair).

In Table A1, $\pi_{i,j}(1)$ and $\pi_{i,j}(0)$ are the frequencies of haplotype pair $(h_i, h_j)$ in the cases and the controls respectively. Likewise, $\pi_{o,o}(1)$ and $\pi_{o,o}(0)$ are the frequencies of the reference haplogenotype in the cases and the controls because haplotype $h_o$ is assigned as the baseline haplotype. Applying formula (7) to the table, we have the $RRR_{i,j}$ as

$$RRR_{i,j}=\frac{\pi_{i,j}(1)}{\pi_{o,o}(1)}\Big/\frac{\pi_{i,j}(0)}{\pi_{o,o}(0)}=\frac{\pi_{i,j}(1)\pi_{o,o}(0)}{\pi_{i,j}(0)\pi_{o,o}(1)}. \tag{A1}$$

Now, for the entire population, the rate of the disease among carriers of haplotype pair $(h_i, h_j)$ is

$$p(D|h_i,h_j)=\frac{\pi_{i,j}(1)pN}{[\pi_{i,j}(1)p+\pi_{i,j}(0)(1-p)]N}$$

$$=\frac{\pi_{i,j}(1)p}{\pi_{i,j}(1)p+\pi_{i,j}(0)(1-p)}. \tag{A2}$$

Similarly, the rate of the disease among carriers of the reference haplotype pair $(h_o,h_o)$ is

$$p(D|h_o,h_o)=\frac{\pi_{o,o}(1)pN}{[\pi_{o,o}(1)p+\pi_{o,o}(0)(1-p)]N}$$

$$=\frac{\pi_{o,o}(1)p}{\pi_{o,o}(1)p+\pi_{o,o}(0)(1-p)}. \tag{A3}$$

From (A2) and (A3), we obtain the relative risk of haplotype pair $(h_i, h_j)$ in a standard prospective

https://doi.org/10.1017/S0016672305007792 Published online by Cambridge University Press

model:

$$RR_{i,j} = \frac{p(D|h_i, h_j)}{p(D|h_o, h_o)} = \frac{\pi_{i,j}(1)p}{\pi_{i,j}(1)p + \pi_{i,j}(0)(1-p)} \Big/ \frac{\pi_{o,o}(1)p}{\pi_{o,o}(1)p + \pi_{o,o}(0)(1-p)} = \frac{\pi_{i,j}(1)}{\pi_{i,j}(1)p + \pi_{i,j}(0)(1-p)} \times \frac{\pi_{o,o}(1)p + \pi_{o,o}(0)(1-p)}{\pi_{o,o}(1)}. \quad \text{(A4)}$$

When $p$ is small, we have in (A4), $\pi_{i,j}(1)p + \pi_{i,j}(0)(1-p) \cong \pi_{i,j}(0)$ and $\pi_{o,o}(1)p + \pi_{o,o}(0)(1-p) \cong \pi_{o,o}(0)$. With this approximation, we have

$$RR_{i,j} \approx \frac{\pi_{i,j}(1)\pi_{o,o}(0)}{\pi_{i,j}(0)\pi_{o,o}(1)} = RRR_{i,j}. \quad \text{(A5)}$$

Equation (A5) shows the connection between the risk parameters in our retrospective model and the general prospective model under the condition that the disease in the population is relatively rare. With this relationship, the relative risk parameter in the prospective model can be approximated by the RRR in our retrospective model.

## References


Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second Internal Symposium on Information Theory* (ed. B. N. Petrov & F. Csake), pp. 267–281. Budapest: Akademiai Kiado.

Akey, J., Jin, L. & Xiong, M. (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics* **9**, 291–300.

Christiansen, L., Bathum, L., Andersen-Ranberg, K., Jeune, B. & Christensen, K. (2004). Modest implication of interleukin 6 promoter polymorphisms in longevity. *Mechanisms of Ageing and Development* **125**, 391–395.

Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genetic Epidemiology* **27**, 321–333.

Comeron, J. M., Kreitman, M. & De La Vega, F. M. (2003). On the power to detect SNP/phenotype association in candidate quantitative trait loci genomic regions: a simulation study. In *Pacific Symposium on Biocomputing*, pp. 478–489.

Epstein, M. P. & Satten, G. A. (2003). Inference on haplotype effects in case–control studies using unphased genotype data. *American Journal of Human Genetics* **73**, 1316–1329.

Excoffier, L. & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12**, 921–927.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edn. New York: Wiley.

Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association* **81**, 471–476.

Lake, S. L., Lyon, H., Tantisira, K., Silverman, E. K., Weiss, S. T., Laird, N. M. & Schaid, D. J. (2003). Estimation and tests of haplotype–environment interaction when linkage phase is ambiguous. *Human Heredity* **55**, 56–65.

Licastro, F., Grimaldi, L. M., Bonafe, M., Martina, C., Olivieri, F., Cavallone, L., Giovanietti, S., Masliah, E. & Franceschi, C. (2003). Interleukin-6 gene alleles affect the risk of Alzheimer's disease and levels of the cytokine in blood and brain. *Neurobiology of Aging* **24**, 921–926.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. & Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70**, 425–434.

Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. *Genetic Epidemiology* **27**, 348–364.

Tan, Q., De Benedictis, G., Yashin, A. I., Bathum, L., Christiansen, L., Dahlgaard, J., Frizner, N., Vach, W., Vaupel, J. W., Christensen, K. & Kruse, T. A. (2004). Assessing genetic association with human survival at multi-allelic loci. *Biogerontology* **5**, 89–97.

Zhao, J. H. & Sham, P. C. (2002). Faster haplotype frequency estimation using unrelated subjects. *Human Heredity* **53**, 36–41.

Zhao, J. H., Lissarrague, S., Essioux, L. & Sham, P. C. (2002). GENECOUNTING: haplotype analysis with missing genotypes. *Bioinformatics* **18**, 1694–1695.