doi: 10.1046/j.1529-8817.2004.00143.x

Haplotype Effects on Human Survival: Logistic Regression Models Applied to Unphased Genotype Data

Q. Tan[1,*], L. Christiansen[1,2], L. Bathum[1,2], J. H. Zhao[3], W. Vach[4], J. W. Vaupel[5], K. Christensen[2] and T. A. Kruse[1]

[1] *Department of Clinical Biochemistry and Genetics, Odense University Hospital, Denmark*
[2] *Institute of Public Health, University of Southern Denmark, Odense, Denmark*
[3] *Department of Epidemiology and Public Health, University College London, UK*
[4] *Department of Statistics, University of Southern Denmark, Odense, Denmark*
[5] *Max-Planck Institute for Demographic Research, Rostock, Germany*

Summary

Haplotype based linkage disequilibrium (LD) mapping exhibits higher power than the single locus approach because it makes use of the LD information contained in the flanking markers. New statistical methods have been proposed to help to infer haplotype effects on human diseases using multi-locus genotype data collected from unrelated individuals. In this paper, we introduce a statistical procedure for measuring haplotype effects on human survival using the popular logistic regression model with haplotype based parameterizations. By modeling haplotype frequency as a function of age, our model infers haplotype effects by estimating and testing the slope parameters under different genetic mechanisms (multiplicative, dominant, or recessive). In addition, by estimating the sex-specific slope parameters, our model allows the detection of sex-specific haplotype effects or haplotype-sex interactions. As an example, we apply our model to an empirical dataset on a stress related gene, *interleukin-6*, to look for haplotypes that affect individual survival and for haplotype-sex interactions. We show that our logistic regression based haplotype model can be a helpful tool for researchers interested in the genetics of human aging and longevity.

Keywords: logistic regression, haplotype, human survival, unphased genotype data

Introduction

Association based linkage disequilibrium mapping (Jorde, 2000; Weiss & Clark, 2002) maybe a useful tool in the genetic study of human aging and longevity (De Benedictis *et al.* 2001). Taking advantage of the newly emerging high-throughput SNP genotyping technique, which enables high-density genome-wide screening of complex trait genes, LD mapping is gaining more popularity (Gray *et al.* 2000). Such methodology challenges the traditional locus-by-locus approach in association studies. As lifespan is a complex trait, multi-locus statistical approaches that take into account the interdependence of genetic variants are crucial in mapping the genes that modulate human survival.

Haplotype based LD mapping represents an increasingly important association approach in localizing human complex disease genes (Collins *et al.* 1999). Because particular DNA variants may remain together on ancestral haplotypes for many generations, groups of neighbouring genetic variants can form haplotypic diversity with distinctive patterns of LD, and this can be exploited in both genetic linkage and association studies (Schork *et al.* 2000). Haplotype analysis is more efficient than a locus-by-locus association test because it makes use of the LD information contained in the flanking markers (Akey *et al.* 2001). As what we observe in practice is multi-locus genotypes instead of haplotypes, directly calculating haplotype frequency is problematic

[*] Author for correspondence: Dr. Qihua Tan, Department of Clinical Biochemistry and Genetics (KKA), Odense University Hospital, Sdr. Boulevard 29, DK-5000 Odense C, Denmark, Tele. 0045 65412822, Fax: 0045 65411911. E-mail: qihua.tan@ouh.fyns-amt.dk

in cases of missing parental genotype information or on "ambiguous triad" (Hodge *et al.* 1999). With the implementation of the EM algorithm, haplotype frequencies can be estimated from data of unrelated individuals (Excofficer & Slatkin, 1995) and used to infer haplotype effects in case-control studies (Zhao & Sham, 2002; Schaid *et al.* 2002; Epstein & Satten, 2003). In longevity studies, the traditional haplotype estimating technique has been applied to analyze multi-locus genotype data collected using a case-control setup, with cases being the long-lived (usually centenarians) and controls being young individuals (Bonafe *et al.* 2002; Ross *et al.* 2003; Geesaman *et al.* 2003; Christiansen *et al.* 2004). Such practice suffers from the power problem, as with the group-wise gene frequency approach in single locus analysis (Yashin *et al.* 1999). In a recent development, Lin (2004) proposed a semiparametric Cox proportional hazard model for estimating relative risk of haplotype on age at disease onset, using unphased genotype data from cohort studies. Differing from age at disease onset, longevity studies using a cohort setup have been rare due to the expense of follow-ups both in term of time and money. Although it is possible to implement the existing model in longevity studies, it is necessary to develop alternative methods to apply to unphased genotype data collected using the popular cross-sectional design.

As a widely used method in the field of epidemiology, the logistic regression model has been applied to estimate the genetic effects on human survival at polymorphic loci in cross-sectional studies (Tan *et al.* 2003). In this paper, we extend the multinomial logistic regression model (Hosmer & Lemeshow, 2000) to deal with multi-locus unphased genotype data. By haplotype based parameterization on the observed multi-locus data, and assuming Hardy-Weinberg equilibrium at birth, the model models haplotype frequency as a function of age to infer haplotype effects on human survival. Our strategic parameterization allows us to investigate the different genetic mechanisms concerning the haplotype function (multiplicative, dominant and recessive), and to estimate sex-specific haplotype effects or haplotype-sex interactions. Important haplotypes can be grouped into one model for extensive analysis, and the Akaike information criterion (AIC) (Akaike, 1973) can be used for selecting the best performance model from the models with different haplotypes included. We apply the model to an empirical multi-locus genotype dataset collected in an association study on the *interleukin-6* (*IL-6*) gene and human longevity (Christiansen *et al.* 2004), to show how our method can be used to search for important haplotypes that affect human survival. We end with a brief discussion on the significance of the model and on practical issues concerning model applications.

Methods

The Multinomial Logistic Regression Model with Haplotype Based Parameterization

We suppose that complete genotype information is available at a series of m loci. Let G denote the collection of all possible multi-locus genotypes observed and H denote the collection of all the haplotypes. The combinations of haplotypes in H form the haplotype pairs or haplogenotypes that make up G. Assuming at age x the frequency of haplotype pair (h_i, h_j) is $\pi_{i,j}(x)$, and defining the haplotype pair formed by the baseline haplotype h_o as the reference haplogenotype, we obtain the multinomial logistic regression model with polytomous responses (here the haplogenotypes) as

$$\begin{aligned}&\ln[\pi_{i,j}(x)/\pi_{o,o}(x)]\\&\quad=\begin{cases}\alpha_{i,j}+\beta_{i,j}x & i=j\\ \ln 2+\alpha_{i,j}+\beta_{i,j}x & i<j\end{cases}\\&\qquad\alpha_{o,o}=0,\quad \beta_{o,o}=0\quad i,j\in H\end{aligned} \tag{1}$$

In (1), $\pi_{o,o}(x)$ is the frequency at age x for the reference haplogenotype (h_o, h_o). In this model, age related changes in the haplogenotype frequency are represented by the slope parameter $\beta_{i,j}$ while the intercept $\alpha_{i,j}$ is related to the haplogenotype frequency at birth. Equation (1) is parameterized on the haplogenotypes or the pair-wise combinations of the haplotypes in H. The number of such combinations can increase drastically with the number of loci covered and the degrees of their polymorphism. A parsimonious method of parameterization is necessary to ensure the statistical power of the model. Assuming Hardy-Weinberg equilibrium and multiplicative haplotype effects, we introduce the haplotype based parameterization into (1) by letting $\alpha_{i,j} = \alpha_i + \alpha_j$ and $\beta_{i,j} = \beta_i + \beta_j$ for haplogenotype (h_i, h_j). Now we can rearrange (1) such that the haplogenotype frequency can be expressed in terms of the haplotype parameters.

$$\pi_{i,j}(x) = \begin{cases} \exp[(\alpha_i + \alpha_j) + (\beta_i + \beta_j)x] \Big/ \sum_{i',j' \in H} \exp[(\alpha_{i'} + \alpha_{j'}) + (\beta_{i'} + \beta_{j'})x] & i = j \\ 2\exp[(\alpha_i + \alpha_j) + (\beta_i + \beta_j)x] \Big/ \sum_{i',j' \in H} \exp[(\alpha_{i'} + \alpha_{j'}) + (\beta_{i'} + \beta_{j'})x] & i < j \end{cases} \quad i, j \in H \qquad (2)$$

Given Hardy-Weinberg equilibrium, we easily obtain the frequency of haplotype h_i at age x from the square root of $\pi_{i,i}(x)$ as

$$\pi_i(x) = \exp(\alpha_i + \beta_i x) \Big/ \sqrt{\sum_{i',j' \in H} \exp[(\alpha_{i'} + \alpha_{j'}) + (\beta_{i'} + \beta_{j'})x]} \quad i \in H \qquad (3)$$

For the baseline haplotype h_o, because $\alpha_{o,o} = 2\alpha_0 = 0$ and $\beta_{o,o} = 2\beta_0 = 0$, we have $\alpha_o = 0$ and $\beta_o = 0$. Then the frequency of the baseline haplotype at age x is

$$\pi_o(x) = 1 \Big/ \sqrt{\sum_{i',j' \in H} \exp[(\alpha_{i'} + \alpha_{j'}) + (\beta_{i'} + \beta_{j'})x]}. \qquad (4)$$

Based on (3) and (4), we can estimate the odds ratio for measuring the effect of haplotype h_i for an increase over k years in age x as

$$\begin{aligned} OR_i(k) &= [\pi_i(x)/\pi_o(x)]/[\pi_i(x-k)/\pi_o(x-k)] \\ &= \exp(\beta_i k) \qquad i \in H. \end{aligned} \qquad (5)$$

(5) means that when β_i is significantly different from zero, the frequency of the haplotype goes up if $\beta_i > 0$ or down if $\beta_i < 0$ with increasing age. Alternatively, we can calculate the odds ratio over k years in age x for carriers of haplotype pair (h_i, h_j) from (1) as

$$\begin{aligned} OR_{i,j}(k) &= [\pi_{i,j}(x)/\pi_{o,o}(x)]/[\pi_{i,j}(x-k)/\pi_{o,o}(x-k)] \\ &= \exp(\beta_{i,j}k) = \exp(\beta_i k)\exp(\beta_j k) \\ &= OR_i(k)\, OR_j(k) \qquad i, j \in H \end{aligned} \qquad (6)$$

(6) shows clearly the multiplicative haplotype effects in determining the odds ratio for the corresponding haplogenotype.

The Likelihood Function

For each multi-locus genotype$g$, there is a set of haplotype pairs, denoted as $S(g)$, that are consistent with g. With this relationship, the frequency of multi-locus genotype g at age x can be calculated as the sum of frequencies for all the haplogenotypes $S(g)$ as expressed in terms of the haplotype parameters in (2), i.e.

$$\pi_g(x) = \sum_{i,j \in S(g)} \pi_{i,j}(x). \qquad (7)$$

(7) is important because, in practice, what we observe are multi-locus genotypes instead of haplotypes. Equation (7) links the observed multi-locus genotypes with the ambiguous haplotypes which we don't observe. Denoting the number of individuals carrying multi-locus genotype$g$with $n_g(x)$, we construct the likelihood function at age x using the multinomial distribution of the multi-locus genotype frequencies in the population as

$$\log L_{\text{data}}(x) \propto \sum_{g \in G} n_g(x) \log \pi_g(x). \qquad (8)$$

The likelihood of the entire data is simply the sum of (8) over all the observed ages.

Sex-specific Haplotype Effects

Similar to Tan *et al.* (2003), we can modify our logistic regression model to account for sex-specific haplotype effects by assigning different slope parameters to the two sexes, such that (1) becomes

$$\ln[\pi_{i,j}(x)/\pi_{o,o}(x)] = \begin{cases} \alpha_{i,j} + {}_m\beta_{i,j}xU + {}_f\beta_{i,j}x(1-U) & i = j \\ \ln 2 + \alpha_{i,j} + {}_m\beta_{i,j}xU + {}_f\beta_{i,j}x(1-U) & i < j \end{cases}$$

$$\alpha_{o,o} = 0, \quad {}_m\beta_{o,o} = 0, \quad {}_f\beta_{o,o} = 0 \quad i, j \in H \qquad (9)$$

where U is an indicator of sex with $U = 1$ for males and $U = 0$ for females. Based on the law of segregation, and assuming the gene of interest does not affect *in utero* survival, we assign the same intercept parameter for both sexes to reduce the number of parameters in the model. By constructing the Wald test using the variance-covariance matrix, statistical significance can be assessed to infer if the slopes are different for the two sexes. When no haplotype-sex interaction exists, the same slope parameter can be assigned so that (9) reduces to (1).

Dominant or Recessive Effects

In (2), we assume the effects of haplotypes are multiplicative such that $\beta_{i,j} = \beta_i + \beta_j$ for haplogenotype (h_i,h_j) and $\beta_{i,i} = 2\beta_i$ for (h_i, h_i). When the effect of haplotype h_i is dominant, we have $\beta_{i,j} = \beta_{i,i} = \beta_i$. While in the case of a recessive effect, we have $\beta_{i,j} = 0$ and $\beta_{i,i} = \beta_i$. Note that in the multiplicative model, at any age x, $\pi_{i,j}(x) = 2\pi_i(x)\pi_j(x)$ for heterozygous and $\pi_{i,i}(x) = \pi_i(x)^2$ for homozygous haplogenotypes. This amounts to the requirement of Hardy-Weinberg equilibrium over all the ages in the sampled data. However, in the non-multiplicative models, the symmetry in the slope parameters no longer exists. In this case, haplotype frequencies are no longer in Hardy-Weinberg proportion, which in turn means that the Hardy-Weinberg assumption can be relaxed when fitting a non-multiplicative model.

Data Analyzing Strategies

In order to carry out the analysis, we first collect all the unique multi-locus genotypes occurring in the data to form G, and then count the numbers of each of the multi-locus genotypes at each age; for example multi-locus genotype g at age x, to use as $n_g(x)$ in (8). Then for each multi-locus genotype g, find the collection of all the haplotype pairs that are consistent with g to form $S(g)$ to use in (7). In the likelihood function (8), the frequency of each multi-locus genotype g is expressed in terms of the haplotype parameters through the linkage provided by (7) and (2). Once the relationship is established, parameters can be estimated by maximizing (8) with the observed data.

Because the number of haplotypes goes up exponentially with the number of loci, the model can be weakly powered due to the large number of parameters to be estimated. We suggest, in an initial analysis, estimating the slope parameter for each single haplotype separately, by assuming no effect from the other haplotypes, with their slopes set to zero. This can be done for the different models assuming multiplicative, dominant or recessive haplotype effects. For each estimation, we record the AIC for selecting the top performance haplotypes. The selected haplotypes, together with their corresponding modes, can be put into one model for an extensive analysis. Models with different combinations of the selected haplotypes can be fitted and compared again, according to their newly recorded AICs, to find the best performance model. Parameter estimates in the best performance model are reported as the final results.

As the number of haplotypes increases with the number of loci and degree of polymorphism at the loci, so does the number of rare haplotypes. It has been shown that the power for detecting association with rare haplotypes is very low (Comeron *et al.* 2003). Depending on the sample size and the number of possible haplotypes, a frequency threshold could be set up such that low frequency haplotypes can be pooled together to form a combined haplotype (Lake *et al.* 2003). The combined haplotype could be used as the baseline haplotype in the analysis, although other alternatives such as the most frequent or wild-type haplotype may also be a good choice (Lake *et al.* 2003). Combining the rare haplotypes could improve the power of the model due to the reduced number of parameters to be estimated for the same data. At the same time, the reduction in multiple testing can help to reduce the type 1 error rate as well. In the parameter estimation, as just mentioned, only the slope parameter for the haplotype of interest is estimated while the slopes of the other haplotypes are set to zero. This means that the baseline to the slope parameter is formed by all the other haplotypes. This not only reduces the number of parameters in the model but should also help to increase stability in the estimates.

Application

The *interleukin-6* gene (*IL 6*) has been associated with stress conditions that are characterized by the aging process, such as Alzheimer's disease (Licastro *et al.* 2003), cardiovascular events (Cesari *et al.* 2003) and type 2 diabetes (Vozarova *et al.* 2003). In a recent study, Christiansen *et al.* (2004) investigated the influence of *IL 6* on human survival in the Danish population. Haplotype analysis was carried out on a total of 1143 Danes genotyped at two single-point polymorphisms (-572G/C and -174G/C) and one AT-stretch polymorphism (-373AnTm, 4 alleles) in the promoter region. Of

Table 1 Parameter estimates in the logistic regression model in the initial analysis

	Fitted logistic regression model				
		Slope			
Haplotype	α	β	S.E.	$\text{p}_{-\text{value}}$	AIC
Multiplicative					
G/A$_8$T$_{12}$/C	2.806	-0.004	0.002	0.040	5405.459
G/A$_9$T$_{11}$/G	1.560	0.002	0.003	0.369	5408.883
G/A$_{10}$T$_{11}$/G	1.448	0.004	0.003	0.161	5408.726
G/A$_{10}$T$_{10}$/G	0.736	-0.002	0.004	0.656	5409.499
C/A$_{10}$T$_{10}$/G	0.071	-0.001	0.005	0.815	5409.657
C/A$_9$T$_{11}$/C	-1.774	0.013	0.009	0.139	5407.302
Dominant					
G/A$_8$T$_{12}$/C	2.472	0.000	0.001	0.778	5409.619
G/A$_9$T$_{11}$/G	1.767	-0.001	0.002	0.739	5409.588
G/A$_{10}$T$_{11}$/G	1.522	0.003	0.002	0.104	5407.003
G/A$_{10}$T$_{10}$/G	0.473	0.002	0.004	0.628	5409.435
C/A$_{10}$T$_{10}$/G	0.181	-0.003	0.005	0.560	5409.401
C/A$_9$T$_{11}$/C	-1.774	0.013	0.009	0.139	5407.302
Recessive*					
G/A$_8$T$_{12}$/C	2.559	-0.003	0.002	0.078	5406.553
G/A$_9$T$_{11}$/G	1.691	0.003	0.002	0.225	5408.266
G/A$_{10}$T$_{11}$/G	1.738	-0.002	0.002	0.476	5409.181
G/A$_{10}$T$_{10}$/G	0.647	-0.015	0.010	0.147	5406.651
C/A$_{10}$T$_{10}$/G	-0.034	0.007	0.008	0.435	5409.195

*Estimation on C/A$_9$T$_{11}$/C haplotype was not possible due to low frequency.

the 16 possible haplotypes arising from the three loci, only 10 are present in the Danish population. Hardy-Weinberg equilibrium was observed for the overall and the age-grouped data (Christiansen *et al.* 2004). Haplotype frequencies in the young (<70 years, 567 individuals) and old (93 years, 576 individuals) age groups were compared for the 6 most common haplotypes. A noticeable decrease with age in the frequency of the -572G/-373A$_8$T$_{12}$/-174C haplotype (indicated as G/A$_8$T$_{12}$/C) was reported (Christiansen *et al.* 2004). Taking the *IL 6* data as an example, we show how our logistic regression model can be applied to infer the haplotype effects on individual survival, as well as to estimate the haplotype frequencies over the observed ages. Similar to Christiansen *et al.* (2004), we combine the 4 rare haplotypes to form one haplotype group, and assign it as the baseline haplotype in the analysis. Following the analyzing strategy, we first conduct an initial analysis on each haplotype and estimate the parameters by introducing the multiplicative, dominant, and recessive models (Table 1). The Wald test statistics are calculated for the slope parameters to assess their statistical significance.

In Table 1, haplotype G/A$_8$T$_{12}$/C has the lowest AIC in the multiplicative model. The p-value for its slope is 0.040. The negative slope for the haplotype indicates that it is a harmful haplotype. The second and third well performing models are the recessive model for the G/A$_{10}$T$_{10}$/G haplotype and the dominant model for the G/A$_{10}$T$_{11}$/G haplotype, respectively. However, the p-values for the slope parameters of the two haplotypes are all above 0.05. In addition to the assessment of haplotype effects, with the estimated haplotype parameters, we calculate frequencies for all the haplotypes in the multiplicative model by using (3) (Table 2). The estimated frequency for the G/A$_8$T$_{12}$/C haplotype decreases from 0.481 at age 46 (the lowest observed age) to 0.431 at age 93 (the highest observed age), due to the increased rate of death for carriers of the haplotype. Most importantly, our logistic regression model produces haplotype frequency estimates comparable to those obtained by the EM algorithm (Christiansen *et al.* 2004).

In Table 3, we use the AIC for selecting the best performance model from the three models, built up by consecutively adding the above three top haplotypes in the models according to their corresponding AICs in Table 1. Adding the recessive G/A$_{10}$T$_{10}$/G haplotype to model 1, which only includes a multiplicative G/A$_8$T$_{12}$/C haplotype, results in a smaller AIC in model 2. However, adding the third haplotype to model 2 does not improve performance in model 3 (AIC increases). We thus choose the two-haplotype model (model 2) as the best model. From model 2, we estimate, using formula (5), the odds ratio for haplotype G/A$_8$T$_{12}$/C over the observed age range as 0.83, indicating the modest and deleterious effect of the haplotype on individual survival as reported by Christiansen *et al.* (2004).

By using (9), we also fit the logistic regression model with sex-specific slope parameters to each haplotype. The sex-specific slope parameters for the G/A$_8$T$_{12}$/C haplotype ($\beta = -0.005$, p-value $= 0.025$ for males; $\beta = -0.004$, p-value $= 0.078$ for females) show different statistical significance, with males more significant than females. We then assess the difference in the two slope parameters by calculating the Wald test statistic using the covariance information, which re-

Table 2 Comparison of haplotype frequencies estimated by the logistic regression model and by the EM algorithm

	Haplotype					
Age	$G/A_8T_{12}/C$	$G/A_9T_{11}/G$	$G/A_{10}T_{11}/G$	$G/A_{10}T_{10}/G$	$C/A_{10}T_{10}/G$	$C/A_9T_{11}/C$
Logistic regression*						
46	0.481	0.201	0.192	0.072	0.038	0.012
93	0.431	0.218	0.220	0.066	0.036	0.021
EM algorithm**						
<70	0.470	0.203	0.196	0.071	0.038	0.013
93	0.432	0.217	0.222	0.066	0.035	0.020

*Multiplicative effect model.

**From Christiansen *et al.* (2004).

Table 3 Comparison of the top performance models using AIC in the extensive analysis

	$G/A_8T_{12}/C$		$G/A_{10}T_{10}/G$		$G/A_{10}T_{11}/G$		
	Multiplicative		Recessive		Dominant		
Model	α_1	$\beta_1(p_{-\text{value}})$	α_2	$\beta_2(p_{-\text{value}})$	α_3	$\beta_3(p_{-\text{value}})$	AIC
1	2.806	−0.004(0.040)					5405.459
2	2.809	−0.004(0.037)	0.649	−0.015(0.144)			5404.306
3	2.760	−0.004(0.085)	0.644	−0.014(0.149)	1.580	0.002(0.273)	5405.094

sults in a p-value of 0.597. We thus conclude that our data cannot yet confirm that there is a significant sex-dependent effect for this haplotype.

Discussion

Similar to genetic association studies in human complex diseases (Risch, 2000; Botstein & Risch, 2003), association based LD mapping is more powerful than linkage approaches in localizing genes that contribute to human survival (Tan *et al.* 2004). With the completion of the human genome project and newly emerging high throughput SNP genotyping techniques, abundant genetic information is becoming available for mapping human complex trait genes. New statistical methods for accommodating this situation are appealing. In this paper, we have presented a logistic regression approach to estimate haplotype effects on human survival using multi-locus genotype data collected from cross-sectional studies. Different from the group-wised approach, our model makes full use of individual phenotype information by modeling haplotype frequency as a function of age. In our model, haplotype effects on survival can be detected by estimating and testing the corresponding slope parameters under different genetic mechanisms (multiplicative, dominant, or recessive). By specifying sex-specific slope parameters, our model also allows the investigators to infer sex-dependent haplotype effects or haplotype-sex interactions. With the fitted model, the haplotype frequency at any given age can be easily calculated to examine the age pattern in haplotype frequency. Moreover, the AIC can be used for discriminating the models fitted under different modes of haplotype function, and for selecting the best performance model when multiple haplotypes are involved.

The observed individual age in our data is the age at participation, or the age when biological sample was taken. This means that, though we are dealing with a survival trait, we don't actually observe the individual's life span. Although completely censored, such data is not a problem for our model because we are modeling the haplotype frequency by age. In survival modeling, new statistical methods have been proposed to analyze single locus data collected using the cross-sectional design to estimate allele or genotype relative risks (Yashin *et al.* 1999). Recently, we have extended the single locus model to estimate haplotype relative risks using unphased multi-locus genotype data. It is interesting that application of the survival model to our *IL 6* data has produced consistent results. As the logistic regression model is widely used in epidemiologic studies, we think our logistic regression based approach may be an important alternative to those who are unfamiliar with survival modeling.

In fitting the logistic regression model, we assume haplotype frequencies at birth follow the Hardy-Weinberg law. As long as the genes we are interested in do not affect *in utero* survival, and there is no preferential transmission of a particular genetic variant in the region under investigation, such an assumption is sensible: differential survival driven by the association between the haplotypes and hazard of death has not yet imposed survival selection on the subjects. With this assumption, genotype frequency information at other ages can contribute to the estimation of haplotype frequencies at birth. As long as Hardy-Weinberg equilibrium holds at birth, we can relax the assumption on haplotype frequencies at the other ages, except in the multiplicative model. This is important because different genetic mechanisms of haplotype function in human survival can be tested without the requirement for Hardy-Weinberg equilibrium at advanced ages.

Although new haplotype based approaches have been proposed for mapping binary (Epstein & Satten, 2003), categorical, and continuous (Schaid *et al.* 2002) disease traits, and even survival traits in cohort studies (Lin, 2004) using unphased genotype data, to our knowledge there has been no statistical method derived for analyzing survival traits using a cross-sectional setup. Given the popularity of the cross-sectional design in genetic studies on human aging and longevity, we hope that application of our model can help to promote haplotype based analysis in this field. Moreover, by modeling haplotype frequency as a function of the disease status, or of the disease trait, our logistic regression model can easily be applied to infer haplotype effects on human diseases (binary or categorical traits). Although our model is proposed for analyzing unrelated data collected using a cross-sectional design, the same setting also applies to cohort data from unrelated individuals if available. In such a case, we are modeling the haplotype frequency change in the aging cohort. In all situations, similar to any association design that deploys data from unrelated individuals, efforts should be taken to assess (Freedman *et al.* 2004), and to account for (Satten *et al.* 2001), population substructure in the sampling population whenever necessary.

Acknowledgements

This work was supported by the US National Institute on Aging (NIA) research grants NIA-P01-AG08761, -AG13196 and by "The MicroArray Center" project under the Biotechnological Research Program financed by the Danish Research Agency. The authors want to thank Shuxia Li for technical support.

References

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

Akey, J., Jin L. & Xiong, M. (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* **9**, 291–300.

Bonafe, M., Marchegiani, F., Cardelli, M., Olivieri, F., Cavallone, L., Giovagnetti, S., Pieri, C., Marra, M., Antonicelli, R., Troiano, L., Gueresi, P., Passeri, G., Berardelli, M., Paolisso, G., Barbieri, M., Tesei, S., Lisa, R., De Benedictis, G. & Franceschi, C. (2002) Genetic analysis of Paraoxonase (PON1) locus reveals an increased frequency of Arg192 allele in centenarians. *Eur J Hum Genet* **10**, 292–296.

Botstein, D. & Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33** Suppl., 228–37.

Cesari, M., Penninx, B. W., Newman, A. B., Kritchevsky, S. B., Nicklas, B. J., Sutton-Tyrrell, K., Rubin, S. M., Ding, J., Simonsick, E. M., Harris, T. B. & Pahor M. (2003) Inflammatory markers and onset of cardiovascular events: results from the Health ABC study. *Circulation* **108**, 2317–2322.

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

Collins, A., Lonjou, C. & Morton, N. E. (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci U S A* **96**, 15173–15177.

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. *Pac Symp Biocomput* 478–489.

De Benedictis, G., Tan, Q., Jeune, B., Christensen, K., Ukraintseva, S. V., Bonafe, M., Franceschi, C., Vaupel, J. W. & Yashin, A. I. (2001) Recent advances in human gene-longevity association studies. *Mech Ageing Dev* **122**, 909–920.

Epstein, M. P. & Satten, G. A. (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* **73**, 1316–1329.

Excoffier, L. & Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**, 921–927.

Freedman M. L., Reich D., Penney K. L., McDonald G. J., Mignault A. A., Patterson N., Gabriel S. B., Topol E. J., Smoller J. W., Pato C. N., Pato M. T., Petryshen T. L., Kolonel L. N., Lander E. S., Sklar P., Henderson B., Hirschhorn J. N. & Altshuler D. (2004) Assessing the impact of population stratification on genetic association studies. *Nat Genet* **6**, 388–393.

Geesaman, B. J., Benson, E., Brewster, S. J., Kunkel, L. M., Blanche, H., Thomas, G., Perls, T. T., Daly, M. J. & Puca, A. A. (2003) Haplotype-based identification of a microsomal transfer protein marker associated with the human lifespan. *Proc Natl Acad Sci USA* **100**, 14115–14120.

Gray, I. C., Campbell, D. A. & Spurr, N. K. (2000) Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet* **9**, 2403–2408.

Hodge, S. E., Boehnke, M. & Spence, M. A. (1999) Loss of information due to ambiguous haplotyping of SNPs. *Nat Genet* **21**, 360–361.

Hosmer, D. W. & Lemeshow, S. (2000) *Applied Logistic Regression*, Second edition, Wiley, USA.

Jorde, L. B. (2000) Linkage disequilibrium and the search for complex disease genes. *Genome Res* **10**, 1435–1444.

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. *Hum Hered* **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. *Neurobiol Aging* **24**, 921–926.

Lin, D. Y (2004) Haplotype-based association analysis in cohort studies of unrelated individuals. *Genet Epidemol* **26**, 255–264.

Risch, N. J. (2000) Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856.

Ross, O. A., Curran, M. D., Rea, I. M., Hyland, P., Duggan, O., Barnett, C. R., Annett, K., Patterson, C., Barnett, Y. A. & Middleton, D. (2003) HLA haplotypes and TNF polymorphism do not associate with longevity in the Irish. *Mech Ageing Dev* **124**, 563–567.

Satten G. A., Flanders W. D. & Yang Q. (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* **68**, 466–477.

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. *Am J Hum Genet* **70**, 425–434.

Schork, N. J., Fallin, D. & Lanchbury J. S. (2000) Single nucleotide polymorphisms and the future of genetic epidemiology. *Clin Genet* **58**, 250–264.

Tan, Q., Bathum, L., Christiansen, L., De Benedictis, G., Bellizzi, D., Rose, G., Frizner, N., Dahlgaard, J., Vach, W., Vaupel, J. W., Yashin, A. I., Christensen, K. & Kruse, T. A. (2003) Logistic regression models for polymorphic and antagonistic pleiotropic gene action on human aging and longevity. *Ann Hum Genet* **67**, 598–607.

Tan, Q., Zhao, J. H., Iachine, I., Hjelmborg, J., Vach, W., Vaupel, J. W., Christensen, K. & Kruse, T. A. (2004) Power of non-parametric linkage analysis in mapping genes contributing to human longevity in long-lived sib-pairs. *Genet Epidemiol* **26**, 245–253.

Vozarova, B., Fernandez-Real, J. M., Knowler, W. C., Gallart, L., Hanson, R. L., Gruber, J. D., Ricart, W., Vendrell, J., Richart, C., Tataranni, P. A. & Wolford J. K. (2003) The interleukin-6 (-174) G/C promoter polymorphism is associated with type-2 diabetes mellitus in Native Americans and Caucasians. *Hum Genet* **112**, 409–413.

Weiss, K. M. & Clark, A. G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* **18**, 19–24.

Yashin, A. I., De Benedictis, G., Vaupel, J. W., Tan, Q., Andreev, K. F., Iachine, I. A., Bonafe, M., DeLuca, M., Valensin, S., Carotenuto, L. & Franceschi, C. (1999) Genes, demography, and lifespan: The contribution of demographic data in genetic studies on aging and longevity. *Am J Hum Genet* **65**, 1178–1193.

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

Received: 13 October 2003
Accepted: 2 August 2004