doi: 10.1046/j.1529-8817.2003.00051.x

Logistic Regression Models for Polymorphic and Antagonistic Pleiotropic Gene Action on Human Aging and Longevity

Qihua Tan[1,*], L. Bathum[1], L. Christiansen[1], G. De Benedictis[2], J. Dahlgaard[1], N. Frizner[1], W. Vach[3], J. W. Vaupel[4], A. I. Yashin[4], K. Christensen[5] and T. A. Kruse[1]

[1]*Department of Clinical Biochemistry and Genetics, KKA, Odense University Hospital, Odense, Denmark*
[2]*Department of Cell Biology, University of Calabria. Rende, Italy*
[3]*Department of Statistics and Demography, University of Southern Denmark*
[4]*Max-Planck Institute for Demographic Research, Rostock, Germany*
[5]*Institute of Public Health, University of Southern Denmark*

Summary

In this paper, we apply logistic regression models to measure genetic association with human survival for highly polymorphic and pleiotropic genes. By modelling genotype frequency as a function of age, we introduce a logistic regression model with polytomous responses to handle the polymorphic situation. Genotype and allele-based parameterization can be used to investigate the modes of gene action and to reduce the number of parameters, so that the power is increased while the amount of multiple testing minimized. A binomial logistic regression model with fractional polynomials is used to capture the age-dependent or antagonistic pleiotropic effects. The models are applied to *HFE* genotype data to assess the effects on human longevity by different alleles and to detect if an age-dependent effect exists. Application has shown that these methods can serve as useful tools in searching for important gene variations that contribute to human aging and longevity.

Introduction

The traditional case-control design has been popular in genetic association studies on human aging and longevity (Kervinen *et al*. 1994; Schachter *et al*. 1994; De Benedictis *et al*. 1997, 1998a,b; Bathum *et al*. 1998, 2001; Wang *et al*. 2001). As aging is a continuous process, some important trajectory in and between the cases and controls could be missed by the traditional case-control approach. A proper model should be able to cover the aging process, instead of simplifying it into a couple of stages such as young controls and old cases (Pletcher & Stumpf, 2002). This is important not only in consideration of statistical power, but also as it is demanded by the complexity of the aging process. Because the force of natural selection diminishes after the age of reproduction, genes that become deleterious only at later ages can survive selection (Rose, 1991). As a consequence, antagonistic pleiotropic effects could play an important role in the biology of aging (Schachter *et al*. 1994; De Benedictis *et al*. 1998b). Another complexity of studying aging is that the genes involved can be highly polymorphic, for example, the *HLA-DRB1* (Ivanova *et al*. 1998), HUMTHO1.STR (in the tyrosine hydroxylase *TH* gene) (De Benedictis *et al*. 1998a; Tan *et al*. 2002a), *CYP2D6* (cytochrome p450 genes) (Bathum *et al*. 1998), 3′*APOB*-VNTR (De Benedictis *et al*. 1998b) and *APOE* (Kervinen *et al*. 1994; Gerdes *et al*. 2000; Schwanke *et al*. 2002; Slooter *et al*. 2001; Wang *et al*. 2001) polymorphisms. The polymorphic situation imposes another power problem onto the

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

case-control approach because of the large number of alleles, w, and consequently the large number of genotypes, $w(w+1)/2$, to be tested in a limited sample size. Recently developed statistical methods based on survival analysis have been applied to cope with some of the problems (Toupance *et al.* 1998; Yashin *et al.* 1999). Tan *et al.* (2001a) introduced a robust non-parametric approach that combines the individual genetic data with population survival information to infer the effects of genes. Although it is intuitive to deploy survival analysis in this case, an obvious limitation is the proportional hazard assumption which takes for granted a fixed or constant genetic effect on hazard of death over the ages. In this paper, we introduce the logistic regression model as an alternative to survival analysis to assess gene-longevity associations. The logistic regression model is popular in epidemiological studies. However, our approach here has the following three features: (a) although we are looking at the genetic effect on individual survival, instead of modelling survival as a function of genotype, we model genotype frequency as a function of age. This is because, even though we know that our samples are composed of young controls and old cases, we don't actually know their life spans without a tedious and expensive follow-up. The survival information we get is completely censored in a cross-sectional study. Modelling genotype frequency as a function of age avoids this problem. (b) as will be shown later, modelling genotype frequency as a function of age facilitates an elegant way to handle highly polymorphic genes by using genotype and/or allele-based parameterization. (c) modelling genotype frequency as a function of age offers the opportunity to model age-dependent pleiotropic effects. Given the complexity in the aging process and the character of the phenotype, an efficient statistical method is appealing. In this sense we hope our methods, characterized by the above three features, will be of help.

In this paper, we start with a description of the logistic regression models, including the motivations of model specification and parameterization as well as their implications. This is followed by an application of the models to an empirical data set collected in an association study on *HFE* (haemochromatosis) gene polymorphism and human longevity (Bathum *et al.* 2001). Finally, we briefly compare the models with other approaches and discuss the pros and cons of our methods in practical applications.

Methods

A Multinomial Logistic Regression Model for Polymorphic Genes

Suppose there is one polymorphic locus hosting w alleles. Then we could expect to observe $n = w(w+1)/2$ distinct genotypes at the locus. As mentioned before, because we don't observe an individual's life span, we turn our interest to the age pattern in genotype frequencies. If the gene is associated with survival, genotype frequencies will change with increasing age as a result of differential survivals. In this case, it is natural to define the nominal genotypes as polytomous responses and the continuous age as an explanatory variable. By assigning one genotype as baseline, we can model the baseline-category logits for $n-1$ genotypes as linear functions of age, x. If allele w is the wild type (or the most frequent) allele, we can assign homozygotes of the wild-type allele, A_wA_w, as the baseline genotype. With this parameterization, we obtain the multinomial logistic regression model with n polytomous responses as

$$\ln[\pi_{i,j}(x)/\pi_{w,w}(x)] = \begin{cases} \alpha_{i,j} + \beta_{i,j}x & i = j \\ \ln 2 + \alpha_{i,j} + \beta_{i,j}x & i < j \end{cases}$$

$$\alpha_{w,w} = 0, \qquad \beta_{w,w} = 0 \qquad i, j = 1, 2, \ldots\ldots w \quad (1)$$

Here $\pi_{i,j}(x)$ is frequency at age x for genotype A_iA_j and $\pi_{w,w}(x)$ is frequency at age x for homozygous genotype of the wild-type allele. Similar to any linear regression model, age related changes in genotype frequency are presented by the slope parameter. A $\beta_{i,j}$ significantly different from zero means that frequency of genotype A_iA_j goes up if $\beta_{i,j} > 0$ or down if $\beta_{i,j} < 0$. Because at any age x, $\sum_{i,j}\pi_{i,j}(x) = 1$, rearranging (1) we have

$$\pi_{i,j}(x) = \begin{cases} \exp(\alpha_{i,j} + \beta_{i,j}x)/[1 + H(x)] & i = j \\ 2\exp(\alpha_{i,j} + \beta_{i,j}x)/[1 + H(x)] & i < j \end{cases} \quad (2)$$

In (2) $H(x)$ is the sum of odds ratios, $\pi_{i,j}(x)/\pi_{w,w}(x)$ as expressed in (1), over all genotypes except the baseline genotype at age x. Based on the

multinomial distribution and the observed genotype frequency by age, a likelihood function can be constructed to estimate the corner and the slope parameters in (1).

It is interesting to see that, by setting age x to zero, we obtain genotype frequency at birth as

$$\pi_{i,j}(0) = \begin{cases} \exp(\alpha_{i,j})/[1 + H(0)] & i = j \\ 2\exp(\alpha_{i,j})/[1 + H(0)] & i < j \end{cases} \quad (3)$$

Here we see that the intercept, unimportant in most traditional regression analysis, becomes important in (1) as it represents genotype frequency at age zero. Moreover, (3) offers an opportunity to reduce drastically the number of intercepts in our model. Since it is sensible to assume Hardy-Weinberg equilibrium for gene frequency at birth if the population is homogeneous (a precondition for any association study), genotype frequency at birth can be predicted from allele frequency, as no differential survival yet exists. With this in mind, instead of each genotype we assign in (1) one intercept for each allele and let $\alpha_{i,j} = \alpha_i + \alpha_j$. Given Hardy-Weinberg equilibrium, we have

$$\begin{aligned} \pi_i(0) &= \exp(\alpha_i)/\sqrt{1 + H(0)}, \\ \pi_j(0) &= \exp(\alpha_j)/\sqrt{1 + H(0)} \\ \pi_{i,j}(0) &= 2\exp(\alpha_i + \alpha_j)/[1 + H(0)] \\ &= 2\pi_i(0)\pi_j(0) \qquad i < j \end{aligned} \quad (4)$$

In (4), $\pi_i(0)$ and $\pi_j(0)$ are allele frequencies at birth for alleles A_i and A_j. Replacing $\alpha_{i,j}$ in (1) with $\alpha_i + \alpha_j$, we only have $w - 1$ instead of $w(w + 1)/2 - 1$ intercepts to be estimated. Here we also have $\alpha_w = 0$ so that, similar to (1), $\alpha_{w,w} = \alpha_w + \alpha_w = 0$.

For a highly polymorphic gene, the number of genotypes to observe, n, increases rapidly with the number of alleles, w. Assigning each genotype one slope parameter will result in low power of the model especially when the sample size is small. One parsimonious way to circumvent the problem is to assume effects of the alleles, in term of odds ratio that are multiplicative so that the genotype-specific slope parameters can be decomposed into allele-specific slope parameters, $\beta_{i,j} = \beta_i + \beta_j$ with β_i for allele A_i and β_j for allele A_j. This further simplifies (1) into an extra parsimonious model with only allele-specific parameters. Similar to the intercepts, we have only $w - 1$ slope parameters. The total number of parameters to be estimated in (1) is now only $2(w - 1)$. For allele A_w, we have $\beta_w = 0$ so that for the baseline genotype A_wA_w, $\beta_{w,w} = \beta_w + \beta_w = 0$ which is consistent with (1). The allele based parameterization requires that, at any age x, the alleles in a given individual are independent, or similarly the Hardy-Weinberg equilibrium holds. In this situation, Sasieni (1997) showed that the statistic using allele-based parameterization is most powerful as long as the allele effect is multiplicative.

As an important phenomenon, gene-sex interaction or sex dependent effect has been reported in gene-longevity association studies (Ivanova *et al*. 1998; Tan *et al*. 2001b, 2002b). To capture the sex-dependent effects, we can specify in our model sex-specific slope parameters so that sex-specific effects, can be measured separately. The intercepts remain unchanged because according to the law of segregation, allele frequency for an autosomal gene should be equal at birth in both sexes. To model the sex-dependent effect, we extend and rewrite (1) as

$$\begin{aligned} &\ln[\pi_{i,j}(x)/\pi_{w,w}(x)] = \\ &\begin{cases} \alpha_{i,j} + \beta_{i,j,m}xU + \beta_{i,j,f}x(1 - U) & i = j \\ \ln 2 + \alpha_{i,j} + \beta_{i,j,m}xU + \beta_{i,j,f}x(1 - U) & i < j \end{cases} \\ &i, j = 1, 2, \ldots . w \quad \alpha_{w,w} = 0, \qquad \beta_{w,w,m} = 0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \beta_{w,w,f} = 0 \end{aligned} \quad (5)$$

Here U is an indicator for sex, $U = 1$ for males and 0 for females. Statistical tests can be applied to infer if the slopes are different in the two sexes. One can certainly fit (1) to males and females separately, but that doubles not only the number of slope parameters but also the number of intercepts.

Unlike in survival analysis, we don't estimate genotype specific hazard functions in our logistic regression model. However, as an equivalent measurement we can calculate genotype specific odds ratios to show how carrying one specific genotype or allele can increase or decrease the probability of surviving to a certain age. Studying odds ratios can help us to understand the implication of the slope parameters in our model. From (1), we first calculate the log odds ratio between ages x and $x - 1$ for genotype A_iA_j as

$$\begin{aligned}(OR_{ij,x/x-1}) &= \ln\{[\pi_{i,j}(x)/\pi_{w,w}(x)]/[\pi_{i,j}(x-1)/\\ &\quad \pi_{w,w}(x-1)]\}\\ &= \beta_{i,j}x - \beta_{i,j}(x-1)\\ &= \beta_{i,j}\end{aligned} \tag{6}$$

In (6), the intercepts are not shown because they cancel each other out. (6) shows that the slope parameter represents the risk of surviving one more year for genotype A_iA_j carriers, and such risk is independent of age which is equivalent to the proportional hazard model. In the multiplicative model, we have $OR_{ij,x/x-1} = \exp(\beta_i + \beta_j) = \exp(\beta_i)\exp(\beta_j)$. In this case, it is interesting to calculate the odds ratio for the heterozygous genotype of the wild type allele, say A_iA_w, for age x and $x-1$. Similar to (6) we have

$$\begin{aligned}\ln(OR_{iw,x/x-1}) &= \ln\{[\pi_{i,w}(x)/\pi_{w,w}(x)]/[\pi_{i,w}(x-1)/\\ &\quad \pi_{w,w}(x-1)]\}\\ &= \beta_i x - \beta_i(x-1)\\ &= \beta_i\end{aligned} \tag{7}$$

(7) means the risk of surviving one more year for heterozygous genotype A_iA_w is only determined by the effect of allele i, because we set the wild type allele as the reference or baseline allele with $\beta_w = 0$.

For a polymorphic locus, it is important to have an overall statistic to summarize the significance of the association with survival at the locus as has been alone in other association studies (Sham & Curtis, 1995). This can be done by the standard likelihood ratio test, by comparing the likelihood of the parameter estimates and that obtained by setting all the slope parameters to zero. A chi-squared statistic where the degrees of freedom equals the number of slope parameters (the number of alleles or genotypes minus one depending on the model fitted) can be calculated for statistical inference.

A Binomial Logistic Regression Model with Fractional Polynomials for Pleiotropic Genes

Schachter *et al.* (1994) reported a pleiotropic age-dependent effect on longevity by a variant of the *ACE* (angiotensin-converting enzyme) gene. Although the gene variant is associated with coronary heart disease, it is also more frequent in centenarians. In another study, a significant age-dependent frequency trajectory for one 3′*APOB*-VNTR polymorphism was observed (De Benedictis *et al.* 1998). According to the concept of antagonistic pleiotropy, genes that have deleterious effects at later ages can survive selection if they convey beneficial effects at early ages. Since the observed age-dependent gene frequency pattern can imply important biological mechanisms in the process of aging, modelling the antagonistic effects is appealing. In order to accommodate the age-dependent frequency pattern, we apply a logistic regression model with fractional polynomials (Royston & Altman, 1994). Suppose we observe an age-dependent frequency pattern for one genotype or allele at a locus; our task is to test and find out if the observed pattern is significantly different from random. In a binomial logistic regression model with fractional polynomials (Hosmer & Lemeshow, 2000), we have

$$\ln\{\pi(x)/[1-\pi(x)]\} = \alpha + \sum_{i=1}^{k} F_i(x)\beta_i \tag{8}$$

In (8) $F_i(x)$ is a power function for age x. The first term in $F_i(x)$ is x^{p_1}, and the rest are defined as

$$F_i(x) = \begin{cases} x^{p_i}, & p_i \neq p_{i-1} \\ F_{i-1}(x)\ln(x), & p_i = p_{i-1} \quad i = 2, \ldots, k \end{cases} \tag{9}$$

Although the power for $F_i(x)$, P_i, can be any number, Royston & Altman (1994) suggested restricting it within a set, $\varphi = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Here $p_i = 0$ means the log of age x. The number of covariates in (8), k, is not restricted, but in most cases it is adequate to set k to 2. With $k = 2$, the model with the largest likelihood is chosen as the best model from the 36 models, each fitted to one of the distinct pairs of power. The partial likelihood ratio test (Royston & Altman, 1994) can be applied to compare (a) the best $k = 1$ model with the linear model with 1 degree of freedom; (b) the best $k = 2$ model with the best $k = 1$ model with 2 degrees of freedom; (c) the best $k = 2$ model with the linear model but with 3 degrees of freedom.

From (8), we can calculate the log odds ratio between ages x and $x-1$ as

$$\begin{aligned}\ln(OR_{x/x-1}) &= \ln\{\pi(x)/[1-\pi(x)]\} - \ln\{\pi(x-1)/\\ &\quad [1-\pi(x-1)]\}\\ &= \sum_{i=1}^{k}[F_i(x) - F_i(x-1)]\beta_i\end{aligned} \tag{10}$$

(10) means that, unlike (6) and (7), the risk of surviving past age x is dependent on x. Later we show how the fractional polynomials model the age-dependent pattern of gene action during the aging process.

Applications

Located in the major histocompatibility complex region on chromosome 6, the *HFE* gene mutation C282Y (a cysteine to tyrosine mutation at amino acid 282) has been identified as the main genetic basis of hereditary haemochromatosis (HH). Recently, association of the *HFE* gene mutation with human longevity has been reported (Bathum *et al*. 2001; Lio *et al*. 2002). Two exons of this gene were screened in the study by Bathum *et al*. (2001), with exon 2 hosting the two mutations H63D and S65C, and exon 4 the C282Y mutation. Here we use an update of the data in Bathum *et al*. (2001) as an example to show how our models can be applied to handle polymorphic and pleiotropic situations. The updated data contains blood samples from 953 unrelated singletons from the middle aged Danish twin study, 400 singletons from the Longitudinal Study of Aging Danish Twins (LSADT), 601 individuals from the Danish 1905 cohort, and 183 centenarians from all over Denmark. Genotyping was carried out in two stages: in stage 1, both exons 2 and 4 were screened but only in 599 individuals (200 from LSADT, 200 from the middle aged Danish twin study, 199 from the Danish 1905 cohort). Bathum *et al*. (2001) found no significant association between *HFE* gene mutation and life span using the stage 1 data (Table 1). In the second stage, the genotyping continued but only in exon 4. Analysis of exon 4 data showed a significant age effect on C282Y heterozygous genotype frequency, which declines with increasing age (Bathum *et al*. 2001). However, the frequency rises in the centenarian group, suggesting an antagonistic effect of the gene at extreme ages.

Table 1 HFE genotype frequency in 599 individuals screened in exons 2 and 4

	Age group					
Genotype	45–54	55–64	65–74	75–84	85–94	Total
Wt/Wt	48	53	61	71	141	374
Wt/H63D	14	19	26	23	52	134
H63D/H63D	0	2	1	1	4	8
Wt/S65C	2	4	4	1	5	16
H63D/S65C	0	1	0	0	1	2
Wt/C282Y	16	7	8	9	20	60
H63D/C282Y	0	3	1	0	1	5
Total	80	89	101	105	224	599

Table 2 Parameter estimates by the multinomial logistic regression model on stage 1 data*

	Intercept			Slope		
Allele	Est.	Std	p-value	Est.	Std	p-value
H63D	−2.139	0.444	0.000	0.004	0.006	0.446
S65C	−3.123	1.147	0.007	−0.012	0.015	0.456
C282Y	−1.375	0.633	0.030	−0.017	0.009	0.047

*The wild type allele is assigned as the baseline allele

In the stage 1 data, 7 genotypes were observed in the 599 individuals from age 45 to 93. Bathum *et al*. (2001) observed no significant deviation from Hardy-Weinberg equilibrium in this data. After grouping the data by age and genotype, Table 1 becomes a sparse table with many small cell counts, which causes problems in conventional statistical analysis. By allele-based parameterization and assigning the wild type allele as the baseline allele, we applied our parsimonious model to the stage 1 data and estimated intercept and slope parameters for each of the 3 mutant alleles, H63D, S65C and C282Y (Table 2). Of all the slope parameters, only β_3 for C282Y showed a borderline significance. Slope parameters for H63D and S65C are not statistically different from zero, which means their frequencies are independent of age. Since β_3 is negative, the frequency of C282Y carriers tends to decrease with increasing age. In Figure 1 we plotted the observed and estimated frequencies for the C282Y allele by age. As expected, the C282Y allele frequency shows a decreasing pattern as age increases. However, after performing the likelihood ratio test, we found that a significant association with survival cannot be established for this locus using the stage 1 data ($\chi^2_{(3)} = 2.228$, $p = 0.527$).

In Table 3, we present the frequency of C282Y carriers by age in the exon 4 data, as calculated by Bathum *et al*. (2001) but with males and females combined. The frequency of C282Y carriers again shows a declining pattern with age, but with a considerable increase in the centenarians. Here we fit a binomial logistic regression model with fractional polynomials to see if there is

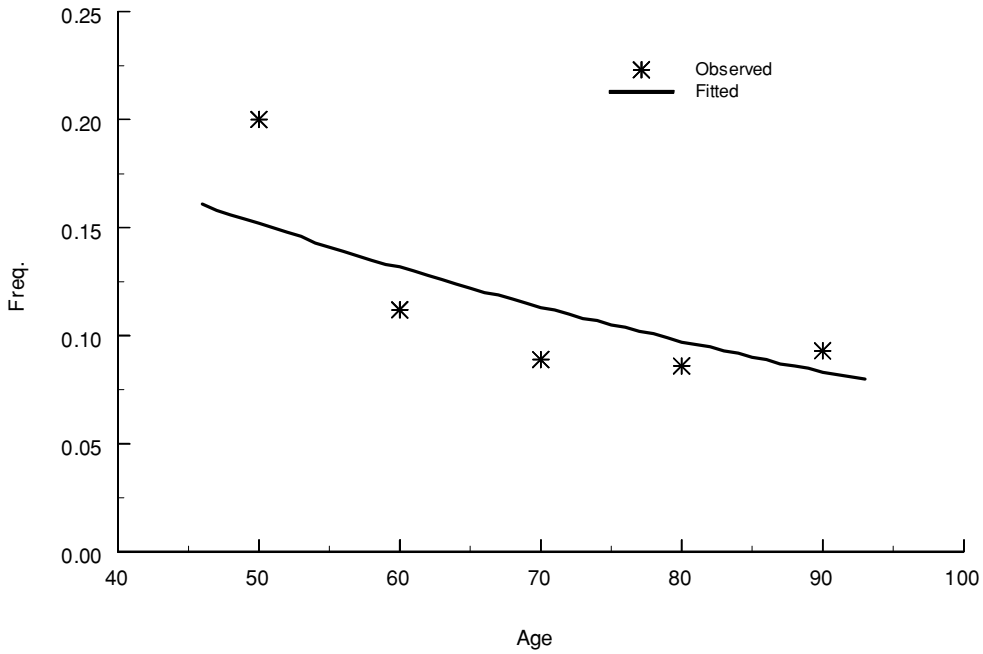

Figure 1 Observed (599 individuals) and fitted age patterns of C282Y allele frequency by the multinomial logistic regression model. As expected, both decrease as age advances.

Table 3 C282Y allele carriers frequency by age in the update

C282Y	Age					
Allele	45–54	55–64	65–74	75–84	85–94	100
(+)	93	33	23	20	53	25
(−)	571	181	192	223	565	158
Freq.	0.140	0.154	0.107	0.082	0.086	0.137

Table 4 Parameter estimates by the fractional polynomial model with $k = 2$

Parameter	Est.	Std	p-value	95% CI	
β_1	−0.018	0.007	0.009	−0.032	−0.005
β_2	0.007	0.003	0.011	0.002	0.013
α	−2.285	0.134	0.000	−2.549	−2.022

a significant age-dependent effect for this mutation. Table 4 shows the results for the best model among the 44 models, by setting k to 2 (8 models for $k = 1$ and 36 models for $k = 2$). In the best model, $F_1(x) = (x/10)^3 - 385.1$ and $F_2(x) = (x/10)^3 \ln(x/10) - 764.2$. All the coefficients are highly significant. In Figure 2, we show the observed and estimated frequencies of C282Y carriers. The fitted curve (dashed) from the fractional polynomial model indicates that the frequency of C282Y allele carriers decreases with age until around age 80, and then starts to increase at later ages. In order to make sure that the best fitted pattern is significantly different from the best $k = 1$ and the linear model, we compared the three models by the likelihood ratio test described above. The fits of both the $k = 2$ and $k = 1$ models are not significantly better than the linear model. The partial log likelihood ratio statistics for the best $k = 2$ model to the linear model is 4.727, which failed to reach the 5% significance level designated by $\chi^2_{(3)0.05} = 7.81$. In the linear model, we obtain the intercept $\alpha = -1.318$ ($Std = 0.259$, $p = 0.000$) and the slope parameter $\beta = -0.010$ ($Std = 0.004$, $p = 0.005$). In Figure 2, we also plot the fitted frequency for C282Y allele carriers (solid) by the linear model. Again, we see a constantly declining pattern. Our results from different models applied to the *HFE* gene data suggest that the mutant allele C282Y is a deleterious allele that increases carriers' risk of death. Although it tends to convey survival advantage at extreme ages, our data can't confirm yet that such a pleiotropic effect exists for this mutation. The above significant finding is also supported by

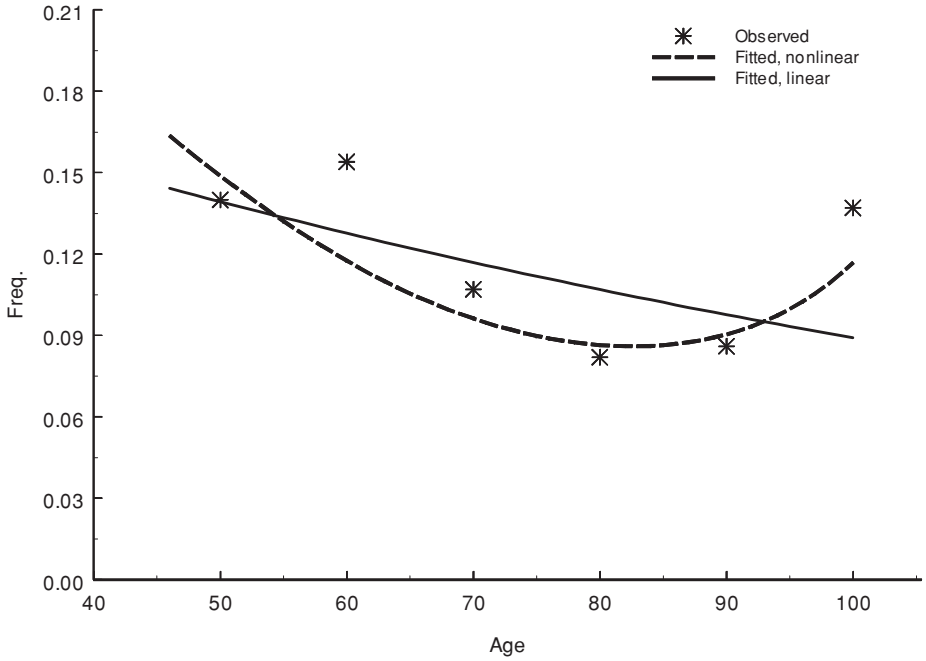

Figure 2 Observed and fitted age patterns of C282Y allele frequency in the updated data (2137 individuals). The linear model shows a constantly declining pattern but the nonlinear model gives a declining and then going up pattern. However, the nonlinear fit does not differ significantly from the linear one.

a χ^2 test in Table 3 ($\chi^2 = 15.97$, p-value $= 0.007$). However, such a simple test can in no way capture the nonlinear frequency pattern, and runs into problems as in Table 1.

Discussion

Modelling genotype frequency as a function of age, by fitting a logistic regression model with polytomous responses, offers an elegant way to handle polymorphic data in longevity studies. By allele-based parameterization, we show through application how our parsimonious model can make use of sparse data (Table 1). Our example showed that the model picked up important clues from a relatively small data set (599 individuals) that were missed in previous analyses. To some, the multiplicative assumption in the allele-based parsimonious model may appear risky. However we think, as a trade-off, it is useful when sample size is small and other models are inadequate. The multiplicative assumption as used by Risch & Merikangas (1996) is popular for summarising epistatic risks in mapping genes for complex diseases (Risch, 1990; Clayton & Jones, 1999; Wright *et al*. 1999; Koeleman *et al*. 2000). As a biological support, Dubois *et al*. (2002) reported multiplicative genetic effects by the prion protein gene polymorphism in scrapie disease susceptibility. Nevertheless, we think the genotype-based analysis should be carried out whenever feasible, because such analysis can help to establish whether the allele effect is recessive, dominant, or codominant (Sasieni, 1997).

Toupance *et al*. (1998) proposed a parametric survival model for analyzing antagonistic pleiotropic genes. A Gompertz-Makeham model was used to model genotype-specific survival functions by incorporating population survival. We think that there are several difficulties with their approach. First of all, it is not a good idea to impose a specific form of survival distribution on a subgroup in a limited sample, because both the choice of a parametric form and the sample size limitation will result in considerable error in estimating the genotype-specific survival distributions. Consequently, the age-dependent frequency pattern resulting from differential survival will be unreliable. This is more

serious at extreme ages when sample collection becomes very difficult. In addition, at old ages, the validity of the Gompertz-Makeham model becomes more questionable. Any pattern based on these uncertainties can be doubious if not arbitrary. Antagonistic pleiotropic effects can also be modelled by applying frailty modelling (Vaupel & Yashin, 1985; Yashin *et al.* 1999). However, estimation of the heterogeneity parameter is problematic in small scale studies. In this case, it is our experience that a very small change in the heterogeneity parameter can lead to a big difference in the fitted frequency pattern. After all, we think that the application of a parametric survival model and the interpretation of the results should be carried out with caution. By simply testing if an age-dependent gene frequency pattern exists using a logistic regression model, all these complications can be avoided. Whenever a significant trend is established, we leave room for biological explanations.

The multinomial logistic regression model introduced in this paper also facilitates a measurement of an overall association with survival at a given locus by the standard likelihood ratio test. This is important when investigating polymorphic loci because the high number of alleles or genotypes creates multiple testing problems which have been ignored by the recent approaches based on survival analysis (Yashin *et al.* 1999; Tan *et al.* 2001a). Adjusting for multiple comparisons is problematic, as the alleles residing at a given locus are not independent. In our analysis, a conclusion based on a significant association for an allele can only be made when the overall statistic is significant.

Although our approach avoids the complexity of survival modelling, there are also problems connected with the simplification. As mortality is low at younger ages, one cannot expect a rapid change of allele or genotype frequency in the population. Because the logistic regression model only models the frequency trend observed in the sample, which may not be representative of the population in general, a steep slope might be fitted to the early ages which does not make much sense biologically. Since our goal is to study aging and longevity this may not be a problem. In any case, efforts should be taken to make sure that the samples are representative and the age structure of the sample is reasonable. Moreover, similar to other association studies, efforts are also needed to ensure ethnic homogeneity in sample collection so that spurious results due to population stratification can be avoided.

As a complex trait, longevity is likely to involve the interplay of many loci (De Benedictis *et al.* 2001). Multilocus approaches have been proved to have more power in linkage disequilibrium studies using case-control and family-based control designs (Akey *et al.* 2001; Risch, 2001; Fallin *et al.* 2001). In the multilocus approach, haplotype-based analyses are applied to detect unique regions that harbour disease genes. Though in longevity studies missing parental genotypic information prevents us from determining phase and assigning haplotypes, approaches based on the EM (expectation-maximization) algorithm can be applied to estimate haplotype frequency using unrelated individuals (Xie & Ott, 1993; Zhao & Sham, 2002). Estimating haplotype frequencies is beyond the scope of this paper. However we think, by incorporating haplotype construction, future work should be able to expand our model to multilocus genotype data.

Conclusions

We have shown through an example that, as an alternative to other approaches, the logistic regression model can be used to measure gene-longevity associations. Modelling genotype frequency as a function of age by fitting logistic regression models: (a) offers us a good way to model genetic association with longevity at polymorphic loci; (b) enables us to model age-specific or pleiotropic effects, which the relative risk or proportional hazard models cannot accommodate; and most importantly (c) measures genotype or allele effects. In addition, as a popular method in epidemiology, most statistical packages offer procedures for fitting logistic regression models. Hence, useful information can be obtained through easily performable data analyses. We believe the methods presented in this paper will serve as useful tools when looking for important genetic variations that modulate human life span.

Acknowledgement

This research was partly supported by the US National Institute on Aging research grant NIA-P01-AG08761. The authors are grateful to the anonymous referees whose comments

helped to improve the text. Dr. Qihua Tan wants to extend his gratitude to Kirsten Pagh for her help in preparing the paper and to the constant support by the Max-Planck Institute of Demographic Research in Rostock, Germany.

References

Akey, J., Jin, L. & Xiong, M. (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* **9**, 91–300.

Bathum, L., Andersen-Ranberg, K., Boldsen, J., Brosen, K. & Jeune, B. (1998) Genotypes for the cytochrome P450 enzymes CYP2D6 and CYP2C19 in human longevity. Role of CYP2D6 and CYP2C19 in longevity. *Eur J Clin Pharmacol* **54**, 427–30.

Bathum, L., Christiansen, L., Nybo, H., Ranberg, K.A., Gaist, D., Jeune, B., Petersen, N.E., Vaupel, J. & Christensen, K. (2001) Association of mutations in the hemochromatosis gene with shorter life expectancy. *Arch Intern Med* **61**, 2441–4.

Clayton, D. & Jones, H. (1999) Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum* **65**, 1161–9.

De Benedictis, G., Falcone, E., Rose, G., Ruffolo, R., Spadafora, P., Baggio, G., Bertolini, S., Mari, D., Mattace, R., Monti, D., Morellini, M., Sansoni, P. & Franceschi, C. (1997) DNA multiallelic systems reveal gene/longevity associations not detected by diallelic systems: The APOB locus. *Hum Genet* **99**, 312–318.

De Benedictis, G., Carotenuto, L., Carrieri, G., De Luca, M., Falcone, E., Rose, G., Cavalcanti, S., Corsonello, F., Feraco, E., Baggio, G., Bertolini, S., Mari, D., Mattace, R., Yashin, A.I., Bonafe, M. & Franceschi, C. (1998a) Gene/longevity association studies at four autosomal loci (REN, THO, PARP, SOD2). *Eur J Hum Genet* **6**, 534–541.

De Benedictis, G., Carotenuto, L., Carrieri, G., De Luca, M., Falcone, E., Rose, G., Yashin, A.I., Bonafe, M. & Franceschi, C. (1998b) Age-related changes of the 3′APOB-VNTR genotype pool in aging cohorts. *Ann Hum Genet* **62**, 115–122.

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* **1**, 909–920.

Dubois, M.A., Sabatier, P., Durand, B., Calavas, D., Ducrot, C. & Chalvet-Monfray, K. (2002) Multiplicative genetic effects in scrapie disease susceptibility. *C R Biol* **325**, 565–570.

Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohen, D. & Schork, N.J. (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res* **11**, 143–51.

Gerdes, L.U., Jeune, B., Ranberg, K.A., Nybo, H. & Vaupel, J.W. (2000) Estimation of apolipoprotein E genotype-specific relative mortality risks from the distribution of genotypes in centenarians and middle-aged men: apolipoprotein E gene is a "frailty gene," not a "longevity gene." *Genet Epidemiol* **19**, 202–10.

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

Ivanova, R., Henon, N., Lepage, V., Charron, D., Vicaut, E. & Schachter, F. (1998) HLA-DR alleles display sex-dependent effects on survival and discriminate between individual and familial longevity. *Hum Mol Genet* **7**, 187–194.

Kervinen, K., Savolainen, M.J., Salokannel, J., Hynninen, A., Heikkinen, J., Ehnholm, C., Koistinen, M.J. & Kesaniemi, Y.A. (1994) Apolipoprotein E and B polymorphisms–longevity factors assessed in nonagenarians. *Atherosclerosis* **105**, 89–95.

Koeleman, B.P., Dudbridge, F., Cordell, H.J. & Todd, J.A. (2000) Adaptation of the extended transmission/disequilibrium test to distinguish disease associations of multiple loci: the Conditional Extended Transmission/Disequilibrium Test. *Ann Hum Genet* **64**, 207–13.

Lio, D., Balistreri, C.R., Colonna-Romano, G., Motta, M., Franceschi, C., Malaguarnera, M., Candore, G. & Caruso, C. (2002) Association between the MHC class I gene HFE polymorphisms and longevity: a study in Sicilian population. *Genes Immun* **3**, 20–24.

Pletcher, S.D. & Stumpf, M.P. (2002) Population genomics: ageing by association. *Curr Biol* **12**, 328–30.

Risch, N. (1990) Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* **46**, 222–8.

Risch, N. & Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Risch, N. (2001) Implications of multilocus inheritance for gene-disease association studies. *Theor Popul Biol* **60**, 215–20.

Rose, M.R. (1991) *Evolutionary biology of aging*, London: Oxford University Press.

Royston, P. & ltman, D. (1994) Regression using fractional polynomials of continuous: Parsimonious parametric modeling. *Applied Statistics* **43**, 429–467.

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

Schachter, F., Faure-Delaneff, L., Guenot, F., Rouger, H., Froguel, P., Lesueur-Ginot, L. & Cohen, D. (1994) Genetic associations with human longevity at the APOE and ACE loci. *Nature Genetics* **6**, 29–32.

Schwanke, C.H., da Cruz, I.B., Leal, N.F., Scheibe, R., Moriguchi, Y. & Moriguchi, E.H. (2002) Analysis of the association between apolipoprotein E polymorphism and

cardiovascular risk factors in an elderly population with longevity. *Arq Bras Cardiol* **78**, 561–79.

Sham, P.C. & Curtis, D. (1995) Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann Hum Genet* **59**, 97–105.

Slooter, A.J., Cruts, M., Van Broeckhoven, C., Hofman, A. & van Duijin, C.M. (2001) Apolipoprotein E and longevity: the Rotterdam Study. *J Am Geriatr Soc* **49**, 1258–9.

Tan, Q., De Benedictis, G., Yashin, A.I., Bonafe, M., De Luca, M., Valensin, S., Vaupel, J.W. & Franceschi, C. (2001a) Measuring the genetic influence in modulating human life span: Gene-environment and gene-sex interactions. *Biogerontology* **2**, 141–153.

Tan, Q., Yashin, A.I., Bladbjerg, E.M., de Maat, M., Andersen-Ranberg, K., Jeune, B., Christensen, K. & Vaupel, J.W. (2001b) Variations of cardiovascular disease associated genes exhibit sex-dependent influence on human longevity, *Exp Gerontol* **36**, 1303–1315.

Tan, Q., Bellizzi, D., Rose, G., Garasto, S., Franceschi, C., Kruse, T., Vaupel, J., De Benedictis, G. & Yashin, A.I. (2002a) The influences on human longevity by HUMTHO1.STR polymorphism (Tyrosine Hydroxylase gene). A relative risk approach. *Mech Ageing Dev* **123**, 1403–1410.

Tan, Q., Yashin, A.I., Bladbjerg, E.M., de Maat, P.M., Andersen-Ranberg, K., Jeune, B., Christensen, K. & Vaupel, J.W. (2002b) A case-only approach for assessing gene by sex interaction in human longevity. *J Gerontol* **57A**, B129–B133.

Toupance, B., Godelle, B., Gouyon, P.H. & Schachter, F. (1998) A model for antagonistic pleiotropic gene action for mortality and advanced age. *Am J Hum Genet* **62**, 1525–34.

Vaupel, J.W. & Yashin, A.I. (1985) Heterogeneity's ruses: some surpring effects of selection on population dynamics. *Am Stat* **39**, 176–185.

Wang, X., Wang, G., Yang, C. & Li, X. (2001) Apolipoprotein E gene polymorphism and its association with human longevity in the Uygur nationality in Xinjiang. *Chin Med J (Engl)* **114**, 817–20.

Wright, A.F., Carothers, A.D. & Pirastu, M. (1999) Population choice in mapping genes for complex diseases. *Nature Genetics* **23**, 397–404.

Xie, X. & Ott, J. (1993) Testing linkage disequilibrium between a disease gene and marker loci. *Am J Hum Genet* **53**, 1107.

Yashin, A.I., De Benedictis, G., Vaupel, J.W., Tan, Q., Andreev, K.F., Iachine, I.A., Bonafe, M., De Luca, M., Valensin, S., Carotenuto, L. & Franceschi, C. (1999) Genes, demography, and life span: 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: 21 January 2003
Accepted: 23 April 2003