*Biogerontology* **5:** 89–97, 2004.



*Research article*

# Assessing genetic association with human survival at multi-allelic loci

Qihua Tan[1,*], G. De Benedictis[2], A.I. Yashin[3], L. Bathum[1], L. Christiansen[1], J. Dahlgaard[1], N. Frizner[1], W. Vach[4], J.W. Vaupel[3], K. Christensen[5] & T.A. Kruse[1]
[1]*Department of Clinical Biochemistry and Genetics, KKA, Odense University Hospital, Odense, Denmark;*
[2]*Cell Biology Department, University of Calabria, Rende, Italy;*
[3]*Max-Planck Institute for Demographic Research, Rostock, Germany;*
[4]*Department of Statistics and Demography, University of Southern Denmark;*
[5]*Epidemiology, Institute of Public Health, and Ageing Research Center, University of Southern Denmark-Odense University, Denmark*
**Author for correspondence (e-mail: qihua.tan@ouh.fyns-amt.dk; fax: +45-6541-1911)*

Received 6 August 2003; accepted in revised form 15 September 2003

*Key words:* association study, gene, Hardy–Weinberg equilibrium, longevity, polymorphism

**Abstract**

Genetic variation plays an important role in natural selection and population evolution. However, it also presents geneticists interested in aging research with problems in data analysis because of the large number of alleles and their various modes of action. Recently, a new statistical method based on survival analysis (the relative risk model or the RR model) has been introduced to assess gene–longevity associations [Yashin et al. (1999) Am J Hum Genet 65: 1178–1193] which outperforms the traditional gene frequency method. Here we extend the model to deal with polymorphic genes or gene markers. Assuming the Hardy–Weinberg equilibrium at birth, we first introduce an allele-based parameterization on gene frequency which helps to cut down the number of frequency parameters to be estimated. We then propose both the genotype and allele-based parameterizations on risk parameters to estimate genotype and allelic relative risks (the GRR and ARR models). While the GRR model allows us to investigate whether the alleles are recessive, dominant or codominant, the ARR model further minimizes the number of parameters to be estimated. As an example, we apply the methods to empirical data on *Renin* gene polymorphism and longevity. We show that our models can serve as useful tools in searching for important genetic variations implicated in human aging and longevity.

## Introduction

Maintenance of genetic variation and genetic polymorphism is one of the key issues in evolutionary biology. This is because, apart from the contribution of newly arisen mutations, genetic variation *per se* conveys heterozygous advantages and makes natural selection possible. It has been shown that the genetically polymorphic population is advantageous for survival in a changing environment (Lee et al. 1998), and that increased homozygosity may affect individual and population negatively in stressful and fluctuating environments by disturbing gene expression (Kristensen et al. 2002), reducing individual fitness and survival and increasing the probability of population extinction (Charlesworth and Charlesworth 1987; Dahlgaard and Hoffman 2000; Bijlsma et al. 2000). Genetic polymorphism thus plays an important role both in maintaining the long-term evolutionary potential of populations and in preserving individual fitness and survival.

Genetic variations at highly polymorphic loci have been associated with human longevity, for example the HLA-DRB1 (Ivanova et al. 1998), HUMTHO1.STR (tyrosine hydroxylase gene) (De Benedictis et al. 1998a; Tan et al. 2002), CYP2D6 (cytochrome p450 genes) (Bathum et al. 1998),

3′APOB-VNTR (De Benedictis et al. 1997, 1998b) and APOE (Kervinen et al. 1994; Gerdes et al. 2000; Slooter et al. 2001; Wang et al. 2001; Schwanke et al. 2002) polymorphisms. These observations, together with the presence of ample genetic variations for human longevity, underline the importance of monitoring the polymorphic genetic influences in human survival. However, highly polymorphic genes present statistical problems in association studies due to the large number of alleles, $n$, and thus the large number of genotypes, $n(n + 1)/2$, to be tested within a limited sample size for which traditional statistical methods run into power problems.

Although the family based linkage analysis takes advantage of high polymorphism, unfortunately it is not applicable in longevity studies because parental genotypes for the long-lived individuals are usually missing. The candidate gene approach in the framework of association study is more feasible (Daly 2003). The statistical methods for analyzing gene–longevity association data have been dominated by the so-called gene frequency approach (Yashin et al. 1999), which simply compares allele frequency differences between cases (centenarians) and controls (young individuals) (Kervinen et al. 1994; Schachter et al. 1994; Bathum et al. 1998, 2001; De Benedictis et al. 1998a, b; Ivanova et al. 1998). A new method that combines individual genetic and population survival information to estimate the risk for gene carriers has been proposed (Yashin et al. 1999; Tan et al. 2001a). Tan et al. (2002) extended the method to estimate genotype relative risk (GRR), which can help to infer if the allele is recessive, dominant or codominant. However, all the survival models deal with genetic polymorphism in such a manner that they test either each allele or genotype separately against the rest (Toupance et al. 1998; Yashin et al. 1998, 1999, 2000; Gerdes et al. 2000; Tan et al. 2001a; Varcasia et al. 2001). Such a practice has the following problems. First, for different alleles or genotypes tested, the reference or baseline effect (the sum of the rest) differs. These effects also overlap each other. This means that the multiple tests carried out are not independent. In this case, correcting the significance level for multiple testing using the Bonferroni adjustment is problematic. Second, because the baseline effect is a mixture of all the others, it could happen that the significant result for the allele tested is simply because there is a risky allele included in the reference group. Third, because the tests are done on each allele or genotype separately, there is no statistic to show an overall significance level on survival for the gene. This, together with the first problem, renders the model impossible to summarize the association at the locus as a whole. To counteract all the problems, we propose a generic but parsimonious model that estimates genotype or allelic relative risks (GRR or ARR) similar to that in gene-disease association studies (Risch and Merikangas 1996; Risch 2000). The strategic ways of parameterization allow us (a) to estimate allele frequency at birth based on the Hardy–Weinberg assumption; (b) to investigate the mode of the gene action (recessive, dominant or codominant) by estimating GRR and (c) to minimize the number of parameters in the model by estimating ARR based on a multiplicative assumption of the allelic effects.

We start with a short description of the RR model for gene–longevity association. Then, we introduce our parsimonious models for polymorphic genes. A simulation follows with the aim of investigating whether the model can retrieve the parameters used in generating the data and comparing the performances of different models. As an example, we apply the methods to evaluate the influence on human longevity by *Renin* gene to show how our model can handle gene polymorphism while capturing important information that can be used to make inferences. We end the presentation with a brief discussion on the significance and implication of our approach in longevity studies.

## The RR model

In the RR model (Yashin et al. 1999; Tan et al. 2001a), we define the RR $r$, for carrying one observed gene allele or genotype as the ratio of hazard of death for carriers, $\mu(x)$, to that for the non-carriers or the baseline hazard, $\mu_0(x)$. Then, in a proportional hazard model, $\mu(x) = r\mu_0(x)$. The corresponding survival function for the carriers is

$$s(x) = e^{-\int_0^x \mu(t)\,dt} = e^{-\int_0^x r\mu_0(t)\,dt} = e^{-r\int_0^x \mu_0(t)\,dt} = e^{-rH_0(x)} = s_0(x)^r. \quad (1)$$

Here, $s_0(x)$ is the survival distribution corresponding to the baseline hazard function, $H_0(x)$ is the cumulative baseline hazard at age $x$. Although $r$ can take any value greater than zero, an allele with $r$ larger than 1 increases the hazard of death, while a gene allele with $r$ smaller than 1 reduces it. Since all individuals can be grouped into carriers and non-carriers of an allele or genotype, one can introduce the simple two-point distribution for the frequency of allele or genotype carriers (Vaupel and Yashin 1985). The average survival at age $x$ for a mixed population consisting of both carriers and non-carriers is

$$\bar{s}(x) = ps_0(x)^r + (1-p)s_0(x). \tag{2}$$

Here, $p$ is the proportion of carriers at birth, $\bar{s}(x)$ is the survival rate at age $x$ obtainable from population statistics. From (2), frequency for carriers at age $x$ is $p(x) = \frac{ps_0(x)^r}{\bar{s}(x)}$. Based on the frequency distribution, a likelihood function can be constructed as

$$L \propto \prod_x p(x)^{n(x)}(1-p(x))^{N(x)-n(x)}, \tag{3}$$

where $n(x)$ is the number of carriers at age $x$, $N(x)$ is the total number of participants at age $x$. The model estimates both RR $r$ and frequency at birth $p$ for allele carriers. Tan et al. (2001a) used a robust EM algorithm to estimate a non-parametric baseline hazard function so that proportional hazard becomes the only assumption in the model.

## Parsimonious models for polymorphic genes

In the case of diallelic loci or SNPs data, the model described above circumvents problems mentioned in the introduction because of the symmetric nature of the tests. However, in the case of a polymorphic gene, all the problems come up. A generic but parsimonious model for dealing with the polymorphic situation is appealing.

Suppose there is one locus hosting $n$ alleles (alleles $A_1, A_2, \ldots, A_n$). Then, one can expect to observe $n(n+1)/2$ distinct genotypes at the locus. For each genotype $A_iA_j$, assume its frequency at birth is $P_{i,j}$ and risk on the baseline hazard is $R_{i,j}$. Then, similar to (2), mean survival in the population is a weighted average survival for individuals carrying the $n(n+1)/2$ genotypes.

$$\begin{aligned}\bar{s}(x) &= \sum_{i,j} P_{i,j}s_{i,j}(x) \\ &= \sum_{i,j} P_{i,j}s_0(x)^{R_{i,j}}, \quad i,j = 1,2,\ldots,n,\ j \geq i, \\ &\quad \sum_{i,j} P_{i,j} = 1.\end{aligned} \tag{4}$$

Similar to the RR model, genotype frequency at age $x$ for genotype $A_iA_j$ can be calculated as

$$P_{i,j}(x) = P_{i,j}s_{i,j}(x)/\bar{s}(x) = P_{i,j}s_0(x)^{R_{i,j}}/\bar{s}(x). \tag{5}$$

But instead of using the binomial function as in (3), we build up the likelihood function on a multinomial distribution

$$L \propto \prod_x \prod_{i,j} P_{i,j}(x)^{n_{i,j}(x)}, \quad j \geq i,\ \sum_{i,j} P_{i,j}(x) = 1, \tag{6}$$

where $n_{i,j}(x)$ is the number of individuals with genotype $A_iA_j$ at age $x$. In order to make the parameters identifiable, we assign homozygous genotype of the most frequent allele (or the wild-type allele) as the reference genotype by setting its risk to 1. Frequency of the baseline genotype can be calculated as one minus the sum of the frequencies of all the other genotypes. In (4), we use a numerical procedure to estimate a distribution-free baseline survival function, $s_0(x)$. Parameters are estimated using a two-step procedure proposed by Tan et al. (2001a). In this model, there are $n(n+1)-2$ parameters, half are frequency and half are risk parameters. For a highly polymorphic gene, there will be a large number of genotypes and thus a large number of parameters to be estimated. Parameter estimation becomes infeasible when only limited data are available.

We designate frequency at birth for any allele $A_i$ with $p_i$. Because at birth, gene frequency is not altered by genotype differential survival, it is natural to assume that the Hardy–Weinberg law holds. Then, for genotype $A_iA_j$ formed by alleles $A_i$ and $A_j$, its frequency at birth is

$$\begin{aligned}&P_{i,j} = 2p_ip_j, \quad j > i, \\ &P_{i,j} = p_i^2, \quad i = j, \quad i,j = 1,2,\ldots,n.\end{aligned} \tag{7}$$

Replacing the frequency parameters in (4) by (7) largely cuts down the number of frequency parameters. As long as the Hardy–Weinberg equilibrium holds at birth, such a way of parameterization should not affect the performance of our model at all. As a constraint on the allele frequencies, frequency for the wild-type allele is calculated as one minus the sum of the frequencies of the other alleles. In this way, we only have $n-1$ frequency parameters in our model instead of $n(n+1)/2-1$. Since we have no assumption on the mode of gene action, inspecting the GRR parameters can help us to figure out whether the alleles are recessive, dominant or codominant.

While it is advantageous to estimate GRR, one big drawback is it still has too many risk parameters. When analyzing a highly polymorphic gene on a limited sample, one immediately encounters the power problem. In this situation, an alternative is to apply an even more parsimonious approach using allele-based parameterization on the risk parameters. Assuming effects of the alleles are multiplicative, we can estimate ARR instead of GRR. If risks for alleles $A_i$ and $A_j$ are $r_i$ and $r_j$, then the risk for genotype $A_iA_j$ is

$$R_{i,j} = r_i r_j, \quad j \geq i, \;\; i,j = 1,2,\ldots,n. \tag{8}$$

To be consistent, we assign the wild-type allele as the reference allele with risk 1 such that the risk for the homozygous genotype of the wild-type allele also becomes 1. By estimating ARR, the number of parameters is drastically cut down to $2(n-1)$ with half for the frequencies and half for the risks.

Similar to the other association studies on highly polymorphic loci (Sham and Curtis 1995), it is necessary to have an overall statistic to summarize the significance of the association with survival for the gene or locus under investigation. In our models, this can be done easily by the standard likelihood ratio test. Since our interest is in testing whether the risk parameters are neutral, we can calculate $\chi^2_{(l)} = -2[L(1) - L(R)]$, a $\chi^2$ statistic with $l$ degrees of freedom to measure the overall association at the locus. Here $L(R)$ is the log maximum likelihood for the risk parameter estimates $R$, and $L(1)$ is the log of the maximum likelihood when setting $R$ to 1. $R$ can either be GRR or ARR depending on the model one has fitted. However, the degree of freedom $l$, which is the number of risk parameters, differs largely between GRR and ARR models. This shows that while we gain power by fitting the ARR model, we have to rely on the multiplicative assumption, which may not always hold.

For a well defined candidate gene, the above method offers a nice way to evaluate the association with survival at the given polymorphic locus and avoids multiple testing. When data on multiple loci are collected, the Boneferroni adjustment can be applied if the genes are independent. Otherwise, empirical locus-wise $P$-value can be obtained by a simulation study but assuming no association.

## Simulation

The aims of conducting a simulation study are (a) to check whether our models can capture the parameters used in generating the genetic association with survival, (b) to compare performances of the different models. In the simulation, we assume multiplicity of the allelic effects (Risch and Merikangas 1996; Wright et al. 1999) at a polymorphic locus with five alleles. Allele frequencies for the five alleles are {0.1, 0.1, 0.1, 0.1, 0.6} and RRs are {1, 0.75, 1, 1, 1}. Among all the five alleles, the second allele has a beneficial effect that reduces the hazard of death by 0.75. Population survival function is borrowed from the 1994 Italian life table (Annuario Statistico Italiano 1997). With the given parameters and the population survival function, we solve (4) to obtain a nonparametric baseline survival function (Tan et al. 2001a). With the above ARR and assuming multiplicative effect of the alleles, the baseline survival is used to calculate genotype specific survival functions. The so-obtained genotype specific survival functions are then used to generate life spans for individuals with corresponding genotypes. Individual genotypes are randomly assigned using the above allele frequency parameters and assuming the Hardy–Weinberg equilibrium at birth. We simulated 1000 samples containing subjects aged from 51 to 100 with 20 individuals for each age. In different empirical studies, age structure of the collected sample may differ due to different reasons. Effect of the sample age structure on parameter estimation has been

investigated by Tan (2000) which showed no strong influence except when extreme age structure is deployed.

After simulating the data, we applied the GRR and ARR models to retrieve the frequency and risk parameters. In Table 1, we show the estimated GRRs for the homozygous and heterozygous genotypes of allele 2 together with their 95% ranges from the 1000 replicates. In the estimation, we choose the homozygous genotype of the most frequent allele, allele 5 as the reference by setting its risk to 1. All the medians of the estimated GRRs for allele 2 heterozygous genotypes are close to 0.75 but with different 95% ranges. The 95% range for 2/5 genotype is the narrowest due to its higher genotype frequency. We also see that the median of GRR for 2/2 genotype is exactly $0.75^2 = 0.563$ because of the multiplicative assumption. As expected, GRRs for non-carriers of allele 2 are all close to 1 (not shown in Table 1). Medians for allele frequency estimates by the GRR model are 0.099 for allele 1 (95% range 0.087–0.114), 0.100 for allele 2 (95% range 0.087–0.113), 0.100 for allele 3 (95% range 0.088–0.115), 0.100 for allele 4 (95% range 0.087–0.114) and 0.601 for allele 5 (95% range 0.544–0.651), which means that all are well captured.

We next apply the ARR model to estimate frequency and risk parameters for each allele. Since we assume multiplicity in generating the data, the ARR model should be the best choice to conduct the data analysis. This is true as can be seen in Table 2; the medians for both the estimated frequency and risk parameters are close to their true values. The 95% range for ARR of allele 2 is narrower than that for the GRR of 2/5 genotype in Table 1.

*Table 1.* Medians of the estimated GRRs for allele 2 genotypes (1000 replicates).

| Genotype[a] | GRR | 95% range |
|---|---|---|
| 2/1 | 0.757 | 0.587–1.009 |
| 2/2 | 0.563 | 0.404–0.820 |
| 2/3 | 0.756 | 0.587–0.962 |
| 2/4 | 0.755 | 0.590–1.003 |
| 2/5 | 0.750 | 0.661–0.849 |

[a]Genotype 5/5 is baseline genotype.

*Table 2.* Medians of the estimated allele frequencies and ARRs (1000 replicates).

| Allele | Frequency | | ARR | |
|---|---|---|---|---|
| | Median | 95% range | Median | 95% range |
| 1 | 0.099 | 0.087–0.114 | 0.997 | 0.898–1.127 |
| 2 | 0.100 | 0.087–0.113 | 0.751 | 0.673–0.825 |
| 3 | 0.100 | 0.087–0.115 | 1.001 | 0.904–1.114 |
| 4 | 0.100 | 0.087–0.113 | 1.001 | 0.899–1.121 |
| 5[a] | 0.601 | 0.545–0.652 | – | – |

[a]Allele 5 is baseline allele.

## Application

### *Renin gene polymorphism and life span*

Started in 1995, the Italian multi-centric longevity study has collected genotype data on both young individuals as controls and cases of advanced ages. Among the genes tested, several have been reported as being associated with human survival (De Benedictis et al. 1997, 1998a, b, 2001; Yashin et al. 1998, 1999; Tan et al. 2001a). One important gene typed in the study was the *Renin* gene which codes for renin, a rate limiting factor in angiotensin II synthesis. By comparing genotype frequencies in controls and in cases, no association with survival was found in an early analysis (De Benedictis et al. 1998a). Here we apply our new approach to see if variations of the gene can influence individual survival. The sample for this gene consists of 375 subjects (Table 3), 157 centenarians (38 males and 119 females) and 218 young controls (88 males and 130 females). The mean age in the control group is 37 years with a range of 13–

*Table 3.* Observed genotype frequency in cases and controls for *Renin* gene.

| Genotype | Young control | | Centenarian | | Sum |
|---|---|---|---|---|---|
| | Count | Prop. | Count | Prop. | |
| 7/8 | 1 | 0.005 | 0 | 0.000 | 1 |
| 8/8 | 111 | 0.509 | 99 | 0.631 | 210 |
| 8/10 | 39 | 0.179 | 26 | 0.166 | 65 |
| 8/11 | 56 | 0.257 | 23 | 0.146 | 79 |
| 8/12 | 1 | 0.005 | 2 | 0.013 | 3 |
| 10/11 | 2 | 0.009 | 2 | 0.013 | 4 |
| 10/12 | 2 | 0.009 | 1 | 0.006 | 3 |
| 11/11 | 6 | 0.027 | 4 | 0.025 | 10 |
| Total | 218 | 1.000 | 157 | 1.000 | 375 |

71 years. The large age range in the control group could have helped to reduce the genotype frequency difference in the two groups. This is no longer a problem for our models because, by fitting the survival model to the data, we make use of each individual's exact age at the time when the blood sample was drawn.

Five polymorphic alleles of the *Renin* gene (alleles 7, 8, 10, 11, 12) were detected at the locus designated by the number of short tandem repeats (STR). A total of eight genotypes were observed from the data. After categorizing the individuals by their genotypes, Table 3 becomes a sparse table with many cell counts less than 5. In this case, traditional $\chi^2$ test based on the asymptotic sampling distribution of the test statistics is no longer reliable. However, our survival model makes use of such data because as in (6) our model makes inferences on the likelihood of the data. Since there is only one allele 7 carrier found in the data, we combined the allele with its adjacent allele, which is also the most frequent allele, allele 8, to form the combined allele 8′. To analyze the data, we use population survival function from the Italian life table for 1994 (Annuario Statistico Italiano 1997). We first apply the GRR model to the data with the hope to see whether some important genotype can be detected even though our sample has only 375 individuals. Table 4 is the estimated GRRs by assigning the homozygous genotype of allele 8′ as the reference genotype. Frequency estimates for the four alleles are 0.723 for allele 8′ (95% CI: 0.636–0.809), 0.109 for allele 10 (95% CI: 0.073–0.146), 0.160 for allele 11 (95% CI: 0.119–0.201), 0.008 for allele 12 (95% CI: 0.000–0.018). Among all the seven genotypes, only 8′/11 showed a *P*-value of 0.016 with a GRR of 1.140 indicating that it could be harmful in terms of survival. We use the log likelihood ratio test to measure the overall significance level on all the GRRs with $\chi^2_{(6)} = -2(-447.853 + 443.133) = 9.440$. Unfortunately, the overall *P*-value is 0.150. However, it is interesting to see that, in Table 4, all the other genotypes with allele 11, 10/11 and 11/11, have GRR above 1 although not significant. This together with the risk estimate for genotype 8′/11, could indicate that allele 11 might be the only risky allele. In another estimation, we apply the ARR model by assigning 8′ as the reference allele. We specify for each of the other alleles, 10, 11 and 12, one risk and one frequency parameter. In Table 5, we show the parameter estimates together with the significance levels for the RRs. One can see that the ARR and GRR models give very similar estimates on the allele frequencies although with completely different ways of parameterization on the risk parameters. It is also interesting to see that, among all the alleles, only allele 11 exhibits a harmful effect that increases the carrier's hazard of death ($P = 0.029$). The estimated risk for allele 11 (ARR = 1.103) is very close to the estimated risk for genotype 8′/11 (GRR = 1.140) in Table 4. We again deploy the log likelihood ratio test to summarize the overall significance of the mutant alleles on survival. We have $\chi^2_{(3)} = -2(-452.595 + 450.052) = 5.085$ with a *P*-value of 0.166. Our analysis indicates that, although both GRR and ARR models point to allele 11 as a potentially risky allele, effect on survival from the gene as a whole does not reach the significance level. More data are needed in order to clarify the association.

*Table 4*. Estimated genotype relative risk on survival for *Renin* gene.

| Genotype[a] | GRR | Std | *P*-value |
|---|---|---|---|
| 8′/10 | 1.030 | 0.061 | 0.620 |
| 8′/11 | 1.140 | 0.058 | 0.016 |
| 8′/12 | 1.024 | 0.213 | 0.909 |
| 10/11 | 1.261 | 0.146 | 0.112 |
| 10/12 | 0.763 | 0.259 | 0.360 |
| 11/11 | 1.041 | 0.127 | 0.744 |

[a]8′/8′is baseline genotype.

*Table 5*. Frequency and relative risk estimates for *Renin* gene alleles.

| Allele | Frequency | | ARR | | |
|---|---|---|---|---|---|
| | Est. | 95% CI | Est. | Std | *P*-value |
| 8′[a] | 0.734 | 0.652–0.814 | 1.000 | – | – |
| 10 | 0.099 | 0.067–0.131 | 1.033 | 0.056 | 0.547 |
| 11 | 0.160 | 0.119–0.202 | 1.103 | 0.047 | 0.029 |
| 12 | 0.007 | 0.000–0.015 | 0.948 | 0.175 | 0.765 |

[a]8′ is baseline allele.

## Discussion

We have presented survival analysis models to assess genetic association with human survival at

multi-allelic loci using genotype data from cross-sectional studies. While the GRR model offers the opportunity for researchers to investigate the different modes of the allelic effects, the ARR model minimizes the number of parameters in the model such that it can easily handle highly polymorphic genes. Application to *Renin* gene data has shown that both models can help to pinpoint important genetic variations that influence human survival. In our model, all parameters are estimated simultaneously in one likelihood function by choosing one single genotype or allele as reference which avoids the dependency problem in the previous approaches (Yashin et al. 1999; Tan et al. 2001a). In addition, the likelihood ratio test offers a nice way to measure the significance of an overall association with survival which the previous approaches failed to do. However, as in the previous approaches, both the GRR and ARR models are proportional hazard models in nature which assume genetic effects on the hazard of death are constant over ages. Proportionality enables us to parameterize on genotypes or alleles to infer the modes of gene function or to fit ARR models for highly polymorphic genes. On the other hand, the proportional hazard model is incapable of capturing genes that show age-dependent effects (Schachter et al. 1994; De Benedictis et al. 1998b) such as antagonistic pleiotropy. Yashin et al. (1999) introduced non-parametric and semi-parametric models to model the age-specific effects. However, a large sample size is required to ensure reliable estimations. In the case of small scale studies, we think the proportional hazard model should be recommended to be on the safe side.

Although multiplicity of allelic effects in the ARR model may sound a bit heavy, we think, as a trade-off, it is useful when sample size is small and the gene of interest is highly polymorphic, a situation that renders other models helpless. The multiplicative assumption is popular in summarizing epistatic risks in mapping genes for complex diseases (Risch 1990; Clayton and Jones 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 done whenever it is feasible because such analysis can help to investigate whether the gene is recessive, codominant, or dominant (Sasieni 1997). It would also be a good idea to apply both models so that results can be compared as done in the example application.

In both the GRR and ARR models, we introduce the Hardy–Weinberg equilibrium to largely cut down the number of frequency parameters. We assume the Hardy–Weinberg equilibrium at birth is sensible because there has been no selection yet imposed at birth as long as the locus we are interested does not influence *in utero* survival and there is no preferential transmission of a particular genetic variant at the locus. In addition, sex-specific allele frequency can easily be introduced into our model to account for possible situations that bring up sex-specific gene frequency at birth. Moreover, as in any gene-disease association study, violation of the Hardy–Weinberg law and linkage disequilibrium can result from population stratifications. In order to avoid spurious results, it is necessary to make sure that the sampling population is ethnically homogeneous.

As longevity is a rare event, studies on this phenotype has always been confronted with problems in obtaining sufficient samples especially when the gene or gene marker under investigation is highly polymorphic. However, accompanied by its high efficiency in fine mapping (Cardon and Bell 2001) and together with the help of new statistical approaches, we think that association study will be a powerful tool in searching for genetic variations that contribute to human aging and survival.

## Acknowledgements

This work was partially financed by the US National Institute on Aging (NIA) research grant NIA-P01-AG08761 and by the Italian Ministry of Health (IRCCS project 2000–2002 Marcatori Genetici e Biologici di invecchiamento Normale e Patologico to Prof G. De Benedictis). Dr Tan wants to thank the Max-Planck Institute for Demographic Research in Rostock, Germany, for constant support in his research.

## References

Annuario statistico Italiano (1997)

Bathum L, Andersen-Ranberg K, Boldsen J, Brosen K and 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(5): 427–430

Bathum L, Christiansen L, Nybo H, Ranberg KA, Gaist D, Jeune B, Petersen NE, Vaupel J and Christensen K (2001) Association of mutations in the hemochromatosis gene with shorter life expectancy. Arch Intern Med 61(20): 2441–2444

Bijlsma R, Bundgaard J and Boerema AC (2000) Does inbreeding affect the extinction risk of small population? Predictions from Drosophila. J Evol Biol 13: 502–514

Cardon LR and Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2(2): 91–99

Charlesworth D and Charlesworth B (1987) Inbreeding depression and its evolutionary consequences. Annu Rev Ecol Syst 18: 237–268

Clayton D and Jones H (1999) Transmission/disequilibrium tests for extended marker haplotypes. Am J Hum Genet 65(4): 1161–1169

Dahlgaard J and Hoffmann A (2000) Stress resistance and environmental dependency of inbreeding depression in *Drosophila melanogaster*: general versus Specific Effects. Conserv Biol 14: 1187–1192

Daly AK (2003) Candidate gene case-control studies. Pharmacogenomics 4(2): 127–139

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 and 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 AI, Bonafe M and 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 AI, Bonafe M and Franceschi C (1998b) Age-related changes of the 3′APOB-VNTR genotype pool in ageing cohorts. Ann Hum Genet 62: 115–122

De Benedictis G, Tan Q, Jeune B, Christensen K, Ukraintseva SV, Bonafe M, Franceschi C, Vaupel JW and Yashin AI (2001) Recent advances in human gene–longevity association studies. Mech Ageing Develop 1(2): 909–920

Dubois MA, Sabatier P, Durand B, Calavas D, Ducrot C and Chalvet-Monfray K (2002) Multiplicative genetic effects in scrapie disease susceptibility. C R Biol 325(5): 565–570

Gerdes LU, Jeune B, Ranberg KA, Nybo H and Vaupel JW (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(3): 202–210

Ivanova R, Henon N, Lepage V, Charron D, Vicaut E and 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 MJ, Salokannel J, Hynninen A, Heikkinen J, Ehnholm C, Koistinen MJ and Kesaniemi YA (1994) Apolipoprotein E and B polymorphisms–longevity factors assessed in nonagenarians. Atherosclerosis 105(1): 89–95

Kristensen TN, Dahlgaard J and Loeschcke V (2002) Inbreeding affects the Hsp70 expression level in two species of *Drosophila* even at benign temperatures. Evol Ecol Res 4: 1209–1216

Koeleman BP, Dudbridge F, Cordell HJ and Todd JA (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(Pt 3): 207–213

Lee HY, Kim D and Choi MY (1998) Genetic polymorphism in an evolving population. Phys Rev E 57(4): 4842–4845

Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46(2): 222–228

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

Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405(6788): 847–856

Sasieni PD (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 and Cohen D (1994) Genetic associations with human longevity at the APOE and ACE loci. Nat Genet 6: 29–32

Schwanke CH, da Cruz IB, Leal NF, Scheibe R, Moriguchi Y and Moriguchi EH (2002) Analysis of the association between apolipoprotein E polymorphism and cardiovascular risk factors in an elderly population with longevity. Arq Bras Cardiol 78(6): 561–579

Sham PC and Curtis D (1995) Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. Ann Hum Genet 59(Pt 1): 97–105

Slooter AJ, Cruts M, Van Broeckhoven C, Hofman A and van Duijin CM (2001) Apolipoprotein E and longevity: the Rotterdam Study. J Am Geriatr Soc 49(9): 1258–1259

Tan Q (2000) How genes affect longevity in heterogeneous populations: binomial frailty models and applications. PhD thesis. Faculty of Health Science, University of Southern Denmark

Tan Q, De Benedictis G, Yashin AI, Bonafe M, DeLuca M, Valensin S, Vaupel JW and Franceschi C (2001a) Measuring the genetic influence in modulating human life span: gene-environment and gene-sex interactions. Biogerontology 2(3): 141–153

Tan Q, Yashin AI, Bladbjerg EM, de Maat M, Andersen-Ranberg K, Jeune B, Christensen K and Vaupel JW (2001b) Variations of Cardiovascular Disease Associated Genes Exhibit Sex-dependent Influence on Human Longevity, Exp Gerontol 36(8): 1303–1315

Tan Q, Bellizzi D, Rose G, Garasto S, Franceschi C, Kruse T, Vaupel J, De Benedictis G and Yashin A (2002) The influences on human longevity by HUMTHO1.STR polymorphism (Tyrosine Hydroxylase gene). A relative risk approach. Mech Ageing Dev 123(10): 1403–1410

Toupance B, Godelle B, Gouyon PH and Schachter F (1998) A model for antagonistic pleiotropic gene action for mortality and advanced age. Am J Hum Genet 62(6): 1525–1534

Varcasia O, Garasto S, Rizza T, Andersen-Ranberg K, Jeune B, Bathum L, Andreev K, Tan Q, Yashin AI, Bonafe M, Franceschi C and De Benedictis G (2001) Replication studies in longevity: puzzling findings in Danish centenarians at the 3′APOB-VNTR locus. Ann Hum Genet 65(Pt 4): 371–376

Vaupel JW and Yashin AI (1985) Heterogeneity's ruses: some surprising effects of selection on population dynamics. Am Stat 39: 176–185

Wang X, Wang G, Yang C and Li X (2001) Apolipoprotein E gene polymorphism and its association with human longevity in the Uygur nationality in Xinjiang. Chin Med J (Engl) 114(8): 817–820

Wright AF, Carothers AD and Pirastu M (1999) Population choice in mapping genes for complex diseases. Nat Gen 23: 397–404

Yashin AI, Vaupel JW, Andreev KF, Tan Q, Iachine IA, Carotenuto L, De Benedictis G, Bonafe M, Valensin S and Franceschi C (1998) Combining genetic and demographic information in population studies of aging and longevity. J Epidemiol Biostat 3: 289–294

Yashin AI, De Benedictis G, Vaupel JW, Tan Q, Andreev KF, Iachine IA, Bonafe M, DeLuca M, Valensin S, Carotenuto L and 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

Yashin AI, De Benedictis G, Vaupel JW, Tan Q, Andreev KF, Iachine IA, Bonafe M, DeLuca M, Valensin S, Carotenuto L and Franceschi C (2000) Genes and longevity: lessons from studies on centenarians. J Gerontol 55a: B1–B10