Detecting genes contributing to longevity using twin data

*Alexander Begun**

Institute of Medical Informatics and Statistics, Brunswiker Strasse 10, D-24105 Kiel, Germany
**Correspondence to:* Tel.: +49 431 5973198; Fax: +49 431 5973193; E-mail: alexander.begun@gmx.de

Date received (in revised form): 26th August 2009

Abstract

Searching for genes contributing to longevity is a typical task in association analysis. A number of methods can be used for finding this association — from the simplest method based on the technique of contingency tables to more complex algorithms involving demographic data, which allow us to estimate the genotype-specific hazard functions. The independence of individuals is the common assumption in all these methods. At the same time, data on related individuals such as twins are often used in genetic studies. This paper proposes an extension of the relative risk model to encompass twin data. We estimate the power and also discuss what happens if we treat the twin data using the univariate model.

Keywords: *Heterogeneity, longevity genes, maximum likelihood method, relative risk, twin data*

Introduction

Most common diseases and traits have a complex structure, for which the phenotype is determined by interactions between genetic and environmental factors. As any individual genetic variant can have a relatively modest effect on a disease or trait, linkage analysis has less power than association analysis. Classical association studies in their simplest form compare the frequency of alleles or genotypes for candidate genes between cases and controls. These candidate genes are usually chosen on the basis of biological hypotheses or from previous linkage analyses.

To identify genes associated with longevity, information on genotype frequencies for two or more age groups is needed. A significant trend of genotype frequencies being associated with age can indicate a gene–longevity association. In the basic 'gene frequency method', only the genotype frequencies in different age groups are compared.[1–3] Some extensions of this method involve the use of demographic information about the population under study and allow the estimation of initial frequencies, relative risks and the age trajectories of mortality for candidate genes. These methods are known in the literature as the 'parametric method', the 'semi-parametric method', the 'non-parametric method' and the 'relative risk method'.[4] The use of these methods, however, has two limitations. First, the initial gene frequencies in all cohorts represented in the study must be the same. Secondly, the mortalities for genotypes do not depend on the birth year of the cohort. In two recently published papers,[5,6] the authors exclude the first limitation, assuming a time trend in the genetic frequencies of subsequent birth cohorts. In principle, the time and the cohort covariates influencing mortality can be incorporated into the models too. The flexible parameterisation in the extended relative risk model[6] also allows detection of the antagonistic pleiotropic effect.

The methods mentioned above have been developed for datasets consisting of independent individuals. In this paper, we propose a method for

detecting longevity genes for the dataset consisting of twin pairs. This method retains all the advantages of the relative risk model for univariate data described previously.[6]

Materials and methods

To analyse the gene–longevity association, two datasets are needed: the genotype data and the univariate survival data for the individuals involved in the study. To improve the accuracy and power of the study, the longevity data for twins can additionally be analysed. Denoting longevity and non-longevity alleles at an autosomal locus by a and A, respectively, assume that the frequencies P_g of genotypes AA, aA or Aa, and aa at the moment of birth are P_{AA}, P_{Aa} and P_{aa}, respectively. If the Hardy–Weinberg equilibrium holds, then $P_{AA} = (1 - P_a)^2, P_{Aa} = 2P_a(1 - P_a)$ and $P_{aa} = P_a^2$, where P_a is the frequency of the allele a at the moment of birth. We parameterise P_a as follows:

$$P_a = 1 - 1/(1 + e^{\nu+\delta x+R\varphi(x,x_0)}), x = T - t. \quad (1)$$

In accordance with (1), the logit of P_a is a linear function of unknown parameters R, ν and δ with domain of definition $\mathbf{R}^3$. This parameterisation includes the sudden change in the allele frequency by the value $R\varphi(x,x_0)$ in the cohort $T-x_0$ and the slow linear cohort effect $\nu+\delta x$ of the allele frequency. Here, T stands for the year of data collection, x for the age, and t for the cohort. We assume that the value of x_0 is known. The step function $\varphi(x,x_0)$ is defined by the interval equations $\varphi(x,x_0)=1$ for $0 \le x \le x_0$ and $\varphi(x,x_0)=0$ for $x > x_0$.

To estimate the genotype frequencies for twin pairs, we need to calculate the bivariate survival functions. One possible approach to doing so is to use the correlated gamma-frailty model, which provides simple analytical expressions for the bivariate survival functions.[7] Assume that that individual's instantaneous risk of death μ for genotype $g \in \{aa,Aa,AA\}$ at age x, as measured by the hazard of mortality, is $\mu(x, Z, g) = Z\mu_{0,g}(x)$, where Z is the gamma distributed frailty (non-observed risk of mortality) with mean 1 and variance σ^2, and $\mu_{0,g}(x)$ is the baseline hazard. The univariate survival function $S_g(x) = Ee^{-ZH_g(x)} = (1 + \sigma^2 H_g(x))^{-1/\sigma^2}$ is the Laplace transform for the gamma probability density function at the point $H_g(x) = \int_0^x \mu_{0,g}(t)dt$ (cumulative hazard function). For related individuals, we assume that life spans T_1 and T_2 are conditionally independent, given frailties Z_1, Z_2 and genotypes g_1, g_2. In general, frailties Z_1 and Z_2 have unequal variances. Below, we shall assume, for simplicity, that Z_1 and Z_2 are identically distributed. If $Corr(Z_1, Z_2) = \rho$, $E(Z_1) = E(Z_2) = 1$ and $Var(Z_1) = Var(Z_2) = \sigma^2$, then:

$$P\{T_1 > x_1, T_2 > x_2\} = S_{g1,g2}(x_1, x_2) = \frac{S_{g1}(x_1)^{1-\rho} S_{g2}(x_2)^{1-\rho}}{(S_{g1}(x_1)^{-\sigma^2} + S_{g2}(x_2)^{-\sigma^2} - 1)^{\rho/\sigma^2}} \quad (2)$$

Here, $S_{g1,g2}(x_1, x_2)$ is the bivariate survival function at ages x_1 and x_2 for twins with genotypes g_1 and g_2, respectively. We relate cumulative hazard functions with some unknown function $H_0(x)$ as follows:

$$H_g(x) = c_g x + a_g H_0(x)^{b_g} \quad (3)$$

with unknown $a_g \ge 0, b_g \ge 0$ and $c_g \ge 0$. Such parameterisation, where cumulative hazards $H_g(x)$ rather than survival functions $S_g(x)$ for different genotypes are parametrically related (eg $S_g(x) = S_0(x)^{b_g}$), is more flexible and allows us to detect the antagonistic pleiotropic effect.[6] Without loss of generality, we can assume that $a_{AA} = b_{AA} = 1$.

For univariate and bivariate survival functions in the whole population, it holds that:

$$S(x) = \sum_g P_g S_g(x),$$
$$S^{MZ}(x_1, x_2) = \sum_{g,g} P^{MZ}_{g,g} S^{MZ}_{g,g}(x_1, x_2) \quad (4)$$
$$S^{DZ}(x_1, x_2) = \sum_{g1,g2} P^{DZ}_{g1,g2} S^{DZ}_{g1,g2}(x_1, x_2)$$

Here, P_g, $P^{MZ}_{g,g}$ and $P^{DZ}_{g1,g2}$ are the univariate and the bivariate genotype frequencies for monozygotic (MZ) and dizygotic (DZ) twin pairs, respectively, at the moment of birth. Since the frailty correlation

ρ can be different for MZ and DZ twins, we use the upper index MZ or DZ in the notation for bivariate survival. Given univariate survival $S(x)$ and parameters, we calculate the baseline cumulative hazard $H_0(x)$ using the simple bisectional procedure.[6] For univariate genotype frequencies, we will use the values given above. To calculate the bivariate genotype frequencies, note that for MZ twin pairs, $g_1{=}g_2{=}g$ and $P^{MZ}_{g,g} = P_g$. Assuming independent transmission of the maternal and paternal alleles to the offspring, we can use the standard formulae for DZ twin pairs:

$$
\begin{aligned}
P^{DZ}_{aa,aa} &= P^2_{aa} + (1/2)P_{aa}P_{Aa} + (1/16)P^2_{Aa}\\
P^{DZ}_{aa,Aa} &= P^{DZ}_{Aa,aa} = (1/2)P_{aa}P_{Aa} + (1/8)P^2_{Aa}\\
P^{DZ}_{aa,AA} &= P^{DZ}_{AA,aa} = (1/16)P^2_{Aa}\\
P^{DZ}_{Aa,AA} &= P^{DZ}_{AA,Aa} = (1/2)P_{Aa}P_{AA} + (1/8)P^2_{Aa}\\
P^{DZ}_{Aa,Aa} &= (1/2)P_{aa}P_{Aa} + 2P_{aa}P_{AA}\\
&\quad + (1/2)P_{Aa}P_{AA} + (1/4)P^2_{Aa}\\
P^{DZ}_{AA,AA} &= P^2_{AA} + (1/2)P_{AA}P_{Aa} + (1/16)P^2_{Aa}
\end{aligned}
\tag{5}
$$

The frequencies $\pi^{MZ}_{g,g}(x)$ and $\pi^{DZ}_{g1,g2}(x)$ of the genotype (g,g) and (g_1,g_2) at any age x for MZ and DZ twin pairs can be calculated from the formulae:

$$
\pi^{MZ}_{g,g}(x) = \frac{P_g S^{MZ}_{g,g}(x,x)}{S^{MZ}(x,x)} \tag{6}
$$

$$
\pi^{DZ}_{g1,g2}(x) = \frac{P^{DZ}_{g1,g2}\, S^{DZ}_{g1,g2}(x,x)}{S^{DZ}(x,x)} \tag{7}
$$

Assuming that the variance σ^2 does not depend on genotype and zygosity, we have the following unknown vector parameter:

$$
\begin{aligned}
\theta = (&R, \delta, \nu, a_{aa}, a_{aA+Aa}, b_{aa}, b_{aA+Aa}, c_{aa}, c_{aA+Aa},\\
&c_{AA}, \sigma^2, \rho_{MZ}, \rho_{DZ}).
\end{aligned}
$$

Here, ρ_{MZ} and ρ_{DZ} are the frailty correlations for MZ and DZ twins. We estimate unknown vector parameter θ maximising the likelihood function:

$$
Lik_g = \left(\prod_{i=1}^{N^{MZ}_g} \pi^{MZ}_{g_i,g_i}(x_i,\theta)\right)\left(\prod_{i=1}^{N^{DZ}_g} \pi^{DZ}_{g_{i1},i2}(x_i,\theta)\right) \tag{8}
$$

(the maximum likelihood estimates [MLE]), where x_i is the age of twin pair i at the moment of data collection, N^{MZ}_g and N^{DZ}_g are the observed numbers of MZ and DZ twin pairs in the genetic dataset (twin pairs with known genotypes and ages), respectively. To choose the optimal model, we can use the likelihood ratio test for nested models and either the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) for non-nested models. Under the null hypothesis, we assume that $a_{aa} = a_{Aa} = b_{aa} = b_{Aa} = 1$ and $c_{aa} = c_{Aa} = c_{AA} = 0$. Significant deviation from this hypothesis can indicate a gene–longevity association.

If, in addition to genetic data, the data on the longevity of related individuals such as twins are also available, we can use this information to improve the accuracy of statistical estimates and to increase the power. Denote the life spans of the twin pair i in the demographic dataset by (x_{i1},x_{i2}), where $I = 1,\ldots,N^{MZ}_d$ for MZ twin pairs and $I = 1,\ldots,N^{DZ}_d$ for DZ twin pairs. We assume that twin pairs in the sample are chosen at random and that all twins are deceased. Although the censored data are less informative than non-censored data, they can be also included in the analysis. The bivariate probability density function for a twin pair with longevities x_{i1} and x_{i2} can be calculated as follows:

$$
\begin{aligned}
\frac{\partial^2 S^j(x_{i1},x_{i2})}{\partial x_{i1}\partial x_{i2}} &= \sum_{g1,g2} P^j_{g1,g2}\frac{S^j_{g1,g2}(x_{i1},x_{i1})}{S_{g1}(x_1)S_{g2}(x_2)}\frac{\partial S_{g1}(x_1)}{\partial S(x_1)}\\
&\times \frac{\partial S_{g2}(x_2)}{\partial S(x_2)}\frac{\partial S(x_1)}{\partial x_1}\frac{\partial S(x_2)}{\partial x_2}\\
&\times \left(1 - \rho_j^2 + \frac{\rho_j(1-\rho_j)\left(S_{g1}(x_1)^{-\sigma^2} + S_{g2}(x_2)^{-\sigma^2}\right)}{\left(S_{g1}(x_1)^{-\sigma^2} + S_{g2}(x_2)^{-\sigma^2} - 1\right)}\right.\\
&\left. + \frac{\rho_j(\rho_j+\sigma^2)S_{g1}(x_1)^{-\sigma^2}S_{g2}(x_2)^{-\sigma^2}}{\left(S_{g1}(x_1)^{-\sigma^2} + S_{g2}(x_2)^{-\sigma^2} - 1\right)^2}\right)
\end{aligned}
\tag{9}
$$

with $j=MZ,DZ$. We can write the likelihood function for the demographic dataset in the form:

$$Lik_d = \left(\prod_{i=1}^{N_d^{MZ}} \frac{\partial^2 S^{MZ}(x_{i1},x_{i2})}{\partial x_{i1}\partial x_{i2}}(x_{i1},x_{i2},\theta) \right) \times \left(\prod_{i=1}^{N_d^{DZ}} \frac{\partial^2 S^{DZ}(x_{i1},x_{i2})}{\partial x_{i1}\partial x_{i2}}(x_{i1},x_{i2},\theta) \right) \quad (10)$$

Now, unknown parameters can be found through maximising the joint likelihood function $Lik_g \times Lik_d$.

Results

To carry out the numerical experiments, we used simulated data. To generate datasets with a sample size of $N_g^{MZ}=1000$, $N_g^{DZ}=2000$ for genotype data and of $N_d^{MZ}=150$, $N_d^{DZ}=300$ for longevity data we assumed that:

- The action of the dominant allele a on longevity can be characterised by parameters $a_{AA}=b_{AA}=1$, $c_{aa}=c_{AA}=0$, $a_{aa}=a_{Aa}=0.8$, $b_{aa}=b_{Aa}=1.2$;
- The survival function for genotype AA has a form

$$\tilde{S}(x) = (1+s^2\tilde{H}(x))^{-1/s^2},$$
$$\tilde{H}(x) = \tilde{c}x + \tilde{a}(e^{\tilde{b}x}-1)/\tilde{b} \quad (11)$$

  with $\tilde{a}=2.5\cdot 10^{-5}$, $\tilde{b}=0.1$, $\tilde{c}=0$ and $\ln s^2=-4.5$;
- Individual frailty for twins are gamma-distributed, with mean 1 and variance $\sigma^2=1$. Frailty correlations ρ_{MZ} and ρ_{DZ} are equal to 0.5 and 0.25, respectively;
- The Hardy–Weinberg equilibrium at the moment of conception holds. There is no genotype selection before birth;
- The slow continuous component of the cohort effect has parameters $\nu=-2$ and $\delta=0.005$. This corresponds to the frequency $P_a\approx 0.182$ for individuals born in year T (the year of data collection) and decreases in the frequency by 0.4 per cent per year. The sudden jump of P_a with parameter $R=0.5$ occurred in the cohort T-50;
- The birth dates of all twin pairs from the longevity dataset are uniformly distributed over the cohort interval $[T\text{-}110, T\text{-}100]$. The ages of the twins from the genetic dataset at the moment of data collection are uniformly distributed over the age interval [0,105] years.

Nearly one in every 100 deliveries is a twin birth, and the DZ/MZ ratio is approximately equal to 2. From this, it follows that in the stationary population consisting of 300,000 individuals with crude birth and death rates q_0 equal to 15 per 1,000, the life expectancy at birth e_0 is equal to $1{,}000/15\approx 66.7$ years and we will find approximately $(1/300)\times(300{,}000\times q_0e_0)=1{,}000$ MZ and $(2/300)\times(300{,}000\times q_0e_0)=2{,}000$ DZ twin pairs. We will also find 150 MZ and 300 DZ newborn twin pairs over the ten-year cohort interval. Since the influence of a decrease in child mortality before the age of 11–13 years on the univariate survival and, therefore, selection is relatively small, we have not included this effect in the simulated data. In general, chosen simulation parameters produce a bivariate lifespan distribution which is similar to the true one.

The estimates of unknown parameters and of the power for 1,000 simulations are given in Table 1. The power was calculated at the 5 per cent significance level. We have used the bivariate and the univariate models applied to the joint bivariate genetic and longevity data or to the bivariate genetic data only. The age dynamics of the hazard functions for genotypes with/without allele a and the age dynamics of the frequencies for the longevity allele/genotypes with the longevity allele are shown in Figures 1 and 2. To establish how often the true bivariate model applied to the bivariate genetic data turns out to be optimal compared with the false univariate model, we used the likelihood ratio test. Significant differences between two these models

Table 1. Parameter estimates (sample means) and their standard deviations (in brackets) for 1,000 simulations, calculated using the bivariate (univariate) model applied to the joint bivariate genetic and longevity data* (***) or to the bivariate genetic data **(****).

	True	Est.*	Est.**	Est.***	Est.****
a_{aa}	0.800	0.775 (0.219)	0.693 (0.431)	0.605 (0.736)	0.614 (0.517)
b_{aa}	1.200	1.198 (0.039)	1.196 (0.062)	1.261 (0.070)	1.252 (0.065)
ν	−2.000	−2.009 (0.178)	−2.016 (0.180)	−1.996 (0.183)	−1.996 (0.182)
$10^3 \cdot \delta$	5.000	5.066 (2.642)	5.121 (2.660)	4.762 (2.876)	4.934 (2.736)
R	0.500	0.509 (0.126)	0.514 (0.129)	0.505 (0.131)	0.501 (0.130)
σ	1.000	1.096 (0.520)	1.368 (0.998)	1.654 (1.008)	1.538 (1.060)
ρ_{MZ}	0.500	0.558 (0.245)	0.539 (0.392)	–	–
ρ_{DZ}	0.250	0.293 (0.212)	0.358 (0.393)	–	–
Power	–	0.833	0.628	0.874	0.719

at a significance level of $p < 0.05$ were observed in 100 per cent of cases.

Discussion

The maximum likelihood method yields correct estimates if the model is correctly specified. In this case, the MLE of unknown parameters under certain regularity conditions are asymptotically unbiased, normal and efficient. If we treat the bivariate data in the same way as the univariate data, and the marginal model is correctly specified, then the robust Hubert–White 'sandwich' estimator of the covariance matrix of parameter estimates yields an asymptotically consistent covariance matrix.[8–10] As we see in Table 1, there is an increase in statistical power when using the more robust univariate model compared with the bivariate model. Nevertheless, the estimates of parameters a_{aa} and b_{aa} for the relative risk of the longevity genotype and the estimate of σ for the standard deviation of frailty are closer to their true values if we use the bivariate model. Including the information on longevity in the dataset, however, can

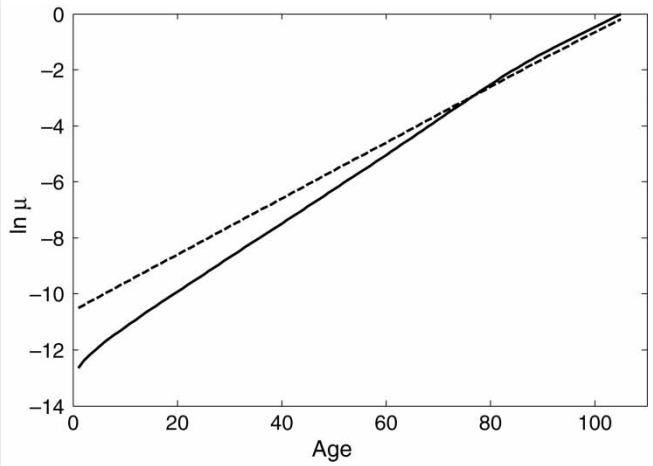

Figure 1. Hazard function for genotypes with/without allele *a* (solid line/dashed line).

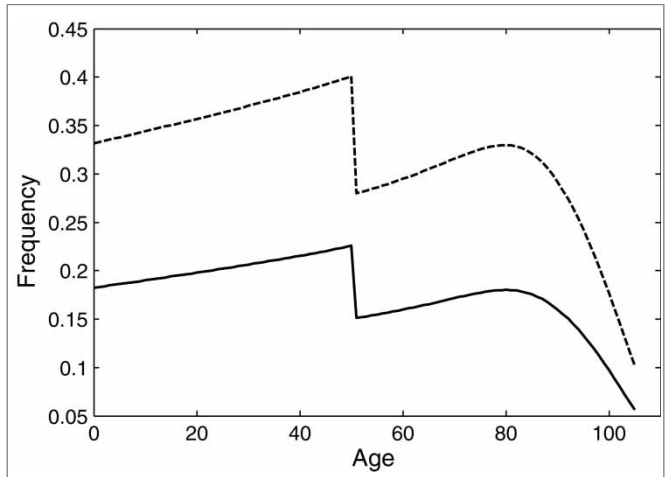

Figure 2. Frequency of allele *a*/genotypes containing allele *a* (solid line/dashed line).

substantially improve statistical estimates, increase the power and decrease the variance. It seems that implementation of the approach based on the more robust univariate model, compared with the bivariate model, is preferable for the sample sizes used in this study. Based on the correlation estimates in the MZ and DZ twins, we are able to estimate the contribution of the candidate gene to the heritability.[11] Under the null hypothesis (no heritability), we put $\rho_{MZ} = \rho_{DZ}$. The effect of antagonistic pleiotropy is clearly seen in Figure 1. The presence of allele *a* in an individual's genotype guarantees the lower hazard of mortality only up to the age of approximately 76 years. The hazard of individuals with genotype *AA* is then lower than that of individuals with allele *a* in the genotype. Similar to the univariate model, the bivariate model effectively identifies not only the slow cohort trend of P_g, including the antagonistic pleiotropic effect, but also the sudden change in this parameter. As expected, the frequencies of allele *a* and of the genotypes containing allele *a* increase continuously in the age intervals [0,50] and [50,80], fall abruptly at the age of 50 and decrease continuously after the age of 80 (see Figure 2). Univariate and bivariate (for twins) genotype frequencies at the longevity locus at the moment of conception depend on the genotype frequencies in the parental population and the transmission probabilities. In the model we have used, two assumptions were made relating to the longevity locus. First, that the Hardy–Weinberg equilibrium holds for the parental population. Secondly, that the segregation ratio does not deviate from 0.5.[12] In principle, we can dispense with both of these assumptions and include them as null hypotheses in the study. Significant deviation from the null hypotheses can be tested using the likelihood ratio test. Rejection of the hypothesis about the Hardy–Weinberg equilibrium can indicate possible genotype selection during the gestation period. Significant deviation from Mendelian transmission can mean, for example, that longevity is not governed by the alleles at a single locus. Population admixture and stratification can lead to linkage disequilibrium between longevity and marker loci. In such situations, the study may reveal evidence for ('spurious') association with the marker, even if it is unlinked to the longevity locus. If the sub-population factors influencing the allele frequencies in the marker and longevity loci are identified (eg ethnicity, geographical origin, etc), they can be included in the study. Another solution for this problem is to partition the association effects into between- and within-family components.[13,14] It was shown that admixture impacts the between-family component estimate, and that the within-family component estimate is independent of any 'spurious' effects when samples from a number of population strata are combined.

References

1. De Benedictis, G., Carotenuto, L., Carrieri, G., Carrieri, G. *et al.* (1998), 'Gene/longevity association studies at four autosomal loci (REN, THO, PARP, SOD2)', *Eur. J. Hum. Genet.* Vol. 6, pp. 534–541.
2. Tan, Q., Bathum, L., Christiansen, L., De Benedictis, G. *et al.* (2003), 'Logistic regression models for polymorphic and antagonistic pleiotropic gene action on human aging and longevity', *Ann. Hum. Genet.* Vol. 67, pp. 598–607.
3. Garasto, S., Rose, G., Derango, F., Berardelli, M. *et al.* (2003), 'The study of APOA1, APOC3, and APOA4 variability in healthy ageing people reveals another paradox in the oldest old subjects', *Ann. Hum. Genet.* Vol. 67, pp. 54–62.
4. Yashin, A.I., De Benedictis, G., Vaupel, J.W., Tan, Q. *et al.* (1999), 'Genes, demography, and life span: The contribution of demographic data in genetic studies on aging and longevity', *Am. J. Hum. Genet.* Vol. 65, pp. 1178–93.
5. Yashin, A.I., Arbeev, K.G. and Ukraintseva, S.V. (2007), 'The accuracy of statistical estimates in genetic studies of aging can be significantly improved', *Biogerontology* Vol. 8, pp. 243–255.
6. Begun, A. (2008), 'A modification of the relative risk model with heterogeneity component for detecting genes contributing to longevity', *Ann. Hum. Genet.* Vol. 72, pp. 111–114.
7. Yashin, A.I., Vaupel, J.W. and Iachine, I.A. (1995), 'Correlated individual frailty: An advantageous approach to survival analysis of bivariate data', *Math. Popul. Stud.* Vol. 5, pp. 145–159.
8. Huber, P.J. (1967), 'The behavior of maximum likelihood estimation under nonstandard conditions', Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley, CA, pp. 221–223.
9. White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica* Vol. 50, pp. 1–25.
10. Williams, R.L. (2000), 'A note on robust variance estimation for cluster-correlated data', *Biometrics* Vol. 56, pp. 645–646.
11. Sham, P. (1998), *Statistics in Human Genetics (Arnold Applications of Statistics Series)*, Edward Arnold, London.
12. Lalouel, J.M., Rao, D.C., Morton, R.E. and Elston, R.C. (1983), 'A unified model for complex segregation analysis', *Am. J. Hum. Genet.* Vol. 35, pp. 816–826.
13. Fulker, D.W., Cherny, S.S., Sham, P.C. and Hewitt, J.K. (1999), 'Combined linkage and association analysis for quantitative traits', *Am. J. Hum. Genet.* Vol. 64, pp. 259–267.
14. Abecasis, G.R., Cardon, L.R. and Coocson, W.O.C. (2000), 'A general test of association for quantitative traits in nuclear families', *Am. J. Hum. Genet.* Vol. 66, pp. 259–292.