Biogerontology **2:** 141–153, 2001.
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

Research article

Measuring the genetic influence in modulating the human life span: gene–environment interaction and the sex-specific genetic effect

Qihua Tan[1], G. De Benedictis[2], A.I. Yashin[1,3,*], M. Bonafe[4], M. DeLuca[2], S. Valensin[4], J.W. Vaupel[1] & C. Franceschi[4]

[1]*Max-Planck Institute for Demographic Research Doberaner Str. 114, 18057 Rostock, Germany;* [2]*Cell Biology Department, University of Calabria, Rende, Italy;* [3]*Center for Demographic Studies and Sanford Institute, Duke University, Durham, North Carolina, USA;* [4]*Department of Experimental Pathology, University of Bologna, Italy;* [*]*Author for correspondence (e-mail: Yashin@demogr.mpg.de; fax: +49-381-2081169)*

Received 16 October 2000; accepted in revised form 1 December 2000

Key words: gene, heterogeneity, longevity, survival analysis, two-step MLE

Abstract

New approaches are needed to explore the different ways in which genes affect the human life span. One needs to assess the genetic effects themselves, as well as gene–environment interactions and sex dependency. In this paper, we present a new model that combines both genotypic and demographic information in the estimation of the genetic influence on life spans. Based on Cox's proportional hazard assumption, the model measures the risks for each gene as well as for gene–environment and gene–sex interactions, while controlling for confounding factors. A two-step MLE is introduced to obtain a non-parametric form of the baseline hazard function. The model is applied to genotypic data from Italian centenarian studies to estimate relative risks of candidate genes, risks due to interactions and initial frequencies of different genes in the population. Results from models that either do or do not take into consideration individual heterogeneity are compared. It is shown that ignoring the existence of heterogeneity can lead to a systematic underestimation of genetic effects and effects due to interactions.

Introduction

The genetics of inter-individual variability in the human life span have been explored in correlation studies in twins and families (McGue et al. 1993; Bocquet-Appel et al. 1990; Herskind et al. 1996), and in association studies of candidate genes in centenarians and younger people from the same population (Schachter et al. 1994; De Benedictis et al. 1997, 1998a; Bathum et al. 1998; Ivanova et al. 1998; Bonafè et al. 1999a, 1999b; Bladbjerg et al. 1999). Moreover, methods that combine genetic information with demographic covariants of the population from which the genetic sample is taken have been proposed (Yashin et al. 1998, 1999; Toupance et al. 1998).

However, new approaches are needed to take into account heterogeneity elements (for example, gene–environment interaction and sex-dependent effects), which are probably crucial in aging and longevity but which have previously been ignored due to methodological limitations (Yashin et al. 1999a). Based on data from Italian centenarian studies, De Benedictis et al. (1998b, 1999) reported that gene–environment interaction and sex-specific effects play a role in individual survival both at the THO locus and in the mitochondrial genome. By applying survival analysis to the same data source, Yashin et al. (2000) found a strong gene–area interaction for the THO10 allele but without incorporating a gene–sex interaction. Likewise, Ivanova et al. (1998) found that HLA-DR7 and HLA-DR11 display significant sex-specific effects on longevity. Gene–environment interactions in the case of complex traits are also reported in twin studies (Martin 2000) and in genetic studies on cancer

(Bennett et al. 1999; Chen et al. 1999), hypertension (Gavras et al. 1999), osteoporosis (Sambrook and Nguyen 1999), and in animal experiments (Clare and Luckinbill 1985; Arking 1987).

The case-control design and case-only study based on the same principle as the case-control study are popular in genetic epidemiology for assessing gene–environment interaction (Andrieu and Goldstein 1999). However, when they are applied to the genetics of longevity, the disadvantages are the same as those addressed by Yashin et al. (1999). First, important variables such as participants' ages are not fully utilized since the aging process is continuous. Second, the demographic background is crucial in assessing the influence of genes on survival (Yashin et al. 1998, 1999). Third, as we show here, the practice of estimating gene–environment interactions is also important for evaluating the effect of genes properly. And it is even more important and relevant in the area of public health, since the discovery of gene–environment interaction and sex-specific effects can help to create a more efficient preventive strategy and to improve the cost-effectiveness of efforts to prolong individual lives.

The relative computational convenience that has resulted from rapidly developing new techniques allows us to think about new approaches in which more information can be combined in measuring the effects of candidate genes and thus more aspects of how genes function in the process of aging can be understood. In this paper, we present a new approach aimed at detecting gene–environment and gene–sex interactions. The approach is based on the relative risk method proposed by Yashin et al. (1999) and is applied to empirical data from Italian centenarian studies.

Materials and methods

Samples

To study the association between genetic variation and longevity, a multicentric longevity study was started in Italy in 1995. Genetic information was collected from individuals in two groups: centenarians and a younger group of people aged 7 to 84. Individuals are recorded by sex and region (southern or northern Italy, respectively). The distribution of participants by sex and area is shown in Table 1. The centenarian group consists of people who had reached the age of 100 or older at the time when blood samples were taken. The oldest individual in this group was a 109-year-old woman. All participants were clinically healthy. The number of males and females in the control, (i.e., the younger) group is well balanced, but this is not the case for the centenarian group, where there are more than twice as many females as males. In the control group, there are more people at younger ages from the south. However, since one can control for confounding factors like area and sex in our model, this data structure is not problematic for the analysis.

Table 1. Observations by sex and area.

Group	Male	Female	Total
Young			
South	311	302	613
North	54	82	136
Total	365	384	749
Centenarian			
South	36	67	103
North	26	83	109
Total	62	150	212

Genotypic data

We examined the eleven autosomal genes and mitochondrial DNA markers shown in Table 2 (apolipoprotein B (APOB): De Benedictis et al. 1998a; Renin (REN), tyrosine hydroxylase (THO), superoxide dismutase 2 (SOD2), poly(ADP-ribose) polymerase (PARP): De Benedictis et al. 1998b; mitochondrial DNA (MtDNA) haplogroups: De Benedictis et al. 1999; superoxide dismutase 1 (SOD1), apolipoprotein A-I (APOA1), apolipoprotein C-III (APOC3), apolipoprotein A-IV (APOA4), insulin (INS), D-Loop MtDNA: unpublished data). The valid number of observations for each locus varies (Table 2) due to the problem of missing values.

Demographic data

Male and female survival distributions are taken from the Italian life table for 1994 (Annuario Statistico Italiano 1997).

Statistical method

Cox's proportional hazard model (Cox 1972) is a widely used tool for doing survival analysis in

Table 2. Genes and markers analyzed.

LOCUS	Biological role	Chromosome	Marker	Alleles	Number of individuals[a]
APOB	Major protein of LDL	2p24-23	3′APOB-VNTR	23, 26, 31, 33, 34, 35, 36, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55	787
REN	Angiotensin II synthesis	1q32	HUMREN4 (STR)	7, 8, 10, 11, 12	375
THO	Catecholamine synthesis	11p.15	HUMTHO1 (STR)	6, 7, 8, 9, 10^{-1}, 10	555
SOD1	Oxygen free radicals scavenging	21q22.1	D21S223 (STR)	1, 2, 3, 4, 5, 6, 7, 8, 10	386
SOD2	Oxygen free radicals scavenging	6q25	(T/C) 401nt	T, C	354
APOA1	Major protein of HDL. Activator of LCAT	11q13-qter	RFLP-MspI (–78nt)	+, –	328
APOC3	Chylomicrons and VLDL	11q13	RFLP-SstI (3′ter)	+, –	328
APOA4	Newly secreted chylomicrons	11q13	RFLP-HincII (ex3)	+, –	328
INS	Codes insulin	11p.15	RFLP-FokI (1428nt)	+, –	438
PARP	DNA repair	1q41-42	STR (ex1)	83, 85, 87, 89, 93, 95, 97, 99	315
Haplo-group	Oxidative phosphorylation	mtDNA	Associated RFLPs	H, I, J, K, T, U, V, W, X, Others	547
D-Loop	Oxidative phosphorylation	mtDNA	STR	132, 134, 136, 138, 140	393

[a] For whom information on both gene typization and age at participation was available.

epidemiology. But it cannot be applied when one is interested in measuring the effect of a certain gene allele or genotype on data from cross-sectional studies because participants are censored as regards their life spans. However, the idea of proportional hazard can be borrowed to construct new models for estimating the relative risks of the gene alleles or genotypes of interest. We define the hazard of death as the instantaneous probability of dying given that an individual has survived to a particular time. We then define the relative risk of a gene allele r as the ratio of hazard of death for carriers of the gene allele, $\mu(x, r)$, to the hazard of death for the non-carriers, which is the baseline hazard $\mu_o(x)$. The proportional hazard model assumes that the relative risk r is constant over time on the baseline hazard so that $\mu(x, r) = r\mu_0(x)$. The corresponding survival function is

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

Although r can take any value greater than zero, a gene allele with r larger than one (frailty allele) increases the hazard of death, while a gene allele with r smaller than one (longevity allele) reduces it. One good example of frailty allele is the ApoE4. Studies have confirmed that it is a frailty allele such that carriers of the allele have lower survival than the non-carriers (Schachter et al. 1994; Zhang et al. 1998).

When considering unobserved individual heterogeneity, or so-called frailty, a frailty model (Vaupel et al. 1979; Vaupel and Yashin 1985; Aalen 1988; Hougaard 1991) should be introduced. If an individual with a gene allele has frailty z, based on the proportional hazard assumption, the hazard of death at age x is $\mu(x, r, z) = xr\mu_0(x)$. The mean hazard of death for a heterogeneous population carrying the gene allele is

$$\begin{aligned} \bar{\mu}(x, r) &= \int_0^\infty \mu(x, r, z) f_x(z)dz \quad (2)\\ &= \int_0^\infty z\mu(x, r) f_x(z)dz\\ &= \mu(x, r)\int_0^\infty z f_x(z)dz\\ &= \mu(x, r)\bar{z}(x) \end{aligned}$$

Table 3. Proportions and risks for different sub-groups.

	South		North	
	+	–	+	–
Proportion	$P_s P_{gs}$	$P_s\,(1 - P_{gs})$	$(1 - P_s)\,P_{gn}$	$(1 - P_s)\,(1 - P_{gn})$
Risk, males	$r r_{area} r_{g\times a} r_{g\times s}$	r_{area}	$r r_{g\times s}$	1
Risk, females	$r r_{area} r_{g\times a}$	r_{area}	r	1

Although several distribution forms can be assumed for frailty z, a gamma-distribution with mean one and variance σ^2 is traditionally preferred (Vaupel et al. 1979; Aalen 1988; Hougaard 1991). Given the gamma-distribution of $z, {}^{-}_{z}\,(x)$ in (2) can be derived as ${}^{-}_{z}(x) = [1 + \sigma^2 r H_0(x)]^{-1}$ (Vaupel et al. 1979) so that (2) becomes

$$\begin{aligned} {}^{-}_{\mu}(x, r) &= \mu(x, r)/[1 + \sigma^2 H(x, r)] \\ &= r\mu_0(x)/[1 + \sigma^2 r H_0(x)] \end{aligned}$$

and its corresponding survival function is

$$\begin{aligned} {}^{-}_{s}(x, r) &= [1 + \sigma^2 r H_0(x)]^{-1/\sigma^2} \qquad (3) \\ &= [1 - \sigma^2 r 1\mathrm{n}(s_0(x))]^{-1/\sigma^2} \end{aligned}$$

Here $H_0(x)$ is the cumulative baseline hazard at age x. Frailty models can help to explain the leveling-off of the death rate at advanced ages as a consequence of selection (Vaupel and Yashin 1985; Aalen 1988). As we can see, the mean frailty ${}^{-}_{z}(x)$ decreases with increasing age as selection takes place in a heterogeneous population. It will be shown later that the risks of observed gene alleles could be underestimated in this situation if one ignores the existence of unobserved heterogeneity.

Since all individuals can be grouped as carriers and non-carriers of a gene allele or genotype, one can introduce the simple two-point distribution for the allele or genotype (Vaupel and Yashin 1985; Hougaard 1991). Then the average survival at age for the mixed population consisting of both carriers and non-carriers is (Vaupel and Yashin 1985)

$$ {}^{=}_{s}(x) = p^{-}_{s}(x, r) + (1 - p)^{-}_{s}(x) \qquad (4)$$

In (4), p is the proportion of carriers at birth, and ${}^{-}_{x}(x)$ is the average survival of non-carriers. From (4) we see that ${}^{=}_{x}(x)$ is the weighted mean survival of carriers and non-carriers. (4) can be extended to include more than two sub-groups on the right-hand side with risk compositions of the observed genetic and non-genetic covariates so that

$$ {}^{=}_{s}(x) = \sum_{i=1}^{k} P_i s^{-}(x, r_i) \qquad (5)$$

In (5), k is the total number of compositions with $\sum_{i=1}^{k} P_i = 1$, and r_i is the risk for sub-group i.
The proportion of sub-group i at age x is

$$ p_i(x) = p_i s^{-}(x, r_i)/{}^{=}_{s}(x) \qquad (6)$$

Based on multinomial distribution (Hastings and Peacock 1975), the likelihood function can be written as

$$ L \propto \prod_{x=x_0}^{\infty} \prod_{i=1}^{k} p_i(x)^{n_i(x)} \qquad (7)$$

where x_0 is age of the youngest participant in the study, and $n_i(x)$ is the number of individuals at age x in sub-group i. $\sum_{i=1}^{k} n_i(x) = N(x)$, the total number of observations at age x.

Extending (4) to include multiple groups enables us to incorporate confounding factors as well as interactions into the model. Our data includes an individual's sex and region, in addition to the genetic covariates. The proportions and total risks of different sub-groups are shown in Table 3. The proportion of individuals from the south is P_s and from the north is $(1 - P_2)$. The proportion of carriers in the south is P_{gs} but that in the north is P_{gn}. r is the risk of carrying the gene allele or genotype, which is defined as the relative risk for carriers in reference to non-carriers. r_{area} is the risk of the confounding factor area, which is defined as the relative risk of being from the south in reference to that of being from the north. $r_{g\times a}$ is the risk of gene–area interaction, which is defined as the relative risk for carriers from the south in reference to carriers from the north. $r_{g\times s}$ is the risk of gene–sex interaction, which is defined as the relative risk for male carriers in reference to female carriers. To detect gene–sex interaction, risk compositions are specified for males and females separately but with shared parameters (Table 3). There are two things to be considered in such an arrangement. First, male and female survivals are different (Hazzard 1986; Holden 1987; Keyfitz and Flieger 1990): death rates for males are usually higher than for females. Second, there could be mortality crossover at late ages (Kannisto 1994), which would indicate that the relative risk of sex itself is not proportional. In the estimation process, separate likelihood functions are constructed for males and for females, respectively, but with shared parameters.

The estimation is carried out by maximizing the product of all the likelihood functions based on each gene allele with the same risk for area r_{area} and the same variance parameter for the unobserved heterogeneity σ^2. The one r_{area} for all genes is necessary because it enables the model to capture the risks of gene–area interaction for different genes. The same σ^2 for all sub-populations is used in order to reduce the number of parameters to be estimated and to increase the efficiency of the estimation. Although variances in unobserved heterogeneity may not be the same among the sub-populations, they cannot differ dramatically since we only observe a small part of the total frailty.

In our estimation strategy we apply a two-step MLE by which male and female baseline survival functions are estimated from (5) for the given parameters (frequencies and risks) in step one. The calculated baseline survivals are then introduced into the likelihood function to estimate the parameters by maximization in step two. This process reiterates until the maximum likelihood function converges (Figure 1). The major advantage of the two-step MLE is that baseline hazard functions for males and females obtained in this way are non-parametric.

All calculations in this paper were performed with the program GAUSS (Aptech Systems 1996), and graphic presentations were constructed using AXUM software (MathSoft 1996).

Results

The model was first applied to each single allele at different loci to find candidate alleles that may have potential influence on individual survival. Irrelevant genes were selected out by testing their statistical significance for their relative risks (risk of the gene allele, risks of the gene–area, and gene–sex interactions). Since this is done for each gene allele separately, the estimate of the risk of area is different for different alleles due to missing values. Twelve alleles at 5 loci (APOB, THO, SOD2, INS, mtDNA, both haplogroups and D-loop markers) were selected from the data as showing potential influence on the life span. We then put them together into one estimation, with the restriction that they had the same risk of area. Significant levels for risk parameters (relative risks for the genes and for interactions) were determined by testing the statistical differences between the estimated risks and one, with the null hypothesis H_0: $r = 1$. The probability of a type I error is $\alpha = 0.05$. The results are shown in Table 4. There are 3 gene alleles (APOB39, THO10, mtDNAhapl-J) with a potential influence on survival, with risks smaller than one, and there are two frailty alleles (THO7 and mtDNAhapl-U), with risks larger than one. There are 3 genes that have significant gene–environment interactions (APOB35, APOB39, SOD2-T). For carriers of APOB35 and 39, southerners have higher risks than

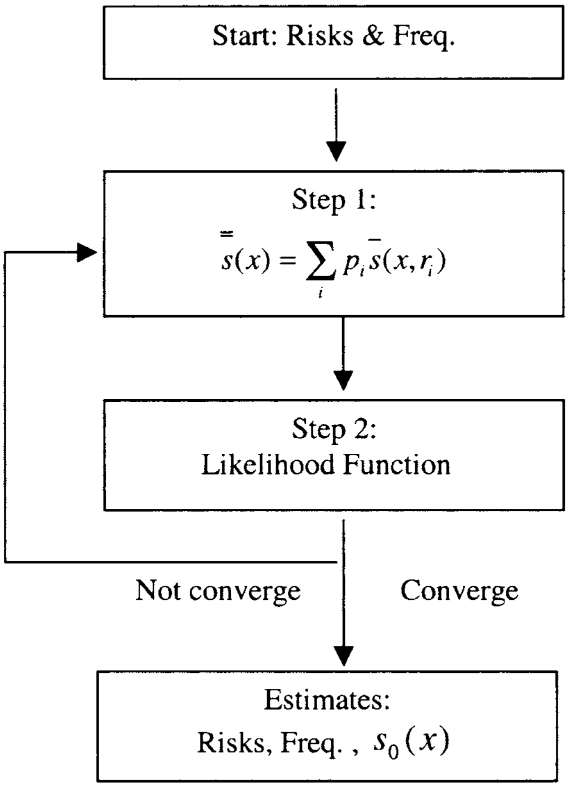

Figure 1. The two-step MLE used for estimating genetic parameters and the baseline survival function. The procedure begins with starting points for the parameters and ends with estimated parameters and baseline survival function when maximum likelihood converges.

Table 4. Parameter estimates without heterogeneity.[a]

Genes	Gene frequency in south Italy		Gene frequency in north Italy		Risk of gene			Risk of $g \times a$			Risk of $g \times s$		
	est.	*sd*	*est.*	*sd*	*est.*	*sd*	P_{-value}	*est.*	*sd*	P_{-value}	*est.*	*sd*	P_{-value}
Apob35	0.402	0.017	0.356	0.017	0.902	0.056	0.079	1.143	0.072	0.046	0.991	0.061	0.883
Apob39	0.084	0.010	0.072	0.009	0.754	0.092	0.008	1.408	0.198	0.039	1.164	0.144	0.255
THO7	0.321	0.020	0.362	0.020	1.117	0.065	0.070	0.936	0.060	0.293	0.906	0.057	0.101
THO8	0.240	0.018	0.142	0.015	0.980	0.087	0.813	1.044	0.095	0.645	0.881	0.064	0.062
THO10	0.330	0.020	0.396	0.021	0.864	0.052	0.009	1.105	0.071	0.135	1.110	0.071	0.122
SOD2-T	0.829	0.020	0.800	0.021	0.986	0.068	0.837	0.898	0.044	0.021	1.033	0.091	0.715
INS–	0.985	0.006	0.965	0.009	1.210	0.155	0.175	0.932	0.041	0.098	0.774	0.166	0.175
INS+	0.258	0.021	0.346	0.023	0.899	0.061	0.095	1.063	0.080	0.429	1.128	0.080	0.107
mtDNAhapl-J	0.045	0.009	0.051	0.009	0.761	0.112	0.033	1.233	0.207	0.261	0.981	0.110	0.867
mtDNAhapl-U	0.138	0.015	0.224	0.018	1.162	0.084	0.053	0.850	0.077	0.053	1.051	0.088	0.563
mtDNAstr-136	0.014	0.006	0.060	0.012	0.933	0.125	0.590	0.637	0.188	0.053	1.059	0.150	0.692
mtDNAstr-138	0.034	0.009	0.014	0.006	0.618	0.201	0.057	1.513	0.521	0.324	1.075	0.184	0.683

[a] r_{area} = 1.162 (*sd* = 0.014, $P_{-value} \approx$ 0.000).

northerners (Table 4). But for carriers of SOD2-T, southerners have lower risk than northerners. There is no allele with sex-specific influences although the P_{-value} of $r_{g\times s}$ for THO8 allele is 0.062. The overall risk of r_{area} is 1.162 (sd = 0.014, ≈ 0.000), which means that southerners have a higher risk of death than northerners.

In another estimation, we took into account unobserved individual heterogeneity. By introducing different variances of hidden frailty σ^2, we arrived at different values of the likelihood function. The highest likelihood was reached when σ^2 is around 0.575 (Figure 2) and when the best fit to the data is obtained (Table 5). Among the major changes, the P_{-value} for the INS+ allele decreased from 0.095 to 0.049, for APOB35 from 0.079 to 0.047, and for mtDNAstr-138 from 0.057 to 0.001. The effect of gene–area interaction for APOB39 becomes less significant although the risk is higher than in Table 4. Meanwhile, the risks of gene–environment interaction for mtDNAhapl-U and mtDNAstr-136 become significant with P_{-value} 0.017 and P_{-value} 0.005 respectively. In the heterogeneity model, THO8 allele shows a strong sex dependent influence on survival which reduces the hazard of death for males but not for females (P_{-value} = 0.013). The estimates of relative risks in Table 5 are all higher when individual heterogeneity is considered. This indicates that if one does not consider heterogeneity, the effect associated with a given gene allele can be systematically underestimated. In Figure 3 we present the hazard functions for female northerners with and without the APOB39 allele. The risk of death is substantially reduced when APOB39 is present.

As concerns allele–area interaction, SOD2-T and mtDNAstr-136 have beneficial effects for southerners although they are neutral genes for northerners. In Figure 4 we plot the mortality curves for mtDNAstr-136 carriers in the south and the north for the two sexes. The risk of death is dramatically lower for southerners than for northerners as a result of gene–environment interaction. In the model that considers unobserved heterogeneity, the estimated r_{area} increases from 1.162 in the model without heterogeneity to 1.611 (sd = 0.066, P_{-value} 0.000).

THO8 is the only gene that shows a sex-specific influence. The gene is neutral in females but it reduces the risk of death by almost half (to 0.644) for males. The mortality curves for southerners are plotted in Figure 5 for both males and females. But only males exhibit a difference between carriers and non-carriers of this allele. The sex-dependent influence of THO8 indicates that the effect of a gene on multifactorial trait depends on the physiological background in which the gene is expressed. Therefore, if the age-related physiological scenario changes in males and females differently, the effect of a certain gene on survival could vary between the sexes. In Figure 5, the female mortality curves overtake those of males at later ages. The necessity of introducing male and female survival functions in the model is obvious.

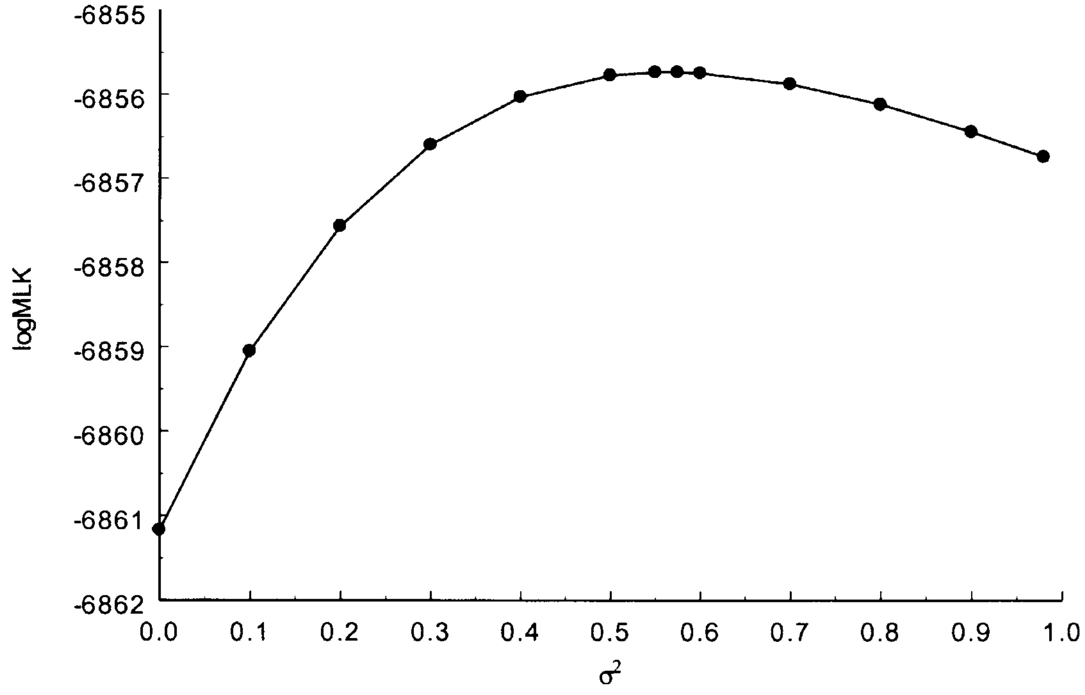

Figure 2. The log likelihood plotted against σ^2. The highest likelihood is reached at the point σ^2 = 0.575.

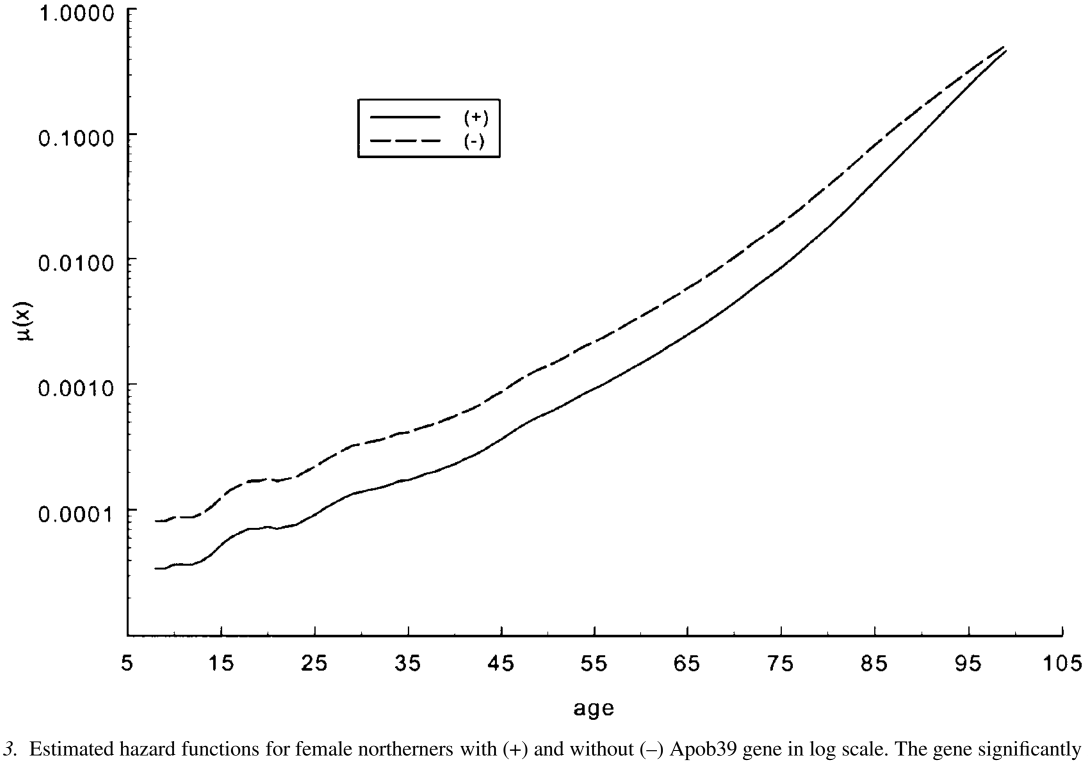

Figure 3. Estimated hazard functions for female northerners with (+) and without (–) Apob39 gene in log scale. The gene significantly reduces risk of death over all ages.

Table 5. Parameter estimates with heterogeneity.[a]

Genes	Gene frequency in south Italy		Gene frequency in north Italy		Risk of gene			Risk of $g \times a$			Risk of $g \times s$		
	est.	*sd*	*est.*	*sd*	*est.*	*sd*	P_{-value}	*est.*	*sd*	P_{-value}	*est.*	*sd*	P_{-value}
Apob35	0.404	0.017	0.354	0.017	0.740	0.131	0.047	1.487	0.269	0.070	1.012	0.306	0.968
Apob39	0.088	0.010	0.069	0.009	0.444	0.132	0.000	2.936	1.201	0.107	1.842	0.747	0.260
THO7	0.322	0.020	0.364	0.020	1.366	0.242	0.130	0.831	0.171	0.323	0.781	0.162	0.177
THO8	0.238	0.018	0.143	0.015	0.972	0.241	0.908	1.091	0.294	0.758	0.644	0.143	0.013
THO10	0.328	0.020	0.391	0.021	0.656	0.106	0.001	1.308	0.246	0.211	1.347	0.266	0.192
SOD2-T	0.830	0.020	0.79	0.021	0.939	0.192	0.752	0.734	0.113	0.019	1.125	0.313	0.689
INS–	0.986	0.006	0.966	0.009	1.675	0.638	0.290	0.815	0.107	0.083	0.589	0.382	0.283
INS+	0.260	0.021	0.344	0.023	0.737	0.134	0.049	1.202	0.275	0.462	1.450	0.325	0.167
mtDNAhapl-J	0.045	0.009	0.050	0.009	0.497	0.172	0.003	1.666	0.759	0.380	0.867	0.258	0.606
mtDNAhapl-U	0.136	0.015	0.228	0.018	1.587	0.375	0.118	0.584	0.175	0.017	1.212	0.343	0.536
mtDNAstr-136	0.016	0.006	0.057	0.012	0.826	0.310	0.573	0.352	0.231	0.005	0.982	0.355	0.959
mtDNAstr-138	0.035	0.009	0.012	0.006	0.275	0.212	0.001	3.154	2.791	0.440	1.158	0.532	0.766

[a] $r_{area} = 1.611$ ($sd = 0.066$, $P_{-value} \approx 0.000$).

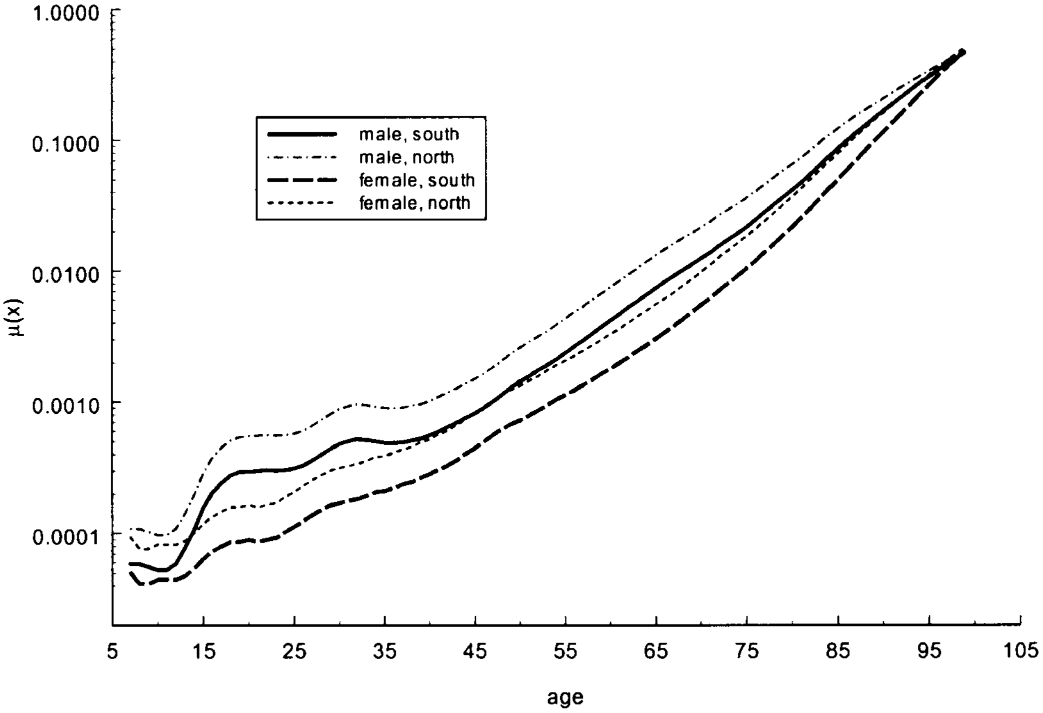

Figure 4. Estimated hazard functions for mtDNAstr-136 male and female carriers by area. Southerners have a lower risk of death than northerners for both sexes.

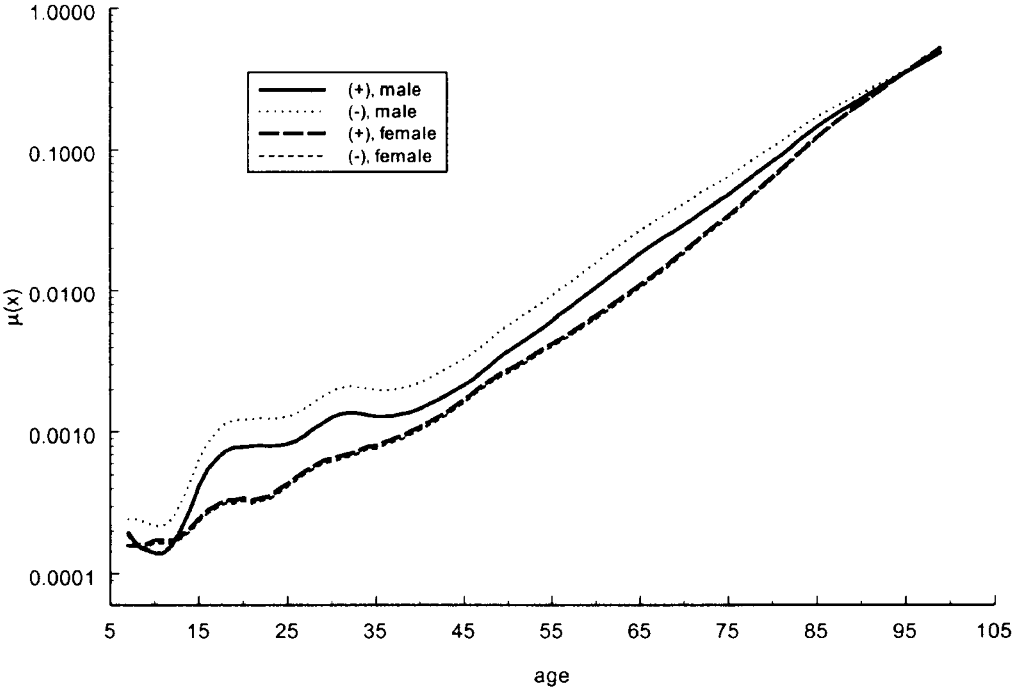

Figure 5. Estimated hazards for southerners with (+) and without (–) THO8 for the two sexes. While the death rate does not differ between female carriers and non-carriers, males with the gene have a lower risk of death than that of those without it.

While some genes may manifest gene–environment interaction, there are others that exhibit different initial frequencies in different geographic regions. Frequencies for THO8 and mtDNAstr-138 are significantly higher in southern than in northern Italy, while the frequencies of INS+, mtDNAhapl-U, mtDNAstr-136 are higher in the north (Table 5). Differences in gene frequencies by area are not unexpected, and they may be due to the differing genetic origins of the southern and northern Italian populations (Cavalli Sforza et al. 1994). In Tables 4 and 5, the reported frequencies are the proportions of carriers of the genes. The corresponding allele frequencies can be calculated since the estimated proportions include individuals carrying one or two gene alleles. Let us assume that is the allele frequency. The proportion of carriers of the allele in the population is $P'^2 + 2P' (1 - P') = 1 - (1 - P')^2$, which equals the estimated proportion P in the model. With this relationship, we can calculate allele frequency as $P' = 1 - \sqrt{1 - P}$. Taking APOB35, for example, the allele frequency is $P' = 1 - \sqrt{1 - 0.404} = 0.228$ in the south. Comparing the frequency estimates in Tables 4 and 5, we see that the introduction of heterogeneity does not seem to affect the estimation significantly.

Conclusions

The present study demonstrates the feasibility of analyzing genotype data in combination with demographic information to estimate the relative risks associated with both a gene by itself and a gene–environment interaction, as well as to estimate sex-specific effects on survival. The estimation of gene–environment interaction is crucial for the following reasons. First, as the results show, gene–environment interaction exists as a common phenomenon in modulating a complex trait such as life span, where the environment has an important role to play. Thus, the study of gene–environment interaction is an important aspect of genetic research on longevity. Second, ignoring these interactions can result in an incorrect assessment of allele effects. If a gene is beneficial in the south but neutral in the north, for example, it could be assessed as a universally beneficial gene if its interaction with geographic area is ignored (simulated result). In this paper, the strategy for detecting gene–environment interaction is also applied to the investigation of the sex dependency of the influence of genes on life span. When investigating sex interaction, it is important to take into account the existing mortality difference between the two sexes, which results in a very high proportion of females among the oldest-old (Vaupel et al. 1998). The two distinctive features, the inclusion of gene–sex interaction and the introduction of sex-specific survival, provide a feasible and efficient way of measuring the sex dependency in gene expression and regulation.

In addition to gene–environment interaction, the environment itself plays a major role in survival (Christensen and Vaupel 1996; Herskind et al. 1996; McGue et al. 1993; Harris et al. 1992) and can thus act as a confounding factor that influences the evaluation of genetic effects (Sellers et al. 1998). Only when the interference of the environment is properly controlled for can the genetic and interactive terms be measured correctly. Sex is another confounding factor that matters. In this application, it is successfully avoided when male and female survival functions available from population life tables are introduced and the interaction of sex incorporated. The sex-specific effect is measured as an extra risk for carriers of the gene of one sex when other parameters are controlled for. With this strategy, it is possible to include other confounding factors when necessary, and it is possible to extend this model to explore gene–gene interaction as well. The life span as a complex trait is a polygenic phenotype that involves the co-effect of multiple genes (Vaupel and Tan 1997; Martin 1997). It will be interesting and necessary to discover whether the genes function together, independently, and/or dependently. If there is any dependency, then the biological significance can be ascertained. Given the fact that there is usually a considerable amount of polymorphism at each locus, a better strategy is to combine the simple gene frequency method with our new approach. In this way, possible interactions can be screened by simply comparing frequencies among different age groups and then examining them carefully and in more detail afterwards by applying the new methods.

Since some of the genes have different frequencies in different geographic areas (THO8, INS+, mtDNAhapl-U, mtDNAstr-136, mtDNAstr-138), ignoring the regional differences might introduce bias into the estimates of the risks associated with them. In a simple simulation, we assume that one gene is neutral, with risk $r = 1$, but that there are different gene frequencies in the south (0.1) and in the north (0.2). We also assume that the area risk for being from the south is 1.5. The estimated risk r for the gene is 0.95 when we impose the same gene frequency for the two areas. This bias stems from an overestimated gene frequency in the south and an underestimated frequency in the north (0.15 for both areas if one assumes that 50% are from the south at birth). Due to the higher risk of death in the south (1.5 to 1), there are more people from the north who reach old age. This high proportion of northerners at old ages is artificially related to a lower risk of the gene occurring when the frequency of the gene is underestimated in the north.

Another important aspect of this application is the introduction of unobserved individual heterogeneity. Its influence on the estimates of risks is explicitly demonstrated in Tables 4 and 5. The risks are systematically underestimated and thus lead to conservative conclusions when heterogeneity is not taken into consideration. When heterogeneity is introduced, however, this dramatically improves the likelihood of the estimation and thus produces better estimates (Figure 2). The likelihood values from heterogeneity and from homogeneity models are comparable because σ^2 in the heterogeneity model is set to different values, whereupon the parameters are then estimated. The number of parameters estimated does not change at all, regardless of whether or not one considers heterogeneity. However, it is not true that all sub-populations had the same variance in their unobserved frailties. The one-σ^2 model offers a feasible way of including heterogeneity with a limited sample size. As a consequence, hazard functions for different sub-populations merely converge and do not cross – which may not necessarily be the case. On the other hand, the convergence phenomenon (Figures 3, 4, 5) raises an important question regarding the influence of genes on survival at very old ages. It seems that a certain gene becomes unimportant as the hazards of death for populations with and without it converge. As we know, the risk associated with the gene, as it is assumed, does not change with age at the individual level since we are using a proportional hazard model as described by Cox (1972). The convergence is almost certainly due to unobserved heterogeneity, which compensates for the genetic effect as selection continues with increasing age in a heterogeneous population with the same genotype (Vaupel and Yashin 1985). We hope that this problem can be addressed in more detail when more data are available. In addition, the assumption of gamma-distributed frailty is only an arbitrary condition for identifiability and mathematical convenience. One could assume that our model could be sensitive to the assumptions of the distribution of frailty. However, it was shown recently (Yashin et al. 1999b) in a comparative analysis of different frailty models that the gamma-frailty model is rather flexible for working with survival data.

Furthermore, one must note that all risk parameters in the model are defined as being multiplicatively proportional to the baseline hazard, as is the case in the Cox model. Like any other assumptions involved in defining a model, the proportional hazard approach may not reflect the real situation, especially for the genetic parameters. As a result, some special patterns of genetic influence on survival that deviate from being constant could be missed. However, the simple assumption can serve as the first step in solving the problem.

One concern is that adjustments of the significance level for the statistical tests might be needed since we are doing multiple comparisons. However, our interest here is focused on each gene allele separately, so we test multiple hypotheses rather than performing multiple tests of a single hypothesis. As is often done, such adjustment is called for (De Benedictis et al. 1999; Weir 1996; Rothman 1990) in the latter situation. If one is interested in making an overall conclusion on a single locus with multi-alleles, then each test on one allele can be treated as one instance of a repeated test that contributes to the final result on the hypothesis on the locus. Adjustment is required in this case because the existence of any significant allele will result in a positive conclusion.

The fact that significant genes were discovered in the present study is not surprising since the candidate genes selected play central roles in crucial metabolic pathways. The APOB gene variations could affect the efficiency in cholesterol metabolism and thus associate with individual's susceptibility to coronary artery disease (Hegele et al. 1986; Myant et al. 1989; Paulweber et al. 1989, 1990; Kervinen et al. 1994) and survival. The significant effects of THO and INS alleles could be relevant to the complex relationship existing between insulin and catecholamins (Natali et al. 1998) in glucose metabolism, whose regulation in turn affects life span from yeast (Jiang et al. 2000) to humans (Paolisso et al. 1996). The beneficial effect of SOD2-T allele could support the finding that SOD2 polymophisms affect the efficiency of mitochondrial transport (Shimoda-Matsubayashi et al. 1996). Lastly, the biological background for the association between mtDNA variation and longevity is probably relevant to mtDNA haplogroup-specific oxidative phosphorylation efficiency (Ruiz-Pesini et al. 2000).

The application of this model on data collected from genetic studies on aging and longevity should help to detect additional relevant genes that contribute to the process of aging both by prolonging or shortening an individual's life span.

Acknowledgements

This work was partly supported by the Italian Ministero Università Ricerca Scientifica Technologica 1998–2000 project, 'Longevity Determinants in Humans: The Model of Centenarians' by Istituto Nazionale Riposo e Cura Anziani, Ancona, Italy, and by grant P01 AG08761-01 from the National Institute of Health/National Institute on Aging. The authors are thankful to Prof. Kaare Christensen from Southern Denmark University (Odense University) for useful discussions and to Dr Karl Brehmer for help in preparing the paper.

References

Aalen O (1988) Heterogeneity in survival analysis. Stat Med 7: 1121–1137

Andrieu N and Goldstein AM (1999) Epidemiologic and genetic approaches in the study of gene–environment interaction: an overview of available methods. Epidemiol Rev 20: 137–147

Aptech System (1996) Gauss: Mathematical and Statistical System. Vol I: System and Graphics manual. Aptech Systems, Maple Valley, Washington

Arking R (1987) Genetic and environmental determinants of longevity in Drosophila. Basic Life Sci 42: 1–22

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: 427–430

Bennett WP, Hussain SP, Vahakangas KH, Khan MA, Shields PG and Harris CC (1999) Molecular epidemiology of human cancer risk: gene–environment interactions and p 53 mutation spectrum in human lung cancer. J Pathol 187: 8–18

Bladbjerg EM, Andersen-Ranberg K, de Maat MP, Kristensen SR, Jeune B, Gram J and Jespersen J (1999) Longevity is independent of common variations in genes associated with cardiovascular risk. Thromb Haemost 82: 1100–1105

Bocquet-Appel JP and Jakobi L (1990) Familial transmission of longevity. Ann Hum Biol 17: 81–95

Bonafe M, Olivieri F, Mari D, Baggio G, Mattace R, Sansoni P, De Benedictis G, De Luca M, Bertolini S, Barbi C, Monti D and Franceschi C (1999a) p 53 variants predisposing to cancer are present in healthy centenarians. Am J. Hum Genet 64: 292–295

Bonafe M, Olivieri F, Mari D, Baggio G, Mattace R, Berardelli M, Sansoni P, De Benedictis G, De Luca M, Marchegiani F, Cavallone L, Cardelli M, Giovagnetti S, Ferrucci L, Amadio L, Lisa R, Tucci MG, Troiano L, Pini G, Gueresi P, Morellini M, Sorbi S, Passeri G, Barbi C and Valensin S (1999b) p 53 codon 72 polymorphism and longevity: additional data on centenarians from continental Italy and Sardinia. Am J Hum Genet 65: 1782–1785

Cavalli Sforza LL, Menozzi P and Piazza A (eds) (1994) The History and Geography of Human Genes, pp 277–280. Princeton University Press, Princeton, New Jersey

Chen J, Giovannucci EL and Hunter DJ (1999) MTHFR polymorphism, methyl-replete diets and the risk of colorectal carcinoma and adenoma among US men and women: an example of gene–environment interactions in colorectal tumorigenesis. J Nutr 129(2S Suppl): 560S–564S

Christensen K and Vaupel JW (1996) Determinants of longevity: genetic, environmental and medical factors. J Intern Med 240: 333–341

Clare MJ and Luckinbill LS (1985) The effects of gene–environment interaction on the expression of longevity. Heredity 55: 19–26

Cox DR (1972) Regression models and life-tables. J R Stat Sco B 34: 187–220

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, Rose G, Carrieri G, De Luca M, Falcone E, Passarino G, Bonafè M, Monti D, Baggio G, Bertolini S, Mari D, Mattace R and Franceschi C (1999) Mitochondrial DNA inherited variants are associated with successful aging and longevity in humans. FASEB J 13: 1532–1536

Gavras I, Manolis A and Gavras H (1999) Genetic epidemiology of essential hypertension. J Hum Hypertens 13: 225–229

Grube K and Burkle A (1992) Poly(ADP-ribose) polymerase activity in mononuclear leukocytes of 13 mammalian species correlates with species-specific life span, Proc Natl Acad Sci USA 89(24): 11759–11763

Harris JR, Lippman ME, Veronesi U and Willett W (1992) Breast cancer (1). N Engl J Med 327: 319–328

Hastings N and Peacock JB (1974) Statistical Distributions. Butterworths, London, 90 pp

Hazzard WR (1986) Biological basis of the sex differential in longevity. J Am Geriatr Soc 34: 455–471

Hegele RA, Huang LS, Herbert PN, Blum CB, Buring JE, Hennekens CH and Breslow JL (1986) Apolipoprotein B-gene DNA polymorphism associated with myocardial infarction. N Engl J Med 315(24): 1509–1515

Herskind AM, McGue M, Holm NV, Sorensen TI, Harvald B and Vaupel JW (1996) The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870–1900. Hum Genet 97: 319–323

Holden C (1987) Why do women live longer than man? Science 238: 158–160

Hougaard P (1991) Modeling heterogeneity in survival analysis. J Appl Prob 28: 695–701

ISTAT (1997) Annuario statistico italiano, pp 48–49. ISTAT, Rome

Jiang JC, Jaruga E, Repnevskaya MV, Jazwinski SM (2000) An intervention resembling caloric restriction prolongs life span and retards aging in yeast. FASEB J 14: 2135–2137

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

Kannisto V (1994) Development of oldest-old mortality, 1950–1990: evidence from 28 developed countries, pp 59–66. Odense University Press, Odense, Denmark

Kervinen K, Savolainen MJ, Salokannel J, Hynninen A, Heikkinen J, Ehnholm C, Koistinen MJ, Kesaniemi YA (1994) Apolipoprotein E and B polymorphisms – longevity factors assessed in nonagenarians, Atherosclerosis 105(1): 89–95

Keyfitz N and Flieger W (1990) World Population Growth and Aging. University of Chicago Press, Chicago

Martin GM (1997) Genetics and the pathobiology of ageing. Philos Trans R Soc London B Biol Sci 1997 352: 1773–1780

Martin N (2000) Gene–environment interaction and twin studies. In: Spector TD, Snieder H, MacGregor (eds) Advances in Twin and Sib-Pair Analysis, pp 143–150. Oxford University Press, London

MatchSoft. (1996) Axum Technical Graphics and Data Analysis. Cambridge, Massachusetts

McGue M, Vaupel JW, Holm N and Harvald B (1993) Longevity is moderately heritable in a sample of Danish twins born 1870–1880. J Gerontol 48: B237–244

Natali A, Gastaldelli A, Galvan AQ, Sironi AM, Ciociaro D, Sanna G, Rosenzweig P and Ferrannini E (1998) Effects of acute alpha 2-blockade on insulin action and secretion in humans. Am J Physiol 274: E57–E64

Paolisso G, Gambardella A, Ammendola S, D'Amore A and Varricchio M (1996) Glucose tolerance and insulin action in healthy centenarians. Am J Physiol 270: E890–E896

Paulweber B, Friedl W, Holzl B, Sandhofer F (1989) Genetics of coronary heart disease. Lancet 2(8659): 384

Paulweber B, Friedl W, Krempler F, Humphries SE, Sandhofer F (1990) Association of DNA polymorphism at the apolipoprotein B gene locus with coronary heart disease and serum very low density lipoprotein levels, Arteriosclerosis 10(1): 17–24

Rothman KJ (1990) No Adjustments are needed for multiple comparisons. Epidemiology 1: 43–46

Ruiz-Pesini E, Lapena AC, Diez-Sanchez C, Perez-Martos A, Montoya J, Alvarez E, Diaz M, Urries A, Montoro L, Lopez-Perez MJ and Enriquez JA (2000) Human mtDNA haplogroups associated with high or reduced spermatozoa mobility. Am J Hum Genet 67: 682–696

Sambrook P and Nguyen T (1999) Bone mineral density and gene–environment interactions in the search for osteoporosis genes. Environ Health Perspect 107: A130-A131

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. Nature Genetics 6: 29–32

Sellers TA, Weaver TW, Phillips B, Altmann M and Rich SS (1998) Environmental factors can confound identification of a major gene effect: results from a segregation analysis of a simulated population of lung cancer families. Genet Epidemiol 15: 251–262

Shimoda-Matsubayashi S, Matsumine H, Kobayashi T, Nakagawa-Hattori Y, Shimizu Y and Mizumo Y (1996) Structural dimophism in the mitochondrial targeting sequence in the human manganese superoxide dismutase gene. Biochem Biophys Res Comm 226561–226565

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: 1525–1534

Vaupel JW and Tan Q (1998) How many longevity genes are there? Paper presented at annual meeting of Population Association of America. Chicago

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

Vaupel JW, Carey JR, Christensen K, Johnson TE, Yashin AI, Holm NV, Iachine IA, Kannisto V, Khazaeli AA, Liedo P, Longo VD, Zeng Y, Manton KG and Curtsinger JW (1998) Biodemographic trajectories of longevity. Science 280: 855–860

Vaupel JW, Manton KG and Stallard E (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16: 439–454

Weir BS (1996) Genetic Data Analysis II, pp 133–135. Sinauer Associates, Massachusetts

Yashin AI, Iachine IA and Harris JR (1999) Half of the variation in susceptibility to mortality is genetic: findings from Swedish twin survival data. Behav Genet 29: 11–19

Yashin AI, Manton KG and Vaupel JW (1985) Mortality and aging in a heterogeneous population: a stochastic process model with observed and unobserved variables. Theor Popul Biol 27: 154–175

Yashin AI, De Benedictis G, Vaupel JW, Tan Q, Andreev KF, Iachine IA, Bonafe M, DeLuca M, Valensin S, Carotenuto L and Franceschi C (1999a) 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, Begun AZ and Iachine IA (1999b) Genetic factors in susceptibility to death: comparative analysis of bivariate survival models. J Epidemiol Biostat. 7: 223–224

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 (2000) Genes and longevity: Lessons from studies on centenarians. J Gerontol 55a: B1–B10

Zhang JG, Ma YX, Wang CF, Lu PF, Zhen SB, Gu NF, Feng GY and He L (1998) Apolipoprotein E and longevity among Han Chinese population. Mech Aging Dev 104: 159–167