Journal of Epidemiology and Biostatistics (1998) Vol. 3, No. 3, 289–294

Combining genetic and demographic information in population studies of aging and longevity

AI YASHIN[1,2], JW VAUPEL[1,3], KF ANDREEV[1], Q TAN[1], IA IACHINE[4], L CAROTENUTO[5], G DE BENEDICTIS[6], M BONAFE[7], S VALENSIN[8] and C FRANCESCHI[7,8]

[1]*Max Planck Institute for Demographic Research, Rostock, Germany*
[2]*Duke University Centre for Demographic Studies, USA*
[3]*Sanford Institute, Duke University, USA*
[4]*Medical School, Odense University, Denmark*
[5]*System Science Department University of Calabria, Italy*
[6]*Cell Biology Department, University of Calabria, Italy*
[7]*Istituto Nazionale Riposo e Cura Anziani, Ancona, Italy*
[8]*Biomedical Science Department University of Modena, Italy*

Background Some genes may play a more important role in human longevity than others. Genetic markers data from cross-sectional population studies are collected for testing hypotheses about the contribution of 'candidate' genes to longevity.

Method A new method, based on the evaluation of relative risks for individuals carrying candidate alleles, is applied to the combination of genetic and demographic data.

Results Candidate genes from six loci (five nuclear loci and mitochondrial DNA) are categorised as frailty, longevity and neutral genes. Area- and sex- related relative risks are evaluated, as well as allele-specific risks.

Conclusion The comparison between the relative risk method and gene frequency method shows that the relative risk method seems to be more robust; its classification is based on estimates of hazards rates, rather than on comparisons between gene frequencies.

Keywords gene frequencies, genetics of longevity, aging, candidate alleles, survival.

Introduction

In genetic studies of human aging and survival, the contribution of candidate genes in the survival process is analysed by using methods based on gene frequency[1,2]. In one strategy used currently to identify genes in multifactorial diseases, allele pools from sample group of extremely old individuals (cases) and younger people (controls) from the same population are compared. The observed case/control differences in allele frequencies are associated with the influence of a respective candidate gene on survival[3–8]. Standard statistical methods which identify differences in observed frequencies between case and control groups for different candidate alleles are used to make proper classifications. For example, they may include methods of multiple comparison, based on Bonferroni inequality method or Sheffe's method. By such an approach, frailty and longevity genes, which decrease and increase respectively the chances of survival, can be identified.

Although very useful, the gene frequency (GF) method does not permit us to evaluate important characteristics, such as relative risks (RR) and survival functions, for individuals carrying a certain allele. Furthermore, this method implicitly assumes that:

- The initial proportions of genotypes in all cohorts represented in cross-sectional study are the same.
- The survival functions of genotypes do not depend on the birth year of the cohorts.

Although the first assumption is well established, the second assumption is controversial. Indeed, due to trends in social, economic and living conditions, different cohorts of individuals experience different environments. Consequently, the birth year of the cohort could affect the survival functions of a certain genotype.

The use of demographic information, together with data on genetic markers, opens a new avenue in cross-sectional genetic studies. This strategy is realised in the RR method suggested and explored in this paper. The method is based on constraint maximisation of the likelihood function of genetic data and allows for the integration of genetic and demographic information. Although both GF and RR methods give a similar

Correspondence to: Anatoli Yashin, Max Planck Institute for Demographic Research, Doberaner Strasse 114, 18057 Rostock, Germany.

classification of candidate genes, the RR method provides more opportunities for the analysis of genetic data. It permits us to calculate initial proportions, RR and survival distributions for individuals with respective genes from cross-sectional data. It allows us to analyse the sensitivity of the results to violations of the assumptions used in the GF method. The influence of secular trends in cohort survival of genotypes, as well as the effects of possible differences in the initial allele proportions in populations represented by different birth cohorts on the results of the analysis, are discussed. The quality of the estimation procedure is tested with simulated data.

The data

Five autosomal loci (*APOB, REN, D21S, SOD2, THO*) and the mitochondrial locus (*mtDNA*) were considered. The polymorphic systems were the following: *3'APOB-VNTR* (15 alleles[9]), *HUMREN 4* (five alleles[10]), *D21S223* (nine alleles[11]), *SOD2* (two C/T alleles[12]), *HUMTHO1* (six alleles[10,13]), *mtDNA* haplogroups (nine alleles[14]).

The data on genetic markers for the group of centenarians (with ages above 100) and the group of younger individuals (with ages 5–80) were obtained from samples collected in both northern and southern Italy. Altogether, 662 individuals were involved in the study, among them 54 male and 143 female centenarians and 220 male and 245 female younger individuals. 26 male centenarians were from northern Italy and 28 were from southern Italy. The number of female centenarians from the north was 83 and from the south 60. The younger group contained 75 males and 87 females from the north and 145 males and 158 females from the south. The ages of the subjects ranged from 5 to 109 years (5–19 year olds were school-children, 20–29 year olds were university undergraduate and graduate students, the subjects over 100 years old were gathered into a larger research project in progress in Italy, the others were volunteer donors). The samples were collected by eight institutions and research groups in Italy. For technical reasons, the number of individuals participating in the analysis of some loci is less than mentioned above.

RR method

The changes in gene frequencies with age within one cohort are produced by differences in survival functions (risk of death) associated with respective genes. This property suggests a strategy for identification of frailty and longevity alleles. Instead of comparing gene frequencies between centenarians and younger individuals, one can evaluate and compare RR and survival distributions associated with different alleles. These characteristics can be identified when additional information on survival in respective cohorts is available. Such information can be taken from standard demographic life tables. Observed risk factors, such as area and sex, may also be included in the likelihood. For example, the hazard rate for an x years old individual may be represented in a Cox-form[15] as $\mu_0(x)e^{\sum_{i=1}^{3}\beta_i U_i}$, where variable U_1 refers to region (0 for the North and 1 for the South), variable U_2 refers to the sex (0 for female and 1 for male), variable U_3 refers to the presence ($U_3 = 1$) or absence ($U_3 = 0$) of the candidate allele on a chromosome, and variable U_4 refers to the presence ($U_4 = 1$) or absence ($U_4 = 0$) of the same allele on the homologous chromosome. Thus, the survival function of an x years old individual can be represented as $S_0(x)^{r_i r_j e^{\beta_1 U_1 + \beta_2 U_2}}$, where $r_i = e^{\beta_3 U_3}$, and $S_0(x)=e^{-\int_0^x \mu_0(s)ds}$. In our study we assume that $\beta_3 = \beta_4$, so the relative risk in individuals homozygous for the candidate allele is RR_3^2, where $RR_3 = e^{\beta_3}$ is the risk in individuals heterozygous for the candidate allele. The RR method can control the situation when the survival functions of genotypes depend on the birth year of the cohorts. Note that Cox's partial likelihood method cannot be used here since all data are censored.

Results

The results of the data analysis are summarised in Table 1. The RR for the area ($RR_1 = e^{\beta_1}$, South/North), the sex ($RR_2 = e^{\beta_2}$, male/female) and the candidate allele ($RR_3 = e^{\beta_3}$, presence/absence) are shown, together with respective 95% confidence intervals (CI), estimated initial allele frequencies and p values. For example, the second line in this table shows that the chances of death for female individuals from southern Italy not carrying allele 1 at the *D21S* locus, are 1.08 times higher than for the same-sex individuals from the North; and the p value for testing risk difference in the area is 0.0013. In the North, males without this allele are 1.18 more likely to die than females without this allele and the p value for testing risk difference in gender is <0.001. (Since all p values for the risk related to gender are <0.001, they are not shown in the table.) The initial frequency of allele 1 at *D21S* locus is 0.02. The chances of death for a female from the North with one allele 1 at the *D21S* locus are 85% of the respective chances of death for a female from the same area without this allele. The p value for testing the difference in risks between those with and without allele is 0.0025. The 95% CI for respective risks are shown in parentheses under the estimates. Both p values and CI are calculated using the bootstrap procedure, with 100 repeats. The '+' sign means that this is a longevity allele. The '–' sign refers to the frailty allele. The CI for the risks associated with the neutral alleles include value 1. Such alleles are not shown in the table. For example, *SOD2* locus is not rep-

Table 1 Estimated RR for the area, sex and candidate allele from the data on Italian genotypes with 95% CI: only frailty (–) and longevity (+) alleles are shown

Allele	Area		Sex		Allele		
	RR_1=S/N	p value	RR_2=M/F	Initial frequency	RR_3 with/ without	p value	Frailty longevity
APOB-31	0.92 (0.85–0.98)	0.0119	1.27 (1.18–1.36)	0.17	1.13 (1.06–1.19)	0.0001	–
D21S-1	1.08 (1.03–1.13)	0.0013	1.18 (1.16–1.24)	0.02	0.85 (0.75–0.95)	0.0025	+
D21S-6	1.10 (1.04–1.16)	0.0011	1.17 (1.12–1.23)	0.07	1.10 (1.03–1.18)	0.0068	–
mtDNA HAPL-J	1.11 (1.06–1.15)	0.0000	1.19 (1.13–1.24)	0.05	0.84 (0.75–0.92)	0.0002	+
mtDNA HAPL-V	1.12 (1.07–1.16)	0.0000	1.19 (1.12–1.26)	0.01	0.72 (0.49–0.96)	0.0224	+
REN-8	1.19 (1.12–1.27)	0.0000	1.22 1.16–1.29)	0.73	0.93 (0.89–0.97)	0.0009	–
REN-11	1.20 (1.13–1.26)	0.0000	1.22 (1.14–1.29)	0.16	1.11 (1.03–1.18)	0.0039	–
THO-10	1.19 (1.14–1.24)	0.0000	1.14 (1.09–1.19)	0.20	0.95 (0.91–0.98)	0.0044	+

resented in the table since all its alleles are found to be neutral.

The data used in the present paper were already used to identify gene/longevity associations by the GF method by De Benedictis *et al.*[8]. The following results were obtained. Of the six analysed loci (*APOB*, *D21S*, *REN*, *THO*, *SOD2*, *mtDNA*), three (*APOB*, *THO* and *mtDNA*) were found to affect life expectancy. As for *APOB*, pooled alleles with less than 35 repeats were found to be frailty alleles[8] and this result agrees with that obtained by the RR method (*APOB-31* allele is frailty). With *THO*, the 10 repeat allele was classified as a longevity allele by both the GF approach (De Benedictis *et al.*, unpublished observations) and the RR approach. With *mtDNA*, both the GF approach (De Benedictis *et al.* unpublished observations) and the RR method showed that J and V haplogroups are longevity alleles. As for the *D21S* and *REN* loci, while the GF method did not reveal an effect of these loci on survival, the RR method did reveal frailty and longevity alleles at these loci, thus indicating that the RR method is more reasonable than the GF method. Lastly, neither method revealed a gene/longevity association at the *SOD2* locus.

Discussion

The analyses show that both methods can be used for classification of genes as robust, frail and neutral. However, the RR method gives a straightforward and comprehensive approach, which requires neither complex heterogeneity tests between samples relevant to sex or to geographic areas, nor multiple comparisons.

Furthermore, the use of the GF method may result in some spurious allele classifications because of age-related non-monotonic trajectories of gene frequencies, as the following example shows. Suppose that there are three alleles in the population with the initial proportions p_0={0.1, 0.3, 0.6} and relative risks RR = {0.5, 1.0, 2.0} respectively. Assume that survival is driven by Gompertz mortality ae^{bt} with a = 0.001 and b = 0.04. The dynamics of allele proportions is given by Figure 1.

The proportion of individuals with neutral (RR = 1.0) allele first increases, reaching its maximum at age about 105 years and then declines. If samples of centenarians and controls are taken from this population, then the neutral gene may mistakenly be classified as a longevity gene by the GF method. Such an error will not happen if the classification is based on estimates of RR. The RR method is focused on estimation and comparison of RR rather than allele frequencies. Hence, if the estimate of RR associated with some allele is close to 1.0, this allele will be classified as a neutral one.

In addition, such important characteristics as initial frequencies and survival functions for individuals with

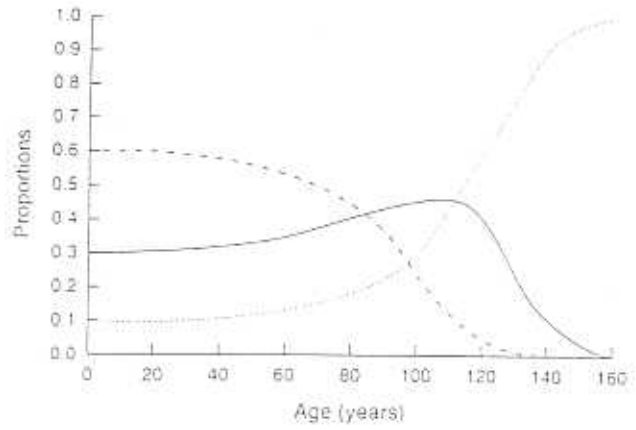

Fig. 1 The dynamics of gene frequencies by age for the allele with RR(1) = 0.5 and initial frequency $p_0(1) = 0.1$ (dashed line); for the allele with RR(2) = 1.0 and initial frequency $p_0(2) = 0.5$ (solid line); and for the allele with RR(3) = 2.0 and initial frequency $p_0(3) = 0.6$ (dotted–dashed line).

specific alleles can be evaluated. In the RR method the data on epidemiological risk factors (observed covariates) can be analysed, together with genetic data. Several other methodological difficulties of the GF method can also be avoided. In particular, the RR method has enough flexibility to describe the survival of homozygous individuals. It can take into account the exact age of individuals in a sample, thus making unnecessary the aggregation of individuals into two groups with inevitable loss of information.

Initial frequency

Among factors which may contribute to an erroneous classification of candidate alleles we investigated the influence of possible changes in initial allele frequencies and secular trends in mortality. Let us assume, for instance, that the initial frequency $p_0(t)$ of the robust allele in the cohort born in year t decreases with time. As a result, the rate of increase of respective frequency with age observed in a cross-sectional study goes up. If, on the contrary, $p_0(t)$ increases with time, then the rate of increase in respective frequency with age observed in a cross-sectional study goes down. Note that meaningful changes in the initial gene frequencies in different cohorts may be caused by migration flows, which may bias the results of analysis in a cross-sectional study. Historical demographic analysis should accompany such a genetic study to avoid confusion.

Secular trends in mortality

Due to changes in social, economic and living conditions different cohorts of individuals experience different environmental influence. These conditions may affect the chances of survival of individuals with different alleles, thereby violating the assumption that survival functions of genotypes do not depend on the birth year of the cohort. This assumption is crucial in the GF method. To calculate the effects of such secular trends in human survival on gene frequencies at age x, measured in year T, we assume that observed survival trends are solely attributed to the changes in the baseline survival for longevity alleles, and that RR are the same for each cohort. Note that, in accordance with demographic data, the annual rate of improvement in survival is age-dependent. Our simulation–estimation studies show that the age-specific curve of annual rate of mortality improvement seems to have little influence on the frequency of individuals carrying longevity alleles in the cohorts of different age. Hence, the classification of genes into frail, robust and neutral is not sensitive to such dependence. Moreover, analyses based on the use of average survival functions for several cohorts and on average survival functions for synthetic cohorts, which correspond to period life-table data, give the same results of allele classification.

Simulation studies show that, unless the risk parameter is very high, an appreciable change in the frequency occurs only in very old cohorts: this result justifies the relevance of studying groups of centenarians. Estimation procedures, performed using simulated data with different sample sizes, show that all parameters in this model are statistically identifiable.

Despite evident progress in establishing connection between specific genes and longevity[16], some aspects of genetic studies using cross-sectional data deserve additional efforts. For example, to better understand the results of the joint influence of several genes on survival, methods for simultaneous analysis of data on multilocus genotypes need to be developed. To better understand the gene–environment interaction, the use of additional demographic information (e.g. mortality by cause) may be helpful. New survival models are needed to combine genetic and epidemiological data in the presence of random effects (hidden heterogeneity in mortality). New approaches are needed to describe the influence of unobserved genes (i.e. genes which are not included in the list of candidates) on age-trajectories of gene frequencies.

Acknowledgments

This work was supported by grants given to GDB from Region Calabria (POP 94–99) and from INRCA (Ancona, Italy). The authors thank anonymous reviewers for valuable comments.

Appendix

We have $N = 12$ groups of individuals specified by two areas (0 for North and 1 for South), two sexes (0 for

female and 1 for male) and three genotypes (0 for the absence of the candidate allele, 1 for the presence of this allele in one chromosome and 2 for the presence of this allele at both chromosomes). Let p_{0n} be the initial proportion of individuals from northern Italy, p_{0f} be the initial proportion of females and p_{0g} be the initial frequency of the candidate allele in a population. Let $p_i(x,T-x)$, $i = 1,2,...,N$ be the proportion of x years old individuals from i^{th} group in some cross-sectional study performed in year T, and let $N_{xi}(T-x)$ be respective numbers of individuals observed in this study. Then the likelihood function of the data is

$$L \sim \prod_{x} \prod_{i=1}^{N} p_i(x,T-x)^{N_{xi}(T-x)} \qquad (1)$$

here $p_N(x,T-x) = 1 - \sum_{i=1}^{N-1} p_i(x,T-x)$, and

$$p_i(x,T-x) = \frac{p_{0i}(T-x)S_{0i}(x,T-x)}{\sum_{j=1}^{N} p_{0j}(T-x)S_{0j}(x,T-x)} \qquad (2)$$

where $p_{0i}(T-x)$ represents the initial proportion of individuals in group i. Note that each of these 12 initial proportions may be represented in terms of the three parameters p_{0n}, p_{0f} and p_{0g} specified above. For each i=1,2,...12, the survival function $S_{0i}(x, T-x)$ is given in terms of the Cox proportional hazard model (1972) with conditional hazards $\mu(x,U_1,U_2,U_3,U_4) = \mu_0(x)e^{\sum_{i=1}^{4}\beta_i U_i}$ (we assume here that $\beta_3 = \beta_4$) and with the respective combination of values for U_1, U_2, U_3 and U_4.

Estimation procedure

The likelihood (1) must be maximised with respect to parameters p_{0n}, p_{0f}, p_{0g}, $S_0(x,T-x)$, $x = x_0, x_1, \ldots X$ [i.e. survival function associated with $\mu_0(x)$] and risks $RR_i = e^{\beta_i}$, i = 1, 2, 3, 4, under constraint

$$S(x,T-x) = \sum_{j=1}^{N} p_{0j}(T-x)S_{0j}(x,T-x) \qquad (3)$$

The values of survival functions $S(x,T-x)$ are taken from the official demographic life tables for Italian population. The values of $S_{0j}(x,T-x)$ depend on $S_{0j}(x,T-x) = e^{-\int_{x_0}^{x}\mu_{0j}(u)du}$ and $RR_i = e^{\beta_i}$, i = 1, 2, 3, 4. For example, let us consider one of the 12 groups for the *APOB* locus (group 'k'). Assume that this number refers to the group of males from the south of Italy which have one allele *APOB*-31. This group is characterised by survival function $S_{0k}(x,T-x) = S_0(x,T-x)^{RR_1RR_3RR_4}$. Respective initial frequency will be $p_{0k}(T-x) = 2(1-p_{0n})(1-p_{0f})p_{0g}(1-p_{0g})$. The group 'j' of females from the north of Italy which have one *APOB*-31 allele has survival function $S_{0j}(x,T-x) = S_0(x,T-x)^{RR_3}$ and initial frequency $p_{0j}(T-x) = 2p_{0n}p_{0f}p_{0g}(1-p_{0g})$. Similar representations can be written for each of the other 10 groups represented in the likelihood (1). The estimation procedure which takes into account constraint (3) starts with maximization of the likelihood (1) with respect to initial proportions p_{0n}, p_{0f}, p_{0g} and risks RR_i, i = 1, 2, 3, 4 taking an initial estimate of $S_0(x,T-x)$ equals, for example, $S(x,T-x)$ (which is known). Then the estimates of p_{0n}, p_{0f}, p_{0g} and risks RR_i, i = 1, 2, 3, 4 are substituted into (3), from where the second estimate of $S_0(x,T-x)$ is calculated. This is substituted in equation (1) with unknown parameters p_{0n}, p_{0f}, p_{0g} and RR_i, i = 1, 2, 3, 4. The likelihood (1) is maximized again to produce a second guess at these parameters and the procedure is repeated until convergence.

References

1 Schachter F, Cohen D, Kirkwood T. Prospects for the genetics of human longevity. *Hum Gen* 1993;91:519–26.

2 De Benedictis G. Genes and longevity. *Aging Clin Exp Res* 1996;8:367–9.

3 Proust J, Moulias R, Fumeron F *et al.* HLA and longevity. *Tissue Antigens* 1982;19:168–73.

4 Takata H, Suzuki M, Ishii T *et al.* Influence of major histocompatibility complex region genes on human longevity among Okinawan-Japanese centenarians and nonagenarians. *Lancet* 1987;2:824–6.

5 Kervinen K, Savolainen MJ, Sallokannen J *et al.* Apolipoprotein E and B polymorphisms — longevity factors assessed in nonagenarians. *Atherosclerosis* 1994;105:89–95.

6 Louhija J, Miettinen HE, Kontula K *et al.* Ageing and genetic variation of plasma apolipoproteins. Relative loss of the apolipoprotein E4 phenotype in centenarians. *Atheroscler Thromb* 1994;14:1084–9.

7 Schachter F, Faure-Delanef L, Guenot F et al. Genetic association with human longevity at the *APOE* and *ACE* loci. *Nature Gen* 1994;6:29–32.

8 De Benedictis G, Falcone E, Rose G *et al.* DNA multiallelic systems reveal gene/longevity associations not detected by diallelic systems. The *APOB* locus. *Hum Gen* 1997; 99:312–18.

9 Boerwinkle E, Xiong W, Fourest E, Chan L. Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: application to the apolipoprotein B3 hypervariable region. *Proc Natl Acad Sci USA* 1989;86:212–16.

10 Edwards A, Hammond HA, Jin L *et al.* Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* 1992;12:241–53.

11 Rosen DR, Sapp PC, O'Regan J *et al.* Dinucleotide repeat polymorphisms (*D21S223* and *D21S224*). *Hum Mol Gen* 1992;1:547.

12 Rosenblum JS, Gilula NB, Lerner RA. On signal sequence polymorphisms and diseases of distribution. *Proc Natl Acad Sci USA* 1996;93:4471–3.

13 Puers C, Hammond HA, Jin L *et al.* Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus *HUMTHO1 (AATG)n* and reassignment of alleles in population analysis by using a locus specific allelic ladder. *Am J Hum Gen* 1993;53:953–8.

14 Torroni A, Huoponen K, Francalacci P *et al.* Classification of European *mtDNAs* from an analysis of three European populations. *Genetics* 1996;144:1835–1850.

15 Cox DR. Regression models and life-tables [with discussion]. *J Roy Stat Soc B* 1972;34:187–220.

16 Jazwinski SM. Longevity, genes and aging. *Science* 1996; 273:54–9.