Original Paper

Human Heredity

Hum Hered 2005;59:88–97
DOI: 10.1159/000085223

Received: October 13, 2004
Accepted after revision: February 23, 2005
Published online: April 18, 2005

Estimating Haplotype Relative Risks on Human Survival in Population-Based Association Studies

Qihua Tan[a, b] Lene Christiansen[b] Lise Bathum[a, b] Jing Hua Zhao[c]
Anatoli I. Yashin[d] James W. Vaupel[e] Kaare Christensen[b] Torben A. Kruse[a]

[a]Department of Clinical Biochemistry and Genetics, KKA, Odense University Hospital, and [b]Institute of Public Health, University of Southern Denmark, Odense, Denmark; [c]Department of Epidemiology and Public Health, University College London, London, UK; [d]Center for Demographic Studies, Duke University, Durham, N.C., USA; [e]Max Planck Institute for Demographic Research, Rostock, Germany

Key Words
Population-based association analysis · Haplotype relative risk · Human survival · Unphased genotype data

Abstract
Association-based linkage disequilibrium (LD) mapping is an increasingly important tool for localizing genes that show potential influence on human aging and longevity. As haplotypes contain more LD information than single markers, a haplotype-based LD approach can have increased power in detecting associations as well as increased robustness in statistical testing. In this paper, we develop a new statistical model to estimate haplotype relative risks (HRRs) on human survival using unphased multilocus genotype data from unrelated individuals in cross-sectional studies. Based on the proportional hazard assumption, the model can estimate haplotype risk and frequency parameters, incorporate observed covariates, assess interactions between haplotypes and the covariates, and investigate the modes of gene function. By introducing population survival information available from population statistics, we are able to develop a procedure that carries out the parameter estimation using a nonparametric baseline hazard function and estimates sex-specific HRRs to infer gene-sex interaction. We also evaluate the haplotype effects on human survival while taking into account individual heterogeneity in the unobserved genetic and nongenetic factors or frailty by introducing the gamma-distributed frailty into the survival function. After model validation by computer simulation, we apply our method to an empirical data set to measure haplotype effects on human survival and to estimate haplotype frequencies at birth and over the observed ages. Results from both simulation and model application indicate that our survival analysis model is an efficient method for inferring haplotype effects on human survival in population-based association studies.

Introduction

Although multidisciplinary approaches have been used in the search of genes implicated in human aging and longevity [1], the association-based linkage disequilibrium (LD) mapping exhibits more power than the linkage-based methods [2], a situation that mimics the mapping of complex or non-Mendelian disease genes [3]. With the completion of the human genome sequence and the newly emerging high-throughput single nucleotide

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/05/0592–0088$22.00/0

Accessible online at:
www.karger.com/hhe

Dr. Qihua Tan
Department of Clinical Biochemistry and Genetics (KKA)
Odense University Hospital, Sdr. Boulevard 29
DK–5000 Odense C (Denmark)
Tel. +45 65412822, Fax +45 65411911, E-Mail qihua.tan@ouh.fyns-amt.dk

polymorphism genotyping techniques which enable the high-density whole-genome screening of complex trait genes, LD mapping is gaining more popularity [4]. At the same time, instead of the traditional single-locus model, multi-locus statistical approaches [5] that take into account the interdependence of genetic variants important in complex disease etiology are appealing.

Because particular DNA variants may remain together on ancestral haplotypes (set of ordered markers) for many generations, groups of neighboring genetic variants can form haplotypic diversity with distinctive patterns of LD that can be exploited in both genetic linkage and association studies [6]. Haplotype analysis is more efficient than the single-locus association test because it makes use of the LD information contained in the flanking markers [7]. The 'haplotype relative risk' (HRR) approaches have been applied to detect allelic associations when parental genotypes are available for phase inference and for constructing the controls [8, 9]. Unfortunately, such methods are not applicable in longevity studies because parental genotype information is unavailable for the long-lived. In order to reconstruct the missing phases in the genotype data, different algorithms have been proposed. These include the rule-based algorithm [10], the E-M algorithm [11] and the recent Bayesian approaches [12–14]. Model comparison [15] has shown that the Bayesian approach, which uses the MCMC algorithm and Gibbs sampling [12], can be regarded as an efficient tool for estimating haplotypes [16]. It is necessary to point out that although a haplotype association analysis of disease traits can be conducted by directly treating the inferred haplotypes per subject as if they were observed, such practice can result in biased estimates of haplotype effects with possibly increased errors in the estimation [17, 18]. The E-M algorithm provides maximum likelihood estimates and therefore allows hypothesis testing using the likelihood ratio statistic [19, 20]. However, the method is confined to case-control data and provides no estimate or test of individual haplotypes. Epstein and Satten [21] introduced a retrospective likelihood method to estimate and test the effects of individual haplotypes on binary traits, but their method is again restricted to case-control data. Using the generalized linear model, Schaid et al. [22] proposed a score test for haplotype inference. The model can be generalized to a variety of different disease traits and performs efficient tests on individual haplotypes. In the context of human longevity studies, the traditional E-M based haplotype-estimating technique has been implemented in data analysis [23–25]. In these applications, the study designs are consequently limited to a simple case-control or two-group setup with cases consisting of the long-lived or centenarians and controls of young individuals. Although popular in use, the case-control design has low power when applied to longevity studies as the phenotype of interest (i.e. age) is a continuous trait [26]. Furthermore, in cross-sectional studies, the observed ages are those at participation which are all censored and cannot be modeled by the traditional survival analysis models. It is thus necessary that efficient haplotype inference methods be derived to accommodate the situation.

For the single-locus analysis, new statistical methods have been developed to model genotype-specific survivals [27–29]. These methods make full use of individual phenotype information and are thus inherently more powerful [26]. In this paper, we propose a new survival analysis method to apply to unphased multi-locus genotype data to evaluate haplotype effects on human survival and to estimate haplotype frequencies at birth and over the observed ages. By incorporating population survival information in the analysis and based on the proportional hazard assumption, we show how our model can estimate sex-specific haplotype effects, incorporate observed covariates, assess haplotype-environment interactions, examine modes of haplotype function (multiplicative, dominant and recessive) and model heterogeneity in the unobserved individual frailty while using a nonparametric baseline hazard function. Data-analyzing strategies are also suggested to optimize the throughput of the data. After model validation using computer simulation, the model is applied to an empirical multi-locus genotype data set collected in an association study on the interleukin 6 (IL-6) gene and longevity [30] to estimate the relative risks and frequencies of the haplotypes. Finally, we discuss the significance of our method in mapping genes that modulate human survival and some practical issues in model application.

Methods

Population and Haplogenotype-Specific Survivals

First we denote the collection of all the observed multi-locus genotypes over the typed loci with G and the collection of all the haplotypes that make up the genotypes with H. When haplotype frequencies are in Hardy-Weinberg equilibrium (HWE), the frequency of the haplotype pair or haplogenotype (h_i, h_j) is

$$P(h_i, h_j) = \begin{cases} 2p_i p_j & i < j \\ p_i p_j & i = j \end{cases} \quad (1)$$

where p_i and p_j are haplotype frequencies at birth for haplotypes h_i and h_j. Assuming that the risks of the haplotypes (denoted as r_i and r_j, respectively) are multiplicative, in a proportional hazard model the hazard of death at age x for the haplogenotype made up of haplotypes h_i and h_j is

$$\mu_{i,j}(x) = r_i r_j \mu_0(x) \tag{2}$$

where $\mu_0(x)$ is the baseline hazard function. Correspondingly, the survival of carriers of the haplogenotype is

$$s_{i,j}(x) = e^{-\int_0^x \mu_{i,j}(t)dt} = e^{-r_i r_j H_0(x)} = s_0(x)^{r_i r_j} . \tag{3}$$

Giving HWE at birth, the mean survival of the population is the weighted haplogenotype-specific survival, i.e.

$$\bar{s}(x) = \sum_{i,j \in H} P\left(h_i, h_j\right) s_{i,j}(x) = 2 \sum_{i<j,\ i,j \in H} p_i p_j s_{i,j}(x) + \sum_{i=j,\ i,j \in H} p_i p_j s_{i,j}(x) \tag{4}$$

The Likelihood Function

The parameterization in equation 4 is haplotype based, i.e. we assume that haplotypes are known explicitly for each individual. In the practical situation, what we observe are unphased multi-locus genotypes instead of haplotypes. However, for each multi-locus genotype g, there is a set of haplotype pairs denoted as $S(g)$ that are consistent with g. With this relationship, the frequency of the observed genotype g at age x can be expressed in terms of haplotype frequencies and haplogenotype-specific survivals by using equation 4,

$$p_g(x) = \frac{\sum_{i,j \in S(g)} P\left(h_i, h_j\right) s_{i,j}(x)}{\bar{s}(x)} . \tag{5}$$

With equation 5, we construct the likelihood function at age x using the multinomial distribution of the multi-locus genotype frequencies in the population as

$$\log L_{data}(x) \propto \sum_{g \in G} n_g(x) \log p_g(x) . \tag{6}$$

In equation 6, $n_g(x)$ is the number of individuals carrying the multi-locus genotype g. The log likelihood of the entire data is simply the sum of equation 6 over all the observed ages. The covariance matrix obtained by inversing the information matrix can be used to calculate the univariate Wald statistic for significance inferences of the risk parameters.

Similarly to the calculation of allele frequencies from genotype data, with the maximum likelihood estimates from equation 6 and using equation 4, we can calculate the frequency of any haplotype h_i at age x as

$$p_i(x) = \frac{p_i^2 s_{i,i}(x) + 0.5 \sum_{i<j} p_i p_j s_{i,j}(x)}{\bar{s}(x)} . \tag{7}$$

Modeling Heterogeneity

As a complex trait, human survival is modulated by the interplay of both genetic and nongenetic factors which form competing risks (frailty) that contribute to the individual hazard of death [31]. Ignoring the existence of heterogeneity in the unobserved individual frailty can lead to a substantial underestimation of the risks of genetic factors [28, 29]. Under the proportional hazard assumption, if an individual carrying a haplotype pair (h_i, h_j) has the frailty z, the hazard of death at age x is

$$\mu_{i,j}(x \mid z) = z\mu_{i,j}(x) = z r_i r_j \mu_0(x).$$

The mean hazard of death for a heterogeneous population carrying the haplogenotype is

$$\bar{\mu}_{i,j}(x) = \int_0^\infty \mu_{i,j}(x|z) f_x(z)\, dz = \mu_{i,j}(x) \int_0^\infty z f_x(z)\, dz = \mu_{i,j}(x)\, \bar{z}(x) . \tag{8}$$

Following the traditional approach [32–34], we assume that the frailty z is gamma distributed with mean 1 and variance σ^2. Then $\bar{z}(x)$ in equation 8 can be derived as

$$\bar{z}(x) = \left[1 + \sigma^2 \int_0^x \mu_{i,j}(s)\, ds\right]^{-1} = \left[1 + \sigma^2 H_{i,j}(x)\right]^{-1} .$$

Substituted into equation 8, we get

$$\bar{\mu}_{i,j}(x) = \frac{\mu_{i,j}(x)}{1 + \sigma^2 H_{i,j}(x)} = \frac{r_i r_j \mu_0(x)}{1 + \sigma^2 r_i r_j H_0(x)} \tag{9}$$

where $H_0(x)$ is the cumulative baseline hazard function. Correspondingly, we have the mean survival for the haplogenotype,

$$\bar{s}_{i,j}(x) = \left[1 + \sigma^2 r_i r_j H_0(x)\right]^{-\frac{1}{\sigma^2}} . \tag{10}$$

In order to fit a frailty model, we replace equation 3 with equation 10 in the analyses. Estimating the variance parameter σ^2 requires a large sample size [35]. In small-scale investigations, σ^2 can be determined by a grid search for the peak of the likelihood for tentatively assigned values of σ^2 [29, 36]. Based on our experiences in fitting the gamma frailty model to large population data sets, one can alternatively fit a frailty model by simply setting σ^2 to 0.1. This can be conservatively compared with some empirical results [29, 35, 36]. However, we think that it is applicable to small data sets.

The Baseline Hazard

The baseline hazard function can take a parametric form such as that of the Gompertz model, $\mu_0(x) = ae^{bx}$. However, by introducing the population survival information available in population statistics into equation 4, our model allows the estimation of a nonparametric baseline hazard function. This is done by using a two-step procedure described by Yashin et al. [28] and in more detail by Tan et al. [29]. The idea is that, with a known population survival, equation 4 can be solved by using a numerical procedure to get a nonparametric $s_0(x)$ for the given haplotype risk and frequency parameters. In the estimation procedure, we start with an initial guess of the haplotype risks and frequencies and apply it to equation 4 to calculate $s'_0(x)$. This $s'_0(x)$ is introduced into equation 6 to estimate a new set of haplotype parameters which are then used to calculate an updated $s'_0(x)$. This process iterates until the likelihood function converges [28, 29].

Sex-Specific HRRs

As a well-known phenomenon in demography, a sex difference in human mortality exists in all populations. Such a difference is

crucial in longevity studies because the majority of centenarians are females [37, 38]. To take this into account, we introduce the sex-specific population survival functions $\bar{s}_m(x)$ and $\bar{s}_f(x)$ from the population statistics into equation 4 and rewrite it as

$$\bar{s}_m(x) = \sum_{i,j \in H} P(h_i, h_j)\, {}_m s_0(x)^{\,{}_m r_i\, {}_m r_j}\,,$$
$$\bar{s}_f(x) = \sum_{i,j \in H} P(h_i, h_j)\, {}_f s_0(x)^{\,{}_f r_i\, {}_f r_j}\,. \qquad (11)$$

By calculating the sex-specific baseline survival functions ${}_m s_0(x)$ and ${}_f s_0(x)$, we are able to estimate sex-specific HRRs to capture the sex-dependent effects or gene-sex interaction in human survival. When no sex-specific effect exists, the same risk parameters can be assigned to reduce the number of parameters in the model. Note that, in any case, the same haplotype frequency parameters are specified for both sexes.

Incorporating Covariates and Interactions

In our proportional hazard model, it is possible to incorporate nongenetic or environmental covariates to account for effects of the observed confounding factors as well as gene-environment interactions. If there is one environmental factor, geographical location (south and north) that affects the mean population survival with relative risk r_e, and in addition has an interaction with one of the haplotypes (haplotype h_i) in our data, the risk of the haplotype is ${}_n r_i$ in the north and ${}_s r_i$ in the south. And if the proportion of northerners is p, we can rewrite equation 4 as

$$\begin{aligned}\bar{s}(x) &= p\bar{s}_n(x) + (1-p)\bar{s}_s(x)\\ &= p\sum_{i,j\in H} P(h_i,h_j) s_0(x)^{\,{}_n r_i r_j r_e} + (1-p)\sum_{i,j\in H} P(h_i,h_j) s_0(x)^{\,{}_s r_i r_j}\\ &\quad + p\sum_{i,j\in H} P(h_{-i},h_j) s_0(x)^{\,{}_n r_{-i} r_j r_e} + (1-p)\sum_{i,j\in H} P(h_{-i},h_j) s_0(x)^{\,{}_s r_{-i} r_j}\,. \end{aligned} \qquad (12)$$

In equation 12, the environmental effect r_e is defined as the risk of being a northerner as opposed to being a southerner. Relative risks for haplotype h_i are estimated separately to allow for area-specific effects or gene-environment interaction.

Nonmultiplicative Effects

Up to now, we have been assuming that the risks of haplotypes are multiplicative. By strategic parameterization, our model can be applied to detect effects of haplotypes that are dominant or recessive. If the effect of haplotype h_i is dominant over the others, then equation 4 can be expressed as

$$\bar{s}(x) = \sum_{i,j\in H} P(h_i,h_j) s_0(x)^{r_i} + \sum_{-i,j\in H} P(h_{-i},h_j) s_{-i,j}(x) \qquad (13)$$

where the same risk parameter r_i is imposed on carriers of the haplotype regardless of their haplogenotypes. In the same manner, when the effect of haplotype h_i is recessive, we have

$$\bar{s}(x) = 2\sum_{i<j,\ i,j\in H} p_i p_j s_0(x)^{r_j} + p_i^2 s_0(x)^{r_i} + \sum_{-i,j\in H} P(h_{-i},h_j) s_{-i,j}(x)\,. \qquad (14)$$

In equation 14, the risk of haplotype h_i is only assigned to those who have two copies of the haplotype. Note that, for the last terms in both equation 13 and 14, the specification of the risk parameters is ambiguous. This has to do with the data-analyzing strategies in the next section.

Model Selection

In our model, we assign one risk and one frequency parameter to each haplotype. Because the number of haplotypes increases exponentially with the number of typed loci, there will be too many parameters to be estimated, thus reducing the power of the model. Similar to Epstein and Satten [21], we recommend testing the association between each haplotype and survival by setting the relative risk for each of the other haplotypes to 1. To further reduce the numbers of parameters, we can group the rare haplotypes [22] and similarly set the group risk to 1. In the analysis, different modes of haplotype function can be assumed and tested. The subset of haplotypes exhibiting a potential association with survival can be selected and put together into the model for an extensive analysis. In both the single-haplotype and the extensive analyses, the haplotypes with risk set to 1 (including the grouped rare haplotypes) serve as the reference or the baseline to ensure that the model is identifiable. The Akaike information criterion (AIC) [39] can be applied to select a model with the maximum number of important haplotypes [21]. Once the best performance model is selected, we suggest using the log likelihood ratio test to obtain an overall significance for the haplotype effects.

Simulation

We conduct a limited simulation study to examine the performance of our model. Data sets of different sizes are generated (1,000 replicates) for the given parameters using population survival data in the 2001 Danish life table [40]. We take the haplotype frequencies estimated from an empirical data set (3 single nucleotide polymorphisms, 8 haplotypes) [unpubl. data from our laboratory] to generate the haplotypes. Among the 8 haplotypes, we choose one with a frequency of 0.145 as a beneficial haplotype and set its HRR to 0.8. Besides the haplotype parameters, we also assume that geographical location has an effect on survival with the relative risk of 1.25 for the north and 1 for the south. The frequency of northerners is assigned as 0.65 in the simulation. In addition, we assume that the unobserved frailty is gamma distributed with mean 1 and variance 0.1. Multi-locus genotypes over the 3 single nucleotide polymorphism loci are simulated for individuals from the ages of 50 to 99 with 10, 20 and 40 individuals at each age (corresponding to sample sizes of n = 500, 1,000 and 2,000).

We specify 4 models to validate our method (models 1 and 3) and evaluate the effects of heterogeneity (model 2) and the observed nongenetic covariate (model 4) on

Table 1. Estimated risk and frequency parameters by different models in the simulation study (1,000 replicates for each model)

Model and parameter	True	n = 500			n = 1,000			n = 2,000		
		medium	percentile		medium	percentile		medium	percentile	
			2.5%	97.5%		2.5%	97.5%		2.5%	97.5%
Model 1 ($\sigma^2 = 0.1$)										
Haplotype freq.	0.145	0.144	0.113	0.177	0.145	0.124	0.167	0.145	0.125	0.162
HRR, north	0.800	0.800	0.614	1.057	0.801	0.666	0.960	0.801	0.707	0.918
HRR, south	1.000	1.006	0.735	1.385	1.008	0.795	1.253	1.007	0.856	1.169
Freq., north	0.650	0.652	0.598	0.701	0.650	0.611	0.686	0.650	0.624	0.678
Risk, north	1.250	1.254	1.113	1.420	1.250	1.147	1.360	1.252	1.178	1.325
Model 2 ($\sigma^2 = 0$)										
Haplotype freq.	0.145	0.144	0.115	0.177	0.145	0.123	0.166	0.145	0.130	0.161
HRR, north	0.800	0.825	0.652	1.021	0.819	0.705	0.965	0.827	0.733	0.923
HRR, south	1.000	1.014	0.791	1.345	1.007	0.827	1.212	1.003	0.873	1.141
Freq., north	0.650	0.646	0.589	0.696	0.646	0.609	0.683	0.646	0.621	0.673
Risk, north	1.250	1.214	1.094	1.346	1.211	1.125	1.306	1.211	1.150	1.277
Model 3 ($\sigma^2 = 0.1$)										
Haplotype freq.	0.145	0.143	0.115	0.176	0.145	0.125	0.167	0.145	0.130	0.160
HRR	0.800	0.794	0.640	0.986	0.801	0.675	0.939	0.801	0.712	0.895
Freq., north	0.650	0.648	0.594	0.699	0.651	0.611	0.686	0.650	0.625	0.675
Risk, north	1.250	1.248	1.125	1.393	1.249	1.165	1.351	1.251	1.189	1.317
Model 4 ($\sigma^2 = 0.1$) Covariate ignored										
Haplotype freq.	0.145	0.144	0.114	0.174	0.145	0.124	0.166	0.145	0.130	0.160
HRR	0.800	0.806	0.643	1.020	0.809	0.698	0.942	0.807	0.721	0.896

parameter estimation. For models 1 and 2, we generate our data by assuming that, in addition to the risk of area, the risk of haplotype is area dependent with an HRR = 0.8 in the north and an HRR = 1 in the south. Simulation results in table 1 show that all parameters used in generating the data are well retrieved by our model that takes into account the unobserved frailty (model 1). In model 2, however, the estimated relative risks for area and the beneficial haplotype are all biased towards 1. The results indicate that ignorance of individual heterogeneity in unobserved frailty can lead to conservative estimates of the risk parameters [41]. In addition, one can see that a sample size of $n > 500$ is required to ensure that the range of the 2.5–97.5% percentile for the estimated HRRs is exclusive of the null risk of 1. In generating the data for models 3 and 4, the same risk and frequency parameters are assumed for the haplotype of interest and for area but the HRR is no longer area dependent because here our interest is to examine how ignorance of the existing risk covariate (area) can affect the estimated HRR. Similar to model 1, model 3 captures all the parameters adequately. However, the median of the estimated HRRs by model 4 is biased towards 1 for all the three different sample sizes (table 1) which reminds us of the situation of model 2. Moreover, by comparing the distribution of the estimated HRRs in models 3 and 4 for n = 500, we see that the ranges of the 2.5–97.5% percentile of HRRs for model 3 is well beyond 1 but that for model 4 includes the null risk of 1. These results suggest that the exclusion of existing risk covariates not only leads to biased estimates on the relative risk parameters but could also result in reduced power in haplotype effect inferences.

In figure 1, we show the simulated and the estimated age patterns of haplotype frequency for the beneficial haplotype in model 1 in the north (fig. 1a) and the south (fig. 1b; 1,000 individuals at each age). The estimated haplotype frequency (solid) using unphased genotype data captures the haplotype frequency trajectory in the simulated data (dotted), which again validates our model. Furthermore, the area-dependent haplotype effect or haplotype-environment interaction is clearly shown by the different haplotype frequency patterns revealed in figure 1a and b.

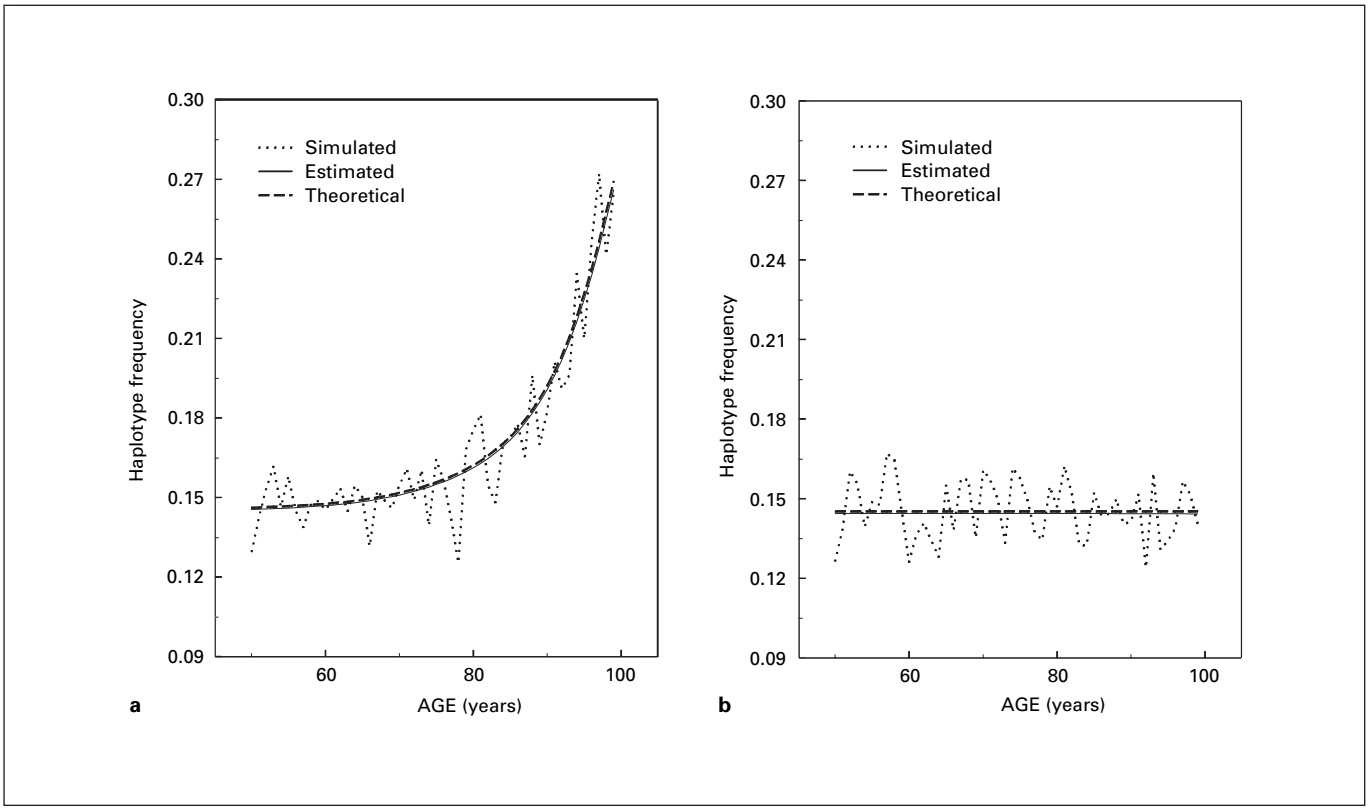

Fig. 1. The age patterns of the theoretical (dashed), the simulated (dotted) and the estimated (solid) haplotype frequencies in the north (**a**) and the south (**b**). The simulated age pattern is based on a large sample of 1,000 individuals at each age from 50 to 99 years. The figure shows that the model correctly captures the true haplotype frequency trajectory by age as well as gene-environment interaction or area-dependent HRR.

Application

The increased level of IL-6 gene activity has been linked to stress conditions that characterize the aging process. Previous studies have revealed associations of IL-6 with age-related diseases such as Alzheimer's disease [42], cardiovascular events [43] and type 2 diabetes [44]. The influence of IL-6 on human aging and survival was investigated by Christiansen et al. [30], who carried out haplotype analysis on a total of 1,143 subjects genotyped at 2 single-point polymorphisms (–572G/C and –174G/C) and 1 AT stretch polymorphism (–373AnTm, 4 alleles) in the promoter region. In their study, haplotype frequencies in the young (<70 years, 567 individuals) and the old (93 years, 576 individuals) age groups were compared for the 6 most common haplotypes (table 2) by the E-M algorithm. A slight decrease with age in the frequency of the –572G/–373A_8T_{12}/–174C haplotype (denoted as G/A_8T_{12}/C in table 2) was found. Instead of dividing the subjects into young and old groups, we applied our HRR model to the same data to estimate HRRs and infer the effects of IL-6 on human survival. To fit the model, we introduced the population survival data from the 2001 Danish life table [40]. Because our preliminary analysis showed no sex-dependent haplotype effect in the data, we assigned the same haplotype risk parameters for both sexes to reduce the number of parameters in the model. In the analyses (table 2), effects of the haplotypes were assumed to be multiplicative, dominant (equation 13) and recessive (equation 14). In table 2, the HRR for each haplotype was estimated by setting the HRRs for the other haplotypes to 1. For each haplotype tested, we calculated AIC for selecting the best-fitting model under the different modes of haplotype function (multiplicative, dominant, recessive). Among the 6 haplotypes, haplotype G/A_8T_{12}/C showed the lowest AIC (5,413.244) in a multiplicative model with an HRR = 1.087 (p = 0.050) suggesting the harmful effect of the haplotype on human sur-

Table 2. Paremeter estimates and model comparison by single-haplotype models fitted to IL-6 data

Haplotype	Frequency at birth	Relative risk[a]				AIC
		HHR	SE	95% CI	p value	
Multiplicative						
$G/A_8T_{12}/C$	0.473	1.087	0.044	1.000–1.173	0.050	5,413.244
$G/A_9T_{11}/G$	0.200	0.949	0.050	0.851–1.047	0.313	5,415.939
$G/A_{10}T_{11}/G$	0.193	0.925	0.053	0.821–1.029	0.158	5,414.931
$G/A_{10}T_{10}/G$	0.072	1.046	0.082	0.885–1.207	0.575	5,416.666
$C/A_{10}T_{10}/G$	0.038	1.031	0.106	0.823–1.237	0.773	5,416.870
$C/A_9T_{11}/C$	0.013	0.758	0.162	0.440–1.075	0.136	5,414.771
Dominant						
$G/A_8T_{12}/C$	0.460	1.068	0.059	0.952–1.183	0.250	5,415.617
$G/A_9T_{11}/G$	0.209	0.991	0.057	0.879–1.102	0.876	5,416.921
$G/A_{10}T_{11}/G$	0.194	0.911	0.056	0.801–1.020	0.110	5,414.424
$G/A_{10}T_{10}/G$	0.070	1.017	0.086	0.848–1.186	0.841	5,416.906
$C/A_{10}T_{10}/G$	0.039	1.066	0.117	0.837–1.294	0.570	5,416.626
$C/A_9T_{11}/C$	0.013	0.758	0.162	0.440–1.075	0.136	5,414.771
Recessive[b]						
$G/A_8T_{12}/C$	0.461	1.097	0.067	0.967–1.227	0.145	5,414.726
$G/A_9T_{11}/G$	0.202	0.826	0.093	0.645–1.006	0.060	5,413.795
$G/A_{10}T_{11}/G$	0.208	1.008	0.061	0.889–1.127	0.889	5,416.939
$G/A_{10}T_{10}/G$	0.070	1.613	0.587	0.463–2.758	0.296	5,415.127
$C/A_{10}T_{10}/G$	0.036	0.556	0.333	0.000–1.205	0.182	5,415.659

[a] Heterogeneity model with $\sigma^2 = 0.1$.
[b] No estimate on $C/A_9T_{11}/C$ haplotype due to low frequency.

vival. Consistent with Christiansen et al. [30], our model pointed to $G/A_8T_{12}/C$ as the only haplotype exhibiting potential influence on human lifespan. Even though more than half of the subjects in our sample were of the same age (93 years), our model produced a higher significance level for the effect of the $G/A_8T_{12}/C$ haplotype as compared with the two-group method [30]. However, as the haplotype was only of marginal significance and considering the multiple haplotypes tested in table 2, we cautiously conclude that our result of the $G/A_8T_{12}/C$ haplotype is only suggestive. The second lowest AIC was observed for the recessive model of haplotype $G/A_9T_{11}/G$. In contrast to the $G/A_8T_{12}/C$ haplotype, its HRR (0.826) indicates that it might be a beneficial haplotype that reduces the hazard of death for homozygous carriers of the haplotype. By calculating the Wald statistic, we obtain a p value of 0.060 which is insignificant.

In table 2, AIC was merely applied to single-haplotype models to evaluate the haplotype effects under different modes of haplotype function. To illustrate how AIC can be used to select the best model of the models with different subsets of haplotypes, we also fitted a 2-haplotype model by adding the recessive $G/A_9T_{11}/G$ haplotype to the $G/A_8T_{12}/C$ dominant model. We got a higher AIC (5,413.270) from the 2-haplotype model as compared with the AIC (5,413.244) from the 1-haplotype model ($G/A_8T_{12}/C$ dominant). As expected, the 2-haplotype model which includes an insignificant haplotype does not outperform the previous 1-haplotype model.

Assuming that the effects of the 6 haplotypes are multiplicative, we also fitted a multiple-haplotype model (log likelihood: –2,694.759) to illustrate how the likelihood ratio test can be used to assess the overall statistical significance of the effects of the 6 haplotypes included in the model. Since the log likelihood is –2,698.472 in the multiple-haplotype model that assumes no haplotype effect, we obtain a likelihood ratio test statistic with 6 degrees of freedom of –2[–2,698.472 – (–2,694.759)] = 7.427. This leads to an overall p value of 0.283 which again means that there is no association between the haplotypes and survival. In figure 2, we show the estimated haplotype frequency trajectories over the observed ages from the multiple-haplotype model. The modest effect of the $G/A_8T_{12}/C$ haplotype is shown by its frequency pattern that slightly decreases with increasing age. Due to the low death rate at early ages and low risks of the haplotypes,

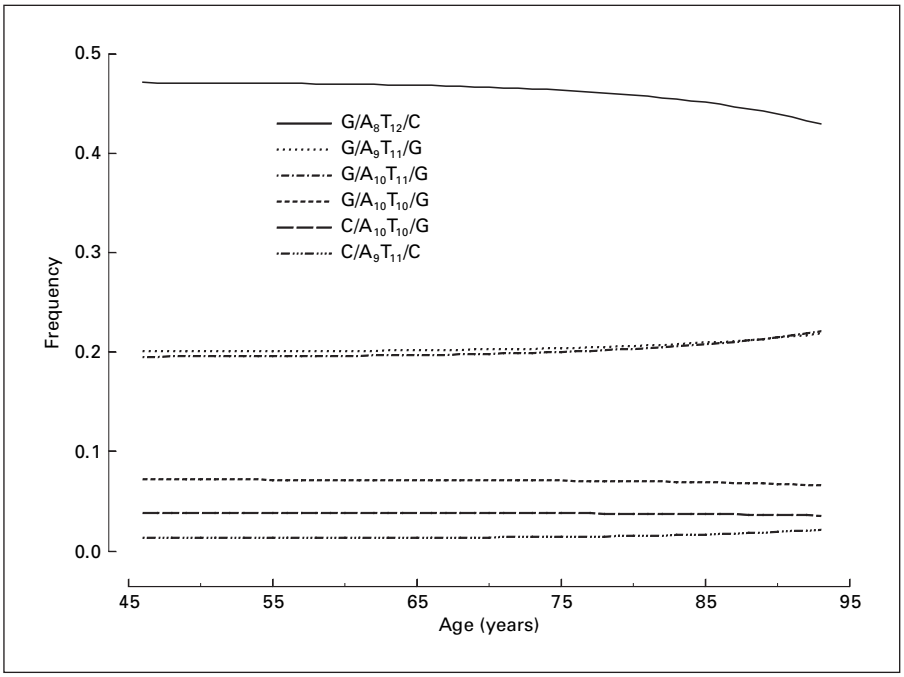

Fig. 2. Estimated haplotype frequency trajectories over the observed ages for the IL-6 data. The modest effect of the $G/A_8T_{12}/C$ haplotype is shown by its frequency pattern that slightly decreases with increasing age. Due to the low death rate at early ages and the low risk of the haplotypes, frequency changes are mainly observable at high ages.

Table 3. Comparison of the estimated haplotype frequencies by the HRR model and the E-M algorithm assuming multiplicative haplotype effects

Haplotype	HRR model		E-M algorithm[a]	
	46	93	<70	93
$G/A_8T_{12}/C$	0.472	0.430	<0.470	0.432
$G/A_9T_{11}/G$	0.201	0.219	<0.203	0.217
$G/A_{10}T_{11}/G$	0.195	0.221	<0.196	0.222
$G/A_{10}T_{10}/G$	0.072	0.066	<0.071	0.066
$C/A_{10}T_{10}/G$	0.038	0.035	<0.038	0.035
$C/A_9T_{11}/C$	0.013	0.021	<0.013	0.020

[a] From Christiansen et al. [30].

frequency changes (although insignificant) are mainly observed at high ages. In table 3, we compare the calculated haplotype frequencies at the ages of 46 (the youngest age in our subjects) and 93 with the frequencies calculated by the E-M algorithm [30]. It is interesting to see that both models produced consistent frequency estimates for the same data. However, it is important to point out that our HRR model not only estimates the frequency for each haplotype, but also provides point and interval estimates on its relative risk.

Discussion

We have shown that our HRR model can be applied to multi-locus genotype data from unrelated individuals to estimate frequencies and risks of haplotypes while incorporating additional covariates. In addition, our proportional hazard model facilitates the estimation of sex-specific HRRs and the assessment of interactions between haplotypes and covariates as well as examination of the modes of gene function. By introducing the gamma-distributed frailty, our model can also infer the haplotype effects on human survival with consideration of individual heterogeneity in the unobserved frailty which is important in the context of longevity studies because of the complex nature of the human lifespan. Given the crucial role of association studies in the genetics of human aging and longevity, we think that our HRR model may serve as a useful tool for researchers in this field.

The basic assumption in our model is that haplotype frequencies at birth follow the Hardy-Weinberg law. As we have mentioned, such an assumption is sensible as differential survival driven by the association between the haplotypes and hazard of death has not yet imposed survival selection on the subjects as long as the haplotypes we are interested in do not affect in utero survival and there is no preferential transmission of a particular genetic variant in the region under investigation. Under this

assumption, genotype frequency information at other ages can contribute to the maximum likelihood estimation of the haplotype frequencies at birth. Most importantly, as long as HWE holds at birth, we can relax with regard to the HWE assumption on haplotype frequencies at other ages except when a multiplicative effect model is preferred. This is important because different genetic mechanisms or modes in the haplotype function in human survival can be tested without imposing HWE at the advanced ages. Here, it is necessary to point out that HWE might not be a reasonable assumption in case of a subdivided human population as it can destroy HWE at any age (including at birth). In family-based association studies, such a problem can be solved by conducting the transmission/disequilibrium test [45]. However, in genetic association studies of human aging and longevity, parental genotype information is usually missing which means that one has to stick to the population-based association approaches. Using unlinked markers, Pritchard and Rosenberg [46] proposed a statistical method to detect population stratification. Furthermore, statistical tests that account for population substructure have been developed for case-control association studies [47, 48]. More work is needed for implementing these ideas in the genetic association analysis of human survival traits.

Because we assume that the risk of haplotype on the hazard of death is constant over the ages, our model is a proportional hazard model [49] in nature. Antagonistic pleiotropic effects have occasionally been reported in the genetic studies on human longevity [50, 51]. To deal with this situation, parametric survival models were proposed in the analysis of single-locus data [27]. Such approaches model the antagonistic effect as an intersection of the mortality curves for different genotypes. Although they can easily be implemented in our model, there are important issues to be considered regarding the parametric modeling. Firstly, when the sample size is limited, there will be a considerable error in estimating the genotype-specific survival distributions. Consequently, the age-dependent effect modeled by the differential survival between the genotypes is unreliable. This becomes more problematic at advanced ages when sample collecting is difficult. Secondly, the choice of a proper parametric form for the survival function can be crucial in determining the results. At extreme ages, the validity of the parametric survival function, such as the Gompertz or the Gompertz-Makeham models, has been questioned recently [52]. On the other hand, when the proportional hazard model is applicable, our method works without imposing any parametric form on the baseline hazard function when population survival from population statistics is introduced.

Although our model is capable of incorporating covariate, it must be pointed out that because the likelihood function is based on the age pattern in the frequency changes of the subgroups formed by the combination of haplogenotypes and the covariate, the covariate has to be an attribute fixed early in life. We refer the fixed attributes or covariates to factors that characterize an individual's social class, education or persistent living environment. Studies on Danish twins have shown that such fixed attributes are important factors in determining an individual's lifespan [53, 54]. Most importantly, the capability of assessing the interaction effect between haplotype and the observed covariate (fig. 1) could help us to better understand the mechanisms in human aging and survival.

Acknowledgements

This work was jointly supported by the US National Institute on Aging (NIA) research grants NIA-P01-AG08761 and -AG13196 and by 'The MicroArray Center' project under the Biotechnological Research Program financed by the Danish Research Agency. We thank Shuxia Li for technical support.

References

1 Tan Q, Yashin AI, Christensen K, Jeune B, De Benedictis G, Kruse TA, Vaupel JW: Multidisciplinary approaches in genetic studies on human aging and longevity. Curr Genomics 2004;5:409–416.

2 Tan Q, Zhao JH, Iachine I, Hjelmborg J, Vach W, Vaupel JW, Christensen K, Kruse TA: Power of non-parametric linkage analysis in mapping genes contributing to human longevity in long-lived sib-pairs. Genet Epidemiol 2004;26:245–253.

3 Risch NJ: Searching for genetic determinants in the new millennium. Nature 2000;405:847–856.

4 Gray IC, Campbell DA, Spurr NK: Single nucleotide polymorphisms as tools in human genetics. Hum Mol Genet 2000;9:2403–2408.

5 Hoh J, Ott J: Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 2003;4:701–709.

6 Schork NJ, Fallin D, Lanchbury JS: Single nucleotide polymorphisms and the future of genetic epidemiology. Clin Genet 2000;58:250–264.

7 Akey J, Jin L, Xiong M: Haplotypes vs single marker linkage disequilibrium tests: What do we gain? Eur J Hum Genet 2001;9:291–300.

8 Terwilliger JD, Ott J: A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum Hered 1992;42:337–346.

9 Clayton D: A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 1999;65: 1170–1177.
10 Clark AG: Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 1990;7:111–122.
11 Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995; 12:921–927.
12 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001; 68:978–989.
13 Niu T, Qin ZS, Xu X, Liu JS: Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 2002; 70:157–169.
14 Lin S, Cutler DJ, Zwick ME, Chakravarti A: Haplotype inference in random population samples. Am J Hum Genet 2002;71:1129–1137.
15 Stephens M, Donnelly P: A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 2003;73:1162–1169.
16 Orzack SH, Gusfield D, Olson J, Nesbitt S, Subrahmanyan L, Stanton VP Jr: Analysis and exploration of the use of rule-based algorithms and consensus methods for the inferral of haplotypes. Genetics 2003;165:915–928.
17 Zhao LP, Li SS, Khalid N: A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 2003;72:1231–1250.
18 Tanck MW, Klerkx AH, Jukema JW, De Knijff P, Kastelein JJ, Zwinderman AH: Estimation of multilocus haplotype effects using weighted penalised log-likelihood: Analysis of five sequence variations at the cholesteryl ester transfer protein gene locus. Ann Hum Genet 2003; 67:175–184.
19 Long JC, Williams RC, Urbanek M: An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 1995;56: 799–810.
20 Zhao JH, Sham PC: Faster haplotype frequency estimation using unrelated subjects. Hum Hered 2002;53:36–41.
21 Epstein MP, Satten GA: Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 2003; 73:1316–1329.
22 Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002; 70:425–434.
23 Bonafe M, Marchegiani F, Cardelli M, Olivieri F, Cavallone L, Giovagnetti S, Pieri C, Marra M, Antonicelli R, Troiano L, Gueresi P, Passeri G, Berardelli M, Paolisso G, Barbieri M, Tesei S, Lisa R, De Benedictis G, Franceschi C: Genetic analysis of paraoxonase (PON1) locus reveals an increased frequency of Arg192 allele in centenarians. Eur J Hum Genet 2002; 10:292–296.
24 Ross OA, Curran MD, Rea IM, Hyland P, Duggan O, Barnett CR, Annett K, Patterson C, Barnett YA, Middleton D: HLA haplotypes and TNF polymorphism do not associate with longevity in the Irish. Mech Ageing Dev 2003; 124:563–567.
25 Geesaman BJ, Benson E, Brewster SJ, Kunkel LM, Blanche H, Thomas G, Perls TT, Daly MJ, Puca AA: Haplotype-based identification of a microsomal transfer protein marker associated with the human lifespan. Proc Natl Acad Sci USA 2003;100:14115–14120.
26 Pletcher SD, Stumpf MP: Population genomics: Ageing by association. Curr Biol 2002;12: 328–330.
27 Toupance B, Godelle B, Gouyon PH, Schachter F: A model for antagonistic pleiotropic gene action for mortality and advanced age. Am J Hum Genet 1998;62:1525–1534.
28 Yashin AI, De Benedictis G, Vaupel JW, Tan Q, Andreev KF, Iachine IA, Bonafe M, DeLuca M, Valensin S, Carotenuto L, Franceschi C: Genes, demography, and lifespan: The contribution of demographic data in genetic studies on aging and longevity. Am J Hum Genet 1999;65:1178–1193.
29 Tan Q, De Benedictis G, Yashin AI, Bonafe M, DeLuca M, Valensin S, Vaupel JW, Franceschi C: Measuring the genetic influence in modulating human lifespan: Gene-environment and gene-sex interactions. Biogerontology 2001;2: 141–153.
30 Christiansen L, Bathum L, Andersen-Ranberg K, Jeune B, Christensen K: Modest implication of interleukin 6 promoter polymorphisms in longevity. Mech Ageing Dev 2004;125:391–395.
31 Christensen K, Vaupel JW: Determinants of longevity: Genetic, environmental and medical factors. J Intern Med 1996;240:333–341.
32 Vaupel JW, Manton KG, Stallard E: The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 1997; 16:439–54.
33 Aalen O: Heterogeneity in survival analysis. Stat Med 1998;7:1121–1137.
34 Hougaard P: Modeling heterogeneity in survival analysis. J Appl Prob 1991;28:695–701.
35 Ewbank DC: Mortality differences by APOE genotype estimated from demographic synthesis. Genet Epidemiol 2002;22:146–155.
36 Yashin AI, De Benedictis G, Vaupel JW, Tan Q, Andreev KF, Iachine IA, Bonafe M, DeLuca M, Valensin S, Carotenuto L, Franceschi C: Genes and longevity: Lessons from studies on centenarians. J Gerontol 2000;55A:B1–B10.
37 Hazzard WR: Biological basis of the sex differential in longevity. J Am Geriatr Soc 1986; 34:455–471.
38 Holden C: Why do women live longer than men? Science 1987;238:158–160.
39 Akaike H: Factor analysis and AIC. Psychometrika 1987;52:317–332.
40 Agerskov U, Bisgaard MP: Statistical Yearbook 2002. Copenhagen, Statistics Denmark, 2002, p 67.
41 Vaupel JW, Yashin AI: Heterogeneity's ruses: Some surprising effects of selection on population dynamics. Am Stat 1985;39:176–185.
42 Licastro F, Grimaldi LM, Bonafe M, Martina C, Olivieri F, Cavallone L, Giovanietti S, Masliah E, Franceschi C: Interleukin-6 gene alleles affect the risk of Alzheimer's disease and levels of the cytokine in blood and brain. Neurobiol Aging 2003;24:921–926.
43 Cesari M, Penninx BW, Newman AB, Kritchevsky SB, Nicklas BJ, Sutton-Tyrrell K, Rubin SM, Ding J, Simonsick EM, Harris TB, Pahor M: Inflammatory markers and onset of cardiovascular events: Results from the Health ABC study. Circulation 2003;108:2317–2322.
44 Vozarova B, Fernandez-Real JM, Knowler WC, Gallart L, Hanson RL, Gruber JD, Ricart W, Vendrell J, Richart C, Tataranni PA, Wolford JK: The interleukin-6 (-174) G/C promoter polymorphism is associated with type-2 diabetes mellitus in Native Americans and Caucasians. Hum Genet 2003;112:409–413.
45 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–516.
46 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65:220–228.
47 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. Am J Hum Genet 2000;67:170–181.
48 Satten GA, Flanders WD, Yang Q: Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 2001;68:466–477.
49 Cox DR: Regression models and life-tables. J R Stat Soc B 1972;34:187–220.
50 Schachter F, Faure-Delaneff L, Guenot F, Rouger H, Froguel P, Lesueur-Ginot L, Cohen D: Genetic associations with human longevity at the APOE and ACE loci. Nat Genet 1994;6: 29–32.
51 De Benedictis G, Carotenuto L, Carrieri G, De Luca M, Falcone E, Rose G, Yashin AI, Bonafe M, Franceschi C: Age-related changes of the 3′-APOB-VNTR genotype pool in aging cohorts. Ann Hum Genet 1998;62:115–122.
52 Driver C: The Gompertz function does not measure ageing. Biogerontology 2001;2:61–65.
53 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, Curtsinger JW: Biodemographic trajectories of longevity. Science 1998;280:855–860.
54 Yashin AI, Iachine IA: How frailty models can be used for evaluating longevity limits: Taking advantage of an interdisciplinary approach. Demography 1997;34:31–48.