Genetic Markers Data in Survival Studies of Twins: The Results of a Simulation Study

Alexander Z. Begun[1] and Anatoli I. Yashin[2]

1 Universität der Bundeswehr Hamburg, Germany

2 Center for Demographic Studies, Duke University, Durham, North Carolina, United States of America

Previous longevity studies of related individuals such as twins or siblings based on the major gene model have shown that the frequency and the relative risk of mortality of a beneficial allele in the population could be estimated. If, in addition to survival data for related individuals, the genetic markers data are available, one could try to locate the longevity allele in the genome. In the case where the phenotypic trait is life span or age at onset of disease, a two-step procedure can be used. First, the parameters of bivariate survival functions must be estimated from bivariate survival data for twins without markers. The second step is focused on determining the position of longevity genes between respective markers. To calculate the joint distribution of inheritance vector and genetic markers, the hidden Markov chain technique is used. This approach is illustrated with a simulation example for one longevity gene.

Let us assume that in addition to survival data, the genetic markers data are available for twins. How can this data be used in genetic studies of longevity? Firstly, genetic markers can be considered as observed covariates. The influence of these covariates on survival can be estimated by standard techniques that involve a Cox-type proportional hazards model and its extensions specified for univariate or multivariate survival analyses. If some coefficient of regression, say in standard Cox's regression univariate survival analysis, is significantly different from zero, and all loci are in linkage disequilibrium, then respective genetic markers are involved in life span determination. Univariate analysis does not always reveal the direct influence of genetic markers on survival, but bivariate or multivariate analyses do. In the case of linkage disequilibrium, this result could mean the possibility of the longevity or frailty gene located in the neighborhood of respective genetic markers at the chromosome. The real strength of bivariate and multivariate survival analyses with genetic markers is that they allow detection not only of the presence of longevity or frailty genes but also determine the location of these genes at the chromosome, even in the case of observed genetic markers in linkage equilibrium. The methods which address this involve linkage analysis. Some of these methods are based on regression models (Haseman & Elston, 1972). Recent approaches use maximization of likelihood (Kruglyak et al., 1996; Kruglyak & Lander, 1995). In the latter, the most dif-ficult element is deriving the likelihood function of the data. This usually involves both the calculation and maximization of the likelihood function. In the case of survival data, that is, when the phenotypic trait is life span or age at onset of disease or disability, the two-step procedure can be used. First, the parameters characterizing the bivariate survival function must be estimated from bivariate survival data without genetic markers. The second step involves linkage analysis, that is, determining the position of the longevity or frailty gene between respective markers at the chromosome. The procedures involved in the first step are described in Begun, Iachine and Yashin, (2000), and Yashin and Iachine (1994). The linkage procedure generally involves calculation of the distribution of inheritance vector data (Kruglyak & Lander, 1995) and of the conditional distribution of life span as an intermediate step, and the consequent averaging of likelihood with respect to this distribution.

Morton (1955) suggested linkage analysis using logarithm of the odds (LOD) score (Ott, 1991). This method was modified and adjusted to different pedigree structures data (Kruglyak & Lander, 1995; Lander & Green, 1987; Lathrop et al., 1986). Clerger-Darpoux et al. (1986) showed that parametric linkage analysis could be extremely sensitive to misspecification of the model. Kruglyak et al. (1996) suggested an approach to both parametric and nonparametric analyses, the key feature being the separation of two issues: 1) extracting information about the inheritance pattern in a pedigree; and 2) defining a statistic for assessing linkage for a given inheritance pattern.

Received 22 June, 2004; accepted 28 September, 2004.

Address for correspondence: Alexander Begun, Universität der Bundeswehr Hamburg, Fachbereich WOW, 22039 Hamburg, Germany. E-mail: Alexander.Begun@unibw-hamburg.de

In this paper this approach is combined with methods of multivariate survival analysis (Begun, Desjardins et al., 2000). Let M_1, M_2, …, M_l be genetic markers for two related individuals, and let V_1, V_2, …V_l be respective inheritance vectors. The hidden Markov chain method is used to calculate the joint distribution of inheritance vectors and genetic markers. For this purpose, the extended vector of genetic markers and extended inheritance-vectors are introduced by adding markers and inheritance vectors corresponding to the major gene with the possible influences on longevity that we are looking for. Since the location of the major gene at the chromosome is not known, it is placed between markers M_1 and M_2, then M_2 and M_3 and so forth. The recombination distance estimate between the observed marker and the major gene locus is then calculated as well as the LOD score for each case. Assume that the locus with the major gene influencing life span is located between markers M_i and M_{i+1}. It is convenient to include this locus in the sequence of markers and consider the extended marker vector. Let θ be the recombination distance between M'_i and the major gene locus (i.e., the distance between M'_i and M'_{i+1} in the extended marker vector assuming that M'_{i+1} is the major gene locus for the longevity or frailty alleles). Here it is assumed that observed markers are in linkage equilibrium and that they do not influence longevity directly — only the unobserved major gene may possess this property. Note that the location of this gene does not influence survival — only the genotype value has this property. The location, however, may influence joint distribution of the extended markers vector and life span values for genetically related individuals. This distribution is used in the specification of joint likelihood of life span and markers data. Equation [3] shows the joint probability of observed markers and survival function. To calculate the likelihood of the data, one must take an appropriate number of derivatives with respect to life span arguments and replace these arguments with respective data.

Thus the main problem with this approach is the probability calculation of the extended vector of markers used in [3]. To calculate this probability, a method which uses the Markov property of a pair (V'_i, M'_i), i = 1, 2, …, l+1 is suggested. This pair is considered as a random process with respect to discrete index i. The inheritance vector contains information on recombination which did or did not happen with parental chromosomes during the meiosis resulting in the birth of a child. One way of coding this information is to know whether the grandmother's or the grandfather's allele was transmitted to the grandchild. Since every individual has four grandparents, and each grandparent has two opportunities for allele transmission to their grandchildren, there are 16 (4 × 4) possible combinations for grandchildren locus in the case of two siblings. Each recombination in the parents' chromosome during meiosis produces change in respective combination, that is, in the inheritance vector. It is assumed that: a) parents' genotypes are independent at each locus with known probabilities in markers' loci and unknown probabilities in longevity locus; b) $V_1(j)$ = 1/16 for all j = 1, …, 16. Given parents' genotypes and probabilities of recombination for each pair of loci (j, j + 1), the stepwise probability of a sibling, twin pair or a group of related individuals having the extended genotype can be calculated. Since life spans and all markers except the major gene locus are observed, the probability of extended genotype must be multiplied by the respective conditional survival functions (or their derivatives). The final likelihood is obtained by averaging the result with respect to all possible parental genotypes and twins' genotypes in the longevity locus. Positioning the longevity locus in different places of a chromosome between respective markers and calculating in each case $\log_{10}$ (Likelihood) – $\log_{10}$ (*Likelihood*) where *Likelihood* is the value of the likelihood function with the condition that the longevity locus is situated out of the chromosome (i.e., the recombination probabilities of the longevity gene is equal to 0.5), the LOD score profile for the longevity gene can be constructed. The accuracy of the estimating θ can be assessed by constructing a specific support interval. This support interval must contain all the points where the LOD score is higher than or equal to 3 as recommended by Ott (1991). For this interval the linkage is significant (except when this interval contains θ = 0.5). All values of θ at which the LOD score is less than or equal to –2, however, are excluded.

There is no principal difference in applying this technique to the case with two or more longevity loci. The extended inheritance vector and the likelihood function can be constructed in the same way. In this article, the implementation of this method to simulated data for one longevity locus located on the same chromosome as the markers is discussed.

Materials and Methods

Suppose that it is known that the $n \times 2$ matrix of life spans X for n dizygotic twins (sibling pairs) and the $n \times l \times 4$ matrix M of markers, where n is the number of twin pairs, l is the number of markers. Given twin pair i, i=1, 2, …, n, and the marker number j, j=1, 2,…l, $M_{i,j,1}$ is the marker-allele inherited by the first twin from the mother, $M_{i,j,2}$ is the marker-allele inherited by the first twin from the father, $M_{i,j,3}$ and $M_{i,j,4}$ are the marker-alleles inherited by the second twin from the mother and the father, respectively. Let marker number j have $K(j)$ possible alleles with probabilities p_{j,j_1}, j_1 = 1, …, $K(j)$,

$$\sum_{j_1=1}^{K(j)} p_{j,j_1} = 1$$

We assume that parental alleles are inherited by offspring independently, that parental genotype frequencies are in Hardy–Weinberg equilibrium and that the

https://doi.org/10.1375/twin.8.1.34 Published online by Cambridge University Press

parents are chosen independently. Distances between observed markers are known and equal $\theta(j)$, $0 < \theta(j) < 0.5$, $j = 1, \ldots, l-1$. We also assume that all observed markers are in linkage equilibrium and do not influence longevity directly. Only the longevity gene, which is in linkage equilibrium with observed markers, influences longevity. These assumptions allow the analysis to be carried out in two steps. First, the bivariate survival model is defined relating longevity with the major (or longevity) gene and the parameters of this model estimated. The position of the major gene is located in the second step.

Major Gene Model With Multiplicative Action of Longevity Allele

We assume that an individual's instantaneous risk of death μ at age t, as measured by the hazard of mortality, depends linearly on frailty Z. Namely, $\mu(t,Z) = Z\mu_0(t)$, where $\mu_0(t)$ is the underlying hazard. The random variable Z does not depend on the age but depends on the number of major gene alleles in the major gene locus and is equal to r^{i-1}, where i is the number of the major gene alleles in the genotype. It is clear that $0 < r < 1$ and that the major gene allele acts multiplicatively. Let p be the frequency of the major gene allele. For univariate survival function and autosomal locus, the following parameterization is used:

$$S(x) = (1 + s^2\tilde{H}(x))^{-1/s^2} = p^2 e^{-r^2H(x)} + 2p(1-p)e^{-rH(x)} + (1-p)^2 e^{-H(x)} \quad [1]$$

where $\tilde{\mu}_0(x) = d\tilde{H} / dx = a + be^{cx}$, $\tilde{H}(30) = 0$, s, a, b, c are unknown parameters, and $H(x)$ is the cumulative hazard for the unit risk of mortality (i.e., in the absence of a longevity allele in genotype). Note that this survival is equal to survival in a population with underlying hazard $\tilde{\mu}_0(x)$ and Gamma-distributed frailty with a mean of 1 and variance s^2 at the age of 30 years. In formula [1], the different possible combinations of alleles in genotypes and respective frequencies of genotypes are used. If longevity locus is autosomal, the bivariate survival function $S(x_1,x_2)$ can be calculated as follows (Begun, Desjardins et al., 2000):

$$\begin{aligned} S(x_1,x_2) &= p^4 e^{-r^2H(x_1)-r^2H(x_2)} + p^3(1-p)(e^{-r^2H(x_1)} + e^{-rH(x_1)}) \\ &(e^{-r^2H(x_2)} + e^{-rH(x_2)}) + p^2(1-p)^2(0.5e^{-r^2H(x_1)} + \\ &e^{-rH(x_1)} + 0.5e^{-H(x_1)})(0.5e^{-r^2H(x_2)} + e^{-rH(x_2)} + 0.5e^{-H(x_2)}) + \\ &p(1-p)^3(e^{-rH(x_1)} + e^{-H(x_1)})(e^{-rH(x_2)} + e^{-H(x_2)}) + \\ &2p^2(1-p)^2 e^{-rH(x_1)-rH(x_2)} + (1-p)^4 e^{-H(x_1)-H(x_2)} \end{aligned} \quad [2]$$

For simplicity we assume that life span data is not censored. However, the analysis can be extended easily to censored data. Unknown parameters of the frailty p, r and of the univariate fit a, b, c, s can be estimated through maximization of the likelihood function.

Location of the Major Gene

Assume that the major gene is situated between the marker number j_0 and the marker number j_0+1 at the recombination distance θ'_{j_0} from the marker number j_0. Our goal is to calculate the full likelihood function.

$$L(M,X) = \prod_{i=1}^{n} f(X_{i,1}, X_{i,2}, M_i) = \prod_{i=1}^{n} \frac{\partial^2 S(X_{i,1}, X_{i,2}, M_i)}{\partial X_{i,1} \partial X_{i,2}}$$

Here, the bivariate survival $S(x_1, x_2, m)$ for a dizygotic twin-pair with longevity values x_1, x_2, and $l \times 4$ marker matrix of observed m is:

$$\begin{aligned} S(x_1,x_2,m) &= P_{1,1}(m)e^{-r^2H(x_1)-r^2H(x_2)} + P_{1,2}(m)e^{-r^2H(x_1)-rH(x_2)} + \\ &P_{1,3}(m)e^{-r^2H(x_1)-H(x_2)} + P_{2,1}(m)e^{-rH(x_1)-r^2H(x_2)} + \\ &P_{2,2}(m)e^{-rH(x_1)-rH(x_2)} + P_{2,3}(m)e^{-rH(x_1)-H(x_2)} + \\ &P_{3,1}(m)e^{-H(x_1)-r^2H(x_2)} + P_{3,2}(m)e^{-rH(x_1)-rH(x_2)} + \\ &P_{3,3}(m)e^{-H(x_1)-H(x_2)} \end{aligned} \quad [3]$$

where $P_{g1,g2}(m)$ is the probability of a twin pair having the marker set m and longevity genotypes $g1$ for the first and $g2$ for the second twin. Each genotype may take one of three possible values. Genotype 1 has two beneficial alleles in longevity loci, genotype 2 one beneficial allele and genotype 3 no beneficial allele.

Let $\nu_j = (\nu_{j,1},\nu_{j,2},\nu_{j,3},\nu_{j,4})^T$ be the inheritance vector for the marker number j, $j = 1, 2, \ldots, l$ with components equaling 0 or 1. The first and third ones denote the alleles inherited from the mother for the first and for the second twin respectively (0 if from the grandmother and 1 if from the grandfather). Analogously, the second and the fourth components denote alleles inherited from the father.

Example 1

Let Aa be maternal and aA paternal genotypes at some locus k_0. For the inheritance vector $\nu_{k_0} = (0,1,1,1)^T$ the first twin will have genotype AA and the second one genotype aA. If the inheritance vector at this loci were $(1,0,0,1)^T$, twins would have genotypes aa and AA correspondingly.

To calculate $P_{g1,g2}(M_i)$ for a twin-pair number i, the recurrent procedure based on the algorithm of the hidden Markov chain (Lander & Green, 1987) will be used as follows.

Initialization

Assume that ν_1 is uniformly distributed and that the probability of each possible value of ν_1 is equal to 1/16. Set j=1, $P_{g1,g2}(M_i) = 0$ and define the set of possible combinations of the major gene alleles corresponding to the genotype $(g1,g2)$ by G (for genotypes 1 or 3 there exists a single combination of the major gene alleles, but for genotype 2 either the maternal or parental allele in the major gene locus is the major gene allele).

Step 1

Select an element from G, denote it by M'_{j_0} and set $G = G \setminus M'_{j_0}$. M'_{j_0} is included in the set of markers M_i between the marker number j_0 and the marker number j_0+1. The extended $(l+1) \times 4$ matrix M'_i of markers, extended l-dimensional vector θ' with unknown recombination probabilities $\theta'(j_0) = \theta'_{j0}$

https://doi.org/10.1375/twin.8.1.34 Published online by Cambridge University Press

and $\theta'(j_0 + 1) = (\theta(j_0) - (\theta'(j_0)))/(1 - 2\theta'(j_0))$ are constructed. Let $M'_{i,1:j} = (M'_{i,1}, \ldots, M'_{i,j})$ for $j = 1, \ldots, l + 1$.

Step 2

Calculate conditional probability

$$P(v_{j+1} \mid v_j, M'_{i,1:j}) = (\theta'_j)^{d(v_{j+1}, v_j)}(1 - \theta'_j)^{4 - d(v_{j+1}, v_j)}$$

where the Hammer distance $d(v_{j+1}, v_j)$ is the number of noncoinciding components in the vectors v_j and v_{j+1}.

Step 3

Calculate probability

$$P(M'_{i,j+1} \mid v_{j+1}, M'_{i,1:j}) = P(M'_{i,j+1} \mid v_{j+1}, v_j, M'_{i,1:j})$$

by firstly reconstructing parental marker genotypes $M_{i,j+1}{}^1$, $M_{i,j+1}{}^2$ for the first and the second siblings:

$M_{i,j+1,1+v_{j+1,1}}{}^1 = M'_{i,j+1,1}$, $M_{i,j+1,3+v_{j+1,2}}{}^1 = M'_{i,j+1,2}$, $M_{i,j+1,1+v_{j+1,3}}{}^2 = M'_{i,j+1,3}$, $M_{i,j+1,3+v_{j+1,4}}{}^2 = M'_{i,j+1,4}$.

If estimated $M_{i,j+1}{}^1$, $M_{i,j+1}{}^2$,do not contradict each other, $P(M'_{i,j+1} \mid v_{j+1}, M'_{i,1:j})$ is equal to the product of frequencies for defined parental alleles. Otherwise, $P(M'_{i,j+1} \mid v_{j+1}, M'_{i,1:j}) = 0$ (see Example 2 below). If $j = 1$, calculate $P(M'_{i,1} \mid v_1)$ and set

$$R = P(M'_{i,1}) = \sum_{v_1} P(M'_{i,1} \mid v_1)P(v_1)$$

Step 4

Calculate

$$P(v_{j+1} \mid M'_{i,1:j}) = \sum_{v_j} P(v_j \mid M'_{i,1:j})P(v_{j+1} \mid v_j, M'_{i,1:j})$$

and

$$P(v_{j+1} \mid M'_{i,1:j+1}) = P(v_{j+1} \mid M'_{i,1:j})P(M'_{i,j+1} \mid v_{j+1}, M'_{i,1:j}) / P(M'_{i,j+1} \mid M'_{i,1:j})$$

where

$$P(M'_{i,j+1} \mid M'_{i,1:j}) = \sum_{v_{j+1}} P(M'_{i,j+1} \mid v_{j+1}, M'_{i,1:j})P(v_{j+1} \mid M'_{i,1:j})$$

Step 5

Calculate $R = P(M'_{i,j+1} \mid M'_{i,1:j})$ R. Set $j = j + 1$. If $j = l + 1$ and $G = \varnothing$, then set $P_{g1,g2}(M_i) = P_{g1,g2}(M_i) + R$ and go to *End*. If $j = l + 1$ and $G \neq \varnothing$, then set $j = 1$, $P_{g1,g2}(M_i) = P_{g1,g2}(M_i) + R$, and go to *Step 1*. Otherwise, go to Step 2.

End

This procedure is applied for all possible genotypes $(g1,g2)$ and all sibling pairs i, $i = 1, \ldots, n$. The full likelihood function and the value of LOD score for each given location of the major gene is then calculated. Finally, the LOD score profile is obtained, depending on the location of the major gene.

Example 2

Let aA and AA be genotypes of the first and the second twin respectively. For the inheritance vector $(0,1,1,0)^T$ maternal genotypes related to the first and the second twin must be aX and XA respectively (X means unknown allele here). That is, the maternal genotype must be aA. Analogously it can be proved that paternal genotype must be AA. The probability of this event can be calculated as a product of the probabilities of genotypes aA and AA. If the inheritance vector is $(0,1,0,0)^T$, one would have contradicting maternal genotypes aX and AX and noncontradicting paternal genotypes XA and AX. The probability of such an event is naturally equal to 0.

Remark

In the case of two longevity genes, the set of markers are extended by two new ones at locations under study and the hidden Markov chain technique is applied. Bivariate survivals can be calculated in accordance with accepted hypotheses about genotype-frailty correspondence.

Results

As previously mentioned, the identification of the parameters of univariate fit and frailty and the identification of the recombination value can be separated. In the first step, the parameters of bivariate survival function are estimated without markers. The LOD score profile is calculated in the second step. Empirical data, including the survival data for 1000 sibling pairs where both siblings were alive at the age of 30 years, and the data on 10 genetic markers for each sibling pair, were simulated. Each gene at marker locus can be characterized by a pair from the set of 10 different alleles and each allele can be met in the population with a frequency equal to 0.1. A longevity gene is situated in the middle between the 5th and 6th markers. The markers were distributed uniformly over a chromosome with a distance of 5cm between the neighboring markers. The distance between the first marker and the longevity gene was approximately 22.5cm. The univariate survival function in the absence of the longevity allele was parameterized by

$$S(x) = (1 + s^2\tilde{H}(x))^{-1/s^2}$$

where

$$\tilde{H}(x) = (b/c)(e^{cx} - e^{30c}), \quad s^2 = 0.01, b = 0.000025$$

and $c = 0.1$. The frequency of the longevity allele with multiplicative action was 0.5. The age dynamics of the marginal estimated and the empirical hazards are shown in Figure 1. The LOD score profile averaged over all simulations for given parameters of univariate fit, for given frequency of longevity allele and for two given different risks of longevity allele is shown in the Figure 2. In the first case, the risk was .1 and the coefficient of correlation between life spans of siblings was equal to .29. In the second case, these values were .35 and .12 respectively. As expected, the maximum LOD score was observed near the real position of the longevity gene and this profile has a symmetrical form. The greater the risk of the beneficial allele, the less expressed the maximum.

https://doi.org/10.1375/twin.8.1.34 Published online by Cambridge University Press

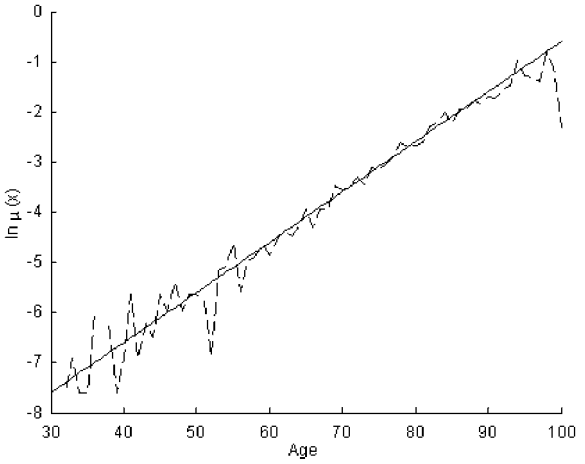

Figure 1

Age dynamics of marginal estimated hazard (solid line) and empirical hazard (dashed line).

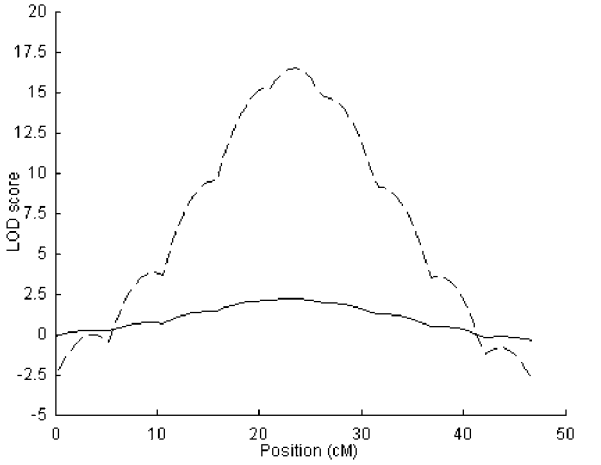

Figure 2

LOD score profile for known frailty parameters. The action of longevity allele is 0.1 (dashed line) and 0.35 (solid line).

Discussion

Data on markers and the longevity of related individuals makes it possible not only to estimate the parameters of univariate fit and frailty distribution, but also to test the hypothesis of genetic influence on longevity, and to estimate the possible location of the longevity gene. The technique used consisted of two parts. First, unknown fit and frailty parameters were estimated using survival data on monozygotic and dizygotic twin pairs. The likelihood function for the unknown parameter of the longevity gene location can be presented as a weight sum of bivariate survivals multiplied by genotype frequency. At the second stage, these genotype frequencies given markers were calculated using a hidden Markov chain algorithm. As shown in the figures, the possibility of such localization of the longevity gene in high degree depends on the relative mortality risk of longevity allele. The clear peak of the LOD score profile can be obtained only if this risk is not too large.

References

Begun, A. Z., Desjardins, B., Iachine, I. A., & Yashin, A. I. (2000). Multivariate frailty model with a major gene: Application to genealogical data. *Studies in Health Technology and Informatics, 77,* 412–416.

Begun, A. Z., Iachine, I. A., & Yashin, A. I. (2000). Genetic nature of individual frailty: Comparison of two approaches. *Twin Research, 3,* 51–57.

Clerget-Darpoux, F., Bonaiti-Pellie, C., & Hochez, J. (1986). Effects of misspecifying genetic parameters in lod score analysis. *Biometrics, 42,* 393–399.

Haseman, J. K., & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics, 2,* 3–19.

Kruglyak, L., Dali, M. J., Reeve-Daly, M. P., & Lander, E., S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics, 58,* 1347–1363.

Kruglyak, L., & Lander, E. S. (1995). Complete multipoint sib pair analysis of quantitative and qualitative traits. *American Journal of Human Genetics, 57,* 439–454.

Lander, E. S., & Green, P. (1987). Construction of multilocus genetic maps in human. *Proceedings of the National Academy of Science, 84,* 2363–2367.

Lathrop, G. M., Lalouel, J. M., & White, R. L. (1986). Calculation of human linkage maps: Likelihood calculations for multilocus linkage analysis. *Genetic Epidemiology, 3,* 39–52.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics, 7,* 277–318.

Ott, J. (1991). *Analysis of human genetic linkage.* London: The John Hopkins University Press.

Yashin, A. I., & Iachine, I. A. (1994). *Environment determines 50% of variability in individual frailty: Results from Danish twin study. Population study of aging.* Odense, Denmark: Center for Health and Social Policy, Odense University.

https://doi.org/10.1375/twin.8.1.34 Published online by Cambridge University Press