© 2008 Cambridge University Press
doi:10.1017/S0016672308009348 Printed in the United Kingdom

Estimation of additive genetic and environmental sources of quantitative trait variation using data on married couples and their siblings

SANGITA KULATHINAL[1], DARIO GASBARRA[2], SANJAY KINRA[3], SHAH EBRAHIM[3] AND MIKKO J. SILLANPÄÄ[2]* FOR THE INDIAN MIGRATION STUDY GROUP

[1] *Indic Society for Education and Development, Nashik, India*
[2] *Department of Mathematics and Statistics, University of Helsinki, FIN-00014 Helsinki, Finland*
[3] *Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK*

(*Received 20 February 2008 and in revised form 9 April 2008*)

Summary

Twin studies have been used to understand the sources of genetic and environmental variation in body height, body weight and other common human quantitative traits. However, it is rather unclear whether these two sources of variation could be really separated in practice. Here, we consider a special study design where phenotype data from married couples and their siblings have been collected. The marital status gives information about the shared environment, while siblings give information about both genetic and environmental variation. To dissect sources of variation and to allow some deviations and pedigree errors in the data, we model such data using a robust polygenic model with finite genome length assumption. As a summary, we provide the estimates for age-dependent proportions of total variation which are due to polygenic and environmental effects. Here, these estimates are provided for body height, weight, systolic blood pressure and total serum cholesterol measured from subjects of the Indian Migration Study.

1. Introduction

Both genetic and common environmental effects are important in understanding sources of variation in quantitative phenotypes (e.g., risk or susceptibility factors for diseases). However, due to lack of evidence for the effects of shared environment, family studies or twin studies may result in over-estimation of genetic effects (Hopper, 2000). Because of this, an extended twin design is often preferred (e.g., Stoel *et al.*, 2006) including phenotypes of the spouses (Eaves *et al.*, 1999). To account for the shared environmental effects, a large number of different types of individuals, their siblings and spouses need to be studied and information on the number of years spent together in a specific/shared environment, age up to which the environment (household) was shared and phenotypes need to be collected. To summarize, the estimated division between the genetics and environment in a population-specific manner, heritability is usually used as a summary statistics to describe the proportion of phenotypic variation attributable to genetic factors.

For quantitative traits, heritability estimation is often performed using a polygenic model (Henderson, 1984; Lynch and Walsh, 1998; Abney *et al.*, 2000). A polygenic model (or infinite locus or infinitesimal model) has been defined (Fisher, 1918) and used by assuming infinite number of additive unlinked loci (and infinite genome length), which implies 0·5 as a constant degree of genetic relationship between full siblings (Fisher, 1918). However, as is well known, the length of the genome is finite and specific to different species. The size of the human genome corresponds to an equivalent of 85 independent loci (Visscher *et al.*, 2006). Because of this, Visscher *et al.* (2006) first estimated the actual values of relationship coefficients (entries of the covariance matrix) using a set of neutral markers and then incorporated these values into the relationship matrix instead of a fixed 0·5 for

* Corresponding author. Department of Mathematics and Statistics, University of Helsinki, PO Box 68 (Gustaf Hällströmin katu 2b), FIN-00014 University of Helsinki, Finland. Tel: (358) 9-191-51512. Fax: (358) 9-191-51400. e-mail: mjs@rolf.helsinki.fi.

https://doi.org/10.1017/S0016672308009348 Published online by Cambridge University Press

full siblings. This practice resulted in much higher heritability estimates, than otherwise expected, based on collected phenotypes for human height, which is a typical quantitative trait (see Visscher *et al.*, 2006; Xu, 2006). The obtained heritability of $h^2 \approx 0{\cdot}8$ was close to estimates from twin studies (Silventoinen *et al.*, 2003).

For computational easiness, a finite polygenic model has been proposed as a finite locus approximation for polygenic model (Thompson and Skolnick, 1977; Du *et al.*, 1999; Du and Hoeschele, 2000). However, the drawback in such models is that the heritability estimates are clearly dependent on the number of loci assumed in the model. Here, we propose another kind of finite approximation for a polygenic model, where instead of assuming infinite genome length, elements of the relationship matrix corresponding to full siblings (degree of genetic relationship between siblings) are treated as random variables and estimated simultaneously with the variance components. Unlike Visscher *et al.* (2006), we do not utilize molecular marker information in estimation, but we assume the mean 0·5 and known standard deviation 0·0384 (cf. Guo, 1996; Visscher *et al.* 2006) for the unknown coefficients of genetic relationship between full siblings. This formulation provides a more flexible model, which should be robust for deviations and pedigree errors in the data.

In the Indian Migration Study (IMS) (Lyngdoh *et al.*, 2006), factory workers from four different parts of India are examined for various anthropometric and biochemical measurements apart from other characteristics. The spouses of the factory workers and siblings or relatives (in a very few cases when it was not possible to involve siblings) of both the factory workers and their spouses are examined. The main aim of IMS is to study the migration pattern in India; data on the place of origin, age when the place of origin was abandoned, age at the time of marriage and the number of years spent with the spouse in a common household are collected. It is to be noted that the siblings or relatives are taken from the place of origin of the respective factory worker or the spouse. Hence, IMS provided a unique opportunity to study the genetic as well as the shared environmental effects by modelling the dependency between the factory worker and his sibling/relative, factory worker and his spouse and the spouse and her sibling/relative (see Fig. 1).

The present work is motivated by IMS. In the setting of a simple linear mixed model, the effects of shared environment is modelled as a function of the number of years the environment was shared. In the present context, an environment refers to a common household in either a rural or an urban region.

In section 2, we construct a model describing the genetic and environmental effects on the phenotypes.

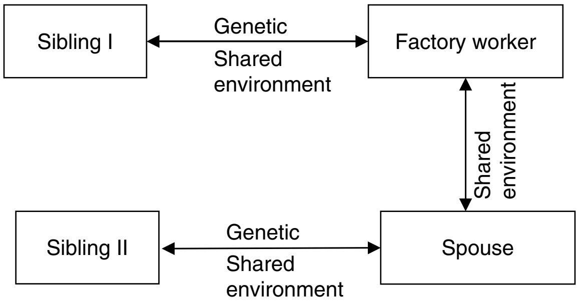


Fig. 1. IMS study subjects and their dependency due to genetic and environmental sharing.

We also define age-dependent proportions of total variations that are due to polygenic and environmental effects as meaningful summary statistics in the present context. In section 3, a Bayesian approach to the estimation of parameters is discussed and, in section 4, the proposed approach is applied to the IMS data. Finally, we conclude with the discussion section.

2. Model

In this section, we construct a model under the setting of the IMS, where the observation unit of the analysis is a group of four individuals defined by a male factory worker, his spouse, his sibling and his spouse's sibling. However, the idea presented here is very general and can be extended easily to other data types by appropriately specifying the relationships between the individuals within the unit of the analysis.

In the sequel, subscripts (1, 2, 3, 4) are used for the factory worker, his spouse, his sibling and his spouse's relative, respectively. Let a 4 × 1 vector of quantitative phenotypes corresponding to the factory worker i, $i = 1, 2, \ldots, n$ and related observations be $y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})^t$, where t denotes the transpose. That is, y_{i1} corresponds to the measurement of the factory worker i, y_{i2} corresponds to the measurement of the spouse of the factory worker i and so forth. Let $x_2 = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^t$ be the dummy variables or continuous measurements of the covariates of interest.

Consider a simple linear mixed model

$$y_i = \mu + \beta x_i + G_i + E_i + \varepsilon_i, \tag{1}$$

where μ is the overall mean and β is the effect of covariates x_i. Further, the additive genetic value or the polygenic effect vector, $G_i = (G_{i1}, G_{i2}, G_{i3}, G_{i4})^t$, is assumed to be jointly (multivariate) normally distributed with the mean vector of zeros and variance–covariance matrix

$$\Sigma_{gi} = \sigma_g^2 A_i. \tag{2}$$

Here, A_i is the additive genetic relationship matrix (Henderson, 1984), which characterizes the degree to

https://doi.org/10.1017/S0016672308009348 Published online by Cambridge University Press

which related individuals jointly share their genomes identical-by-descent (IBD) and is given by

$$A_i = \begin{pmatrix} 1 & 0 & c_i(1,3) & 0 \\ 0 & 1 & 0 & c_i(2,4) \\ c_i(1,3) & 0 & 1 & 0 \\ 0 & c_i(2,4) & 0 & 1 \end{pmatrix},$$

and sibling covariances $c_i(j, j') \in [0, 1]$, $(j, j') = (1, 3)$, (2, 4) and $i = 1, 2, \ldots, n$ are either taken from the common literature (when there is no inbreeding) as fixed constants equal to 0·5 or here are assumed to be independent and identically distributed random variables with expectation 0·5 and known variance $(0{\cdot}0384)^2$. Note that the fixed value 0·5 was originally derived as an asymptotic limiting value by assuming infinitely many loci (infinite genome length). Because all the genomes are of finite length in practice, the amount of IBD-sharing varies from one pair of siblings to another. Substantial deviations may also be expected in the presence of pedigree errors in the data. The detailed motivation for such a covariance structure is provided in Appendix A.

Further, the environment effect is modelled through the cumulative effects of the locations $E_i = (E_{i1}, E_{i2}, E_{i3}, E_{i4})^t$. The effects E_i are assumed to be drawn randomly from a multivariate normal distribution with mean vector zero and variance–covariance matrix Σ_{ei}, which has elements

$$\begin{aligned}\Sigma_{ei}(j,j') &= \mathrm{cov}\,(E_{ij}, E_{ij'}) \\ &= \sigma_e^2 D_i,\end{aligned} \tag{3}$$

where the sum $D_i = \sum D_i^l(j, j')$ is over all the locations or households shared by j and j', $D_i^l(j, j')$ is the time (in years) (j, j') spent together at location l for observation i, and σ_e^2 is the constant multiplier of the duration matrix D_i. Note that if (j, j') did not share any location, then $D_i^l(j, j') = 0$ and the corresponding observations on E_i would be independent. This is the case for (j, j') equal to $\{(1, 4), (4, 1), (3, 4), (4, 3)\}$. Similar identity is expected between observation (y_{i2}, y_{i3}), but in many Indian families siblings share the household even after they are married and hence the dependency between (y_{i2}, y_{i3}) and (y_{i2}, y_{i3}). This property is lacking from the settings of studies from developed countries. The rationale behind such a covariance structure is explained in Appendix A.

Excluding the interaction between G and E and under the assumption of independent and identical errors ε_i, the overall variance–covariance of y_i, *viz.* Σ_i is then the sum of appropriate covariance terms of the polygenic effect, shared environmental effects and the model (1) error variance (σ_m^2).

$$\Sigma_i = \sigma_g^2 A_i + \Sigma_{ei} + \sigma_m^2 I = \sigma_g^2 A_i + \sigma_e^2 D_i + \sigma_m^2 I, \tag{4}$$

where I is a 4×4 identity matrix. The mean vector of y_i is $\mu_i = \mu + \beta x_i$.

In twin studies the heritability is traditionally defined directly from the observed differences between the measured correlations among monozygotic and dizygotic twins (Falconer, 1989). Here, we assume no dominance variance and thus focus on 'narrow sense' heritability (σ_g^2/σ_y^2). It is clear from (4) that the phenotypic variability (σ_y^2) is modelled as an age-dependent quantity $(\sigma_g^2 + \text{age}\ \sigma_e^2 + \sigma_m^2)$ and this corresponds to the diagonal elements of the Σ_i. Note that each individual has lived with oneself for the number of years which are identical to his or her current age. In the present situation, the phenotypic variation turns out to be a function of the age when the phenotype was measured.

For a given age, we define the proportion of the total variation due to polygenic effect as

$$v_g(\text{age}) = \frac{\sigma_g^2}{\sigma_g^2 + \text{ age}\ \sigma_e^2 + \sigma_m^2}, \tag{5}$$

and the proportion of total variation due to environmental effect as

$$v_e(\text{age}) = \frac{\text{age}\ \sigma_e^2}{\sigma_g^2 + \text{ age}\ \sigma_e^2 + \sigma_m^2}. \tag{6}$$

It is clear that heritability $v_g(\text{age})$ is a decreasing function of age, but the rate at which it decreases with age depends on the environment variability σ_e^2. In general, the heritability estimates of the same trait assessed in an environment with larger variability would be lower. Similarly, $v_e(\text{age})$ is an increasing function of age and the rate is dependent on $(\sigma_g^2 + \sigma_m^2)$.

Omitting the age-dependency, a rather crude estimator of the heritability can also be given as

$$\tilde{v}_g = \frac{\hat{\sigma}_y^2 - (\sigma_e^2 + \sigma_m^2)}{\hat{\sigma}_y^2}, \tag{7}$$

$$\tilde{v}_e = 1 - \tilde{v}_g, \tag{8}$$

where $\hat{\sigma}_y^2$ is the empirical phenotypic variability, which can be estimated using sample variance of the observed phenotype data. Note that in expressions (5) and (6), σ_g^2 is estimated under the proposed model, while in expressions (7) and (8), $\hat{\sigma}_y^2$ is estimated externally from the observed data. The other two variances (σ_e^2, σ_m^2) are as defined earlier under the proposed model.

3. Bayesian computation

The likelihood function is simply the product of n-independent 4-variate normal density

$$L(\theta; y, x) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{4/2}\sqrt{|\Sigma_i|}} \exp \times \left\{ -\frac{1}{2}(y_i - \mu - \beta x_i)^t \Sigma_i^{-1} (y_i - \mu - \beta x_i) \right\}, \tag{9}$$

where $\theta=(\mu, \beta, \sigma^2_{\mathrm{g}}, \sigma^2_{\mathrm{e}}, \sigma^2_{\mathrm{m}})$ and $|\Sigma_i|$ is the determinant of the matrix Σ_i. *A priori* independent normal priors are assigned to μ and βs, while mutually independent Gamma priors are assigned to the inverses of the three variance components. For discussions of informative and uninformative priors, see Gelman (2006) and Dongen (2006). The posterior distribution is proportional to the product of the likelihood and the prior densities

$$p(\theta|y,x) \propto L(\theta;\, y,x)N(\mu;\, 0,10000)\prod_{j=1}^{k} N(\beta_j;\, 0,10000)$$
$$\times\Gamma(1/\sigma^2_{\mathrm{g}};\, 0\cdot01, 0\cdot01)\Gamma(1/\sigma^2_{\mathrm{e}};\, 0\cdot01, 0\cdot01)\Gamma(1/\sigma^2_{\mathrm{m}};\, 0\cdot01, 0\cdot01), \quad (10)$$

where k is the number of covariates, $N(z;\, 0{,}10000)$ is univariate normal density with mean 0 and variance 10 000 evaluated at z and $\Gamma(z;\, 0\cdot01,\ 0\cdot01)$ is the Gamma density with shape and scale parameters equal to 0·01 evaluated at z. The Gibbs sampler, applied to the posterior distribution of θ, involves specification of the following full conditional distributions:

$$p(\mu|.) \propto N(\mu;\, 0,10000)\prod_{i=1}^{n} \exp\left\{-\frac{1}{2}(y_i-\mu-\beta x_i)^t\,\Sigma_i^{-1}(y_i-\mu-\beta x_i)\right\},$$
$$p(\beta_j|.) \propto N(\beta_j;\, 0,10000)\prod_{i=1}^{n} \exp\left\{-\frac{1}{2}(y_i-\mu-\beta x_i)^t\,\Sigma_i^{-1}(y_i-\mu-\beta x_i)\right\},$$
$$p(\sigma^2_{\mathrm{g}}|.) \propto \Gamma(1/\sigma^2_{\mathrm{g}};\, 0\cdot01, 0\cdot01)L(\theta;\, y,x),$$
$$p(\sigma^2_{\mathrm{e}}|.) \propto \Gamma(1/\sigma^2_{\mathrm{e}};\, 0\cdot01, 0\cdot01)L(\theta;\, y,x),$$
$$p(\sigma^2_{\mathrm{m}}|.) \propto \Gamma(1/\sigma^2_{\mathrm{m}};\, 0\cdot01, 0\cdot01)L(\theta;\, y,x),$$

where all the variables are being conditioned on their most recently sampled values. If sampling from any of these full conditional distributions is difficult, then a Metropolis–Hastings step can be used. The samples from the posterior distribution of θ (10) are generated by using the OpenBUGS 2.2.0 software (Spiegelhalter *et al.*, 2005; Thomas *et al.*, 2006).

Let $(\sigma^{2(l)}_{\mathrm{g}}, \sigma^{2(l)}_{\mathrm{e}}, \sigma^{2(l)}_{\mathrm{m}})$, $l=1,\ldots, M$ be the Markov Chain Monte Carlo (MCMC) samples from the posterior distribution. The sample for the proportions at MCMC round l, $l=1,\ldots, M$ is

$$v^{(l)}_{\mathrm{g}}(\mathrm{age})=\frac{\sigma^{2(l)}_{\mathrm{g}}}{\sigma^{2(l)}_{\mathrm{g}}+\ \mathrm{age}\ \sigma^{2(l)}_{\mathrm{e}}+\sigma^{2(l)}_{\mathrm{m}}},$$
$$v^{(l)}_{\mathrm{e}}(\mathrm{age})=\frac{\mathrm{age}\ \sigma^{2(l)}_{\mathrm{e}}}{\sigma^{2(l)}_{\mathrm{g}}+\ \mathrm{age}\ \sigma^{2(l)}_{\mathrm{e}}+\sigma^{2(l)}_{\mathrm{m}}}.$$

The posterior median and credible intervals are then estimated using these samples.

When $c_i(1,\ 3)$ and $c_i(2{,}4)$ are allowed to be randomly distributed according to $N(0\cdot5,\ 0\cdot0384^2)$, the above posterior distribution (10) is multiplied by

$$\prod_{i=1}^{n} N(c_i(1,3);\, 0\cdot5, 0\cdot0384^2)N(c_i(2,4);\, 0\cdot5, 0\cdot0384^2).$$

The full conditional distributions of $c_i(1,\ 3)$ and $c_i(2,\ 4)$ are

$$p(c_i(1,3)|.) \propto N(c_i(1,3);\, 0\cdot5, 0\cdot0384^2)L_i(\theta;\, y_i, x_i),$$
$$p(c_i(2,4)|.) \propto N(c_i(2,4);\, 0\cdot5, 0\cdot0384^2)L_i(\theta;\, y_i, x_i),$$

where

$$L_i(\theta;\, y_i, x_i)=\frac{1}{(2\pi)^{4/2}\sqrt{|\Sigma_i|}}\exp$$
$$\times\left\{-\frac{1}{2}(y_i-\mu-\beta x_i)^t\Sigma_i^{-1}(y_i-\mu-\beta x_i)\right\}.$$

These unknown factors then get updated at each iteration. The missing data in y (under the missing at random assumption) are handled naturally in the Bayesian computation and they get updated or predicted at each iteration. The missing data in the covariates can be handled by introducing a prior distribution for covariates.

4. Data analyses

The IMS data are analysed here using three different models: (1) assuming a model $y_i=\mu+\varepsilon_i$ and a general variance–covariance matrix Σ for each i and using the Wishart prior for it, (2) the proposed model $y_i=\mu+\beta x_i+G_i+E_i+\varepsilon_i$ with 0·5 as the constant degree of genetic relationship, that is, $C_i(1,\ 3)$ and $C_i(2,\ 4)$ are equal to 0·5 for all i, and (3) the proposed robust model $y_i=\mu+\beta x_i+G_i+E_i+\varepsilon_i$ when the degree of genetic relationship is random and distributed according to the normal distribution with mean 0·5 and variance $(0\cdot0384)^2$. We also estimate the proportions $v_{\mathrm{g}}(\mathrm{age})$ and $v_{\mathrm{e}}(\mathrm{age})$.

https://doi.org/10.1017/S0016672308009348 Published online by Cambridge University Press

Table 1. *Estimated variance components. Posterior median (95% credible intervals) of three variances (σ^2_e, environment; σ^2_g, genetic; σ^2_m, error) for various quantitative phenotypes using models with 0·5 as the constant degree of genetic relationship (2) and when the degree of genetic relationship is random (3). The abbreviation 'SystBP' is used for systolic blood pressure*

Phenotype	Model	σ^2_e	σ^2_g	σ^2_m
Height (m)	(2)	8E-05 (7E-05, 9E-05)	7E-04 (5E-04, 0·0010)	7E-04 (5E-04, 9E-04)
	(3)	8E-05 (7E-05, 9E-05)	7E-04 (5E-04, 0·0010)	7E-04 (5E-04, 9E-04)
Weight (kg)	(2)	0·716 (0·372, 1·170)	49·43 (25·13, 71·73)	42·40 (27·60, 58·47)
	(3)	0·727 (0·472, 1·212)	48·18 (24·78, 72·56)	41·84 (24·95, 58·13)
SystBP (mmHg)	(2)	2·60 (1·76, 3·51)	101·95 (50·81, 156·25)	62·89 (30·61, 97·65)
	(3)	2·57 (1·79, 3·56)	101·85 (45·95, 153·84)	64·00 (30·64, 100·78)
Cholesterol (mg/dl)	(2)	10·50 (4·98, 16·08)	734·21 (433·08, 1039·17)	379·93 (189·07, 584·79)
	(3)	11·53 (7·76, 17·08)	703·23 (376·22, 1068·60)	363·43 (138·29, 575·37)

Table 2. *Estimated proportion of the total variation due to polygenic effect (v_g(age)) and due to environment effect (v_e(age)) at the age of 30 and crude estimates of these proportions $\tilde{v}_g$ and $\tilde{v}_e$. Posterior median (95% credible intervals) of the two proportions for various quantitative phenotypes using models with 0·5 as the constant degree of genetic relationship (2) and when the degree of genetic relationship is random (3). The abbreviation 'SystBP' is used for systolic blood pressure*

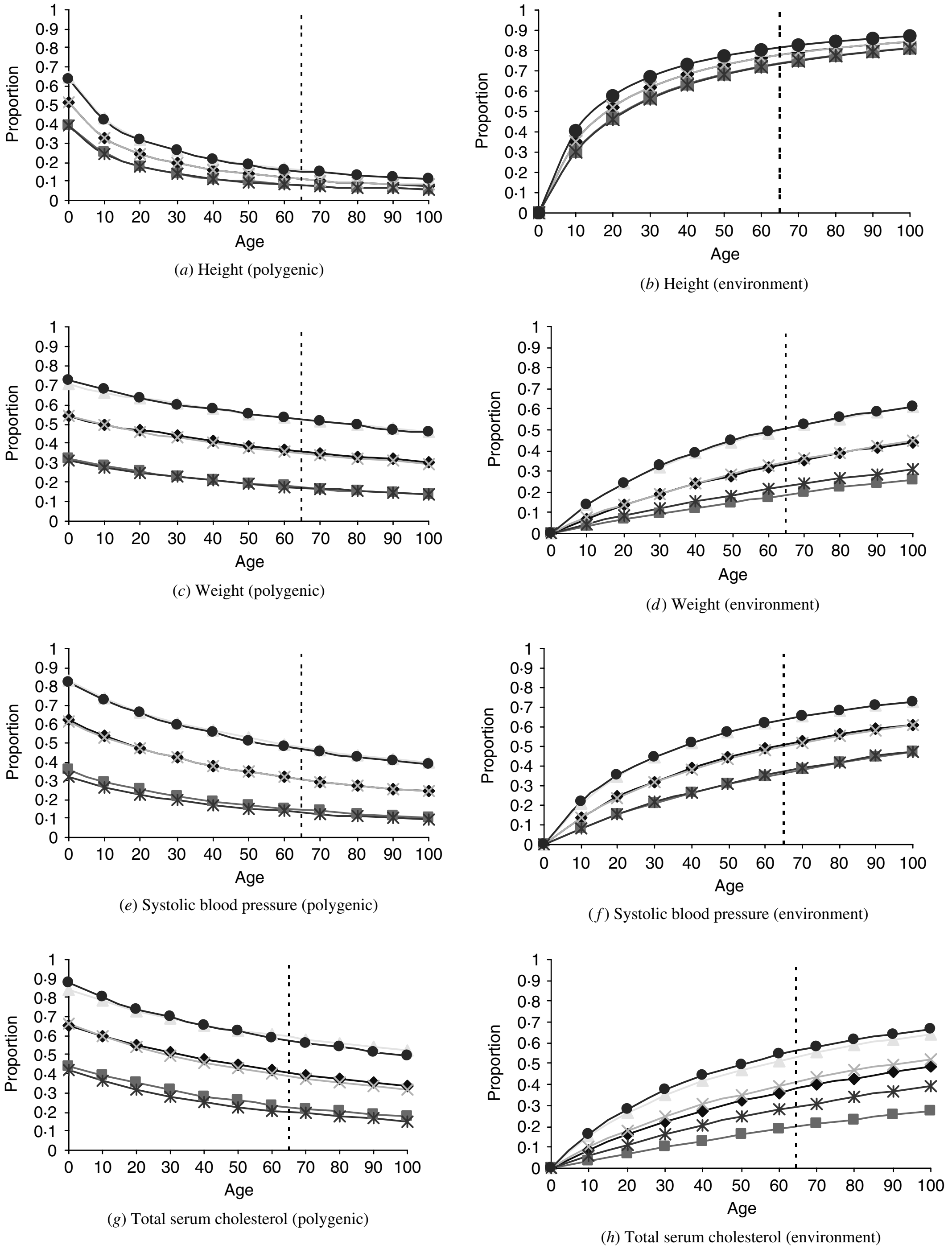
Phenotype	Model	$v_g(30)$	$v_e(30)$	$\tilde{v}_g$	$\tilde{v}_e$
Height (m)	(2)	0·19 (0·14, 0·25)	0·61 (0·56, 0·67)	0·89 (0·86, 0·92)	0·10 (0·08, 0·13)
	(3)	0·19 (0·14, 0·25)	0·61 (0·56, 0·67)	0·89 (0·86, 0·91)	0·10 (0·08, 0·13)
Weight (kg)	(2)	0·43 (0·23, 0·60)	0·18 (0·095, 0·31)	0·71 (0·60, 0·80)	0·29 (0·19, 0·39)
	(3)	0·42 (0·22, 0·59)	0·19 (0·11, 0·32)	0·71 (0·60, 0·82)	0·28 (0·17, 0·39)
SystBP (mmHg)	(2)	0·42 (0·21, 0·60)	0·32 (0·21, 0·44)	0·80 (0·70, 0·90)	0·19 (0·09, 0·29)
	(3)	0·42 (0·19, 0·59)	0·31 (0·21, 0·44)	0·80 (0·69, 0·90)	0·19 (0·09, 0·30)
Cholesterol (mg/dl)	(2)	0·51 (0·31, 0·69)	0·22 (0·10, 0·34)	0·76 (0·64, 0·88)	0·23 (0·11, 0·35)
	(3)	0·49 (0·27, 0·69)	0·24 (0·16, 0·37)	0·77 (0·65, 0·91)	0·22 (0·08, 0·34)

We analyse 372 ($=n$) household data each of which consists of four observations as shown in Figure 1. For the purpose of illustration, four phenotypes are analysed, *viz*. body height, weight, systolic blood pressure and total serum cholesterol. The covariates used are age when the phenotypes were measured, sex (1 = man, 0 = woman) and location (categorized as urban and rural). The total number of observations on the phenotypes is 1488. For cholesterol, 167 observations were missing, while for systolic blood pressure, four observations were missing. All the observations were available for height and weight. A natural question that arises is whether the correlation between phenotypes of spouses is larger for spouses that have been together for longer? To check this, a regression analysis of the phenotype of the factory worker y_{i1} on the age when the phenotype was collected, the phenotype of the spouse y_{i2}, the duration they have lived together $D_i(1, 2)$ and the interaction $y_{i2} * D_i(1, 2)$ gave positive regression coefficients for the phenotype of the spouse, and the duration they have lived together. Hence, it may be concluded that the spouse correlation is larger for spouses that have been together longer. We also looked at the phenotypic variances among those below and above 40 years of age to see whether phenotypic variances increase with age and by grouping the pairs of observations, $((y_{i1}, y_{i2}), (y_{i1}, y_{i3}), (y_{i2}, y_{i4}))$, according to the number of years spent in 5-year groups. We noticed that in our data, the phenotypic variances among those who are below 40 years of age are smaller than those above 40 years of age. There was a clear increase in the phenotypic covariations in the initial years of living together (up to 10 years) and for the later years no clear pattern was observed. This may indicate a more general model, where the shared environmental effect decreases with age (see the Discussion section).

In Appendix B, OpenBUGS code for analysing the proposed model is given. The model (1) gave an estimate of Σ with positive covariance between individuals 1 and 4. This is not desirable since these individuals neither are genetically related nor have a shared environment. Hence, this may indicate a possible lack of fit for models (2) and (3), which assume a specific structure of dependency under random mating (see the Discussion section).

Table 1 gives the posterior median and 95% credible intervals for three variance components and for four phenotypes using models (2) and (3) and Table 2 gives corresponding estimates of the

https://doi.org/10.1017/S0016672308009348 Published online by Cambridge University Press



(*a*) Height (polygenic)

(*b*) Height (environment)

(*c*) Weight (polygenic)

(*d*) Weight (environment)

(*e*) Systolic blood pressure (polygenic)

(*f*) Systolic blood pressure (environment)

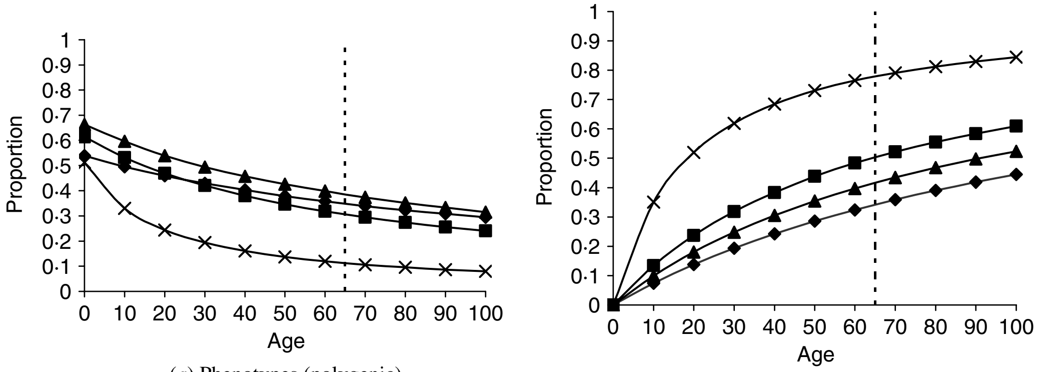
(*g*) Total serum cholesterol (polygenic)

(*h*) Total serum cholesterol (environment)

Fig. 2. Posterior median (middle curves) and 95 % credible intervals (upper and lower curves) of the proportion of total variation due to polygenic and shared environmental sources for various phenotypes using models with 0·5 as the constant degree of genetic relationship (2) and when the degree of genetic relationship is random (3). (*a*) Height – polygenic, (*b*) height – environmental, (*c*) weight – polygenic, (*d*) weight – environmental, (*e*) systolic blood pressure – polygenic, (*f*) systolic blood pressure – environmental, (*g*) total serum cholesterol – polygenic and (*h*) total serum cholesterol – environmental. The 2·5 % limit, median and the 97·5 % limit are denoted by square, polygon and triangle symbols, respectively, for model (2) and by star, cross and polygon symbols, respectively, for model (3). The vertical line corresponds to the age limit of 65 in the data.

proportions of variability due to polygenic effect and environmental effect at the age of 30 years along with their crude estimates. The models show similar results except for slight differences for cholesterol. This is due to the missing data for cholesterol. It is our observation that the convergence was faster with model (3) when there were missing data for phenotypes. For all the phenotypes considered, the

https://doi.org/10.1017/S0016672308009348 Published online by Cambridge University Press



(*a*) Phenotypes (polygenic)

(*b*) Phenotypes (environment)

Fig. 3. Posterior median of the proportion of total variation due to (*a*) polygenic and (*b*) shared environment sources for various phenotypes when the degree of genetic relationship is random (cross, height; polygon, weight; triangle, total serum cholesterol; square, systolic blood pressure). The vertical line corresponds to the age limit of 65 in the data.

variation due to the shared environment is less compared to the polygenic and residual variability. This is because the environmental variation is not simply σ_e^2 but with a multiplier D_i. The latter are rather close for height and weight, while polygenic variability is higher than the residual variability for systolic blood pressure and cholesterol.

Figures 2(a)–(h) show the posterior median and 95% credible intervals for the proportions v_g(age) and v_e(age) for different phenotypes and for different values of age. The curves due to the two models are overlapping in most cases. It is clear from Figure 3(a), which shows the posterior medians of v_g(age), that the trend in the proportion is similar between weight, systolic blood pressure and cholesterol, while for height there is sudden decrease with age. This is because participants reach their maximal height in early adult life and from then onwards there is no further growth. Till age 20, the proportion of total variability due to polygenic effect is higher in systolic blood pressure compared to weight, but after the age of 30 the trend reverses. Similarly, v_e(age) are shown in Figure 3(b) where height behaves differently compared to other three phenotypes. We also carried out the analysis using log-transformation for height and the proportions of total variations due to polygenic and environmental effects were close to the results obtained without log-transformation.

The posterior median and 95% credible intervals are given for some of the regression parameters in Table 3. Only the sex effect is seen on height with men tending to be taller than women. Body weight is affected by age and sex. With age, body weight increases and men have a higher weight compared to women. A similar trend is seen for systolic blood pressure with respect to age and sex. The regression coefficient of sex is negative for cholesterol, indicating that women have higher cholesterol compared to men.

5. Discussion

The household data with a typical structure of a married couple and their siblings are analysed so that the built-in dependency structure is used in specifying the variance–covariance structure. The dependency structure allows separation of overall variability into polygenic and shared environmental components. Like McGregor *et al.* (2003), our analysis represents multivariate analysis of heritability in the sense that all age strata are analysed jointly compared to age-stratified heritability estimation (Brown *et al.*, 2003). In our analyses, a general model that does not take into account the dependency structure but allows any positive definite Σ_i showed weak association between the observations (y_{i1}, y_{i4}), which may indicate a small amount of assortative mating present among the study subjects. In the proposed model, we assume that these two individuals are neither related nor have lived in the same household and the variability is directly proportional to the number of years lived in the same household. This reflects our understanding of how environmental exposures operate.

The model used in this paper is the first possible simplest model for the sibling pairs and spouse pairs association separating environmental and genetic variations. The unexplained variation is expressed in terms of the residual variance σ_m^2. A more complex model allowing different dependencies between the sibling pairs and spouse pairs based on the age when lived together and allowing covariate information like socio-economic status (SES) influencing the traits/phenotypes might describe the data better. One such model could be developed under the assumption that the environmental dependency is stronger if the environment is shared early in life (that is, at a younger age). For example, a more complex modelling that allows the effect of the shared environment on the phenotypes to be larger in the initial years of

https://doi.org/10.1017/S0016672308009348 Published online by Cambridge University Press

Table 3. *Estimates of regression coefficients. Posterior median* (*95% credible intervals*) *estimates for the regression parameters* (*overall mean, age and male*) *for various quantitative phenotypes. The abbreviation* '*SystBP*' *is used for systolic blood pressure*

	Height (m)	Weight (kg)	SystBP (mmHg)	Cholesterol (mg/dl)
μ	1·57	51·58	92·7	140·7
	(1·55, 1·58)	(48·8, 54·47)	(88·57, 96·7)	(130·5, 151·2)
Age	−0·0009	0·20	0·69	1·08
	(−0·001, −0·0005)	(0·14, 0·27)	(0·60, 0·78)	(0·85, 1·32)
Male	0·13	5·27	4·99	−3·19
	(0·12, 0·14)	(4·21, 6·36)	(3·44, 6·58)	(−7·09, 0·82)

life and to decrease with age could be carried out using appropriate functions, for example $\exp\{-\lambda t\}$. The shared environment variance is then computed as the area under the function with respect to the Gaussian process. We may also need different models for different traits since the proposed model gives counter-intuitive results for height and the development of phenotypes with age could be different. The heritability obtained without adjusting for age (Table 2, $\tilde{v}_g$) is more comparable to the results reported from the twin studies, but the age-adjusted proportions are rather different. However, we must remember that modelling polygenic and environmental variations as time-dependent quantities is very rare and we must not compare mere numbers in the absolute sense. We think that the proposed approach opens up methodological questions and provides scope for further research in this area.

The estimates of heritability obtained here are not directly comparable to the ones reported in the literature. In the context of cholesterol, Mastropaolo *et al.* (2001) pointed out that previous studies consist of older children and adult twin pairs have indicated environmental contribution to the variation of cholesterol among individuals of 7–68% in the populations evaluated. Our results are to be compared with the previously reported results keeping this in mind. In this paper, we defined heritability as a decreasing function of age, where its value is decreasing at a rate depending on the ratio of environmental and genetic variances. Thus, it is not surprising that height seems to be less genetic at age 100 than at age 20. However, this trend may be more pronounced if the information relevant to early life environment comes from siblings who have less variability in trait and later-life information comes more from spouses who will have greater variability.

The proposed method allows a comparison of various phenotypes between rural and urban areas by incorporating an appropriately defined covariate based on migration status. A more thorough analysis focusing on the rural and urban comparisons would be reported elsewhere since the main interest here is in estimating the relative genetic and environmental variations.

In summary, we have developed a model that uses the dependency between sibling pairs and spouse pairs to estimate the age-dependent proportions of total variation that are due to polygenic and environmental effects. Using the Bayesian approach, the variation of IBD-sharing could be incorporated into the model without using marker information and this is the first time such an idea has been tried. The approximate $N(0{\cdot}5,\ 0{\cdot}038^2)$ for IBD shared by siblings is prior information based on genomewide analysis from previous studies. The Bayesian method can sample the realized IBD from the conditional posterior distribution. This explains why the variation of IBD can be incorporated without using marker information. This idea is novel and appears to be useful in separating the genetic from environmental variances. The model would benefit from testing in populations who have experienced different exposures. As pointed out to us by the Editor, using spouse information to quantify the common environmental variance may be better than using the twin data because the twin data are hard to collect, while spouse and sibling data are easier to collect. We may have very large samples for spouse–sibling data compared to the twin data.

This work was done while the first author was a visiting researcher at the Department of Mathematics and Statistics, University of Helsinki. This work was supported by research grants (numbers 114786 and 202324) from the Academy of Finland. We would like to thank Dr Bijoy Joseph, Indic Society for Education and Development (INSEED), India for his help in preparing the data in the required format. We are grateful to the Editor and two anonymous referees for their constructive comments on the manuscript.

The IMS Group comprises: Professor K. Srinath Reddy, Dr Dorairaj Prabhakaran, Professor Tulsi Patel, Dr Lakshmy Ramakrishnan, Dr Ruby Gupta and Dr Tanica Lyngdoh (New Delhi, India); Professor R. C. Ahuja and Professor R. K. Saran (Lucknow, India); Dr Prashant Joshi and Dr N. M. Thakre (Nagpur, India); Dr K. V. R. Sarma, Professor S. Mohan Das, Dr R. K. Jain and Dr S. S. Potnis (Hyderabad, India); Professor Anura V. Kurpad, Dr Mario Vaz, Ms A. V. Barathi and Dr Murali Mohan (Bangalore,

https://doi.org/10.1017/S0016672308009348 Published online by Cambridge University Press

India); Dr Chittaranjan Yajnik (Pune, India); Dr Ian Baker, Professor George Davey Smith, Professor Yoav Ben Shlomo and Dr Kate Tilling (Bristol); and Professor Shah Ebrahim and Dr Sanjay Kinra (London School of Hygiene and Tropical Medicine).

The IMS is funded by Wellcome Trust project grant GR070797MF. We are grateful to the field staff conducting the migration study and to the participants.

Appendix A: Rationale behind the covariance structures

(i) *Covariance structure of polygenic effects*

Consider a single locus with s allele effects $(\eta_1,\ldots,\eta_s)$ We assume that these are independent and identically distributed normal variates with zero mean and variance σ_1^2. Let us suppose that the human genome contains K such independent loci. Then G_{i1} can be expressed as the sum of the effects of the $2K$ alleles transmitted by parents and let this be $\sum_{k=1}^{K}(\eta^1_{i1_k}+\eta^2_{i1_k})$. Similarly, let $G_{i3}=\sum_{k=1}^{K}(\eta^1_{i3_k}+\eta^2_{i3_k})$ be the genetic effect for a sibling of $i1$. It is easy to check that for $j=1, 3$,

$$\begin{aligned}E(G_{ij}|\text{parents}) &= 0 = E(G_{ij}),\\ \operatorname{var}(G_{ij}|\text{parents}) &= 2K\sigma_1^2=\sigma_g^2=\operatorname{var}(G_{ij}),\\ \operatorname{cov}(G_{i1},G_{i3}|\text{parents}) &= (\text{number of alleles shared between } i1 \text{ and } i3)\sigma_1^2\\ &= (\text{shared genome between } i1 \text{ and } i3)\sigma_g^2.\end{aligned}$$

The conditional distribution of (G_{i1}, G_{i3}) is the normal distribution with mean and covariance structure as specified above. The shared genome between the sibling pair is the proportion of IBD genes actually shared by them. The value 0·5 was originally obtained as an asymptotic limiting value by assuming infinite genome length. As stated in Xu (2006), the actual amount of IBD-sharing between human full siblings is not a constant but a variable with expectation equal to 0·5 and variance $(0{\cdot}0384)^2$. Here, the variance is specified by the length of the human genome. Hence, in our model, we allow the covariance between the polygenic effects of full siblings ij and ij' to vary by a factor $c_i(j, j')\in[0, 1]$ with mean 0·5 and variance $(0{\cdot}0384)^2$. For simplicity, during estimation we have assumed that $c_i(j, j')$ has a normal distribution with mean 0·5 and variance $(0{\cdot}0384)^2$.

(ii) *Covariance structure of environmental effects*

Under the assumption that the environment changes over time, we define

$$E_{ij}=\sum_l\int_0^\infty f(a(s))I_{ij}(s;\, l)\,\mathrm{d}W_{ijl}(s),$$

where $a(s)$ is the age at time s, $f(a(s))$ is a positive, monotone nonincreasing function of the age $a(s)$, $I_{ij}(s;\, l)$ is an indicator function assuming the value 1 if the individuals i and j have lived together at location l at time s and is zero otherwise, $\{W_{ijl}(s), s\in[0,\infty]\}$ is a Gaussian process, with expected value $E(W_{ijl}(s))=0$ for all (i,j,l) and s and covariance function $\operatorname{cov}(W_{ijl}(s_1), W_{ij'l}(s_2))=\sigma_e^2\min(s_1,s_2)$ and $\operatorname{cov}(W_{ijl}(s_1), W_{i'j'l'}(s_2))=0$, if $i\neq i'$, $l\neq l'$. Here, we take $f(a(s))=1$ but a more general function describing the model assumption of stronger environment if shared at younger ages can be modelled using, for example, $\exp\{-\lambda a(s)\}$ for $\lambda>0$.

The variance–covariance matrix is then

$$\begin{aligned}\Sigma_{ei}(j,j') &= \operatorname{cov}(E_{ij},E_{ij'})\\ &= \sum_l\int_0^\infty I_{ij}(s;\, l)I_{ij'}(s;\, l)\,\mathrm{d}\langle W_{ijl}(s), W_{ij'l}(s)\rangle\\ &= \sigma_e^2\sum_l D_i^l(j,j'),\end{aligned}$$

where $D_i^l(j, j')$ is time spent together by (j, j') at the location l.

https://doi.org/10.1017/S0016672308009348 Published online by Cambridge University Press

Appendix B: OpenBUGS code for analysing the proposed model

```
model
    {
        for( i in 1 : N ) {
            Y[i, 1 : 4] ~ dmnorm(mu[i, 1 : 4], T_m[i,1 : 4,1 : 4])
            for(j in 1 : 4) {
                mu[i, j] <- beta[1] + beta[2] * x1[i,j] + beta[3] * x2[i,j]
                          + beta[4] * equals(x3[i,j],1) + beta[5]
                          * equals(x3[i,j],2) + beta[6] * equals(x3[i,j],3)
            } # x1=age, x2=sex, x3=type of migration
        }
        for (i in 1:N) {
        T_m[i,1 : 4, 1 : 4] <- inverse (S_m[i,1 : 4,1 : 4])
# Uncomment following to use constant factor
#          c13[i] <- 0.5
#          c24[i] <- 0.5
# D is the duration matrix
           S_m[i,1,1] <- sigma_g2 + sigma_m2 + sigma_e2*D[4*i-3,1]
           S_m[i,1,2] <- sigma_e2*D[4*i-3,2]
           S_m[i,1,3] <- sigma_g2*c13[i] + sigma_e2*D[4*i-3,3]
           S_m[i,1,4] <- sigma_e2*D[4*i-3,4]
           S_m[i,2,1] <- sigma_e2*D[4*i-3+1,1]
           S_m[i,2,2] <- sigma_g2+sigma_m2+sigma_e2*D[4*i-3+1,2]
           S_m[i,2,3] <- sigma_e2*D[4*i-3+1,3]
           S_m[i,2,4] <- sigma_g2*c24[i] + sigma_e2*D[4*i-3+1,4]
           S_m[i,3,1] <- sigma_g2*c13[i] + sigma_e2*D[4*i-3+2,1]
           S_m[i,3,2] <- sigma_e2*D[4*i-3+2,2]
           S_m[i,3,3] <- sigma_g2+sigma_m2+sigma_e2*D[4*i-3+2,3]
           S_m[i,3,4] <- sigma_e2*D[4*i-3+2,4]
           S_m[i,4,1] <- sigma_e2*D[4*i-3+3,1]
           S_m[i,4,2] <- sigma_g2*c24[i] + sigma_e2*D[4*i-3+3,2]
           S_m[i,4,3] <- sigma_e2*D[4*i-3+3,3]
           S_m[i,4,4] <- sigma_g2 + sigma_m2 + sigma_e2*D[4*i-3+3,4]
        }

        for( i in 1 : N ) {
           c13[i] ~ dnorm(0.5, tau)
           c24[i] ~ dnorm(0.5, tau)
        }
        tau <- 1/(0.0384*0.0384)

        for (j in 1 : 6) {
                beta[j] ~ dnorm(0.0, 0.0001)
        }
        tau_m2 ~ dgamma(0.01,0.01)
        sigma_m2 <- 1 / tau_m2

        tau_g2 ~ dgamma(0.01,0.01)
        sigma_g2 <- 1 / tau_g2

        tau_e2 ~ dgamma(0.01, 0.01)
        sigma_e2 <- 1 / tau_e2
}
```

https://doi.org/10.1017/S0016672308009348 Published online by Cambridge University Press

References

Abney, M., McPeek, M. S. & Ober, C. (2000). Estimation of variance components of quantitative traits in inbred populations. *American Journal of Human Genetics* **66**, 629–650.

Brown, W. M., Beck, S. R., Lange, E. M., Davis, C. C., Kay, C. M., Langefelt, C. D. & Rich, S. S. (2003). Age-stratified heritability estimation in the Framingham Heart Study families. *BMC Genetics* **4** (Suppl 1), S32.

Dongen, S. V. (2006). Prior specification in Bayesian statistics: three cautionary tales. *Journal of Theoretical Biology* **242**, 90–100.

Du, F.-X. & Hoeschele, I. (2000). Estimation of additive, dominance and epistatic variance components using finite locus models implemented with a single-site Gibbs and a descent graph sampler. *Genetical Research* **76**, 187–198.

Du, F.-X., Hoeschele, I. & Gage-Lahti, K. M. (1999). Estimation of additive and dominance variance components in finite polygenic models and complex pedigrees. *Genetical Research* **74**, 179–187.

Eaves, L. J., Martin, N. G., Meyer, J. M. & Corey, L. A. (1999). Biological and cultural inheritance of stature and attitudes. In *Personality and Psychopathology* (ed. C. R. Cloninger). Washington, DC: American Psychiatric Press.

Falconer, D. S. (1989). *Introduction to Quantitative Genetics*, 3rd edn. Harlow: Longman.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515–533.

Guo, S. W. (1996). Variation in genetic identity among relatives. *Human Heredity* **46**, 61–70.

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. Guelph, ON, Canada: University of Guelph.

Hopper, J. L. (2000). Why 'common environmental effects' are so uncommon in the literature. In *Advances in Twin and Sib-Pair Analysis* (ed. T. M. Spector, H. Snieder & A. J. MacGregor), Chapter 13. London: Greewich Medical Media Ltd.

Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.

Lyngdoh, T., Kinra, S., Shlomo, Y. B., Reddy, S., Phabhalaran, D., Smith, G. D. & Ebrahim, S. for the Indian Migration Study Group (2006). Sib-recruitment for studying migration and its impact on obesity and diabetes. *Emerging Themes in Epidemiology* **3**, 2.

Mastropaolo, W., Matheny, A. & Lang, C. A. (2001). Plasma cholesterol concentrations in twin children: estimates of genetic and environmental influences. *Clinical Chemistry* **47**, 771.

McGregor, S., Knott, S. A., White, I. & Visscher, P. M. (2003). Longitudinal variance – components analysis of the Framingham Hearth Study data. *BMC Genetics* **4** (Suppl 1), S22.

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., Dunkel, L., De Lange, M., Harris, J. R., Hjelmborg, J. V., Luciano, M., Martin, N. G., Mortensen, J., Nisticò, L., Pedersen, N. L., Skytthe, A., Spector, T. D., Stazi, M. A., Willemsen, G., Kaprio, J. (2003). Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Research* **6**, 399–408.

Spiegelhalter, D. A., Thomas, A., Best, N. & Lunn, D. (2005). *WinBUGS User Manual. Version 2.10*. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Stoel, R. D., De Geus, E. J. C. & Boomsma, D. I. (2006). Genetic analysis of sensation seeking with an extended twin design. *Behavior Genetics* **36**, 229–237.

Thomas, A., O'Hara, R. B., Ligges, U. & Stuartz, S. (2006). Making BUGS open. *R News* **6/1**, 17–21.

Thompson, E. A. & Skolnick, M. H. (1977). Likelihoods on complex pedigrees for quantitative traits. In *Proceedings of the International Conference on Quantitative Genetics* (ed. E. Pollack, O. Kempthorne & T. B. Bailey, Jr), pp. 815–818. Ames, IA: Iowa State University Press.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., Martin, N. G. (2006). Assumption-free estimation of heritability from genomewide identity-by-descent sharing between full siblings. *PLoS Genet* **2**, 0316–0325.

Xu, S. (2006). Separating nurture from nature in estimating heritability. *Heredity* **97**, 256–257.

https://doi.org/10.1017/S0016672308009348 Published online by Cambridge University Press