Stat Biosci (2009) 1: 181–198
DOI 10.1007/s12561-009-9010-5

A Variance-Component Framework for Pedigree Analysis of Continuous and Categorical Outcomes

Michael P. Epstein · Jessica E. Hunter · Emily G. Allen · Stephanie L. Sherman · Xihong Lin · Michael Boehnke

Received: 25 June 2009 / Accepted: 24 September 2009 / Published online: 8 October 2009

Abstract Variance-component methods are popular and flexible analytic tools for elucidating the genetic mechanisms of complex quantitative traits from pedigree data. However, variance-component methods typically assume that the trait of interest follows a multivariate normal distribution within a pedigree. Studies have shown that violation of this normality assumption can lead to biased parameter estimates and inflations in type-I error. This limits the application of variance-component methods to more general trait outcomes, whether continuous or categorical in nature. In this paper, we develop and apply a general variance-component framework for pedigree analysis of continuous and categorical outcomes. We develop appropriate models using generalized-linear mixed model theory and fit such models using approximate maximum-likelihood procedures. Using our proposed method, we demonstrate that one can perform variance-component pedigree analysis on outcomes that follow any exponential-family distribution. Additionally, we also show how one can modify the method to perform pedigree analysis of ordinal outcomes. We also discuss extensions of our variance-component framework to accommodate pedigrees ascertained based on trait outcome. We demonstrate the feasibility of our method using both simulated data and data from a genetic study of ovarian insufficiency.

Keywords Variance component model · Linkage analysis · Generalized linear mixed model

M.P. Epstein (✉) · J.E. Hunter · E.G. Allen · S.L. Sherman
Department of Human Genetics, Emory University, 615 Michael Street, Suite 301, Atlanta, GA 30322, USA
e-mail: mpepste@emory.edu

X. Lin
Department of Biostatistics, Harvard University, Boston, MA, USA

M. Boehnke
Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA

1 Introduction

Mixed-model procedures (see [34] for an overview) have an impressive history in the statistical analysis of clustered, hierarchical, and spatial data. Researchers also have applied mixed models extensively to perform genetic analyses of correlated trait data from relatives within pedigrees. Initially, studies applied such mixed models to familial trait data to assess whether a quantitative trait of interest contained a significant genetic component and was heritable [22, 30]. Such analyses consist of partitioning the trait variance within a family into estimated components due to separate genetic and environmental effects and subsequently examining whether the estimated component of variance due to the genetic effect has a significant impact on the trait of interest. Amos [5] and Almasy and Blangero [4] later extended this variance-component (VC) method to perform linkage analysis of quantitative trait data by modeling a separate component of variation due to a putative major-gene locus of interest. Additional studies demonstrated further extensions of the VC framework to allow for interaction testing [35] as well as outcome data from longitudinal studies [11].

The VC method has many appealing practical features for pedigree analysis of quantitative trait data. Compared to relative-pair based methods (e.g. [20, 26]), VC methods utilize a framework that is more flexible and permits easier modeling of covariates and interaction effects. More importantly, many studies have shown that variance-component methods often have more power than relative-pair methods for detecting linkage [41, 48, 49]. Nevertheless, VC methods require strong assumptions for valid inference. In particular, traditional VC methods typically rely on the assumption that the familial trait data follow a multivariate normal distribution in order to facilitate analyses. One can test this assumption using diagnostic tools such as those described in [22] and [10]. If the distributional assumption is violated, the traditional variance-component method may yield biased parameter estimates and elevated type-I error rates for testing different effects. For example, Allison et al. [3], Blangero et al. [6], and Epstein et al. [17] demonstrated these problems for quantitative data derived from Laplace, χ^2, and censored normal distributions, respectively. Moreover, the normality assumption severely hinders the potential application of variance-component pedigree analysis to categorical outcomes, such as presence/absence of disease (binary outcome) or a disease-severity scale (ordinal outcome).

Even if one can transform the trait data to approximate normality, Allison et al. [3] and Blangero et al. [6] demonstrated the traditional variance-component method still yields inference problems when the transformed data have a standardized kurtosis greater than zero, which indicates more probability in the tails of the distribution than for a normal distribution. This finding impedes our interpretation of results for data that naturally follow a gamma distribution (Fig. 1). Typically, variance-component analysis of gamma-distributed data proceeds by first applying an appropriate normality transformation (such as a logarithmic transformation) and then applying the traditional variance-component method to the transformed data. However, the transformed data may still have a positive kurtosis (as shown in Fig. 2). In such a situation, one could perhaps find a better transformation for the data that reduces the kurtosis; however, such a transformation could be difficult to obtain.

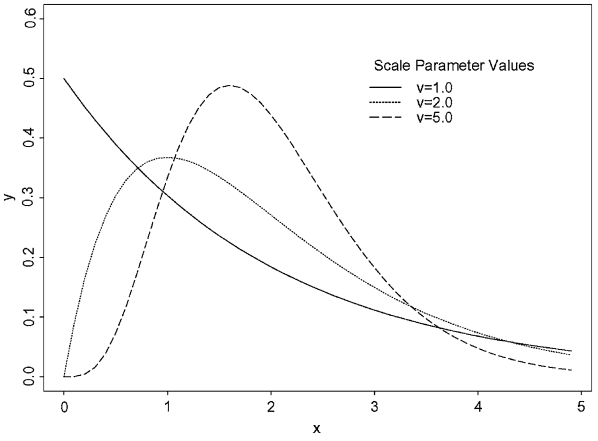

Fig. 1 Gamma(2, v) distribution

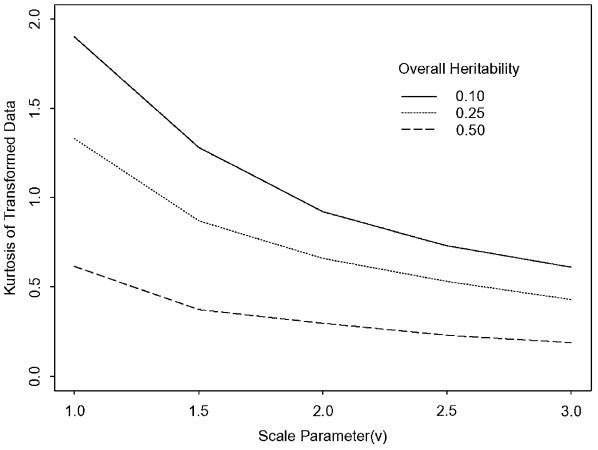

Fig. 2 Kurtosis of Gamma(2, v) data transformed to approximate normality

If the assumptions of the traditional variance-component method are violated, one can apply robust variance-component methods that allow for trait distribution misspecification [5, 6, 12] although it is unclear whether such methods are applicable to categorical outcomes. Moreover, if the trait distribution is approximately known, modeling the distribution could lead to increased power and more efficient parameter estimates relative to the robust methods. In an earlier paper [17], we extended the variance-component method to test for linkage of a major gene that influences censored normal data using generalized linear mixed model (GLMM) theory [7]. Results indicated that linkage analyses of censored normal data using our tobit variance-component method had improved efficiency and more appropriate type-I error rates compared to application of traditional variance-component methods that explicitly assumed trait normality.

Given this success, we now develop a general variance-component framework for pedigree analysis of non-normal data that follow an exponential-family distribution, which includes Bernoulli, Poisson, and gamma distributions as special cases. Additionally we also develop a VC framework for pedigree analysis of ordinal categorical

data. Like the tobit VC method, we base this framework on the GLMM framework proposed by Breslow and Clayton [7]. Our proposed method generalizes the work of Duggirala et al. [13, 14] and Hasstedt [21], who analyzed binary and polychotomous traits, respectively, using VC methods, but assumed trait values were determined by liability values being above or below specific thresholds. Our method is also related to the work of Burton et al. [8], who applied GLMM theory to analyze binary trait data but did not test for linkage. Further, the authors employed Gibbs–Sampling procedures for inference whereas we employ approximate maximum-likelihood procedures for this purpose.

We organize the rest of this article as follows. We first describe the general VC framework for pedigree analysis of an arbitrary trait with an exponential-family distribution and subsequently extend the framework to accommodate ordinal data. We next discuss approximate maximum-likelihood procedures for fitting our VC models followed by discussion of appropriate hypothesis testing of heritability and linkage for an outcome of interest. We use simulated pedigree data to evaluate our VC framework and further illustrate our approach with an application to a genetic study of ovarian insufficiency. We then make some concluding remarks about our approach and discuss potential extensions.

2 Generalized Linear Mixed Models for Pedigree Analysis

2.1 Likelihood Derivation Using Generalized Linear Mixed Models

Consider a family of n relatives. Let y_j be the trait value of the jth relative and $y = (y_1, y_2, \ldots, y_n)$ be the trait data for the family. We assume y_j originates from the sum of independent effects due to observed and unobserved factors both genetic and environmental in nature. Observed factors consist of fixed covariates such as age and gender. We denote X_j as a vector of such fixed factors (covariates) for the jth relative. We assume the unobserved random factors that influence y_j consist of a major-gene locus of interest and a number of independent genes of small effect (polygenes). While we will not do so here, we easily could assume other random effects such as two or more major genes, gene–gene interaction, or unobserved shared environment. We assume the alleles of the major gene and polygenes act additively on the trait; the additive-alleles assumption is easily relaxed. We define MG_j and PG_j, to be the additive allelic effects of the unobserved major gene and polygenes for the jth relative. We assume MG_j and PG_j, are independent and normally distributed with means zero and variances σ^2_{mg} and σ^2_{pg}, respectively. Finally, we denote $U_j = MG_j + PG_j$ as the total unobserved genetic effects for the jth relative and $U = (U_1, U_2, \ldots, U_n)$ as the set of unobserved genetic effects for the family. Conditional on U, the familial trait values $y = (y_1, y_2, \ldots, y_n)$ are independent.

We construct the likelihood of the family's trait data $L(y_1, y_2, \ldots, y_n)$ as a function of the fixed and unobserved effects. We first integrate this likelihood across the unobserved genetic effects U such that

$$L(y_1, y_2, \ldots, y_n) = \int \prod_{j=1}^{n} L(y_j|U)L(U)\,dU. \tag{1}$$

We assume a subject's trait value conditional on the unobserved genetic effects follows an exponential-family distribution such that $L(y_j|U) = f(\mu_j, \varphi)$, where $f(\cdot)$ is an exponential-family density function, $\mu_j = E[y_j|U]$ is the conditional trait mean, and φ denotes nuisance parameters. The exponential-family distributions include gamma, binomial, Poisson, and normal distributions [33].

To model the relationship between $y_j|U$ and both the fixed and random effects, we employ a link function g that models μ_j on X_j and U_j such that

$$g(\mu_j) = X_j^t \beta + U_j. \tag{2}$$

Here, β denotes a vector of regression coefficients for the covariates X_j. For simplicity, we assume X_j contains an intercept. Once we specify the link function g in (2), we can reparameterize the likelihood in (1) as a function of β and U_j.

The specification of both $f(\mu_j, \varphi)$ and $g(\mu_j)$ depends on the trait distribution. For a normally-distributed trait, $f(\mu_j, \varphi)$ is a probability density function for a normal distribution with mean μ_j and variance σ_e^2, while the link function is the identity link $g(\mu_j) = \mu_j$. For a binary trait, $f(\mu_j, \varphi)$ is a probability mass function for a Bernoulli distribution with mean μ_j and the link function is the logit link $g(\mu_j) = \text{logit}(\mu_j)$. For a gamma-distributed trait, $f(\mu_j, \varphi)$ is a probability density function for a gamma distribution with mean μ_j and scale parameter ν, while $g(\mu_j)$ is either a log link function $g(\mu_j) = \log(\mu_j)$ or a reciprocal link function $g(\mu_j) = \mu_j^{-1}$. One typically applies the log link function, since the reciprocal link function can lead to unstable estimation procedures. We note that examples of $f(\mu_j, \varphi)$ and $g(\mu_j)$ for other distributions are found in [33].

The final step in constructing likelihood (1) is specification of the distribution of the unobserved genetic effects U. As the genetic effects will induce similarity among relatives, there will be covariance among the U_j for different relatives in a family. As shown by [24] and [22], we can write the covariance for two non-inbred relatives j and k as

$$\text{Cov}(U_j, U_k) = \sigma_{mg}^2 \cdot \pi_{jk} + \sigma_{pg}^2 \cdot 2\phi_{jk}. \tag{3}$$

Here, π_{jk} denotes the proportion of alleles shared identical by descent (IBD) at the major gene by j and k. A relative pair shares two alleles IBD at a locus if the alleles are physical copies of the same ancestral allele. For autosomal loci, the proportion of alleles shared IBD by a relative pair at a particular locus is equivalent to the number of alleles shared IBD by the pair divided by 2. Generally, we cannot observe π_{jk} but can estimate it using a multipoint algorithm based on available genetic marker data (for example, the algorithms of [27] and [18]). $2\phi_{jk}$ is the expected proportion of genes shared IBD by the relative pair, where ϕ_{jk} is called the kinship coefficient [24]. For any relative pair, $2\phi_{jk}$ is a known function of the relationship of the pair. For example, $2\phi_{jk} = 1$ for monozygotic-twin pairs, 0.5 for full-sib pairs and parent-offspring pairs, 0.25 for half-sib pairs and avuncular pairs, and 0.125 for full-cousin pairs.

Using the covariance structure in (3), we assume that the random genetic effects U follow a multivariate normal distribution with mean vector 0 and covariance matrix

$$\Sigma = \sigma_{mg}^2 \cdot \Pi + \sigma_{pg}^2 \cdot 2\Phi. \tag{4}$$

Here, Π and 2Φ are $n \times n$ matrices with (j,k)th elements π_{jk} and $2\phi_{jk}$, respectively. VC studies typically assume a multivariate normality assumption for U and this assumption certainly will hold if the underlying genes act additively and independently on the trait [28, 29].

2.2 Extension to Ordinal Data

While ordinal data do not follow one of the traditional exponential-family distributions described in the previous section, we can still implement a variance-component procedure for pedigree analysis of such categorical data. Suppose y_j denotes an ordinal outcome that takes one of M possible values. We assume there is a clear ordering of these values such that, for example, larger outcome values would denote increased severity compared to smaller outcome values. We model the ordinal outcome using a variation of the proportion-odds model as described in [32]. Let $p_j^{(m)} = P[y_j = m|U_j]$ denote the probability that relative j has ordinal outcome $m (m = 1, \ldots, M)$ conditional on the random genetic effects and define $\gamma_j^{(m)} = P[y_j \leq m|U_j] = \sum_{t=1}^{m} p_j^{(t)}$ as the corresponding cumulative probability that relative j has an ordinal outcome in the range between 1 and m. Using these definitions, we can model the ordinal data by fitting $M-1$ proportional-odds models

$$\log\left(\frac{\gamma_j^{(m)}}{1-\gamma_j^{(m)}}\right) = \theta_m + X_j^t\beta + U_j \quad (1 \leq m \leq M-1), \tag{5}$$

where θ_m denotes a specific intercept for category m. Here, we assume the same slope β for each of the $M-1$ proportional-odds models. If interest exists, we could also generalize the model in (5) to model different slopes $\beta^{(m)}$ for each category, which corresponds to fitting $M-1$ cumulative logit models.

Using the model in (5), we can write the likelihood in (1) by noting that the subject-specific piece of the likelihood corresponds to

$$L(y_j = m|U_j) = \begin{cases} \gamma_j^{(m)} - \gamma_j^{(m-1)} & m > 1, \\ \gamma_j^{(1)} & m = 1, \end{cases} \tag{6}$$

which is a function of parameters (β, θ) in (5). We complete specification of the likelihood (1) for ordinal data by assuming (as done previously) that the random genetic effects U follow a multivariate normal distribution with the covariance structure shown in (4).

2.3 Inference Methods

For a sample of I independent families, we construct the full likelihood as $L_F = \prod_{i=1}^{I} L_i$, where L denotes the likelihood in (1) and i indexes family. We use L_F to obtain estimates of fixed and random effects that maximize the likelihood. If the trait follows a normal distribution, then the integral in likelihood (1) has a closed-form solution and maximum-likelihood inference [30] is straightforward. However,

if the trait follows a non-normal distribution, the integral does not have a closed-form solution, which complicates inference.

To resolve the intractability of the integral for non-normal data, statisticians have developed inference methods that maximize an approximate version of L_F. Such procedures include the Solomon–Cox [44] approximation, penalized quasi-likelihood [7], and Gibbs sampling [8, 50]. Here, we apply an approximate maximum-likelihood approach for inference called adaptive Gaussian quadrature [38] that is implemented in the SAS [39] procedure NLMIXED. This method approximates the integral in L_F using a weighted sum over predefined weights for U_i (Gauss–Hermite integration). Adaptive Gaussian quadrature then maximizes the approximate likelihood using one of several optimization algorithms. We use a quasi-Newton maximization procedure, which is the default algorithm implemented in NLMIXED.

Accurate likelihood approximation in adaptive Gaussian quadrature requires a suitable set of evaluation points (known as quadrature points) and their corresponding weights. In SAS NLMIXED, one can either directly choose the number of quadrature points for analysis or let the program adaptively select the appropriate number. For the latter procedure, NLMIXED selects the number of quadrature points by evaluating the log-likelihood function at its initial parameter values for increasing number of quadrature points until two successive evaluations has a relative difference smaller than 0.0001. Unless otherwise noted, we use this adaptive procedure to select the number of quadrature points for pedigree analysis to ensure a likelihood that is well approximated. We note that adaptive Gaussian quadrature with one quadrature point corresponds to a Laplace approximation [7], which is computationally fast but may not accurately approximate the likelihood. As the number of quadrature points increases, the likelihood approximation becomes more accurate, but the complexity of the maximization algorithm also increases, leading to longer computer run times.

2.4 Testing the Heritability and Linkage Hypotheses

To test whether a trait outcome is heritable within the family sample, we perform the hypothesis test $H_0 : \sigma^2_{pg} = 0$ vs. $H_A : \sigma^2_{pg} > 0$. We test this hypothesis by calculating the likelihood ratio statistic, $2\log_e(\hat{L}_A/\hat{L}_0)$, where $\hat{L}_A$ and $\hat{L}_0$ are the maxima of L_F fit under the alternative and null hypotheses. Under the null hypothesis, σ^2_{pg} is set to 0 while, under the alternative hypothesis, σ^2_{pg} is estimated together with the other unknown parameters (σ^2_{mg} is not modeled under either hypothesis). As the value of σ^2_{pg} under the null hypothesis is on the boundary of the parameter space, the likelihood ratio statistic is asymptotically distributed as a 1/2:1/2 mixture of χ^2_1 and a point mass of zero under the null hypothesis [40]. As an alternative to a likelihood-ratio test, we also can employ a score statistic [31] that has the appealing feature of robustness in the presence of random-effect distribution misspecification.

The hypothesis test for linkage at the major gene is $H_0 : \sigma^2_{mg} = 0$ vs. $H_A : \sigma^2_{mg} > 0$. We test this hypothesis by calculating a likelihood ratio statistic, which is also asymptotically distributed as a 1/2:1/2 mixture of χ^2_1 and a point mass of zero under the null hypothesis. Under the null hypothesis, σ^2_{mg} is set to 0 while, under the alternative hypothesis, σ^2_{mg} is estimated together with the other unknown parameters.

2.5 Adjusting for Non-random Sampling

One often employs non-random sampling in genetic studies to increase the information content of the sample. For a family-based analysis of a rare genetic disease, a common ascertainment-sampling scheme is to collect families with at least one or at least two affected relatives. For a quantitative trait, a common sampling scheme is to collect families with a proband whose trait value is found in one or the other tail of the population distribution.

If one applies the random-sampling L_F to ascertained data, results often will be biased [9]. Therefore, given non-random sampling, we adjust L_F to account for the ascertainment criterion by dividing the unconditional likelihood by the probability of the ascertainment event. Let ASC_i denote the ascertainment criterion for family i. The proper ascertainment-adjusted likelihood for this family takes the form:

$$L(y_{i1}, y_{i2}, \ldots, y_{in_i}|ASC_i) = \frac{L(y_{i1}, y_{i2}, \ldots, y_{in_i})}{L(ASC_i)} = \frac{\int \prod_{j=1}^{n_i} L(y_{ij}|U_i)L(U_i)\,dU_i}{\int L(ASC_i|U_i)L(U_i)\,dU_i} \tag{7}$$

and the ascertained-adjusted likelihood for I sampled families takes the form $L_{F,ASC} = \prod_{i=1}^{I} L_{i,ASC}$, where $L_{i,ASC}$ denotes the ascertained-adjusted likelihood for the ith family shown in (7). Using this ascertainment-adjusted likelihood, we expect to obtain unbiased population-based estimates of $(\underline{\beta}, \sigma^2_{mg}, \sigma^2_{pg}, \varphi)$ [16, 19], although certain situations may occur where the estimates are not identifiable; an example of this for binary data can be found in [15].

3 Simulation Studies

3.1 Simulation Design

We used simulated pedigree data to examine bias, type-I error, and power of our general VC framework for genetic analysis of both continuous and categorical outcomes. We assumed variable number of simulated pedigrees consisting of sibships of various sizes. For a given sibship, we simulated marker data by inserting a major gene at 55 cM on a 110 cM chromosome. We simulated a 10 cM map of 12 genetic markers each with four equally-frequent alleles. For each locus, we randomly assigned alleles to the parents of the sibship, after which we created offspring genotypes using the Haldane mapping function. We then removed the parental genotypes from the data set.

We examined the performance of our VC framework for pedigree analysis of both continuous and categorical outcomes. For continuous outcomes, we simulated gamma-distributed data using the model in (2) where we assumed $g(\mu) = \log(\mu)$ and the covariance matrix Σ in (4). For fixed effects, we assumed an intercept with $\beta = \log(2)$, a categorical covariate with two equally-frequent outcomes that has an

effect of $\beta_{\text{cat}} = 0.5$, and a continuous covariate derived from a standard normal distribution that has an effect of $\beta_{\text{con}} = 0.1$. We assumed the scale parameter of the gamma distribution to be $\nu = 2.0$ and varied $(\sigma^2_{mg}, \sigma^2_{pg})$ among different combination of values to investigate the performance of our gamma VC model to detect heritability and linkage.

We analyzed the gamma-distributed data using two different procedures. First, we analyzed the untransformed data using the gamma variance-component method. Then, we applied a logarithmic transformation to the data to obtain approximate normality and applied the traditional variance-component method for analysis. For linkage analyses, we estimated IBD sharing at the major-gene locus using the Lander–Green [27] algorithm as implemented in Genehunter [25]. We performed pedigree analysis for each procedure by maximizing the appropriate likelihood in the SAS procedure NLMIXED.

For categorical pedigree analysis, we simulated a three-level ordinal outcome (with possible values 0, 1, and 2) using the proportional-odds model in (5) and the likelihood in (6). Within (5), we set $\theta_0 = -1$ and $\theta_1 = 1$, such that we would expect to observe levels of 0, 1, and 2 in approximately 25%, 50%, and 25% of our sample, respectively. We assumed a categorical covariate with two equally-frequent outcomes that has an effect of $\beta_{\text{cat}} = 0.2$, and a continuous covariate derived from a standard normal distribution that has an effect of $\beta_{\text{con}} = 0.05$. We then varied $(\sigma^2_{mg}, \sigma^2_{pg})$ among different combination of values. For linkage analysis, we again used the Lander–Green algorithm to estimate IBD sharing. We implemented the procedure in SAS procedure NLMIXED.

3.2 Simulation Results: Continuous Outcomes

Tables 1 and 2 summarize results of heritability analyses of gamma-distributed data for 200 sibpairs using both our proposed gamma VC model and the traditional VC model (after a log transformation). Within Table 1, we show the power of the two models to detect a genetic effect under different levels of σ^2_{pg}. We chose values of σ^2_{pg} corresponding to approximate heritability values of 0, 0.10, 0.20, and 0.30 for the log-transformed data. Based on these results, we see that both approaches have appropriate type-I error when the null hypothesis of $\sigma^2_{pg} = 0$ holds. However, under alternative models, we observe that the gamma VC method is substantially more

Table 1 Power of heritability analysis for gamma-distributed data

	Power to Reject $H_0 : \sigma^2_{pg} = 0$ at $\alpha = 0.05$			
	Heritability(σ^2_{pg})			
VC Model	0.0 (0.0)	0.1 (0.08)	0.2 (0.17)	0.3 (0.29)
Gamma	0.043	0.501	0.890	0.990
Traditional	0.049	0.207	0.371	0.688

Results are based on 1000 replicates of data sets consisting of 200 sibpairs. Gamma data were generated using scale parameter of 2.0. Heritability refers to the approximate proportion of variance explained by genetic factors for log-transformed data. Traditional VC model was applied to trait data after a log transformation

Table 2 Heritability analysis of gamma-distributed data: parameter estimates

Parameter	True Value	Gamma VC Model			Traditional VC Model		
		Mean	Empirical SE	Mean Model SE	Mean	Empirical SE	Mean Model SE
Heritability = 0.10							
σ^2_{pg}	0.08	0.080	0.052	0.054	0.093	0.087	0.103
β_{con}	0.10	0.099	0.040	0.039	0.100	0.044	0.042
β_{cat}	0.50	0.502	0.081	0.078	0.502	0.088	0.085
Heritability = 0.20							
σ^2_{pg}	0.17	0.166	0.067	0.068	0.161	0.106	0.116
β_{con}	0.10	0.102	0.043	0.042	0.102	0.047	0.045
β_{cat}	0.50	0.500	0.085	0.084	0.500	0.090	0.090
Heritability = 0.30							
σ^2_{pg}	0.29	0.285	0.089	0.087	0.281	0.133	0.133
β_{con}	0.10	0.098	0.046	0.046	0.097	0.048	0.048
β_{cat}	0.50	0.502	0.097	0.092	0.502	0.100	0.095

Results are based on 1000 replicates of data sets consisting of 200 sibpairs. Gamma data were generated using scale parameter of 2.0. Heritability refers to the approximate proportion of variance explained by genetic factors for log-transformed data. Traditional VC model was applied to trait data after a log transformation

powerful for testing for heritability compared to traditional VC analysis of the log-transformed data. This result is further supported by Table 2, which shows parameter estimates and estimated standard errors for model parameters. Both VC approaches yielded unbiased parameter estimates and estimated standard errors that closely mirrored the empirical standard errors. However, we observe that the estimated standard errors for $\hat{\sigma}^2_{pg}$ under the gamma VC model are substantially smaller than those from the traditional VC model, which further demonstrates the improved efficiency of the gamma VC model for testing heritability.

While the gamma VC yields improved performance for testing for an overall genetic effect compared to traditional VC analysis of log-transformed data, we interestingly observe that both procedures yield similar inference for testing linkage. We simulated data sets comprised of 400 sibpairs and generated data under the gamma VC model using simulated values of $(\sigma^2_{mg}, \sigma^2_{pg})$ that, for log-transformed data, corresponded to major-gene heritability values (defined as the proportion of trait variance explained by the major gene) of 0 and 0.10 and overall heritability values (defined as the proportion of trait variance explained by major gene and polygenes) of 0.30 and 0.50. As observed in Table 3, we see that both VC methods have appropriate type-I error when the major-gene heritability is 0 and have similar power for testing a major-gene locus whose major-gene heritability is 0.10. We observed these results for overall heritability values of both 0.30 and 0.50.

Table 4 shows parameter estimates derived from the gamma and traditional VC models for simulation models where the major-gene heritability values are 0.10. For each VC model, we observe that mean parameter estimates are unbiased and the

Table 3 Power of linkage analysis for gamma-distributed data

	Power to Reject $H_0 : \sigma^2_{mg} = 0$ at $\alpha = 0.05$			
	Overall Heritability = 0.30		Overall Heritability = 0.50	
	Major-Gene Heritability = 0.00	Major-Gene Heritability = 0.10	Major-Gene Heritability = 0.00	Major-Gene Heritability = 0.10
VC Model	$(\sigma^2_{mg}, \sigma^2_{pg}) = (0.00, 0.29)$	(0.10, 0.20)	(0.00, 0.66)	(0.14, 0.52)
Gamma	0.054	0.196	0.053	0.196
Traditional	0.041	0.148	0.055	0.196

Results are based on 1000 replicates of data sets consisting of 400 sibpairs. Gamma data were generated using scale parameter of 2.0. Overall (major-gene) heritability refers to the approximate proportion of variance explained by all genetic factors (major gene) for log-transformed data. Traditional VC model was applied to trait data after a log transformation

Table 4 Linkage analysis of gamma-distributed data: parameter estimates

		Gamma VC Model			Traditional VC Model		
Parameter	True Value	Mean	Empirical SE	Mean Model SE	Mean	Empirical SE	Mean Model SE
Heritability = 0.30							
σ^2_{mg}	0.10	0.119	0.102	0.112	0.137	0.115	0.136
σ^2_{pg}	0.20	0.187	0.117	0.114	0.190	0.137	0.142
β_{con}	0.10	0.103	0.032	0.032	0.103	0.033	0.034
β_{cat}	0.50	0.498	0.067	0.065	0.497	0.069	0.067
Heritability = 0.50							
σ^2_{mg}	0.14	0.155	0.134	0.160	0.155	0.134	0.168
σ^2_{pg}	0.52	0.504	0.167	0.168	0.507	0.187	0.200
β_{con}	0.10	0.101	0.038	0.038	0.101	0.039	0.039
β_{cat}	0.50	0.499	0.076	0.076	0.498	0.077	0.090

Results are based on 1000 replicates of data sets consisting of 400 sibpairs. Gamma data were generated using scale parameter of 2.0. Heritability refers to the approximate proportion of variance explained by genetic factors for log-transformed data. Major-gene heritability for both models was approximately 0.10. Traditional VC model was applied to trait data after a log transformation

mean standard errors generally match the empirical standard errors (although the mean model-based standard errors for the variance parameters slightly overestimate the empirical values). As in Table 2, we again observe that the estimated standard errors for σ^2_{pg} are smaller for the gamma VC model than for the traditional VC model. For σ^2_{mg}, we observe that the gamma VC model yields smaller standard errors for estimates of this parameter compared to the traditional VC model when the overall heritability of 0.30. For overall heritability of 0.50, the two VC models yield similar standard-error estimates for σ^2_{mg}.

Table 5 Pedigree analysis of ordinal data: parameter estimates

Parameter	True Value	Ordinal VC Model		
		Mean	Empirical SE	Mean Model SE
Model 1				
σ^2_{pg}	2.00	1.970	0.650	0.708
β_{con}	0.05	0.049	0.082	0.080
β_{cat}	0.20	0.192	0.161	0.162
Model 2				
σ^2_{mg}	0.50	0.521	0.482	0.539
σ^2_{pg}	1.50	1.450	0.747	0.827
β_{con}	0.05	0.049	0.085	0.081
β_{cat}	0.20	0.202	0.164	0.161

Results are based on 1000 replicates of data sets consisting of 300 sibtrios

3.3 Simulation Results: Ordinal Categorical Outcomes

Table 5 shows parameter estimates from application of the ordinal VC model to 3-level ordinal outcomes from simulated data sets comprised of 300 sibtrios that were generated under one of two models. Consistent with our results from the gamma VC simulations, we found that mean parameter estimates are unbiased. For fixed effects, we also found that model-based standard errors corresponded well with empirical values. However, for variance parameters, we found that model-based standard errors for the variance parameters differed somewhat from empirical values. This result suggests that one should use score or likelihood-ratio tests, rather than Wald tests, for testing heritability and linkage under the ordinal VC model.

Using likelihood-ratio statistics, we found that our ordinal VC model had excellent power for testing heritability $H_0 : \sigma^2_{pg} = 0$ vs. $H_A : \sigma^2_{pg} > 0$ in data sets consisting of 3-level ordinal data on 300 sibtrios. In choosing simulation values for σ^2_{pg}, we note that $\exp(\sqrt{\sigma^2_{pg}})$ corresponds roughly to the average spread of risk of being in level 1 (compared to level 0) or being in level 2 (compared to levels 0–1) due to within-family polygenic effects [37]. We assumed values of σ^2_{pg} of 0.5, 1.0, and 2.0, which corresponded to ∼2, ∼2.7, and ∼4-fold increase or decrease in per-family risk compared to the overall risk. For σ^2_{pg} of 0.5, 1.0, and 2.0, we found that we had power of 0.57, 0.90, and 0.99, respectively, at $\alpha = 0.05$ to detect the overall genetic effect. We also found our approach had appropriate type-I error under the null hypothesis (empirical size of 0.047 at $\alpha = 0.05$). For testing linkage $H_0 : \sigma^2_{mg} = 0$ vs. $H_A : \sigma^2_{mg} > 0$, we found that our ordinal VC model had appropriate size at $\alpha = 0.05$ and $\alpha = 0.01$ for data sets consisting of 300 sibtrios assuming values of σ^2_{pg} ranging between 0.5 and 2.0 (results not shown). Power to detect linkage under these simulated models increased with increasing values of σ^2_{mg}. For example, assuming a total genetic variance $\sigma^2_{mg} + \sigma^2_{pg} = 2.0$, we had power of 0.25 to detect linkage when $\sigma^2_{mg} = 0.5$ and power of 0.53 to detect linkage when $\sigma^2_{mg} = 1.0$ at $\alpha = 0.05$.

4 Application to Genetic Study of Ovarian Insufficiency

Ovarian insufficiency is a complex disorder that encompasses a variety of conditions related to reproductive dysfunction within women. Such reproductive issues include premature ovarian failure (cessation of menses prior to age 40), altered menstrual-cycle characteristics, and infertility. Many researchers have shown that a premutation within the fragile X mental retardation 1 gene (*FMR1*) leads to an increased risk for ovarian insufficiency [2, 36, 46]. The *FMR1* premutation is clinically defined as having 55–199 unmethylated CGG repeats in the 5′ untranslated region of the gene [43]. When the premutation further expands to over 200 repeats during transmission from mother to child, it becomes a full mutation that hypermethylates the *FMR1* gene and leads to the separate disorder of fragile X syndrome. Interestingly, studies have shown that the *FMR1* full mutation does not associate with ovarian insufficiency; hence, only premutation carriers have increased risk for this condition [42]. This disorder is now commonly referred to as fragile X-associated primary ovarian insufficiency (FXPOI).

The *FMR1* premutation plays a substantial genetic role in the development of ovarian insufficiency. For example, the prevalence of premature ovarian failure in *FMR1* premutation carriers is 20% whereas it is only 1% in the general population [42]. However, once adjusted for *FMR1* premutation status, it is unclear how much of the remaining residual variation in ovarian-insufficiency outcomes are explained by additional genetic factors. If significant residual heritability of these outcomes exists, this would motivate the formation of additional studies for further gene mapping of these outcomes.

To investigate whether these reproductive outcomes have significant residual heritability after adjusting for *FMR1* premutation status, we applied our VC framework to a reproductive-history data set consisting of 680 women from 225 families who have a history of fragile X syndrome (and, hence, are enriched for being premutation carriers) and 321 women from 219 families from the general population. We focused on the outcome of menstrual-cycle length in the last year of natural cycling. The distribution of this outcome in the data set is shown in Fig. 3. The estimated skewness of the outcome was 8.48 and the estimated kurtosis was 93.31. We obtained a p-value of $p < 0.0001$ for a test of normality using a Shapiro–Wilk test.

We were unsuccessful in our attempts to transform menstrual-cycle length to approximate normality using log, square-root, or reciprocal transformations. Therefore, as done previously in [2], we formed a 3-level ordinal variable for menstrual-cycle length based on the 25th and 75th percentiles of the outcome distribution. We defined a short cycle as a menstrual length less than or equal to 25 days, an average cycle to be a length between 25 and 30 days, and a long cycle to be a length greater or equal to 30 days. Using this ordinal variable, we examined whether menstrual-cycle length had significant residual heritability after adjusting for *FMR1* premutation status by applying the ordinal VC model in (5) to this outcome. Using the likelihood in (1) and the model in (5), we tested the hypothesis $H_0 : \sigma^2_{pg} = 0$ vs. $H_A : \sigma^2_{pg} > 0$ after adjusting for the fixed effects of *FMR1* repeat size, age when last cycled, smoking, and ethnicity (Caucasian or non-Caucasian). As done previously [2], we modeled *FMR1* repeat size as a categorical variable with four levels: <59 repeats (normal), 59–79 repeats (low-premutation group), 80–100 repeats (medium-premutation group), and >100 repeats (high-premutation group).

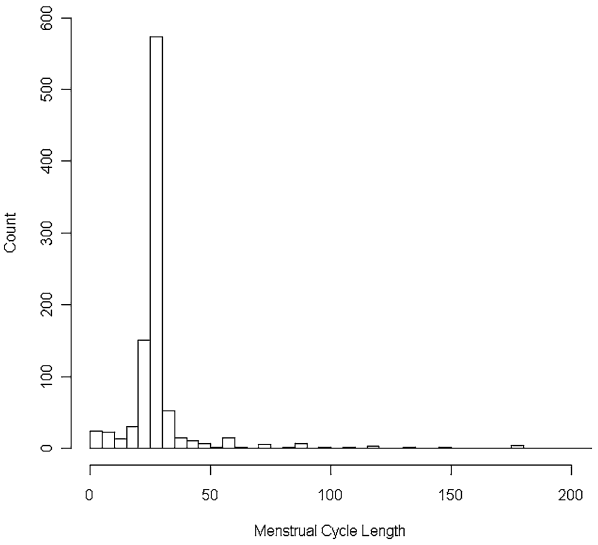

Fig. 3 Distribution of menstrual cycle length

Table 6 Heritability analysis of ordinal cycle data

Parameter	Estimate	SE	P-value
FMR1 Premutation Status[a]			
Low	−0.590	0.235	0.012
Medium	−0.537	0.191	0.005
High	−0.196	0.291	0.501
Smoking	−0.192	0.160	0.231
Age of interview	0.001	0.007	0.865
Race[b]	0.590	0.222	0.008
σ^2_{pg}	0.949	0.690	0.027[c]

[a]Reference is normal *FMR1* group (<59 repeats)

[b]Race category: 1 = Caucasian, 0 = Non-Caucasian

[c]Likelihood-Ratio Statistic

Table 6 shows results of the application of the proportional-odds ordinal VC model to the reproductive-history data set. Supporting the work in the previous studies of [46] and [2], we found that *FMR1* premutation status was significantly associated with menstrual-cycle length. In particular, compared to *FMR1* normal carriers, we observe that low- and medium-premutation carriers showed decreased risk for medium cycles (compared to short cycles) and long cycles (compared to medium and short cycles) ($p = 0.012$ for low-premutation carriers, $p = 0.005$ for medium-premutation carriers). Interestingly, high-premutation carriers showed no change in risk compared to normal carriers ($p = 0.501$). Using a likelihood-ratio statistic, we tested for the existence of a residual polygenic effect by considering $H_0 : \sigma^2_{pg} = 0$ vs. $H_A : \sigma^2_{pg} > 0$ and found significant evidence of additional genetic factors that influence the outcome ($p = 0.027$). Our ordinal VC model yields a variance estimate of $\hat{\sigma}^2_{pg} = 0.949$, which corresponds to a ∼2.6-fold change in per-family risk compared to the average risk. This finding supports previous work by Hunter et al. [23], who found evidence for additional genetic factors influencing age of menopause (another indicator

of ovarian function) after adjusting for *FMR1* premutation status in the same set of samples.

5 Discussion

In this paper, we show a general VC framework for pedigree analysis of continuous and categorical outcomes that do not follow the multivariate normal distribution assumed by traditional variance-component procedures. Using simulated data, we demonstrate the procedure returns unbiased estimates of both fixed and random effects and has appropriate size for testing heritability and linkage of a trait of interest. For continuous non-normal outcomes, such as those that follow a gamma distribution, we also show that our proposed approach can lead to improved power for testing certain genetic hypotheses relative to traditional VC methods. We also demonstrate that our approach is applicable to real genetic data sets, as shown by our heritability analysis of ordinal menstrual-cycle data from a genetic study of ovarian insufficiency.

We developed our general VC framework for the purpose of heritability or linkage analysis of a complex trait of interest. We could also extend our framework to test whether a specific genetic marker is in linkage disequilibrium (LD) with the trait of interest; that is, associated with the trait in the presence of linkage. We would base this LD extension on the work of Abecasis et al. [1], who extended the traditional VC framework to test for marker association at a linked gene locus for nuclear families. The authors assumed the trait data conditional on offspring and parental genotypes at the marker of interest follow a multivariate normal distribution with a given mean structure and covariance matrix. They modeled the parentally-transmitted alleles of the offspring as a covariate in the mean structure, while modeling the effects of the linked gene locus and polygenes in the covariance structure (as shown in equations (3) and (4)). The authors partitioned the offspring genotype effects into between-family and within-family components and showed that a test of the within-family genotype effect was a valid test of LD in the presence of potential confounders such as population stratification. Using these ideas, we could modify the general VC framework in similar fashion to develop a procedure for LD mapping of general continuous and categorical outcomes. We will explore this idea in a subsequent paper.

For non-normally distributed outcomes, our general VC framework yields a likelihood that is intractable due to the fact that the integration of the likelihood over the random genetic effects does not yield a simple closed-form solution. To resolve such intractability in analysis, we used approximate maximum-likelihood procedures as implemented in SAS PROC NLMIXED. However, we can use other statistical algorithms to circumvent this issue. We could maximize the likelihood using the Gibbs sampler [8, 50] as implemented in the computer program WinBUGS [45]. However, the procedure is computationally slower than maximum-likelihood procedures and cannot, under ascertainment sampling, maximize the correct ascertainment-adjusted likelihood (6) [16]. One could employ the method of penalized quasi-likelihood [7]. For this method, one approximates the integral using Laplace's method [47] to obtain an approximate quasi-likelihood that can be modified into a penalized quasi-likelihood for maximization [27]. While this method works well in many situations,

it tends to underestimate variance parameters when the trait of interest is categorical [7]. Also, one likely will have difficulty constructing the appropriate ascertainment-adjusted penalized quasi-likelihood for non-random samples.

For non-normal data, our proposed general variance-component framework maximizes the appropriate likelihood using adaptive Gaussian quadrature. As this maximization procedure requires numerical integration, it is computationally more intensive than the typical maximization procedures (Fisher Scoring or Newton–Raphson) used for the traditional variance-component method. The degree of the computational complexity for our method depends on both the number of quadrature points used for likelihood approximation (which we selected using the adaptive procedures available in SAS NLMIXED) and the dimension of the integral in the likelihood. As both of these quantities increase, so does the amount of computer time required for successful likelihood maximization. The number of quadrature points needed for accurate likelihood approximation depends on the type of data analyzed as fewer points are needed for continuous data relative to discrete data. The dimension of the likelihood integral will increase with the size of the family. Therefore, our proposed framework may have prohibitive computer-run times for linkage analysis of discrete data sets from large families given current computational resources.

To reduce the amount of computer time for linkage analysis using our general variance-component method, one could perform the analyses using a Laplace approximation, which corresponds to 1 quadrature point. For a Laplace approximation, maximization of the likelihood for our general variance-component method typically only requires a few seconds. For continuous non-normal data such as gamma data or censored normal data, the resulting variance-component analyses typically have the same linkage power and type-I error as compared with the same variance-component analyses using adaptive procedures for quadrature-point selection, although we observe bias in the variance estimates [17]. Therefore, we can use a Laplace approximation to perform an efficient genome scan for identifying regions linked to these continuous traits. We would then recommend repeating the analyses on these linkage regions using more quadrature points to obtain unbiased estimates of these variance parameters. While the Laplace approximation is suitable for linkage tests for continuous non-normal data, the same cannot be said for discrete binary traits. Simulation results (not shown) reveal the Laplace approximation yields tests for binary traits that have little or no power to detect linkage. Therefore, a Laplace approximation appears to be useful for variance-component linkage analysis only when one is studying continuous data.

Acknowledgements This research was supported by the University Research Committee of Emory University (to M.P.E.) and National Institutes of Health Grants R01 HG03618 (to M.P.E.), R01 HD29909 (to J.E.H., E.G.A. and S.L.S.), R29 CA76404 (to X.L.), and R01 HG00376 (to M.B.).

References

1. Abecasis GR, Cardon LR, Cookson WOC (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66:279–292
2. Allen EG, Sullivan AK, Marcus M, Small C, Dominguez C, Epstein MP, Charen K, He W, Taylor KC, Sherman SL (2007) Examination of reproductive aging milestones among women who carry the FMR1 premutation. Hum Reprod 22:2142–2152

3. Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J (1999) Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am J Hum Genet 65:531–544
4. Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211
5. Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535–543
6. Blangero J, Williams JT, Almasy L (2001) Variance component methods for detecting complex trait loci. In: Rao DC, Province MA (eds) Genetic dissection of complex traits. Academic Press, San Diego, pp 151–182
7. Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:9–25
8. Burton PR, Tiller KJ, Gurrin LC, Cookson WO, Musk AW, Palmer LJ (1999) Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. Genet Epidemiol 17:118–140
9. Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olson JM, Elston RC (2000) Ascertainment adjustment: where does it take us? Am J Hum Genet 67:1505–1514
10. de Andrade M, Fridley B, Boerwinkle E, Turner S (2003) Diagnostic tools in linkage analysis of quantitative traits. Genet Epidemiol 24:302–308
11. de Andrade M, Gueguen R, Visvikis S, Sass C, Siest G, Amos CI (2002) Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. Genet Epidemiol 22:221–232
12. Diao G, Lin DY (2005) A powerful and robust method for mapping quantitative trait loci in general pedigrees. Am J Hum Genet 77:97–111
13. Duggirala R, Blangero J, Almasy L, Dyer TD, Williams KL, Leach RJ, O'Connell P, Stern MP (1999) Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. Am J Hum Genet 64:1127–1140
14. Duggirala R, Williams JT, Williams-Blangero S, Blangero J (1997) A variance component approach to dichotomous trait linkage analysis using a threshold model. Genet Epidemiol 14:987–992
15. Epstein MP (2002) Comment on "Ascertainment adjustment in complex diseases". Genet Epidemiol 23:209–213
16. Epstein MP, Lin X, Boehnke M (2002) Ascertainment-adjusted parameter estimates revisited. Am J Hum Genet 70:886–895
17. Epstein MP, Lin X, Boehnke M (2003) A tobit variance-component method for linkage analysis of censored trait data. Am J Hum Genet 72:611–620
18. Fulker DW, Cherny SS, Cardon LR (1995) Multipoint interval mapping of quantitative trait loci using sib pairs. Am J Hum Genet 56:1224–1233
19. Glidden DV, Liang K-Y (2002) Ascertainment adjustment in complex diseases. Genet Epidemiol 23:201–208
20. Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19
21. Hasstedt SJ (1993) Variance components/major locus likelihood approximation for quantitative, polychotomous, and multivariate data. Genet Epidemiol 10:145–158
22. Hopper JL, Mathews JD (1982) Extensions to multivariate normal models for pedigree analysis. Ann Hum Genet 46:373–383
23. Hunter JE, Epstein MP, Tinker SW, Charen KW, Sherman SL (2008) Fragile X-associated primary ovarian insufficiency: evidence for additional genetic contributions to severity. Genet Epidemiol 32:553–559
24. Jacquard A (1974) The genetic structure of populations. Springer, New York
25. Kruglyak L, Daly M, Reeve-Daly M, Lander E (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363
26. Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454
27. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367
28. Lange K (1978) Central limit theorems for pedigrees. J Math Biol 6:59–66
29. Lange K, Boehnke M (1983) Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. Am J Med Genet 14:513–524

30. Lange K, Westlake J, Spence MA (1976) Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet 39:485–491
31. Lin X (1997) Variance component testing in generalised linear models with random effects. Biometrika 84:309–326
32. McCullagh P (1980) Regression models for ordinal data. J R Stat Soc, Ser B 42:109–142
33. McCullagh P, Nelder JA (1983) Generalized linear models. Chapman and Hall, London
34. McCulloch CE, Searle SR (2000) Generalized, linear, and mixed models. Wiley-Interscience, New York
35. Mitchell BD, Ghosh S, Schneider JL, Birznieks G, Blangero J (1997) Power of variance component linkage analysis to detect epistasis. Genet Epidemiol 14:1017–1022
36. Murray A, Webb J, MacSwiney F, Shipley EL, Morton NE, Conway GS (1999) Serum concentrations of follicle stimulating hormone may predict premature ovarian failure in FRAXA premutation women. Hum Reprod 14:1217–1218
37. Pankratz VS, de Andrade M, Therneau TM (2005) Random effects Cox proportional hazards model: general variance components methods for time-to-event data. Genet Epidemiol 28:97–109
38. Pinheiro JC, Bates DM (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. J Comput Graph Stat 4:12–35
39. SAS Institute (1999) SAS version 8, Cary, NC
40. Self SG, Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. J Am Stat Assoc 82:605–610
41. Sham PC, Purcell S (2001) Equivalence between Haseman–Elston and variance-components linkage analyses for sib pairs. Am J Hum Genet 68:1527–1532
42. Sherman SL (2000) Premature ovarian failure in the fragile X syndrome. Am J Med Genet 97:189–194
43. Sherman S, Pletcher BA, Driscoll DA (2005) Fragile X syndrome: diagnostic and carrier testing. Genet Med 7:584–587
44. Solomon PJ, Cox DR (1992) Nonlinear component of variance models. Biometrika 79:1–11
45. Spiegelhalter D, Thomas A, Best N (2000) WinBUGS version 1.3 user manual. MRC Biostatistics Unit, Cambridge, UK
46. Sullivan AK, Marcus M, Epstein MP, Allen EG, Anido AE, Paquin JJ, Yadav-Shah M, Sherman SL (2005) Association of FMR1 repeat size with ovarian dysfunction. Hum Reprod 20:402–412
47. Tierney L, Kadane JB (1986) Accurate approximations for posterior moments and marginal densities. J Am Stat Assoc 81:82–86
48. Williams JT, Blangero J (1999) Comparison of variance components and sibpair-based approaches to quantitative trait linkage analysis in unselected samples. Genet Epidemiol 16:113–134
49. Yu X, Knott SA, Visscher PM (2004) Theoretical and empirical power of regression and maximum-likelihood methods to map quantitative trait loci in general pedigrees. Am J Hum Genet 75:17–26
50. Zeger SL, Karim MR (1991) Generalized linear models with random effects: a Gibbs sampling approach. J Am Stat Assoc 86:79–86