frontiers in Genetics

ORIGINAL RESEARCH
published: 31 March 2015
doi: 10.3389/fgene.2015.00120

A re-formulation of generalized linear mixed models to fit family data in genetic association studies

Tao Wang[1], Peng He[1,2], Kwang Woo Ahn[1], Xujing Wang[3], Soumitra Ghosh[4] and Purushottam Laud[1]*

[1] Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI, USA, [2] Global Biostatistical Science, Amgen Inc., Thousand Oaks, CA, USA, [3] Bioinformatics and Systems Biology Core, National Heart, Lung, and Blood Institute, Bethesda, MD, USA, [4] Department of Genetics, Quantitative Sciences, GlaxoSmithKline, King of Prussia, PA, USA

The generalized linear mixed model (GLMM) is a useful tool for modeling genetic correlation among family data in genetic association studies. However, when dealing with families of varied sizes and diverse genetic relatedness, the GLMM has a special correlation structure which often makes it difficult to be specified using standard statistical software. In this study, we propose a Cholesky decomposition based re-formulation of the GLMM so that the re-formulated GLMM can be specified conveniently via "proc nlmixed" and "proc glimmix" in SAS, or OpenBUGS via R package BRugs. Performances of these procedures in fitting the re-formulated GLMM are examined through simulation studies. We also apply this re-formulated GLMM to analyze a real data set from Type 1 Diabetes Genetics Consortium (T1DGC).

Keywords: family data, generalized linear mixed models (GLMM), genetic correlation, genetic variance components, random genetic effects, re-parameterization, Cholesky decomposition, Bayesian methods

OPEN ACCESS

Edited by:
Eduardo Manfredi, Institut National de la Recherche Agronomique, France

Reviewed by:
Solomon K. Musani, Egerton University, Kenya
Ana I. Vazquez, University of Alabama at Birmingham, USA
Dalin Li, Cedars-Sinai Medical Center, USA

****Correspondence:***
Tao Wang, Division of Biostatistics, Institute for Health and Society, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226-0509, USA
taowang@mcw.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *18 November 2014*
Accepted: *12 March 2015*
Published: *31 March 2015*

Citation:
Wang T, He P, Ahn KW, Wang X, Ghosh S and Laud P (2015) A re-formulation of generalized linear mixed models to fit family data in genetic association studies. Front. Genet. 6:120. doi: 10.3389/fgene.2015.00120

1. Introduction

Generalized linear mixed model (GLMM) provides a rich class of statistical models to model correlated data with responses from the exponential family of distributions including Gaussian, Binomial, Poisson, etc. (see McCulloch and Searle, 2001). The mixed model approach, which is also called the variance component approach, has long been used in genetic studies to estimate genetic parameters, predict breeding values and model correlated family or pedigree data (Henderson, 1963, 1975; Amos, 1994; Falconer and Mackay, 1996; Almasy and Blangero, 1998; Abecasis et al., 2000; Sham and Purcell, 2001). Due to the diverse genetic correlation structures among families or pedigrees, currently fitting this type of GLMM often relies on special genetic software packages, such as SOLAR (Almasy and Blangero, 1998), Multic (de Andrade et al., 1998, 2006). More recently, as an extension of the R package *lme4*, Vazquez et al. (2010) also developed an R package *pedigreemm* following the method of Harville and Callanan (1989). *Pedigreemm* can handle the additive genetic correlation among all sampled individuals via a Cholesky decomposition of the coancestry (or so-called numerator relationship) matrix. However, despite the popularity of these software packages, they often lack many options such as choosing different algorithms for maximizing the likelihood or specifying various particular type of the covariance structures that are available in standard statistical software packages such as procedures "proc mixed," "proc glimmix," and "proc nlmixed" in SAS (SAS Institute, Inc.) or OpenBUGS in R (e.g., via R package BRugs). One major obstacle in using these standard software packages to fit the GLMM is the requirement of the same correlation structures across all the families (or clusters). A few recent studies also suggested fitting

the GLMM by using SAS (Feng et al., 2009; Wang et al., 2011). But it appears that these studies have only considered the cases where all the families had the same genetic correlation structure.

The genetic correlation structure among family members for additive effects, dominance and epistasis has been well described (see Lynch and Walsh, 1998). In this study, for the special type of GLMM that models family data with varied family sizes or diverse genetic relatedness, we propose a Cholesky decomposition based re-formulation of the GLMM so that the re-formulated GLMM can be specified conveniently when using some standard statistical software packages. First, a standard GLMM is presented, which can account for the genetic correlation among family members. We briefly discuss the identifiability issue of the variance components from their possible confounding perspective in the GLMM. Next, we explain how the GLMM can be re-formulated into a GLMM with random regression coefficients of equal variances. Assuming that there is no genetic correlation between different families, we start by applying separate Cholesky decompositions on the genetic kinship matrices within each family. Then we stack these Chelosky decomposition matrices from different families into one column and treat the columns of the stacked matrix as fixed covariates. Unlike the regular Cholesky decomposition performed on the whole covariance matrix of the random genetic effects (e.g., Vazquez et al., 2010), here the stacked matrix has its column size being the maximum family size instead of the summation of all family sizes. With the reduced number of columns, this re-formulated GLMM can then be specified conveniently in "proc nlmixed" or "proc glimmix" using SAS, and OpenBUGS via R package BRugs. We provide detailed codes on fitting this re-formulated GLMM by using either "proc nlmixed" or "proc glimmix" in SAS, and OpenBUGS with R (via R package BRugs). The performances of these procedures on fitting the re-formulated GLMM are also examined through some simulation studies. Finally, we apply this re-formulated GLMM to a real data set from Type 1 Diabetes Genetics Consortium (T1DGC).

2. Methods

2.1. A Generalized Linear Mixed Model (GLMM)

Suppose that we have a randomly collected sample of N families from a study population. In the i-th family of size n_i, let y_{ij} be a (binary or continuous) response variable for a disease phenotype; z_{ij} be some fixed environmental covariates that need to be adjusted for; g_{ij} denote the observed genotypes at certain targeted genetic marker loci, for family members $j = 1, 2, \cdots, n_i$ and $i = 1, 2, \cdots, N$. To model the family data and test for the association of g_{ij} with the phenotypic response y_{ij}, we need to account for the genetic correlation among family members induced by identity-by-descent (IBD) alleles shared by the family members at some putative disease susceptible loci (DSL). In addition, family members may share certain common environmental factors which could also contribute to the disease phenotypes. Let v_{ij} denote the random genetic effect from those putative (unobserved) DSL on the phenotypic response y_{ij}, and e_i be the shared environmental effect such as diets for members in a family i. Define $\mu_{ij} = \mathrm{E}(y_{ij}|v_{ij}, e_i)$. Then a generalized linear mixed model (GLMM) to model the family data can be written as

$$\begin{cases} y_{ij}|v_{ij}, e_i \sim f_{y_{ij}|v_{ij},e_i}(y|v_{ij}, e_i),\ j = 1, \cdots, n_i \\ \mathrm{E}(y_{ij}|v_{ij}, e_i) = \mu_{ij} \\ g(\mu_{ij}) = m + z_{ij}\alpha + x(g_{ij})\beta + v_{ij} + e_i,\ \ j = 1, \cdots, n_i \\ \mathbf{v_i} = (v_{i1}, \cdots, v_{in_i})^T \sim N(\mathbf{0}, \Sigma_\mathbf{i}) \\ e_i \sim N(0, \sigma_c^2),\ e_i \perp \mathbf{v_i} \end{cases} \quad (1)$$

where $g(\cdot)$ is a known link function, m is an intercept for the baseline, α is a p-dimensional vector of parameters for the fixed effects of environmental covariates z_{ij}, $x(g_{ij})$ is a q-dimensional vector with its components being defined by certain coding functions for marker genotypes (see Wang, 2011), and β is a q-dimensional vector of parameters for the fixed genetic effects contributed by the observed genotypes g_{ij}. Typically, the random effects $\{\mathbf{v_i}, i = 1, \cdots, n\}$ from different families are assumed to be independent. Given the random effects $\mathbf{v_i}$ and e_i, $y_{ij}, j = 1, \cdots, n_i$, are also assumed to be conditionally independent. As each v_{ij} represents an aggregated polygenic effect from multiple putative DSLs, v_{ij} tends to be normally distributed based on the Central Limit Theorem. Therefore, $\Sigma_\mathbf{i}$ denotes the genetic covariance matrix among the g-transformed conditional means $g(\mu_{i1}), \cdots, g(\mu_{in_i})$, which is often induced by identity-by-descent (IBD) alleles shared by the i-th family members at the putative DSL.

For a quantitative phenotypic trait, the link function g is often chosen as an identity function (i.e., $g(x) = x$). Assuming that there is no inbreeding between parents, it has been known that the genetic covariance between a pair of relatives j, k within a family i can be expressed as Kempthorne (1955), Amos (1994), Lynch and Walsh (1998), and Yu et al. (2006)

$$\mathrm{Cov}(v_{ij}, v_{ik}) = 2\phi_{jk}\sigma_A^2 + \delta_{jk}\sigma_D^2 \quad (2)$$

where σ_A^2 and σ_D^2 are the so-called additive and dominant genetic variance components, which are contributed by the additive allelic effects and allelic interactions from those unknown DSL, respectively; ϕ_{jk} and δ_{jk} are the so-called kinship and double coancestry coefficients between the two relatives. Similarly, for a general link function g, we can also define σ_A^2 and σ_D^2 as the additive and dominant variance components contributed by the allelic effects and allelic interactions from those unknown DSL to the variation of the g-transformed conditional means $g(\mu_{ij})$. Let $\mu_\mathbf{i} = (\mu_{i1}, \cdots, \mu_{in_i})$. From model Equations (1), (2), we have

$$\mathrm{Cov}(g(\mu_\mathbf{i})) = 2\Phi_i\sigma_A^2 + \Delta_i\sigma_D^2 + \sigma_c^2\mathrm{J}_{n_i}$$

where J_{n_i} is a $n_i \times n_i$ matrix of 1's, $\Phi_i = (\phi_{jk})$ and $\Delta_i = (\delta_{jk})$ are $n_i \times n_i$ kinship and double coancestry matrices, respectively, for the i-th family members. In the absence of inbreeding, the expected kinship and double coancestry coefficients for various common relatedness have been well established (e.g., see Table 7.1 in Lynch and Walsh, 1998). In practice, the actual kinship and double coancestry coefficients could deviate from their expected values because the realized IBD status could vary for a particular pair of family members due to the randomness in their parents' inheritance segregation. Two parents could also be related to each

other due to possible inbreeding from their common ancestry. Therefore, the genetic covariance matrix $\Sigma_{\mathbf{i}} = 2\Phi_i\sigma_A^2 + \Delta_i\sigma_D^2$ may have varied sizes across different families. For families with the same size and relatedness, their actual kinship and double coancestry matrices could also be different. Nowadays, with the high density of genome-wide genetic markers available, such as single nucleotide polymorphism (SNPs), it is possible to estimate the actual genome-wide kinship and double coancestry coefficients among family members using external programs such as PLINK (Purcell et al., 2007). The genome-wide kinship matrix can also be estimated as half of the genomic relationship matrix (see VanRaden, 2008).

The above GLMM provides a rich class of statistical models, which is applicable to both quantitative and qualitative traits. Depending on the family structures, however, not all the variance components in a GLMM are always estimable. For example, for a quantitative trait with $g(\cdot)$ being the identity link function, the GLMM becomes a linear mixed model (LMM).

$$y_{ij} = m + z_{ij}\alpha + x(g_{ij})\beta + v_{ij} + e_i + \epsilon_{ij},$$

where $i = 1, ..., N$, $j = 1, ...n_i$, $\epsilon_{ij} \sim N(0, \sigma^2)$ are the model residuals, with σ^2 being the residual variance of random effects from other risk factors not captured by z_{ij}, g_{ij}, v_{ij}, and e_i. When each family comprises two parents and one offspring, the expected $\Delta_i = I_3$, which is a 3×3 identity matrix for any family i. We have

$$\mathrm{Cov}(\mathbf{y_i}) = 2\Phi_i\sigma_A^2 + (\sigma_D^2 + \sigma^2)I_3 + \sigma_c^2 J_3,\ i = 1, \cdots, N$$

Thus, the dominant variance component σ_D^2 and the residual variance σ^2 are completely confounded. Similarly, if each family consists of only siblings (e.g., in a sib-pair design), then

$$\begin{aligned}\mathrm{V}(y_{ij}) &= \sigma_A^2 + \sigma_D^2 + \sigma_c^2 + \sigma^2,\ i = 1, \cdots, N\\ \mathrm{Cov}(y_{ij}, y_{ik}) &= \frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2 + \sigma_c^2,\ j \neq k\end{aligned}$$

In this case, we cannot distinguish the four variance components from each other unless some of them are negligible (e.g., assuming $\sigma_D^2 = \sigma_c^2 = 0$). For a sample of unrelated individuals, it is also easy to see that all the four variance components σ_A^2, σ_D^2, σ_c^2, and σ^2 are inseparable.

2.2. In Fitting the GLMM

The main goal in fitting GLMM Equation (1) is to make statistical inference on the fixed effects α, β as well as assessing the variance components σ_A^2, σ_D^2, and σ_c^2. Based on model Equation (1), the full likelihood is

$$L(\alpha, \beta|\{y_{ij}, z_{ij}, g_{ij}\}) = \prod_{i=1}^{N} \int_{\mathbf{v_i}} \int_{e_i} f_{\mathbf{y_i}|\mathbf{v_i}, e_i}(\mathbf{y_i}|\mathbf{v_i}, e_i) f_{\mathbf{v_i}}(\mathbf{v_i}) f_{e_i}(e_i) de_i d\mathbf{v_i}$$

We need to calculate a multi-dimensional integration for each family, which in most cases cannot be analytically evaluated in closed forms. Various methods in fitting a standard GLMM have been proposed based on either numerical approximations to the integrations or a linearization of the regression model. Traditional methods for numerical integral approximation include Laplace approximation (Wolfinger, 1993), the adaptive Gauss-Hermite quadrature (Pinheiro and Bates, 1995), Monte Carlo integration, and Bayesian method via Markov chain Monte Carlo. The model linearization is often made via Taylor expansion on the inverse of the nonlinear link function $g(\cdot)$, based on which the pseudo-likelihood or restricted pseudo-likelihood for optimization can then be derived (Wolfinger and O'Connell, 1993). It has been known that the numerical integral approximations could become computationally intractable when a large number of random effects are involved. On the other hand, the model linearization approach could encounter severe unconvergence problems especially for binary outcomes with small cluster (family) sizes.

In practice, several common statistical software packages are available in fitting a standard GLMM. These include but not limited to "proc nlmixed" and "proc glimmix" procedures in SAS, and OpenBUGS for Bayesian approach. The "proc nlmixed" mainly conducts integral approximations using an adaptive Gauss-Hermite quadrature as default, and then directly maximizes the approximately integrated likelihood. In contrast, the "proc glimmix" primarily performs several model linearization based pseudo-likelihood methods, although it can also fit the GLMM using Laplace or adaptive Gauss-Hermite quadratures for integral approximations (see documentations supported by SAS Institute Inc, Raleigh, NC). In addition, the model fitting algorithm in R package "lme4" consists of an iteration between two sub-optimization procedures. One is to determine the conditional modes of the fixed and random effects, given the current deviance and variance components using penalized iteratively re-weighted least squares (PIRLS). The other is to obtain MLE of the deviance and variance components based on a profile likelihood from Laplace approximation, given the current conditional modes of the fixed and random effects (Bates et al., 2012). Nonetheless, these software packages typically require the random effects to have the same covariance (or correlation) structures across all the clusters. When families have different sizes or varied kinship or double coancestry matrices, it is difficult to directly specify GLMM Equation (1) using these software packages. It should be pointed out that the R package *pedigreemm* can fit a GLMM to family data with complex family structures without separating families into clusters. It also allows individuals from different families to be correlated. Here, we focus on fitting GLMM (1) using "proc nlmixed" and "proc glimmix" procedures in SAS, or OpenBUGS as an alternative choice.

To deal with different families sizes or varied kinship or double coancestry matrices, we apply a Cholesky decomposition based re-parameterization of the random genetic effects $\mathbf{v_i}$ to re-formulate GLMM Equation (1) into a GLMM with random regression coefficients. First, we apply separate Cholesky decompositions on the kinship and double coancestry matrices: $2\Phi_i = L_{\Phi i}L_{\Phi i}^T$, $\Delta_i = L_{\Delta i}L_{\Delta i}^T$, for $i = 1, \cdots, N$. We then re-parameterize the random genetic effects within family i as $\mathbf{v_i} = L_{\Phi i}\mathbf{a_i} + L_{\Delta i}\mathbf{d_i}$, where $\mathbf{a_i} = (a_{i1}, \cdots, a_{ir_i})^T \sim N(0, \sigma_A^2 I_{r_i})$ with $r_i = rank(\Phi_i)$, and $\mathbf{d_i} = (d_{i1}, \cdots, d_{is_i})^T \sim N(0, \sigma_D^2 I_{s_i})$ with $s_i = rank(\Delta_i)$.

Besides, we assume that $\mathbf{a_i}$'s are independent of $\mathbf{d_i}$'s. By replacing $\mathbf{v_i}$ by $L_{\Phi i}\mathbf{a_i} + L_{\Delta i}\mathbf{d_i}$ in model Equation (1), we have

$$\mathrm{Cov}(\mathbf{v_i}) = \mathrm{Cov}(L_{\Phi i}\mathbf{a_i} + L_{\Delta i}\mathbf{d_i}) = 2\Phi_i\sigma_A^2 + \Delta_i\sigma_D^2 = \mathbf{\Sigma_i}$$

Note that the random effects $\mathbf{a_i}$ and $\mathbf{d_i}$ are orthogonal within families as well as across families, even though the dimension of these random effects may still vary from family to family because the families may have different sizes.

Next, to deal with possible different family sizes, let $r = max_i\{r_i\}$ and $s = max_i\{s_i\}$ be the maximum number of columns in $\{L_{\Phi i}, i = 1, \cdots, N\}$ and $\{L_{\Delta i}, i = 1, \cdots, N\}$, respectively. For those families with $r_i < r$ (or $s_i < s$), we further expand their design matrices $L_{\Phi i}$ (or $L_{\Delta i}$) to r (or s) columns by adding $r - r_i$ (or $s - s_i$) columns of 0's at the right end (or any places). Then, we obtain a re-formulated GLMM with all the families having the same dimension r (or s) for their random effects $\mathbf{a_i}$ (or $\mathbf{d_i}$) as the following.

$$g(\mu_\mathbf{i}) = \mathbf{1_{n_i}}m + Z_i\alpha + X(\mathbf{g_i})\beta + L_{\Phi i}\mathbf{a_i} + L_{\Delta i}\mathbf{d_i} + \mathbf{1_{n_i}}e_i$$

where $Z_i = (z_{i1}, \cdots, z_{in_i})^T$, $X(\mathbf{g_i}) = (x(g_{i1}), \cdots, x(g_{in_i}))^T$, $\mathbf{a_i} \sim N(0, \sigma_A^2 I_r)$, $\mathbf{d_i} \sim N(0, \sigma_D^2 I_s)$ and $e_i \sim N(0, \sigma_c^2)$. Finally, we construct the design matrix L_Φ of $\mathbf{a_i}$ from $L_{\Phi i}$, $i = 1, \cdots, N$, by stacking them one above the other; and similarly build the design matrix L_Δ of $\mathbf{d_i}$ by stacking $L_{\Delta i}$, $i = 1, \cdots, N$, one above the other. Now, the columns of the two matrices L_Φ and L_Δ can be treated as $(r + s)$ ordinary fixed continuous covariates with $\mathbf{a_i}$'s and $\mathbf{d_i}$'s being their random regression coefficients. Within each family, all the family members share the same set of slopes. Across different families, the $\mathbf{a_i}$, $i = 1, \cdots, N$ (or $\mathbf{d_i}$, $i = 1, \cdots, N$), are independent but share the same variance component σ_A^2 (or σ_D^2).

The re-formulated GLMM above can be easily specified by "proc nlmixed" or "proc glimmix" procedures in SAS. With "proc glimmix," we can use three separate "random" commands "random *e*/subject = famid," "random La_1 ... La_r/subject = famid type = TOEP(1);" and "random Ld_1 ... Ld_s/subject = famid type = TOEP(1);" to specify the correlation structures for the random effects e_i, $\mathbf{a_i}$, and $\mathbf{d_i}$, respectively. Here, La_1, ... , La_r represent the columns of the stacked matrix L_Φ; Ld_1 ... Ld_s denote the columns of the stacked matrix L_Δ. The option "TOEP(1)" can force all the elements in $\mathbf{a_i}$ (or $\mathbf{d_i}$) to share the same variance component σ_A^2 (or σ_D^2). For "proc nlmixed," currently it only allows to have one "random" command. But we can specify a joint multivariate normal distribution for e_i, $\mathbf{a_i}$, and $\mathbf{d_i}$ via "random e a_1 ... a_r d_1 ... d_s ~ normal(mu, v) subject=famid," where v is a diagonal matrix with one σ_c^2, r elements of σ_A^2's and s elements of σ_D^2's on its diagonal. As an example, the SAS codes for specification of a GLMM using both "proc glimmix" and "proc nlmixed" for families with two parents and two full sibs (i.e., $r = s = 4$) are provided in Appendices A, B (Supplementary Material), respectively. We also explored using the R package *lme4* to fit the re-formulated GLMM. But it appears that the functions "lmer" and "glmer" provided in *lme4* do not have an option that can force all the elements in $\mathbf{a_i}$ (or $\mathbf{d_i}$) to share the same variance component σ_A^2 (or σ_D^2).

This re-formulation also makes it more convenient to fit GLMM Equation (1) using the Markov chain Monte Carlo (MCMC) based Bayesian approach. We use R package BRugs to get access to OpenBUGS software (Christensen et al., 2011), which has the Bayesian approach implemented. It is noticed that OpenBUGS has a weak support for matrix operations. In the original GLMM Equation (1), the genetic covariance matrix $\mathbf{\Sigma_i} = \mathbf{2\Phi_i\sigma_A^2} + \mathbf{\Delta_i\sigma_D^2}$ involves two unknown variance components σ_A^2 and σ_D^2. As a result, we cannot directly specify the covariance matrix $\mathbf{\Sigma_i}$ in OpenBUGS. With the re-formulated GLMM, we can pass L_Φ and L_Δ as fixed covariates to OpenBUGS with $\mathbf{a_i}$ and $\mathbf{d_i}$ being their random regression coefficients. Since the Bayesian approach often treats all the model parameters as random, it appears especially suitable for fitting the re-formulated GLMM. We can also extract from the MCMC the posterior distributions of random effects for each individual, which allow us to assess the variation contribution from the putative random genetic effects $\{v_{ij}\}$ to the total variance of $\{g(\mu_{ij})\}$. The R codes for specification of a re-formulated GLMM using BRugs + OpenBUGS are also provided in Appendix C (Supplementary Material).

3. Simulation Study

In this section, we examine the performances of procedures "proc nlmixed" and "proc glimmix" in SAS (version 9.3) as well as R packages BRugs + OpenBUGS in fitting the re-formulated GLMM through simulation. We consider three biallelic genetic markers with alleles "0" or "1," and one fixed explanatory covariate $Z \sim Bernoulli(0.5)$. The three genetic markers are assumed to be unlinked (i.e., independent) and have allele frequencies $p_1 = 0.5, p_2 = 0.2, p_3 = 0.1$ for alleles "1" at locus 1, 2, and 3, respectively. We first generate a pairs of haplotypes independently for each parent, where each haplotype is comprised of three alleles randomly generated from $Bernoulli(p_j)$ for $j = 1$, 2, and 3, respectively. Each child inherits one haplotype from father, and the other from mother. The haplotype from father (or mother) consists of three alleles with each allele being selected from the two paternal (or maternal) alleles at the same locus with 50% chance. In case where a family has more than one child, the above random process is repeated for each child independently. For simplicity, we use the expected values to construct the kinship and double coancestry matrices for each family.

We first consider simulating quantitative traits from a LMM. Let $\mathbf{g}_{ij} = (g_{ij1}, g_{ij2}, g_{ij3})$ be the genotypes of the j-th subject in family i, where $g_{ijk} \in \{0, 1, 2\}$ counts the number of alleles "1" at locus $k = 1, 2, 3$. The quantitative trait values are simulated from the following LMM.

$$y_{ij} = m + z_{ij}\alpha + x(g_{ij1})\beta_1 + x(g_{ij2})\beta_2 + x(g_{ij3})\beta_3 + v_{ij} + e_i + \epsilon_{ij},\ i = 1, ..., n,\ j = 1, ...n_i$$

where $x(g_{ijk}) = (x_a(g_{ijk}), x_d(g_{ijk}))\beta_k$ with $x_a(g_{ijk}) = g_{ijk}$, $x_d(g_{ijk}) = 1$ (or 0) if $g_{ijk} = 2$ (or otherwise) based on the allele-coding (see Wang, 2011), and $\beta_k = (\beta_{ka}, \beta_{kd})^T$ being the effects of marker k for $k = 1, 2, 3$. We further set $\sigma_c^2 = \sigma^2 = \sigma_A^2 = 1$ and $\sigma_D^2 = 0.5$. The other true values of parameters are listed in **Table 1**. Each simulation data set contains n_1 families with two parents and one child and n_2 families with two parents and two

TABLE 1 | The true values of model parameters in simulation.

Parameter	Definition	True value
n_1	The number of families with one child	0, 500, 1000
n_2	The number of families with two children	500, 1000, 2000
m	The intercept	10
α	The fixed effect of Z	2
β_1	The genetic effects of marker 1	(1, -2)
β_2	The genetic effects of marker 2	(1, 0)
β_3	The genetic effects of marker 3	(1, -1)

children. We consider three cases: (a) $n_1 = 500$, $n_2 = 500$; (b) $n_1 = 1000$, $n_2 = 1000$; (c) $n_1 = 0$, $n_2 = 2000$.

For each simulation data set, we fit the GLMM by using two methods: (1) adaptive Gaussian quadrature (AGQ) via "proc nlmixed" in SAS; (2) Bayesian approach via BRugs + OpenBUGS. Based on 200 simulation data sets, the parameter estimates are summarized in **Table 3**. We also explored using "proc glimmix" to fit the simulation data but abandoned it due to some severe unconvergence problems. In running BRugs + OpenBUGS, we choose the following priors for initialization of the parameters: $m \sim N(0, 10)$, $\alpha \sim N(0, 10)$, $\beta_{ka} \sim N(0, 10)$ and $\beta_{kd} \sim N(0, 10)$ for $k = 1, 2, 3$, $\tau_a \sim Gamma(1, 1)$, $\sigma_D \sim Uniform(0, 5)$, $\tau_c \sim Gamma(1, 1)$, and $\tau \sim Gamma(1, 1)$, where $\tau_a = 1/\sigma_A^2$, $\tau_d = \sigma_D^2$, $\tau_c = 1/\sigma_c^2$, and $\tau = 1/\sigma^2$. We use the first 10,000 updates as burn-in, and another 10,000 updates to estimate parameters as posterior means. The parameter estimates and their standard deviations (SD) from the 200 simulations are summarized in **Table 2**. For each simulation data, we also calculate the length of the 95% confidence (or probability) interval for each parameter in running "proc nlmixed" (or BRugs + OpenBUGS). The average length of these intervals and their coverage rate for each true parameter value from the 200 simulations are summarized in **Table 3**.

From **Table 2** we can see that both methods can provide reasonable estimates of the fixed effects α, β_{ka}, and β_{kd} ($k = 1, 2, 3$) as well as the variance components with improved accuracy as sample size increases. As expected, the estimate of allelic interaction β_{kd} has a larger SD than that of the additive allelic effect β_{ka} at each locus k, for $k = 1, 2, 3$. The estimates of σ_D^2 and σ^2 are slightly biased, which could be caused by the potential confounding between these two variance components. Besides, it appears that "proc nlmixed" tends to over-estimate the dominant variance σ_D^2 and under-estimate the residual variance σ^2. Meanwhile, the Bayesian method heads to the opposite way. From **Table 3** the coverage probabilities are close to the nominal level of 95% for most parameters except σ_D^2 and σ^2. The average lengths of the 95% confidence (or probability) intervals for σ_D^2 and σ^2 are also greater than the ones for other two variance components. In addition, the average length of the 95% confidence (or probability) intervals for β_{3d} is substantially larger than that for β_{1d} and β_{2d}, which is likely caused by the fact that the allele "1" at marker locus 3 is rare and the homozygous genotypes "11" at marker locus 3 are much less present in a simulation data set than those at the other two loci.

Overall, the results from "proc nlmixed" and BRugs + OpenBUGS are quite comparable in all three cases. As we have mentioned before, the variance components σ^2 and σ_D^2 are confounded in single-child families, and only two-child families are informative for distinguishing them. In case (b), where there is an increased number of 2-child families, the estimates of these two variance components are improved with both methods. In case (c), the accuracy in estimates of these two variance components is further improved. In terms of the computational speed, it appears that the "proc nlmixed" runs much faster than BRugs + OpenBUGS. We ran BRugs + OpenBUGS on one laptop installed with Intel(R) Core(TM)i7-3520M CPU @ 2.90GHz. It took about 19 h, 50 h and 38 min, and 59 h and 15 min of CPU time to complete the 200 simulations (including data generation and MCMC iterations) in cases (a), (b), and (c), respectively. The "proc nlmixed" procedure in SAS was performed on a UNIX workstation which has a compatible speed with the laptop, and it took about 2 and a half hours, 6 and a half hours, and 7 h 18 min of CPU time to complete the 200 simulations in cases (a), (b), and (c), respectively. By contrast, the Bayesian approach can provide the posterior distributions for all the parameters rather than just the modes (MLE) and their variance or covariance estimates. When σ^2 and σ_D^2 are almost completely confounded, we found that the "proc nlmixed" may give unreliable estimates of the model parameters or encounter unconvergence problem caused by the nearly singular Hessian matrix, while the Bayesian approach can still provide reasonable estimates at least for other model parameters except σ^2 and σ_D^2.

We also consider binary traits and simulate phenotypic values from the following mixed logistic regression model.

$$\begin{aligned}\text{logit}P(y_{ij} = 1|z_{ij}, g_{ij1}, g_{ij2}, g_{ij3}, v_{ij}, e_i) &= m + z_{ij}\alpha \\ &+ x(g_{ij1})\beta_1 + x(g_{ij2})\beta_2 + x(g_{ij3})\beta_3 + v_{ij} + e_i,\end{aligned}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. In order to have enough number of events, we choose $m = -3$. Meanwhile, we keep the same true values as before for other model parameters. We explored various options on using "proc nlmixed" and "proc glimmix" but failed to fit the above model appropriately due to severe unconvergence problems. By contrast, BRugs + OpenBUGS can still provide reasonable estimates for most of the model parameters. Using the similar priors in the previous setting, 10,000 burn-in and 10,000 updates for parameter estimation, we compute the means and standard deviations (SD) of the parameter estimates and the average lengths (AL) and coverage rates (CR) of the 95% probability intervals from 200 simulations as summarized in **Table 4**.

Under our simulation setting, it appears that the SAS procedures "proc nlmixed" and "proc glimmix" perform unexpectedly poorly especially for binary outcomes. But we have to admit that, although we have explored many options that are available in running "proc nlmixed" and "proc glimmix," our exploration is surely not exclusive based on our limited knowledge. Besides, we use SAS version 9.3 on our Linux computer system. Recently, a newer version 9.4 of SAS has become available for PC users, which may provide improved performance in fitting the GLMM. Unfortunately, we do not have access to this new version of SAS

TABLE 2 | Mean and standard deviation (SD) of the parameter estimates from 200 simulations for linear mixed models.

No. families	(a) n_1=500, n_2 = 500		(b) n_1 = 1000, n_2 = 1000		(c) n_1 = 0, n_2 = 2000	
Parameters	**AGQ**	**Bayesian**	**AGQ**	**Bayesian**	**AGQ**	**Bayesian**
m	10.01 (0.07)	10.00 (0.07)	10.00 (0.06)	9.99 (0.05)	10.00 (0.05)	10.00 (0.05)
α	2.00 (0.06)	2.01 (0.06)	2.00 (0.04)	2.00 (0.04)	2.00 (0.03)	2.00 (0.03)
β_{1a}	1.00 (0.07)	1.00 (0.07)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)
β_{1d}	−2.00 (0.11)	−2.00 (0.11)	−2.01 (0.08)	−2.00 (0.07)	−1.99 (0.07)	−2.00 (0.07)
β_{2a}	0.99 (0.06)	1.01 (0.06)	1.01 (0.05)	1.00 (0.05)	1.00 (0.04)	1.01 (0.05)
β_{2d}	0.04 (0.17)	−0.01 (0.18)	−0.00 (0.11)	−0.02 (0.11)	−0.00 (0.10)	0.00 (0.11)
β_{3a}	1.00 (0.08)	1.00 (0.08)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)	1.00 (0.05)
β_{3d}	−0.98 (0.30)	−0.97 (0.33)	−1.00 (0.21)	−0.99 (0.21)	−1.00 (0.21)	−0.98 (0.17)
σ_A^2	1.01 (0.18)	0.98 (0.18)	1.01 (0.12)	1.00 (0.12)	1.00 (0.11)	1.00 (0.11)
σ_D^2	0.54 (0.36)	0.46 (0.31)	0.53 (0.30)	0.49 (0.27)	0.51 (0.22)	0.49 (0.22)
σ_c^2	0.99 (0.11)	1.01 (0.12)	0.99 (0.08)	1.00 (0.08)	0.99 (0.08)	1.01 (0.08)
σ^2	0.95 (0.36)	1.06 (0.32)	0.97 (0.31)	1.02 (0.27)	0.99 (0.23)	1.03 (0.23)

TABLE 3 | The average length of the 95% confidence (or probability) intervals and the coverage rate for true parameters from 200 simulations for linear mixed models.

No. families	(a) n_1 = 500, n_2 = 500		(b) n_1 = 1000, n_2 = 1000		(c) n_1 = 0, n_2 = 2000	
Parameters	**AGQ**	**Bayesian**	**AGQ**	**Bayesian**	**AGQ**	**Bayesian**
m	0.31, 96.0%	0.31, 96.0%	0.22, 94.5%	0.22, 96.0%	0.21, 96.0%	0.21, 96.5%
α	0.21, 94.5%	0.21, 94.5%	0.15, 95.0%	0.15, 97.0%	0.14, 96.0%	0.14, 95.0%
β_{1a}	0.28, 96.0%	0.28, 94.0%	0.20, 95.0%	0.20, 96.5%	0.18, 95.0%	0.18, 96.5%
β_{1d}	0.43, 95.5%	0.43, 95.0%	0.30, 93.5%	0.30, 97.0%	0.28, 93.0%	0.28, 92.0%
β_{2a}	0.26, 97.5%	0.26, 96.5%	0.18, 95.5%	0.18, 94.5%	0.17, 97.0%	0.17, 90.5%
β_{2d}	0.67, 93.5%	0.67, 92.0%	0.47, 96.0%	0.47, 97.5%	0.44, 96.0%	0.44, 96.0%
β_{3a}	0.32, 95.0%	0.32, 96.5%	0.23, 98.5%	0.23, 96.5%	0.21, 95.5%	0.21, 94.5%
β_{3d}	1.22, 96.5%	1.20, 92.5%	0.85, 94.5%	0.85, 97.0%	0.78, 95.0%	0.78, 98.5%
σ_A^2	0.68, 94.5%	0.68, 92.0%	0.49, 96.0%	0.48, 95.0%	0.47, 97.0%	0.46, 95.0%
σ_D^2	1.61, 87.0%	0.77, 74.0%	1.16, 89.5%	0.66, 75.0%	0.84, 93.5%	0.62, 84.0%
σ_c^2	0.44, 94.5%	0.44, 95.0%	0.32, 94.5%	0.31, 94.0%	0.31, 96.0%	0.31, 94.0%
σ^2	1.54, 87.0%	0.87, 74.5%	1.12, 88.5%	0.72, 75.5%	0.86, 95.0%	0.67, 85.0%

TABLE 4 | Means (SD) of the parameter estimates and AL (CR) of the 95% probability intervals from 200 simulations for mixed logistic regression models.

No. families	(a) n_1 = 500, n_2 = 500		(b) n_1 = 1000, n_2 = 1000		(c) n_1 = 0, n_2 = 2000	
Parameters	**Mean (SD)**	**AL (CR)**	**Mean (SD)**	**AL (CR)**	**Mean (SD)**	**AL (CR)**
m	−2.69(0.20)	0.82, 69.0%	−2.69(0.16)	0.63, 56.0%	−2.69(0.14)	0.58, 47.0%
α	1.80(0.13)	0.58, 75.5%	1.79(0.11)	0.44, 58.0%	1.80(0.10)	0.41, 58.5%
β_{1a}	0.89(0.14)	0.55, 85.5%	0.89(0.10)	0.40, 80.5%	0.90(0.09)	0.37, 77.5%
β_{1d}	−1.78(0.22)	0.90, 85.5%	−1.79(0.16)	0.66, 78.5%	−1.81(0.16)	0.62, 74.0%
β_{2a}	0.90(0.12)	0.49, 87.5%	0.89(0.11)	0.36, 78.0%	0.91(0.09)	0.34, 83.5%
β_{2d}	0.01(0.31)	1.24, 95.5%	0.02(0.23)	0.87, 95.0%	−0.01(0.22)	0.82, 94.5%
β_{3a}	0.88(0.15)	0.58, 85.0%	0.89(0.11)	0.42, 83.0%	0.88(0.10)	0.39, 77.0%
β_{3d}	−0.84(0.55)	2.20, 95.0%	−0.85(0.40)	1.56, 93.0%	−0.84(0.36)	1.46, 94.5%
σ_A^2	1.11(0.45)	2.00, 95.5%	1.08(0.42)	1.56, 93.0%	1.02(0.32)	1.45, 98.5%
σ_D^2	0.49(0.18)	0.80, 94.5%	0.52(0.20)	0.73, 93.5%	0.57(0.17)	0.72, 94.5%
σ_c^2	0.74(0.19)	0.83, 78.0%	0.72(0.14)	0.64, 62.0%	0.75(0.14)	0.62, 70.0%

yet based on our current Linux computer system. The performance of "proc nlmixed" and "proc glimmix" in SAS 9.4 on fitting the re-formulated GLMM needs further exploration.

While the frequencist approach implemented in "proc nlmixed" and "proc glimmix" often require numerical approximations to the full likelihood, the Bayesian approach directly maximizes the full likelihood via MCMC. From our simulation study, it seems that the Bayesian approach can better handle the re-formulated GLMM especially for binary traits. However, it should be pointed out that the choice of priors in using the Bayesian approach could have a significant impact on convergence of an MCMC procedure. Besides the gamma priors for all the τ's, we also tested using uniform priors on σ's and obtained similar results. Throughout our simulation, the Monte-Carlo errors for all the parameter estimates appear to be acceptable. But some auto-correlations in certain Markov chains (e.g., the ones for σ_D^2 and σ^2) are noticed. Typically, the auto-correlation could be reduced by thinning the Markov chains. Otherwise, appropriate adjustment is needed in computing the SD of parameter estimates.

4. Analysis of T1DGC data for Type I Diabetes

As an example, we consider fitting a real family data set obtained from the Type 1 Diabetes Genetics Consortium (T1DGC). The data set includes five cohorts: Asia-Pacific (AP), Danish Steno Diabetes Center (DAN), European (EUR), Sardinian (SAR) and United Kingdom Genetic Resource Investigating Diabete (UK). For simplification, we only adopt nuclear families and exclude some grand-parents or grand children (less than 1% of the total subjects). Most of the families (about 80.4%) consist of 2 parents and 2–3 children. There are 13 families that have more than 9 family members, and they all belong to the DAN cohort. The actual numbers of subjects and families we used in the five cohorts are listed in **Table 5**.

Our research interest is to test for association of HLA-DQB1 locus with Type 1 Diabetes (T1D) incidence, while appropriately controlling for other potential genetic risk factors on the incidence of T1D. The adjustment for random additive and dominance effects is important because it has been known that T1D is a polygenic disease. Some studies have suggested that other genes such as INS and CTLA4 could be implicated with T1D (Anjos and Polychronakos, 2004; McGinnis et al., 2009). One article "Genetics and Diabetes" from the World Health Organization (WHO) web site "*http://www.who.int/genomics/about/Diabetis-fin.pdf*" also provides a nice review of the T1D. From our previous study (Glisic et al., 2009), we classify the subjects into 4 groups: low risk (DQrisk = 0), moderate risk (DQrisk = 1), high risk (DQrisk = 2), and very high risk (DQrisk = 3) based on the HLA-DQB1 genotypes and CD4 + CD25 + ^{high}T-cell apoptosis. Gender and age are also known risk factors for T1D. We categorize age into 6 categories: age≤18, 18<age≤30, 30<age≤40, 40<age≤50, 50<age≤60, and age>60. By choosing the "high risk" (the largest group across all cohorts) at HLA-DQB1, male and "age≤18" as a baseline, we fit each cohort separately using the following mixed logistic regression model.

$$\begin{aligned}&\text{logit}P(y_{ij}=1|v_{ij},e_i)=\mu+\alpha_1*1_{(18<age\le 30)}+\\&\alpha_2*1_{(30<age\le 40)}+\alpha_3*1_{(40<age\le 50)}+\\&\alpha_4*1_{(50<age\le 60)}+\alpha_5*1_{(age>60)}+\alpha_6*1_{(sex=F)}+\\&\beta_1*1_{(DQrisk=0)}+\beta_2*1_{(DQrisk=1)}+\\&\beta_3*1_{(DQrisk=3)}+v_{ij}+e_i\end{aligned}$$

where $e_i \sim N(0,\sigma_c^2)$, $e^{\alpha_i}(i=1,\cdots,5)$ are odds ratios of having T1D in different age groups comparing with the youngest age group of age ≤18, e^{β_j} $(j=1,2,3)$ are odds ratios of having T1D in different HLA-DQB1 risk groups comparing with the high risk group, e^{β_6} is the odds ratio of T1D in females versus males, and $\{v_{ij}\}$ have the covariance structures as specified in GLMM Equation (1).

To simplify the calculation, we use the expected values to construct the kinship and double coancestry matrices for each family. As the actual kinship and double coancestry matrices should be close to their expected matrices, we would expect only minor deviations from the fitted GLMM. In running "proc glimmix" and "proc nlmixed" to fit the GLMMs, we encountered some severe unconvergence problems (data not shown). In running BRugs + OpenBUGS, we use 10,000 burn-in and another 10,000 updates to estimate the parameters. For the DAN cohort, the current BRugs + OpenBUGS cannot accommodate more than 12 additive or dominant random coefficients in the model specification—see part (2) of Appendix C in Supplementary Material, where "maxsize" (i.e., the number of columns of the stacked matrices L_Φ and L_Δ) cannot exceed 12 even though the total number of family members can still exceed 12. We also find that some of the parameter estimates become unstable when we actually use 10–12 columns. So we adopt using 8 columns of the stacked matrices L_Φ and L_Δ in fitting the Dan cohort. The estimates of odds ratios and variance components in the fitted models from running BRugs + OpenBUGS are summarized in **Table 6**.

The results have confirmed that age is significantly associated with T1D incidence. In most cohorts except AP, the odds of T1D occurrence reaches the highest in the youngest group of age < 18 and then decreases quickly as age increases. The females appear to have a less chance of having T1D than males in DAN and EUR cohorts. For HLA-DQB1, it appears that in each cohort there is a significant increase of T1D risk in the HLA-DQB1 very high risk group comparing with the high risk group, and the high risk group is also significantly different from the low risk and moderate risk groups after adjusting for the age and gender effects. The odds of T1D for the high or very high risk groups appear significantly higher than that for the low or moderate risk groups, which are consistent with the relative risk of 3–45 of the HLA-DQB1 susceptibility variant reported in "Genetics and Diabetes" from WHO.

TABLE 5 | Number of subjects and families in T1DGC by cohorts.

Cohort	AP	DAN	EUR	SAR	UK
No. of subjects	741	664	1936	347	465
No. of families	184	147	475	77	113
Maximum family size	6	14	8	6	6

TABLE 6 | Posterior means and 2.5%, 97.5% percentiles of the odds ratios and variance components for type I diabetes in five cohorts of the T1DGC data set.

Cohorts	AP	DAN	EUR	SAR	UK
Baseline intercept (μ)	2.32 (1.58, 3.29)	2.64 (1.72, 3.67)	3.70 (3.05, 4.48)	1.11 (0.12, 2.14)	2.22 (1.44, 3.20)
18<age≤30 vs. age≤18 ($e^{\alpha}1$)	1.23 (0.56, 2.82)	0.50 (0.19, 1.39)	0.35 (0.20, 0.59)	0.69 (0.21, 2.32)	0.45 (0.15, 1.35)
30<age≤40 vs. age≤18 ($e^{\alpha}2$)	0.10 (0.036, 0.25)	0.25 (0.09, 0.64)	0.047 (0.024, 0.086)	0.48 (0.15, 1.45)	0.014 (0.003, 0.045)
40<age≤50 vs. age≤18 ($e^{\alpha}3$)	0.01 (0.003, 0.04)	0.07 (0.02, 0.19)	0.007 (0.002, 0.014)	0.06 (0.01, 0.22)	0.004 (0.001, 0.013)
50<age≤60 vs. age≤18 ($e^{\alpha}4$)	0.005 (0.001, 0.022)	0.04 (0.01, 0.11)	0.002 (0.0004, 0.004)	0.009 (0.001, 0.063)	0.005 (0.001, 0.025)
age>60 vs. age≤18 ($e^{\alpha}5$)	0.012 (0.002, 0.055)	0.009 (0.002, 0.032)	0.0004 (0.0001, 0.0015)	0.003 (0.0001, 0.023)	0.027 (0.0003, 1.25)
Female vs. Male ($e^{\alpha}6$)	1.27 (0.75, 2.21)	0.60 (0.35, 0.97)	0.60 (0.43, 0.83)	0.60 (0.24, 1.38)	1.36 (0.64, 2.92)
DQrisk = 0 vs. 2 ($e^{\beta}1$)	0.09 (0.025, 0.26)	0.02 (0.005, 0.06)	0.04 (0.02, 0.08)	0.11 (0.01, 0.71)	0.02 (0.002, 0.12)
DQrisk = 1 vs. 2 ($e^{\beta}2$)	0.14 (0.05, 0.30)	0.30 (0.14, 0.59)	0.25 (0.15, 0.39)	0.19 (0.04, 0.63)	0.24 (0.08, 0.71)
DQrisk = 3 vs. 2 ($e^{\beta}3$)	2.26 (1.17, 4.80)	5.84 (2.75, 14.47)	5.91 (3.48, 10.61)	15.89 (4.80, 82.60)	8.36 (3.30, 26.13)
Additive variance (σ_A^2)	0.71 (0.20, 1.90)	0.69 (0.21, 1.73)	0.47 (0.17, 1.05)	0.68 (0.19, 2.11)	0.71 (0.17, 2.05)
Dominant variance (σ_D^2)	1.64 (0.28, 6.35)	1.44 (0.21, 4.93)	0.66 (0.23, 1.85)	2.96 (0.41, 10.51)	1.25 (0.28, 4.32)
Family-shared variance (σ_c^2)	0.62 (0.20, 1.45)	0.66 (0.22, 1.42)	2.36 (1.33, 3.92)	0.56 (0.17, 1.42)	0.63 (0.19, 1.67)

Regarding the variance components, the estimates of the additive and family shared variances appear reasonably well, although the dominance variances in SAR and AP cohorts have relatively large variation due to perhaps the lack of information in their family data. To see whether we should not include σ_D^2 in the models, we also fit the GLMM models without σ_D^2 and compare them with our previous models. Based on the deviance information criterion (DIC), which is an estimate of the expected predictive error (lower deviance is better) and it can account for both model fitness and model complexity, the models with both additive and dominance variances included have their DIC values of 658.6, 647.3, 1525, 342.2, and 316.6 which are lower than the DIC values of 669.5, 658.9. 1543. 354.1, and 322.3 in the reduced models without σ_D^2 in AP, DAN, EUR, SAR, and UK cohorts, respectively. So the models with both additive and dominance variances included are preferable. We also estimate the variations contributed by the random putative genetic effects $\{v_{ij}\}$, which account for 22%, 26%, 9%, 27%, and 12% of the total variation in $\text{logit}P(y_{ij} = 1|v_{ij}, e_i)$ for the AP, DAN, EUR, SAR, and UK cohorts, respectively.

The T1DGC family data was collected from the observational retrospective sampling, which likely had over-sampled families with T1D children. It is well known that fitting a mixed logistic regression model with random effects to a retrospective data set may no longer provide the equivalent maximum likelihood estimates of the model parameters as the ones defined in the same model for a prospective cohort. Therefore, the results above are only applicable to the families we obtained from T1DGC. For general populations, an adjustment for the retrospective sampling strategy is needed in order to avoid the bias in the parameter estimates.

5. Discussion

In this study, we propose a Cholesky decomposition based re-formulation of the GLMM to fit family data with varied sizes and diverse genetic relatedness. Assuming that there is no genetic correlation between different families, we first apply separate Cholesky decompositions on the genetic kinship (or double coancestry) matrices within each family. Next, we stack these Chelosky decomposition matrices from different families into columns to form two stacked matrices. The columns of the stacked matrices can then be treated as fixed covariates with random regression coefficients of equal variances. It should be pointed out that applying separate Cholesky decompositions on each family does not provide computational or storage benefits comparing with the regular Cholesky decomposition on the whole covariance matrix of the random genetic effects because the non-diagonal blocks in the sparse matrix decomposition are actually not saved. However, with the reduced number of columns in our stacked matrices, this re-formulated GLMM can be specified more conveniently using "proc nlmixed" and "proc glimmix" in SAS, or OpenBUGS via R package BRugs. Theoretically, the re-formulated GLMM is equivalent to the original GLMM. Therefore, it should retain the same validity and power in testing the fixed and random genetic effects as the original GLMM. From our simulation, it appears that this re-formulated GLMM can be fitted reasonably well by "proc nlmixed" and BRugs + OpenBUGS for quantitative traits with moderate family sizes. For binary traits, our simulation and real data example shows that at least BRugs + OpenBUGS can appropriately fit the re-formulated GLMM for families of sizes not exceeding 12.

This Cholesky decomposition based re-formulation of GLMM in fitting family data is somewhat analogous to Haseman and Elston's regression method for sib-pairs (Haseman and Elston, 1972). While Haseman and Elston's algorithm regresses the squared sib-pair's phenotypic difference on the kinship and double coancestry coefficients at a targeted locus with fixed effects, our Cholesky decomposition based GLMM regresses all family members phenotypic values on the square roots of the kinship and double coancestry coefficient matrices with random regression coefficients across families.

Except SOLAR, Multic, and *pedigreemm*, many other special software packages are currently available for analyzing pedigree data. For examples, VITESSE implemented the well-known Elston–Stewart's peeling algorithm for computing the likelihood of pedigrees in linkage analysis (Elston and Stewart, 1971; O'Connell

and Weeks, 1995). Mendel can run pedigree analysis for quantitative traits (Lange et al., 2013). SAGE SAGE (2012) can be used for pedigree analysis of binary traits. Based on the score statistics and generalized estimating equations, FBAT and its extension PBAT can also be used in association testing for both quantitative and binary traits (Rabinowitz and Laird, 2000; Lange and Laird, 2002; Lange et al., 2004; Laird and Lange, 2006). Nevertheless, our study provides an alternative choice for analyzing the pedigree data by using standard statistical software, which could be useful for statisticians who are not very familiar with these special genetic software packages. The standard statistical software packages often provide many options on choosing different optimization techniques to maximize the likelihood. As long as the optimization procedure converges appropriately, the standard statistical software packages can provide reliable results in most of the cases.

Sometimes we may want to perform hypothesis tests on the existence of certain variance components. For non-Bayesian methods in fitting the GLMM, it has been known that the likelihood ratio statistics (LRS) usually do not asymptotically follow the standard Chi-square distributions under the null because the zeros under the null hypothesis are located on the boundary of the parameter space for the variance components, where the standard regularity conditions no longer hold. As pointed out in Wang et al. (2011), in the hypothesis testing of a single variance component $H_0 : \sigma_A^2 = 0$, the asymptotic distribution of LRS has $0.5\chi_0^2 + 0.5\chi_1^2$ under the null. In the hypothesis testing of more than one variance components such as $H_0 : \sigma_A^2 = \sigma_D^2 = 0$, the asymptotic distribution of LRS could be a mixture of several chi-square distributions with their weights of the mixture depending on the number of family types. Therefore, directly using LRS to test for the existence of variance components could be incorrect. On the other hand, the Bayesian method can always provide appropriate estimates of the variance components as well as their variances without relying on the asymptotic results.

It should be pointed out that this Cholesky decomposition based re-formulation in fitting the GLMM has some limitations. For example, in order to synchronize the varied family sizes, the number of random effects in smaller families needs to be expanded to match that in the largest family. The Cholesky decomposition also requires that the kinship and double coancestry matrices be positive. When the kinship matrix is calculated from genome-wide genotypes, the kinship or double coancestry matrices could become singular for some of the families. One possible solution to this problem is to apply a different type of decomposition to the kinship and double coancestry matrices for these families. Note that the decomposition matrices can have a reduced number of columns as long as a good approximation to the kinship and double coancestry matrices is maintained. Finally, it appears that the computational speed in fitting the re-formulated GLMM via SAS or OpenBUGs is slow. The proposed GLMM re-formulation is probably more suitable for a refined analysis on certain targeted loci rather than a genome-wide scan for a large number of genetic markers.

Acknowledgments

This work is partially supported by the National Institute of Diabetes and Digestive and Kidney Diseases under grant number R01 DK080100. This research also uses the data resource provided by T1DGC, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Allergy and Infectious Diseases, National Human Genome Research Institute, National Institute of Child Health and Human Development, and Juvenile Diabetes Research Foundation International (JDRF), which is supported by U01 DK-062418.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2015.00120/abstract

References

Abecasis, G. R., Cardon, L. R., and Cookson, W. O. (2000). A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* 66, 279–292. doi: 10.1086/302698

Almasy, L., and Blangero, J. (1998). Multipoint quantitative trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* 62, 1198–1211. doi: 10.1086/301844

Amos, C. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* 54, 535–543.

Anjos, S., and Polychronakos, C. (2004). Mechanisms of genetic susceptibility to type i diabetes: beyond HLA. *Mol. Genet. Metab.* 81, 187–195. doi: 10.1016/j.ymgme.2003.11.010

Bates, D., Maechler, M., and Bolker, B. (2012). *Linear Mixed-Effects Models Using S4 Classes. lme4 Version 0.999999-0.*

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2011). *Bayesian Ideas and Data Analysis.* Boca Raton, FL: CRC Press; Taylor and Francis Group, LL.

de Andrade, M., Amos, C. I., and Foulkes, W. D. (1998). Segregation analysis of squamous cell carcinoma of the head and neck: evidence for a major gene determining risk. *Ann. Hum. Genet.* 62(Pt 6), 505–510. doi: 10.1017/S0003480099007204

de Andrade, M., Atkinson, E. J., Lunde, E., Amos, C. I., and Chen, J. F. (2006). *Estimating Genetic Components of Variance for Quantitative Traits in Family Studies Using the MULTIC.* Technical Report 78, Mayo Foundation.

Elston, R. C., and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Hum. Hered.* 21, 523–542. doi: 10.1159/000152448

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics, 4th Edn.* Harlow: Longman.

Feng, R., Zhou, G., Zhang, M., and Zhang, H. (2009). Analysis of twin data using SAS. *Biometrics* 65, 584–589. doi: 10.1111/j.1541-0420.2008.01098.x

Glisic, S., Klinker, M., Waukau, J., Jailwala, P., Jana, S., Basken, J., et al. (2009). Genetic association of hla dqb1 with cd4+cd25+(high) t-cell apoptosis in type 1 diabetes. *Genes. Immun.* 10, 334–340. doi: 10.1038/gene.2009.14

Harville, D., and Callanan, T. (1989). "Computational aspects of likelihood-based inference for variance component," in *Advances in Statistical Methods for Genetic Improvement of Livestock*, eds D. Gianola and K. Hammond (Berlin, Germany: Springer-Verlag), 136–176. doi: 10.1007/BF01066731

Haseman, J. K., and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* 2, 3–19. doi: 10.1007/BF01066731

Henderson, C. R. (1963). "Selection index and expected genetic advance," in *Statistical Genetics and Plant Breeding*, eds W. D. Hanson and H. F. Robison (Washington, DC: National Academy of Sciences), 141–163.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447.

Kempthorne, O. (1955). The theoretical values of correlations between relatives in random mating populations. *Genetics* 40, 153–167.

Laird, N. M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.* 7, 385–394. doi: 10.1038/nrg1839

Lange, C., and Laird, N. M. (2002). Power calculations for a general class of family-based association tests: dichotomous traits. *Am. J. Hum. Genet.* 71, 575–584. doi: 10.1086/342406

Lange, C., Blacker, D., and Laird, N. M. (2004). Family-based association tests for survival and times-to-onset analysis. *Stat. Med.* 23, 179–189. doi: 10.1002/sim.1707

Lange, K., Papp, J. C., Sinsheimer, J. S., Sripracha, R., Zhou, H., and Sobel, E. M. (2013). Mendel: the swiss army knife of genetic analysis programs. *Bioinformatics* 29, 1568–1570. doi: 10.1093/bioinformatics/btt187

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

McCulloch, C., and Searle, S. (2001). *Generalized, Linear, and Mixed Models*. New York, NY: Wiley & Sons, INC.

McGinnis, R., McLaren, W., Ranganath, V., Whittaker, P., Hunt, S., Deloukas, P., et al. (2009). Haplotype-based search for snps associated with differential type 1 diabetes risk among chromosomes carrying a specific hla drb1-dqa1-dqb1 haplotype. *Diabetes Obes. Metab.* 11 Suppl. 1, 8–16. doi: 10.1111/j.1463-1326.2008.00998.x

O'Connell, J. R., and Weeks, D. E. (1995). The vitesse algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nat. Genet.* 11, 402–408. doi: 10.1038/ng1295-402

Pinheiro, J., and Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J. Comput. Graph. Stat.* 4, 12–35.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Rabinowitz, D., and Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.* 50, 211–223. doi: 10.1159/000022918

SAGE (2012). *Statistical Analysis for Genetic Epidemiology, Release 6.3*. Available online at: http://darwin.cwru.edu.

Sham, P. C., and Purcell, S. (2001). Equivalence between haseman-elston and variance-components linkage analyses for sib pairs. *Am. J. Hum. Genet.* 68, 1527–1532. doi: 10.1086/320593

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Vazquez, A., Bates, D., Rosa, G., Gianola, D., and Weigel, K. (2010). Technical note: an r package for fitting generalized linear mixed models in animal breeding. *J. Anim. Sci.* 88, 497–504. doi: 10.2527/jas.2009-1952

Wang, X., Guo, X., He, M., and Zhang, H. (2011). Statistical inference in mixed models and analysis of twin and family data. *Biometrics* 67, 987–995. doi: 10.1111/j.1541-0420.2010.01548.x

Wang, T. (2011). On coding genotypes for genetic markers with multiple alleles in genetic association study of quantitative traits. *BMC Genet.* 12:82. doi: 10.1186/1471-2156-12-82

Wolfinger, R., and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *J. Stat. Comput. Simul.* 48, 233–243. doi: 10.1080/00949659308811554

Wolfinger, R. (1993). Laplace approximation for nonlinear mixed models. *Biometrika* 80, 791–795. doi: 10.1093/biomet/80.4.791

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.