Comparison of mixed model based approaches for correcting for population substructure with application to extreme phenotype sampling

M.Onifade, K.Burkett

University of Ottawa

Abstract

Mixed models have been useful in correcting for confounding due to population stratification and hidden relatedness in genome wide association studies. This class of models includes linear mixed models (LMM) and generalised linear mixed models (GLMM). Existing mixed model approaches to correct for population substructure have been investigated with both continuous and case/control response variables. However, they have not been investigated in the context of 'extreme phenotype sampling' (EPS), where genetic covariates are only collected on samples having extreme response variable values. In this work, we compare the performance of existing mixed model approaches (LTMLM, GMMAT, CARAT) with EPS data analysed as a binary trait. We use simulation to estimate the type 1 error of all approaches when there is confounding. Since linear mixed models are commonly used even with binary traits, we also analysed the data using a LMM (GEMMA). (Describe results here)

Keywords: Population stratification, Extreme phenotype samples, Mixed models, type 1 error.

1. Introduction

In genetic studies involving human populations, researchers are interested in determining how genetic variation contributes to diseases. Genome Wide Association Studies (GWAS), which involve genotyping a large number of individuals at hundreds of thousands of genetic markers have been useful for discovering the relationships between common variants and complex diseases. Recently, rare variants have been identified as important genetic factors contributing to the risk of disease and human traits. . Exome sequencing has been used to discover rare variation in the human genome; although costs have reduced, it remains a relatively expensive technique . Therefore study designs that are powerful at lower sample sizes are advantageous.

reference??

find a reference

An example of a cost saving design is extreme phenotype sampling: a design where genotyping or sequencing is only done on individuals in the tails of the phenotype distribution. The use of this study design can be traced to the work of [19] where it was used in mapping quantitative trait locis (QTLs) during linkage analysis. Extreme phenotype sampling has since found other uses beyond linkage analysis as other authors have adapted the study beyond linkage analysis. Still in linkage analysis, [6] used EPS and advised that the cutoffs shouldn't be more than the upper and lower 25th percentile. In association studies, [27] used extreme selection techinique to test for the association between a genetic variant and intelligence quotient and [2] assessed the association between general cognitive ability as a behavioral trait and variation in candidate genes. [36] explored the power of the study when using extreme samples compared to the whole population. In rare variant study, [11, 17, 26], EPS has been shown to have sufficient power to detect rare variants.

As with all population based genetic association designs, extreme phenotype sampling is prone to confounding by population structure or stratification. Difference in allele frequencies among members of a strata or subgroup in the population may lead to confounding if there are differences in the phenotype distribution between the subgroups. Confounding is known to leading to spurious associations and an inflation of the type 1 error, which has led to a development of methods that can correct for the effects of population stratification. The earliest methods includes the Genomic control method of Devlin et al. [7] and the STRUCTURE approach of Pritchard et al. [31]. Principal components (PC) based corrections, implemented in the program EIGENSTRAT [29] have also been successfully applied in a number of studies [29, 25]. Very recently, mixed models have become popular due to their robustness in tackling other sources of confounding in the study, in particular cryptic relatedness[30]. Over the years, an impressive number of exact and approximate LMM methods have been developed for use in genetic association studies [18, 22, 41]. Each of these methods incorporate different approaches for making LMM-based analyses feasible at the genome wide level.

However in genetic studies involving humans, the phenotype of interest is often a binary trait, which can be obtained from case-control and cohort study designs, for example.

examples of studies involving case control or cohort

Just like continuous traits, binary traits have also been analysed using linear mixed models [9, 32, 35]. These methods have used an additive polygenic model which allows for transformation of the parameters of the linear model and the logistic model. This practice has resulted in a loss of power as the mean and variance behaviour of the binary trait is ignored [14].

Earlier studies that aimed to correct for confounding rates in binary trait have relied on the method of [28] that derived a direct relationship between linear models and logistic regression. In particular, the authors justified the application of a linear mixed model to binary data by introducing a way of transforming the effect size estimates from the linear to the log-odds scale which is the natural scale by which case-control data is measured. Although widely applied to binary traits, the LMM assumes a continous phenotype where it is reasonable to assume that the trait has a constant residual variance. However, for binary traits in the presence of covariates, this assumption is not valid; therefore fitting a binary response with mixed models may fail to correct the type 1 error rate [5].

Mixed model approaches that are applicable under binary traits have also been developed. Example of an applicable method are based on the use of the liability threshold model that associates with each individual a normally distributed latent variable known as the liability. These methods have been implemented in the softwares LTMLM [12] and LEAP [37] and offers an attractive method for association testing and case-control ascertainment in case/control studies. These methods estimates the latent liabilities and tests for association using these estimates. While LTMLM tests for association using the posterior mean liabilities, LEAP uses a maximum posterior estimation. Another suitable method proposed for analysing binary traits is based on the generalised linear mixed model (GLMM). Specifically, the GMMAT uses the logistic mixed model and first fits a null model for all SNPs in the study and uses this model to compute score test statistics for testing the association between the binary traits and the genetic variant. Another binary trait association method known as CARAT uses a retrospective case-control analysis method to account for the analysis of binary traits and covariates. We desire to state here that all these methods have been examined in cases of population structure.

In this work, we aim to accomplish two goals. First, we present an overview and comparison of methods available for analysing binary traits with or without covariates adjustments using liability models and mixed models. Secondly, we investigate their performance when the binary data comes from an EPS design. We also include an LMM approach, which treats the phenotypes as if it were continous, in our comparison. Here, each of the extremes will be treated as a different category. This is motivated by the fact that mixed model based approaches for correcting confounding has not been tested in the context of the EPS. We focus on whether these methods adequately correct the type 1 error rates due to confounding under an extreme phenotype study design. Finally, we also compare the approaches on a real dataset.

add more detail when the real data phenotype is known.

2. Material and Methods

In this section, we give a brief overview of the mathematical formulations of some of the general methods being considered in this paper.

2.1. Linear Mixed Models

The linear mixed model for a vector of response values y is usually represented as a sum of fixed and random effects and an error term. Specifically, we can represent a standard LMM by the equation:

$$y = X\beta + Zb + \epsilon \tag{1}$$

such that y is a $n \times 1$ vector of response variables (continous or binary), $X_{n\times p}$ denotes the design matrix of known covariates, β is a vector of unknown regression coefficients also known as the vector of fixed effects, $Z_{n\times p}$ is a known matrix, b is a vector of random effects and ϵ is a vector of random errors. Usually in a regression analysis, b and ϵ are unobservable quantities that are assumed to be uncorrelated with a mean 0 and known variance. We represent the variances of b and respectively as $var(b) = G$ and $var(\epsilon) = R$. Hence, $b \sim \mathcal{N}(0, G)$ and $\epsilon \sim \mathcal{N}(0, R)$. The simple linear model is different from the LMM equation in (2) through the inclusion of the random effects components Zb. This enables us to specify a rich class of flexible models that have been found to be important in genetic studies. They are mainly applied in association testing between a genetic variant and a trait of interest, estimating the narrow sense heritability [1], correcting for confounding [23] and phenotype prediction [1]. Fitting LMMs involves evaluating the random effects known as the variance components. This measures the correlation between individuals.

In using LMMs for genetic analysis, the confounding effect is fit as a fixed effect while the random effects is represented as a genetic relationship matrix (GRM). This represents the pairwise genetic similarity between pairs of individuals in the study. The equation is given by:

$$y = X\beta + Zb + \mu + \epsilon \tag{2}$$

Here, we represent y as the vector of phenotype values that is assumed continous, X is the genetic variant being studied, β is the genetic effects, Z is a matrix of covariate values and b is the covariate effects. The first two terms of (2) corresponds to the fixed effects and μ is used to represent the effect of population structure in the data. Using the expressions of the mean and variance, we can then represent the distribution of y as:

$$\begin{aligned} E(y) = X\beta + Zb \ \ , \ \mathrm{Var}(y) = \sigma^2 A + \sigma^2 I \\ y = \mathcal{N}(X\beta + Zb, \sigma^2 A + \sigma^2 I) \end{aligned} \tag{3}$$

We can infer from (3) that the matrix μ imposes a sort of covariance structure on y in the form of A and this forms the basis of using LMM to correct for confounding in GWAS. In order to carry out mixed model analysis in GWAS, there is need for large sample sizes in order to achieve sufficient statistical power. Unfortunately, with increase in sample sizes there is the burden of computational complexity that increases cubicly with the number of individuals [40] in the model. This has motivated several approximate LMM methods designed to increase the speed of LMM computations and inturn make large scale GWAS feasible.

2.1.1. Generalised Linear Mixed Models

Given the vector of random effects b and the independent responses $y_1 \dots y_n$, we define the generalised linear mixed model as the conditional distribution of y_i given b using the exponential family of distributions. The probability density function $f_i(y_i|b)$ is given as:

$$f_i(y_i|b) = \exp\{\frac{y_i\varphi - b^*(\varphi)}{a_i(\phi)} + c_i(y_i, \phi)\} \tag{4}$$

where $b^*(.), a_i(.), c_i(.,.)$ are known functions, ϕ is a dispersion parameter which may or may not be known, φ is a quantity which is associated with the conditional mean $\mu_i = E(y_i|b)$, which is also associated with a linear predictor $\eta = x_i\beta + z_ib$. x_i and z_i are known vectors and β is a vector of unknown parameters i.e the fixed effects. Since the distribution of y is not normal, the mean μ_i is related to the linear predictors via a link function $g(.)$ such that

$$g(\mu_i) = \eta_i.$$

Unlike linear regression models where the variance of the observation is a constant, the variance of b depends on a vector of unknown variance components i.e $b \sim \mathcal{N}(0, G)$ such that the covariance matrix G depends on a vector θ of unknown variance components.

As a special case of the GLMM, we consider the mixed logistic model defined for binary responses $y_1 \ldots y_n$ which are conditionally independent bernoulli and $p_i = P(y_i = 1|b)$, then

$$\text{logit}(p_i) = x_i\beta + Z_ib$$

is the logistic mixed model, and x_i, z_i are as defined above. The link function here is canonical given by $g(\mu) = \text{logit}(\mu)$ and the dispersion parameter $\phi = 1$. In obtaining the parameters of estimation in a GLMM model, the traditional methods of maximum likelihood estimation and restricted maximum likelihood are not of great use here. This is because the likelihood function for a full glmm model with random effects usually involves high dimensional integrals with no closed form expressions, hence the need for specialized methods. A common approach has been to use numerical approaches like the Laplace transforms and penalized likelihood methods. The laplace approximation uses approximate integrals to find a gaussian approximation to the conditional distribution of a set of variables [16]. The parameters of the GLMM can now be obtained using traditional or restricted likelihood by considering the Laplace approximations as the true likelihood. Under a more general framework, the laplace transforms have been used in another method known as the Penalized Quasi-likelihood Estimation (PQL). The use of PQL was proposed by Breslow and Clayton (1993) has been adjudged the most popular among the approximations to the likelihood function. PQL approximates the high dimensional integrals found in GLMMs with the laplace approximation such that the approximated likelihood function is a normal distribution.

2.2. Liability Threshold Models

The liability threshold model (LTM) associates with every individual i in a population a latent variable known as the liability whose scale is regarded as arbitrary but can be assumed normal with a mean 0 and variance 1. We define a variable t, known as the threshold for a particular trait as the point on the scale of liability above which all individuals are affected and below which all are normal. We can regard these two divisions as cases and controls respectively. Hence the distribution can be regarded as:

$$Pr(Y = 1|Z) = \begin{cases} 1 & \text{if } Z > t \\ \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where Y is the vector of liability values and Z is the normal cumulative distribution function. In order to generate the classical liability threshold model, the relationship between observations on the observed risk scale and liabilities on the unobserved continous scale is modeled using a probit transformation [10, 20] and can be written as:

$$I = \mu 1_N + g + \epsilon \tag{6}$$

where I is a vector of the liability phenotypes such that $I \sim \mathcal{N}(0,1)$, g is the vector of random additive genetic effects on the liability scale with distribution $\mathcal{N}(0,\sigma_g^2)$ and ϵ is the vector of residual error [20]. Since the total phenotypic variance for the liability is equal to 1, the heritability (liability) defined as the proportion of total variance that is due to genetic factors is given as $h_L^2 = \sigma_g^2$. Liability threshold models make use of the available disease prevalence information in the data and this is given by the expected value of the phenotype values i.e $K = E(Y)$. The liability threshold models have been used in case-control studies to correct the associated loss in power as a result of acsertainment bias [12].

2.3. Extreme Phenotype Sampling (EPS)

The term selective or trait dependent sampling is used to denote genotyping those individuals whose phenotypic values are in the extremes of the phenotype distribution. This idea was motivated by the fact that individuals in the extremes are more likely to provide more linkage information than others and generally, these individuals are those whose genotypes can be clearly inferred from their phenotypes (source??). Although not a popular population sampling design as case control or cohort studies, sampling from the extremes is not a new design. It was first used by Lander and Botstein [19] in animal breeding studies to map QTLs. Subsequently, it's use have been explored both in GWAS (sources) and candidate gene association studies. Recently, it has been proposed as a cost effective design compared to exome sequencing in the detection of rare variants Barnett et al. [3] with high power compared to random sampling Emond et al. [8]. Peloso et al. [26] have used EPS design in detecting rare variants and compared the association in extreme samples with a population based random sample. In extreme phenotype sampling, the selection depends on the phenotype hence standard statistical methods are not applicable. Perhaps the most common method involves treating the extreme groups as a binary trait and applying common methods of assessing associations like the chi-square tests and logistic regression. Although valid, Lin et al. [21] stated that these methods are not optimal as the initial continous trait values are ignored [13]. Other methods of analysing EPS samples includes likelihood based methods of Huang and Lin [13] and Lin et al. [21]. Despite the wide use of EPS design in association testing and rare-variant analysis, the effects of population stratification have not been extensively investigated. Panarella and Burkett (2019) investigated this concern using principal components analysis (PCA).

3. Comparing some existing Approaches for Binary data

In human genetic studies, the phenotype of interest is a binary trait (disease status) obtained from case control sampling or cohort studies. Case-control studies samples diseased individuals (cases) from a study and a comparable group of individuals from the same population who are free of the disease to serve as controls. In such case-control kind of studies, we are interested in the problem of association testing with a known causal variant while accounting for the effects of population structure and/or covariates [15]. The use of mixed model based approaches have been extensively explored in correcting the effects of population stratification and other unknown sources of population structure. These methods have all being applied to quantitative data as application to binary traits have resulted in largely inflated type 1 error rates and loss in power. This is because the assumption of a constant residual variance for all the individuals in the sample might not hold for binary traits in the presence of covariates [5]. Furthermore, a correct analysis of binary traits using mixed model approaches should include methods that are able to account for the selective sampling. Early approaches that have used mixed models in analysing binary traits have used the method of Pirinen et al. [28] that assumes that in the absence of population stratification, linear models can be approximated by a logistic regression. Recent methods for the analysis of binary traits are methods based on the liability threshold models. Liability models estimates the model parameters for each associated genetic variants while also accounting for the case control ascertainment bias. Examples of such methods are the LTMLM, LTSOFT and CARAT. ROADTRIPS is a binary-trait association testing method that accounts for the population structure in the data using association statistics that have been adapted to cases where the population structure is known [34]. However, it is not suited for association testing between a single causal variant and the trait of interest. Very recently, the use of generalised linear mixed models have been explored in binary-trait association testing. GLMMs leverages the advantages of generalised linear models (GLMs) and linear mixed models (LMM) so that we are able to analyse binary trait data without the unrealistic assumption that the covariates have a constant residual variance. This was implemented as the logistic mixed model in a tool known as GMMAT [5].

Similar to the methods that have been primarily designed for analysing quantitative data, quite a number of these binary methods have all recently appeared in research and the differences or similarities between these methods have not been clearly elucidated. Here, we undertake a review of these methods suitable for association testing between a binary-trait and a genetic variant of interest. These methods can be classified broadly into three: (i) approaches using liability threshold models (LTMLM, LEAP, LTSCORE) , (ii) mixed model approaches (CARAT, GMMAT) and (iii) association statistics that have incorporated cases of completely unknown or partially unknown population structure (ROADTRIPS, GCAT (Song et al.)).

- Liability threshold models (LTM) have been proposed as a valid approach to tackle the effects of ascertainment in case-control studies. In ascertained case-control studies, cases are usually oversampled relative to the disease prevalence leading to loss of power when linear mixed models are used. Weissbrod et al. [37] stated that the loss of power was due to the violation of several model assumptions one of which included the dependence between the candidate SNPs and the SNPs used to estimate kinship. Although LMMs are able to resolve the effects of confounding in genetic association studies, their use for binary traits association leads to a different form of confounding. The population stratification in ascertained case control studies is as a result of the unequal case-control ratios from different sampling schemes which results in unequal variances of the binary traits [5]. Using the liability threshold models involves computing liability scores for each individual to be used in testing the association between a phenotype of interest and genetic variants. In this manner, the LTM is able to directly represent the case-control phenotype while taking into account the ascertainment bias [37]. Although an attractive method, the use of liability threshold models is computationally expensive thus rendering whole genome association tests infeasible. To harness the attractive nature of LTM, a number of approaches have been developed. One of the earliest approaches includes the LTSCORE method of Zaitlen and Kraft [38]. By introducing external prevalence data into the liability threshold model, LTSCORE is able to account for the study design and disease prevalence and at the same time test for association with previously identified causal SNPS using a linear regression. A similar method by Zaitlen et al. [39] computes the liability estimates and thereafter tests for the association using the EIGENSTRAT method [29]. Another liability threshold method known as LEAP (Liability Estimator as a Phenotype) Weissbrod et al. [37] computes liability estimates conditional on the phenotypes, genotypes and disease prevalence on the entire genome and tests for association using a LMM method. Unlike the methods of Zaitlen and Kraft [38] and Zaitlen et al. [39], LEAP computes the liabilities using the whole genome and tests for association with the maximum a posteriori (MAP) estimate. Similar to LEAP is the recent method by Hayeck et al. [12] known as the liability threshold mixed linear model (LTMLM). LTMLM computes the posterior mean liabilities of all the individuals under a liability threshold model and tests for association using a chi-square score statistic. The posterior mean liabilities of the individuals are computed dependent on the individual's case control status, every other individuals' case-control status and the genetic relationship matrix. Its performance was benchmarked across other existing mixed model methods and was found to have a well-controlled false positive rate.

- Generalised linear mixed models (GLMMs) have been used in genetic studies involving binary traits due to their ability to take into account the erroneous assumption made in the use of LMM for binary traits. These class of models can be viewed either as an extension of generalised linear models [24] or as an extension of LMM. When viewed in the former way, GLMMs are able to model the distribution of the response variable as a function of non-normally distributed variables. In the other viewpoint, GLMMs can be seen as an extension of LMMs when covariates have been included in the model. The inclusion of covariates implies that the assumption of constant variance made for the random effects in the linear mixed model no longer holds and there is need for an alternative construct to help model the covariate variability. Hence, for binary traits in the presence of covariates, the use of LMMs to test for the association between a causal SNP and a trait of interest will lead to type 1 errors. In the use of GLMMs to model non-normally distributed response variables, the logistic mixed models: a special case of GLMMs have been used to analyse binary traits and have been found to offer better correction of the type 1 error rates compared to logistic regression and ordinary linear models. Despite this, the logistic mixed models have not been seen in widespread use in GWAS due to the computational complexity involved in fitting logistic mixed models for large scale genetic variants. Chen et al. [5] proposed a GLMM tool implementing a logistic mixed model known as GMMAT for large scale GWAS. GMMAT firsts fits a null logistic mixed model including as fixed effects only the covariates while the random effects are used to account for residual population stratification and relatedness. This fitted null model which is the same for all genetic variants in the study is then used in the test for association between a genetic variant and phenotype via a score test. The use of just one null model for testing all genetic variants greatly simplifies the model compared to fitting the full logistic mixed model for a large GWAS. Another mixed model method is known as CARAT (Case Control retrospective association test). It is a binary traits testing approach which accounts for relevant covariate information and control for population structure. The response variable is modeled by using a mixed effects quasi likelihood approach which exploits the binary nature of the trait. Similar to the GMMAT approach, CARAT does not require the knowledge of disease prevalence. For large scale genetic association studies, the use of estimating equations and a score test is used. In assessing the score test statistic, under the null model, the genotypes are regarded as random conditional on the phenotype and covariates [15].

- ROADTRIPS - a binary trait association testing tool proposed by Thornton and McPeek [34]. It can be regarded as an extension of a collection of association statistics that have been used in traditional case control testing in the presence of known structure to the context of unknown or partially known structure. These include the corrected χ^2, armitage trend tests, W_{QLS} [4] and M_{QLS} [33] tests.

size=scriptsize

Criteria	LTMLM	LEAP	GMMAT	CARAT	ROADTRIPS
Type of model	Liability threshold model with focus on ascertained case-control studies assuming known disease prevalence.	Liability threshold model with focus on ascertained case-control studies assuming known prevalence	Mixed model (retrospective and prospective)	Retrospective mixed model	Association Statistics
Model Approach	Liabilities computed using the whole genome	Liabilities computed using the whole genome	Logistic Mixed model via Penalised Quasi likelihood.	Estimating equation approach	-
GRM	Whole genome excluding the candidate SNP(s)	Exclude SNPs involved in liabilities estimation.	Whole genome excluding the candidate SNP(s)	Total number of SNPs.	
Test for association	Posterior mean liabilities in a chi-square score test framework	Liability estimates tested for association via a standard linear mixed model.	Score test or wald tests.	Quasi-likelihood in a score test framework	-
Covariates	Program does not allow for covariate adjustment	-	Allows for covariate adjustment	Allows for covariate adjustment	No adjustment for covariates or polygenic additive effects

Table showing main features of some methods suitable for analysing binary response variable for the purpose of comparison.

4. Application of LTMLM and GMMAT to Extreme Phenotype Sampling design.

In analysing samples from the extremes, the most common method has been to treat the two extreme categories as binary traits [8] and use methods suitable for categorical data to assess the association between a causal variant and the phenotype of interest. With the advent of more suitable methods for analysing binary traits such that the sampling is taken into account, it is worthwhile to explore the performance of some of these identified methods from the literature in correcting the type 1 error rate when we assume that the data comes from the extremes. To this end, we treat the two extremes which make up the whole extreme population as either cases or controls in a typical prospective case-control design. These methods are the liability threshold mixed linear model (LTMLM) and the generalised mixed model association test which incorporates the liability threshold model and the generalised linear mixed model respectively. We used these two methods on simulated extreme phenotype data samples to test for the association between a causal SNP and the phenotype of interest. Although these are not the only existing methods, we choose these methods as a representative of the broad categories of the methods we discovered. Also, we used methods which are capable of performing candidate gene study and not GWAS. We didn't use ROADTRIPS in out simulation experiments as the ROADTRIPS method was not suitable to carry out an association between a single candidate gene and the simulated phenotype values. In all these methods used, we were mainly interested in assessing how each method worked in correcting the type 1 error rate as a result of population stratification. Here, we give a little insight into the LTMLM and GMMAT methods and describe the simulation study in the section that follows.

1. LTMLM (Liability Threshold Mixed Linear model) [12]): The LTMLM association statistic and software was proposed to control the false positive rate in ascertained case-control studies. The method was compared with existing mixed model methods and was found to have a well controlled false-positive rate for diseases with a low prevalence rate in the population. The LTMLM association statistic is given as a chi-square score statistic computed from the posterior mean liabilities under the liability threshold model. Here, for each individual in the study the posterior mean liability is conditional on the case-control status of that individual, the case-control status of every other individual in the study i.e disease prevalence and on the genetic relationship matrix computed from all the SNPs excluding the candidate SNP in the study. For the multivariate model considered, the PMLs are estimated using a multivariate Gibbs sampler: a MCMC algorithm for multivariate random sampling. The phenotypic covariance matrix on the liability scale is based on the GRM and the heritability estimate computed using the Haseman-Elton regression on the case-control phenotypes and then transformed to the liability scale. In order to account for the ascertainment, the liabilities and genotypes are jointly modelled using a retrospective model.
2. GMMAT (Generalised Mixed Model Association Test): The GMMAT tool for association testing firsts fits a null generalised linear mixed model that includes just the covariates and random effects to account

for population structure and other forms of confounding (cryptic or family relatedness.) Unlike LTMLM, the GMMAT tool can be adapted into testing for association both for GWAS and candidate gene studies. Association tests are usually measured using score tests for every genetic variant and wald tests to obtain the effect estimates of each genetic variant for candidate gene studies. By specifying a particular link family such as binary or gaussian, GMMAT will perform mixed model based association tests when the response variable is either categorical such as disease status or quantitative. For a candidate gene study such as we are interested in, the GMMAT method is represented as the GLMM model given in Chen et al. [5].

$$\eta_i = g(\mu_i) = X_i\alpha + G_i\beta + b_i \tag{7}$$

Since the GLMM follows the exponential family of distributions, given the random effects b_i, which are conditionally independent with the responses y_i, the mean and variance is given respectively by $E(y_i|b) = \mu_i$ and $Var(y_i|b) = \Phi a_i^{-1} v(\mu_i)$ respectively. Here Φ is the dispersion parameter defined for exponential families, $v(.)$ is the variance and a_i are known weights. All other parameters are as defined for GLMMs in previous sections. For binary traits, GMMAT uses a logistic mixed model: GLMM which assumes a bernoulli distribution for the responses and a logit link function. The GMMAT tool is implemented as an R package.

[1] Balding, D. J. and Nichols, R. A. (1995). A method for quantifying

differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12.

[2] Ball, D., Hill, L., Eley, T. C., Chorney, M. J., Chorney, K., Thompson, L. A., Detterman, D. K., Benbow, C., Lubinski, D., Owen, M., et al. (1998). Dopamine markers and general cognitive ability. *Neuroreport*, 9(2):347–349.

[3] Barnett, I. J., Lee, S., and Lin, X. (2013). Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genetic epidemiology*, 37(2):142–151.

[4] Bourgain, C., Hoffjan, S., Nicolae, R., Newman, D., Steiner, L., Walker, K., Reynolds, R., Ober, C., and McPeek, M. S. (2003). Novel case-control test in a founder population identifies p-selectin as an atopy-susceptibility locus. *The American Journal of Human Genetics*, 73(3):612–626.

[5] Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., Szpiro, A. A., Chen, W., Brehm, J. M., Celedón, J. C., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4):653–666.

[6] Darvasi, A. and Soller, M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and applied Genetics*, 85(2-3):353–359.

[7] Devlin, B., Roeder, K., and Wasserman, L. (2001). Genomic control, a

new approach to genetic-based association studies. *Theoretical population biology*, 60(3):155–166.

[8] Emond, M. J., Louie, T., Emerson, J., Zhao, W., Mathias, R. A., Knowles, M. R., Wright, F. A., Rieder, M. J., Tabor, H. K., Nickerson, D. A., et al. (2012). Exome sequencing of extreme phenotypes identifies dctn4 as a modifier of chronic pseudomonas aeruginosa infection in cystic fibrosis. *Nature genetics*, 44(8):886.

[9] Fakiola, M., Strange, A., Cordell, H. J., Miller, E. N., Pirinen, M., Su, Z., Mishra, A., Mehrotra, S., Monteiro, G. R., Band, G., et al. (2013). Common variants in the hla-drb1–hla-dqa1 hla class ii region are associated with susceptibility to visceral leishmaniasis. *Nature genetics*, 45(2):208.

[10] Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics*, 29(1):51–76.

[11] Guey, L. T., Kravic, J., Melander, O., Burtt, N. P., Laramie, J. M., Lyssenko, V., Jonsson, A., Lindholm, E., Tuomi, T., Isomaa, B., et al. (2011). Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. *Genetic epidemiology*, 35(4):236–246.

[12] Hayeck, T. J., Zaitlen, N. A., Loh, P.-R., Vilhjalmsson, B., Pollack, S., Gusev, A., Yang, J., Chen, G.-B., Goddard, M. E., Visscher, P. M., Patterson, N., and Price, A. L. (2015). Mixed Model with Correction

for Case-Control Ascertainment Increases Association Power. *American Journal of Human Genetics*, 96(5):720–730.

[13] Huang, B. and Lin, D. Y. (2007). Efficient association mapping of quantitative trait loci with selective genotyping. *The American Journal of Human Genetics*, 80(3):567–576.

[14] Jiang, D., Mbatchou, J., and McPeek, M. S. (2015). Retrospective association analysis of binary traits: overcoming some limitations of the additive polygenic model. *Human heredity*, 80(4):187–195.

[15] Jiang, D., Zhong, S., and McPeek, M. S. (2016). Retrospective binary-trait association test elucidates genetic architecture of crohn disease. *The American Journal of Human Genetics*, 98(2):243–255.

[16] Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

[17] Kang, G., Lin, D., Hakonarson, H., and Chen, J. (2012). Two-stage extreme phenotype sequencing design for discovering and testing common and rare genetic variants: efficiency and power. *Human heredity*, 73(3):139–147.

[18] Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

[19] Lander, E. S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199.

[20] Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305.

[21] Lin, D.-Y., Zeng, D., and Tang, Z.-Z. (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences*, 110(30):12247–12252.

[22] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833.

[23] Lippert, C., Quon, G., Kang, E. Y., Kadie, C. M., Listgarten, J., and Heckerman, D. (2013). The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific reports*, 3:1815.

[24] McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

[25] Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190.

[26] Peloso, G. M., Rader, D. J., Gabriel, S., Kathiresan, S., Daly, M. J., and Neale, B. M. (2016). Phenotypic extremes in rare variant study designs. *European Journal of Human Genetics*, 24(6):924–930.

[27] Petrill, S. A., Plomin, R., McClearn, G. E., Smith, D. L., Vignetti, S., Chorney, M. J., Chorney, K., Thompson, L. A., Detterman, D. K., Benbow, C., et al. (1997). No association between general cognitive ability

and the a1 allele of the d2 dopamine receptor gene. *Behavior Genetics*, 27(1):29–31.

[28] Pirinen, M., Donnelly, P., Spencer, C. C., et al. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics*, 7(1):369–390.

[29] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904.

[30] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459.

[31] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

[32] Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C., Patsopoulos, N. A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S. E., et al. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214.

[33] Thornton, T. and McPeek, M. S. (2007). Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *The American Journal of Human Genetics*, 81(2):321–337.

[34] Thornton, T. and McPeek, M. S. (2010). Roadtrips: case-control association testing with partially or completely unknown population and pedigree structure. *The American Journal of Human Genetics*, 86(2):172–184.

[35] Tsoi, L. C., Spain, S. L., Knight, J., Ellinghaus, E., Stuart, P. E., Capon, F., Ding, J., Li, Y., Tejasvi, T., Gudjonsson, J. E., et al. (2012). Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature genetics*, 44(12):1341.

[36] Van Gestel, S., Houwing-Duistermaat, J. J., Adolfsson, R., van Duijn, C. M., and Van Broeckhoven, C. (2000). Power of selective genotyping in genetic association analyses of quantitative traits. *Behavior genetics*, 30(2):141–146.

[37] Weissbrod, O., Lippert, C., Geiger, D., and Heckerman, D. (2015). Accurate liability estimation improves power in ascertained case-control studies. *Nature methods*, 12(4):332.

[38] Zaitlen, N. and Kraft, P. (2012). Heritability in the genome-wide association era. *Human genetics*, 131(10):1655–1664.

[39] Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A. L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS genetics*, 9(5):e1003520.

[40] Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., et al. (2010).

Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4):355.

[41] Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821.