Mixed modeling with whole genome data

Jing Hua Zhao*, Jian'an Luan

MRC Epidemiology Unit & Institute of Metabolic Science

April 20, 2012

Abstract

Objective: We consider the need for a modeling framework for related individuals and various sources of variations. The relationships could either be among relatives in families or among unrelated individuals in a general population with cryptic relatedness; both could be refined or derived with whole genome data. As with variations they can include oliogogenes, polygenes, single nucleotide polymorphism (SNP) and covariates. **Methods**: We describe mixed models as a coherent theoretical framework to accommodate correlations for various types of outcomes in relation to many sources of variations. The framework also extends to consortium meta-analysis involving both population-based and family-based studies. **Results**: Through examples we show that the framework can be furnished with general statistical packages whose great advantage lies in simplicity and flexibility to study both genetic and environmental effects. Areas which require further work are also indicated. **Conclusion**: Mixed models will play an important role in practical analysis of data on both families and unrelated individuals when whole genome information is available.

Keywords: genomewide association study, mixed models

*Address for correspondence: MRC Epidemiology Unit & Institute of Metabolic Science, Addrenbrooke's Hospital Box 285, Hill's Road, Cambridge CB2 0QQ, United Kingdom, Tel: +44(0)1223-769165, e-mail:jinghua.zhao@mrc-epid.cam.ac.uk

1 Introduction

Genomewide association studies (GWASs) have successfully identified many genetic variants consistently associated with human diseases or other traits. Both unrelated individuals in a population or related individuals in families have been involved in such studies. There is a variety of issues which merit further consideration.

Our concern here is on correlations among individuals, which are "the central piece of information" [1] in detection and characterization of gene-trait association. Consideration of these correlations has traditionally limited to family data whose critical role in genetic epidemiological study ranges from familial aggregation, segregation, linkage to association [2], and special attention is required in the analysis compared to unrelated individuals from a population. Correlations arise naturally among relatives but can be relevant to population-based study as well given that relatedness can also be established among unrelated individuals based on whole-genome data in GWASs [3]. The correlations are linked to a long attempt to model influence of multiple genes on a specific phenotype. Specifically, Fisher [4] assumed that a quantitative trait results from many genes with variable small to moderate effects. Concrete evidence of multiple genetic influence has been revealed by recent waves of GWASs on height [5], blood pressure [6], lipids [7], obesity [8], etc., leading to the note in [9]. Gene-environment interaction and common environment can be considered similarly.

There is a relatively small literature in human genetics to iterate mixed models to account for heterogeneity among groups of individuals compared to the general statistics literature where genetic applications been acknowledged [10, pp190-192][11, pp4864-4871]. This is likely due to the complexity with a generic implementation. We therefore conduct a survey of the framework with exploration of general software environments. As will be seen below, it readily applies to human genetics when correlations within these groups are explicitly modeled. The familiar form accommodate effects of major or oligogenes, polygenes, common environment, and unique environment, which collectively contribute to variance of the trait and known as "variance component models" [12, 13]. For instance, individuals' body weight (kg) divided by height^2(m^2), referred as body mass index (BMI, kg/m^2) and commonly used as surrogate of obesity, varies with the broad heritable background of individuals (polygenes), sex, age, family membership, susceptible genes such as FTO [14] (which has a major effect serving an example of oliogene), where

sex and age can be considered as fixed effects while variability attributable to (expected) correlation between members of family as with FTO are random effects [15, 16, 17, 18]. The flexibility of such a framework may be missing in various computer programs[1]. As for outcome of interest, it is usually quantitative or binary traits, with [19] as an exception. The implementation we consider will be SAS[2] [20] and R[3] [21] with a Cox model counterpart [22]. A note on Bayesian counterpart is also ready [23, 24, 25]; especially for linkage [1], association [26] and implementation in **Morgan**. To save space, we consequently omit reference to programs when they are available from the lists given here.

We attempt to connect various models in our survey paying special attention to their use in data analysis. We show that with generic facilities as available from R, we can accommodate additional outcomes such as count, survival, as well as account for information such as identity-by-descent (IBD) or common environment. We will illustrate with the family data available to genetic analysis workshops (GAWs)[4] 16 and 17. We will also discuss the implications of whole genome data availability via connection to earlier literature.

2 Models

As will soon become clear, the framework is essentially motivated from the usual general linear model (GLM) or generalized linear mixed model (GLMM) allowing for correlated random effects, including the Cox regression model. We will briefly describe the models as an analogy between GLM and GLMM but will not go into details of their estimation procedures, as both are widely available.

2.1 GLM

We start from the usual GLM disregarding familial correlations. Let the phenotypes of n individuals in a family be $(y_1, \ldots, y_n)$, its distribution is

[1] See `http://linkage.rockefeller.edu`

[2] See `http://www.sas.com`

[3] See `http://www.r-project.org`

[4] See `http://www.gaworkshop.org`

exponential

$$f(y_i, \theta_i, \varphi) \;=\; \exp\left[\frac{y_i\theta_i - b_i(\theta_i)}{\varphi} + c(y_i, \varphi)\right] \tag{1}$$

where $b(.)$ and $c(.)$ are known functions, φ a scale or dispersion parameter. Furthermore, let $E[y_i] = \mu_i$ and let this be connected to a linear predictor using link function $g(.)$ by $\eta_i = g(\mu_i) = X_i\beta$ where X_i is a vector of covariates and β the regression coefficient(s). For simplicity, only canonical link is used so that $\theta_i = \mu_i$. It can be shown [27] that the expectation $E(y_i) = \mu_i = b'(\theta_i)$ and variance $V(y_i) = \varphi b''(\theta_i)$. Some special cases as with their properties are well-recognized [28], for which models involving continuous and binary outcomes are most common:

Normal: $y_i \sim N(\mu_i, \sigma_i^2)$, we have $\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2$, $\varphi = \sigma_i^2$, $b'(\theta_i) = \theta_i$, $\varphi b''(\theta_i) = \sigma_i^2$ and an identity link.

Binomial: $y_i \sim Binom(n, \mu_i)$, $\theta(\mu_i) = \ln(\mu_i/(1-\mu_i))$, $b(\theta_i) = \ln(1 + \exp(\theta_i))$, $\varphi = 1/n$, $b'(\theta_i) = \exp(\theta_i)/(1+\exp(\theta_i))$, $\varphi b''(\theta_i) = \mu_i(1-\mu_i)/n$, and a logit link $g(\mu_i) = \ln(\mu_i/(1-\mu_i))$.

Analysis of censored survival data can be molded into the framework [29]. Let t_i denote the event time, c_i the censoring time and $\delta_i = I(t_i \leq c_i)$ the event indicator for unit $i, i = 1, \ldots, n$, the basic Cox model with vector of explanatory variables X_i is specified via a hazard function $\lambda_i(t) = \lambda_0(t)\exp(X_i\beta)$, where $\lambda_0(t)$ is the baseline hazard function. The partial likelihood (PL) for the standard Cox model can be expressed in (2).

$$PL(\beta) \;=\; \prod_{i=1}^{n}\left[\frac{\exp(X_i\beta)}{\sum_{j\in R(t_i)}\exp(X_j\beta)}\right]^{\delta_i} \tag{2}$$

where n failure times have been ordered such that $t_1 < \ldots < t_n$ and $R(t_i)$ is the "risk set", the number of cases that are at risk of experiencing an event at time t_i.

Although GLM lays the foundation in many applications of general statistics, it largely serves a motivating role for models that are capable to account for familial correlations. As shown below, this is achieved with introduction of (correlated) random effects as in GLMM but it is also linked with other models.

2.2 GLMM

We now consider model involving individual i, $i = 1, \ldots, N$ where N is the total number of individuals in our sample.

Polygene

Let P denote the polygene representing independent genes of small effect, which follows a multivariate normal distribution with covariance matrix

$$g(\mu_i) \quad = \quad X_i\beta + P_i \tag{3}$$

The likelihood for all relatives is furnished with specification of the distribution of $P = (P_1, \ldots, P_N)$ with covariance

$$\Sigma_P = 2\Phi\sigma_P^2 \tag{4}$$

where $\Phi \equiv \{\phi_{ij}\}_{n\times n}$ and ϕ_{ij} is the kinship coefficient, defined such that given two individuals, one with genes (g_i, g_j) and the other with genes (g_k, g_l), the quantity is $\frac{1}{4}(P(g_i \equiv g_k) + P(g_i \equiv g_l) + P(g_j \equiv g_k) + P(g_j \equiv g_l))$ where $\equiv$ represents probability that two genes sampled at random from each individual are IBD. The kinship coefficients for MZ twins, DZ twins/full-sibs, parent-offspring, half-sibs and unrelated individuals are 0.5, 0.25, 0.25, 0.125 and 0, respectively.

The likelihood function for model (3) has the following form,

$$L(y_1, \ldots, y_N) \quad = \quad \int L(y|P)L(P)dP \tag{5}$$

where $L(y|P) = \prod_{i=1}^{N} f(y_i|P)$ and $L(P) = \left(\sqrt{2\pi|\Sigma_P|}\right)^{-1} \exp\left[-P'\Sigma_P^{-1}P/2\right]$ only involves with random effects, noting that it is assumed that given random effects in the model, the phenotypic values among n relatives are independent and that the parameters of interest in (4) are the variances involving polygene (σ_P^2). Regarding the statistical inference of random effects, since the parameter under the null hypothesis is on the boundary of the parameter space, the test for a specific $\sigma_k^2 = 0$, likelihood ratio statistic testing for the hypothesis that $H_0 : \sigma_P^2 = 0$ vs $H_A : \sigma_P^2 > 0$ for is referred to a $0.5\chi_0^2 + 0.5\chi_1^2$ distribution or a score statistic as outlined in [19][11, p2961].

Oligogene

Suppose that a major gene M is also involved, independently and normally distributed with mean 0 and variances σ^2_M, then the covariance matrix has the form

$$\Sigma_M = \sigma^2_M \Pi \tag{6}$$

where $\Pi \equiv \{\pi_{ij}\}_{N\times N}$ in which π_{ij} is the proportion of alleles shared (IBD) at the major gene between relatives i and j which can be estimated from a multipoint data. so that it acts additively with polygene P, the likelihood is furnished with an extended covariance

$$\Sigma_{M,P} = \Sigma_M + \Sigma_P \tag{7}$$

For a test of a strictly positive variance associated with a polygene versus polygene and an oligogene, the log likelihood ratio test statistic is referred to $0.5\chi^2_1 + 0.5\chi^2_2$ [30].

Multiple random effects

The framework in (3) includes the common distributions such as normal, gamma, binomial and Poisson as special cases. For simplicity, we consider a quantitative trait, whose probability density function is normal and a statistical model is as follows

$$y \;\; = \;\; X\beta + U + \epsilon \tag{8}$$

and $U \sim N(0, \Sigma)$, $\epsilon \sim N(0, \sigma^2)$, $Cov(U, \epsilon) = 0$. The expression of Σ^{-1} relative to the precision $1/\sigma^2$ of ϵ as a Cholesky factorization $\Delta'\Delta$, i.e., $\Sigma^{-1}/(1/\sigma^2) = \Delta'\Delta$ led to the term *relative precision factor* for Δ [31]. Note that the partition of effects as being fixed and random (H_A: genetic effect) can be compared to a sporadic model (H_0: no genetic effect) $y = X_1\beta_1 + X_2\beta_2 + e$ where both β_1 and β_2 are fixed effects, the involvement of Σ or more specifically Σ^{-1} as a "ridge factor" creates shrinkage in the random effects solutions to the normal equations, i.e., "regression towards the mean".

We will see an example from the GAW17 data below that a quantitative trait Q1 is influenced by polygenic background and specific gene *VEGFC* as captured by kinship or relationship matrix and IBD matrix, respectively. This prompts the need to consider multiple random effects. We therefore

pursue (8) further. As in [32], write $y = X\beta + Z_1 a_1 + \ldots + Z_k a_k + \epsilon$ with the usual assumption that y being $N \times 1$ vector of observations, X an $N \times p$ known matrix, not necessarily of full column rank, β a vector of fixed effects, Z_i a known $N \times r_i$ matrix of rank r_i, a_i random effects with $E(a_i) = 0$, $cov(a_i) = \sigma_i^2 I_{r_i}, cov(a_i, a_j) = 0, i \neq j, cov(a_i, \epsilon) = 0, i, j = 1, \ldots, k$, ϵ an $N \times 1$ vector of errors with $E(\epsilon) = 0, cov(\epsilon) = \sigma^2 I_N$. Then $E(y) = X\beta$ and $cov(y) = \Sigma = \sigma^2 I_N + \sum_{j=1}^{k} \sigma_j^2 Z_j Z_j'$. This turns out to be critical to explore the covariance structure involving more (k) parameters $(\sigma_1^2, \ldots, \sigma_k^2)$ in form

$$\Sigma(\sigma_1^2, \ldots, \sigma_k^2) \quad = \quad \Sigma_1(\sigma_1^2) + \ldots + \Sigma_k(\sigma_k^2) \tag{9}$$

where $\Sigma_i(\sigma_i^2)$ has the form of $\sigma_i^2 H_i, i = 1 \ldots, k$ with σ_i^2 being the unknown parameter and H_i a (known) coefficient matrix. It will also hold when different variance components such as multiple major genes of interest, gene-gene, gene-environment interactions, common shared environment are to be modeled. For significance test, Case 4 in [30] serves as a general guideline.

A closely related model is the so-called **marginal or population-average model** whereby familial relationship can be specified for e, namely generalized estimating equations (GEE) [12, 33]. Given $\mu_i = E(y)$, $V_i = Var(y)$, it has the form

$$\sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right)' V_i^{-1} (y_i - \mu_i) \quad = \quad 0 \tag{10}$$

for which only link function and variance need to be specified. Parameter estimates are consistent even when variance structure is misspecified, but the ability to use (9) is an apparent advantage.

We now turn to the Cox model. First, the consideration of an unobserved family specific random effect is often termed as frailty model, such that families with a larger value of the frailty will experience the event at earlier times and most "frail" individuals will fail early[34]. Now we allow for correlated frailty and in analogy to model (3) and [22], the appropriate model with random effect U_i becomes $\lambda_i(t) = \lambda_0(t) \exp(X_i\beta + U_i)$. Assuming the parameters of interest are β and σ^2 we have

$$PL(\beta, U) \quad = \quad \prod_{i=1}^{N} \left[\frac{\exp(X_i\beta + U_i)}{\sum_{j \in R(t_i)} \exp(X_j\beta + U_i)} \right]^{\delta_i} \tag{11}$$

The so-called integrated log likelihood is derived as.

$$L \quad = \quad \int PL(\beta, U) L(U) dU \tag{12}$$

A more tractable solution is via a Laplace approximation for an approximate marginal log likelihood that can be maximized by a penalized partial likelihood (PPL) for parameters (β, σ^2), $PPL(\beta, U) = \log(PL(\beta, U)) - U^T\Sigma^{-1}U/2$, followed by a profile likelihood function involving only σ^2.

Furthermore, we can take advantage of the generic form of covariance in other types of models as well. A straightforward yet remarkably useful extension is the multivariate model. For instance, consider (8) with m phenotypes. Let $y = (y_{11}, \ldots, y_{1N}, \ldots, y_{mN})^T$ be a vector of m multivariate phenotypes for N individuals. Let β a vector of dimension mp of the regression coefficients for the p covariates including a vector of 1's corresponding to the overall mean, $X = I_m \bigotimes X_{N,p}$ an $mN \times mp$ known matrix of covariate values. An analogy to (7,8) leads to the variance-covariance matrix of the m phenotypes with dimension $mN \times mN$ is

$$\Sigma \quad = \quad A \bigotimes \Pi + B \bigotimes R + C \bigotimes I \tag{13}$$

where R is the $N \times N$ matrix of the coefficients of relationship, Π an $N \times N$ matrix of estimated proportion of alleles IBD, and A, B, C are oligogenic, polygenic and residual variance-covariance matrices each with dimension $m \times m$.

2.3 Meta-analysis

One indispensable element in current GWASs is meta-analysis, typically involving findings from both unrelated individuals in a population and those from family data. While we have seen that mixed models are appropriate for a variety of traits in family-based association studies, broadly models for meta-analysis also fall into the same framework as described above. One can imagine a meta-analysis involving individual participant data (IPD). A good summary of approaches for IPD meta-analysis is available [35]:

In the two-step approach, the individual participant data are first analysed in each separate study independently by using a statistical method appropriate for the type of data being analysed; for example, a linear regression model might be fitted for continuous responses such as blood pressure, or Cox regression might be applied for time to event data. (This step produces aggregate data for each study including effect estimate and its standard error). *These data are then synthesised in the second step using a suitable model for meta-analysis of aggregate data, such as one that weights studies by the inverse of*

the variance while assuming fixed or random effects across studies. In the one-step approach, the individual participant data from all studies are modelled simultaneously while accounting for the clustering of participants within studies. This approach again requires a model specific to the type of data being synthesised, alongside appropriate specification of the assumptions of the meta-analysis (for example, of fixed or random effects across studies).

The two-step approach is the usual one used in various GWAS consortia while a one-step approach for all studies in our context could involve unrelated population-based samples and family data in the meta-model as long as the correlation structure is appropriately specified. The pracitcality of both approaches has been illustrated in the literature [36, 37] but in view of the complexity involving in such a framework, and the practical difficulty that a researcher may not have access to individual data from all studies, for now we refrain ourselves from such a consideration for now but remain focus on family data as illustrated with both simulated and real data.

2.4 Related results and implementations

There have been concerns in the literature regarding large number of units each with bounded sizes [38] and a large number of random effects [39]. In our context large number of families each with bounded members, consistent estimate of the random effect is difficult to obtain though fixed effects and variance components will be consistent. However, Type I error rate and power have been explored before [19, 22, 26, 40], so we will be more on specific examples.

Instead of using purposely written programs, we chose to use *R*, for its wide availability and many other features [41] and in particular procedures to fit models described earlier are to a great extent available, including generic procedures from **nlme**, **lme4** and **gee**, among others, but package designed for family data is **pedigreemm** with *lmekin* for linear mixed models available from **coxme**. We will also compare them to *SAS*, due to its ability to deal with large data, and great flexibility in model specification.

3 Examples

We consider two examples from GAWs 17 and 16, which involve simulated and real data widely available and allow for a lot of experiments to be done.

3.1 GAW17 Data

Data distributed by GAW17 were based on a collection of unrelated individuals and their genotypes were generated from the 1000 Genomes Project[5], from which a sample of 697 individuals in 8 extended families and their genotypes and phenotypes were available. A total of 202 founders in the family data set were chosen at random from the set of unrelated individuals. Replicates of the trait were generated 200 times but the simulated genotypes remain constant over replicates. The traits made available were Q1, Q2, Q4, and AFFECTED (coded 0=no 1=yes) with covariates AGE and SMOKE. The variables describing family structures were ID, FA, MO, SEX (1=men, 2=women). Fully informative IBD information was available for 3205 genes.

We chose to examine traits Q1, Q2 and AFFECTED as representatives of quantitative and qualitative traits. According to [42], vascular endothelial growth factor (*VEGF*) pathway was enriched and here vascular endothelial growth factor C (*VEGFC*[6]) was chosen as a causal variant associated with Q1 but not Q2. Q1 also increased with age, and the fact that AFFECTED is a function of Q1 offers the possibility to furnish a logistic regression model and explore age at onset via a Cox model. For illustration, we used age as surrogate for age onset. Being aware of the fact that this was only an approximation, whenever multipe affected individuals within a sibship are available, their average age was used. Causal variants and associate genes provide information on power of association testing statistics while the non-causal counterparts provide analogous results on Type I error rate.

The statistical significance was assessed according to log likelihood ratio tests between models using relationship only versus using both relationship and IBD information. The computation for this is relatively fast, results for all 200 replicates took 1 hour 48 minutes on our 20-node Linux clusters each with 16GB RAM and 4 CPUs using Sun grid engines. The nominal significance levels are shown in Table 1, which reveal that the tests are both close to the expected level under H_0 and H_A.

Gene-based analysis was also conducted for Q1 involving all 3205 genes and the results are shown with selected candidates highlighted in Figure 1, which agree with the simulated model in which the significant regions were in VEGFC/VEGFA.

As one would be keen to see various parameter estimates in a real analy-

[5]See `http://www.1000genomes.org/`

[6]See `http://en.wikipedia.org/wiki/Vascular_endothelial_growth_factor_C`

sis, we also provide results associated with replicate one. Q1 as based on restricted maximum likelihood (REML) are shown in Table 2. The models with relationship only and with both relationship and IBD information have -2 Res(tricted) Log likelihood being 1789.5 and 1775.2, respetively while Akaike Information Criteria (AIC) being 1793.5 and 1781.2, respectively so that using IBD information improved fit for Q1 (smaller AIC). For AFFECTED the results based on maximum pseudo-likelihood are shown in Table 3 and those from Cox model in Table 4. Note that the improvement in terms of -2 Log Pseudo-Likelihood from 3434.4 to 3445.7 was also substantial. To explore the multivariate model (13) involving the polygenic effects for Q1, Q2 and Q4, the six parameters (σ_{11}, σ_{21}, σ_{22}, σ_{31}, σ_{32}, σ_{33}) in the variance-covariance matrix have been expressed according to (9). The appropriate matrices associated with all parameters are constructed a priori. These are then subject to procedures such as PROC MIXED and *lmekin*. The joint model of Q1, Q2, Q4 are shown in Table 5.

The implementations are provided in **Supplementary information**. While code blocks shown there are appropriate for one instance, it is preferable to use *SAS*'s output delivery system (ODS) to save various results into databases.

3.2 The Framingham Heart Study

The Framingham Heart Study is under the direction of National Heart, Lung, and Blood Institute (NHLBI) which began in 1948 with the recruitment of adults from the town of Framingham, Massachusetts. Data available for GAW16 were 7130 individuals from the original cohort (373), the first generation cohort (2760) and the third generation cohort (3997) with sex, age, height, weight, blood pressure, lipids, smoking and drinking. Data as outlined in [43] was used here, where 6848 had genotype data for at least one of the four specified SNPs (rs1121980, rs9939609, rs17782313 and rs17700633). Data for 96 individuals without any phenotype data but with genotype data and an additional 227 individuals without being assigned a family ID were excluded from analyses. Additionally, four individuals had no data on weight, 86 observations were measured at <18 years of age, and therefore were excluded. The 6,520 remaining individuals were part of 962 families, among which 2073 individuals had completed four visits. Meanwhile, there were also 365 cases of diabetes with their ages of onset.

Kinship information was obtained from family structure and used for

genotype-trait association. Computer program **PLINK** [44] with the *–genome* option was also used to infer correlations ($\hat{\pi}$) using whole genome data. A total of 8485 SNPs on Affymetrix 500K chips were derived from a panel of 45620 informative autosomal SNPs used in our consortium analysis. This led to estimates for $6520(6520-1)/2 = 21251940$ pairs of relationship. The genetic distance according to $|\pi - \hat{\pi}|$ [45], i.e., **sum(abs(EZ-PI_HAT), na.rm=TRUE)**, is 3421.724. Approximately half (10478474) had $\hat{\pi}$ of 0.01 or more. Although there was a good agreement between kinship according to the specified family structures and $\hat{\pi}$, 11207 pairs of individuals deemed to be unrelated had $\hat{\pi}$ between 0.1-0.3 and 12 of which were greater than 0.3.

Both types of relationship matrices were used for the Cox model via *kinship* and *bdsmatrix.ibd* functions in *R*. The frailty and polygenic models had log likelihoods of -1788.53, -1791.93 with variance estimates 0.10^2 and 0.02^2, respectively. However, with inferred relationship the log likelihood turned out to be -1762.69 and variance estimate 0.24^2. Similar model for BMI at wave 1 was also fitted, a family specific random intercept model yielded log likelihood of -19273.26 and variance 3.44, while a correlated random intercept model gave log likelihood -19379.3 and variance 0.01^2 with comparable results from inferred relationship though with a smaller residual error. The results on diabetes might have suggested a substantial genetic effect while for BMI the use of inferred relationship performed equally well with a model using explicit family structures.

4 Discussion

The models we have considered extend counterparts for unrelated sample by taking into account correlation within and heterogeneity between families. To a large extent, we have presented an appreciation of models and implementations for related individuals using mixed models. At the meantime, we have envisaged a whole range of analyses that can be put in the framework. However, compared to [13] and especially [19], our development is more incremental and helps to gain insight into more complicated models. As a key feature of the model specification, oligogenes, polygenes, common environment, gene-environment interaction, and multivariate data are accommodated in a coherent framework via appropriate covariance structure. The generic nature has enabled a range of genetic association studies. Our interpretation of the model also naturally extends the model for quantitative traits out-

lined by [46] and [19]. It has been recognized that for longitudinal data some commonly used covariance structures, such as compound symmetry, can be expressed as "linear covariance of dimension k" [47, p258]. Although it could be more involved, it may be possible in our context. Data as in consortium meta-analysis analysis is also perceived in broader framework consisting of both unrelated and related individuals.

It should be aware that mixed models are quite general and may well be linked to other models. For instance, we noticed that model (10) is reminiscent of an approach proposed for generalized method of moments [48]. An example as with its link with individual empirical Bayes estimates has been provided by [49] and [50]. A reviewer has brought to our attention recent work on nonparametric methods for longitudinal data [51] and the utility of mixed models in controlling for bias of population stratification (e.g., [52]). This paper has limited coverage of literature on longitudinal analysis of family data, mainly owing to the fact that there is greater difficulty in implementation via general software package. However, this is expected to change. To our knowledge, little work has been done on joint analysis of individual data in the GWAS meta-analysis context. In view of the popularity of consortium data analysis, it will be appealing to have the appropriate mechanism to make it possible.

The models and their implementations are connected with whole genome data in several ways. First, the transition from the variance components models in earlier literature becomes more explicit. More specifically, the models described here are appropriate for GWAS where genetic variants coupled with a high resolution map is available. In general, the variance components associated with a major gene as in (7) is a function of the recombination rate (r) [12], i.e., $\sigma^2_M f(r, \pi_{ij})$, where π_{ij} represents identity-by-descent sharing between a pair of individuals i, j for the marker locus; with dense marker, we can assume that $r = 0$ which is also true with (9). Second, as in the Framingham data there is a further benefit with dense genetic markers such that they can be used to infer family structure [53] or (global) IBD information [54]. The availability of the deep sequencing data and a long list of established genes are likely to give greater weight on use of family data [55]. It is also desirable that cryptic relatedness in population-based sample can be appropriately taken into account in association analysis. In our own EPIC-Norolk GWAS, samples with cryptic relatedness have been excluded at the quality control stage [56]. It is interesting to note that, **coxme** was developed for handling large pedigrees involving sparse matrices, the availability of whole

genome data will alter the scenario slightly but nevertheless remain in the same framework. Third, more work is required to shorten computing time. In the literature, it has been proposed to absorb the relationship in the model for quantitative trait by multiplying inverse of the kinship matrix followed by a linear regression, or using residuals from a phenotype-covariate only regression as outcome in a model including SNPs as in **GenABEL**. In principle one can extend the idea to multivariate or longitudinal models where the residuals are obtained only once for GWAS or incorporating regional information before turning to SNP-specific analysis. There are also alternative approaches such as retrospective methods found in **Merlin**. With its greater requirement in computation the "measured genotype" approach here remains intuitive especially for gene-environment characterization. To this point, associate projects such as **BORDICEA**[7] and **BayesMendel**[8] have contributed to the success of work on *R* described here.

A reviewer has expressed interest regarding the Type I error linking to results shown in Table 1. We believe that data as distributed by GAW17 as they were (200 replicates) are not ideal for assessing Type I error and possibly requires a bootstrap procedure. In general, from our experience (and personal communications with Profs Douglas Bates and Terry Therneau), this is a difficult issue and possibly problem specific. In fact, in the recent implementation of GLMM in **lme4**, the associate p values for fixed effects are not shown which nevertheless may leave users with temptation to employ normal approximation. Although we have not conducted extensive numerical experiments, results from GAW17 and the Framingham Study have indicated good performance of these models, and that of the inferred relationship based on whole genome data is impressive. Since only directly genotyped Affy500K SNPs were used, the addition of imputed genotypes, say based on the HapMap, should help to improve the inference. Its use in the usual genomewide association analysis should be considered.

Our attention lies on the implementation by taking advantages of the available implementation in general statistical computing environment. The clarification of the implementation in these should facilitate practical analysis of family data. Although these models are conceptually simple, availability of their implementation vary, notably the ability to allow for both oligogenes and polygenes in a GLMM framework. For *R*, these are at least possible with

[7]See `http://www.srl.cam.ac.uk/genepi/boadicea/boadicea_home.html`

[8]See `http://bcb.dfci.harvard.edu/BayesMendel/`

nlme, **lme4** and additionally **coxme**. At the moment, application of packages in *R* are often restricted with *lmekin* in **coxme** offering outcomes only on continuous outcome but for **pedigreemm** it is unable to handle complex covariance structure. It is desirable that a function called *nlmekin* can be developed as with *pedigreemm* expanded to incorporate additive covariance structures. For *SAS*, MIXED, GLIMMIX and NLMIXED together provide a rich source of practical modeling functionality though the Cox model counterpart is not available. The tackling of various issues has led to efficient algorithm [25]. When the interest is on correlation between multiple traits, the use of **nlme** for multivariate longitudinal data in unrelated individuals has been described [57]. In general, this could be complicate with longitudinal familial data without [58] or with [59] consideration of relationship. In study of obesity-related traits, *FTO* has been shown to be strongly associated with BMI and supported by cross-sectional data as in [14], longitudinal data as in [43] and data across life span as in [60]. Our previous attempt [43] was based on a three-level model and it would be of interest to use kinship information as well.

While the framework we have outlined is comprehensive, we feel that our "proof of concepts" here awaits for extensive testing. It is also desirable that the current implementation can be optimized in computing time. A lot of work has been done for quantitative genetics in plants and aminals. Our experience indicated that the running time with *SAS* was longer time than *R*. However, in an analysis of longitudinal lung function data in the EPIC-Norfolk study, we have shown that although an individual analysis could be slow it is possible to perform an analysis for GWAS using *SAS* and Linux clusters so that ∼2.5M SNPs would finish within 14 hours when running each chromosome on a separate node. It is likely that was benefited from *SAS* caching frequently-used instructions. Greater proportion of coding in *C/C++* should also be helpful. Given the utility of the popular environments can be shown, their take-up in genomewide association studies will be quick and it is very much in line with efforts in other disciplines where large volume of data is involved.

Acknowledgements

A lot of the insights were gained during analysis of GAWs 14, 16, 17 and in particular maintenance of the *R* counterpart of the *S-PLUS* package **kin-**

ship[9] by the first author. We are therefore very grateful of the pioneering work and advices given by Profs Terry Therneau, Beth Atkinson, Mariza de Andrade all at the Mayo Clinic and interactions with many other colleagues elsewhere. The comprehensive *R* archive network (CRAN[10]) as with Profs Kurt Hornik and Brian Ripley has been a constant source of support. The work presented here was partly done for CompBio2011 and useR!2011. We wish to thank Drs Qihua Tan, Fuzhong Xue, Wendi Qian, Luigi Palla for their participation and comments during the GAWs 16 & 17 analysis which led to this work, Dr Wendi Qian's comments on *SAS* PROC GLIMMIX, and Dr Antonis Antoniou's suggestion of using average age within a sibship to approximate age at onset. The example regarding twins was due to a query from my colleague Dr Marcel de van Hoed.

We are also grateful of the Editor for communications which led to the work on the paper and three anonymous reviewers for their insightful comments which led to its improvement. The work reported here also allows us for making minor changes to the syntax shown in [36]. Prof Peter McCullagh from University of Chicago and Dr David Clifford from CSIRO have kindly provided advices regarding the use of **regress**.

References

[1] D. C. Thomas and W. J. Gauderman. Gibbs sampling methods in genetics. In W. R. Gilks, S. Richard, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC., 1996.

[2] D. C. Thomas. *Statistical Methods in Genetic Epidemiology*. Oxford University Press, 2004.

[3] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565–9, 2010.

[4] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Trans Roy Soc Edin*, 52:399–433, 1918.

[9] `http://mayoresearch.mayo.edu/mayo/research/biostat/upload/kinship.pdf`
[10] `http://cran.r-project.org`

[5] H. L. Allen, K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, T. Ferreira, A. R. Wood, R. J. Weyant, A. V. Segre, E. K. Speliotes, E. Wheeler, N. Soranzo, J. H. Park, J. Yang, D. Gudbjartsson, N. L. Heard-Costa, J. C. Randall, L. Qi, A. V. Smith, R. Magi, T. Pastinen, L. Liang, I. M. Heid, J. Luan, G. Thorleifsson, T. W. Winkler, M. E. Goddard, K. S. Lo, C. Palmer, T. Workalemahu, Y. S. Aulchenko, A. Johansson, M. C. Zillikens, M. F. Feitosa, T. Esko, T. Johnson, S. Ketkar, P. Kraft, M. Mangino, I. Prokopenko, D. Absher, E. Albrecht, F. Ernst, N. L. Glazer, C. Hayward, J. J. Hottenga, K. B. Jacobs, J. W. Knowles, Z. Kutalik, K. L. Monda, O. Polasek, M. Preuss, N. W. Rayner, N. R. Robertson, V. Steinthorsdottir, J. P. Tyrer, B. F. Voight, F. Wiklund, J. F. Xu, J. H. Zhao, D. R. Nyholt, N. Pellikka, M. Perola, J. R. B. Perry, I. Surakka, M. L. Tammesoo, E. L. Altmaier, N. Amin, T. Aspelund, T. Bhangale, G. Boucher, D. I. Chasman, C. Chen, L. Coin, M. N. Cooper, A. L. Dixon, Q. Gibson, E. Grundberg, K. Hao, M. J. Junttila, L. M. Kaplan, J. Kettunen, I. R. Konig, T. Kwan, R. W. Lawrence, D. F. Levinson, M. Lorentzon, B. McKnight, A. P. Morris, M. Muller, J. S. Ngwa, S. Purcell, S. Rafelt, R. M. Salem, E. Salvi, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[6] G. B. Ehret, P. B. Munroe, K. M. Rice, M. Bochud, A. D. Johnson, D. I. Chasman, A. V. Smith, M. D. Tobin, G. C. Verwoert, S. J. Hwang, V. Pihur, P. Vollenweider, P. F. O'Reilly, N. Amin, J. L. Bragg-Gresham, A. Teumer, N. L. Glazer, L. Launer, J. H. Zhao, Y. Aulchenko, S. Heath, S. Sober, A. Parsa, J. Luan, P. Arora, A. Dehghan, F. Zhang, G. Lucas, A. A. Hicks, A. U. Jackson, J. F. Peden, T. Tanaka, S. H. Wild, I. Rudan, W. Igl, Y. Milaneschi, A. N. Parker, C. Fava, J. C. Chambers, E. R. Fox, M. Kumari, M. J. Go, P. van der Harst, W. H. Kao, M. Sjogren, D. G. Vinay, M. Alexander, Y. Tabara, S. Shaw-Hawkins, P. H. Whincup, Y. Liu, G. Shi, J. Kuusisto, B. Tayo, M. Seielstad, X. Sim, K. D. Nguyen, T. Lehtimaki, G. Matullo, Y. Wu, T. R. Gaunt, N. C. Onland-Moret, M. N. Cooper, C. G. Platou, E. Org, R. Hardy, S. Dahgam, J. Palmen, V. Vitart, P. S. Braund, T. Kuznetsova, C. S. Uiterwaal, A. Adeyemo, W. Palmas, H. Campbell, B. Ludwig, M. Tomaszewski, I. Tzoulaki, N. D. Palmer, T. Aspelund, M. Garcia, Y. P. Chang, J. R. O'Connell, N. I. Steinle, D. E. Grobbee, D. E. Arking, S. L. Kardia, A. C. Morrison,

D. Hernandez, S. Najjar, W. L. McArdle, D. Hadley, M. J. Brown, J. M. Connell, A. D. Hingorani, I. N. Day, D. A. Lawlor, J. P. Beilby, R. W. Lawrence, R. Clarke, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–9, 2011.

[7] T. M. Teslovich, K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou, M. Koseki, J. P. Pirruccello, S. Ripatti, D. I. Chasman, C. J. Willer, C. T. Johansen, S. W. Fouchier, A. Isaacs, G. M. Peloso, M. Barbalic, S. L. Ricketts, J. C. Bis, Y. S. Aulchenko, G. Thorleifsson, M. F. Feitosa, J. Chambers, M. Orho-Melander, O. Melander, T. Johnson, X. Li, X. Guo, M. Li, Y. Shin Cho, M. Jin Go, Y. Jin Kim, J. Y. Lee, T. Park, K. Kim, X. Sim, R. Twee-Hee Ong, D. C. Croteau-Chonka, L. A. Lange, J. D. Smith, K. Song, J. H. Zhao, X. Yuan, J. Luan, C. Lamina, A. Ziegler, W. Zhang, R. Y. Zee, A. F. Wright, J. C. Witteman, J. F. Wilson, G. Willemsen, H. E. Wichmann, J. B. Whitfield, D. M. Waterworth, N. J. Wareham, G. Waeber, P. Vollenweider, B. F. Voight, V. Vitart, A. G. Uitterlinden, M. Uda, J. Tuomilehto, J. R. Thompson, T. Tanaka, I. Surakka, H. M. Stringham, T. D. Spector, N. Soranzo, J. H. Smit, J. Sinisalo, K. Silander, E. J. Sijbrands, A. Scuteri, J. Scott, D. Schlessinger, S. Sanna, V. Salomaa, J. Saharinen, C. Sabatti, A. Ruokonen, I. Rudan, L. M. Rose, R. Roberts, M. Rieder, B. M. Psaty, P. P. Pramstaller, I. Pichler, M. Perola, B. W. Penninx, N. L. Pedersen, C. Pattaro, A. N. Parker, G. Pare, B. A. Oostra, C. J. O'Donnell, M. S. Nieminen, D. A. Nickerson, G. W. Montgomery, T. Meitinger, R. McPherson, M. I. McCarthy, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–13, 2010.

[8] E. K. Speliotes, C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, A. U. Jackson, H. L. Allen, C. M. Lindgren, J. Luan, R. Magi, J. C. Randall, S. Vedantam, T. W. Winkler, L. Qi, T. Workalemahu, I. M. Heid, V. Steinthorsdottir, H. M. Stringham, M. N. Weedon, E. Wheeler, A. R. Wood, T. Ferreira, R. J. Weyant, A. V. Segre, K. Estrada, L. Liang, J. Nemesh, J. H. Park, S. Gustafsson, T. O. Kilpelainen, J. Yang, N. Bouatia-Naji, T. Esko, M. F. Feitosa, Z. Kutalik, M. Mangino, S. Raychaudhuri, A. Scherag, A. V. Smith, R. Welch, J. H. Zhao, K. K. Aben, D. M. Absher, N. Amin, A. L. Dixon, E. Fisher, N. L. Glazer,

M. E. Goddard, N. L. Heard-Costa, V. Hoesel, J. J. Hottenga, A. Johansson, T. Johnson, S. Ketkar, C. Lamina, S. Li, M. F. Moffatt, R. H. Myers, N. Narisu, J. R. Perry, M. J. Peters, M. Preuss, S. Ripatti, F. Rivadeneira, C. Sandholt, L. J. Scott, N. J. Timpson, J. P. Tyrer, S. van Wingerden, R. M. Watanabe, C. C. White, F. Wiklund, C. Barlassina, D. I. Chasman, M. N. Cooper, J. O. Jansson, R. W. Lawrence, N. Pellikka, I. Prokopenko, J. Shi, E. Thiering, H. Alavere, M. T. Alibrandi, P. Almgren, A. M. Arnold, T. Aspelund, L. D. Atwood, B. Balkau, A. J. Balmforth, A. J. Bennett, Y. Ben-Shlomo, R. N. Bergman, S. Bergmann, H. Biebermann, A. I. Blakemore, T. Boes, L. L. Bonnycastle, S. R. Bornstein, M. J. Brown, T. A. Buchanan, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42(11):937–48, 2010.

[9] R. Plomin, C. M. Haworth, and O. S. Davis. Common disorders are quantitative traits. *Nat Rev Genet*, 10(12):872–8, 2009.

[10] C. E. McCulloch and S. R. Searle. *Generalized, Linear, and Mixed Models*. Wiley, 2001.

[11] SAS Institute Inc. *SAS/STAT 9.3 User's Guide*. SAS Publishing, Cary, NC, USA, 2011.

[12] C. I. Amos. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet*, 54(3):535–43, 1994.

[13] J. Blangero, J. T. Williams, and L. Almasy. Variance component methods for detecting complex trait loci. *Advances in Genetics*, 42:151–81, 2001.

[14] T. M. Frayling, N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, J. R. Perry, K. S. Elliott, H. Lango, N. W. Rayner, B. Shields, L. W. Harries, J. C. Barrett, S. Ellard, C. J. Groves, B. Knight, A. M. Patch, A. R. Ness, S. Ebrahim, D. A. Lawlor, S. M. Ring, Y. Ben-Shlomo, M. R. Jarvelin, U. Sovio, A. J. Bennett, D. Melzer, L. Ferrucci, R. J. Loos, I. Barroso, N. J. Wareham, F. Karpe, K. R. Owen, L. R. Cardon, M. Walker, G. A. Hitman, C. N. Palmer, A. S. Doney, A. D. Morris, G. D. Smith, A. T. Hattersley, and M. I. McCarthy. A common variant in the FTO gene is associated with body

mass index and predisposes to childhood and adult obesity. *Science*, 316(5826):889–94, 2007.

[15] N. E. Morton and C.J. MacLean. Analysis of family resem blance. III. complex segregation of quantitative traits. *Am J Hum Genet*, 26:489–503, 1974.

[16] J. L. Hopper and J. D. Mathew. Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet*, 46:373–83, 1982.

[17] K. Lange and M. Boehnke. Extensions to pedigree analysis. IV. covariance components models for multivariate traits. *Am J Med Genet*, 14:513–24, 1983.

[18] S. J. Hasstedt. A mixed-model likelihood approximation on large pedigrees. *Comput Biomed Res*, 15:295.307., 1982.

[19] M. P. Epstein, J. E. Hunter, E. G. Allen, S. L. Sherman, X. Lin, and M. Boehnke. A variance-component framework for pedigree analysis of continuous and categorical outcomes. *Stat Biosci*, 1(2):181–198, 2009.

[20] A. M. Saxton, editor. *Genetic Analysis of Complex Traits Using SAS*. SAS Publishing, 2004.

[21] A. I. Vazquez, D. M. Bates, G. J. M. Rose, D. Gianola, and K. A. Weigel. Technical note: An r package for fitting generalized linear mixed models in animal breeding. *J Anmi Sci*, 88:497–504, 2010.

[22] V. S. Pankratz, M. de Andrade, and T. M. Therneau. Random-effects cox proportional hazards model: general variance components methods for time-to-event data. *Genet Epidemiol*, 28(2):97–109, 2005.

[23] V. Ducrocq and G. Casella. A Bayesian analysis of mixed survival models. *Gen. Sel. Evol.*, 28:505–29, 1996.

[24] D. Sorensen and D. Gianola. *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. Springer, 2002.

[25] P. Waldmann. Easy and flexible bayesian inference of quantitative genetics parameters. *Evolution*, 63-6:1640–3, 2009.

[26] P. R. Burton, K. J. Scurrah, M. D. Tobin, and L. J. Palmer. Covariance components models for longitudinal family data. *Int J Epidemiol*, 34(5):1063–77; discussion 1077–9, 2005.

[27] J. M. Lachin. *Biostatistical Methods-The Assessment of Relative Risks*. Wiley, second edition, 2011.

[28] A. Skrondal and S. Rave-Hesketh. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC., 2004.

[29] J. Whitehead. Fitting Cox's regression model to survival data using GLIM. *Appl Stat*, 29:268–75, 1980.

[30] G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, 2000.

[31] J. Pinheiro and D. M. Bates. *Mixed Effects Models in S and S-PLUS*. Springer, 2000.

[32] R. B. Bapat. *Linear Algebra and Linear Models*. Springer, third edition, 2012.

[33] P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, second edition, 2002.

[34] J. P. Klein and M. L. Moeschberger. *Survival Analysis-Techniques for Censored and Truncated Data*. Springer, second edition, 2003.

[35] R. D. Riley, P. C. Lambert, and G. Abo-Zaid. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*, 340:c221, 2010.

[36] J. H. Zhao, J. A. Luan, R. J. F. Loos, and N. J. Wareham. On genotype-phenotype association using SAS. In *CompBio 2011*, Proceedings of the IASTED International Conference, pages 428–433, Cambridge, United Kingdom, 2011.

[37] T. D. Pigott. *Advances in Meta-Analysis*. Springer, 2012.

[38] J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32, 1948.

[39] P. Hall, J. S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 67:427–444, 2005.

[40] N. J. Schork. Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. *Am J Hum Genet*, 53(6):1306–19, 1993.

[41] J. H. Zhao and Q. Tan. Integrated analysis of genetic data with R. *Hum Genomics*, 2(4):258–65, 2006.

[42] L. Almasy, T. D. Dyer, J. M. Peralta, J. W. Kent, J. C. Charlesworth, J. E. Curran, and J. Blangero. Genetic analysis workshop 17 mini-exome simulation. *BMC Proc*, 5(Suppl 9):S2, 2011.

[43] J. Luan, B. Kerner, J. H. Zhao, R. J. Loos, S. J. Sharp, B. O. Muthen, and N. J. Wareham. A multilevel linear mixed model of the association between candidate genes and weight and body mass index using the framingham longitudinal family data. *BMC Proc*, 3 Suppl 7:S115, 2009.

[44] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75, 2007.

[45] A. Sanchez, J. Ocana, and F. Utzet. Sampling theory, estimation, and significance testing for prevosti's estimate of genetic distance. *Biometrics*, 51(4):1216–1235, 1995.

[46] M. de Andrade, E. Atkinson, E. Lunde, C. I. Amos, and J. Chen. Estimating genetic components of variance for quantitative traits in family studies using the multic. Technical report, Mayo Clinic, 2006.

[47] E. F. Vonesh and V. M. Chinchilli. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Marcel Dekker, 1997.

[48] G. Yin. Bayesian generalized method of moments. *Bayesian Analysis*, 4:191–208, 2009.

[49] T. A. Moger, O. O. Aalen, K. Heimdal, and H. K. Gjessing. Analysis of testicular cancer data using a frailty model with familial dependence. *Stat Med*, 23(4):617–632, 2004.

[50] O. O. Aalen, O. Borgan, and H. K. Ejessing. *Survival and Event History Analysis-A Process Point of View*. Springer, 2008.

[51] Y. J. Wang, C. H. Huang, Y. X. Fang, Q. Yang, and R. Z. Li. Flexible semiparametric analysis of longitudinal genetic studies by reduced rank smoothing. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 61:1–24, 2012.

[52] J. Yu, G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*, 38(2):203–8, 2006.

[53] A. G. Day-Williams, J. Blangero, T. D. Dyer, K. Lange, and E. M. Sobel. Linkage analysis without defined pedigrees. *Genet Epidemiol*, 35(5):360–70, 2011.

[54] L. Han and M. Abney. Identity by descent estimation with dense genome-wide genotype data. *Genet Epidemiol.*, 35(6):557–567, 2011.

[55] James R. Lupski, Jeffrey G. Reid, Claudia Gonzaga-Jauregui, David Rio Deiros, David C. Y. Chen, Lynne Nazareth, Matthew Bainbridge, Huyen Dinh, Chyn Jing, David A. Wheeler, Amy L. McGuire, Feng Zhang, Pawel Stankiewicz, John J. Halperin, Chengyong Yang, Curtis Gehman, Danwei Guo, Rola K. Irikat, Warren Tom, Nick J. Fantin, Donna M. Muzny, and Richard A. Gibbs. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *New England Journal of Medicine*, 362(13):1181–1191, 2010.

[56] R. J. F. Loos, C. M. Lindgren, S. Li, E. Wheeler, J. H. Zhao, I. Prokopenko, M. Inouye, R. M. Freathy, A. P. Attwood, J. S. Beckmann, S. I. Berndt, K. B. Jacobs, S. J. Chanock, R. B. Hayes, S. Bergmann, A. J. Bennett, S. A. Bingham, M. Bochud, M. Brown, S. Cauchi, J. M. Connell, C. Cooper, G. D. Smith, I. Day, C. Dina, S. De, E. T. Dermitzakis, A. S. Doney, K. S. Elliott, P. Elliott, D. M. Evans, I. Sadaf Farooqi, P. Froguel, J. Ghori, C. J. Groves, R. Gwilliam, D. Hadley, A. S. Hall, A. T. Hattersley, J. Hebebrand, I. M. Heid, C. Lamina, C. Gieger, T. Illig, T. Meitinger, H. E. Wichmann, B. Herrera, A. Hinney, S. E. Hunt, M. R. Jarvelin, T. Johnson, J. D. Jolley,

F. Karpe, A. Keniry, K. T. Khaw, R. N. Luben, M. Mangino, J. Marchini, W. L. McArdle, R. McGinnis, D. Meyre, P. B. Munroe, A. D. Morris, A. R. Ness, M. J. Neville, A. C. Nica, K. K. Ong, S. O'Rahilly, K. R. Owen, C. N. Palmer, K. Papadakis, S. Potter, A. Pouta, L. Qi, J. C. Randall, N. W. Rayner, S. M. Ring, M. S. Sandhu, A. Scherag, M. A. Sims, K. Song, N. Soranzo, E. K. Speliotes, H. E. Syddall, S. A. Teichmann, N. J. Timpson, J. H. Tobias, M. Uda, C. I. Vogel, C. Wallace, D. M. Waterworth, M. N. Weedon, C. J. Willer, Wraight, X. Yuan, E. Zeggini, J. N. Hirschhorn, D. P. Strachan, W. H. Ouwehand, M. J. Caulfield, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet*, 40(6):768–75, 2008.

[57] S. Bandyopadhyay, B. Ganguli, and A. Chatterjee. A review of multivariate longitudinal data analysis. *Stat Methods Med Res*, 20(4):299–330, 2011.

[58] B. C. Sutradhar. *Dynamic Mixed Models for Familial Longitudinal Data*. Springer, 2011.

[59] J. M. Soler and J. Blangero. Longitudinal familial analysis of blood pressure involving parametric (co)variance functions. *BMC Genet*, 4 Suppl 1:S87, 2003.

[60] R. Hardy, A. K. Wills, A. Wong, C. E. Elks, N. J. Wareham, R. J. F. Loos, D. Kuh, and K. K. Ong. Life course variations in the associations between FTO and MC4R gene variants and body size. *Hum Mol Genet*, 19(3):545–52, 2010.

[61] J. H. Zhao. Mixed-effects Cox models of alcohol dependence in extended families. *BMC Genet*, 6(Suppl 1):S127, 2005.

[62] D. Clifford and P. McCullagh. The regress function. *R News*, 6(2):6–10, 2006.

[63] K. Zhao, M. Nordborg, and P. Marjoram. Genome-wide association mapping using mixed-models: application to gaw15 problem 3. *BMC Proc*, 1 Suppl 1:S164, 2007.

Supplementary information

Our focus in this section will be in *R* and compared to *SAS*. Some comments on Bayesian methods will be given at later part of this section.

R implementation

The **kinship** package was used for kinship calculation, linear mixed model as with mixed Cox models. The package was originally implemented in *S-PLUS* and ported to *R* as described in [61]. Some recent initatives have been made to improve the facilities for handling sparse matrices, various tools for family data including pedigree drawing as with kinship calculation, and mixed effects Cox model, so the original **kinship** package was partitioned into three separate packages called **bdsmatrix**, **kinship2** and **coxme**. The **pedigreemm** package [21] is appropriate for modeling polygenes within the GLMM framework.

Kinship calculation

A function *makefamid* from **kinship** will generate a "pedigree" type, which can be used by function *makekinship* to obtain kinship matrices from different families,

```
library(kinship)
pid <- with(fam, makefamid(ID, FA, MO))
kmat <- with(fam, makekinship(pid, ID, FA, MO))
```

Note that with GWAS this only needs to be done once and does not have a big overhead. Interestingly, the models for a collection of monozygotic (MZ) and dizygotic (DZ) twins can be treated as a special case. A model using an exchangeable correlation, say, will not be so desirable compared to those using the kinship information[11]. Consider a study of nMZ MZ and nDZ DZ twins, we can order that data such that MZ twins precede their DZ counterpart, then function *bdsmatrix* is called to generate the kinship matrix to be used by *glm* for a sporadic model or *lmekin* for a linear mixed model.

[11] For twin data, to account for the relationship between twin pairs one can pragmatically specify the correlation structure. In one study of physical activity, we order twin pairs by zygosity such that MZ precede DZ twins in the data, we can then subject the data for analysis with following code.

```
idn <- 1:(2*(nMZ+nDZ))
kmat <- bdsmatrix(rep(2,nMZ+nDZ),c(rep(0.5,3*nMZ),
                      rep(c(0.5,0.25,0.5),nDZ)),
                      dimnames=list(idn,idn))

glmfit <- glm(paee ~ walkability + age + weight + sex)
summary(glmfit)
kfit <- lmekin(paee ~ walkability + age + weight + sex,
               random = ~1|id, varlist=list(kmat))
kfit
```

There are a number of other packages available, e.g., **gap** and **identity**.

Linear mixed model

```
library(kinship)
kkin <- lmekin(Q1 ~ SEX + AGE + SMOKE,
        data=pg, random = ~1|ID, varlist=list(kmat))
kibd <- lmekin(Q1 ~ SEX + AGE + SMOKE,
        data=pg, random = ~1|ID,
        varlist=list(kmat,ibdmat))
```

Generalized linear mixed model

```
library(pedigreemm)
bt <- pedigreemm(AFFECTED ~ SEX + AGE + SMOKE + (1|ID),
      data=pg, family="binomial"(link="logit"),
      pedigree=list(ID=ped))
```

Multivariate model

The package **multic** [46] has facility for multivariate analysis, however, it was bound to particular environments. In principle, this is a computing problem that can be fixed.

Marginal models

A notable implementation is *R* **gee** package.

```
library(gee)
m1 <- gee(Q1 ~ SEX + AGE + SMOKE, id=pid,
          data=pg, corstr="exchangeable")
summary(m1)
```

To ensure maximum compatibility with the GLMM fit, the scale parameter is chosen to be fixed at the default value of 1. The structure "exchangeable" assumes equal correlations between relatives in a pedigree but in principle this could be modified to use kinship matrix as in *SAS* below.

Mixed Cox models

This is a Cox model with correlated frailty.

```
library(kinship)
kcox <- coxme(Surv(age, AFFECTED) ~ SEX + SMOKE,
        data=pg, random = ~1|ID, varlist=list(kmat))
```

More information is available from the package vignette.

kinship2 and coxme

As of 14 March 2012, the current version of **kinship** at CRAN will be archived. This has been due to a recent development which involves splitting the package into three separate ones, namesly **bdsmatrix**, **kinship2**, and **coxme**. One only expects a slight change from a user's perspective, e.g., the way to specify random effects associated with **coxme**. Although this also involves *lmekin*, here only examples with respect to Cox model are given. For the Framingham data contained in **nf**, the diabetes status and age onset were available and the modeling syntax is as follows,

```
library(kinship2)
attach(nf)
f <- makefamid(shareid, fshare, mshare)
k <- makekinship(f, shareid, fshare, mshare)
detach(nf)
library(coxme)
print("Cox model with random intercept")
f1 <- coxme(Surv(agediab, diabetes) ~ sex + (1|pedno), nf)
```

```
f1
print("Cox model with random intercept and additive variance")
k2 <- 2*as.matrix(k)
f2 <- coxme(Surv(agediab, diabetes) ~ sex + (1|shareid), nf,
            varlist=coxmeMlist(k2, rescale=FALSE))
f2
```

The standard Cox model provides a baseline to compare. Note that **kinship2** depends on **Matrix** so k2 is created for *coxme*.

Suppose we intend to read output **k.dat** from **PLINK**, we can use the following code,

```
k <- read.table("k.dat",header=TRUE)
library(bdsmatrix)
attach(k)
ID <- unique(c(IID1,IID2))
t1 <- cbind(IID1,IID2,PI_HAT)
t2 <- cbind(IID1=ID,IID2=ID,PI_HAT=0.5)
trio <- rbind(t1,t2)
k2 <- bdsmatrix.ibd(trio)
detach(k)
save(k,k2,file="k.RData")
```

Note that we add the diagonal elements in the kinship matrix, which can be loaded with *load*("k.RData").

regress and MASS

After submission of the paper, we learned about the work in a similar but alternative context [62]. It turned out that the associated package **regress** yielded comparable results to *lmekin* from **kinship** (data not shown).

We have also become aware of the possibility to use *glmmPQL* available from **MASS** [63] and it appears straightforward to use the *corSymm* function in **nlme** to construct a correlation for data on twins and affected sib pairs as input for the *correlation* option, but for data containing general pedigrees this is more involved and we had limited experience of success.

SAS implementation

SAS has used G and R to indicate the variance-covariance matrices associated with random and fixed effects. The procedures of interest in *SAS* are MIXED, GLIMMIX, NLMIXED. GLIMMIX is an extension to both PROC GENMOD and PROC MIXED.

When individuals in a pedigree is ordered appropriately, the specification should be as follows,

```
proc inbreed data=families covar outcov=kmat;
     var id fa mo;
run;
```

The following block is for the polygenic model.

```
title kinship only;
proc mixed data=pheno covtest asycov noclprint;
     class id;
     model q1=sex age smoke SNP / noint solution covb;
     random id / type=lin(1) ldata=kmat solution;
run;
proc glimmix data=pheno asycov method=mmpl;
     class id;
     model affected(event='1')=sex age smoke SNP
           / dist=binary link=logit solution covb;
     random id / type=lin(1) ldata=kmat solution;
     random _residual_;
run;
```

By default, PROC MIXED employs REML method. For PROC GLIMMIX, maximum or restricted maximum likelihood approach was applied to a pseudo-likelihood (PL) in the sense that a linearization is applied, leading to abbreviations such as M_PL and R_PL, where _ can be subject-specific (S) expansion where linearization is carried out about the current estimate or β and U, or a marginal (M) expansion where the linearization is about a current estimate of β and $E(U) = 0$. If method=QUAD is specified, an adaptive Gauss-Hermite quadrature is used.

The following block is for both oligogenic and polygenic effects

```
title kinship and ibd;
proc mixed data=pheno covtest asycov noprofile;
     class id;
     model q1=sex age smoke / noint solution covb;
     random id / type=lin(2) ldata=kibd solution;
run;
proc glimmix data=pheno asycov method=mmpl;
     class id;
     model affected(event='1')=sex age smoke
           / dist=binary link=logit solution covb;
     random id / type=lin(2) ldata=kibd solution;
     random _residual_;
run;
```

Therefore the method of estimation here is maximum marginal pseudo-likelihood. As can be seen, the second part has extended those available from R and the statement "random _residual_" also allows for overdispersion.

For Cox model, one can take advantage of the PHREG procedure with RANDOM statement to specify a shared frailty model which can be compared with a model using ID statement to identify clusters.

Finally, one can specify the relationship as R part of the variance-covariance matrix as follows,

```
proc mixed data=pheno sandwich;
     class id;
     model q1=sex age smoke / noint;
     repeated / type=lin(1) ldata=kmat sub=pid;
run;
```

Where the PROC statement specifies the data set to be analyzed using a sandwich estimator, MODEL the statistical model, REPEATED the R matrix incorporating kinship information.

Multivariate implementations

Simulation and estimation for a tri-variate normal data

We simulated 500 samples under a tri-variate normal $N(\mu, \Sigma)$ with

$$\mu = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 10 & 1 & 2 \\ 1 & 20 & 3 \\ 2 & 3 & 50 \end{pmatrix}$$

The simulation and estimation are furnished as follows,

```
library(regress)
library("MASS")
set.seed(12345)
n <- 500
m <- c(1,2,3)
S <- matrix(c(10,1,2, 1,20,3, 2,3,50),3,3)
Y <- mvrnorm(n,m,S)
y <- as.vector(t(Y))
c <- kronecker(rep(1,n),diag(1,3))
V1 <- matrix(c(1,0,0, 0,0,0, 0,0,0),3,3,byrow=TRUE)
V2 <- matrix(c(0,1,0, 1,0,0, 0,0,0),3,3,byrow=TRUE)
V3 <- matrix(c(0,0,0, 0,1,0, 0,0,0),3,3,byrow=TRUE)
V4 <- matrix(c(0,0,1, 0,0,0, 1,0,0),3,3,byrow=TRUE)
V5 <- matrix(c(0,0,0, 0,0,1, 0,1,0),3,3,byrow=TRUE)
V6 <- matrix(c(0,0,0, 0,0,0, 0,0,1),3,3,byrow=TRUE)
id <- as.vector(t(cbind(1:n,1:n,1:n)))
s1 <- kronecker(diag(1,n),V1)
s2 <- kronecker(diag(1,n),V2)
s3 <- kronecker(diag(1,n),V3)
s4 <- kronecker(diag(1,n),V4)
s5 <- kronecker(diag(1,n),V5)
s6 <- kronecker(diag(1,n),V6)
results <- regress(y~c-1,~s1+s2+s3+s4+s5+s6,pos=c(1,0,1,0,0,1),
           identity=FALSE,start=c(10,1,20,1,1,30))
apply(Y,2,mean)
cov(Y)
```

which produces results as follows,

```
Likelihood kernel: K = c1+c2+c3

Maximized log likelihood with kernel K is  -3041.732

Linear Coefficients:
    Estimate Std. Error
 c1    0.891      0.144
 c2    2.026      0.201
 c3    3.592      0.313

Variance Coefficients:
    Estimate Std. Error
 s1   10.313      0.653
 s2    1.313      0.649
 s3   20.241      1.281
 s4    3.476      1.017
 s5    2.881      1.414
 s6   48.863      3.093

> apply(Y,2,mean)
[1] 0.8908874 2.0262134 3.5922123
> cov(Y)
          [,1]      [,2]      [,3]
[1,] 10.312879  1.313217  3.475876
[2,]  1.313217 20.240644  2.881214
[3,]  3.475876  2.881214 48.863036
```

Through package **regress** we obtained

$$\hat{X} = \begin{pmatrix} 0.89 \\ 2.03 \\ 3.59 \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} 10.31 & 1.31 & 3.48 \\ 1.31 & 20.24 & 2.88 \\ 3.48 & 2.88 & 48.86 \end{pmatrix}$$

agreeing with the simulated data. We now turn to the GAW17 data using a multivariate model for Q1, Q2, Q4, with a bit of simplication over covariance specification,

```
library(foreign)
pheno <- read.dta("pheno2.dta")
```

```
kid <- read.csv("kmat.csv")
k <- as.matrix(kid[,-1])

library(regress)
library("MASS")

n <- 697
v1 <- v2 <- v3 <- v4 <- v5 <- v6 <- matrix(0,3,3)
v1[1,1] <- 1
v2[1,2] <- v2[2,1] <- 1
v3[2,2] <- 1
v4[1,3] <- v4[3,1] <- 1
v5[2,3] <- v5[3,2] <- 1
v6[3,3] <- 1
s1 <- kronecker(v1,k)
s2 <- kronecker(v2,k)
s3 <- kronecker(v3,k)
s4 <- kronecker(v4,k)
s5 <- kronecker(v5,k)
s6 <- kronecker(v6,k)
c <- kronecker(rep(1,n),diag(1,3))
id <- as.vector(t(cbind(1:n,1:n,1:n)))
results <- regress(q~-1+c+sex+age+smoke,~s1+s2+s3+s4+s5+s6,
                   identity=FALSE,pos=c(1,0,1,0,0,1),
                   start=c(5.546, 2.999, 3.940, -1.260, -0.780, 0.680),
                   data=pheno)
results
```

The results are given as follows,

```
Likelihood kernel: K = c1+c2+c3+sex+age+smoke

Maximized log likelihood with kernel K is  -1393.867

Linear Coefficients:
       Estimate Std. Error
 c1       0.565      0.108
 c2       0.531      0.109
 c3       0.526      0.109
```

```
 sex     -0.005      0.043
 age     -0.013      0.001
 smoke   -0.019      0.051

Variance Coefficients:
       Estimate Std. Error
    s1    4.219      0.227
    s2   -0.103      0.166
    s3    4.542      0.244
    s4    0.601      0.178
    s5   -0.108      0.183
    s6    5.115      0.275
```

SAS implementation

We first revisit the simulated data generated above. Assuming the Y and indicator c are stored in dataset *mv* while the coefficient matrices are stored in *mv_ldata*, then the appropriate syntax in *SAS* is as follows,

```
proc mixed data=mv covtest asycov noclprint;
     class id c;
     model q=c / noint solution;
     random c*id / type=lin(6) ldata=mv_ldata;
run;
```

Although *SAS* complains about *Convergence criteria met but final hessian is not positive definite*, it turns out that the estimats are fairely close.

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr Z
LIN(1)	9.3328	0.6529	14.30	<.0001
LIN(2)	1.3133	0.6494	2.02	0.0432
LIN(3)	19.2612	1.2814	15.03	<.0001
LIN(4)	3.4760	1.0169	3.42	0.0006
LIN(5)	2.8813	1.4138	2.04	0.0415

LIN(6)	47.8845	3.0936	15.48	<.0001
Residual	0.9797	0	.	.

Asymptotic Covariance Matrix of Estimates

Cov Parm	CovP1	CovP2	CovP3	CovP4	CovP5	CovP6
LIN(1)	0.4262	0.05428	0.006913	0.1437	0.01830	0.04843
LIN(2)	0.05428	0.4218	0.1065	0.06870	0.1486	0.04015
LIN(3)	0.006913	0.1065	1.6421	0.01517	0.2338	0.03328
LIN(4)	0.1437	0.06870	0.01517	1.0341	0.1487	0.6808
LIN(5)	0.01830	0.1486	0.2338	0.1487	1.9988	0.5643
LIN(6)	0.04843	0.04015	0.03328	0.6808	0.5643	9.5704

-2 Res Log Likelihood	8853.4
AIC (smaller is better)	8867.4

Solution for Fixed Effects

Effect	c	Estimate	Standard Error	DF	t Value	Pr > |t|
c	1	0.8909	0.1436	1497	6.20	<.0001
c	2	2.0262	0.2012	1497	10.07	<.0001
c	3	3.5922	0.3126	1497	11.49	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
c	3	1497	76.12	<.0001

We now return to the GAW17 data. With the same spcification of Q1, Q2, and Q4 in a single outcome, along with a variable c corresponding to particular traits, the GLMMIX counterpart is as follows,

```
title kinship and multivariate;
proc mixed data=pheno2 covtest asycov noclprint;
     class id c;
     model q=c sex age smoke / noint solution covb;
     random c*id / type=lin(6) ldata=ldata solution;
run;
```

Note that in addition to a comparable estimate to the R implementation, the *REPEATED /group=c* statement also adds trait-specific residual variances. Furthermore, *ldata* contains the coefficient matrix generated from kinship matrix *kmat* via the following code,

```
proc iml;
     use kmat;
     read all var _num_ into kmat;
     k=0;
     do i=1 to 3;
        do j=1 to i;
           j3=j(3,3,0);
           j3[i,j]=1;
           j3[j,i]=1;
           v=j3@kmat;
           k=k+1;
           vp=v||j(nrow(v),1,k);
           if k=1 then vps=vp;
           else vps=vps//vp;
        end;
     end;
     create vps from vps;
     append from vps;
     close vps;
quit;
libname x '.';
data x.ldata;
     set vps (rename=(col2092=parm));
     by parm;
     if first.parm then row=1;
     else row+1;
run;
```

By default variance components can have lower boundary constraint of 0, in cases this is not so one can use the PARMS statement, e.g., for the multivariate example as

```
parms / lowerb=1e-4,.,1e-4,.,.,1e-4,1e-4,1e-4,1e-4;
```

which informs the procedure to use default values (.) as lower boundaries for the the covariances while 0.0001 for the variances.

BUGS[12]

The BUGS (Bayesian inference Using Gibbs Sampling) project is concerned with flexible software for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) methods. Initiatives have been made to make it available to Windows and other platforms and link with the *R* project.

Analysis of family data has been described but we could not access the source code associated with [26]. According to [25], the models we have described can be straightforwardly implemented in software such as *WinBUGS* but the implementation still requires founders precede their offsprings though it is not necessary to do so with *R* in general and *SAS* for the examples used here. It remains to explore the possibility to combine ideas in those implementations. However, analysis via WINBUGS is expected to be slower.

[12]See `http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`

Table 1: Nominal significance according to *VEGFC*

	Q1	Q2	AFFECTED
Significance level	Power	Type I error	Power
.05	.989	.060	.880
.01	.907	.016	.730
.001	.665	0	.555
.0001	.412	0	.420
.00001	.225	0	.305
.000001	.104	0	.200

Table 2: Q1 and *VEGFC* under a linear model

Model/parameter	Estimate	SE	z/t†	-2 Res log likelihood	AIC
kinship				1789.5	1793.5
σ^2_P	0.5488	0.08262	6.64		
SEX	-0.2379	0.04614	-5.16		
AGE	0.01014	0.001345	7.54		
SMOKE	0.36894	0.07280	5.07		
kinship+IBD				1775.2	1781.2
σ^2_P	0.4157	0.08713	4.77		
σ^2_M	0.1076	0.03846	2.80		
SEX	-0.2488	0.04542	-5.48		
AGE	0.01044	0.001334	7.82		
SMOKE	0.3821	0.07181	5.32		

† z is for variance components while t for fixed effects.

Table 3: AFFECTED and *VEGFC* under a logistic model

Model/parameter	Estimate	SE	t	-2 Log pseudo-likelihood
kinship				3434.4
σ^2_P	1.3170	0.4376		
SEX	-0.00822	0.2042	-0.04	
AGE	0.07181	0.006047	11.87	
SMOKE	0.9098	0.2285	3.98	
kinship+IBD				3445.7
σ^2_P	0.6918	0.5989		
σ^2_M	0.4868	0.3698		
SEX	0.006923	0.2048	0.03	
AGE	0.07211	0.006114	11.79	
SMOKE	0.9429	0.2290	4.12	

Table 4: AFFECTED and *VEGFC* under a Cox model

Model/parameter	Estimate	SE	z	Integrated/Penalized likelihoods†
kinship				-998.8/-980.6
σ^2_P	0.2073			
SEX	0.05267	0.1541	0.34	
SMOKE	0.5000	0.1622	3.08	
kinship+IBD				-996.1/-967.3
σ^2_P	0.002690			
σ^2_M	0.3615			
SEX	0.07146	0.1603	0.43	
SMOKE	0.5560	0.1696	3.28	

† The log likelihood under the null is -1003.9

Table 5: Q1, Q2 and Q4 under a multivariate polygenic model

	Estimate	SE	Log likelihood
Linear Coefficients			-1393.867
c1	0.565	0.108	
c2	0.531	0.109	
c3	0.526	0.109	
sex	-0.005	0.043	
age	-0.013	0.001	
smoke	-0.019	0.051	
Variance Coefficients			
σ_{11}	4.219	0.227	
σ_{12}	-0.103	0.166	
σ_{22}	4.542	0.244	
σ_{31}	0.601	0.178	
σ_{32}	-0.108	0.183	
σ_{33}	5.115	0.275	

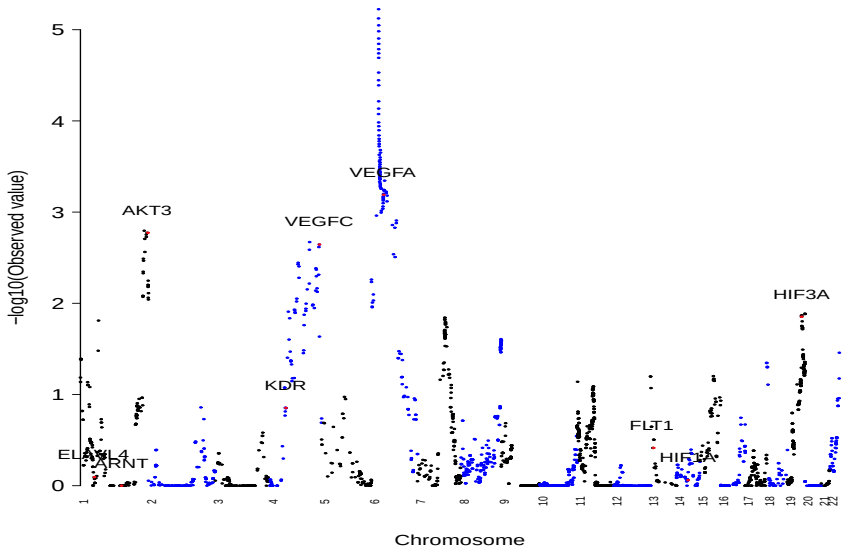

Figure 1: Manhattan plot of Q1 and IBD information where the true loci are highlighted