David Golan and Saharon Rosset

Mixed Models for Case-Control Genome-Wide Association Studies: Major Challenges and Partial Solutions

Introduction – Linear mixed models overview

Background and intuition. A linear mixed model (LMM) is an extension of the standard linear regression model, wherein the variables are divided into two groups: fixed effects and random effects. Fixed effects are modelled as parameters, i.e. fixed, but unknown, quantities, while random effects are modelled as being drawn from a random distribution – typically a Gaussian distribution with mean zero and an unknown variance. Intuitively, this formulation allows accounting for the random effects, while not specifically estimating the value of each random effect. This is done by integrating the random effects out, resulting in a linear regression model with a non-identity covariance matrix (where samples which have similar random-effects values have stronger correlations and vice verse).

To illustrate this idea, consider the following toy example: A researcher is interested in modelling some biometric measure as a function of time (e.g. weight of children). For that purpose she measures the weight of several toddlers once a month over a year, starting at the age of two. However, it is obvious that each subject has a different starting point at time $t = 0$. When using basic linear regression, accounting for this heterogeneity requires adding an indicator variable per individual, dramatically inflating the number of parameters in the model and increasing the standard errors of each estimate. The idea behind LMMs is to realize that the between-individuals differences are not the focus of interest of this specific study, and so it is counter-productive to "waste" information on estimating the per-individual effect. Instead, each individual's intercept term is treated as being drawn from some distribution. Treating the intercept as random captures the between-individual differences without requiring per-individual parameters, thus reducing the overall number of parameters while increasing accuracy and power for the parameters of interest. Instead of adding one parameter for each individual, we add only a single parameter – the variance of the intercept. This variance captures the extent of the heterogeneity. The key is that the observations of each individual share the value of the random effect, and so, after integrating it out, they become highly correlated: If a child is seen as high-weight at the first time point, we expect her to continue being relatively high weight at following time points.

Formulation and estimation. To put these ideas in a more rigorous form, we write down the linear mixed effect model:

$$Y = X\beta + Zu + e,$$

where $Y_{1\times n}$, $X_{p\times n}$, $\beta_{1\times p}$ and $e_{1\times n}$ take their usual roles as in a standard linear regression model (outcome, intercept, covariates, regression coefficients and noise drawn from a $N(0, \sigma_e^2)$, respectively). The important addition is Zu where $Z_{m\times n}$ is another set of covariates, similar to X but for which we are not interested in specific estimates of the coefficients, and $u_{1\times m}$ which are the associated random effects drawn from $N(0, \sigma_u)$. This model can be expressed as a specific distribution of Y:

$$Y \mid u \sim N(\alpha + X\beta + Zu, I_{n\times n}\sigma_e^2).$$

Note how we condition on u but not on β, because only the former is a random variable. Assuming the covariates in Z were standardized to have mean 0 and variance 1, and integrating u out, we get the unconditional distribution of Y:

$$Y \sim MVN(X\beta, ZZ^T\sigma_u^2 + I_{n\times n}\sigma_e^2).$$

Notice how Z no longer affects the mean of the distribution. Instead, it appears as a component of the covariance matrix. This is the reason why this component of the model is also often referred to as a "variance component". Another useful representation of this model is to define $\sigma_g^2 = m\sigma_u^2$, and $G = \frac{1}{m}ZZ^T$, and replace the Zu term by $g \sim MVN(0, G\sigma_g^2)$ so that the model can be written as:

$$Y = \alpha + X\beta + g + e.$$

This representation can be useful when the specific values of Z are unknown, but G can somehow be calculated or estimated using external data, as will be the focus of this chapter. LMMs can be naturally extended to accommodate several variance components, each with a different variance parameter.

The primary uses of LMMs are:

Estimation and testing of fixed and random effects. Once the distribution of Y is specified, one can proceed to estimate the fixed effects (β) and the variance of each group of random effects using maximum likelihood approaches. Specifically, the common approach for estimating variance components is known as restricted maximum likelihood (REML). This is then often used to test hypotheses about either the fixed or random effects, most commonly of the form $H_0 : \beta = 0$.

Prediction Given x and z, the covariates associated with a newly observed individual for whom the outcome y is unknown, we would like to predict y with the greatest possible accuracy. For a simple linear regression model, the answer is simply taking the covariate vector x and multiplying it by the estimated coefficients $\hat{\beta}$ (and adding the intercept): $\hat{y} = \hat{\alpha} + x^T\hat{\beta}$. This practice yields unbiased estimates. However, when attempting prediction in the LMM case, things are not so simple. One could adopt the same approach, but since the effects of the random components are not directly estimated, the vector of covariates z will not contribute directly to the predicted value of y, and will only affect the variance of the prediction, resulting in an unbiased but inefficient estimate. Instead, one can use the correlation between the realized values of Zu, to attempt a better guess at the realization of zu for the new sample. This is achieved by computing the conditional distribution of the outcome of the new sample conditional on the full dataset, by using the following property of the multivariate normal distribution. Assume we sampled n individuals, but the outcome for the i'th individual is unknown. The conditional distribution of y_i given the rest of the outcomes (y_{-i}) is given by:

$$y_i|y_{-i} \sim N\big(\alpha + x_n^T\beta + \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}(y_{-i} - X\beta_{-i}), \Sigma_{n,-n}\Sigma_{-i,-i}^{-1}\Sigma{-i,i}\big), \tag{1}$$

where $\Sigma = ZZ^T\sigma_u^2 + I\sigma_e^2$, and positive/negative indices indicate the extraction/removal of rows or columns, respectively. Intuitively, we use information from different samples that have a high correlation with the new sample, to improve its prediction accuracy. The practice of using the conditional distribution is known as BLUP (Best linear Unbiased Predictor).

LMMs in genetics.

Linear mixed models have been extensively used in genetics, and in particular have been popular in the animal-breeding literature and practice for many years [1]. Historically, the major focus of interest was breeding selection – choosing which sires and dames to mate in order to improve a specific trait, or phenotype, in the next generation (e.g. dairy yield). The phenotype is the outcome y, the covariates X are measured and observed quantities (e.g. nutrition type, the farm where the animal was raised etc.) and the Z matrix is the matrix of genetic variants. In many cases, the phenotype is influenced by many such variants, a situation which is typically referred to as a highly-polygenic phenotype or a complex phenotype. In such situations there are simply too many genetic variants to be included in the model, as they dramatically outnumber the samples. Moreover, even if there were enough samples to allow including all the genetic variants, up until roughly 20 years ago, actually measuring – or genotyping – these variants was either impossible or prohibitively expensive. So the problem posed by such genetic studies is double: the variables are both too many and unobserved. The solution put forward by LMMs is to skip the representation of the model that includes Z, directly to the variance components representation, and use pedigree information to estimate G, and plug the estimate into the equations.

The existence of pedigree information that allows to estimate the correlation accurately is quite unique to this domain, and explains the unique success of LMMs in the context of animal breeding long before genotyping became possible and affordable. In short, the DNA of an offspring is a mix of DNA segments from both parents. On average, each parent contributes 50% of the DNA (note that this is true in expectation only). Hence, when we look at the realizations of the random vector g for parent-offspring pairs, they have a correlation of 0.5 (under additivity assumptions on the genetic architecture, which we will not delve

into here). Similarly, for any known relationship, one can compute the expected correlation: grandparent-grandchild pairs would have a correlation of 0.25, as would avuncular pairs, while second cousins would have a correlation of $\frac{1}{8}$, and unrelated individuals would have a zero correlation. For these reasons g is typically referred to as the "genetic" effect, while G is referred to as the "kinship", or "genetic relationship" matrix. Similarly, e is referred to as the "environmental" effect (which combines unmeasured effects and random noise). Thus animal breeders could utilize their pedigree information to compute an (approximate) genetic correlation matrix $\hat{G}$, and use it for prediction with BLUP, thus predicting which breeding choices would result in the best (predicted) yield.

Applications extend beyond prediction. One could be interested in estimating fixed effects (e.g. nutrition type, local weather) while controlling for genetic effects. For example one could ask whether an observed difference in yield is due to controlable living conditions or accumulated genetic differences. To answer such questions one needs to control the genetic differences in different farms. Again, this is achieved by using LMM with $\hat{G}$, thus controlling for the genetic effect, and estimating or testing the fixed effects.

Lastly, it is often of interest to estimate $h^2 = \sigma_g^2/var(y)$, which is referred to as the (narrow-sense) heritability, the fraction of phenotypic variance which is explained by genetics. This magnitude is specifically important in the context of animal breeding, as it signifies how effective would the breeding actually be: high heritability implies a considerable genetic basis of the trait, so selective breeding would be very effective. Zero heritability means no genetic basis, so random breeding would be just as effective. This is captured by the famous *breeder's equation*: $R = h^2 S$ where R is the expected difference in phenotype between the previous and current generations, h^2 is the heritability, and S is the difference between the average phenotype in the population and the average phenotype of the selected parents [2]. As the selection progresses and the phenotype improves, the relevant variants become more and more frequent in the population, until they are fixated, thus reducing the role of genetic diversity and reducing the heritability.

Moving to GWAS. As genotyping technologies emerged and prices plummeted, genome-wide association studies (GWAS) became more and more affordable and a major dogma for human genetics research. When performing GWAS, one genotypes thousands of individuals at hundreds of thousands of genomic loci and scans the genome for loci which are significantly associated with the measured phenotype. Several major differences exist between the human-centric GWAS and the animal breeding practice. First, the typical goal of the GWAS is not improved breeding, but rather identifying loci which harbor causative variants (hoping to implicate genes near these loci, thus leading to better understanding of a disease and novel therapeutics). Second, when dealing with humans, one has less control over the design of the study compared to cattle. The vast genetic heterogeneity in humans is mostly undocumented, and pedigrees are not as carefully maintained for human populations as they are in the animal breeding business. Lastly, unlike carefully bred animals, human populations are structured, where geographic, ethnic and other factors are correlated with genetic differences. Many of these issues can be addressed by applying LMMs to GWAS.

The use of LMMs in GWAS was pioneered by [3, 4]. They consider the problem of association tests of a polygenic phenotype, in a highly structured population (wild-type and domesticated mice). They note that testing the association of a single SNP (say, x_1) involves assuming a univariate model:

$$y_i = \beta_1 x_{i1} + e_i,$$

to estimate β_1 and test its significance, while in fact the true model is polygenic, and so the fitted model should be:

$$y_i = u_1 x_{i1} + \sum_{k>1} u_k x_{ik} + \eta_i.$$

Due to the population structure, many of the SNPs have different frequencies in each population, resulting in a considerable dependence between them. Running a univariate scheme ignores the effect of the other SNPs, effectively modelling them as part of the error, so we have $e_i = \sum_{k>1} u_k x_{ik} + \eta_i$. Since the SNPs are correlated, due to the major genome-wide differences in allele frequencies between the populations, this results in a correlation between the tested SNP and the noise term, as well as between the noise of same-population samples, resulting in inflated type-1 error rates. To solve this problem, they adopt the LMM framework and treat the SNPs which are not directly tested as random-effects. The effects u_k are treated as identically

and independently distributed random variables drawn from a distribution with mean 0 and variance σ_u^2. Higher values of σ_u^2 imply larger effects. Hence, there is an additional "genetic effect", $g_i = \sum_{k>1} u_k x_{ik}$, with variance σ_g^2, and the model becomes:

$$y_i = \beta_1 x_{i1} + g_i + e_i,$$

where the genetic effects are positively correlated between genetically similar individuals and vice-versa. This is exactly the same LMM formulation used before, with the major difference being the way that G is obtained. Instead of using pedigree data, the correlation between any two individuals is estimated using the observed genotypes (after centering and scaling):

$$G_{ij} = \text{cor}(g_i, g_j) = \frac{1}{m} \sum_{k=1}^{m} \frac{(x_{ik} - 2f_k)(x_{jk} - 2f_k)}{2f_k(1 - f_k)}.$$

Note that this is only an estimate of the true correlation for several reasons. First, the true causal SNPs affecting the phenotype are not necessarily genotyped. Second, the correlation is estimated using all genotyped SNPs, thus including many non-causal SNPs (adding noise to the estimate). However, the estimated correlation is an unbiased estimate of the true kinship [5]. Several works focus on improving the estimation of G, either by modelling LD [6], accounting for cryptic relatedness [7] or trying to pick out only the causal SNPs (or the SNPs which best tag those SNPs) [8]. However, for our discussion, we treat the estimated G as the true G, and note that this is an interesting area for future research.

While this approach was first applied to mouse GWAS, the methodology as described is well suited to human GWAS as well, and indeed the LMM approach to GWAS gained popularity as it was repeatedly shown that accounting for the subtle genetic similarities between individuals in a GWAS increases power and reduces type-1 error rates [9, 10]. In addition, it was shown that accounting for the genetic correlation by using LMMs is appropriate for controlling for population structure (which is a common problem in human GWAS), as well as for cryptic relatedness, and that LMMs outperform the previously preferred principal component analysis (PCA) approach in addressing these issues [10]. At the same time, methods for efficient estimation emerged and enabled applying these methods to progressively larger GWAS (e.g. [4, 11, 12]). The state of the art methods in terms of speed and memory requirements are BOLT-LMM [13] and BOLT-REML [14], which are the first to provide a method to estimate an LMM that is not cubic in the number of individuals, allowing their application to GWAS as large as $50,000$ individuals.

At the same time, LMMs were used to address a different burning question in human genetics: the problem of the missing heritability [15]. Despite the clear evidence from twin and family studies that many traits and diseases are highly genetic (e.g. height, type-1 and type-2 diabetes, multiple sclerosis, schizophrenia and more), GWAS were only able to identify a handful of variants associated with these traits and diseases, and these variants accounted for only a fraction of the heritability expected from twin studies. Specifically for height, commonly cited numbers are 80% heritability from twin studies, and $< 20\%$ heritability explained by discovered genetic variants. One leading theory explaining this gap was that these phenotypes are driven by a large number of variants with small effects, and that GWAS are under-powered to detect most of the causal variants. Yang et al. [16] used the LMM framework to address this question: assuming that the effect of each variant is drawn from a Gaussian distribution, the model of the phenotype is an LMM, and the estimates of the variance components can be used to estimate the heritability: $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$. Their seminal paper showed that the fraction of heritability of height explained by a set of $500K$ genotyped SNPs is considerably larger than the heritability explained by the genome-wide significant hits alone, suggesting that height is indeed driven, to the large part, by a plethora of common variants with small effects.

Their work sparked a wave of follow-up work adopting and adapting the LMM framework for various heritability-related goals. One popular approach is to partition the SNPs to l groups, and compute a correlation matrix G_i for each group. The variances $\sigma_{g1}^2, ..., \sigma_{gl}^2$ are estimated simultaneously, capturing the heritability explained by each group. This practice was used, for example, to partition the heritability by chromosomes [17], by *cis/trans* effects [18] or by functional annotations [19].

The mixed modelling challenge for case-control GWAS.

So far, we discussed LMMs in the context of studies where sampling is random and the phenotype in question is quantitative. However, in many scenarios one or both of these assumptions does not hold. In particular, the most important use of GWAS is for studying human diseases, where the phenotype is usually binary (affected/healthy), and (relative) rarity of the disease requires the adoption of case-control sampling schemes. In this setting, the problems addressed by using LMMs are still present, but application of LMMs is much less straight forward. In this section we describe how LMMs are extended in these situations and what problems arise.

Modelling discrete phenotypes. Quite often, the phenotype in question is discrete rather than quantitative, specifically binary like disease phenotypes. In traditional regression settings (fixed-effects modelling), this is often addressed by moving from linear models to the generalized linear models (GLM, [20]) framework, where instead of assuming $Y = X\beta + e$, we move to assuming $f(P(Y = 1|X)) = X\beta$, for an appropriate link function f. The most common approaches are the probit and logit link functions:

$$\begin{aligned} \text{Probit}: \quad & f(p) = \Phi^{-1}(1-p) \\ \text{Logit}: \quad & f(p) = \log(\frac{p}{1-p}), \end{aligned}$$

where Φ is the standard normal cumulative distribution function. The well known logistic regression approach is simply a GLM with the logit link.

In the generalized linear mixed models (GLMM) literature in statistics, the binary situation is often addressed by the same approach that generalizes the linear mixed model through use of a link function [21]:

$$f(P(Y = 1|X, Z, u)) = X\beta + Zu,$$

where we still assume that $u_i \sim N(0, \sigma_u^2)$. While this model is well defined, estimation and inference in this model is a much more complex task than in the standard LMM setting, since the normally distributed vector is now unobserved. If the link function used is probit, the setting falls under the more general category of Gaussian process regression and classification, which has been widely studied in the machine learning literature [22]. The state of the art solutions developed in this area, including expectation-propagation (EP, [22]) and Markov-Chain Monte Carlo approaches (MCMC, [23]), can offer practical solutions to GWAS-sized problems. Such solutions can address all aspects of the GWAS problem discussed above: fixed effects estimation/testing, variance components (heritability) estimation, and prediction.

In the context of genetics, Wright [24] put forward the liability threshold model (LTM) to address the issue of binary phenotypes. In the mixed model version of LTM, one assumes the existence of a latent phenotype vector $L = X\beta + Zu + e$, which follows the same normal assumptions of the standard LMM:

$$L \mid u \sim N(X\beta, ZZ^t\sigma_u^2 + I\sigma_e^2), \tag{2}$$

$$Y = \mathbb{I}(L > T). \tag{3}$$

The observed phenotype is determined by the latent phenotype crossing or not crossing a threshold T, set such that the probability of crossing T is exactly the prevalence of the disease in the population. In the GLMM context, LTM is simply a GLMM with a probit link function, since it gives:

$$P(Y = 1|X, Z, u) = P(L > T|X, Z, u) = 1 - \Phi(\frac{T - (X\beta + Zu)}{\sigma_e}).$$

An important development in the analysis of genetic data under LTM was the presentation by Dempster and Lerner [25] of a mathematical connection between the heritability on the observed (Y) scale, i.e., $h_o^2 = \text{cov}(Y, g)/\text{var}(Y)$, and the heritability on the liability scale, i.e. $h_l^2 = \frac{\text{cov}(L,g)}{\text{var}(L)} = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$:

$$h_l^2 = \frac{K(1-K)}{\varphi(T)^2} h_o^2,$$

where K is the prevalance of the phenotype, T is the liability threshold, and φ is the density of the standard Gaussian distribution. This implies that any reasonable estimate of the heritability on the observed scale can be transformed to an estimate of the "true" heritability. One such estimate can be obtained using a method known as Haseman-Elston regression [26], whereby the $O(n^2)$ products of binary phenotypes $Y_i \times Y_j$ are regressed on the kinship values G_{ij}. Thus within the LTM we have two competing approaches for estimating heritability: via maximum likelihood GLMM or using the "moment-based" regression estimator, followed by the correction of Dempster and Lerner.

A different approach which is often practiced in the genetics community is to ignore the binary nature of the phenotype and apply regular LMM's to the data, as if Y were a quantitative, normally distributed phenotype. While this practice is clearly unsubstantiated from a probabilistic perspective, it leads in practice to useful methods, especially for prediction. It can be thought of as using Y as a surrogate the unobserved L.

Case-control sampling. So far we have assumed that samples are randomly drawn from the population. However, quite often this is not the case, with the prime example being case-control studies. In case-control studies, one is interested in studying a (relatively) rare outcome, e.g. a disease which affects $< 1\%$ of the population. In such scenarios, a random sample from the population would include a very small fraction of affected individuals (cases), and so the efficiency of the statistical analysis would be compromised. The common strategy to address this problem is to make an effort to sample more cases than their random share in the population, e.g. by recruiting cases using ads or at clinics and hospitals directly, and separately sampling healthy controls from a "similar" population (to mitigate the effects of population structure and other confounders). The resulting sample is then subjected to GWAS.

This seemingly innocent sampling approach has the potential to wreak havoc in the statistical analysis of the resulting data, which has been the subject of a long and storied line of work in the statistics literature [27, 28, 29, 30, 31, 32]. Here we concentrate on some of the aspects of this area which are most relevant to mixed models analysis of case-control GWAS.

It is first important to note one situation where case-control sampling can be mostly ignored. The famous result by Prentice and Pyke [28], building on earlier work by Anderson [27], shows that if we are assuming that $P(Y|X)$ follows a fixed-effects GLM with a logit link function in the population, with an intercept:

$$P(Y = 1|X) = \frac{\exp(\alpha_0 + X\beta)}{1 + \exp(\alpha_0 + X\beta)},$$

then the parameters β can be estimated from a case-control sample from the population, and inference carried out on them, while ignoring the fact that such sampling has taken place. This result is extremely useful and widely used, but it represents the exception rather than the norm in analyzing case-control data. Even if no random effects are assumed, but we move away from the logit link (say, to probit), important aspects of this result no longer hold. Once we also include random effects and move to a mixed model (say, using LTM), we can no longer ignore the case-control sampling and hope to obtain meaningful results.

More concretely, the typical probabilistic mixed model in case-control GWAS still assumes that the GLMM (specifically, probit GLMM through the LTM) holds at the population level without the case-control sampling. This is often justified by a Central Limit Theorem (CLT) type of argument, for example that the genetic effect is the sum of many small effects drawn from the same distribution and therefore follows a normal distribution by virtue of the CLT. Thus, in the population we still assume the model (2,3). Once we move to case-control sampling, then in the sampled population the probabilistic setting is significantly changed. Formally, we add a sampling vector S which encodes the sampling process, i.e., in case-control sampling the liability is sampled not from $P(L)$, but from $P(L \mid S = 1)$, and similarly for all other quantities. In this setting, it is easy to see that [33]:

1. The distributions of the liability L, the genetic effects vector g and the environmental effects vector e are no longer normally distributed
2. The genetic effect and the error vector e are no longer independent, because cases are oversampled, and these tend to have both g and e high, so that L passes the threshold T.

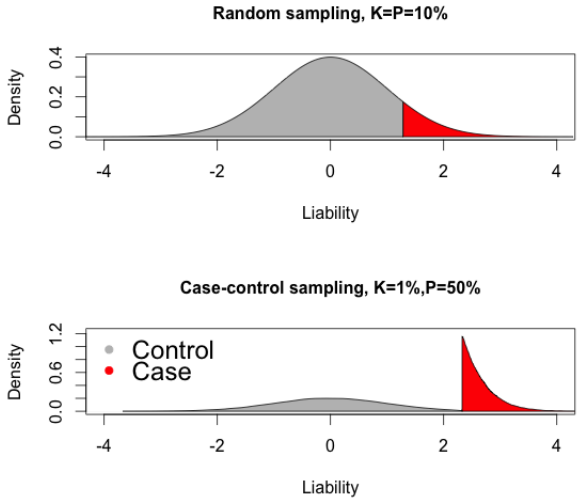

FIGURE 1
Comparison of the distribution of the liability in a random sample and a case-control sample of a discrete phenotype. We simulated two studies using the liability threshold model: a random sampling study of a relatively common disease phenotype ($K = 10\%$, top panel), where the liability follows a normal distribution as expected; and a balanced ($P = 50\%$) case-control study where the phenotype is relatively rare ($K = 1\%$, bottom panel). In the latter case, the over-sampling of cases results in an oversampling of the right tail of the distribution, and the distribution is obviously no longer normal.

These effects are demonstrated in Figures 1 and 2.

Maximum-likelihood based statistical modelling and estimation of GLMMs with case-control sampling in this setting is to our knowledge an unsolved problem. The Gaussian-process literature discussed previously does not offer EP or MCMC solutions to this problem, and we are not aware of other work offering practical computational solutions to this problem (which can be thought of as a high-dimensional integration problem).

However, as discussed before, using LMMs to analyze GWAS offers a unique combination of major benefits, including their ability to control for population structure and model cumulative effect of many small genetic effects, and the efficient computational tools that exist for computing LMMs. Because of this, many researchers have sought to analyze case-control studies by applying standard (normal) LMMs to the data, and using the results in association testing [11], heritabililty estimation [34], and genetic risk prediction [35]. In some of these cases, the results of LMM were "corrected" to account for case-control sampling [34].

The fundamental difficulty in all these efforts is that the probabilistic model assumed by the LMM does not hold at all: As just demonstrated, the distributions of the elements of the LTM (liability, genetic effect and environmental effect) and their correlation structure are fundamentally influenced by the sampling. Not surprisingly, we are not aware of any theory that can describe the distribution of estimates derived by applying standard LMMs to case-control GWAS, and it seems unlikely that such theory is possible. Consequently, we do not believe that tasks concentrated on statistical estimation, inference and testing in case-control GWAS (like association testing and heritability estimation) should be based on LMMs. A slightly different case is presented by genetic risk prediction, where the goal is to predict the phenotype of new individuals, based on their genotypes. Since this task carries with it an objective measure of performance that does not need to be tied to probabilistic inference, methods based on LMMs can be justified. However, they should be thought of as algorithmic predictive modelling approaches, rather than a well founded probabilistic model of case-control GWAS.

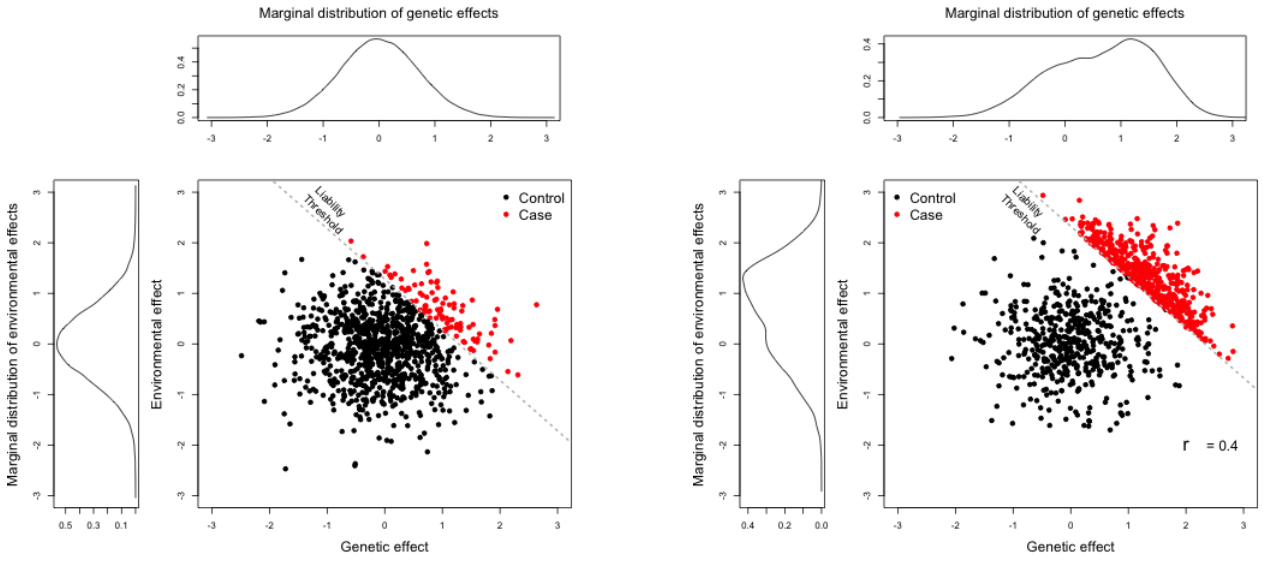

FIGURE 2

The effects of case-control sampling on the marginal and joint distribution of the genetic and environmental effects. We simulated genetic and environmental effects for unrelated individuals with $\sigma_g^2 = \sigma_e^2$ (so $h^2 = 0.5$). Phenotypes were determined using a liability threshold model without fixed effects using for either a common phenotype ($K = 10\%$, left panel) or a rare phenotype ($K = 1\%$, right panel). In the former scenario, random sampling was applied, while in the latter scenario, cases were oversampled to achieve a balanced study ($P = 50\%$). The joint distributions are illustrated in the middle panels while the marginal distributions of the random effects are illustrated in the side panels.

Methods for mixed modelling of case-control GWAS

After reviewing the general challenge, in this section we review some of the recent methods for addressing the major tasks in case-control GWAS under the mixed-models paradigm. We try to emphasize methods that take into account the complex probabilistic model in this setting and offer valid solutions (if partial and suboptimal, compared to the currently impractical alternative of solving the full case-control GLMM problem).

Association testing and estimation of fixed effects. Association testing is the main original intended use of GWAS (as expressed in the term itself), and naturally it has been extensively applied to case-control GWAS. Ultimately, association testing seeks to test a null hypothesis for the association between each variant genotyped in a GWAS (typically counting the number of minor alleles $0, 1, 2$) and the phenotype in question. The simplest approach treats this through standard univariate tests: Armitage's test for trend, Chi-square tests on contingency tables, likelihood-based tests (G test, score test) on logistic regression, etc. All of these tests are statistically valid for case-control GWAS (since under the null, the sampling does not affect the distribution of the statistic) and have been extensively used in GWAS (e.g. [36]).

Already in the early days of GWAS, the statistical genetics community came to the realization that population structure and linkage disequilibrium are critical aspects of the problem, and univariate tests which ignore them give results that are statistically valid but are often of little interest in the application. This is because the presence of the above factors, especially population structure, imply that many non-causative genetic variants will be significantly associated with the phenotype through their correlation with causative variants, and these can be spread throughout the genome. Early efforts to correct this problem (beyond traditional genomic control approaches [37]) concentrated on explicitly modelling structure through the use of principal components (PCs). These were then added as additional fixed effects to a regression model, or regressed out of both the genetic variants and the phenotype, before testing [38]. Under the assumption that a few PCs successfully capture population structure, this approach is reasonable. However, as previously described, the use of mixed models for controlling population and genetic structure has been demonstrated to be the most effective and general approach in many settings.

Thus there is an obvious interest in taking advantage of the mixed models conceptual framework in case-control GWAS as well. The first option is to apply LMMs to this problem "out of the box", ignoring both the discrete phenotype and the sampling (see, e.g., [11, 35]), assuming that the 0/1 phenotype follows the LMM normal distribution. This is difficult to justify as has been discussed, and indeed leads empirically to low power [10].

We are aware of two recent papers that made an effort to adapt the mixed model framework to association testing in case-control GWAS [39, 40]. The common idea to these two methods is to start from replacing the 0/1 case-control status by an estimate of the liability L , and doing so while taking into account the probabilistic structure (i.e., considering the distribution $P(L|S = 1)$).

The paper by Weissbrod et al. [39] takes advantage of the similarity between BLUP and ridge regression prediction (which are equivalent when no fixed effects are present), to formulate the liability estimation problem as a penalized probit regression problem:

$$\begin{aligned}\hat{L} &= Z\hat{u} + \epsilon \\ \hat{u} &= \arg\min_u \sum_{i=1,y_i=1}^{n} \log\left(\Phi(\frac{T - z_i^T u}{\sigma_e})\right) + \sum_{i=1,y_i=0}^{n} \log\left(1 - \Phi(\frac{T - z_i^T u}{\sigma_e})\right) + \frac{1}{2\sigma_u^2}\|u\|^2.\end{aligned}$$

This calculation assumes that the variances $\sigma_g^2 = m\sigma_u^2, \sigma_e^2$ are known in advance, and the resulting $\hat{L}$ is the maximum a-posteriori (MAP) estimate of $L|S = 1$. [39] then plug this $\hat{L}$ into a regular LMM to perform the association testing, and demonstrate that the resulting power is superior to that of standard LMMs or PC-based correction for structure. However, as demonstrated in Figures 1,2 there is no reason to assume that $\hat{L}$ (or indeed the true unobserved liability L) has a normal distribution under case-control sampling, hence the second part of their solution still fails to fully take the case-control sampling into account.

The paper by Hayeck et al. [40] takes a different approach to estimating the liability. They use the fact that given the phenotype vector Y, and all liabilities but one L_{-i}, the distribution of the missing liability is a truncated normal:

$$l_i \mid L_{-i}, Y_i = 1 \ \sim \ TN(\mu_i, \sigma_i^2, T, \infty),$$
$$l_i \mid L_{-i}, Y_i = 0 \ \sim \ TN(\mu_i, \sigma_i^2, -\infty, T),$$

where μ_i, σ_i^2 are the conditional mean and variance, calculated as in Eq. (1). This allows them to design a simple Gibbs sampling algorithm for generating random samples of "representative" liability vectors L for the case-control probabilistic model. These are averaged to calculate a "posterior mean" liability vector $\hat{L}$. This vector is then used as if it were a normally distributed LMM response in a score test of the null of no association for each genetic variant. Hence for this approach too, the second part fails to take the probabilistic structure into account. The superiority of the approach over standard LMMs and trend test in terms of power is demonstrated in both simulations and real data.

Beyond testing, actual estimation of the association parameters (fixed effects) is usually considered a by-product of the process. We are not aware of specific efforts to estimate fixed effects within mixed-model analysis of case-control GWAS. This is in contrast to the problem of estimating variance components and heritability, discussed next.

Estimating variance components (heritability). The first attempt to estimate the variance of the genetic random effect (i.e., the heritability) in the context of case-control GWAS was by Lee. et al. (2011) [34]. They describe a procedure in the spirit of [25]: First code the phenotype as a 0/1 variable, treat it as quantitative and apply a standard LMM method (in their case, REML as implemented in GCTA [41]). Then, apply a post-hoc correction to correct the errors and biases introduced by the fact that, in fact, the method applied was inappropriate. Specifically, Lee et al. obtain an "observed scale" heritability estimate $\hat{h}_o^2$, which is the heritability of the synthetic 0/1 phenotype, and transform it to the desired "liability scale" heritability using the following relationship:

$$\hat{h}_l^2 = \frac{K^2(1-K)^2}{P(1-P)\varphi(T)^2}\hat{h}_o^2,$$

where K is the prevalence of the disease in the population, and P the percentage of cases in the study (typically $P >> K$ in case-control studies).

While Lee's method has become extremely popular, evidence from both simulation studies and actual studies show that it, in fact, produces downwards-biased estimates [10, 33]. Strikingly, this bias appears to increase with sample size, as demonstrated by simulations in [33] and using real data in [14], who used down-sampling of a huge GWAS to demonstrate how the estimates decrease as the size of the sample increases.

Recently, Golan et al. (2014) [33] developed an alternative method that does not suffer from the same problems as the method of Lee et al. They adopt the moments-method approach of Haseman and Elston [26] to obtain estimates that are unbiased despite the complicated underlying probabilistic model. The basic idea is to look at the relationship between two correlations: the correlation between the phenotypes of pairs of individuals (phenotypic correlation) and the correlation between the genotypes of pairs of individuals (the genetic correlation). Higher heritability implies that high genetic correlation should yield high phenotypic correlation, and low heritability implies no such relationship. More formally, Golan et al. express the product of the phenotypes of any two individuals, as a function f of the true underlying heritability, the genetic correlation, and the fixed effects, where f itself depends on the actual design of the study (specifically P) and the properties of the disease (specifically K):

$$\mathbb{E}(y_i y_j) \mid S = 1; G_{ij}, h^2) = f(h^2, G_{ij}),$$

where the conditioning on $S = 1$ indicates the fact that both individuals were selected for the study. Note that we assume that the phenotypes are centered and scaled so $\mathbb{E}(y_i y_j) = \mathrm{cor}(y_i, y_j)$. Next, f is approximated using its Taylor series approximation:

$$f(h^2, G_{ij}) = a_0 + a_1 G_{ij} h^2 + \mathcal{O}(G_{ij}^2).$$

Since individuals in the the GWAS study are typically unrelated, the values of G_{ij} are relatively small, so the first order approximation is satisfactory, resulting in an (approximated) linear relationship between the phenotypic correlation and the genetic correlation. In this situation, one can use linear regression to estimate the slope $a_1 h^2$ (by regressing the products $y_i y_j$ onto G_{ij} for all pairs $i \neq j$). To obtain an estimate of the heritability, all that is left is to compute the constant a_1. Golan et al. show how this can be done for various study designs (e.g. case-control and extreme phenotype sampling). Importantly, the computation explicitly involves the conditioning on the selection variable $S = 1$, so it accounts the effects of the non-random selection. The resulting estimates are unbiased (by virtue of being first moments estimators) and fast to compute ($O(n^2)$ instead of the usual $O(n^3)$ of most REML based methods). The method is named PCGC (for regressing phenoytpe correlations on genetic correlations) and a fast and memory-efficient implementation of PCGC regression can be found in the software reference list.

The application of PCGC regression to a wide range of GWAS in [33] demonstrated that the fraction of heritability explained by common variants is larger than estimated by the method of Lee et al. (for example, the estimated heritability of multiple sclerosis explained by common variants increased from 30% [42] to 45% [33] using the same data), and recent applications of PCGC regression for other phenotypes show similar results [14, 43]. Importantly, a recent paper used a huge GWAS of schizophrenia (involving $50,000$ individuals) to show that estimates using the method of Lee et al. method indeed decrease as the sample size increases, and that PCGC regression yields the correct estimate (i.e. a similar estimate to the estimates obtained when applying Lee et al.'s method to very small subsets of the data which have a very small bias due to their size).

Prediction The prediction problem is essentially different from the problems of estimating fixed or random effects. For these statistical inference problems, one is interested in unraveling some "ground truth" (the true heritability of a disease or the true effect size of a SNP), or making a scientific discovery (identifying a novel causal locus). In contrast, the prediction problem comes with its objective and measurable metric of success – predictive accuracy. In this case, applying methods which are not theoretically justified, but yield good performance is legitimate, as evidenced by the popularity of the application of out-of-the-box machine-learning methods such as support vector machines or elastic-nets for phenotype prediction (e.g. [44]). However, machine learning methods typically make no assumptions regarding the data, and take as input only a feature matrix (the genotyped SNPs) and an outcome vector (the phenotype). In contrast, GWAS in general, and case-control GWAS in particular, do have several unique characteristics: a highly polygenic nature of many phenotypes; a unique structure of correlations between the SNPs (linkage disequilibrium); the existence of population structure, which is captured by the correlation matrix G; and, of course, the artifacts introduced by the non-random sampling scheme in case-control studies. As discussed earlier, LMMs are particularly suited to take advantage of these features in the case of randomly sampled phenotypes, and typically outperform other methods, including "simple" classifiers which are based only on genome-wide significant SNPs (e.g., [45]). Given these favorable results, it is only natural to apply LMMs to the problem of prediction using a case-control GWAS as reference panel.

One popular approach is to code the binary phenotype as 0/1, and use standard LMM methods for prediction (e.g. using BLUP, or its recent extension multiBLUP [35] on the coded phenotype). The application of LMM-based methods aims to utilize their advantages for improved prediction. However, the same logic implies that a BLUP-like method that accounts for the quirks introduced by case-control sampling, should out-perform naive BLUP methods as it takes full advantages of the unique features of the case-control GWAS problem, namely, assumes a highly additive model, captures population structure, *and* accounts for the non-random sampling.

This intuition is captured by GeRSI – a method for genetic risk score inference which accounts for case-control sampling [46]. Here the authors find the conditional distributions of the genetic and environmental effects $e_i \mid g, e_{-i}, Y_i$ and $g_i \mid g_{-i}, e, Y_i$, which turn out to be truncated normal distributions, similarly to the conditional distribution of the liabilities in [40] described earlier. Once the conditional distributions are specified, Gibbs sampling is used to sample the posterior distribution of the genetic effect of an individual with an unknown phenotype, and these samples can be used to compute the posterior risk prediction (intuitively, higher posterior value of the genetic effect translates to higher risk). Simulations and application to real data show that GeRSI outperforms its BLUP equivalent.

Conclusion

The problem of mixed-modelling analysis of case-control studies in general, and case-control GWAS in particular, is unique in its combination of high importance and popularity, extreme difficulty, and paucity of computationally effective and statistically valid approaches. Indeed, while several of the approaches we presented here offer statistically valid solutions to specific aspects, we are aware of no fully valid approaches based on maximum-likelihood principles or Bayesian principles for estimation or testing in case-control GWAS. An intriguing question is whether this is because this problem is simply too difficult from a computational and statistical perspective, or whether it is a matter of getting the right communities and capabilities involved. In particular, the Gaussian processes literature does offer efficient solutions to GWAS-sized problems with binary phenotypes and natural sampling [22, 23]. If these EP and MCMC approaches can be adapted to dealing with case-control sampling, they may present an important opportunity. We note that recent research efforts in our group have been focused on this direction, and we are hopeful that a solution may be found.

One common critique of the mixed effects approach is that while many phenotypes are considered to be highly polygenic, it is not reasonable that all of the SNPs have identically distributed non-zero effects. Several methods try to address this issue by introducing an indicator variable for every SNP, indicating whether the SNP has a non-zero effect, and another parameter p which is the proportion of causal SNPs. Then, p can be jointly estimated with the other parameters of the model using MCMC-based methods [8, 47]. These models were recently extended to allow for a richer distribution of the effect sizes of SNPs, jointly modelling several scales of effect sizes [48]. These models are promising as they allow for a more realistic modelling of the genetic architecture and yield posterior probabilities of causality per SNP, as well as an overall estimate of the proportion of causal SNPs. While some efforts were made to modify these models to address some of the issues discussed here [49], we still view the problem of extending these approaches to account for case-control sampling in a way which is scalable for large GWAS as an open problem of great interest and potential importance.

Software reference list

GCTA A software package [41] containing an implementation of BLUP for standard LMM and an implementation of the biased heritability estimation method of Lee et al. (2011) [34], as well as many other useful functions for data handling (e.g. computing GRMs).
`http://cnsgenomics.com/software/gcta/`

PCGC Regression A memory-efficient implementation of the PCGC method [33] implemented by [50].
`https://github.com/gauravbhatia1/PCGCRegression/`

LTSOFT A software package implementing various liability-threshold related functions, including the computation of posterior liabilities of [40].
`http://www.hsph.harvard.edu/alkes-price/software/`

LEAP Implementation of the MAP liability for case-control GWAS of [39].
`https://github.com/omerwe/LEAP`

GeRSI Prediction of case-control status using LMMs which takes the case-control sampling scheme into account [46].
`https://sites.google.com/site/davidgolanshomepage/software/gersi`

GEMMA A software package which includes the Sparse Bayesian regression (BSLMM) models of [49].
`http://www.xzlab.org/software.html`

Bibliography

[1] Raphael A Mrode. *Linear models for the prediction of animal breeding values*. Cabi, 2014.

[2] Robert Plomin, JC DeFries, and GE McClearn. Behavior genetics: A primer. *WH Freeman and Company, New York*, 1990.

[3] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[4] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[5] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[6] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[7] Andrew Crossett, Ann B Lee, Lambertus Klei, Bernie Devlin, and Kathryn Roeder. Refining genetically inferred relationships using treelet covariance smoothing. *The annals of applied statistics*, 7(2):669, 2013.

[8] David Golan and Saharon Rosset. Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics*, 27(13):i317–i323, 2011.

[9] Jianming Yu, Gael Pressoir, William H. Briggs, Irie Vroh Bi, Masanori Yamasaki, John F. Doebley, Michael D. McMullen, Brandon S. Gaut, Dahlia M. Nielsen, James B. Holland, Stephen Kresovich, and Edward S. Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, 2 2006.

[10] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, 46(2):100–106, 2014.

[11] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

[12] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

[13] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 2015.

[14] Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature genetics*, 2015.

[15] Brendan Maher. The case of the missing heritability. *Nature*, 456(7218):18–21, 2008.

[16] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

[17] Jian Yang, Taeheon Lee, Jaemin Kim, Myeong-Chan Cho, Bok-Ghee Han, Jong-Young Lee, Hyun-Jeong Lee, Seoae Cho, and Heebal Kim. Ubiquitous polygenicity of human complex traits: genome-wide analysis of 49 traits in koreans. *PLoS Genet*, 9(3):e1003355, 2013.

[18] Alkes L Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genetics*, 7(2):e1001317, 2011.

[19] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552, 2014.

[20] P. McCullagh and J.A. Nelder. *Generalized linear models*, volume 37. Chapman & Hall/CRC, 1989.

[21] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

[22] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT press, 2005.

[23] Roger Frigola, Fredrik Lindsten, Thomas B Schön, and Carl Rasmussen. Bayesian inference and learning in gaussian process state-space models with particle mcmc. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3156–3164. Curran Associates, Inc., 2013.

[24] Sewall Wright. An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics*, 19(6):506, 1934.

[25] Everett R Dempster and I Michael Lerner. Heritability of threshold characters. *Genetics*, 35(2):212, 1950.

[26] J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavioral Genetics*, 2:3–19, 1972.

[27] J. A. Anderson. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972.

[28] Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

[29] A. J. Scott and C. J. Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84:57–71, 1997.

[30] A. J. Scott and C. J. Wild. Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, 96:3–27, 2001.

[31] N. Chatterjee and R. J. Carroll. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, 92:399–418, 1972.

[32] D. Clayton. Link functions in multi-locus genetic models: implications for testing, prediction, and interpretation. *Genetic Epidemiology*, 36:409–418, 2012.

[33] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

[34] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305, 2011.

[35] Doug Speed and David J Balding. Multiblup: improved snp-based prediction for complex traits. *Genome research*, 24(9):1550–1557, 2014.

[36] Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[37] B Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

[38] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[39] Omer Weissbrod, Christoph Lippert, Dan Geiger, and David Heckerman. Accurate liability estimation improves power in ascertained case-control studies. *Nature Methods*, 12:332–334, 2015.

[40] Tristan J. Hayeck, Noah A. Zaitlen, Po-Ru Loh, Bjarni Vilhjalmsson, Samuela Pollack, Alexander Gusev, Jian Yang, Guo-Bo Che, Michael E. Goddard, Peter M. Visscher, Nick Patterson, and Alkes L. Price. Mixed model with correction for case-control ascertainment increases association power. *American Journal of Human Genetics*, 96:720–730, 2015.

[41] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[42] S Hong Lee, Denise Harold, Dale R Nyholt, Michael E Goddard, Krina T Zondervan, Julie Williams, Grant W Montgomery, Naomi R Wray, and Peter M Visscher. Estimation and partitioning of polygenic variation captured by common snps for alzheimer's disease, multiple sclerosis and endometriosis. *Human molecular genetics*, 22(4):832–841, 2013.

[43] Long Jiang, Lu Liu, Yuyan Cheng, Yan Lin, Changbing Shen, Caihong Zhu, Sen Yang, Xianyong Yin, and Xuejun Zhang. More heritability probably captured by psoriasis genome-wide association study in han chinese. *Gene*, 573(1):46–49, 2015.

[44] Gad Abraham, Adam Kowalczyk, Justin Zobel, and Michael Inouye. Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic Epidemiology*, 37(2):184–195, 2013.

[45] Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, Pamela Sklar, Douglas M Ruderfer, Andrew McQuillin, Derek W Morris, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.

[46] David Golan and Saharon Rosset. Effective genetic-risk prediction using mixed models. *The American Journal of Human Genetics*, 95(4):383–393, 2014.

[47] Yongtao Guan, Matthew Stephens, et al. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780–1815, 2011.

[48] Gerhard Moser, Sang Hong Lee, Ben J Hayes, Michael E Goddard, Naomi R Wray, and Peter M Visscher. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. 2015.

[49] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet*, 9(2):e1003264, 2013.

[50] Gaurav Bhatia, Alexander Gusev, Po-Ru Loh, Bjarni J Vilhjálmsson, Stephan Ripke, Shaun Purcell, Eli Stahl, Mark Daly, Teresa R de Candia, Kenneth S Kendler, et al. Haplotypes of common snps can explain missing heritability of complex diseases. *bioRxiv*, page 022418, 2015.