Efficient multiple trait association and estimation of genetic correlation using the matrix-variate linear mixed-model

Nicholas A. Furlotte*, Eleazar Eskin†

February 24, 2015

*Department of Computer Science, University of California, Los Angeles
†Department of Computer Science, Department of Human Genetics, University of California, Los Angeles
Running Head: Multiple Trait Association

Key Words: Association Studies, Multivariate Analysis, Genetic Correlation

Corresponding Author:

Eleazar Eskin

Department of Computer Science
Department of Human Genetics
University of California, Los Angeles
Telephone: (310) 594-5112
E-mail: eeskin@cs.ucla.edu
Abstract

Multiple trait association mapping, in which multiple traits are used simultaneously in the identification of genetic variants affecting those traits, has recently attracted interest. One class of approaches for this problem builds on classical variance component methodology, utilizing a multi-trait version of a linear mixed-model. These approaches both increase power and provide insights into the genetic architecture of multiple traits. In particular, it is possible to estimate the genetic correlation which is a measure of the portion of the total correlation between traits that is due to additive genetic effects. Unfortunately, the practical utility of these methods is limited since they are computationally intractable for large sample sizes. In this paper, we introduce a reformulation of the multiple trait association mapping approach by defining the matrix-variate linear mixed model. Our approach reduces the computational time necessary to perform maximum-likelihood inference in a multiple trait model by utilizing a data transformation. By utilizing a well-studied human cohort, we show that our approach provides more than a 10-fold speed up, making multiple trait association feasible in a large population cohort on the genome-wide scale. We take advantage of the efficiency of our approach to analyze gene expression data. By decomposing gene coexpression into a genetic and environmental component, we show that our method provides fundamental insights into the nature of co-expressed genes. An implementation of this method is available at http://genetics.cs.ucla.edu/mvLMM.
AUTHOR SUMMARY

The ability to assess the association between genetic variations and traits has transformed the study of genetics. Genetic variations that are associated with traits give researchers an idea of what parts of the DNA might influence traits. The ability to accurately quantify the association between genetic and phenotypic variation has been enabled by both statistical and computational innovations that allow the computation of association for single traits across millions of genetic markers. Since many traits are correlated, it is natural to attempt to identify genetic variations that are associated with multiple traits simultaneously. In fact, the use of multiple traits has resulted in an increased ability to identify genetic markers that affect each of the traits. Although, techniques exists for evaluating the association between genetic variations and multiple traits, their use is computationally cumbersome and their applicability is limited by computational time. In this paper, we introduce a mathematical and computational construct that overcomes these computational limitations and enables for the first time the large scale analysis of the association between genetic variants and multiple correlated traits.

INTRODUCTION

Classically, genome-wide association studies have been carried out using single traits. However, it is well-known that genes often affect multiple traits, a phenomenon known as pleiotropy, and more recently, it has been shown that performing association mapping with multiple traits simultaneously may increase statistical power (Korol et al. 2001; Ferreira and Purcell 2009; Liu et al. 2009; Avery et al. 2011; Korte et al. 2012). Analysis of multiple traits increases power because intuitively, multiple trait measurements increase sample size relative to a single trait measurement. However, utilizing the additional data is not straightforward as measurements from the same individual are not independent. This issue is analogous to that of association analysis in cohorts of related individuals, where trait measurements between related individuals are not independent. Variance component methods model this correlation structure by assuming that the covariance due to genetics between related individuals is proportional to their kinship coefficient (Kang et al. 2008). This constant of proportionality normalized by the total trait variance is related to narrow-sense heritability of the trait (the variance accounted for by additive genetic effects) (Yang et al. 2010).

When the same genetic variants affect multiple traits, trait values for an individual will tend to be correlated. Similarly, shared environmental effects also introduce some level of correlation between traits. A fundamental problem in understanding the relationship between the traits is determining the proportion of the total correlation due to genetics and the proportion due to environment. Classical approaches originating from animal breeding and agricultural research solve this problem by modeling the statistical relationship
between traits using a linear mixed-model (LMM) (Falconer 1981; Mrode and Thompson 2005). These approaches decompose the between trait correlation into both a genetic component and an environmental component and then use the LMM framework to obtain estimates for these quantities. The LMMs used in these classical approaches can be adapted for use in GWAS by utilizing them to test the association between genetic variants and multiple traits. Multiple trait variance component methods closely follow the approach utilizing kinship values to model the covariance between different traits among different individuals, such that the genetic covariance between two individual’s traits is proportional to their kinship coefficient (Henderson and Quaas 1976). In this case, the constant of proportionality is a function of the two trait heritabilities as well as the genetic correlation. These types of models are widely utilized in the plant breeding (Malosetti et al. 2008; Kelly et al. 2009; Verbyla and Cullis 2012) and animal breeding communities (Ducrocq and Besbes 1993). Similarly, multiple trait models represent the covariance between traits within an individual as a function of both genetics and shared environment.

In order to utilize LMMs for association analysis, an iterative procedure must be employed to identify the maximum-likelihood estimates of the parameters of the statistical model used for association. The use of LMMs for single traits has been limited by the computational complexity of traditional maximum-likelihood procedures: $O(n^3 \cdot t)$, where $n$ is the number of individuals in the study and $t$ is the number of iterations necessary for the maximum-likelihood algorithm to converge. However, recently developed estimation algorithms have made LMMs computationally efficient and feasible for large population cohorts (Kang et al. 2008; Kang et al. 2010; Lippert et al. 2011; Zhou and Stephens 2012), reducing the computational complexity of from $O(n^3 \cdot t)$ to $O(n^3 + n \cdot t)$, enabling genome wide association mapping for single traits using LMMs. Unfortunately, the previous approaches (Kang et al. 2008; Kang et al. 2010; Lippert et al. 2011; Zhou and Stephens 2012) cannot be directly applied to multiple trait LMMs, meaning that the same computational inefficiencies that limited the widespread use of LMMs for single trait GWAS, now hinder the scale at which researchers can perform multiple trait GWAS. More specifically, with $p$ traits measured over $n$ individuals the covariance matrix relating the $p$ traits measured over the $n$ individuals will be of size $np \times np$ and the running time for classical multivariate LMMs is $O(n^3 p^3 \cdot t)$. In other words, even when $p$ is small (eg. $p = 2$), the running time scales as the cube of the number of individuals in the sample, meaning that the use of multiple trait LMMs is not feasible for large sample sizes.

A widely utilized approximation to using the $np \times np$ covariance matrix is to assume that the genetic and environmental effects are independent which allows the decomposition of the $np \times np$ matrix into the Kroncker product of an $n \times n$ matrix and a $p \times p$ matrix. This type of approach is widely utilized in the plant breeding literature (Malosetti et al. 2008; Kelly et al. 2009; Verbyla and Cullis 2012). In our
work we reformulate this decomposition using the matrix-variate normal distribution (Gupta and Nagar 2000). Using this formulation, we show how a simple data transformation leads to a model equivalent to the above mentioned model while allowing maximum-likelihood inference to be performed in computational time essentially linear in the size of the data set, given a one time cost of \(O(n^3)\) and \(O(n^2)\). In a simple case, let us assume that \(p\) is much less than \(n\) (eg. 2 vs. 10,000) and that we only have a global mean for each trait; this leads to a total computational complexity of \(O(n^3 + n^2p + (p^3(n + 1)) \cdot t)\). The iterative part of the algorithm is then essentially linear in the size of the dataset. We call our method the matrix-variate linear mixed-model (mvLMM). Our approach differs from previous approaches in the plant and animal breeding communities in that our inference approach is more closely related to the EMMA algorithm (Kang et al. 2008) while previous inference methods are more closely related to the Average Information REML algorithm as implemented in ASReml (Gilmour et al. 1995). The reason why algorithms such as EMMA (Kang et al. 2008), EMMAX (Kang et al. 2010), FaST-LMM (Lippert et al. 2011) and GEMMA (Zhou and Stephens 2012) and related methods have become popular in human GWAS studies is that they take advantage of the specific formulation of the variance components to allow for efficient estimation compared to methods such as ASReml which can be applied to a more general set of models.

We demonstrate the efficacy of our method by analyzing correlated traits in the Northern Finland Birth Cohort (Sabatti et al. 2008). Comparing to a standard approach (Lee et al. 2012), we show that our method results in more than a 10-fold time reduction for a pair of correlated traits, taking the analysis time from about 35 minutes to about 2.5 minutes for the cubic operations plus another 12 seconds for the iterative part of the algorithm. In addition, the cubic operation can be saved so that it does not have to be re-calculated when analyzing other traits in the same cohort. Finally, we demonstrate how this method can be used to analyze gene expression data. Using a well-studied yeast dataset (Smith and Kruglyak 2008), we show how estimation of the genetic and environmental components of correlation between pairs of genes allows us for to understand the relative contribution of genetics and environment to coexpression.

**METHODS**

**Modeling multiple traits with the matrix-variate linear mixed-model** Given a set of \(p\) traits for \(n\) individuals, a standard statistical model for the \(i\)th trait vector, denoted by \(y_i\), is given by the following linear mixed model (LMM) the model relating phenotypes to genotypes is

\[
y_i = X\beta_i + g_i + e_i
\]

where \(X\beta_i\) represents the mean term for the \(i\)th trait such that \(X\) is an \(n \times q\) matrix encoding \(q\) covariates including the SNP being tested, \(g_i\) represents the population structure or genetic background component and
e_i represents the effect due to environment and error. We use y_{ij} to represent the value of the i\textsuperscript{th} trait for the j\textsuperscript{th} individual. We have assumed that the covariates determining the mean will be shared among traits, but this is not a requirement. The variance of y_i is given by the following, assuming that \text{cov}(g_i, e_i) = 0.

\[
\text{var}(y_i) = \text{var}(g_i) + \text{var}(e_i) = \sigma_{g(i)}^2 K + \sigma_{e(i)}^2 I
\]

where \sigma_{g(i)}^2 represents the genetic variance component for trait i, K represents the n \times n kinship matrix calculated using a set of m known variants and \sigma_{e(i)}^2 represents the environmental and error variance. We note this model assumes i.i.d. environmental errors for a given trait which maybe unrealistic for some applications(Bello et al. 2012). We use K_{jk} to represent the entry of the kinship matrix corresponding to the relation between the j\textsuperscript{th} and k\textsuperscript{th} individuals. From these models (Henderson and Quaas 1976; Mrode and Thompson 2005), it follows that the covariance between measurements for individuals j and k for trait i is given by

\[
\text{cov}(y_{ij}, y_{ik}) = \sigma_{g(i)}^2 K_{jk}.
\]

We now consider models with multiple traits. By letting \rho_{im} represent the correlation between traits i and m due to genetic effect and letting \lambda_{im} represent the correlation due to an individual’s environment, the covariance between the trait measurements i and m for individual j is

\[
\text{cov}(y_{ij}, y_{mj}) = \text{cov}(g_{ij}, g_{mj}) + \text{cov}(e_{ij}, e_{mj}) = \rho_{im} \sigma_{g(i)} \sigma_{g(m)} + \lambda_{im} \sigma_{e(i)} \sigma_{e(m)}.
\]

Assuming that environmental effects are independent between individuals, let the covariance between traits i and m for individuals j and k be

\[
\text{cov}(y_{ij}, y_{mk}) = K_{jk} \rho_{im} \sigma_{g(i)} \sigma_{g(m)}.
\]

In fact, these models are standard models utilized in the animal and plant breeding communities.

A straightforward approach to represent this model is to stack all of the traits for each trait into one long vector of length np and then represent their covariances in a np \times np matrix populated using equations (1), (2), and (3). However, this matrix will have n^2p^2 elements and fitting this model to estimate the parameters for even a small number of phenotypes is computationally intractable.

**Matrix-variate normal distribution** We note that the np \times np covariance matrix above has a significant amount of structure as evident in equations (1), (2), and (3). In fact, this matrix can be represented by the
sum of two matrices, each of which is a Kronecker product of an n x n and p x p matrix. This decomposition is widely utilized in the plant breeding literature (Malosetti et al. 2008; Kelly et al. 2009; Verbyla and Cullis 2012). In our work, the main contribution is that we provide an efficient method for performing inference in these models efficiently by modeling the full set of trait measurements using a matrix-variate normal distribution. The matrix-variate normal distribution is a generalization of the multivariate normal distribution to matrices (Gupta and Nagar 2000). The matrix-variate normal distribution is a very natural way to represent these types of factored models. Unlike in a multivariate normal model where the data is concatenated into a single vector of length np, in a matrix-variate model, the data ($Y$) is a $n \times p$ matrix where each column is a trait. Instead of representing a covariance structure using a single $np \times np$ matrix, the matrix variate normal distribution represents the covariance using two matrices: a $p \times p$ matrix $A$ which represents the covariance between columns of the data and a $n \times n$ matrix $B$ which represents the covariance between rows of the data. In a matrix-variate normal distribution, the mean ($M$) is now a $n \times p$ matrix. We denote a matrix variate normal model using the notation $N_{n \times p}(M, A, B)$.

Using the matrix-variate normal distribution, our model can be represented

$$Y = Z + R$$

where $Y$ is the $n \times p$ matrix of traits, $Z$ follows a matrix-variate normal distribution with mean $X\beta = X[\beta_1 \ldots \beta_p]$ and covariance matrices $\Psi$ and $K$, where $\Psi$ is a $p \times p$ matrix representing the correlation between traits due to genetics and $K$ is the kinship matrix. $R$ follows a matrix variate normal distribution with mean zero and covariance matrices $\Phi$ and $I_n$, where $\Phi$ is a $p \times p$ matrix representing the covariance between traits due to environment and error. The $i$th diagonal component of $\Psi$ is given by $\sigma^2_{g(i)}$ and the $i, j$th component by $\rho_{ij} \sigma_{g(i)} \sigma_{g(j)}$, and similarly $\Phi_{ij} = \lambda_{ij} \sigma_{e(i)} \sigma_{e(j)}$. The distribution for $Y$ is then summarized as follows, where $N_{n \times p}(M, A, B)$ denotes the matrix variate normal distribution with mean matrix $M$ and columns and row covariance matrices $A$ and $B$.

$$Y \sim N_{n \times p}(X\beta, \Psi, K) + N_{n \times p}(0, \Phi, I_n)$$

Efficient Maximum Likelihood Computation Likelihood evaluation for the matrix-variate distribution given by equation (4) is accomplished by evaluating the equivalent multivariate normal distribution. By using the $\text{vec}()$ operator, which creates a vector from a matrix input by concatenating the columns of the matrix, we are able to represent the distribution given in equation (4) in the following way, where $\otimes$ represents the Kronecker product of two matrices.
\[ \text{vec}(Y) \sim N_{np}(\text{vec}(X\beta), \Psi \otimes K + \Phi \otimes I_n) \]

The likelihood computation for this model takes time on the order of \((np)^3\). This computational time becomes prohibitive when maximizing the likelihood function while considering a large cohort with multiple traits. Previous work has shown how similar multivariate models with kronecker product matrices can be utilized efficiently when residual errors are independent (Stegle et al. 2011). However, it is not known how these models may be used efficiently when residual errors are correlated, which is the case for our model. To remedy this problem, we introduce a transformation that results in a reduced computational time.

Let the eigendecomposition of \( K = H_R S_K H_K' \). This decomposition is calculated with a computational complexity of \( O(n^3) \). Let \( L \) be a \( p \times p \) matrix that diagonalizes both \( \Psi \) and \( \Phi \), such that \( L\Psi L' = I \) and \( L\Phi L' = D \), a diagonal matrix. This bi-diagonalization can be accomplished in \( O(p^3) \) (details are found in a later section). We then define the matrix \( M = (L \otimes H_K') \). The transformed data vector \( Y_T \) is defined as \( Y_T = M\text{vec}(Y) \). This transformed vector has the following distribution.

\[ Y_T \sim N(M\text{vec}(X\beta), I \otimes S_k + D \otimes I) \]

The log likelihood of \( Y_T \) is then given as follows.

\[
L(Y_T|X\beta, \Psi, K, \Phi) = -\frac{np}{2} \ln(2\pi) - \frac{1}{2} \ln|I \otimes S_k + D \otimes I| \\
-\frac{1}{2}(M\text{vec}(Y_T - X\beta)')(I \otimes S_k + D \otimes I)^{-1}(M\text{vec}(Y_T - X\beta)) + \log(|M|)
\]

In order to calculate the likelihood given \( \Psi \) and \( \Phi \), we first obtain the transformation matrix \( M \), which is accomplished in \( O(n^3+p^3) \). Next, we compute the transformed data vector \( Y_T \) in \( O(n^2p+p^2n) \). Given \( Y_T \), we obtain an estimate of \( \beta \), denoted by \( \hat{\beta} \), which we show may be accomplished in \( O(np^3q^2 + p^3q^3 + np^2q) \) and given this we calculate the residual vector \( Y_T - M\text{vec}(X\hat{\beta}) \) in \( O(np^2q + np) \). Finally, the likelihood is computed in \( O(np) \). Part of the reason that our approach is efficient is that much of the computations can be reused for many analyses. For example, the matrix \( M \) which is computed in \( O(n^3+p^3) \) requires diagonalizing the \( K \) matrix which requires \( O(n^3) \) time and only needs to be performed once for the complete analysis of the dataset. Similarly, the transformed data vector \( Y_T \) can be computed in \( O(n^2p+p^2n) \) does not depend on which variant is actually being tested and can be computed only once for each set of traits which is being considered. Thus the likelihood computation for each variant is dominated by \( O(np^3q^2) \) utilizing the quantities which were computed once. In addition, in many scenerios we can assume that the effect sizes are...
small as in human studies. Under this assumption, we can fit the variance parameters just once assuming
that $\beta = 0$ and then use this estimate to test each variant. In this case, computing the maximum likelihood
reduces to $O(np)$. This transformation is similar to the approaches in the plant breeding literature to speed
up computations using two eigendecomposition (P. PIEPHO et al. 2012).

This assumption is the same assumption which differentiates EMMAX (KANG et al. 2010) from EMMA
(KANG et al. 2008). While this assumption is appropriate for human studies where most identified genetic
variants have very small effects, this assumption may not be appropriate for plant and animal models where
there are often several loci with very strong effects. An approach to handle this case while avoiding refitting
the variance parameters for each variant is to first identify the variants with strong effects using the above
assumption and then refit the variance parameters after including these strongly associated variants as fixed
effects in the model.

**Restricted Maximum Likelihood Computation** The restricted maximum likelihood (REML) and the
maximum likelihood (ML) solutions should be similar when the model contains no covariates, or only a bias
term. However, when this is not the case, parameter estimates obtained in REML analysis may deviate
significantly from those of ML. We obtain the REML version of the mvLMM by extending the ML solution
(WELHAM and THOMPSON 1997). By denoting the log-likelihood obtained by ML as $L_{ML}$ and similar
for REML, we define the following log-likelihood function. For a standard multivariate normal vector $y$
with distribution $N(T\alpha, \Theta)$, where $T$ is $n \times q$, the REML is $LL_{REML} = LL_{ML} + \frac{1}{2} [q ln(2\pi) + ln(|T'\Theta^{-1}T|) - ln(|T'\Theta^{-1}T|)]$ (KANG et al. 2008). Given this standard result, we define the REML log-likelihood for the
mvLMM in the following.

$$LL_{REML} = LL_{ML} + \frac{1}{2} [q ln(2\pi) + ln(|(L' \otimes (H_k'X)')(L \otimes H_k'X)|) -
ln(|(L' \otimes (H_k'X)')(I \otimes S_k + D \otimes I)^{-1}(L \otimes H_k'X)|)]$$

The computational cost of the operations required to define $LL_{REML}$ do not change the order of the com-
putational complexity.

**Estimating Genetic Correlation** In order to evaluate the likelihood function in equation (5), we obtain
estimates for the parameters $\Psi$ and $\Phi$. We estimate these parameters under the null model, where SNPs are
not included as covariates. This assumption has been used previously and is valid for cases when the effect
due to each SNP is small (KANG et al. 2010; LIPPERT et al. 2011). First, for each trait $i$, we fit the basic
LMM from equation (1), in order to identify the optimal values of the variance parameters \( \sigma^2_{g(i)} \) and \( \sigma^2_{e(i)} \). Holding these parameters constant, we perform a two dimensional global grid search in order to identify the optimal genetic and environmental correlation parameters. With caching the likelihood calculation takes time on the order of \( O(p^3 + np^3) \). This time will be multiplied by a constant \( k^2 \) when searching over a grid of size \( k \) for each correlation parameter. That is, if we evaluate the likelihood for each genetic and environmental correlation combination for a grid size of \( k \), then we need to evaluate the likelihood \( k^2 \) times.

In order to expand this approach to more than two traits, we propose a straightforward pairwise approach to identify the maximum-likelihood parameters. Instead of performing a full grid search over the correlation parameters, we identify the ML estimates of the parameters in each pair of traits. This procedure will be much faster than a full grid search over all pairs of traits and we discuss the in the supplementary why this procedure is also more robust.

**Calculating sampling variance for parameter estimates** We calculate the sampling variance of the variance parameters and the correlation parameters using standard multivariate theory. Generally, the sampling variance of a maximum likelihood (ML) parameter is given by the inverse of the Fisher’s information (or average information) matrix evaluated at the ML parameters (Searle et al. 1992). Using the search technique we describe, we identify the ML parameters for a given set of traits and then use these parameters to estimate the sampling variance using the Fisher’s information matrix.

**Association Analysis** In order to identify genetic variations that have an effect on our traits of interest, we employ a hypothesis testing framework. We first estimate the effect that a particular SNP \( x \) has on each of the traits using the mvLMM model, then we jointly test \( m \) hypotheses, each testing the effect of the SNP on a given trait. Our null hypothesis for this test is that the SNP has no effect on any of our traits and the alternative hypothesis is that it has an effect on one or more of the traits.

To obtain estimates for the SNP effect sizes, we include one SNP in the model at a time and estimate \( \beta \) from equation (5). First, we obtain the maximum likelihood parameters for \( \Psi \) and \( \Phi \) under the null model in which the SNP has no effect, as described in the previous section. Then, given these two parameters, we compute an estimate of the coefficient matrix \( \hat{\beta} \) using the following result.

In the previous section, we defined a transformation \( M = (L \otimes H_k^r) \) and used it to define a transformed data vector \( Y_T \). The mean of the transformed data is given by \( M \text{vec}(X\beta) = (L \otimes H_k^r)\text{vec}(X\beta) \), which can be reduced as follows.
\( (\mathbf{L} \otimes \mathbf{H}_k')\text{vec} (\mathbf{X}\beta) \)

\[ = \text{vec} (\mathbf{H}_k'\mathbf{X}\beta) \]

\[ = \text{vec} (\mathbf{X}^*\beta) \]

\[ = (\mathbf{L} \otimes \mathbf{X}^*)\text{vec} (\beta) \]

Here we have let \( \mathbf{X}^* = \mathbf{H}_k'\mathbf{X} \). By denoting \( \text{vec}(\beta) \) as \( \beta^T \), we obtain an estimate \( \hat{\beta} \) using the following result, where \( \text{unvec}(\cdot) \) represents the reversal of the \( \text{vec}(\cdot) \) operation and we have let \( \mathbf{P} = (\mathbf{I} \otimes \mathbf{S}_k + \mathbf{D} \otimes \mathbf{I}) \), the transformed data covariance matrix.

\[ \hat{\beta}_T = \left[ (\mathbf{L}' \otimes \mathbf{X}'^*)\mathbf{P}^{-1}(\mathbf{L} \otimes \mathbf{X}^*) \right]^{-1}(\mathbf{L}' \otimes \mathbf{X}'^*)\mathbf{P}^{-1}\text{Mvec}(\mathbf{Y}) \]

\[ \hat{\beta} = \text{unvec}(\hat{\beta}_T) \]

Since \( \mathbf{P} \) is a diagonal matrix, \( \hat{\beta}_T \) can be computed in \( O(n^2q^2 + p^3q^2 + np^2q) \) given the one time cost of \( O(n^2q) \) for computing \( \mathbf{X}^* \).

The statistic for testing the proposed hypothesis is obtained by defining a transformation matrix \( \mathbf{R} \) so that \( \mathbf{R}\hat{\beta}_T = [\hat{\beta}_{1x} \hat{\beta}_{2x} \ldots \hat{\beta}_{px}]' \), where \( \hat{\beta}_{ix} \) is the coefficient estimate for the effect of SNP \( x \) on trait \( i \). Therefore, given this matrix, we define the F-statistic for testing association in equation (5), which under the null follows an F-distribution with \( p \) numerator degrees of freedom and \( np - pq \) denominator degrees of freedom, where \( \hat{\sigma}^2 = \text{var}(\mathbf{P}^{-1/2}\mathbf{Y}_T) \) and \( \text{var}(\ldots) \) represents the sample variance. Details on this test can be found in (McCulloch and Neuhaus 1999).

\[ f = (\mathbf{R}\hat{\beta}_T)'(\mathbf{R}'[(\mathbf{L}' \otimes \mathbf{X}'^*)\mathbf{P}^{-1}(\mathbf{L} \otimes \mathbf{X}^*)]^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\beta}_T) \cdot \frac{1}{p\hat{\sigma}^2} \]

**Diagonalizing two matrices** We are given two positive semi-definite matrices \( \Phi \) and \( \Psi \) and we wish to identify a matrix \( \mathbf{L} \) that diagonalizes both of these matrices. This is accomplished in the following way. First, we obtain the eigendecomposition of \( \Psi = \mathbf{H}_\Psi \mathbf{S}_\Psi \mathbf{H}_\Psi' \) and then define a matrix \( \mathbf{R} = \mathbf{S}_\Psi^{-1/2}\mathbf{H}_\Psi \), so that \( \mathbf{R}'\mathbf{R} = \Psi^{-1} \). Next, we obtain an eigendecomposition \( \mathbf{R}\Phi\mathbf{R}' = \mathbf{QDQ}' \) and then define a matrix \( \mathbf{L} = \mathbf{Q}'\mathbf{R} \). With this we see that \( \mathbf{L}\Psi\mathbf{L}' = \mathbf{I} \) and that \( \mathbf{L}\Phi\mathbf{L}' = \mathbf{D} \). The entire procedure has complexity \( O(p^3) \).

**Genotype and phenotype data** We apply our method to the Northern Finland Birth Cohort data (Sabatti et al. 2008) which was used in (Kang et al. 2010) and (Korte et al. 2012). This data set
consisted of 5326 individuals which had been filtered to reduce the presence of family structure. The data set contains 331,450 autosomal SNPs after application of the exclusion criteria of Hardy-Weinberg equilibrium \((p < 10^{-4})\), genotyping completeness \((< 95\%)\), and minor allele frequency \((< 1\%)\). Missing genotypes are replaced with the MAF. Missing phenotypes are replaced with the phenotypic mean.

We use a well-studied yeast dataset (Smith and Kruglyak 2008) consisting of 109 yeast strains each with 5793 gene expression measurements. Bi-variate association mapping is performed on all 2956 available SNPs. Gene expression values were normalized and subjected to quality control by (Smith and Kruglyak 2008) and we utilized the same data as they.

RESULTS

Association and genetic correlation in the Northern Finland Birth Cohort

Association We apply our method to the Northern Finland Birth Cohort, a founder cohort consisting of 5,043 individuals each of which has multiple trait measurements for four different metabolic traits. We analyze a total of six pairs of traits or all combinations of four traits: HDL and LDL cholesterol, C-reactive protein (CRP) and triglycerides (TG). Association between each SNP and each pair of traits is evaluated by assuming that under the null hypothesis the SNP does not effect either trait.

We compare our results to the analysis of Korte et al. (2012) which analyzed the same data using a classically-based multiple trait LMM which they refer to as the multi-trait mixed model (MTMM) method. Our results are highly concordant \((r =0.96 - 0.99)\), indicating that our method is consistent with classical approaches. For example, Figure 1 compares the QQ-plots of the mvLMM and MTMM for the joint analysis of TG and LDL.

Over 99% of associations identified in marginal analysis are also identified when respective pairs of traits are mapped (significance threshold of 1.5e-7). However, the joint mapping uncovers more significant associations; 19 new associations are identified across all trait pairs. For example, in the analysis of TG with CRP, we identify a SNP (rs2000571) with a p-value of 8.58e-7 and with MTMM p-value of 1.7e-6. This SNP was not significant in the marginal analysis of TG (1.7e-5) or CRP (0.03), but belongs to a region on chromosome 11 that has been shown to harbor variants contributing to triglycerides (Braun et al. 2012).

Many of the identified associated SNPs were more significant in mvLMM compared to MTMM which we suspect is because the mvLMM finds the actual parameters which maximize the likelihood. We also apply our method to analyze all four traits simultaneously and the results are shown in 2. For all variants, at least one of the pair of trait p-values is more significant than the all trait p-value. Thus it appears that in this scenario, it is best to follow a single trait analysis with the all pairs analysis. This raises the more general
issue of how one should apply a method such as this and we provide some guidance in the discussion.

*Genetic Correlations* In multiple trait models, the total trait correlation is partitioned into a genetic and an environmental component. The genetic component of the correlation (the genetic correlation) represents the part of the total trait covariance that is attributed to genetics normalized by the genetic variances. This quantity provides insight into the genetic architecture of the relationships between traits. We estimate the genetic correlations for each pair of traits analyzed in the Finland Birth Cohort and compare these estimates with those obtained using a standard implementation of a bi-variate LMM as implemented in GCTA (Genome-wide Complex Trait Analysis) (Lee et al. 2012). Table 1 compares estimates of genetic correlation obtained with GCTA and mvLMM. When we compare our results to those of GCTA we find that the two methods yield similar results, with genetic correlation estimates falling less than one standard deviation from one another. In addition, the running time for the classical approach was around 35 minutes, while the running time for mvLMM was on average roughly 12 seconds, given a one time cost of 2.5 minutes shared across pairs of traits.

**Bi-variate analysis in yeast data** Gene coexpression, defined as the correlation between expression levels of a pair of genes estimated in a set of individuals, is a fundamental quantity that has been utilized for a variety of applications (Lee et al. 2006; Ghanzalpour et al. 2006; Subramanian et al. 2005; Stuart et al. 2003). There are two prevalent views about the meaning of significant coexpression. The first is that coexpression stems from similar environmental conditions such as disease status (Heller et al. 1997). The second comes from the systems genetics literature where it is thought that coexpressed genes have a similar genetic regulatory program and that specific genetic variants drive modules of coexpressed genes (Lee et al. 2006; Ghanzalpour et al. 2006). However, correlation estimates from gene expression levels measures the combined effect of both the genetic and environmental components. Our methodology allows for the first time to decompose the coexpression into a genetic and environmental component.

We utilize the major gain in efficiency of our approach to perform an analysis that is not feasible with current methods. Using a well-studied yeast dataset (Smith and Kruglyak 2008) consisting of 109 yeast strains each with 5793 gene expression measurements, we perform a bi-variate analysis, estimating genetic correlations for all 5793 choose 2 gene expression pairs. Within this dataset several regions of the genome have been implicated to harbor genetic variation that affects many gene expression levels.

Using a set of hotspot locations derived from (Smith and Kruglyak 2008), we define a set of 13,508 hotspot gene pairs by extracting all pairs of genes that lie in each known hotspot. We then compare the phenotypic correlation to the total proportion of covariation accounted for by genetics for each of these
pairs. Assuming that hotspot pairs are under the same genetic regulation, we expect that the phenotypic correlation for any given pair should reflect this by having a high value. However, this might not be the case if the environmental correlation between the pair contributes in such a way to lower the overall phenotypic correlation. Therefore, an estimation of the total phenotypic covariation attributed to genetics may better reflect the fact that the two genes are under the same genetic program.

In Figure 2(a), we plot the histogram of the absolute value of the total phenotypic correlation for all gene pairs and for hotspot gene pairs. We see that the distribution of phenotypic correlations for hotspot pairs is shifted towards higher correlations with respect to all pairs, giving an indication of co-regulation. However, most of the pairs have correlation less than 0.5. Figure 2(b) shows the same plot generated using the total proportion of the phenotypic covariation attributed to genetics. In the figure, we observe that the estimated genetic correlation for hotspot pairs is dramatically skewed towards one. In fact, most of the pairs have a genetic covariance above 0.7. This result suggests that the estimated genetic correlations on average give a stronger indication of co-regulation compared to the phenotypic correlation.

DISCUSSION

In this paper, we introduced a method for performing multi-trait genome-wide association analysis and for the estimation of the genetic correlation. Our method is based on classical theory, but introduces a computational advance that makes it much faster, reducing running time over 10-fold when compared with the classic approach. We have shown that our method achieves similar results to that of the classical approach. In addition, we have shown that the ability to quickly estimate genetic correlation may be of great benefit to researchers, leading to fundamental insights into the architecture of complex traits.

The ability to quickly optimize multiple trait linear mixed-models will have a large impact on the ability to dissect complex traits. For example, multiple expression quantitative trait loci (multi-eQTL) may be discovered by mapping multiple traits to genetic variants across the genome. The ability to perform this type of research is infeasible with current methodologies. In addition, we have shown that the genetic correlation between gene expression measurements may be a better indicator of co-regulation. It stands to reason that these genetic correlations may be used in coexpression analysis and lead to the discovery of gene modules that are truly co-regulated and not in part due to environmental correlations.

We note that in our model, the genetic background component is assumed to have a covariance structure, defined by the matrix K, which is computed using all of the marker genotypes. This model inherently assumes that the effect size of each genetic variant is drawn from a normal distribution with equal variance. This may be inaccurate for several reasons. First, not all of the markers are causal variants and even among the variants which are causal, their effect sizes may vary widely. Second, many of the causal variants themselves
may not be genotyped in the study and the markers are merely proxies for these causal variants. This difference between the estimated covariance structure from the markers and the true covariance structure has been shown to lead to inaccurate heritability estimates (de Los Campos et al. 2013) and may lead to inaccuracies in estimates of genetic correlations. A more appropriate term from the quantities we estimate maybe “genomic heritabilities” and “genomic correlations.”

Our method presents an approach for jointly performing association analysis for multiple traits. However, the question remains on what is the best way to analyze a dataset with multiple traits. Unfortunately, there is no clear answer. If a variant affects only a single trait, then an individual trait by trait analysis is the most powerful to identify such a trait because analyzing more than one trait increases the degrees of freedom of the statistical test. On the other hand, if a variant affects multiple traits, then analyzing all traits together will be more powerful. From a practical perspective, we advocate first analyzing each trait independently and then applying this method to groups of traits where there is suspected shared genetic components and keep increasing the number of traits analyzed until the p-values become less significant. Our estimates of genetic correlation can guide identification of potential groups of traits. Any such sequential strategy complicates issues of controlling Type I errors. Exactly how to control Type I errors in this context is an important avenue of future work.

ACKNOWLEDGEMENTS

N.F. and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448 and 1320589, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-ES021801, R01-MH101782 and R01-ES022282. N.F. was supported in part by the NIH Training Grant T32MH073526. E. E. is supported in part by the NIH BD2K award, U54EB020403. We acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691).
LITERATURE CITED


LIU, Y., Y. PEI, J. LIU, F. YANG, Y. GUO, L. ZHANG, X. LIU, H. YAN, L. WANG, Y. ZHANG, and
others, 2009 Powerful bivariate genome-wide association analyses suggest the SOX6 gene influencing

multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize.

McCULLOCH, C. and J. NEUHAUS, 1999 Generalized linear mixed models. Wiley Online Library.

MRODE, R. and R. THOMPSON, 2005 Linear models for the prediction of animal breeding values (2nd ed.).
Cambridge, MA: Cabi.

PIEPHO, H., J. O. OGUTU, T. SCHULZ-STREECK, B. ESTAGHVIROU, A. GORDILLO, and F. TECHNOW,
2012 Efficient Computation of Ridge-Regression Best Linear Unbiased Prediction in Genomic Selection

SABATTI, C. and others, 2008 Genome-wide association analysis of metabolic traits in a birth cohort from


Biology 6(4): e83.

in matrix-variate gaussian models with iid observation noise. In Advances in Neural Information


SUBRAMANIAN, A., P. TAMAYO, V. K. MOOTHA, S. MUKHERJEE, B. L. EBERT, M. A. GILLETTE,
A. PAULOVICH, S. L. POMEROY, T. R. GOLUB, E. S. LANDER, and J. P. MESIROV, 2005 Gene set
enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc

VERBYLA, A. P. and B. R. CULLIS, 2012 Multivariate whole genome average interval mapping: QTL
analysis for multiple traits and/or environments. Theor Appl Genet 125(5): 933–53.

Common SNPs explain a large proportion of the heritability for human height. Nature Genet 42(7): 565–569. doi:10.1038/ng.608.

Figure 1: **QQ Plot** comparing MTMM and mvLMM p-values obtained when performing analysis with LDL and TG.
Figure 2: Comparison of the phenotypic correlation with the total proportion of the correlation accounted for by genetics for all gene pairs and for gene pairs from regulatory hotspots. We compare the phenotypic correlation with the total proportion of correlation accounted for by genetics in order to assess the ability of the genetic correlation to differentiate gene pairs that are co-regulated. Utilizing a set of known hotspots, we derive a set of hotspot gene pairs, where a hotspot pair is defined as a gene pair in which both genes lie in a given hotspot. We find that the genetic correlation differentiates these co-regulated pairs better than the overall trait correlation.
Table 1: **Genetic correlation estimates in the Finland Birth Cohort.** We compare the maximum likelihood estimates obtained with mvLMM with those obtained with GCTA and find that the results are very similar.
We compare the p-values of the analysis of all four traits to the six possible pairwise trait analyses. In all cases, the pairwise analyses are more significant.

Table 2: Joint analysis of all traits compared to all pairwise combinations. We compare the p-values of the analysis of all four traits to the six possible pairwise trait analyses. In all cases, the pairwise analyses are more significant.
SUPPLEMENTARY MATERIALS

**mvLMM and Bayesian Linear Regression** The standard LMM used in GWAS has been shown to be equivalent to a Bayesian linear regression in which a number of SNPs $m$ are assumed to each have an effect on the trait, such that each effect is sampled IID from a normal distribution (Hayes et al. 2009; Listgarten et al. 2012). By integrating out these effects, one may arrive at a standard LMM using the realized relationship matrix (RRM) as the kinship matrix (Goddard et al. 2009; Yang et al. 2010). Here we briefly summarize this result and show how it extends to multiple trait LMMs.

Let us assume that a set of $m$ SNPs each contribute to the background phenotypic variation for trait $k$. Let $W$ be a $n \times m$ matrix allocating SNP effects to individuals, such that $E[W_{ij}] = 0$ and $\text{var}(W_{ij}) = 1$ and assume that the phenotypic effect attributed to SNP $j$ for trait $k$ is $b_{jk}$, so that individual $i$ will have a total effect due to SNP $j$ of $W_{ij}b_{jk}$. We treat the SNP effect as random and assume that each $b_{jk}$ is sampled IID from distribution $N(0, \frac{1}{m}\sigma^2_{g(k)})$. Let $g_k = Wb_k$, where $b_k = [b_{1k} \ b_{2k} \ldots b_{mk}]'$.

Therefore, the variance of $g_k$ is given by equation (5). Thus, LMM-based population structure correction may be viewed as a basic linear model, while treating the SNP effects as random effects.

$$\text{var}(g_k) = \frac{WW'}{m}\sigma^2_{g(k)} = K\sigma^2_{g(k)}$$

This framework may be extended to multiple traits by assuming that the correlation between SNP effect vectors has the following form, where $\text{cor}(g_{ki}, g_{ji}) = \rho_{ij}$.

$$\begin{bmatrix} b_i \\ b_j \end{bmatrix} \sim N\left(0, \frac{1}{m} \begin{bmatrix} \sigma^2_{g(i)} & \rho_{ij}\sigma_{g(i)}\sigma_{g(j)} \\ \rho_{ij}\sigma_{g(i)}\sigma_{g(j)} & \sigma^2_{g(j)} \end{bmatrix} \right)$$

To obtain the joint distribution of $g_i$ and $g_j$, we apply the following linear transformation.

$$\begin{bmatrix} g_i \\ g_j \end{bmatrix} = \begin{bmatrix} Wb_i \\ Wb_j \end{bmatrix} = \begin{bmatrix} W & 0 \\ 0 & W \end{bmatrix} \begin{bmatrix} b_i \\ b_j \end{bmatrix} \sim N\left(0, \frac{1}{m} \begin{bmatrix} \sigma^2_{g(i)}WW' & \rho_{ij}\sigma_{g(i)}\sigma_{g(j)}WW' \\ \rho_{ij}\sigma_{g(i)}\sigma_{g(j)}WW' & \sigma^2_{g(j)}WW' \end{bmatrix} \right)$$

$$= N\left(0, \begin{bmatrix} \sigma^2_{g(i)}K & \rho_{ij}\sigma_{g(i)}\sigma_{g(j)}K \\ \rho_{ij}\sigma_{g(i)}\sigma_{g(j)}K & \sigma^2_{g(j)}K \end{bmatrix} \right)$$

This result is consistent with the proposed model in the previous section.
The same basic logic is easily applied to derive the \( \text{cov}(e_i, e_j) \). By substituting \( W \) for \( I \) as well as the appropriate variance and correlation parameters we arrive at the equivalent result for the correlation between residuals, given in the equation below.

\[
\begin{pmatrix}
e_i \\
e_j
\end{pmatrix} \sim N \left( 0, \begin{bmatrix}
\sigma^2_{e(i)} I & \lambda_{ij} \sigma_{e(i)} \sigma_{e(j)} I \\
\lambda_{ij} \sigma_{e(i)} \sigma_{e(j)} I & \sigma^2_{e(j)} I
\end{bmatrix} \right)
\]

We note that a similar analysis may be applied when the two sets of causal SNPs are different for each trait. In this case, the between trait genetic covariance will be proportional to the \( W_c W_c' \), where \( W_c \) represents the \( n \times t \) SNP incidence matrix for causal SNPs that are common between the two traits. If this matrix deviates significantly from the full kinship matrix \( K \), then it is possible that the estimated genetic correlation may be biased.

**Robustness of Estimation Procedure** One concern with our approach for identifying the variance parameters is that the ML parameters we identify in the marginal model for a given trait might not be the same as the variance estimates we identify using a traditional method. In the Supplementary Figure (1), we show through simulation that this is not a great concern. In particular, we compare variance estimates between mvLMM and GCTA for a set of 1000 simulated trait pairs. This figure shows that the variance estimates are highly correlated. We also note that for a small number of trait pairs (~50) GCTA estimates extremely large variances for at least 1 trait (e.g. \( 1e16 \)). This is likely due to some numerical issue with their method and to be fair we disregard these cases. In addition to this, in about 1/10th of the cases, GCTA did not converge in 1000 iterations or resulted in an error. In addition to this, we show in Supplementary Table (1) that the genetic and environmental variance estimates for the Finland Cohort are also highly concordant between the two methods.

Another concern is that our pairwise fitting of the genetic and environmental correlations may lead to different estimates than if we fit the full model. In Supplementary Figure (2), we show through simulation that this procedure (Pairwise) results in lower residual error in the genetic and environmental correlation for three traits when compared with a the full grid search approach (Full Grid).