

UCLA

UCLA Previously Published Works

Title

Efficient Multiple-Trait Association and Estimation of Genetic Correlation Using the Matrix-Variate Linear Mixed Model

Permalink

<https://escholarship.org/uc/item/4h6560jp>

Journal

Genetics, 200(1)

ISSN

0016-6731

Authors

Furlotte, Nicholas A
Eskin, Eleazar

Publication Date

2015-05-01

DOI

10.1534/genetics.114.171447

Peer reviewed

Efficient Multiple-Trait Association and Estimation of Genetic Correlation Using the Matrix-Variate Linear Mixed Model

Nicholas A. Furlotte* and Eleazar Eskin^{†,1}

*Department of Computer Science and [†]Department of Computer Science, Department of Human Genetics, University of California, Los Angeles, California 90095

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 multitrait 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 article, 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 speedup, 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 coexpressed genes. An implementation of this method is available at <http://genetics.cs.ucla.edu/mvLMM>.

KEYWORDS association studies; multivariate analysis; genetic correlation

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

Copyright © 2015 by the Genetics Society of America
doi: 10.1534/genetics.114.171447

Manuscript received September 30, 2014; accepted for publication February 16, 2015;
published Early Online February 27, 2015.

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.171447/-/DC1>.

¹Corresponding author: Department of Computer Science, Department of Human Genetics, 3532-J Boelter Hall, University of California, Los Angeles, CA 90095-1596.
E-mail: eeskin@cs.ucla.edu

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 genome-wide association studies (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 individuals' 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.

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, 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, 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 (*e.g.*, $p = 2$), 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 Kronecker 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 abovementioned 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 $p \ll n$ (*e.g.*, 2 vs. 10,000) and that we have only a global mean for each trait; this leads to a total computational complexity of $O(n^3 + n^2 p + (p^3(n+1)) \cdot t)$. The iterative part of the algorithm is then essentially linear in the size of the data set. 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 restricted maximum-likelihood (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 is that they take advantage of the specific formulation of the variance components to allow for efficient estimation compared to methods such as ASReml that 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 it to a standard approach (Lee *et al.* 2012), we show that our method results in a >10 -fold time reduction for a pair of correlated traits, taking the analysis time from ~ 35 min to ~ 2.5 min for the cubic operations plus another 12 sec for the iterative part of the algorithm. In addition, the cubic operation can be saved so that it does not have to be recalculated 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 data set (Smith and Kruglyak 2008), we show how estimation of the genetic and environmental components of correlation between pairs of genes allows us 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 \mathbf{y}_i , is given by the following LMM, the model relating phenotypes to genotypes, which is

$$\mathbf{y}_i = \mathbf{X}\beta_i + \mathbf{g}_i + \mathbf{e}_i,$$

where $\mathbf{X}\beta_i$ represents the mean term for the i th trait such that \mathbf{X} is an $n \times q$ matrix encoding q covariates including the SNP being tested, \mathbf{g}_i represents the population structure or genetic background component, and \mathbf{e}_i represents the effect due to environment and error. We use y_{ij} to represent the

value of the i th trait for the j 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 \mathbf{y}_i is given by the following, assuming that $\text{cov}(\mathbf{g}_i, \mathbf{e}_i) = 0$,

$$\text{var}(\mathbf{y}_i) = \text{var}(\mathbf{g}_i) + \text{var}(\mathbf{e}_i) = \sigma_{g(i)}^2 \mathbf{K} + \sigma_{e(i)}^2 \mathbf{I},$$

where $\sigma_{g(i)}^2$ represents the genetic variance component for trait i , \mathbf{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 is 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 th and k 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

$$\text{cov}(y_{ij}, y_{ik}) = \sigma_{g(i)}^2 K_{jk}. \quad (1)$$

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

$$\begin{aligned} \text{cov}(y_{ij}, y_{mj}) &= \text{cov}(g_{ij}, g_{mj}) + \text{cov}(e_{ij}, e_{mj}) \\ &= \rho_{im} \sigma_{g(i)} \sigma_{g(m)} + \lambda_{im} \sigma_{e(i)} \sigma_{e(m)}. \end{aligned} \quad (2)$$

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

$$\text{cov}(y_{ij}, y_{mk}) = K_{jk} \rho_{im} \sigma_{g(i)} \sigma_{g(m)}. \quad (3)$$

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 1–3. However, this matrix will have $n^2 p^2$ 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 1–3. In fact, this matrix can be represented by the sum of two matrices, each of which is a Kronecker product of an $n \times n$ and $p \times p$ 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 are concatenated into a single vector of length np , in a matrix-variate model, the data (\mathbf{Y}) are an $n \times p$ 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 \mathbf{A} that represents the covariance between columns of the data and an $n \times n$ matrix \mathbf{B} that represents the covariance between rows of the data. In a matrix-variate normal distribution, the mean (\mathbf{M}) is now an $n \times p$ matrix. We denote a matrix-variate normal model, using the notation $N_{n \times p}(\mathbf{M}, \mathbf{A}, \mathbf{B})$.

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

$$\mathbf{Y} = \mathbf{Z} + \mathbf{R},$$

where \mathbf{Y} is the $n \times p$ matrix of traits; \mathbf{Z} follows a matrix-variate normal distribution with mean $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}[\beta_1 \dots \beta_p]$ and covariance matrices $\boldsymbol{\Psi}$ and \mathbf{K} , where $\boldsymbol{\Psi}$ is a $p \times p$ matrix representing the correlation between traits due to genetics; and \mathbf{K} is the kinship matrix. \mathbf{R} follows a matrix-variate normal distribution with mean zero and covariance matrices $\boldsymbol{\Phi}$ and \mathbf{I}_n , where $\boldsymbol{\Phi}$ is a $p \times p$ matrix representing the covariance between traits due to environment and error. The i th diagonal component of $\boldsymbol{\Psi}$ is given by $\sigma_{g(i)}^2$ 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 \mathbf{Y} is then summarized as follows, where $N_{n \times p}(\mathbf{M}, \mathbf{A}, \mathbf{B})$ denotes the matrix-variate normal distribution with mean matrix \mathbf{M} and columns and row covariance matrices \mathbf{A} and \mathbf{B} :

$$\mathbf{Y} \sim N_{n \times p}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Psi}, \mathbf{K}) + N_{n \times p}(0, \boldsymbol{\Phi}, \mathbf{I}_n). \quad (4)$$

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 $\text{vec}(\cdot)$ 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:

$$\text{vec}(\mathbf{Y}) \sim N_{np}(\text{vec}(\mathbf{X}\boldsymbol{\beta}), \boldsymbol{\Psi} \otimes \mathbf{K} + \boldsymbol{\Phi} \otimes \mathbf{I}_n).$$

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 $\mathbf{K} = \mathbf{H}_K \mathbf{S}_K \mathbf{H}'_K$. This decomposition is calculated with a computational complexity of $O(n^3)$. Let \mathbf{L} be a $p \times p$ matrix that diagonalizes both Ψ and Φ , such that $\mathbf{L}\Psi\mathbf{L}' = \mathbf{I}$ and $\mathbf{L}\Phi\mathbf{L}' = \mathbf{D}$, a diagonal matrix. This bidiagonalization can be accomplished in $O(p^3)$ (details are in the *Diagonalizing two matrices* section below). We then define the matrix $\mathbf{M} = (\mathbf{L} \otimes \mathbf{H}'_k)$. The transformed data vector \mathbf{Y}_T is defined as $\mathbf{Y}_T = \mathbf{M} \text{vec}(\mathbf{Y})$. This transformed vector has the following distribution:

$$\mathbf{Y}_T \sim N(\mathbf{M} \text{vec}(\mathbf{X}\beta), \mathbf{I} \otimes \mathbf{S}_k + \mathbf{D} \otimes \mathbf{I}).$$

The log likelihood of \mathbf{Y}_T is then given as follows:

$$\begin{aligned} L(\mathbf{Y}_T | \mathbf{X}\beta, \Psi, \mathbf{K}, \Phi) &= -\frac{np}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{I} \otimes \mathbf{S}_k + \mathbf{D} \otimes \mathbf{I}| \\ &\quad - \frac{1}{2} (\mathbf{M} \text{vec}(\mathbf{Y}_T - \mathbf{X}\beta))' (\mathbf{I} \otimes \mathbf{S}_k + \mathbf{D} \otimes \mathbf{I})^{-1} \\ &\quad \times (\mathbf{M} \text{vec}(\mathbf{Y}_T - \mathbf{X}\beta)) + \log(|\mathbf{M}|). \end{aligned}$$

To calculate the likelihood given Ψ and Φ , we first obtain the transformation matrix \mathbf{M} , which is accomplished in $O(n^3 + p^3)$. Next, we compute the transformed data vector \mathbf{Y}_T in $O(n^2p + p^2n)$. Given \mathbf{Y}_T , we obtain an estimate of β , 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 $\mathbf{Y}_T - \mathbf{M} \text{vec}(\mathbf{X}\hat{\beta})$ 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 \mathbf{M} that is computed in $O(n^3 + p^3)$ requires diagonalizing the \mathbf{K} matrix, which requires $O(n^3)$ time and needs to be performed only once for the complete analysis of the data set. Similarly, the transformed data vector \mathbf{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 that is being considered. Thus the likelihood computation for each variant is dominated by $O(np^3q^2)$, utilizing the quantities that were computed once. In addition, in many scenarios 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 eigendecompositions (Piepho *et al.* 2012).

This assumption is the same assumption that 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 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 similarly for REML, we define the following log-likelihood function. For a standard multivariate normal vector \mathbf{y} with distribution $N(\mathbf{T}\alpha, \Theta)$, where \mathbf{T} is $n \times q$, the REML is $LL_{\text{REML}} = LL_{\text{ML}} + (1/2)[q \ln(2\pi) + \ln(|\mathbf{T}'\mathbf{T}|) - \ln(|\mathbf{T}'\Theta^{-1}\mathbf{T}|)]$ (Kang *et al.* 2008). Given this standard result, we define the REML log-likelihood for the mvLMM in the following:

$$\begin{aligned} LL_{\text{REML}} &= LL_{\text{ML}} \\ &\quad + \frac{1}{2} [q \ln(2\pi) + \ln(|(\mathbf{L}' \otimes (\mathbf{H}'_k \mathbf{X})') (\mathbf{L} \otimes \mathbf{H}'_k \mathbf{X})|) \\ &\quad - \ln(|(\mathbf{L}' \otimes (\mathbf{H}'_k \mathbf{X})') (\mathbf{I} \otimes \mathbf{S}_k + \mathbf{D} \otimes \mathbf{I})^{-1} \\ &\quad \times (\mathbf{L} \otimes \mathbf{H}'_k \mathbf{X})|)]]. \end{aligned}$$

The computational cost of the operations required to define LL_{REML} does not change the order of the computational complexity.

Estimating genetic correlation

To evaluate the likelihood function in Equation 5, we obtain estimates for the parameters Ψ and Φ . 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, to identify the optimal values of the variance parameters $\sigma_{g(i)}^2$ and $\sigma_{e(i)}^2$. Holding these parameters constant, we perform a two-dimensional global grid search 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.

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 in [Supporting Information, File S1, Figure S1, Figure S2, and Table S1](#) 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 ML parameter is given by the inverse of 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 Fisher's information matrix.

Association analysis

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, and 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 β from Equation 5. First, we obtain the maximum-likelihood parameters for Ψ and Φ 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 β , using the following result.

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

$$\begin{aligned} & (\mathbf{L} \otimes \mathbf{H}'_k) \text{vec}(\mathbf{X}\beta) \\ &= \text{vec}(\mathbf{H}'_k \mathbf{X} \beta \mathbf{L}') \\ &= \text{vec}(\mathbf{X}^* \beta \mathbf{L}') \\ &= (\mathbf{L} \otimes \mathbf{X}^*) \text{vec}(\beta). \end{aligned}$$

Here we have let $\mathbf{X}^* = \mathbf{H}'_k \mathbf{X}$. By denoting $\text{vec}(\beta)$ as β_T , we obtain an estimate $\hat{\beta}_T$, 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:

$$\begin{aligned} \hat{\beta}_T &= [(\mathbf{L}' \otimes \mathbf{X}^{*'}) \mathbf{P}^{-1} (\mathbf{L} \otimes \mathbf{X}^*)]^{-1} (\mathbf{L}' \otimes \mathbf{X}^{*'}) \mathbf{P}^{-1} \mathbf{M} \text{vec}(\mathbf{Y}) \\ \hat{\beta} &= \text{unvec}(\hat{\beta}_T). \end{aligned}$$

Since \mathbf{P} is a diagonal matrix, $\hat{\beta}_T$ can be computed in $O(np^3q^2 + p^3q^3 + np^2q)$ 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 $\mathbf{R}\hat{\beta}_T = [\hat{\beta}_{1x}, \hat{\beta}_{2x}, \dots, \hat{\beta}_{px}]'$, 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 $\hat{\sigma}^2 = \widehat{\text{var}}(\mathbf{P}^{-1/2} \mathbf{Y}_T)$ and $\widehat{\text{var}}(\dots)$ represents the sample variance. Details of this test can be found in McCulloch and Neuhaus (1999):

$$\begin{aligned} f &= (\mathbf{R}\hat{\beta}_T)' \left(\mathbf{R} [(\mathbf{L}' \otimes \mathbf{X}^{*'}) \mathbf{P}^{-1} (\mathbf{L} \otimes \mathbf{X}^*)]^{-1} \mathbf{R}' \right)^{-1} \\ &\quad \times (\mathbf{R}\hat{\beta}_T) \cdot \frac{1}{p\hat{\sigma}^2}. \end{aligned}$$

Diagonalizing two matrices

We are given two positive semidefinite matrices Φ and Ψ 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 $\mathbf{R}'\mathbf{R} = \mathbf{\Psi}^{-1}$. Next, we obtain an eigendecomposition $\mathbf{R}\Phi\mathbf{R}' = \mathbf{Q}\mathbf{D}\mathbf{Q}'$ 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 were used in Kang *et al.* (2010) and Korte *et al.* (2012). This data set consisted of 5326 individuals that 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 minor allele frequency. Missing phenotypes are replaced with the phenotypic mean.

We use a well-studied yeast data set (Smith and Kruglyak 2008) consisting of 109 yeast strains each with 5793 gene expression measurements. Bivariate 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 5043 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

TG - LDL

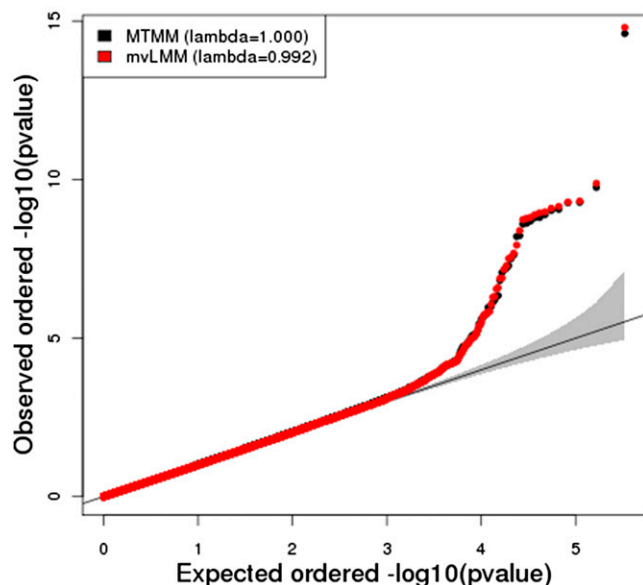


Figure 1 QQ Plot comparing MTMM and mvLMM P -values obtained when performing analysis with LDL and TG.

SNP and each pair of traits is evaluated by assuming that under the null hypothesis the SNP does not affect 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 the MTMM a 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 the mvLMM compared to the MTMM, which we suspect is because the mvLMM finds the actual parameters that maximize the likelihood. We also apply our method to analyze all four traits simultaneously and the results are shown in Table 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*.

Table 1 Genetic correlation estimates in the Finland Birth Cohort

Trait pair	Phenotypic correlation	mvLMM genetic correlation	GCTA genetic correlation
HDL/CRP	-0.19	0.28 ± 0.19	0.26 ± 0.22
HDL/LDL	-0.13	-0.16 ± 0.11	-0.18 ± 0.11
HDL/TG	-0.37	-0.37 ± 0.17	-0.32 ± 0.16
LDL/CRP	0.09	0.03 ± 0.17	-0.02 ± 0.17
TG/CRP	0.21	-0.62 ± 0.26	-0.75 ± 0.41
TG/LDL	0.32	0.33 ± 0.16	0.29 ± 0.14

We compare the maximum-likelihood estimates obtained with the mvLMM with those obtained with GCTA and find that the results are very similar.

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 bivariate LMM as implemented in genome-wide complex trait analysis (GCTA) (Lee *et al.* 2012). Table 1 compares estimates of genetic correlation obtained with GCTA and the mvLMM. When we compare our results to those of GCTA, we find that the two methods yield similar results, with genetic correlation estimates falling <1 standard deviation from one another. In addition, the running time for the classical approach was ~ 35 min, while the running time for the mvLMM was on average ~ 12 sec, given a one-time cost of 2.5 min shared across pairs of traits.

Bivariate 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 (Stuart *et al.* 2003; Subramanian *et al.* 2005; Ghanzalpour *et al.* 2006; Lee *et al.* 2006). 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 (Ghanzalpour *et al.* 2006; Lee *et al.* 2006). However, correlation estimates from gene expression levels measure the combined effect of both the genetic and the environmental components. Our methodology allows us 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 data set (Smith and Kruglyak 2008) consisting of 109 yeast strains each with 5793 gene expression measurements, we perform a bivariate analysis, estimating genetic correlations for all 5793 chose 2 gene expression pairs.

Table 2 Joint analysis of all traits compared to all pairwise combinations

rs ID	All	HDL_CRP	HDL_LDL	HDL_TG	LDL_CRP	TG_CRP	TG_LDL
rs3764261	3.115800e-01	2.610400e-31	6.167100e-31	7.179700e-33	3.857500e-01	2.301000e-01	2.475900e-01
rs1532624	5.934000e-01	7.134300e-24	1.286400e-23	1.477200e-24	4.096300e-01	1.467800e-01	1.844000e-01
rs2794520	2.936500e-13	4.404900e-22	3.241900e-01	2.812900e-01	2.030400e-22	5.021100e-23	8.300500e-01
rs7499892	3.460300e-02	3.303200e-16	1.553200e-16	1.935800e-20	6.642000e-01	3.166200e-01	6.211500e-01
rs2592887	3.707500e-10	6.883400e-17	9.482500e-02	1.104900e-01	5.918400e-17	8.024200e-17	7.710100e-01
rs646776	5.625600e-02	6.116800e-02	2.055800e-14	2.914200e-01	2.389500e-15	2.819600e-01	1.587700e-15
rs1532085	9.795800e-01	2.046400e-11	2.626800e-11	2.063700e-15	8.229500e-01	2.304300e-01	2.445000e-01
rs1811472	6.488000e-10	1.028000e-14	1.075200e-01	1.334100e-01	7.085600e-15	8.336700e-15	7.862900e-01
rs12093699	3.487700e-08	2.803100e-14	8.704700e-01	9.775600e-01	1.324500e-13	5.136100e-14	8.555500e-01
rs2650000	2.150800e-06	3.117800e-11	4.840500e-01	5.395800e-01	1.412700e-11	1.175100e-11	7.312200e-01
rs6728178	1.348000e-02	3.726700e-06	3.718900e-11	8.567700e-09	1.192200e-07	1.192900e-06	5.170100e-10
rs6754295	1.244400e-02	4.028300e-06	3.838300e-11	1.715000e-08	9.608700e-08	2.323300e-06	8.179200e-10
rs693	5.549800e-02	3.253100e-02	4.795900e-11	3.963400e-03	1.410000e-10	1.015900e-02	1.324700e-10
rs7953249	6.533200e-06	2.201500e-10	4.063400e-01	4.969100e-01	1.016400e-10	8.025500e-11	7.893600e-01
rs1169300	1.736100e-05	9.062700e-10	5.325500e-01	7.450100e-01	3.118100e-10	1.515200e-10	6.399000e-01
rs2464196	1.619700e-05	9.053200e-10	6.220200e-01	7.569200e-01	4.075100e-10	1.744700e-10	7.528400e-01
rs673548	4.967500e-02	4.309800e-06	2.014500e-10	3.655500e-09	1.150800e-06	5.222900e-07	1.055400e-09
rs415799	2.000800e-01	2.320400e-07	1.493100e-07	2.216300e-10	7.457500e-01	2.088500e-01	3.640400e-01
rs676210	5.072700e-02	5.251900e-06	2.883700e-10	5.535600e-09	1.364200e-06	7.256100e-07	1.583900e-09
rs174546	1.379500e-01	1.590200e-01	8.556400e-07	9.688800e-03	1.623200e-05	1.427300e-02	4.819700e-10
rs102275	1.678600e-01	1.112900e-01	5.655100e-07	9.205100e-03	1.723700e-05	1.722200e-02	7.111600e-10
rs1260326	4.928500e-01	1.036300e-01	2.940800e-01	7.494900e-10	8.537200e-02	1.110700e-09	1.140400e-09
rs261336	6.096900e-01	1.066600e-04	1.045000e-03	9.195300e-10	7.777300e-02	2.454000e-04	5.129400e-04
rs174537	1.410500e-01	1.549200e-01	1.474200e-06	1.113400e-02	3.039100e-05	1.637500e-02	1.306000e-09
rs1535	1.565800e-01	1.949600e-01	1.776900e-06	1.539500e-02	2.517300e-05	2.155300e-02	1.698300e-09
rs174556	6.854800e-02	3.880900e-01	1.614800e-06	5.138000e-02	5.632100e-06	5.600000e-02	1.846800e-09
rs10096633	7.343600e-01	1.076800e-05	1.327200e-05	2.542900e-09	6.869700e-01	7.121200e-08	2.132500e-08
rs735396	4.019400e-04	3.576200e-08	4.346600e-01	4.457100e-01	1.036900e-08	2.650500e-09	4.197800e-01
rs3923037	1.198500e-03	2.762800e-03	3.162700e-08	1.440400e-06	9.422300e-07	6.535200e-06	4.137700e-09
rs2126259	1.323700e-02	4.359000e-09	9.557500e-08	1.423400e-04	5.889200e-06	2.068100e-04	6.961300e-04
rs9989419	9.182800e-01	1.922700e-08	1.983700e-08	4.919800e-09	9.527800e-01	8.393700e-01	8.597400e-01
rs780094	6.157600e-01	2.610900e-01	6.568900e-01	7.159300e-09	2.491000e-01	2.042200e-08	1.189700e-08
rs11668477	4.263400e-01	4.826000e-02	8.350000e-09	1.810800e-02	2.422300e-08	4.300100e-02	3.161500e-08
rs11265260	7.116800e-06	4.158100e-08	2.523900e-02	2.956100e-02	1.047600e-08	9.316100e-09	3.308700e-01
rs1800961	6.390900e-01	1.094000e-08	1.636900e-07	1.074400e-07	1.981400e-01	1.465700e-03	1.085900e-02
rs754524	4.032900e-02	1.337000e-01	1.529600e-08	1.141300e-01	3.017000e-08	3.648500e-01	2.833300e-08
rs2075650	7.106400e-03	3.324200e-04	5.500700e-04	3.998700e-04	2.922700e-07	2.092800e-08	1.028800e-05
rs255049	8.151200e-01	2.079900e-07	2.294600e-08	1.371900e-07	3.933200e-01	4.514200e-01	8.102700e-02
rs166358	5.721200e-01	7.791800e-07	8.254800e-07	3.621500e-08	4.471400e-02	5.396600e-01	1.480400e-02

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. ID, identification number.

Within this data set 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 2A, 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 toward higher correlations with respect to all pairs, giving an indication of coregulation. However, most of the pairs have correlation <0.5 . Figure 2B shows the same plot generated using the total proportion of the phenotypic covariation attributed to genetics. In Figure 2B, we observe that the estimated genetic correlation for hotspot pairs is dramatically skewed toward 1. In fact, most of the pairs have a genetic covariance >0.7 . This result suggests that the estimated genetic correlations on average give a stronger indication of coregulation compared to the phenotypic correlation.

Discussion

In this article, we introduced a method for performing multitrait genome-wide association analysis and for the estimation of the genetic correlation. Our method is based on classical theory, but

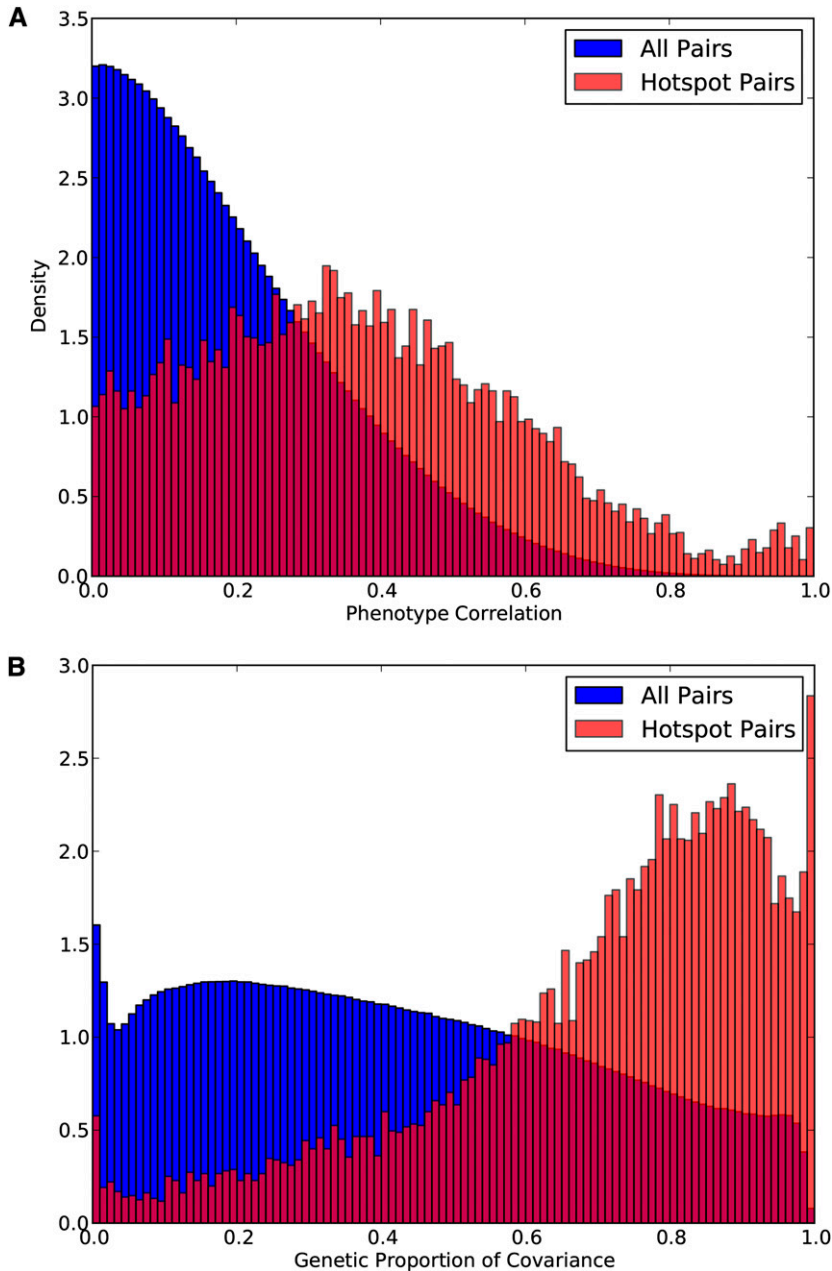


Figure 2 Comparison of the phenotypic correlation with the total proportion of the correlation accounted for by genetics for all gene pairs and for gene pairs from regulatory hotspots. We compare the phenotypic correlation with the total proportion of correlation accounted for by genetics to assess the ability of the genetic correlation to differentiate gene pairs that are coregulated. 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 coregulated pairs better than the overall trait correlation.

introduces a computational advance that makes it much faster, reducing running time >10 -fold when compared with the classic approach. We have shown that our method achieves similar results to those 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 coregulation. 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 coregulated 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 that 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 of what is the best way to analyze a data set 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 are suspected shared genetic components and 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.

Acknowledgments

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 (NIH) 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 NIH training grant T32MH073526. E.E. is supported in part by the NIH BD2K award, U54EB020403. We acknowledge the support of the National Institute of Neurological Disorders and Stroke Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691).

Literature Cited

- Avery, C. L., Q. He, K. E. North, J. L. Ambite, and E. Boerwinkle *et al.*, 2011 A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. *PLoS Genet.* 7(10): e1002322.
- Bello, N. M., J. P. Steibel, and R. J. Tempelman, 2012 Hierarchical Bayesian modeling of heterogeneous cluster- and subject-level associations between continuous and binary outcomes in dairy production. *Biom. J.* 54(2): 230–48.
- Braun, T., L. Been, A. Singhal, J. Worsham, S. Ralhan, and G. Wander *et al.*, 2012 A replication study of GWAS-derived lipid genes in Asian Indians: the chromosomal region 11q23. 3 harbors loci contributing to triglycerides. *PLoS ONE* 7(5): e37056.
- de Los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen, 2013 Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* 9(7): e1003608.
- Ducrocq, V., and B. Besbes, 1993 Solution of multiple trait animal models with missing data on some traits. *J. Anim. Breed. Genet.* 110(1–6): 81–92.
- Falconer, D., 1981 *Introduction to Quantitative Genetics*, Ed. 2. Longman, New York.
- Ferreira, M. A. R., and S. M. Purcell, 2009 A multivariate test of association. *Bioinformatics* 25(1): 132–133.
- Ghanzalpour, A., S. Doss, B. Zhang, S. Wang, and C. Plaisier *et al.*, 2006 Integrating genetic and network analysis to characterize gene related to mouse weight. *PLoS Genet.* 2(8): e130.
- Gilmour, A. R., R. Thompson, and B. R. Cullis, 1995 Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 51(4): 1440–1450.
- Gupta, A., and D. Nagar, 2000 *Matrix Variate Distributions*, Vol. 104. Chapman & Hall/CRC, Boca Raton, FL.
- Heller, R., M. Schena, A. Chai, D. Shalon, and T. Bedilion *et al.*, 1997 Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA* 94(6): 2150–2155.
- Henderson, C. R., and R. L. Quaas, 1976 Multiple trait evaluation using relatives' records. *J. Anim. Sci.* 43(6): 1188.
- Kang, H., N. Zaitlen, C. Wade, A. Kirby, and D. Heckerman *et al.*, 2008 Efficient control of population structure in model organism association mapping. *Genetics* 178: 1709–1723.
- Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, and S.-Y. Y. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42(4): 348–54.
- Kelly, A. M., B. R. Cullis, A. R. Gilmour, J. A. Eccleston, and R. Thompson, 2009 Estimation in a multiplicative mixed model involving a genetic relationship matrix. *Genet. Sel. Evol.* 41: 33.
- Korol, A., Y. Ronin, A. Itskovich, J. Peng, and E. Nevo, 2001 Enhanced efficiency of quantitative trait loci mapping analysis based on multivariate complexes of quantitative traits. *Genetics* 157: 1789–1803.
- Korte, A., B. J. Vilhjálmsson, V. Segura, A. Platt, and Q. Long *et al.*, 2012 A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* 44: 1066–1071.
- Lee, S. H., J. Yang, M. E. Goddard, P. M. Visscher, and N. R. Wray, 2012 Estimation of pleiotropy between complex diseases using SNP-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* 28(19): 2540–2542.
- Lee, S. I., D. Pe'er, A. M. Dudlet, G. M. Church, and D. Koller, 2006 Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc. Natl. Acad. Sci. USA* 103(38): 14062–14067.
- Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, and R. I. Davidson *et al.*, 2011 FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8(10): 833–835.
- Liu, Y., Y. Pei, J. Liu, F. Yang, and Y. Guo *et al.*, 2009 Powerful bivariate genome-wide association analyses suggest the SOX6 gene influencing both obesity and osteoporosis phenotypes in males. *PLoS ONE* 4(8): e6827.
- Malosetti, M., J. M. Ribaut, M. Vargas, J. Crossa, and F. A. v. Eeuwijk, 2008 A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize. *Euphytica* 161(1–2): 241–257.
- McCulloch, C., and J. Neuhaus, 1999 *Generalized Linear Mixed Models*. Wiley Online Library, New York, NY.
- Mrode, R., and R. Thompson, 2005 *Linear Models for the Prediction of Animal Breeding Values*, Ed. 2. CABI, Cambridge, MA.
- Piepho, H. P., J. O. Ogutu, T. Schulz-Streeck, B. Estaghirou, A. Gordillo *et al.*, 2012 Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. *Crop Sci.* 52(3): 1093–1104.

- Sabatti, C., S. K. Service, A. L. Hartikainen, A. Pouta, and S. Ripatti *et al.*, 2008 Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* 41(1): 35–46.
- Searle, S., G. Casella, and C. McCulloch, 1992 *Variance Components*. John Wiley & Sons, New York.
- Smith, E. N., and L. Kruglyak, 2008 Gene-environment interaction in yeast gene expression. *PLoS Biol.* 6(4): e83.
- Stegle, O., C. Lippert, J. M. Mooij, N. D. Lawrence, and K. M. Borgwardt, 2011 Efficient inference in matrix-variate Gaussian models with iid observation noise, pp. 630–638 in *Advances in Neural Information Processing Systems 24 (NIPS 2011)*.
- Stuart, J. M., E. Segal, D. Koller, and S. K. Kim, 2003 Gene-coexpression network for global discovery of conserved genetic modules. *Science* 302(5634): 249–255.
- Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, and B. L. Ebert *et al.*, 2005 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 102(43): 15545–50.
- 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): 933–953.
- Welham, S. J., and R. Thompson, 1997 Likelihood ratio tests for fixed model terms using residual maximum likelihood. *J. R. Stat. Soc. Ser. B Methodol.* 59(3): 701–714.
- Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, and A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42(7): 565–569.
- Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44(7): 821–824.

Communicating editor: S. Sen

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.171447/-/DC1>

Efficient Multiple-Trait Association and Estimation of Genetic Correlation Using the Matrix-Variate Linear Mixed Model

Nicholas A. Furlotte and Eleazar Eskin

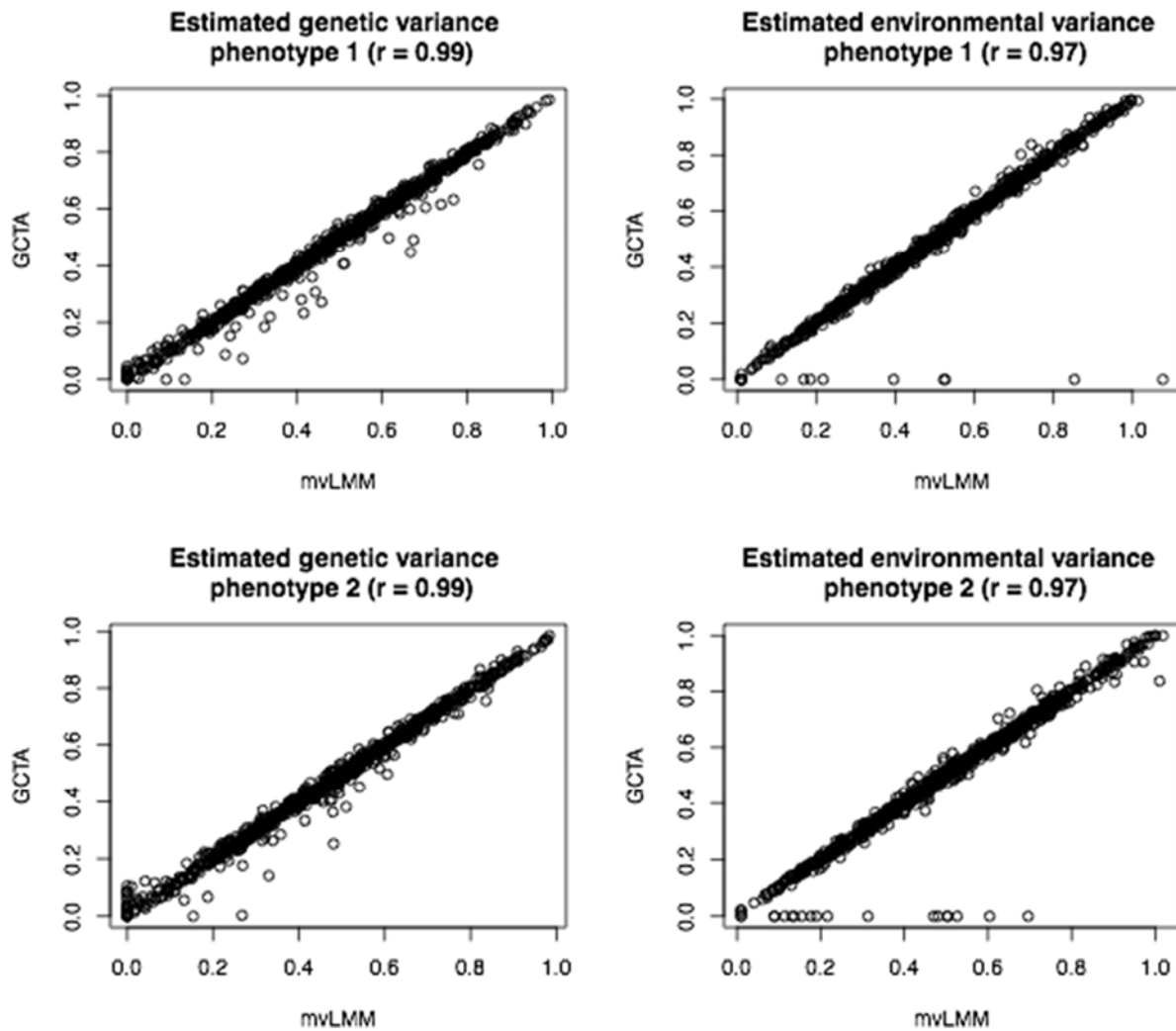


Figure S1 Comparison of variance estimates between mvLMM and GCTA. We simulated 1,000 pairs of phenotypes under the multiple phenotype model assumed by mvLMM and GCTA with the genetic variance for phenotype 1 and 2 set to 0.50. The figure shows that the parameter estimates from both methods are highly concordant. However, we note that GCTA has the potential to predict very large genetic variance such as $1e16$. We have filtered these cases to produce the plots.

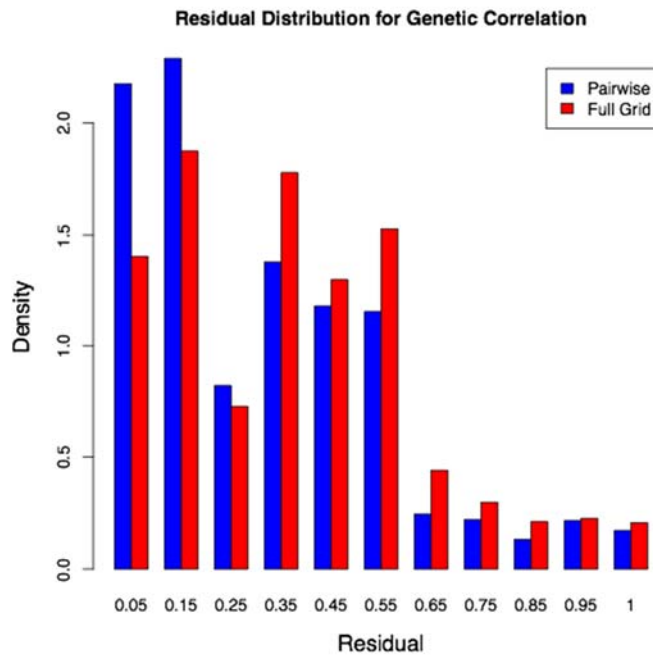
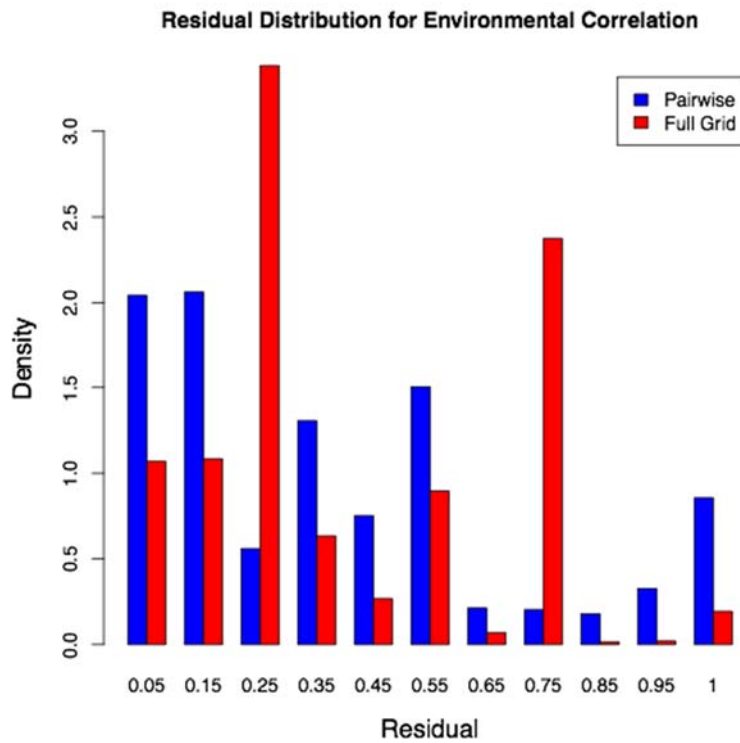


Figure S2 Distribution of residuals for the genetic and environmental correlations. We simulate 1000 data sets each with three correlated phenotypes, with a genetic and environmental correlation of 0.25. We employ both a pairwise approach and a full grid search approach to estimate the correlations and plot the distribution of the absolute value of the residuals here.

Table S1 Comparison of genetic and environmental variance components estimated using mvLMM and GCTA. The variance estimates are highly concordant between the two methods across pairs of traits from the Northern Finland Birth Cohort.

Phenotype Pair	Genetic Variance (trait 1/ trait 2)	Environment Variance (trait1 / trait 2)	Genetic Variance (trait 1/ trait 2)	Environment Variance (trait 1 / trait 2)
HDL-CRP	0.037 / 0.272	0.08 / 2.03	0.038 / 0.276	0.084 / 2.038
HDL-LDL	0.037 / 0.261	0.08 / 0.441	0.0384 / 0.283	0.084 / 0.458
HDL-TG	0.037 / 0.032	0.08 / 0.17	0.038 / 0.030	0.084 / 0.182
LDL-CRP	0.261 / 0.272	0.441 / 2.03	0.282 / 0.272	0.459 / 2.033
TG-CRP	0.032 / 0.272	0.17 / 2.03	0.030 / 0.275	0.182 / 2.03
TG-LDL	0.032 / 0.261	0.17 / 0.441	0.030 / 0.276	0.182 / 0.464

File S1

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 \mathbf{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 $N(0, \frac{1}{m}\sigma_{g(k)}^2)$. Let $\mathbf{g}_k = \mathbf{W}\mathbf{b}_k$, where $\mathbf{b}_k = [b_{1k} \ b_{2k} \ \dots \ b_{mk}]'$. Therefore, the variance of \mathbf{g}_k is given by equation (1). Thus, LMM-based population structure correction may be viewed as a basic linear model, while treating the SNP effects as random effects.

$$\begin{aligned} \text{var}(g_k) &= \frac{\mathbf{W}\mathbf{W}'}{m} \sigma_{g(k)}^2 \\ &= \mathbf{K} \sigma_{g(k)}^2 \end{aligned}$$

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

$$\begin{bmatrix} \mathbf{b}_i \\ \mathbf{b}_j \end{bmatrix} \sim N\left(\mathbf{0}, \frac{1}{m} \begin{bmatrix} \sigma_{g(i)}^2 \mathbf{I} & \rho_{ij} \sigma_{g(i)} \sigma_{g(j)} \mathbf{I} \\ \rho_{ij} \sigma_{g(i)} \sigma_{g(j)} \mathbf{I} & \sigma_{g(j)}^2 \mathbf{I} \end{bmatrix}\right)$$

To obtain the joint distribution of \mathbf{g}_i and \mathbf{g}_j , we apply the following linear transformation.

$$\begin{aligned} \begin{bmatrix} \mathbf{g}_i \\ \mathbf{g}_j \end{bmatrix} &= \begin{bmatrix} \mathbf{W}\mathbf{b}_i \\ \mathbf{W}\mathbf{b}_j \end{bmatrix} = \begin{bmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix} \begin{bmatrix} \mathbf{b}_i \\ \mathbf{b}_j \end{bmatrix} \sim \\ N\left(\mathbf{0}, \frac{1}{m} \begin{bmatrix} \sigma_{g(i)}^2 \mathbf{W}\mathbf{W}' & \rho_{ij} \sigma_{g(i)} \sigma_{g(j)} \mathbf{W}\mathbf{W}' \\ \rho_{ij} \sigma_{g(i)} \sigma_{g(j)} \mathbf{W}\mathbf{W}' & \sigma_{g(j)}^2 \mathbf{W}\mathbf{W}' \end{bmatrix}\right) \\ &= N\left(\mathbf{0}, \begin{bmatrix} \sigma_{g(i)}^2 \mathbf{K} & \rho_{ij} \sigma_{g(i)} \sigma_{g(j)} \mathbf{K} \\ \rho_{ij} \sigma_{g(i)} \sigma_{g(j)} \mathbf{K} & \sigma_{g(j)}^2 \mathbf{K} \end{bmatrix}\right) \end{aligned}$$

This result is consistent with the proposed model in the previous section.

The same basic logic is easily applied to derive the $cov(\mathbf{e}_i, \mathbf{e}_j)$. By substituting \mathbf{W} for \mathbf{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.

$$\begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_j \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_{e(i)}^2 \mathbf{I} & \lambda_{ij} \sigma_{e(i)} \sigma_{e(j)} \mathbf{I} \\ \lambda_{ij} \sigma_{e(i)} \sigma_{e(j)} \mathbf{I} & \sigma_{e(j)}^2 \mathbf{I} \end{bmatrix}\right)$$

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 $\mathbf{W}_c \mathbf{W}'_c$, where \mathbf{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 \mathbf{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 Figure S1, 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 (eg. $1e16$). 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 Table S1 that the genetic and environmental variance estimates for the Finland Cohort and 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 Figure S2, 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).

LITERATURE CITED

- GODDARD, M. E., N. R. WRAY, K. VERBYLA, and P. M. VISSCHER, 2009 Estimating effects and making predictions from genome-wide marker data. *Statistical Science* *24*(4): 517–529.
- HAYES, B. J., P. M. VISSCHER, and M. E. GODDARD, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Res* *91*(1): 47–60.

LISTGARTEN, J., C. LIPPERT, C. M. KADIE, R. I. DAVIDSON, E. ESKIN, D. HECKERMAN, J. LISTGARTEN, C. LIPPERT, C. M. KADIE, R. I. DAVIDSON, E. ESKIN, and D. HECKERMAN, 2012 Improved linear mixed models for genome-wide association studies. *Nature Methods* 9(6): 525.

YANG, J., B. BENYAMIN, B. P. MCEVOY, S. GORDON, A. K. HENDERS, D. R. NYHOLT, P. A. MADDEN, A. C. HEATH, N. G. MARTIN, G. W. MONTGOMERY, M. E. GODDARD, and P. M. VISSCHER, 2010 Common SNPs explain a large proportion of the heritability for human height. *Nature Genet* 42(7): 565–569. doi:10.1038/ng.608.