Published in final edited form as:
Hum Hered. 2013 ; 76(2): 53–63. doi:10.1159/000356016.

A Rapid Genome-wide Gene-based Association Test with Multivariate Traits

Saonli Basu[1,*], **Yiwei Zhang**[1], **Debashree Ray**[1], **Michael B. Miller**[2], **William G. Iacono**[2], and **Matt McGueM**[2]

[1]Division of Biostatistics, School of Public Health, University of Minnesota.

[2]Department of Psychology, University of Minnesota.

Abstract

Objectives: A gene-based genome-wide association study (GWAS) provides a powerful alternative to the traditional single SNP association analysis due to its substantial reduction in the multiple testing burden and possible gain in power due to modeling multiple SNPs within a gene. A gene-based association analysis on multivariate traits is often of interest, but imposes substantial analytical as well as computational challenges to implement it at a genome-wide level.

Methods: We have proposed a rapid implementation of multivariate multiple linear regression approach (RMMLR) in unrelated individuals as well as in families. Our approach allows for covariates. Moreover the asymptotic distribution of the test statistic is not heavily influenced by the linkage disequilibrium (LD) among the SNPs and hence can be used efficiently to perform a gene-based GWAS. We have developed corresponding R package to implement such multivariate gene-based GWAS with this RMMLR approach.

Results: We compare through extensive simulation several approaches for both single and multivariate traits. Our RMMLR maintains correct type-I error level even for set of SNPs in strong LD. It also has substantial gain in power to detect a gene when it is associated with a subset of the traits. We have also studied their performance on Minnesota Center for Twin Family Research dataset.

Conclusions: In our overall comparison, our RMMLR approach provides an efficient and powerful tool to perform a gene-based GWAS with single or multivariate traits and maintains the type I error appropriately.

Keywords

Multivariate regression; Gene-based genome-wide association studies; Multivariate trait

1 Introduction

The rapid progress in genotyping technology has greatly facilitated our understanding of the genetic aspect of various diseases. Several genome-wide association studies (GWAS) have

[*]Corresponding Author: Saonli Basu. Division of Biostatistics, School of Public Health, Universityof Minnesota. 420 Delaware St SE, Minneapolis, MN 55455. Tel. +1 612 624 2135, Fax. +1 612 626 0660, saonli@umn.edu.

been published on various complex diseases, where genotype data on a large number of single nucleotide polymorphisms (SNPs) are collected to study the association between these SNPs and a disease (Visscher et al., 2012). Currently, most existing statistical methods of detecting genetic association focus on analyzing SNPs one-by-one (Roeder et al., 2005).

Although, new SNPs are found to be associated with these diseases, little of the genetic risk or heritability for these diseases is explained by the additive effects of these SNPs. Studying one SNP at a time may not be sufficient to understand complex diseases. Analyses focusing on individual SNPs may have several limitations: (a) A large number of comparisons increases the probability of false positives and a stringent correction could lead to loss of power. (b) each SNP alone may have very little effect on risk of disease (Lesnick et al., 2007), but multiple causal variants together may increase the risk substantially. A single-SNP analysis may not be optimal here; especially in presence of multi-SNP effects.

Hence, gene-based association analysis is often of interest in genome-wide association studies. These approaches substantially reduce the multiple testing burden at a GWAS scale. Moreover by pooling information on multiple SNPs within a gene, one could gain substantial power in detecting association. Current gene-based association approaches are mostly limited to combining single SNP association p-values to propose a gene-based test statistic (Neale and Sham, 2004; Ballard et al., 2010; Wang et al., 2007). One limitation of these gene-based statistics is that combining single SNP p-values ignores the potential joint effect of multiple SNPs within a gene. Joint modeling of the SNPs within a gene could have substantial gain in power over these existing gene-based tests. Moreover, the current approaches typically require permutation or simulation testing to account for the correlation between the SNPs as well as the gene size. VEGAS (Liu et al., 2010) and PLINK set-based test (Purcell et al., 2007) are two such available approaches for performing gene-based tests. They can be computationally very intensive to perform a gene-based GWAS, since they require a large number of permutations to correct for multiple testing. Hence their usefulness to conduct a gene-based GWAS is somewhat limited. Moreover, these approaches have been implemented for univariate phenotypes.

In the study of a complex disease, several correlated traits are often measured as risk factors for the disease. For example, in the study of thrombosis, the intermediate correlated phenotypes such as Factor VII, VIII, IX, XI, XII, and von Willebrand factor jointly predict the risk of developing thrombosis (Souto et al., 2000; Germain et al., 2011). There may be genetic variants affecting several of these traits. Analyzing multiple disease-related phenotypes could potentially increase power to detect association of SNPs/genes with a disease. Moreover this joint analysis could reveal some pleiotropic genes involved in the biological development of the disease. It has been shown that modeling multivariate phenotypes may increase the power over analyzing individual phenotypes separately in genetic association studies (Klei et al., 2008; Lange et al., 2004).

Few approaches have been developed to perform association analysis with multivariate traits at a GWAS level. O'Reilly et al. (2012) proposed MultiPhen to detect association between multivariate traits and a SNP with unrelated individuals. It uses ordinal regression to regress a SNP on a collection of phenotypes and tests whether all regression parameters

corresponding to the phenotypes in the model are significantly different from zero. van der Sluis et al. (2013) proposed a method TATES for testing association between multiple traits and multiple SNPs using extended Simes procedure on the univariate p-values derived from univariate trait and single SNP association analysis. Recently, Maity et al. (2012) proposed a kernel machine method for joint analysis of multiple genetic variants, which treats the SNP effects as a random effect in the model. Their test for association between multiple SNPs and the phenotypes is equivalent to testing the variance components corresponding to the random effect in a multivariate linear mixed model. Implementation of this approach requires parametric bootstrapping to estimate the distribution of the test statistic and could be computationally intensive at a GWAS level. Also this method has been proposed for unrelated individuals.

Recently, the advantage of using canonical correlation analysis to perform gene-based tests on multivariate phenotypes has been elaborately discussed in Tang and Ferreira (2012). Previously, Ferreira and Purcell (2009) proposed a multivariate test of association based on canonical correlation analysis (CCA) to simultaneously test the association between a single SNP and multiple phenotypes. Their CCA approach is equivalent to multivariate analysis of variance (MANOVA) or more generally the Wilk's lambda test in multivariate multiple linear regression (MMLR) approach (Muller et al., 1984). According to O'Reilly et al. (2012), MultiPhen and CCA performed very similarly in terms of power to detect causal variants except in case of low frequency SNPs and non-normal phenotypes. Both O'Reilly et al. (2012) and van der Sluis et al. (2013) found significantly high power of CCA in case of a subset of traits were associated with the causal variant or gene. One major advantage of CCA is that it can easily be extended to incorporate multiple phenotypes as well as multiple SNPs (such as a gene). CCA provides the opportunity to study such multivariate phenotype-genotype association analysis. But CCA cannot accommodate covariates and one needs to perform the computationally intensive permutation testing to implement CCA in family data (Ferreira and Purcell, 2009).

In this paper, we have used the equivalence between the canonical correlation analysis and multivariate multiple linear regression (MMLR) approach to analyze multiple traits as well as multiple SNPs in families. In fact, our approach can be applied to samples involving both families as well as unrelated individuals.The advantage of MMLR approach is that one can easily incorporate the covariates in the model. Another advantage of MMLR approach is that it could provide a gene-based test of association without the need for permutation testing (Tang and Ferreira, 2012). We have proposed here a rapid MMLR (RMMLR) in families. We have developed related R package to perform a multivariate gene-based GWAS with this RMMLR approach. We have compared the performance of this RMMLR approach with several existing gene-based association analysis approaches, all of which require permutation or resampling techniques to estimate the null distribution of the test statistic in presence of correlated SNPs or correlated phenotypes.

2 Methods

2.1 Models

For a set of pedigrees each including one or more related individuals, let *ylij* denote the measured phenotype *l* on individual *j* in pedigree *i* (*i* = 1, … , *m*, *j* = 1 , … , n_i, $\sum_i = n$, and *l* = 1, … , *L*). Let X_{ijk} denote the additive genotype score of *k*-th SNP (*k* = 1, 2, … , *K*) with alleles 'A' and 'a' of individual *j* in pedigree *i*, and X_{ijk} can take values of 0, 1, or 2 depending on the number of minor allele 'A' for individual *j* in *i* th pedigree. Let vector $\boldsymbol{Y}_{li}$ = {*yli*1, … , $ylin_i$} denote the vector of *l*-th phenotype on individuals in *i*-th pedigree, and $\boldsymbol{X}_{ik}$ = {X_{i1k}, … , X_{in_ik}} denote the vector of genotypes of *k*-th SNP from pedigree *i*. Let $\boldsymbol{C}_i$ be a $n_i \times p$ matrix that contains the *p* covariates of pedigree *i*. We are interested in methods to conduct gene-based association analysis in pedigrees with one (unrelated individuals) or more related members with univariate or multivariate quantitative phenotypes.

For a quantitative phenotype *l*, we can do a single SNP association test for the *k*-th SNP using the following linear mixed effect model,

$$\boldsymbol{Y}_{li} = \alpha_l + \boldsymbol{X}_{ik}\beta_{lk} + \boldsymbol{C}_i\beta_{lc} + \epsilon_{li} \quad (1)$$

$$= \boldsymbol{X}_i\beta_l + \epsilon_{li},\ i = 1, \ldots,\ m, \quad (2)$$

where α_l is the population mean, β_{lk} is the additive fixed effect of the *k*-th SNP on *l*-th trait, $\boldsymbol{\beta}_{lc}$ is a size *p* vector of the covariate effects, and $\boldsymbol{\varepsilon}_{li}$ is the random residual term for *l*-th trait which is modeled as

$$\epsilon_{li} \overset{ind}{\sim} MVN\left(0, \boldsymbol{V}_{li}\right), \quad (3)$$

assuming independent pedigrees. The $n_i \times n_i$ matrix $\boldsymbol{V}_{li}$ is the variance-covariance matrix for *l*-th trait. To simplify the notation as shown on the right hand side of equation (1), the observed data on the (*p* + 1) fixed predictors (1 SNP, *p* covariates) and the intercept corresponding to pedigree *i* are contained in the $n_i \times (p + 2)$ design matrix $\boldsymbol{X}_i$, and the (*p* + 2) parameters are contained in the vector $\boldsymbol{\beta}_l$ for *l*-th trait.

A maximum likelihood estimator of $\boldsymbol{V}_{li}$ is calculated by $\hat{\boldsymbol{V}}_{li}$ by assuming multivariate normal distribution of the phenotype and certain parametric structure of $\boldsymbol{V}_{li}$ and considering the above model in Equation 2 but without the SNP term $\boldsymbol{X}_{ik}\beta_{lk}$. As shown in Li et al. (2011b), this approximation technique produces minimally biased estimates of $\boldsymbol{V}_{li}$ under the assumption that a single SNP explains a small part of the total phenotypic variance. Let $\hat{\boldsymbol{V}}_l$ be the block-diagonal matrix with $\hat{\boldsymbol{V}}_{li}$'s on its diagonal blocks, *i* =1,2, … , *m*. We used bdspackage() in R (Hornik, 2012) to compute the inverse of this block-diagonal matrix, denoted by $\hat{\boldsymbol{V}}_l^{-1}$ and a Cholesky decomposition of $\hat{\boldsymbol{V}}_l^{-1}$ as $\hat{\boldsymbol{V}}_l^{-1} = S_l S_l^T$ (Golub and Van Loan, 1996; Horn and Johnson, 1985; Trefethen and Bau, 1997). The Cholesky decomposition of the positive definite covariance matrix $\hat{\boldsymbol{V}}_l^{-1}$ yields a matrix S_l such that $\hat{\boldsymbol{V}}_l^{-1} = S_l S_l^T$, where S_l is lower triangular with positive diagonal elements. The matrix S_l is

called the 'Cholesky factor' of $\hat{\boldsymbol{V}}_l^{-1}$ and is equivalent to the square-root of $\hat{\boldsymbol{V}}_l^{-1}$, denoted by $V_l^{-1/2}$. Hence $V_l^{-1/2}\,\boldsymbol{Y}_l$ would multivariate normal distribution with mean $V_l^{-1/2}\,(\boldsymbol{X}\beta_l)$ and covariance matrix $V_l^{-1/2}V_lV_l^{-1/2}=\boldsymbol{I}$. In other words, the transformed data $V_l^{-1/2}\,\boldsymbol{Y}_l$ or $S_l\boldsymbol{Y}_l$ would produce observations following multivariate normal distribution with mean $S_l\boldsymbol{X}\boldsymbol{\beta}_l$ and covariance matrix $\boldsymbol{I}$.

Then a simple linear regression for these unrelated observations is carried out using the Cholesky-factor-transformed data,

$$Y_l^*=S_l\,\boldsymbol{Y}_l=S_l\boldsymbol{X}\beta_l+\epsilon_l^*, \quad (4)$$

where $\boldsymbol{X}$ is the $n \times (p + 2)$ matrix containing predictor values of all the m pedigrees, $\boldsymbol{Y}_l$ is the size n vector of phenotypes, $\boldsymbol{\beta}_l = (a_l, \beta_{lk}, \boldsymbol{\beta}_{lc})$ as defined in equation (2), and the residual term ϵ_l^* is distributed as $N(0, \boldsymbol{I})$. We make association calls based on the F-test statistics from the simple linear regressions in equation 4. Here we want to test the null hypothesis $H_0 : \beta_{lk} = 0$, for $k = 1, 2, \ldots, K$ and $l = 1, 2, \ldots, L$.

2.2 Combination Methods

We used two combination methods to perform gene-based association analysis. The combination methods combine the single SNP association p-values to generate test statistics for the gene-based tests. At the first stage, p-values for all SNPs within a gene are obtained through single-SNP association tests based on standard regression models adjusting for other covariates. We used a rapid feasible generalized least square (RFGLS) approach (Li et al., 2011b) to perform the single SNP association analysis in pedigrees. The second stage requires the use of a single p-value to represent the significance of each gene. We applied two methods, the Fisher's combination test and minP approach to calculate the significance of each gene.

- **minP approach:** Consider K SNPs within a gene. The minP test statistic is given by $p_{\min}=\min_{i=1}^{K}\{Kp_i\}$ for a single phenotype. We can extend it to consider testing association between genes and multiple phenotypes. For L phenotypes and K SNPs, the test statistic is given by $p_{\min}=\min_{i=1}^{KL}\{KLp_i\}$. Under the assumption of independence among the SNPs as well as among the phenotypes, this test statistic has uniform distribution between [0,1] under null. For positively correlated tests, this test tends to be conservative. There are alternative approaches such as Simes method, Sidak method (Simes, 1986; Li et al., 2011a; Sidak, 1967) that improve the conservativeness of the test. Otherwise permutation procedure or parametric bootstrap could be used to get valid p-values.

- **Fisher's Combination Test:** The test statistic is given by $\chi^2=-2\sum_{i=1}^{K}\log p_i$ for SNPs and a single phenotype, which has a chi-square distribution with $2K$ degrees of freedom under the null hypothesis when the K tests are independent. We can extend this test for multiple phenotypes as well. The test statistic for testing no association between a gene and multiple phenotypes could be given by

$\chi^{2'} = -2\sum_{i=1}^{LK}\log p_i$ for K SNPs and L phenotypes. Assuming we have independent tests here, this test statistic has a chi-square distribution with $2KL$ degrees of freedom under the null hypothesis H_0. The major issue with this test is that it becomes anti-conservative for positively correlated tests, which could be the situation for the SNPs in LD or for positively correlated multiple traits. Hence a permutation procedure or parametric bootstrap technique (equivalent to VEGAS-SUM approach) is needed if a valid p value is to be obtained. The Versatile Gene-Based Test for Genome-wide Association (VEGAS) (Liu et al., 2010) is a recent multivariate method that sums the association signal from all the SNPs within a gene and corrects the sum for LD to generate a p-value. One thing to note here is that VEGAS can only handle a single trait to perform a gene-based GWAS.

2.3 Direct modeling of multiple SNPs and multiple phenotypes

The combination methods provide a convenient way of proposing gene-based tests by utilizing the single SNP association p-values. However by directly modeling multiple correlated traits and multiple SNPs, one could detect pleiotropic effects and could gain power by jointly modeling the effects of multiple SNPs in a single model.

Multivariate multiple linear regression—Below we describe a multivariate multiple regression model (MMLR) with L traits and K SNPs for n unrelated individuals. Consider the following model for n individuals with outcome variable $\boldsymbol{Y}^* = (Y_l^*, l=1, \ldots, L)$ in Equation 4 of length L:

$$\boldsymbol{Y}_i^* = \boldsymbol{Z}_i\beta + \boldsymbol{e}_i;\ i=1, \ldots, n,$$

where $\boldsymbol{e}_i = (e_{i1}, \ldots e_{iL})$ follows multivariate normal distribution; $\boldsymbol{Y}_i^* = (Y_{i1}, \ldots Y_{iL})$ contains transformed data on L phenotypes; $\boldsymbol{Z}_i$ contains the transformed data $S_l\boldsymbol{X}_i$ on K predictors (SNPs) and intercept for the i-th individual. The vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_K)$ represents the parameters corresponding to the intercept and K SNPs. In this case the within-sample sum of squares (SS) matrix (W) is calculated as the covariance matrix of the residuals of the model, and the between-sample SS matrix is calculated as the covariance matrix of the fitted values (B). We use the Wilk's λ test, $\lambda = \frac{|W|}{|T|}$, where $T = W + B$ to test the null hypothesis $\boldsymbol{\beta} = 0$. It has been shown that for unrelated individuals, this MMLR approach is equivalent to the CCA approach mentioned before (Muller et al., 1984).

Compared with CCA, MMLR is much easier to expand to family structure and adjust for covariates. To apply it to family data, we use the computationally efficient RFGLS approach (Li et al., 2011b). We calculate the within-cluster covariance matrix for each trait, and left-multiply both the traits and the predictors separately by the inverse-square-root of the estimated covariance matrix. This transformation gives us the traits individually as independently distributed. We then compute the residuals for each trait from a model with intercept and modified covariates as well as the SNPs. Let us denote the residuals as $\boldsymbol{e}_1^*, \boldsymbol{e}_2^*, \ldots, \boldsymbol{e}_L^*$. We compute the covariance matrix of these residuals. This will be our W. In

order to calculate *T*, we compute the residuals for each trait from a model with only intercept and the modified covariates. The covariance matrix of these residuals will give us *T*. We calculate our test statistic as $\lambda = |\frac{W}{T}|$, which will be approximately distributed as *F* with degrees of freedom *df1* and *df2*, where *df1* = *KL* and $df2 = \left(n - 3/2 - \frac{K+L}{2}\right) s - \frac{KL}{2} + 1$. We refer to this approach as rapid multivariate multiple linear regression (RMMLR).

2.4 Simulations

To compare different methods, we investigated their type I error and power by simulating data on nuclear families with two parents and two offspring. For simulating genotype data on a gene, we used a simulation set-up very similar to that of Li et al. (2011a). First, we generated genotype data on 16 SNPs with different LD structures. We generated 3 independent blocks of SNPs; the first and the third block with 6 SNPs and the second block with 4 SNPs. For low linkage disequilibrium (LD) scenario, pairwise LDs (r^2) were 0.5, 0.4, 0.5 for the 3 blocks respectively; while for high LD scenario, pairwise LDs (r^2) were 0.9, 0.8, 0.9 respectively. The minor allele frequencies (MAFs) for 3 blocks were 0.2, 0.4, 0.25 respectively. Note that X_{ijk} denote the genotype of the *k*-th SNP for the *j*-th member in the *i*-th family. For *j*-th member in *i*-th family, we draw a vector (16 × 1) of latent variables U_{ij}, from multivariate normal with covariance matrix as (1) low LD structure as shown in appendix of Li et al. (2011a); (2) high LD structure as shown in appendix of Li et al. (2011a). Then by assuming HWE, we dichotomize U_{ijk} such that if $U_{ijk} < p_k^2$, then $X_{ijk}=2$; if $U_{ijk} > 1 - (1 - p_k)^2$, then $X_{ijk} = 0$; otherwise $X_{ijk}=1$, $k = 1, \ldots, 16$ and p_k is the MAF of *k*th SNP. We repeat this procedure to simulate genotypes of parents for all the nuclear families. Then following Mendel's Law, we generate genotypes of 2 offspring in each family. From each LD block of SNPs, the first SNP is selected as the causal SNP. We consider 2 different gene-lengths such that the gene (1) contains the first 3 SNPs (2) contains all the 16 SNPs.

We then simulate five correlated traits. The simulation model is

$$\boldsymbol{Y}_{li} = \boldsymbol{G}_i \beta + \alpha_{li} + e_{li},$$

where $Y_{li} = (Y_{li1}, Y_{li2}, Y_{li3}, Y_{li4})$ is the vector of phenotypes for *l*-th trait in *i*-th family. The parameter α_{li} is the polygenic random effect for *i*-th family and *l*-th trait. $\boldsymbol{G}_i$ is the matrix of genotypes of the causal SNPs, which is a 4× r matrix if we have *r* causal SNPs.

We allow correlation among different traits within each family. Suppose we have 5 traits, We then draw $\alpha_i = (\alpha_{1i1}, \alpha_{1i2}, \alpha_{1i3}, \alpha_{1i4}, \ldots, \alpha_{5i1}, \alpha_{5i2}, \alpha_{5i3}, \alpha_{5i4})$ from multivariate normal with mean 0 and covariance matrix $\sum_1$ as:

$$\begin{pmatrix} \sigma_{a1}^2 2\Phi & \rho_{ab}\sigma_{a1}\sigma_{a2} * 2\Phi & \ldots & \rho_{ab}\sigma_{a1}\sigma_{a5}2\Phi \\ \rho_{ab}\sigma_{a1}\sigma_{a2} * 2\Phi & \sigma_{a2}^2 2\Phi & \ldots & \rho_{ab}\sigma_{a2}\sigma_{a5}2\Phi \\ & & \ldots & \\ \rho_{ab}\sigma_{a1}\sigma_{a5}2\Phi & \rho_{ab}\sigma_{a2}\sigma_{a5}2\Phi & \ldots & \sigma_{a5}^2 2\Phi \end{pmatrix},$$

where Φ is the kinship matrix and ρ_{ab} is the genetic correlation between traits, which is assumed in the simulation to be the same for all traits. In our simulation, for nuclear families with two offspring, we use the kinship matrix for offspring 1, offspring 2, parent 1 and parent 2 as

$$2\Phi = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0 \\ 0.5 & 0.5 & 0 & 1 \end{pmatrix}$$

and $\sigma^2_{a1}=\sigma^2_{a2}=\sigma^2_{a3}=\sigma^2_{a4}=\sigma^2_{a5}=60$. We also simulate the environmental component e_i = (e_{1i1}, e_{1i2}, e_{1i3}, e_{1i4}, ….e_{5i1}, e_{5i2}, e_{5i3}, e_{5i4}) from multivariate normal with mean 0 and covariance matrix as:

$$\begin{pmatrix} \sigma^2_{e1}I & \boldsymbol{W} & \dots & \boldsymbol{W} \\ \boldsymbol{w} & \sigma^2_{e2}I & \dots & \boldsymbol{W} \\ & & \dots & \\ \boldsymbol{W} & \boldsymbol{W} & \dots & \sigma^2_{e5}I \end{pmatrix}$$

where σ^2_{el} represent the environmental variance of trait l, l = 1, … , 5. In majority of the simulations, we consider no environmental correlation between the trait values for the same individual, i.e $\boldsymbol{W} = \boldsymbol{I}$. Also we assume $\sigma^2_{e1}=\sigma^2_{e2}=\sigma^2_{e3}=\sigma^2_{e4}=\sigma^2_{e5}=40$. For a 3-snp gene, only the first block is used; while for the 16 snp gene, all 3 blocks are used. We simulate 5 traits with different genetic correlations between the traits: ρ_{ab} = 0.33, 0.67, 1, which corresponds to a phenotypic correlation between the traits of 0.2,0.4 and 0.6 respectively. The first SNP of every block is the causal SNP under the alternative, if the gene has covered that block.

This means that the correlations between different traits are entirely coming from the genetic effect rather than the environment effect. Lastly, we add the causal SNP(s) effect to the mean. The causal effect is obtained as $\beta_k = \sqrt{\frac{h^2_k}{2p_k(1-p_k)}}$, for the k-th causal SNP, where h^2_k is the phenotypic variance explained by the k-th causal SNP. For ρ_{ab} = 0.33 and ρ_{ab} = 0.67, we also perform 5000 simulations to study the effect environmental correlation on power. We allow for environmental correlation of r = 0.25 and r = 0.5 between each pair of traits, i.e we consider $\boldsymbol{W} = r\sigma_{el}\sigma_{el}\boldsymbol{I}$, for each pair of traits l and k, l, k = 1, .., 5.

3 Results

We first simulated data under the null hypothesis to see the impact on type I error. We simulated 100000 datasets on 1000 nuclear families under the null hypothesis of no association between a gene and 5 traits. The type I error was estimated by calculating the proportion of times the asymptotic p-value of the test statistic for each method was 0≤001, 0≤0001 and 0≤00001. We considered situations where the SNPs were in low LD and in high LD and presented here only the results for the gene with 16 SNPs. The minP statistic in low LD situation (Table 1) was not that conservative, but did get conservative in high LD

situation (Table 2). On the other hand, Fisher's test was inflated due to the positive correlation among the traits as well as the SNPs. The type I error for Fisher's test was substantially higher for the high LD situation (Table 2) as compared to the low LD (Table 1) condition due to the higher positive correlation among the SNPs. Moreover the type I error for Fisher's test was increased as the correlation among the traits increased from 0.2 to 0.6. This are known issues with minP approach and Fisher's test, but we still wanted to demonstrate the extent of the impact of correlation among traits as well as SNPs on these two popular approaches and the need to use approaches such as Liu et al. (2010), Li et al. (2011a) to correct for these issues with type-I error. On the other hand, the RMMLR approach maintained correct type-I error under both low LD and high LD situation and for different degrees of correlation among the traits.

We next simulated data under the alternative. We considered situations where the gene was associated with 1, 2, 3, 4 or all 5 traits. For each scenario, we simulated 5000 datasets on 500 nuclear families. We also considered 2 different gene-length and different LD patterns among the SNPs to see the impact of gene-length and LD on power. We reported the empirical power for the minP and Fisher's test in order to correct for any inflatedness or conservativeness in the type I error. The empirical power was calculated by counting the proportion of times the trait SNP was called significant (p-value ≤ empirical 0.001-th cut-off) out of the 5, 000 simulations. Since RMMLR approach maintained correct type I error rate, we reported the asymptotic power for the approach.

We first simulated data on 500 nuclear families for a gene of size 3 SNPs with one causal SNP explaining .6% of the phenotype variance. Table 3 shows the empirical power of minP, Fisher's test and asymptotic power of RMMLR at a type-I error level of .001. We reported here the situations where the causal SNP is associated with all the traits and the situation where 3 of the traits are associated with the SNP. In case of complete association, minP and Fisher's test had comparable power (Table 3) for different inter-trait correlations. The power for RMMLR approach was relatively lower than the other two approaches. All of them lost power as we increased the inter-trait correlation among individuals. On the other hand, the RMMLR approach had significantly higher power than other approaches in case a subset of traits were associated with the gene (Table 3). Moreover the power for gene detection generally went up for the RMMLR approach when the trait correlation was increased in case the gene was associated with a subset of the traits, but the other two approaches lost power as the trait correlation went higher. Especially, Fisher's test performed very poorly when the trait correlation was high and a subset of traits were associated with the gene.

We next considered a 16-SNP gene with 3 causal SNPs explaining a total of 0.9% variability for each of the associated traits and where the gene was associated with 1,2,3,4 or all 5 traits. Figure 1 shows the performance of the different approaches when the correlation among the traits was 0.2,0.4 and 0.6. The SNPs were in moderate LD. In general, the RMMLR approach had significantly better or equivalent power to the other approaches except for the situation when the gene was associated with all the traits. It also appeared to perform better when the residual trait correlation was moderate or high (0.4 or 0.6). Especially when the trait correlation was 0.6, the power was substantially higher in case the gene was associated with a subset of traits (Figure 1). On the other hand, the power for both minP and Fisher's

test decreased as the trait correlation was increased. Fisher's test gained little power for the 16-SNP gene compared to the 3-SNP gene situation (Table 3) due to the fact that multiple SNPs were associated with the traits. On the other hand, minP lost significant power for a 16-SNP gene (Figure 1) as compared to the 3-SNP gene study for not accounting for multiSNP association. As expected, Fisher's test had very good power when most or all the traits were associated with the gene, but its inflated type I error due to high correlation among traits or SNPs significantly limited its usefulness to perform gene-based association analysis.

We also checked the behavior of these approaches for a 16-SNP gene when the LD among the SNPs was high and the gene was associated with 3 out of the 5 traits. After correcting for type I error for minP and Fisher's test, all the approaches maintained similar power for both the low LD (Table 4) and high LD situations (Figure 1). We also studied the effect of increasing environmental correlation for this 16-SNP gene with 3 causal SNPs and when the LD among the SNPs was high and the gene was associated with 3 out of the 5 traits. For a genetic correlation of ρ_{ab} =0.33 and 0.67, we considered environmental correlation (r) of 0.25 and 0.5. The overall residual trait correlation was 0.3 and 0.4 for ρ_{ab} = 0.33 and r= 0.25 and 0.5 respectively. The overall residual trait correlation was 0.5 and 0.6 for ρ_{ab} = 0.67 and r= 0.25 and 0.5 respectively. Our findings were very similar to Table 4. The power substantially went up for the RMMLR approach with the increase in residual trait correlation, but both minP and Fisher's test lost power with the increase in residual trait correlation (results not shown).

In summary, the RMMLR approach performed really well when some but not all the correlated traits were associated with the gene. Moreover, the RMMLR approach maintained the correct type I error even for highly correlated traits and for markers in high LD. Hence this RMMLR approach provides a computationally efficient technique for gene-based GWA studies using unrelated individuals or families. Under complete association, Fisher's test performed really well for detection of association, but we had to correct for the inflatedness in the type I error by simulating data under the null. The minP approach performed well for a 3 SNP gene, but the power of the minP approach was substantially lower than the other approaches for the 16-SNP gene.

3.1 Real Data Analysis

A genome-wide gene-association analysis was carried out using our proposed RMMLR approach and other existing tests on data from the Minnesota Center for Twin and Family Research (MCTFR), which consists of an interrelated series of longitudinal twin, family and adoption studies. The MCTFR sample and genotyping methods have been described by Miller et al. (2012); the MCTFR phenotypes have been described by Hicks et al. (2011); and an initial SNP-based GWAS analysis has been reported by McGue et al. (2013). Briefly, the sample used here includes 7188 individuals of European ancestry clustered in 2300 nuclear families. Included are 2072 mothers (mean [SD] age at assessment of 42.8 [5.3]), 1780 fathers (44.9 [5.7]), 1788 daughters (18.0 [0.8]) and 1548 sons (17.9 [0.7]). Families are distinguished by the relationship of the two participating offspring, which were either monozygotic twins, same-sex dizygotic twins, full biological siblings, adoptive siblings, or

mixed adopted/biological siblings. The sample was genotyped using Illumina's Human 660W-Quad Array (Illumina, Inc., San Diego, CA) and a description of the quality control (QC) procedures is given in Miller et al. (2012). Of the 561,490 SNP markers on the array, 527,829 (94.0%) passed all QC filters. QC filters on samples resulted in removal of fewer than 2% of genotyped participants. European ancestry was based on self-report and was confirmed through the principal component analysis of the covariance matrix of marker data as implemented in the Eigensoft package (Price et al., 2006).

The four quantitative clinical phenotypes used in this analysis are all indicators of a higher-order construct known as externalizing or disinhibitory psychopathology (Iacono et al., 2008). These phenotypes were derived using a higher-order factor analytic approach (Hicks et al., 2011) and included: 1) Nicotine factor (NIC_FAC), a measure of symptoms of nicotine dependence, 2) Alcohol Consumption factor (ALC_FAC), a measure of quantity and frequency indicators of alcohol consumption, 3) Illicit Drug Consumption factor (DRG_FAC), a measure of use of and symptoms of abuse and dependence on illicit substances, and 4) Behavioral Disinhibition factor (BD_FAC), a measure of indicators of non-substance use behavioral disinhibition. In the MCTFR, parents are typically assessed a single time, while offspring are assessed longitudinally, spanning adolescence through early adulthood. The factor scores here for the offspring are based on the assessment completed nearest their 17th birthday so long as that assessment was completed when they were between 16.5 and 21 years. The average (range) correlation among these four phenotypes was .56 (.46 to .68). We conducted both univariate and multivariate gene-based GWAS scans for these 4 phenotypes using our RMMLR approach. Covariates in these analyses included sex, age, generation (i.e., parent versus offspring) and the first 10 principal components of the genotype covariance matrix derived by the Eigensoft package (Price et al., 2006). For comparison, we also report findings from VEGAS-SUM gene-based GWAS findings. One important issue to note is that RMMLR is much more computationally efficient than the VEGAS-SUM approach. Moreover, RMMLR can perform multivariate gene-based GWAS, which could have more power to detect multi-SNP association within a gene.

We first used our RFGLS approach (Li et al., 2011b) to perform single SNP association tests. The stepparents in the sample were treated as unrelated individuals and the rest of the sample was classified into different family types. We assumed a 12 parameter structure for the residual covariance matrix to account for the difference in variances and correlation among the pedigree members in different family types. We then assumed normal distribution of the traits to estimate the covariance matrix through maximum likelihood estimation. The VEGAS-SUM approach was then used to perform a gene-based GWAS using the single SNP association test statistics. Next we used the same gene definition as used by VEGAS to implement the RMMLR approach. We estimated the pair-wise correlations among the SNPs within a gene and removed all except one of the SNPs in perfect LD with each other for our gene-based analysis.

The VEGAS-SUM approach is similar to the Fisher's test we used in our Simulation study. The positive correlation in SNPs within a gene can cause inflatedness in the VEGAS test statistic. The bootstrapping in the VEGAS-SUM approach adjusts for such inflatedness. We

performed a maximum of 10^7 such bootstrap iterations to estimate the null distribution of the VEGAS test statistic. Note that it took us 6.02 hours of CPU time to do the 17,600 genes using 520,478 SNPs and it used up to 20 GB of RAM to conduct this gene-based GWAS.

On the other hand, RMMLR was computationally very efficient and took us significantly less time (2.35 hours of CPU time) and memory (approximately 1.95 GB of RAM) to complete the analysis. Moreover, we were able to perform both univariate and multivariate gene-based GWAS with this approach (Figure 2 and Figure 3). The multi-trait multi-SNP association analysis did produce several more significant findings as compared to individual univariate multi-SNP association analysis, supporting our hypothesis of possible association between a subset of phenotypes and genes (Figure 3). Through our multivariate analysis, we found the gene RPL7 on Chromosome 8q21.11 to be genome-wide significant (p value 6.135530×10^{-7}) even after applying Bonferroni correction for multiple testing ($.05/17600 = 2.84 \times 10^{-6}$). The corresponding univariate p values produced by the RMMLR approach were 0.007322907 for BD_FAC, 2.023403×10^{-5} for ALC_FAC, 0.001710981 for DRG_FAC and 0.004259516 for NIC_FAC. None of them were significant at the genome-wide level. Other than RPL7, we found gene MYL1 on Chromosome 2q33-q34 and RDH10 on 8q21.11 to have p-values <.00001 in the multivariate gene-based association analysis.

4 Discussion

Very few methods are available to model association between multiple traits and multiple SNPs. In the study of a complex disease, several correlated traits are often measured as risk factors for the disease. Analyzing multiple disease-related phenotypes could potentially increase power to detect association of SNPs/genes with a disease. Moreover this joint analysis could reveal some pleiotropic genes involved in the biological development of the disease. Our flexible RMMLR approach could provide gene-based association analysis with univariate as well as multivariate traits. Our approach could be applied to secondary continuous traits collected as part of a case-control study using the retrospective likelihood method developed for mapping secondary phenotypes using regression models (SPREG) in case control studies (Lin and Zeng, 2009). One could also use this approach for interaction detection by adding higher order interaction terms in the regression model if there is any prior evidence or interest in modeling interaction between certain SNPs in a gene. The RMMLR approach is computationally efficient and simulation studies showed substantial power gain when the gene is associated with a subset of traits, especially in presence of moderate or high residual trait correlation. The method remains valid even when there is high LD among the SNPs and hence the asymptotic Wilk's test can be used to assess the association. We have developed an R package to perform gene-based GWAS analyses with this RMMLR approach.

The CCA or MMLR approach has already been successfully implemented for unrelated individuals. Tang and Ferreira (2012) demonstrated the advantage of this CCA approach in multi-trait and multiSNP association analysis. Many genome-wide association studies involve family samples.

Ferreira and Purcell (2009) suggested technique for family data is computationally intensive and can only be implemented for single SNP association analysis. Here we propose a technique that can be implemented in families as well as in unrelated individuals. It is computationally efficient and captures the desirable features of the CCA algorithm. It has much higher power compared to the existing approaches, when a gene is associated with only a subset of traits.

It could get computationally intensive to compute the variance-covariance matrix in the MMLR approach (Equation 4) for every gene in the gene-based GWAS analysis. Our RMMLR approach is an approximation technique that avoids such intensive computation. The method works based on the principle that most of the genes in the GWAS study have very little or no effect on the traits. Hence estimating the variance matrix without including the gene effect does not cause much loss in power. On the other hand, if a gene is associated with the traits, our approximation technique can detect it. In our simulation studies, the RMMLR approach had more than 80% power to detect a gene explaining .6% of the phenotypic variance and in case the gene is associated with a subset of the traits with a sample size of 500 families with a familial correlation of 0.6.

Our univariate gene-based association testing with VEGAS was computationally intensive and it took us 6.02 hours of CPU time to test 17,600 genes using 520,478 SNPs and it used up to 20 GB of RAM to conduct this gene-based GWAS. On the other hand, RMMLR approach was easy to implement and we were able to use the the asymptotic p-values of the gene-based test statistic to assess the significance of the genes.

One disadvantage of this RMMLR approach over the combination methods is that it requires availability of genotype and phenotype data on individuals in order to perform the multivariate analysis. On the other hand, these combination methods require only the p-values from single SNP association analysis of univariate traits and hence might be more preferable to the user in some situations. We also need to remove SNPs in complete LD from the regression model to control for perfect multicollinearity for our RMMLR approach. Another issue is that the RMMLR approach assumes multivariate normality of the trait. Ferreira and Purcell (2009) found that the CCA approach remains valid even for non-normal trait. O'Reilly et al. (2012) found that in unrelated samples, CCA approach was sensitive to non-normality for low frequency SNPs. We expect similar behavior of the RMMLR approach in presence of non-normality, but intend to investigate it extensively through simulation studies in future.

In conclusion, our proposed RMMLR approach provides a useful framework for detection of gene-based association with single or multivariate traits, which could outperform the commonly-used permutation-based approaches, especially in case the gene is associated with a subset of traits. Our gene-based association analysis with multiple traits from the MCTFR dataset had some improved p-values as compared to the Univariate RMMLR tests, indicating evidence of association between the gene and a subset of these multiple correlated traits.

Acknowledgments

This research was supported in part by NIH grant U01-DA024417, NIH grant R21-DK089351 and NIH grant R01-DA033958. The authors would like to thank the two anonymous reviewers for their helpful comments and suggestions.

References

1. Ballard DH, Cho J, Zhao H. Comparisons of multi-marker association methods to detect association between a candidate region and disease. Genet Epidemiol. 2010; 34(3):201–212. [PubMed: 19810024]
2. Ferreira MAR, Purcell SM. A multivariate test of association. Bioinformatics. 2009; 25:132–133. [PubMed: 19019849]
3. Germain M, Saut N, Greliche N, Dina C, Lambert J, Perret C, et al. Genetics of venous thrombosis: Insights from a new genome wide association study. PLoS One. 2011; 6(9):e25581. [PubMed: 21980494]
4. Golub, GH.; Van Loan, CF. Matrix Computations. The Johns Hopkins University Press; Maryland, USA: 1996.
5. Hicks BM, Schalet BD, Malone SM, Iacono WG, McGue M. Psychometric and genetic architecture of substance use disorder and behavioral disinhibition measures for genetic association studies. Behavioral Genetics. 2011; 41(4):459–475.
6. Horn, RA.; Johnson, CR. Matrix Analysis. Cambridge University Press; Cambridge, UK: 1985.
7. Hornik, K. The R FAQ. 2012. ISBN 3-900051-08-9, URL http://CRAN.R-project.org/doc/FAQ/R-FAQ.html.
8. Iacono WG, Malone SM, McGue M. Behavioral disinhibition and the development of early-onset addiction: Common and specific influences. Annual Review of Clinical Psychology. 2008; 4:325–348.
9. Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet Epidemiol. 2008; 32:9–19. [PubMed: 17922480]
10. Lange C, Van Steen K, Andrew T, Lyon H, DeMeo DL, et al. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. Stat Appl Genet Mol Biol. 2004; 3 Article17.
11. Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, et al. A genomic pathway approach to a complex disease: axon guidance and parkinson disease. PLoS Genet. 2007; 3:e98. [PubMed: 17571925]
12. Li MX, Gui HS, Kwan JS, Sham PC. Gates: a rapid and powerful gene-based association test using extended simes procedure. Am J Hum Genet. 2011a; 88(3):283–293. [PubMed: 21397060]
13. Li X, Basu S, Miller MB, Iacono WG, McGue M. A rapid generalized least squares model for a genome-wide quantitative trait association analysis in families. Hum Hered. 2011b; 71(1):67–82. [PubMed: 21474944]
14. Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. Genet Epidemiol. 2009; 33(3):256–65. [PubMed: 19051285]
15. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, et al. A versatile gene-based test for genome-wide association studies. Am J Hum Genet. 2010; 87:139–145. [PubMed: 20598278]
16. Maity A, Sullivan PF, Tzeng JY. Multivariate phenotype association analysis by marker-set kernel machine regression. Genet Epidemiol. 2012:686–695. [PubMed: 22899176]
17. McGue, M.; Zhang, Y.; Miller, MB.; Basu, S.; Vrieze, S., et al. A genome-wide association study of behavioral disinhibition. Behavioral Genetics. 2013. (in press)
18. Miller, MB.; Basu, S.; Cunningham, JM.; Eskin, E.; Malone, SM., et al. The minnesota center for twin and family research genome-wide association study. Twin Research and Human Genetics. 2012. (in press)
19. Muller KE, Peterson BL. Practical methods for computing power in testing the multivariate general linear hypothesis. Computational Statistics and Data Analysis. 1984; 2(2):143–158.

20. Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. Am J Hum Genet. 2004; 75:353–362. [PubMed: 15272419]
21. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FCF, Elliott P, et al. Multiphen: Joint model of multiple phenotypes can increase discovery in gwas. Plos One. 2012 Doi: 10.1371/journal.pone.0034861.
22. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–9. [PubMed: 16862161]
23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC. Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–75. [PubMed: 17701901]
24. Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B. Analysis of single-locus tests to detect gene/disease associations. Geneti Epidemiology. 2005; 28:207–219.
25. Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. JASA. 1967; 62:626–633.
26. Simes RJ. An improved bonferroni procedure for multiple tests of significance. Biometrika. 1986; 73:751–754.
27. Souto JC, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, et al. Genetic susceptibility to thrombosis and its relationshipto physiological risk factors: The GAIT study. genetic analysis of idiopathic thrombophilia. Am J Hum Genet. 2000; 67(6):1452–1459. [PubMed: 11038326]
28. Tang CS, Ferreira MAR. A gene-based test of association using canonical correlation analysis. Bioinformatics. 2012; 28(6):845–850. [PubMed: 22296789]
29. Trefethen, LN.; Bau, D. Numerical linear algebra. SIAM: Society for Industrial and Applied Mathematics; Pennsylvania, USA: 1997.
30. Van Der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS Genet. 2013
31. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of gwas discovery. Am J Hum Genet. 2012; 90(1):7–24. [PubMed: 22243964]
32. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet. 2007; 81(6):1278–1283. [PubMed: 17966091]

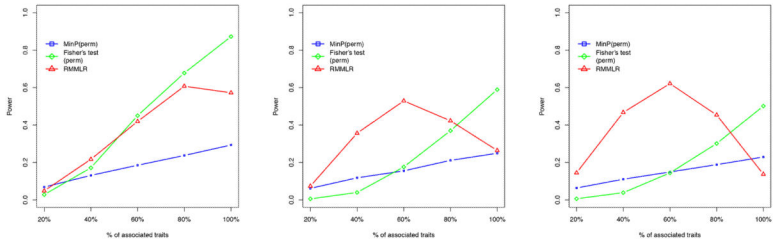

Figure 1.

Figure shows the performance of the multivariate test of association. The powers for minP, Fisher's test and RMMLR approach are shown for different number of traits being associated with a gene with 16 SNPs. The three figures (left to right) corresponds to the residual trait correlation of 0.2, 0.4 and 0.6 respectively.

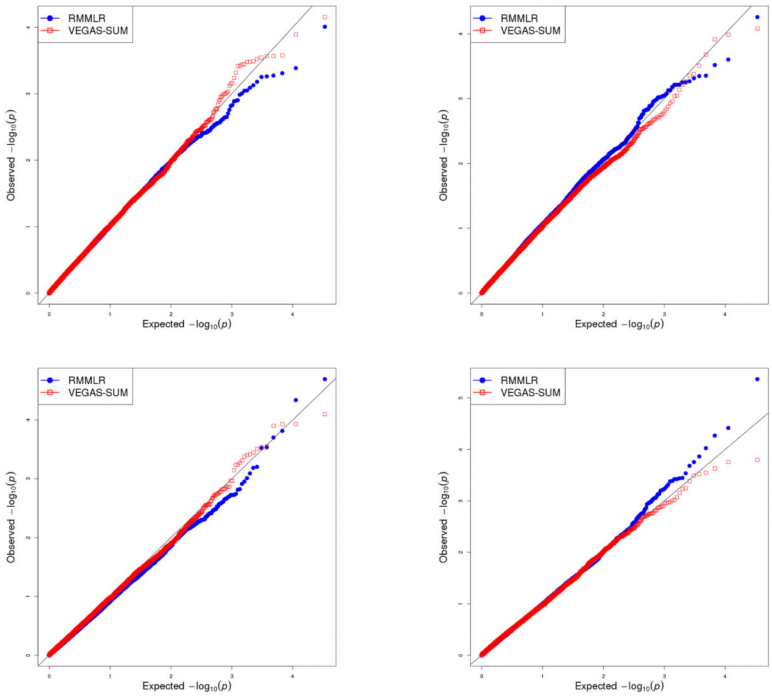

Figure 2.

Figure shows -log10(P-values) comparing the VEGAS-SUM and RMMLR approach for each of the 4 phenotypes NIC_FAC, DRG_FAC, ALC_FAC and BD_FAC respectively (top left, top right,bottom left, bottom right). There were 17600 genes. The filled squares show the qqplot of -log10(p-value)s for VEGAS-SUM approach. The filled circles show the qqplot of -log10(p-value)s from the RMMLR approach.

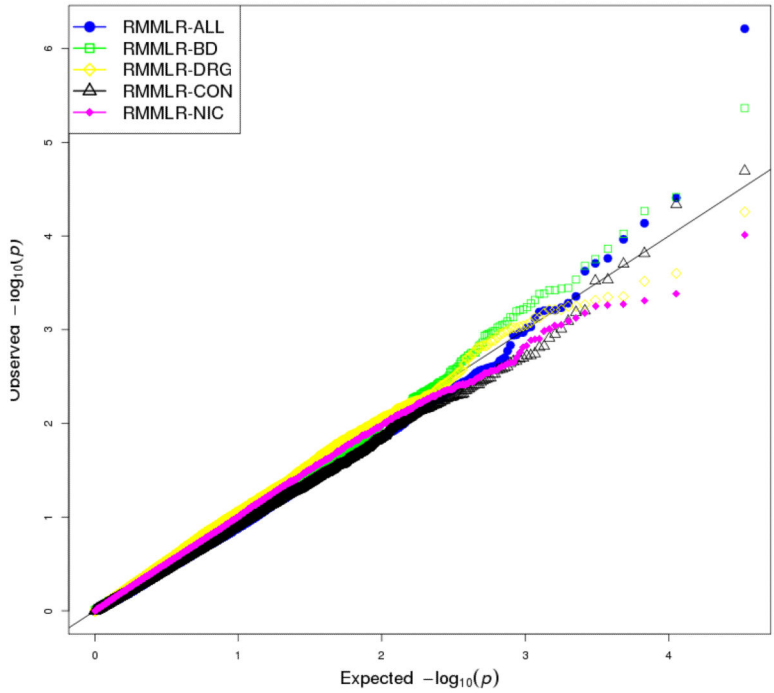

Figure 3.

Figure shows -log10(P-values) using RMMLR approach for each of the 4 phenotypes as well as for the 4 phenotypes analyzed together. The filled circles show the qqplot for -log10(p-value)s of RMMLR while analyzing the 4 phenotypes together.

Table 1

Type I error for testing association between 5 traits and a gene of length of 16 SNPs. There is moderate LD among the set of SNPs.

	$\alpha = 0.001$			$\alpha = 0.0001$			$\alpha = 0.00001$		
Trait cor	**minP**	**Fisher**	**RMMLR**	**minP**	**Fisher**	**RMMLR**	**minP**	**Fisher**	**RMMLR**
0.2	0.0012	0.0071	0.00095	$9.8e^{-5}$	0.0018	$9.75e^{-5}$	$9.4e^{-6}$	$4.5e^{-4}$	$9.9e^{-6}$
0.4	0.0011	0.026	0.00095	$9.5e^{-5}$	0.0012	$9.5e^{-5}$	$9.7e^{-6}$	$5.8e^{-3}$	$9.9e^{-6}$
0.6	0.0011	0.035	0.00096	$9.8e^{-5}$	0.0017	$9.5e^{-5}$	$9.6e^{-6}$	$8.7e^{-3}$	$9.7e^{-6}$

Table 2

Type I error for testing association between 5 traits and a gene of length of 16 SNPs. There is high LD among the set of SNPs

	$\alpha = 0.001$			$\alpha = 0.0001$			$\alpha = 0.00001$		
Trait cor	**minP**	**Fisher**	**RMMLR**	**minP**	**Fisher**	**RMMLR**	**minP**	**Fisher**	**RMMLR**
0.2	0.0007	0.026	0.0011	$8.03e^{-5}$	0.0016	$9.97e^{-5}$	0.000	$5.5e^{-3}$	$1.004e^{-5}$
0.4	0.0008	0.048	0.00095	$5.1e^{-5}$	0.026	$1.02e^{-4}$	0.000	0.0015	$1e^{-5}$
0.6	0.0008	0.076	0.001	0.000	0.048	$9.5e^{-5}$	0.000	0.031	$9.9e^{-6}$

Table 3

Power of different methods for 3-SNP gene (1 causal SNP): moderate LD

Trait cor	All 5 traits associated			3 traits associated		
	minP	Fisher's test	RMMLR	minP	Fisher's test	RMMLR
0.2	0.687	0.815	0.653	0.574	0.425	0.596
0.4	0.596	0.535	0.463	0.536	0.247	0.639
0.6	0.476	0.475	0.285	0.406	0.149	0.801

Table 4

Power of different methods for 16-SNP gene (3 causal SNPs): LD is high and 3 of these 5 traits are associated with the causal snps.

Trait cor	minP	Fisher's test	RMMLR
0.2	0.171	0.450	0.419
0.4	0.170	0.250	0.523
0.6	0.152	0.144	0.675