Bioinformatics Advance Access published March 24, 2014

Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression

Pekka Marttinen[1,2], Matti Pirinen[3], Antti-Pekka Sarin[3,4], Jussi Gillberg[1], Johannes Kettunen[3,4], Ida Surakka[3,4], Antti J. Kangas[5], Pasi Soininen[5,6], Paul F. O'Reilly[7], Marika Kaakinen[8,9], Mika Kähönen[10], Terho Lehtimäki[11], Mika Ala-Korpela[5,6,12], Olli T. Raitakari[13,14], Veikko Salomaa[15], Marjo-Riitta Järvelin[7,8,9,16,17], Samuli Ripatti[3,4,18,19,*], Samuel Kaski[1,20,*]

[1]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Finland, [2]Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, MA, [3]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland, [4]Unit of Public Health Genomics, National Institute for Health and Welfare, Helsinki, Finland, [5]Computational Medicine, Institute of Health Sciences, University of Oulu and Oulu University Hospital, Oulu, Finland, [6]NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland, [7]Department of Epidemiology and Biostatistics, MRC Health Protection, Agency (HPA) Centre for Environment and Health, School of Public Health, Imperial College, London, UK, [8]Institute of Health Sciences, University of Oulu, Finland, [9]Biocenter Oulu, University of Oulu, Finland, [10]Department of Clinical Physiology, Tampere University Hospital and University of Tampere, [11]Department of Clinical Chemistry, Fimlab Laboratories, University of Tampere School of Medicine, Tampere, Finland, [12]Computational Medicine, School of Social and Community Medicine and the Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK, [13]Department of Clinical Physiology and Nuclear Medicine, University of Turku and Turku University Hospital, Turku, Finland, [14]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku and Turku University Hospital, Turku, Finland, [15]Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland, [16]Unit of Primary Care, Oulu University Hospital, Finland, [17]Department of Children and Young People and Families, National Institute for Health and Welfare, Oulu, Finland, [18]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK, [19]Hjelt Institute, University of Helsinki, Helsinki, Finland, [20]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Finland.

Associate Editor: Dr. Jeffrey Barrett

ABSTRACT

Motivation: A typical genome-wide association study searches for associations between single nucleotide polymorphisms (SNPs) and a univariate phenotype. However, there is a growing interest to investigate associations between genomics data and multivariate phenotypes, for example in gene expression or metabolomics studies. A common approach is to perform a univariate test between each genotype-phenotype pair, and then to apply a stringent significance cutoff to account for the large number of tests performed. However, this approach has limited ability to uncover dependencies involving multiple variables. Another trend in the current genetics is the investigation of the impact of rare variants on the phenotype, where the standard methods often fail due to lack of power when the minor allele is present in only a limited number of individuals.

Results: We propose a new statistical approach based on Bayesian reduced rank regression to assess the impact of multiple SNPs on a high-dimensional phenotype. Due to the method's ability to combine information over multiple SNPs and phenotypes, it is particularly suitable for detecting associations involving rare variants. We demonstrate the potential of our method and compare it with alternatives using the Northern Finland Birth Cohort with 4,702 individuals, for whom genome-wide SNP data along with lipoprotein profiles comprising 74 traits are available. We discovered two genes (*XRCC4* and *MTHFD2L*) without previously reported associations, which replicated in a combined analysis of two additional cohorts: 2,390 individuals from the Cardiovascular Risk in Young Finns study and 3,659 individuals from the FINRISK Study.

Availability and Implementation: R-code freely available for download at http://users.ics.aalto.fi/pemartti/gene_metabolome/.

Contact: samuli.ripatti@helsinki.fi, samuel.kaski@aalto.fi

1 INTRODUCTION

Concentrations of human metabolites are associated with risk of many common diseases; for example, low- and high-density lipoprotein cholesterol (LDL, HDL) levels are associated with coronary artery disease. For this reason, human metabolism has been under intensive investigation and over the past few years several genome-wide association studies have successfully uncovered a part of its genetic basis (Sabatti *et al.*, 2008; Teslovich *et al.*, 2010; Suhre *et al.*, 2011; Kettunen *et al.*, 2012). For example, a large meta-analysis (Teslovich *et al.*, 2010) identified 95 loci influencing the levels of total cholesterol, LDL, HDL and triglycerides. More recent studies utilized finer sub-classifications of metabolites and discovered dozens of novel loci (Suhre *et al.*, 2011; Kettunen *et al.*, 2012).

*to whom correspondence should be addressed

Despite these advances, the variance explained by all reported SNPs falls far below the suggested heritability of the common metabolites, as estimated either from twin studies (Kettunen *et al.*, 2012) or from more distantly related individuals (Vattikuti *et al.*, 2012). This motivates us to develop new approaches for association testing that could better utilize all information available to us.

The present-day cohort studies often come with a rich set of phenotypic features. Examples in addition to metabolomics (Soininen *et al.*, 2009; Kettunen *et al.*, 2012; Suhre *et al.*, 2011) include studies of gene expression (Ackermann *et al.*, 2013) and 3D-facial imaging (Hammond and Suttie, 2012). As a consequence, we need statistical methods that increase the power to uncover genotype-phenotype dependencies by combining information over several related phenotypes (Ferreira and Purcell, 2009; OReilly *et al.*, 2012; Inouye *et al.*, 2012). The underlying idea is that if a genetic variant affects a trait, then it is likely to affect other traits that are related to the first one and, by testing for association with the two traits jointly, power may be increased. This reasoning can be taken a step further by testing all traits in high-dimensional omics data simultaneously, for example all metabolites in comprehensive metabolomic profiles. A comparison of different statistical methods available for joint testing of complete metabolomics profiles was recently conducted (Marttinen *et al.*, 2013).

Besides testing several phenotypes simultaneously, the ability to detect certain kinds of associations may be boosted by combining statistical evidence over several SNPs. Usually this is done in a supervised manner, such that SNPs related by location or function, for example, are tested simultaneously. Combining information over multiple SNPs is particularly crucial with rare variants, i.e., SNPs where the minor allele is present in a small proportion of the population. Testing such SNPs individually is unlikely to yield significant findings due to limited power. Most approaches for handling rare variants are based on collapsing several rare SNPs into a single variable (Bansal *et al.*, 2010). For example, one can simply collapse several rare variants into a single indicator telling whether any of the rare variants is present in the individual (Morgenthaler and Thilly, 2007) or to count the number of rare variants present in the individual (Morris and Zeggini, 2010). The problem with the collapsing methods is the implicit assumption that the effects are in the same (or a pre-defined) direction. A more sophisticated variance component method avoiding this assumption is able to investigate the impact of several rare variants on a univariate trait (Wu *et al.*, 2011).

Even if there exists a large number of methods for analyzing rare variants, none of those has been tailored for multivariate phenotypes. On the other hand, standard methods for multivariate phenotypes (Ferreira and Purcell, 2009) have not been thoroughly investigated in the context of rare variants, and we will see in the following that severe overfitting may occur. In summary, methods for dealing with rare variants in the context of multivariate phenotypes are clearly lacking.

In this paper, we derive a novel formulation of the Bayesian reduced rank regression model (Geweke, 1996) to detect multivariate associations between pre-defined groups of SNPs and a high-dimensional phenotype. In particular, our approach is suitable for analyzing both common and rare variants. Our formulation incorporates prior knowledge about effect sizes to increase the power to detect associations. Furthermore, it is capable of correcting for the number of SNPs considered, which is important when testing a large number of SNP groups of different sizes. We validate our method by assessing associations between SNPs in all human genes, one gene at a time, and metabolic profiles comprising fine-scale lipoprotein measurements for 4,702 individuals from the Northern Finland Birth Cohort 1966 (Rantakallio, 1969; Sabatti *et al.*, 2008). Among the top-scoring genes without known associations to the traits studied, two genes (*XRCC4* and *MTHFD2L*) replicated in a combined analysis of 2,390 individuals from the Cardiovascular Risk in Young Finns study (Raitakari *et al.*, 2008) and 3,659 individuals from the FINRISK Study (Vartiainen *et al.*, 2010). Additional analyses of the same data confirmed that alternative methods discovered only one of these associations and did not identify any further associations that were not known before.

2 METHODS

2.1 Model

To build our model, we assume that the phenotypes may be affected by three kinds of variables, (i) known factors, such as age, sex, or population structure, (ii) unknown factors, such as experimental conditions and other batch effects, and (iii) SNPs under consideration, as schematically presented in Figure 1. In contrast to standard regression, where each SNP-phenotype pair has a parameter representing the effect of the SNP on the phenotype, here we assume that a combination of several SNPs is influencing several phenotypes through some unknown factors. This assumption is compactly expressed in terms of the reduced rank regression, where the SNPs are first projected onto a low-dimensional sub-space, and the projections are then used as regressors when predicting the phenotypes. The reduced-rank regression formulation immediately implies some structural assumptions deemed sensible in the current setting: first, if a SNP has an effect on a phenotype, then the SNP is likely to have an effect on other phenotypes; second, if a phenotype is affected by a SNP, then the phenotype is more likely to be affected by other related SNPs as well.

Let N denote the number of individuals, S the number of SNPs, P the number of phenotypes, and C the number of other covariates. Formally, we consider the Bayesian reduced rank regression model

$$Y = X\Psi\Gamma + ZA + H\Lambda^T + E, \tag{1}$$

where $Y_{N\times P}$ contains the phenotypes, $X_{N\times S}$ contains the SNPs, $\Psi_{S\times K_1}$ and $\Gamma_{K_1\times P}$ represent a low-rank approximation for the regression coefficient matrix $\Theta = \Psi\Gamma$, $Z_{N\times C}$ represents other covariates with the corresponding coefficient matrix $A_{C\times P}$, $H_{N\times K_2}$ contains hidden confounding factors with the corresponding coefficient matrix $\Lambda_{P\times K_2}$, and $E_{N\times P} = [e_1,\ldots,e_N]^T$, with $e_i \sim N(0,\Sigma)$, where $\Sigma = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_P^2)$. Note that by integrating over the hidden factors H, the model is equivalent to

$$y_i \sim N(\Theta^T x_i + A^T z_i, \Lambda\Lambda^T + \Sigma), \quad i = 1,\ldots,N, \tag{2}$$

where it is assumed that the factors have independent standard normal prior distributions. Therefore, we see that having the latent variable part $H\Lambda^T$ in the model corresponds to assuming a low-rank approximation to the full covariance matrix, important when analyzing high-dimensional data sets.

For computational reasons, we restrict the rank of the regression model, K_1, to unity in our genome-wide analysis (w.l.o.g. in the detection task, see below), and show results with $K_1 = 1, 2, 3$ for a few representative examples. However, here we present a general infinite-dimensional framework available in our implementation, which does not necessitate the selection of a fixed rank. In general, in order to use the Bayesian reduced rank regression model, the rank of the model, K_1, and the rank of the low-rank approximation for the covariance matrix, K_2, must be selected. A recent Bayesian infinite sparse factor analysis model (Bhattacharya and Dunson, 2011) circumvents the selection of a fixed rank for K_2 by assuming in principle an infinite number of columns in the Λ matrix; however, the columns shrink progressively such that only the first ranks are influential in practice. We

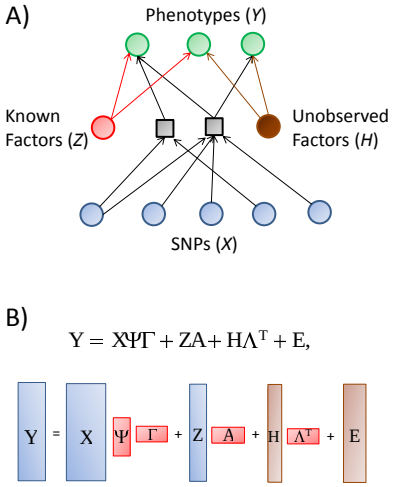

Fig. 1. Graphical illustration of the model. A) The variables and dependencies between them. The phenotypes Y are assumed to be affected by known factors, such as age or sex, unknown factors, such as batch effects caused by varying experimental conditions, and the SNPs. The influence of the SNPs is mediated by unknown combinations of the original SNPs, represented by black squares. B) The same model using matrix notation. Matrices containing the observed variables, Y (the phenotypes), X (the SNPs), and Z (known factors), are blue. The regression coefficient matrices are red. Note that the coefficient matrix for the SNP effects is written as a product of two matrices, Ψ and Γ, corresponding to a low-rank approximation to an unconstrained coefficient matrix. The brown matrices comprise unobserved variables, H (unknown factors) and E (noise terms).

assume this prior formulation for our noise model $H\Lambda^T + E$. We exploit the idea further by allowing also the rank K_1 to be infinite in principle, and enforcing the low-rank nature by shrinking the columns of Ψ and the rows of Γ increasingly as the column/row index grows. In practice, one needs to specify upper bounds for K_1 and K_2. We select the upper bound for K_2 using the adaptive procedure of Bhattacharya and Dunson (2011), and we have implemented an analogous method for learning the upper bound for K_1. In practice we wanted to minimize the model complexity to maximize the power to detect associations and to speed up the computations. Therefore, we decided to run the genome-wide analysis (see Section 3.1) using a fixed rank $K_1 = 1$. We note that the model with $K_1 = 1$ is sufficient for our purposes of detecting whether a gene is unrelated to the phenotypes (in which case rank zero would already be sufficient), although for prediction a higher rank might be more suitable. We experimented with a selected set of known genes with upper bounds 2 and 3 (see below) and noticed that the effect of increasing the upper bound from unity had only a small impact on the amount of variation that is explained by the model. From the biological perspective this means that the influence of a gene on the phenotypes can mostly be described in terms of a single latent factor mediating the effect. A detailed model description is given in Supplementary Section 1.

The model is related to many published methods, and a thorough comparison can be found in Supplementary Section 2. Correction for unknown factors has recently been considered with the standard regression model, typically for just one SNP at a time (Stegle *et al.*, 2010; Fusi *et al.*, 2012). The difference to the original Bayesian reduced rank regression formulation (Geweke, 1996) is that our model utilizes low-rank approximation to the covariance matrix, making it more suitable for high-dimensional phenotypes, and informative prior distributions accommodating problem-specific knowledge. Furthermore, we use the model in a new way, as described in the subsequent sections. The Bayesian infinite sparse factor analysis model has been used for high-dimensional data (Bhattacharya and Dunson, 2011); here, we use it to represent the multivariate noise. Canonical correlation analysis (CCA) is a classical tool for modeling multivariate dependencies that has recently been introduced in the association study context (Hotelling, 1936; Ferreira and Purcell, 2009). Sparse canonical correlation analysis (Waaijenborg *et al.*, 2008; Witten *et al.*, 2009; Parkhomenko *et al.*, 2009) is more suitable to very high-dimensional data sets; however, introducing prior distributions for CCA that would be intuitive in the association study context does not seem straightforward.

2.2 Proportion of total variation explained (PTVE)

A commonly used measure of the impact of multiple SNPs, say $x_1, \ldots, x_S$, on a univariate trait y is the proportion of variance explained (PVE) by the SNPs:

$$\text{PVE} = 1 - \frac{\text{Var}(y - \sum_{i=1}^{S} \hat{a}_i x_i)}{\text{Var}(y)} = \frac{\text{Var}(\sum_{i=1}^{S} \hat{a}_i x_i)}{\text{Var}(y)}.$$

Here, $\sum_{i=1}^{S} \hat{a}_i x_i$ is a linear prediction for the phenotype y, given the SNPs. Analogously, as a measure of the overall impact of multiple SNPs on a high-dimensional phenotype, we propose to use the proportion of total variation of the phenotypes explained (PTVE) by the model, namely

$$\text{PTVE} = \frac{\text{Tr}(\text{Cov}(\hat{Y}))}{\text{Tr}(\text{Cov}(Y))}, \quad (3)$$

where $\hat{Y}$ is a prediction for a high-dimensional phenotype Y from the model and Tr denotes the trace, i.e., the sum of the diagonal elements of the matrix. In (3) and in general, the total variation of a multivariate random variable is defined as the trace of the covariance matrix, i.e., the sum of the variances of the individual variables. Therefore, PTVE measures the joint impact of the SNPs on several phenotypes, and hence is expected to yield high scores to such dependencies in which many phenotypes are affected by the SNPs, even if none of the effects is very large by itself.

In the Bayesian statistical framework, the inferences are based on posterior probability distributions of the quantities of interest (Gelman *et al.*, 2004). With the Bayesian reduced rank regression model, samples from the posterior distribution of the PTVE can be obtained from

$$\text{PTVE}^{(i)} = \frac{\text{Tr}\left(\text{Cov}\left(X\Psi^{(i)}\Gamma^{(i)}\right)\right)}{\text{Tr}\left(\text{Cov}\left(Y\right)\right)}, \quad (4)$$

where $\Psi^{(i)}$ and $\Gamma^{(i)}$ are samples from the posterior distribution of the parameters Ψ and Γ. It is similarly straightforward to estimate the posterior distribution for the proportion of variation explained by the rare variants, by dividing the variation of the prediction into two components, one corresponding to the rare variants, the other to the common variants. The score obtained in this way is referred to as PTVE-rare in the sequel and its exact definition is given in Supplementary Section 3. In practice, we approximate the posterior distribution by using a mode-based point estimate for one of the parameters, Γ, and estimating the joint distribution of the other parameters using an MCMC-algorithm, as described thoroughly in Supplementary Sections 4 and 5.

2.3 Informative prior

In the Bayesian analysis, external background knowledge may be incorporated in the statistical analysis through prior probability distributions. As our analysis is focused on estimating the posterior distribution of PTVE, a sensible prior is obtained by making the prior distribution of the PTVE represent our true beliefs about this quantity. The prior distribution of the PTVE appears not to be available in a closed form; however, in Supplementary Section 6, we derive results which show how the distribution of the mean of the PTVE, μ_{PTVE}, depends on model hyperparameters. Using these results we set the prior distributions to satisfy the following properties

$$\text{Median}(\mu_{\text{PTVE}}) = 10^{-6} \quad (5)$$

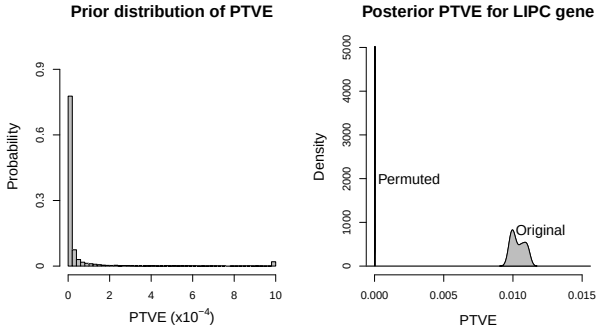

Fig. 2. Prior and posterior distributions for the proportion of total variation explained (PTVE) by the model. The panel on the left shows the prior distribution imposed on the proportion of total variation of the phenotypes explained by the SNPs under consideration (here, the SNPs from the *LIPC* gene). The median of the prior distribution is located at approximately 4e-6. The characteristic features of the prior distribution include the peak at values very close to zero, effectively removing noise unless there is strong evidence about a possible association, and the long tail allowing a small percentage of genes to explain larger proportions of the phenotype variation. The rightmost bin on the x-axis contains the total probability of values exceeding the maximum value on the axis. The panel on the right shows the posterior distribution of the PTVE for the same SNPs. Two posterior densities are shown, one showing the distribution for the original data, the other showing the distribution for data in which the rows of the phenotype matrix have been permuted. Notice the differing scales on the x-axes of the two panels.

and

$$P(\mu_{\text{PTVE}} > 0.001) = 0.01. \quad (6)$$

Equation (5) means that the prior median of μ_{PTVE} is very close to zero, as we expect most of the genes to be unrelated to the phenotypes. Equation (6) says that with a small probability, here equal to 0.01, the gene may explain more than 0.1 per cent of the total variation, a value deemed resonable based on the background knowledge. The resulting prior distribution for the PTVE is obtained by integrating over the distribtion of μ_{PTVE}, and we used Monte Carlo simulation to investigate the distribution. Figure 2 shows PTVE values sampled from the prior distribution. The following desirable characteristics can be seen: first, a peak close to zero, shrinking the coefficients when no effect is present; second, a long tail, corresponding to the genes with a non-neglible impact on the phenotypes, without imposing strong beliefs about the actual size of these non-zero effects.

Another important property that follows from using the informative prior is that the number of SNPs under consideration can be accounted for. Detailed examination of Corollary 1 in Supplementary Section 6 reveals that with fixed hyperparameters the expected PTVE is proportional to $\sum_{i=1}^{S} \text{Var}(x_i)$, i.e., the total variation of the SNPs. Roughly, this means that if the number of SNPs doubles, the expected PTVE doubles as well, if the hyperparameters are kept fixed. Using the Corollaries 1 and 2 in Supplementary Section 6, it is straightforward to modify the hyperparameter distributions to assert the prior conditions (5) and (6), implying in our setting that all genes are expected to explain the same amount of the variation of the phenotypes, regardless of how many SNPs they contain.

2.4 Data

As a data set for detecting associations we consider a sample of 4,702 individuals from the Northern Finland Birth Cohort 1966 (NFBC1966), a birth cohort study of children born in 1966 in the two northernmost provinces of Finland (Rantakallio, 1969). The blood samples for the DNA extraction and phenotype data were collected at a follow-up visit when the participants were 31 years of age. For replication we consider two cohorts. The Cardiovascular Risk in Young Finns Study (YFS) is a population-based prospective cohort study (Raitakari *et al.*, 2008) conducted in Finland, the purpose of which was to investigate the levels of cardiovascular risk factors in children and adolescents in different parts of the country. The FINRISK Study comprises cross-sectional population surveys that have been carried out every 5 years since 1972, to assess the risk factors of chronic diseases, with emphasis on cardiovascular risk factors (Vartiainen *et al.*, 2010). The individuals analyzed in this study belong to the sample from the year 1997. The blood samples for the two replication cohorts were collected when the participant were 30-45 and 25-71 years of age, respectively. The study protocols of all data sets have been approved by the local ethics committees.

The samples were genotyped with Illumina arrays (Illumina, Inc. San Diego, CA, USA) and imputed with IMPUTE 2 (Howie *et al.*, 2009, 2011) using a 1000 Genomes Project reference panel (The 1000 Genomes Project Consortium, 2012). Of the resulting good-quality autosomal SNPs (info > 0.4), we extracted SNPs in 24,025 human genes by adding 50kb flanking regions on both sides of the endpoints of the genes given in NCBI Gene database (genome assembly GRCh37.p10, NCBI annotation 104, Nov 2012). As a pre-processing step, we reduced the genotype space within each gene to the most promising 200 SNPs (at most) that had the highest canonical correlation test score (Ferreira and Purcell, 2009) with the metabolites; however, to prevent overfitting, the priors were specified as described above using the unpruned SNP set. In preliminary experiments, decreasing the number of SNPs in this way from 800 to 200 had no visible effect on the results. Finally, the SNPs were scaled to have unit variance.

Phenotype data came from the serum NMR metabolomics platform described earlier (Soininen *et al.*, 2009). As a pre-processing step the traits were quantile-normalized to have standard normal distribution. Individuals with 20 percent missing values were removed and the remaining missing values were imputed by sampling them from the multivariate normal distribution. In this work, we analyzed a subset of 74 lipoprotein subclass measures (see Supplementary Table 3). The empirical correlation matrix of the traits is shown in Supplementary Figure 1. Linear regression was used to correct the phenotypes for age, sex, and population structure using 10 principal components (Price *et al.*, 2006).

3 RESULTS

3.1 Genome-wide analysis of NFBC1966 data

We computed the PTVE and PTVE-rare scores for each of the 24,025 human genes in the NFBC1966 data, one gene at a time, using the Bayesian reduced rank regression. Table 1a shows the top-5 genes with the highest PTVE scores, all of which are well-known lipid-associated loci. More generally, all 43 top-scoring genes were located within 1Mb from previously reported genome-wide significant lipid associations (Teslovich *et al.*, 2010; Kettunen *et al.*, 2012; The Global Lipids Genetics Consortium, 2013), and detailed listing of the top genes is given in Supplementary Table 4. Furthermore, of the top-100 genes, which correspond approximately to FDR=0.2 (see the next section), 73 genes had known associations. Together, these findings serve as a validation of the method and its implementation.

To illustrate the effect of the model rank on the results, we carried out a detailed analysis for three known lipid genes: *LIPC*, *APOB* and *PLTP*, with rank $K_1 = 1, 2, 3$. The genes were selected such that they were located in different chromosomes and had dissimilar association profiles (*APOB* associated most strongly to VLDL, IDL and LDL, *PLTP* to HDL, and *LIPC* to VLDL, IDL and HDL (Tukiainen *et al.*, 2012)). In addition to considering the genes separately, we repeated a joint analysis for all three possible pairwise gene combinations. The results are shown in Supplementary Figure

Table 1. Summary of results from the genome-wide analysis of the real data. a) Reference results for genes with five highest PTVE scores. b) Replicated genes from the PTVE-rare score (out of 6 genes tested for replication). c) Replicated genes from the PTVE score (out of 167 genes). Five other replicated genes (*PPBP, CXCL5, CXCL2, PF4, CXCL3*) are not shown as they were located within 1Mb from *MTHFD2L*, which had the strongest effect. Column **PTVE(sd)** shows the proportion of total variation explained and its standard deviation, **rare** specifies the proportion of the variation explained by the gene attributed to the rare variants, **p-value** specifies the p-value pooled over YFS and FINRISK replication data sets (unless stated otherwise), and the last four columns specify the ranking of the gene among all genes with different methods. * denotes genes which replicated significantly in only one of the two replication data sets. $^{a/b}$ replication p-value in FINRISK/YFS.

a)					Gene rank			
Chr	Locus	PTVE(sd)	rare	p-value	PTVE	pairwise	S-CCA	CCA-single
15	*LIPC*	0.01(5e-04)	0.015	5e-19	1	1	1	132
19	*APOC1*	0.0046(3e-04)	0.017	1.4e-26	2	8	18	275
19	*PVRL2*	0.0045(1e-04)	0.0015	7e-35	3	9	14	276
2	*APOB*	0.0044(3e-04)	0.017	1.1e-17	4	45	41	3244
11	*APOA5*	0.0043(2e-04)	0.021	5e-10	5	26	33	2433

b)					Gene rank			
Chr	Locus	PTVE(sd)	rare	p-value	PTVE-rare	pairwise	S-CCA	CCA-single
16	*SPIRE2**	0.0015(9e-05)	0.89	0.00091^{a}	5	8344	5490	4973
5	*XRCC4*	0.0024(2e-04)	0.55	0.0016	6	2706	5163	2155

c)					Gene rank			
Chr	Locus	PTVE(sd)	rare	p-value	PTVE	pairwise	S-CCA	CCA-single
2	*DTNB**	0.0015(2e-04)	0.15	2.6e-04^{b}	138	4715	338	7652
4	*MTHFD2L*	0.0015(2e-04)	0.019	7e-06	163	102	1444	1552

2, and are summarized as follows: 1) the rank of the model had a minor effect on the results for a single gene, 2) increasing the rank from $K_1 = 1$ to $K_1 = 2$ increased the variance explained in all joint pairwise analyses, and the increase was significant in two of the three cases, 3) a combination of two genes always had a larger effect than either gene individually; however, the sum of the individual effects was slightly larger than the effect of the combination. We attribute this difference mainly to the stronger shrinkage in the second than the first genotype component. 4) The first component of a joint model was always strongly correlated with the single-gene model with the larger effect, the second component with the single-gene model with the smaller effect. This is as expected as the prior distribution was designed to identify the strongest associations using the first genotype component.

In order to check whether novel associations could be detected by any method, we carried out a replication experiment with the YFS and FINRISK data sets for the most promising genes, after excluding genes located within 1Mb from previously reported associations. As promising we considered from each method all genes with FDR<0.4. For the PTVE-rare score, 6 genes were tested, none of which had previously reported associations to lipids. For the PTVE score 305 genes had FDR<0.4; however, only 167 were not located close to known associations, and these 167 genes were selected for replication.

For the purposes of replication, the multivariate dependency between the multiple SNPs and phenotypes was reduced into a univariate test by using parameters estimated in the NFBC1966 data to construct univariate genotype and phenotype combinations (see Supplementary Section 7 for further details). Then, the standard linear model was used to test for positive correlation between the genotype and phenotype combinations. A pooled linear regression coefficient combining YFS and FINRISK data sets was formed using a fixed effect model over the two data sets (Thompson *et al.*, 2011) and the p-value was obtained by relating the pooled estimate to its standard deviation. A one-tailed test was used as we were only interested in findings in which the effects were in the same direction in the replication data sets as in the NFBC1966 data. Bonferroni correction was utilized to account for the number of genes tested with each method, such that corrected p-value threshold corresponding to the nominal 0.05 level of significance was equal to $p < 0.0083$ for the 6 putative novel genes from PTVE-rare score and $p < 0.00030$ for the 167 putative genes from PTVE score.

We considered a replication significant if the test was nominally significant (p<0.05) in both replication data sets, and the p-value for the pooled estimate was significant after correcting for the multiple tests. Information about genes which replicated significantly is provided in Tables 1b and 1c and in Figure 3. In total 1 gene with PTVE-rare and 6 genes with PTVE were detected, corresponding to two independent genes: *XRCC4*, and *MTHFD2L*. We note that *MTHFD2L* is located within 1Mb from two SNPs (rs2168889 and rs16850360) associated with 'metabolic networks' containing some lipoprotein traits from our data (Inouye *et al.*, 2012). However, the top-metabolite in the associations was Albumin, and when we repeated the multivariate test with the lipoprotein traits considered here, these associations were no longer genome-wide significant (p=2.5e-4 and p=6.4e-7). In addition, two genes, *SPIRE2* and *DTNB*, replicated significantly (after the multiple testing correction) in one, but not in the other replication data set. Table 1 also presents the rankings of these genes by alternative methods (see below). We see that *XRCC4*, *SPIRE2*, and *DTNB* were completely missed by the other methods. Supplementary Table 1 shows the SNPs contributing to the reported associations. We see that with *XRCC4*, *SPIRE2* and

DTNB, many SNPs, some of which are rare, are required to represent the overall association. This explains why these genes did not receive any signal from the standard testing with the pairwise linear model. Supplementary Figure 3 shows graphically how the phenotypes are affected by the significant genes and the known *LIPC* gene. We see that in neither of the new genes is the effect focused on any single trait, but rather a small effect is seen on many lipoprotein measures. This is not surprising as the PTVE score is expected to give high scores to precisely this kind of associations. Supplementary Figure 4 shows the estimated SNP coefficients for the genes, and demonstrates the usefulness of analyzing all SNPs in a gene simultaneously in order to reduce noise resulting from the correlation between the SNPs. Further background information on these genes is presented in Supplementary Table 2; however, a more thorough biological interpretation of the genes remains for future work.

Finally, we repeated a similar analysis with three alternative methods: 1) exhaustive pairwise search with a linear model, where the minus logarithm of the smallest pairwise p-value over all SNPs in the gene and metabolites was taken as the test score for the gene, 2) canonical correlation analysis, applied to a single SNP vs. all metabolites at a time as by Ferreira and Purcell (2009) and Inouye *et al.* (2012), and 3) sparse canonical correlation analysis, which was used to compute the canonical correlation between all SNPs in a gene and all phenotypes (Parkhomenko *et al.*, 2009). The methods 2) and 3) were found to be the most powerful in a recent comparison of approaches for a multivariate metabolomics phenotype (Marttinen *et al.*, 2013); however, here the SNPs were not pruned using the minor allele frequency before applying the methods as by Marttinen *et al.* (2013), because here the focus is in the rare variant setting. Similarly to PTVE and PTVE-rare methods, we tried to replicate the top-scoring genes up to FDR=0.4. The last column in Table 2 shows the numbers of genes included in the replication with detailed listings given in Supplementary Tables 4-8. The association in *MTHFD2L* was confirmed to be genome-wide significant using the standard pairwise linear regression (SNP rs185567543, trait S.LDL.P, p=2.5e-15, where the p-value is pooled over all three data sets). All other replicated associations were located within 1Mb of this or previously known associations, thus yielding no additional novel detections.

3.2 Power comparison

To investigate the power of the introduced method and compare it with alternative methods, we estimated false discovery rates (FDR) corresponding to different threholds d for declaring a gene as detected. FDR estimates were obtained by permuting the rows of the phenotype matrix and analyzing the permuted data in exactly the same way as the original data (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003; Xie *et al.*, 2005). In detail, we computed the average number of genes in the permuted data sets with scores exceeding a given threshold d (false positives, $\mathrm{FP}(d)$), the number of genes in the original data with scores exceeding the same threshold d (total positives, $\mathrm{TP}(d)$), and considered the ratio of the two

$$\mathrm{FDR}(d) = \frac{\mathrm{FP}(d)}{\mathrm{TP}(d)}.$$

We note that because of computational burden, only one permutation was used with the Bayesian reduced rank regression. With the other methods, three permutations were used. Therefore, the FDR estimates are approximate, which we consider to be sufficient for our purposes, especially as the most extreme quantiles are not considered.

The numbers of genes declared detected with different FDR thresholds are presented in Table 2, and the overall concordance of the scores from different methods is shown in Supplementary Figure 5. In summary, the results show that the method with which most associations were found was exhaustive pairwise search and the second-most powerful method was the Bayesian reduced rank regression with PTVE as the test score. These two methods also had the highest agreement in scoring genes. CCA applied to test for association between individual SNPs and the multivariate phenotype performed badly. These results are somewhat different from what has been reported before by us and others (Marttinen *et al.*, 2013; Inouye *et al.*, 2012). In particular, the CCA applied to individual SNPs performed much worse that the pairwise testing, although previously it has been reported to have clearly higher power. The explanation is that here we applied the methods to all SNPs, including the rare variants. Specifically, the single-SNP-CCA starts to overfit when applied to rare variants. For example, when we investigated the results more carefully, we discovered SNPs in which the minor allele was present in very few individuals, and such individuals could be almost perfectly identified by a seemingly random combination of phenotypes, leading to a very large spurious canonical correlation test score (or, equivalently, a highly significant p-value). Bayesian reduced rank regression and sparse CCA were less affected by the overfitting because they exploit ways to control the complexity of the model, the former using the informative priors, the latter cross-validation.

Combining information over several SNPs using CCA or related methods has recently been demonstrated to improve power to detect associations under certain conditions (Tang and Ferreira, 2012; Zhang *et al.*, 2011; Marttinen *et al.*, 2013). Here we see that the sparse CCA, and also the Bayesian reduced rank regression, which utilize multi-SNP information, have lower power in the genome-wide analysis than the simple pairwise testing. The difference from the earlier experiments is that here the methods are applied to genotype data with a much higher SNP density, as obtained through careful imputation. The conclusion is that if the multi-SNP information has already been utilized within the imputation protocol, the power to detect associations cannot in general be expected to improve by utilizing multi-SNP models.

For a more detailed comparison between the Bayesian reduced rank regression and the exhaustive pairwise testing, Supplementary Figure 6 shows a Q-Q plot of the scores (both PTVE and PTVE-rare) against the expected scores that have been obtained by permutation. The Supplementary Figure 6 also shows genes detectable by the simple exhaustive pairwise search. Both Q-Q plots, but PTVE in particular, indicate an excess of large test-scores, reflecting the fact that at least some true associations are detected by the model. We further see that although with PTVE-rare score fewer genes can be detected than with PTVE, none of the top-scoring genes from PTVE-rare are flagged by the pairwise approach, making PTVE-rare an attractive score for mining associations that might be missed by the standard method.

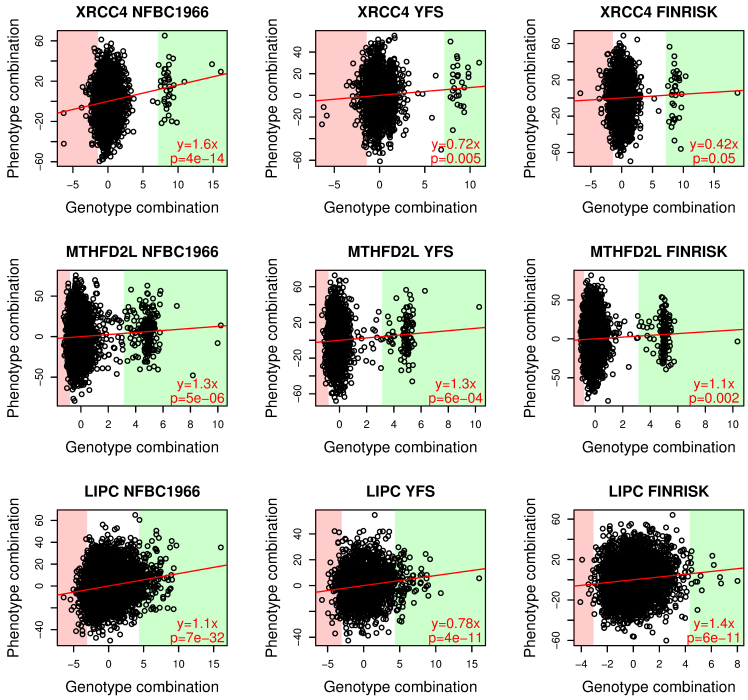

Fig. 3. Results for genes with significant replication in both test sets: *XRCC4*, and *MTHFD2L*; for reference, the well-known *LIPC* lipid locus is also shown. Each panel shows the identified phenotype combination plotted against the genotype combination. The left column shows results in the NFBC1966 data set, in which the associations were detected. The center and right columns show results with the YFS and FINRISK data sets, where coefficient matrices learned with the NFBC1966 data were used to form the variable combinations. The green and red background colorings mark the individuals with the highest/lowest genotype combination values, and the phenotype values in these extreme groups are investigated in more detail in Supplementary Figure 3.

Table 2. Power comparison of the different methods. The table shows the numbers of gene-metabolome associations that had false discovery rate below the specified threshold. The last column shows the number of putative novel associations within genes with FDR=0.4 after removing the known associations as described in the main text.

	FDR=0	FDR=0.1	FDR=0.2	FDR=0.4	Novel
PTVE	36	55	103	305	167
PTVE-rare	3	3	3	6	6
Pairwise	103	176	243	651	300
CCA, single SNP	7	7	7	11	11
Sparse CCA	37	50	66	117	51

4 DISCUSSION

We have presented a new statistical method for investigating associations in GWAS data sets with multivariate phenotypes. The method can combine information over multiple SNPs, making it particularly suitable for studying rare variants in the high-dimensional phenotype setting. For this setup no methods known to the authors have been presented before. Our method is based on estimating the proportion of total variance of the phenotypes that is explained by the SNPs under consideration. For this purpose, we have derived a Bayesian formulation of the reduced rank regression model, which enables us to incorporate our knowledge of the expected effects sizes in the analysis.

We used the new method to analyse a real GWAS data set with a multivariate lipoprotein phenotype. Two novel loci not previously associated with the phenotype were discovered and replicated in an analysis combining two additional data sets. Furthermore, two more loci were found which replicated significantly in one, but not in the other test data set. Possible reasons for the lack of success in replicating the findings in both the data sets include: (i) the associations were false positive in the first place, (ii) the associations involved rare variants which were not present in sufficient numbers to see the effects, (iii) the imputation accuracy of the rare variants was not sufficient in all data sets, (iv) the phenotype data, although pre-processed in exactly the same way with all the data sets, have not been fully equivalent. For example, the scaling of the phenotypes during pre-processing has been done using factors not exactly equal, and the parameters learned in one data may thus not represent the effects adequately in another data. Only further studies will help to distinguish between the alternative explanations.

For doing inference with the model, the current implementation uses MCMC sampling, the computation time of which is approximately half an hour per gene on a 2.3GHz processor. Thus, analyzing all human genes requires a cluster computer to parallelize the computations over the genes. Analytical approximations, such as the variational or Laplace approximations, see, e.g., Bishop *et al.* (2006), could be used to speed up the computations and, based on our experiments with the current method, are worth doing in the

future. Alternative ways to use the model might also be considered. For example, focusing the analysis on variants with a predicted function could improve power to detect associations and lessen the computational burden. As another example, we have used 0.01 as the threshold for defining the rare variants when computing their impact on the phenotypes. Results based on different thresholds could readily be extracted from the output of a single MCMC run, and are likely to highlight different sets of genes.

ACKNOWLEDGEMENT

Funding: A complete list of funding is given in the Supplementary material.

REFERENCES

Ackermann, M., Sikora-Wohlfeld, W., and Beyer, A. (2013). Impact of natural genetic variation on gene expression dynamics. *PLoS Genetics*, **9**(6), e1003514.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, **11**(11), 773–785.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Bhattacharya, A. and Dunson, D. (2011). Sparse Bayesian infinite factor models. *Biometrika*, **98**(2), 291–306.

Bishop, C. M. *et al.* (2006). *Pattern recognition and machine learning*. Springer, New York.

Ferreira, M. A. and Purcell, S. M. (2009). A multivariate test of association. *Bioinformatics*, **25**(1), 132–133.

Fusi, N., Stegle, O., and Lawrence, N. (2012). Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Computational Biology*, **8**(1), e1002330.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Florida, 2nd edition.

Geweke, J. (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics*, **75**(1), 121–146.

Hammond, P. and Suttie, M. (2012). Large-scale objective phenotyping of 3d facial morphology. *Human mutation*, **33**(5), 817–825.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**(3/4), 321–377.

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, **1**(6), 457–470.

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, **5**(6), e1000529.

Inouye, M., Ripatti, S., Kettunen, J., Lyytikäinen, L.-P., Oksala, N., Laurila, P.-P., Kangas, A. J., Soininen, P., Savolainen, M. J., Viikari, J., *et al.* (2012). Novel loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS Genetics*, **8**(8), e1002907.

Kettunen, J., Tukiainen, T., Sarin, A.-P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.-P., Kangas, A. J., Soininen, P., Würtz, P., Silander, K., *et al.* (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature Genetics*, **44**(3), 269–276.

Marttinen, P., Gillberg, J., Havulinna, A., Corander, J., and Kaski, S. (2013). Genome-wide association studies with high-dimensional phenotypes. *Statistical Applications in Genetics and Molecular Biology*, **12**(4), 413–431.

Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **615**(1), 28–56.

Morris, A. P. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic epidemiology*, **34**(2), 188–193.

OReilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M.-R., and Coin, L. J. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PloS ONE*, **7**(5), e34861.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, **8**(1), 1–34.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**(8), 904–909.

Raitakari, O. T., Juonala, M., Rönnemaa, T., Keltikangas-Järvinen, L., Räsänen, L., Pietikäinen, M., Hutri-Kähönen, N., Taittonen, L., Jokinen, E., Marniemi, J., *et al.* (2008). Cohort profile: the cardiovascular risk in Young Finns Study. *International journal of epidemiology*, **37**(6), 1220–1226.

Rantakallio, P. (1969). Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatrica Scandinavica*, **193**, Suppl. 193:1+.

Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., *et al.* (2008). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics*, **41**(1), 35–46.

Soininen, P., Kangas, A. J., Würtz, P., Tukiainen, T., Tynkkynen, T., Laatikainen, R., Järvelin, M.-R., Kähönen, M., Lehtimäki, T., Viikari, J., *et al.* (2009). High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst*, **134**(9), 1781–1785.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology*, **6**(5), e1000770.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**(16), 9440–9445.

Suhre, K., Shin, S.-Y., Petersen, A.-K., Mohney, R. P., Meredith, D., Wägele, B., Altmaier, E., Deloukas, P., Erdmann, J., Grundberg, E., *et al.* (2011). Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, **477**(7362), 54–60.

Tang, C. S. and Ferreira, M. A. (2012). A gene-based test of association using canonical correlation analysis. *Bioinformatics*, **28**(6), 845–850.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., *et al.* (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**(7307), 707–713.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

The Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*. doi:10.1038/ng.2797.

Thompson, J. R., Attia, J., and Minelli, C. (2011). The meta-analysis of genome-wide association studies. *Briefings in Bioinformatics*, **12**(3), 259–269.

Tukiainen, T., Kettunen, J., Kangas, A. J., Lyytikäinen, L.-P., Soininen, P., Sarin, A.-P., Tikkanen, E., O'Reilly, P. F., Savolainen, M. J., Kaski, K., *et al.* (2012). Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. *Human Molecular Genetics*, **21**(6), 1444–1455.

Vartiainen, E., Laatikainen, T., Peltonen, M., Juolevi, A., Männistö, S., Sundvall, J., Jousilahti, P., Salomaa, V., Valsta, L., and Puska, P. (2010). Thirty-five-year trends in cardiovascular risk factors in Finland. *International journal of epidemiology*, **39**(2), 504–518.

Vattikuti, S., Guo, J., and Chow, C. C. (2012). Heritability and genetic correlations explained by common snps for metabolic syndrome traits. *PLoS Genetics*, **8**(3), e1002637.

Waaijenborg, S., Verselewel de Witt Hamer, P. C., and Zwinderman, A. H. (2008). Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, **7**, 1–29.

Witten, D. M., Tibshirani, R., *et al.* (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, **8**(1), 1–27.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, **89**(1), 82–93.

Xie, Y., Pan, W., and Khodursky, A. B. (2005). A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, **21**(23), 4280–4288.

Zhang, F., Guo, X., and Deng, H.-W. (2011). Multilocus association testing of quantitative traits based on partial least-squares analysis. *PLoS ONE*, **6**(2), e16739.