Abstract

Recent progress in

Key words: ; ; ; ; ; ;

Introduction

There is strong evidence that genetic mechanisms account for a large part of the etiology of many complex disorders, such as cardiovascular diseases (CVD), cancers, schizophrenia, autism and others. Although genome scans have identified a number of candidate regions of interest for several complex diseases, most of these successful studies have explained only a small proportion of the disease heritability through the identified loci. Statistical approaches that take into account the complexity of disease etiology are thus of interest to understand the genetics of complex diseases.

Classical genome-wide association studies (GWAS) test for the association between one SNP and one trait at-a-time and correct for multiple testing. This is known as the univariate approach. Recently, some methods have been developed to simultaneously analyze multiple correlated traits in genome-wide association analysis such as MultiPhen [10], PCHAT [8], TATES [16] and GWAS for Multiple Continuous Secondary Phenotypes [12]. These methods are either performed one SNP at-a-time or based on replacing a gene (or a functional region) by a representative single SNP according to a particular criteria. However, realistically, many genetic factors contribute to the quantitative variation of many traits of interest. Moreover, most of these methods are based on the normality assumption of the measured traits which is not appropriate for analyzing data with different scales. Such approaches have, then, limited utility in our ultimate goal of understanding the paths leading from genetic variants to phenotypic variation.

Given that, for many diseases, a substantial amount of information has already been gathered on biological pathways involved in disease aetiology, we propose to model complex relationships between genetic variants in candidate genes and measured correlates traits, taking advantage of prior knowledge on clinical/molecular pathways (traits) and genetic pathways (genomic data). Indeed, analyzing genomic data through functional pathways offers the potential of greater power for discovery of connections to biological mechanisms. In fact, small but coordinated changes in sets of functionally related measures (i.e. pathways) can have significant effects. Therefore, broadening the analysis focus from gene to pathway, thus incorporating more complexity into the analysis, can address at least in part the challenges presented by pleiotropy and heterogeneity. Note that the term pathway has been used in very broad contexts in the literature. In this paper, sets or pathways are defined, for genomic data, as any latent functional structure within which data are correlated or co-regulated. For traits or phenotypes, we use the word pathways to represent a latent biological process relating traits to each other. Over a few years, dozens of different methods have been proposed to summarize the significance of a biological pathway from a collection of SNPs and to adjust for multiple testing at the pathway level. Some of the related issues have been discussed and reviewed in [17] and [7]. At the genome-wide level, several studies have used information from annotation databases to improve the power of GWAS or prioritize results ([9], [15], [14]). Although several methods are available in the literature for pathway analysis using prior information, none of these methods can fully address these aspects of complexity:

pleiotropy and heterogeneity of genetic effects and simultaneously take into account the correlation between the phenotypes on one hand and between the genetic variants on the other. Furthermore, current methods rely heavily on strong distributional assumptions. We believe that SEM models provide the ideal solution to this complex problem.

Over the past several decades, structural equation modeling (SEM) has been employed for the specification and testing of complex, path-analytic relationships between observed variables and underlying theoretical constructs, often called latent variables. SEM has become a remarkably popular method in social, natural, and health sciences for many reasons including the analytic flexibility and generality imparted by the procedure. In our context, the genotyped SNPs and the measured traits (or phenotypes) are the observed variables and the genetic pathways (genes or genomic regions) and the clinical (or molecular) pathways are the latent variables. A structural equation model constitutes two sub-models: measurement and structural models. The former specifies the relationships between observed and latent variables, whereas the latter expresses the relationships between latent variables. Traditionally, two statistical approaches have been proposed for SEM: covariance structure analysis (CSA) [1, 2] and partial least squares path modeling (PLSPM)[3]. While, CSA is based on the normality assumption and often results in improper solutions (e.g., negative variance estimates or correlations greater than one in absolute value), PLSPM does not rely on a stringent distributional assumption and avoids improper solutions because latent variables are treated as weighted composites of observed variables. Nonetheless, PLSPM estimation does not involve a global optimization criterion which is consistently optimized to estimate parameters. As a result, PLSPM can provide measures of overall model fit, but such measures are not suitable for model validation.

Generalized structured component analysis (GSCA) was recently introduced as another approach to SEM in the psychometric literature [4]. As in PLSPM, GSCA treats latent variables as weighted composites of observed variables. Moreover GSCA employs a global least-squares optimization criterion for parameter estimation and thus permits the calculation of measures of overall model fit. Consequently, GSCA becomes appealing to substantive researchers and practitioners for the following reasons: (i) GSCA does not require the multivariate normality assumption of observed variables; (ii) GSCA is free from improper solutions that tend to occur with high frequency in practice; (iii) it enables the provision of overall model-fit measures for theory testing and model comparison and (iv) GSCA has been extended to accommodate more advanced analyses that may be of great interest to a wide cadre of researchers, for instance, regularized estimation [5], analyses of higher-order latent variables [4], time-series data [6], and multiple group comparison [4]. A simulation study was also conducted to evaluate the relative performance of GSCA with respect to the traditional approaches under a variety of experimental conditions [4]. The results of the Monte Carlo analysis provide guidelines with respect to the conditions under which GSCA is to be preferred over the other approaches. Specifically, it was recommended that GSCA be used in lieu of PLSPM for general SEM purposes. Moreover, GSCA was recommended over CSA unless the researcher is confident that his/her model is correctly specified.

Despite its significant technical and empirical advantages, GSCA and SEM in general have been mainly used in psychology and other social sciences and remains novel to researchers in genetic studies. In the proposed research, GSCA will be adopted for the first time as a statistical framework for the joint analysis of multiple correlated traits and genomic mea-

sures. The proposed approach is set within a hypothesis-driven context and is not a pathway exploratory analysis. However, tuning such models to meet the specific needs of the clinicians and generic researchers will require the development of specific statistical tests, designed to identify key elements of influence on multiple other elements. The development of such tests constitute the main methodological contribution of this paper.

Method

GSCA framework

Consider the case where J candidate SNPs $(X_1, \cdots, X_J)$ and a set of K phenotypes $(X_{J+1}, \cdots, X_{J+K})$ are observed. Let $I = J + K$ denote the total number of observed variables. The J SNPs are mapped to G different genes $(\gamma_1, \cdots, \gamma_G)$ and the K phenotypes are involved in T different clinical pathways $(\gamma_{G+1}, \cdots, \gamma_{G+T})$. The G genes and T clinical pathways are unobserved latent variables. Figure 1(a) illustrates an example with $J = 5$ observed SNP genotypes $X_1, \cdots, X_5$ and $K = 3$ observed phenotypes X_6, X_7, X_8. The two latent variables γ_1 and γ_2 represent a gene and a clinical pathway respectively.

GSCA defines latent variables as weighted composites of observed variables

$$\gamma_\ell = \sum_{i \in S_\ell} w_{i\ell} X_i, \quad \ell = 1, \cdots, L$$

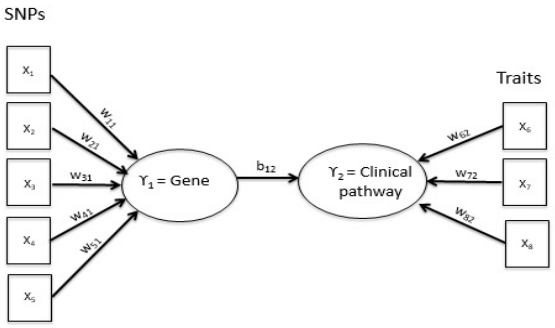
where S_ℓ denotes the set of observed variables mapped to the ℓ^{th} latent structure, $w_{i\ell}$ denotes the weight associated with the observed variable X_i in the definition of the latent variable ℓ and $L = G + T$ the total number of latent variables. This is the measurement model. The effect of a gene is, thus, modelled as the additive effect of Snps and a clinical pathway is defined as the weighted sum of the underlying phenotypes. In the example of Figure 1(a), the set S_1 includes the SNPs mapped to the gene γ_1, $S_1 = \{X_1, \cdots, X_5\}$ and $S_2 = \{X_6, X_7, X_8\}$ corresponds to phenotypes involved in the clinical pathway γ_2.

The relationships between latent variables are established given prior biological knowledge

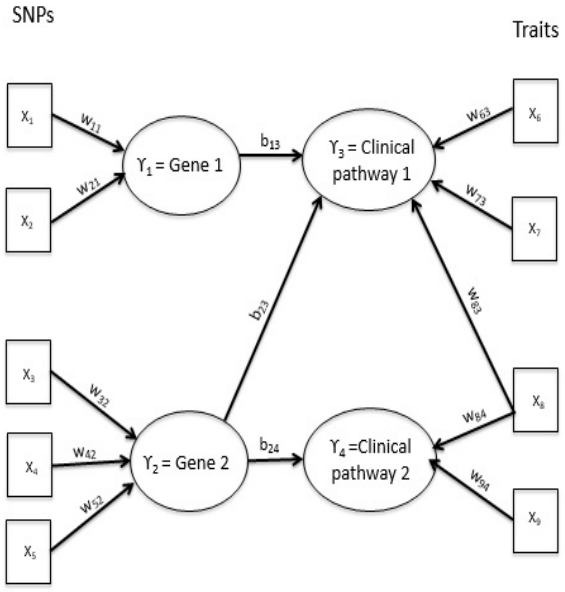
$$\gamma_\ell = \sum_{\ell'=1, \ell' \neq \ell}^{L} b_{\ell'\ell} \gamma_{\ell'\ell} + \epsilon_\ell, \quad \ell = 1, \cdots, L$$

where ϵ_ℓ represents the error term and $b_{\ell'\ell}$ represents the path coefficient linking two latent variables. This corresponds to the structural model. In Figure 1(a), b_{12} is the path coefficient linking the gene variable to the clinical pathway. In this example, the path coefficient b_{21} linking the clinical pathway to the gene variable is set to 0. Figure 1(b) provides an example of a more complex network, where SNPs are mapped to two genes, and traits are mapped to two clinical pathways. Note that a trait can map to several pathways and genes can also be associated with more than one clinical pathway. More generally, genes can be replaced by candidate functional regions.

In such a framework, GSCA provides a tool for the joint analysis of all genotypes and phenotypes and allows estimation of the model parameters using a global least squares optimization criterion.



(a) Five SNPs mapped to one gene; three traits mapped to a common clinical pathway



(b) Five SNPs mapped to two genes; four traits mapped to two clinical pathways. Gene 2 two affects both clinical pathways

Figure 1: Examples of path models.

Path models allow one to discern between direct and indirect relationships between variables. Direct effects go directly from one variable to another. Indirect effects are those characterizing a relationship between two variables mediated by one or more variables. In fact, the effect of a direct path from a gene to a clinical pathway is represented by the corresponding path coefficient. This coefficient quantifies the joint effect of the SNPs mapped to the gene on the multiple phenotypes together. The indirect effect of a Snp on a clinical pathway is the product of all the coefficients along the path relating them. In the example illustrated in Figure 1 (a) b_{11} quantifies the joint effect of $X_1, \cdots, X_5$ on the phenotypes $X_6, \cdots, X_8$ together and the indirect effect of Snp X_1 on clinical pathway γ_2 is given by $w_{11} \times b_{12}$.

Within the GSCA framework, a bootstrap method is used for estimating the standard errors of parameter estimates. However, constructing confidence intervals for the path coefficients b and weight coefficients w based on the bootstrap estimate of their respective standard

errors requires assumptions on the distributions of their respective estimators. On the other hand, constructing a bootstrap confidence interval based on the generated bootstrap samples can ... An alternative would be to perform permutation tests, first because such tests don't require any assumption on the distributions of the estimators and second because of their ability to handle simultaneously data with different scales. A permutation test procedure for the gene effect on a clinical pathway, or more generally the joint effect of multiple SNPs on multiple phenotypes, is described in next section*.

Test procedure

Structural equations models allow one to represent and measure different directed effects between observed and latent variables as well as between latent variables. Consider the model described in section* . Based on the GSCA estimation method, we want to test for the null hypothesis $H_0^{\ell,\ell'} : b_{\ell,\ell'} = 0$ of no effect of gene γ_ℓ on clinical pathway $\gamma_{\ell'}$. We propose to perform a permutation test as follows: SNP genotypes mapping to the gene tested are permuted by randomly re-assigning the subject's ID to genotypes. This destroys any association between the gene and the phenotypes. Note that it is important to apply the same permutations on all the tested SNPs not to break their LD pattern. The permutation distribution is obtained by calculating, for each permuted dataset, the estimate of the path coefficient $\hat{b}_{\ell,\ell'}$. This is an approximation of the distribution of the test statistic $\hat{b}_{\ell,\ell'}$ under the null hypothesis. The permutation p-value is, then, estimated by the proportion of permuted datasets for which the path coefficient $\hat{b}_{\ell,\ell'}$ is larger than its value calculated on the original dataset.

Results

Simulation study

We conducted a simulation study to assess the performance of the permutation test procedure. We considered the 9 scenarios illustrated in Figure 2. The scenarios involve one or two genes and one or two clinical pathways. One or two SNPs are mapped to each gene and two or three correlated traits are related to each clinical pathway. In each new scenario, a new latent variable or a new association is added. The associations simulated are controlled by the heritability coefficients of the SNPs on each trait. This coefficient is defined as the proportion of a trait variance explained by a particular SNP. Note that the datasets are not simulated according to the GSCA models (**??**) and (**??**) but by fixing the values of the different heritability coefficients as well as the value of the correlation coefficient r between traits involved in the same clinical pathway. Genotypes were simulated using HAPGEN2 program [13]. Based on a reference panel of known haplotypes and an estimate of the fine-scale recombination rate across the region, HAPGEN2 simulates datasets over large regions. The simulated data has the same linkage disequilibrium patterns as the reference data. The genome-wide haplotype data, minor allele frequencies, and recombination rates were downloaded from the 1000 genomes project website. The genotypes were, then, coded according to the additive model, i.e. a SNP's value is equal to the number of minor allele.

Scenario (a)

This scenario involves one Snp X_1 from gene γ_1 and two phenotypes X_2 and X_3 involved in one clinical pathway γ_2. We simulated 1000 samples of size $N = 1000$ of SNP *rs2070744* (on gene eNOS, chr7, minor allele frequency=0.375), one of the SNPs of interest in the QCAHS study denoted T786C. For each dataset, the two phenotypes (X_2, X_3) were simulated using a bivariate normal distribution such that h_2 and h_3, respectively the heritabilities of the SNP on the traits X_2 and X_3, take the values in $\{0, 0.007, 0.01, 0.03\}$ and such that the correlation coefficient r between the two traits is in $\{0.3, 0.6\}$. In this scenario, we have a single SNP so $w_{11} = 1$. To test the effect of the gene on the clinical pathway i.e. on the two traits jointly we used the GSCA-based permutation test as well as the Multiphen method to compare their performances.

Figure 3 shows the empirical power functions for the two methods. The two tests have roughly equal powers with a slight uniform advantage for the proposed GSCA-based procedure. Both tests are powerful enough to detect small effects corresponding to small heritabilities. The results also show that, unsurprisingly, a stronger correlation between the trait reduces the power of the test.

Unlike the GSCA-based approach, MultiPhen can only be applied to cases with a single SNP. In the remaining scenarios only GSCA-based test is investigated.

Scenario (b)

We simulated 1000 samples of size $N = 1000$ of SNPs *rs2070744* and *rs1799983*, denoted X_1 and X_2 respectively, mapped to gene eNOs (γ_1). SNP *rs1799983* has a minor allele frequency of 0.401 and is also of interest in the QCAHS study (denoted Glu298Asp). The correlation coefficient between the two SNPs is equal to 0.345.

For each simulated dataset, we randomly selected one from the two SNPs, then simulated two phenotypes (X_3, X_4) associated with the selected SNP. In this scenario we want to investigate the case where two traits are associated to the same gene with different degrees. For this purpose, the phenotypes were simulated using a bivariate normal distribution with a correlation coefficient r in $\{0.3, 0.6\}$, such that $h_{3,1}$ and $h_{4,1}$, respectively the heritabilities of the traits X_3 and X_4 with respect to the SNP selected from gene γ_1, take the values in $\{0, 0.007, 0.01, 0.03\}$.

Table 1 reports the probability of rejecting the null hypothesis for different values of $h3, 1$, $h_{4,1}$ and r. If the null hypothesis of no association is true i.e the corresponding heritability is zero, this probability must be close to the level of the test. Under alternatives, it is an estimate of the power of the test. As one can see, the test respects the used level of 5%. Furthermore, a non associated trait slightly reduces the power of the test. For instance, for $h_{3,1} = h_{4,1} = 0.01$ the power is equal to 0.896 while for $h_{3,1} = 0$ and $h_{4,1} = 0.01$ the obtained power is 0.756. Nonetheless, the power of the test remains satisfactory. The power decreases for stronger correlation between the phenotypes except when one of the traits is not associated with the gene ($h_{3,1} = 0$).

Scenario (c)

In addition to the SNPs *rs2070744* and *rs1799983* (X_1 and X_2) we simulated SNPs *rs1042713* and *rs1042714* respectively denoted X_3 and X_4 (Gly16Arg and Gln27Glu in the QCAHS dataset). These SNPs are mapped to gene B2ADR on chromosome 5. Their minor allele frequencies are equal to 0.235 and 0.380 respectively.

We want to investigate the case where two genes are associated with different degrees to the same clinical pathway. For each simulated dataset, we randomly selected one from the two SNPs mapped to each gene, then simulated two phenotypes (X_5, X_6) associated with the two selected SNPs (one from gene γ_1 and the other from gene γ_2). Let $h_{5,1}$, $h_{6,1}$, $h_{5,2}$ and $h_{6,2}$ denote the heritabilities of the traits X_5 and X_6 with respect to the chosen SNP from genes γ_1 and γ_2 respectively. The phenotypes were simulated using a bivariate normal distribution such that $h_{5,1} = h_{6,1} = h_{.,1}$ and $h_{5,2} = h_{6,2} = h_{.,2}$ take the values in $\{0, 0.007, 0.01, 0.03\}$ with a correlation coefficient r in $\{0.3, 0.6\}$.

The results presented in Table 2 show that the proposed test is able to distinguish genes with null and moderate effects in a scenario involving more than one gene. It respects the level of test and is powerful enough to detect small heritabilities. As one can notice, adding a non associated gene to the model slightly decreases the power of detecting the other (associated) gene. Furthermore, the test on a gene is more powerful when the other gene has a larger heritability coefficient. For instance, a gene with a heritability of 0.7% is detected with a probability of almost 80% when the other gene has the same heritability coefficient on the phenotypes, and with probabilities of almost 83% and 84% when the heritability coefficient of the other gene is equal to 1% and 2% respectively. Finally, the power also decreases for stronger correlation between the phenotypes.

Scenario (d)

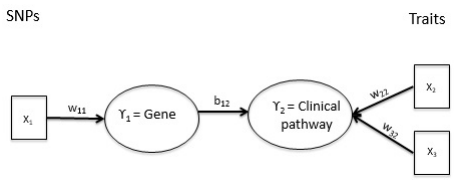
The same SNPs as in scenario (c) were simulated. Also, one from the two SNPs mapped to each gene were randomly selected. Then, four phenotypes were simulated, (X_5, X_6) associated with the selected SNP from gene γ_1 and (X_7, X_8) associated with the one from γ_2, following a multivariate normal distribution such that $h_{5,1} = h_{6,1} = h_1$, $h_{7,2} = h_{8,2} = h_2$. Note that in this case $h_{5,2} = h_{6,2} = h_{7,1} = h_{8,1} = 0$. The correlation coefficient between phenotypes involved in the same clinical pathway is equal to 0.3. It is null otherwise.

In this scenario, we compare the test's performances in two situations: correctly specified model and misspecified model. In the latter, we specify additional associations between γ_1 and γ_4 and between γ_2 and γ_3, as indicated by the dashed lines on Figure 2 (d). When the model is wrongly specified, in addition to the significance of b_{13} and b_{14}, we test for the association between γ_1 and γ_4 and between γ_2 and γ_3 which are, in reality, null. In Table 10, we report the results for the two situations. For this sceanrio, when the model is correctly specified, a non associated gene doesn't affect the power of detecting the other associated gene. This is due to the fact that, in the true model, there are two separated submodels each one involving one gene and one clinical pathway. In this case the GSCA estimates of the path and weight coefficients are the same when the two submodels are estimated together (as in our scenario) or separately. In the misspecified model case, our test recognizes the wrongly added associations and reject them while keeping the same power for detecting the true associations as in the correctly specified model case.

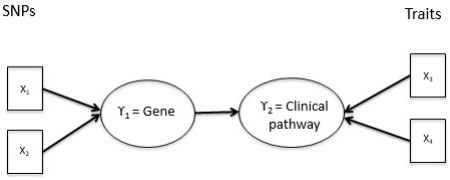
Scenarios (e)-(i)

Now, we want to assess the effect of the complexity of the model on the performance of the test. Starting from scenario (d), we progressively add new associations between genes and pathways or a connection between phenotype X_7 and clinical pathway γ_4. The results are reported in Table 4 for scenarios (e) to (g) and in Table 5 for scenarios (h) and (i). Comparing the results of scenarios (e), (g) and (i) with those of (d), (f) and (h) respectively, we see that the additional connection between X_7 and γ_3 increases the power for detecting the gene(s) associated with γ_3. Then, the more phenotypes involved in the clinical pathway are considered in the model, the more powerful is our test. Now, the comparison of the results of scenarios (e), (f) and (h) shows that the number of associations between the genes and the clinical pathways doesn't affect the power of the test.

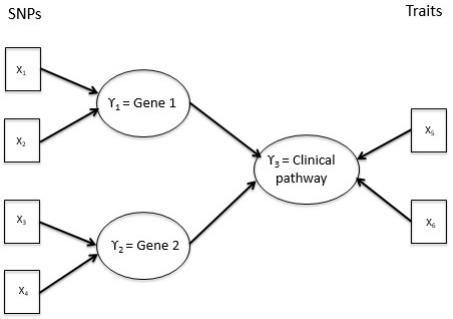
Figure 2: Scenarios for the simulation study



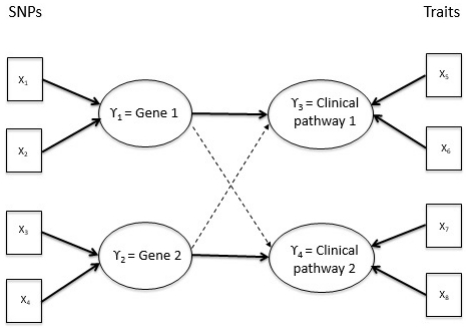
(a) One SNP mapped to one gene; two traits mapped to a common clinical pathway



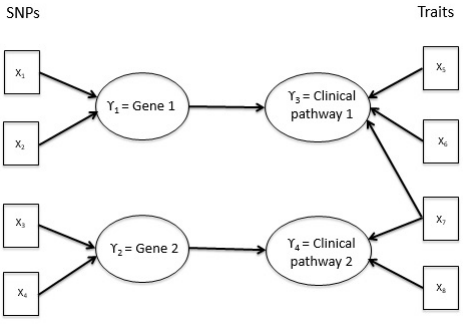
(b) Two SNPs mapped to one gene; two traits mapped to a common clinical pathway



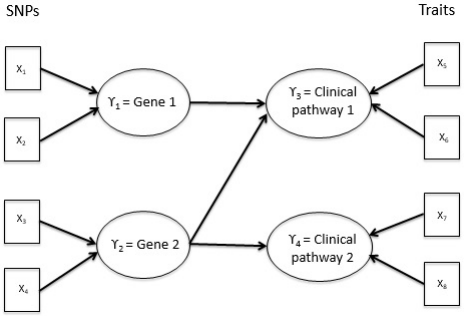
(c) Two SNPs mapped to two genes; two traits mapped to a common clinical pathway



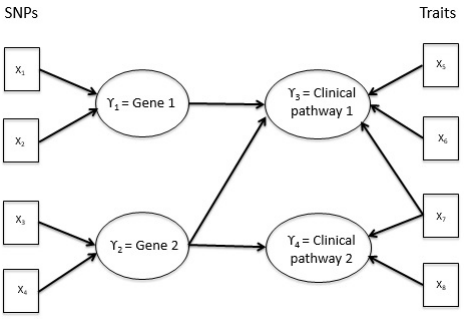
(d) Two SNPs mapped to two genes; four traits mapped to two clinical pathways; each gene is associated to one clinical pathway



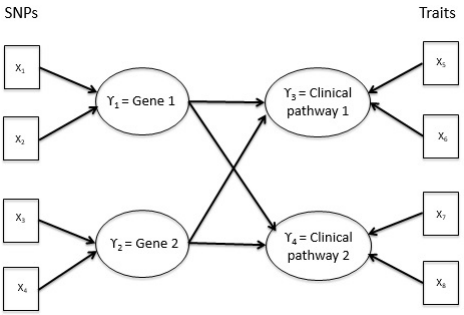
(e) Scenario (d) with a common trait shared by both clinical pathways



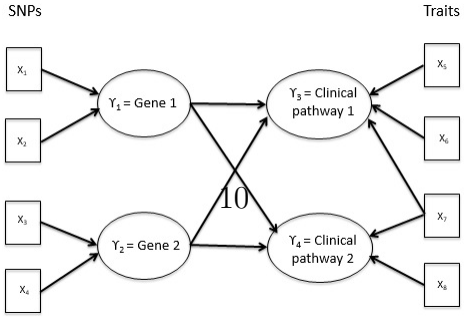
(f) Scenario (d) with gene γ_2 associated to both clinical pathways



(g) Scenario (f) with a common trait shared by both clinical pathways



(h) Scenario (f) with gene γ_1 also associated to both clinical pathways



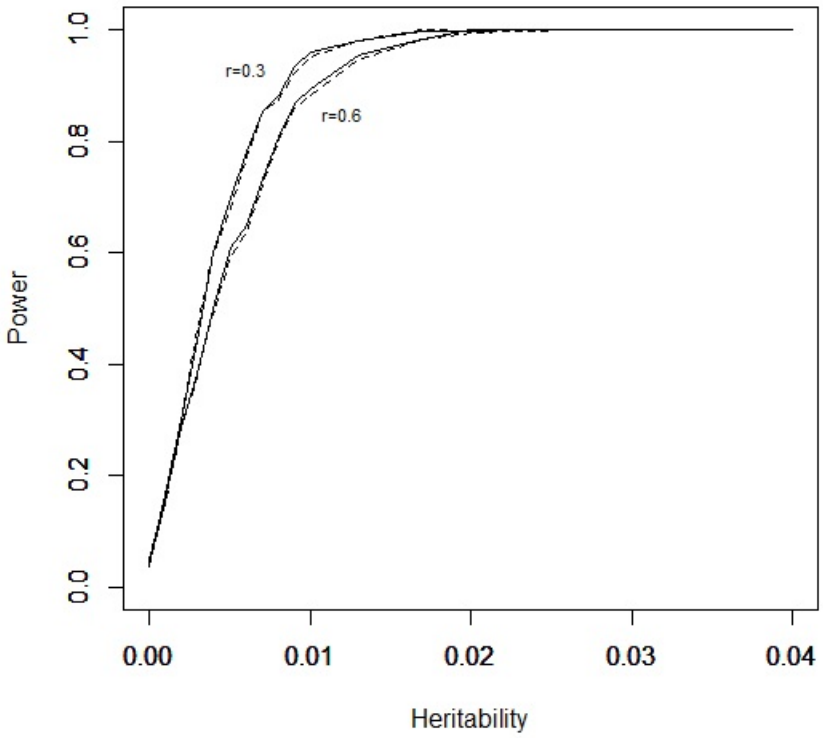


Figure 3: Empirical power comparison of the GSCA-based test (solid line) and MultiPhen test (dashed line) for scenario (a) with $r = 0.3$ and $r = 0.6$.

Table 1: Results for scenario (b).Probability of rejecting the null hypothesis of no gene effect on the clinical pathway. The coefficients $h_{3,1}$ and $h_{4,1}$ denote respectively the heritabilities of the traits X_3 and X_4 with respect to the SNP selected from gene 1. The correlation between the traits is denoted by r.

$h_{3,1}$	$h_{4,1}$	$r = 0.3$	$r = 0.6$
0	0	0.050	0.061
0	0.007	0.606	0.768
0	0.01	0.756	0.910
0	0.03	0.999	1
0.007	0.007	0.758	0.668
0.007	0.01	0.851	0.763
0.007	0.03	0.997	0.998
0.01	0.01	0.896	0.809
0.01	0.03	1	0.998
0.03	0.03	1	1

Table 2: Results for scenario (c). Probability of rejecting the null hypotheses $H^0_{\gamma_\ell,\gamma_3}$, $\ell = 1, 2$ no gene effect on the clinical pathway. The coefficients $h_{5,1} = h_{6,1} = h_{.,1}$ and $h_{5,2} = h_{6,2} = h_{.,2}$ denote the heritabilities of the traits X_5 and X_6 with respect to the SNPs selected, respectively, from gene 1 and gene 2. The correlation between the traits is denoted by r.

		$r = 0.3$		$r = 0.6$	
$h_{.,1}$	$h_{.,2}$	$H^0_{\gamma_1,\gamma_3}$	$H^0_{\gamma_2,\gamma_3}$	$H^0_{\gamma_1,\gamma_3}$	$H^0_{\gamma_2,\gamma_3}$
0	0	0.048	0.051	0.049	0.054
0	0.007	0.045	0.748	0.059	0.657
0	0.01	0.048	0.911	0.045	0.849
0	0.03	0.040	1	0.045	1
0.007	0.007	0.796	0.794	0.697	0.732
0.007	0.01	0.829	0.937	0.704	0.876
0.007	0.03	0.837	1	0.79	1
0.01	0.01	0.943	0.940	0.876	0.873
0.01	0.03	0.946	1	0.911	1
0.03	0.03	1	1	1	1

Table 3: Results for scenario (d). Probability of rejecting the null hypotheses. The coefficients $h_{5,1} = h_{6,1} = h_{.,1}$ and $h_{7,2} = h_{8,2} = h_{.,2}$ respectively denote the heritability of the traits X_5 and X_6 with respect to the SNPs selected from gene 1 and the heritability of the traits X_7 and X_8 with respect to the SNPs selected from gene 2. The correlation between each two traits from the same clinical pathway is equal to 0.3. $H^0_{\gamma_\ell,\gamma_{\ell'}}$ corresponds to the null hypotheses of no effect of gene γ_ℓ on the clinical pathway $\gamma_{\ell'}$.

		Correctly specified		Misspecified			
$h_{.,1}$	$h_{.,2}$	$H^0_{\gamma_1,\gamma_3}$	$H^0_{\gamma_2,\gamma_4}$	$H^0_{\gamma_1,\gamma_3}$	$H^0_{\gamma_2,\gamma_4}$	$H^0_{\gamma_1,\gamma_4}$	$H^0_{\gamma_2,\gamma_3}$
0	0	0.054	0.048	0.053	0.048	0.061	0.051
0	0.007	0.054	0.749	0.051	0.752	0.053	0.031
0	0.01	0.047	0.923	0.046	0.923	0.054	0.021
0	0.03	0.051	1	0.045	1	0.063	0.013
0.007	0.007	0.771	0.763	0.765	0.756	0.039	0.036
0.007	0.01	0.776	0.905	0.775	0.903	0.021	0.028
0.007	0.03	0.763	1	0.766	1	0.024	0.012
0.01	0.01	0.885	0.907	0.884	0.909	0.036	0.026
0.01	0.03	0.916	1	0.909	1	0.022	0.023
0.03	0.03	1	1	1	1	0.012	0.026

Table 4: Results for scenarios (e), (f) and (g). Probability of rejecting the null hypotheses. The coefficients $h_{.,1}$ denotes the heritability of the traits mapped to clinical pathway γ_3 with respect to the SNPs selected from gene γ_1 and $h_{.,2}$ the heritability of the traits mapped to clinical pathway γ_4 with respect to the SNPs selected from gene γ_2. The correlation between each two traits from the same clinical pathway is equal to 0.3. $H^0_{\gamma_\ell,\gamma_{\ell'}}$ corresponds to the null hypotheses of no effect of gene γ_ℓ on the clinical pathway $\gamma_{\ell'}$.

		Scenario (e)		Scenario (f)			Scenario (g)		
$h_{3,1}$	$h_{4,2}$	$H^0_{\gamma_1,\gamma_3}$	$H^0_{\gamma_2,\gamma_4}$	$H^0_{\gamma_1,\gamma_3}$	$H^0_{\gamma_2,\gamma_3}$	$H^0_{\gamma_2,\gamma_4}$	$H^0_{\gamma_1,\gamma_3}$	$H^0_{\gamma_2,\gamma_3}$	$H^0_{\gamma_2,\gamma_4}$
0	0	0.046	0.047	0.060	0.046	0.050	0.053	0.045	0.056
0	0.007	0.055	0.744	0.059	0.754	0.715	0.048	0.781	0.72
0	0.01	0.055	0.890	0.059	0.882	0.883	0.038	0.929	0.915
0	0.03	0.045	1	0.052	1	1	0.059	1	1
0.007	0.007	0.796	0.760	0.793	0.786	0.735	0.848	0.845	0.771
0.007	0.01	0.797	0.902	0.820	0.924	0.888	0.868	0.950	0.876
0.007	0.03	0.790	1	0.859	1	1	0.925	1	1
0.01	0.01	0.931	0.909	0.940	0.923	0.880	0.960	0.964	0.895
0.01	0.03	0.924	1	0.949	1	1	0.978	1	1
0.03	0.03	1	1	1	1	1	1	1	1

Table 5: Results for scenarios (h) and (i). Probability of rejecting the null hypotheses. The coefficients $h_{.,1}$ and $h_{.,2}$ respectively denote the heritability of all the traits with respect to the SNPs selected from gene γ_1 and gene γ_2. The correlation between each two traits from the same clinical pathway is equal to 0.3. $H^0_{\gamma_\ell,\gamma_{\ell'}}$ corresponds to the null hypotheses of no effect of gene γ_ℓ on the clinical pathway $\gamma_{\ell'}$.

		Scenario (h)				Scenario (i)			
$h_{3,1}$	$h_{4,2}$	$H^0_{\gamma_1,\gamma_3}$	$H^0_{\gamma_1,\gamma_4}$	$H^0_{\gamma_2,\gamma_3}$	$H^0_{\gamma_2,\gamma_4}$	$H^0_{\gamma_1,\gamma_3}$	$H^0_{\gamma_1,\gamma_4}$	$H^0_{\gamma_2,\gamma_3}$	$H^0_{\gamma_2,\gamma_4}$
0	0	0.061	0.038	0.053	0.049	0.040	0.049	0.053	0.042
0	0.007	0.048	0.053	0.736	0.716	0.047	0.056	0.82	0.762
0	0.01	0.045	0.051	0.878	0.897	0.055	0.050	0.932	0.890
0	0.03	0.040	0.043	1	1	0.048	0.053	1	1
0.007	0.007	0.779	0.764	0.766	0.785	0.868	0.796	0.848	0.796
0.007	0.01	0.790	0.771	0.902	0.908	0.868	0.802	0.947	0.906
0.007	0.03	0.828	0.826	1	1	0.906	0.813	1	1
0.01	0.01	0.938	0.910	0.914	0.929	0.916	0.923	0.974	0.924
0.01	0.03	0.931	0.932	1	1	0.99	0.957	1	1
0.03	0.03	1	1	1	1	1	1	1	1

Analysis of the Quebec and Adolescent Health and Social Survey (QCAHS) data

The QCAHS was designed to characterize CVD risk factors in a representative sample of Quebec youth and targeted all youth aged 9, 13 and 16 years attending public or private schools in Quebec. Detailed descriptions of the QCAHS design and methods can be found in [11]. DNA is available on 1707 French Canadian participants of the QCAHS study (860 boys and 847 girls). Table 6 lists of 8 traits measured in QCAHS that are of interest in our study. We use the z-score transformations of these traits to standardize for...? These traits are grouped into three pathways: lipid metabolism, energy metabolism and blood pressure control. Available genotypic data on 35 variants within 25 genes are listed in Table 7 along with biological pathways within which they fall. Among the considered genetic variants 33 are SNPs coded according to the additive model and two polymorphisms. The first polymorphism belongs to gene APOE. It has three isoforms and then admits 6 different genotypes. It is coded using 5 indiactors APOE1,. . ., APOE5. The second is variant of gene PCSK9? I need more information about this morphism, is it a SNP or a polymorphism. How was is coded?

Our new approach provides a tool to investigate complex relationships between genetic variants in these candidate genes and the CVD-related traits, taking advantage of prior knowledge on CVD related pathways available in the literature. Figure 4 shows a diagram illustrating the path model we fitted to the data, along with the weights and path coefficients estimated by GSCA.

In table **??** we report the results obtained by 3 approaches. the univariate test, the GSCA procedure applied on one gene and one metabolism at-a-time, and the GSCA procedure applied on all data at once. The first is the univariate approach along with a Bonferroni correction for the $41*8$ performed tests. The second is the GSCA-based approach applied one gene and one metabolism at-a-time. As the Blood Pressure Control pathway is independent from the other pathways of the model, the multiple tests correction is then applied separatly for each submodel i.e. considers only the number of path coefficients tested in each submodel. Thus, for the path coefficients involved in the Lipid and/or Energy metabolisms and for the path coefficients involved in the Blood Pressure Control pathway, a Bonferroni correction for respectively19 and 9 tests is used. The third approach is the GSCA-based procedure is applied on all the data simultaneously. For this latter, no multiple test correction is needed.

The results are reported in table **??**. The GSCA-based test allowed us to discover genetic association to CVD-related phenotypes that are not discovered by the univariate approach. Except for the SNP C514T from gene HL that was detected as associated to HDL phenotype by the univariate test but the GSCA-based test for the association between gene HL and the Lipid metabolism gave a *p-value* of 0.08.

Table 6: Traits of interest available in QCAHS.

Lipid	Blood pressure	Energy
HDL cholesterol	Systolic (SBP)	Fasting glucose
LDL peak particle size	Diastolic (DBP)	Fasting insulin
Triglycerides (TG)		
Apolipoprotein B (ApoB)		
Apolipoprotein A1 (ApoA1)		

Table 7: Candidate genotypes available in QCAHS possibly related to the considered pathways.

Gene	Variant	Lipid metabolism	Energy Metabolism	Blood pressure control
CETP	TaqIB	X		
ApoC3	C-482T	X		
ABCA1	Arg219Lys	X		
FABP-2	T54A	X		
ApoA1	G-75A	X		
ApoE	E1E2E3	X		
HL	C-514T	X		
LPL	Hind III	X		
MTP	G-493T	X		
PON1	A192G	X		
PON2	C311G	X		
PCSK9	R46L L15-L16insL	X		
PGC	G1564A G-1302A	X	X	
Adiponectin	T45G G276T -11391 -11377	X	X	
PPARg2	Pro12Ala	X	X	
TNFa	G-308A G-238A		X	
eNOS	T-786C Glu298Asp			X
a23-AR	DelGlu301-303			X
b1-AR	Gly389Arg			X
b2-AR	Gly16Arg Gln27Glu			X
b3-AR	Trp64Arg			X
ACE	Ins/Del			X
AGT	Met235Thr			X
AGTR1	A1166C			X
LEPR	Lys656Asn Gln223Arg Lys109Arg			X

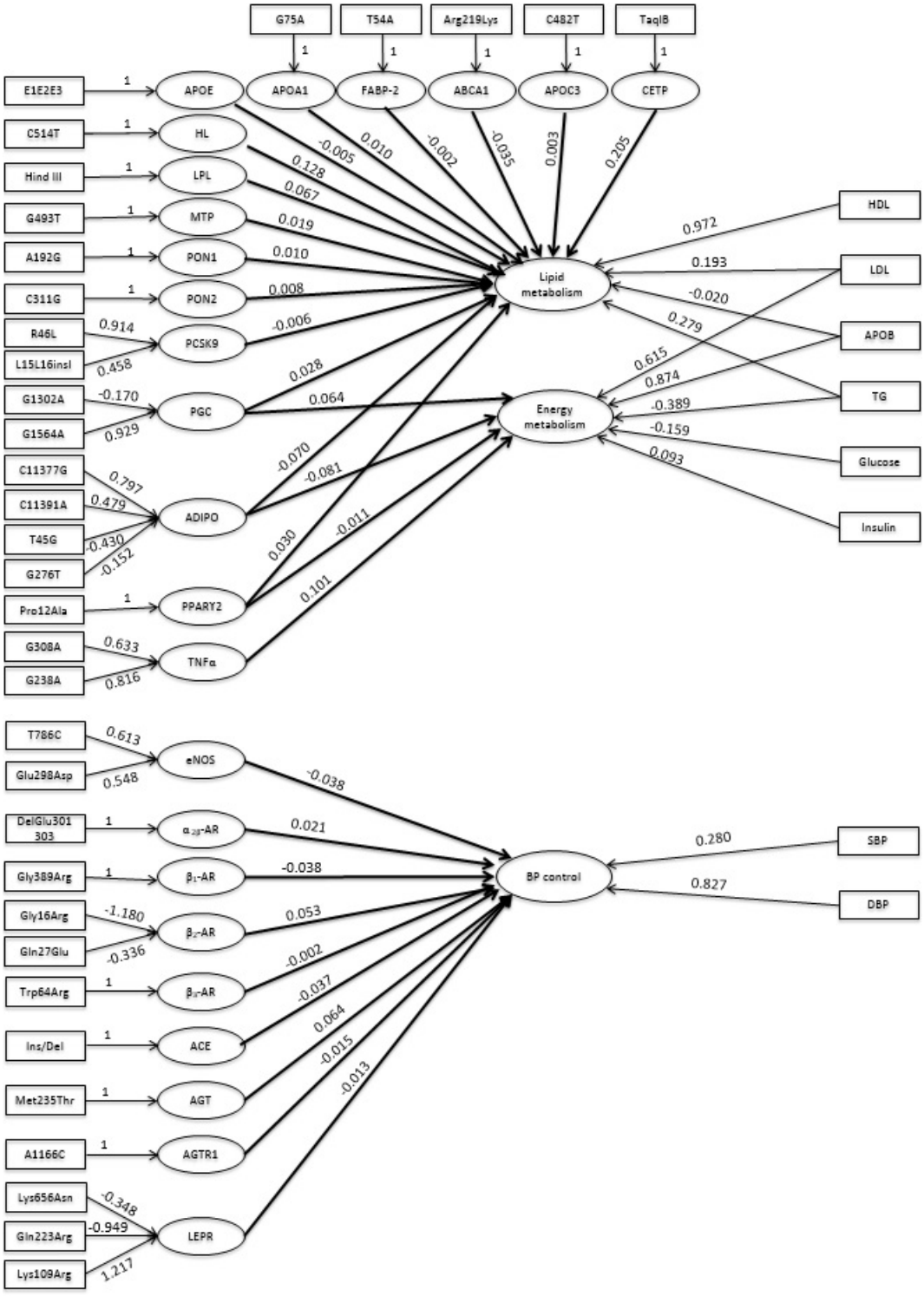


Figure 4: Path model for the QCAHS data. Weights and path coefficients are included.

Table 8: Associations detected by the univariate test with a Bonferroni correction.

Gene	Variant	HDL	LDL	APOB	TG	Glucose	Insulin	SBP	DBP
CETP	TaqIB	X	X						
APOC3	C482T								
ABCA1	Arg219Lys								
FABP-2	T54A								
APOA1	G75A								
	APOE1		X						
	APOE2			X					
APOE	APOE3								
	APOE4								
	APOE5			X					
HL	C514T	X							
LPL	Hind3								
MTP	G493T								
PON1	A192G								
PON2	C311G								
	R46L			X					
PCSK9	L15-L16insl1								
	L15-L16insl2								
	L15-L16insl3								
PGC	G1302A								
	G1564A								
	G11377C								
ADIPO	G11391A								
	T45G								
	G276T								
PPARg2	Pro12Ala								
TNFa	G308A								
	G238A								
eNOS	T786C								
	Glu298Asp								
a2b-AR	DelGlu301303								
b1-AR	Gly389Arg								
b2-AR	Gly16Arg								
	Gln27Glu								
b3-AR	Trp64Arg								
ACE	Ind/Del								
AGT	Met235Thr								
AGTR1	A1166C								
	Lys656Asn								
LEPR	Gln223Arg								
	Lys109Arg								

Table 9: Results for QCAHS data analysis. Gene par gene, on compare les pvalues avec $0.005/19 = 0.0026$ pour les genes des metab Lipid et Energy et avec $0.05/9 = 0.005$ pour le BP control

	All genes simultaneously			Gene by gene		
Gene	Lipid	Energy	BP	Lipid	Energy	BP
1 CETP	0.0006			0		
2 APOC3						
3 ABCA1						
4 FABP-2						
5 ApoA1				0.0205		
6 ApoE	0			0		
7 HL	0.0815			0		
8 LPL	0.0144			0.0308		
9 MTP						
10 PON1						
11 PON2	0.0109			0.0162		
12 PCSK9	0			0.0001		
13 PGC		0.071				
14 Adiponectin		0.0499				
15 PPARg2						
16 TNFa		0.0044			0	
17 ENOs						
18 a23AR						
20 b1AR						
21 b2AR						
22 b3AR						
23 ACE						
24 AGT			0.032			0.0202
25 AGTR1						
26 LEPR						

Table 10: Results for QCAHS data analysis. Gene par gene, on compare les pvalues avec $0.005/19 = 0.0026$ pour les genes des metab Lipid et Energy et avec $0.05/9 = 0.005$ pour le BP control

	All genes simultaneously			Gene by gene		
Gene	Lipid	Energy	BP	Lipid	Energy	BP
1 CETP	X			X		
2 APOC3						
3 ABCA1						
4 FABP-2						
5 ApoA1						
6 ApoE	X			X		
7 HL				X		
8 LPL	X					
9 MTP						
10 PON1						
11 PON2	X					
12 PCSK9	X			X		
13 PGC		X				
14 Adiponectin		X				
15 PPARg2						
16 TNFa		X			X	
17 ENOs						
18 a23AR						
20 b1AR						
21 b2AR						
22 b3AR						
23 ACE						
24 AGT			X			
25 AGTR1						
LEPR						

Discussion

model selection strategy will need to be developed (see Method) in order to draw conclusions about the relationship between all the variables.

Conclusion

Our main objective is to develop a non-parametric statistical framework for the joint analysis of multiple correlated traits and multiple genomic measures from one or more candidate functional regions in genetic studies. Our approach will be based on structural equation modeling. This will allow us, within a single model, to include the following features: (i) develop statistical tests for the association between sets or pathways and between individual

observed measures and pathways; (ii) characterization of the individual contribution of each observed measure on its pathway; (iii) use of prior biological information about clinical and genomic pathways; iv) modeling of pleiotropic effects and genetic heterogeneity; v) accounting for the correlation between measures within pathways; vi)

(4) GSCA has been extended to accommodate more advanced analyses that may be of great interest to a wide cadre of researchers; for instance, regularized estimation ([10]), analyses of higher-order latent variables [8], time-series data [11], and multiple group comparison [8].

References

[1] R.D. Bock. Components of variance analysis as a structural and discriminal analysis for psychological tests. *British Journal of Statistical Psychology*, 13:151–163, 1960.

[2] R.D. Bock and R.E. Bargmann. Analysis of covariance structures. *Psychometrika*, 31:507–533, 1966.

[3] V. Esposito Vinzi, L. Trinchera, S. Esposito Vinzi, L. Trinchera, and S. Amato. Pls path modeling: From foundations to recent developments and open issues for model assessment and improvement. In V. Esposito Vinzi, editor, *Handbook of Partial Least Squares*. Springer, 2010.

[4] H. Hwang. Generalized structured component analysis. *Psychometrika*, 69:81 – 99, 2004.

[5] H. Hwang and Y. Takane. Regularized generalized structured component analysis. *Psychometrika*, 74:517 – 530, 2009.

[6] K. Jung, Y. Takane, H. Hwang, and T. S. Woodward. Dynamic gsca (generalized structured component analysis) with applications to the analysis of effective connectivity in functional neuroimaging data. *Psychometrika*, 77:827–848, 2012.

[7] P. Khatri, M. Sirota, and A. Butte. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375, 2012.

[8] R. Klein, R. Klein, C. Zeiss, E. Chew, A. Henning, J. Sangiovanni, S. Mane, J. Ott, C. Barnstable, and J. Hoh. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 385(2005), 2012.

[9] L. Luo, G. Peng, Y. Zhu, H. Dong, C.I. Amos, and M. Xiong. Genome-wide gene and pathway analysis. *Eur J Hum Genet*, 18:1045–1053, 2010.

[10] P.F. O'Reilly, C.J. Hoggart, Y. Pomyen, P.E. Calboli, M. Jarvelin, and L.J. Coin. MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. *PLoS One*, 7(5):e34861, 2012.

[11] G. Paradis, M. Lambert, J. O'Loughlin, C. Lavallee, J. Aubin, P. Berthiaume, M. Ledoux, E.E. Delvin, E. Levy, and J.A. Hanley. The quebec child and adolescent health and social survey: design and methods of a cardiovascular risk factor survey for youth. *Can J Cardiol*, 19:523 – 531, 2003.

[12] ElizabethD. Schifano, Lin Li, DavidC. Christiani, and Xihong Lin. Genome-wide association analysis for multiple continuous secondary phenotypes. *The American Journal of Human Genetics*, 92(5):744 – 759, 2013.

[13] Z. Su, J. Marchini, and P. Donnelly. Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305, 2011.

[14] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, and et al. Lander, E.S. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102:15545–15550, 2008.

[15] A. Torkamani, E.J. Topol, and N.J. Schork. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 92:265–272, 2008.

[16] Sophie van der Sluis, Danielle Posthuma, and Conor V. Dolan. Tates: Efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet*, 9(1):e1003235, 01 2013.

[17] K. Wang, M. Li, and H. Hakonarson. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*, 11:843 – 854, 2010.