UPPSALA
UNIVERSITET

Mediation modeling and analysis for high-throughput omics data

Zheng Ning

Department of Statistics

Uppsala University

Supervisors: Xia Shen, Yudi Pawitan and Fan Yang-Wallentin

June 2015

Abstract

There is a strong need for powerful unified statistical methods for discovering underlying genetic architecture of complex traits with the assistance of omics information. In this paper, two methods aiming to detect novel association between the human genome and complex traits using intermediate omics data are developed based on statistical mediation modeling. We demonstrate theoretically that given proper mediators, the proposed statistical mediation models have better power than genome-wide association studies (GWAS) to detect associations missed in standard GWAS that ignore the mediators. For each of the modeling methods in this paper, an empirical example is given, where the association between a SNP and BMI missed by standard GWAS can be discovered by mediation analysis.

Keywords. Mediation model, metabolite, SNP, BMI

Contents

1 Introduction

1.1 Biological background

A general goal of genetic studies in biology is to discover the genetic architectures underlying complex traits such as height, body mass index (BMI), disease endpoints, etc. Such work is becoming more and more feasible owing to the completion of the human genome project (Adams et al., 1991) and developments of high-throughput technology. The most popular strategy to detect the associations between genetic variants and a complex trait is genome-wide association study (GWAS). Generally speaking, GWAS performs a simple regression to test the association between each genetic variant and the complex trait. Nowadays, the most commonly used genetic markers in GWAS are single-nucleotide polymorphisms (SNPs), which refers to a DNA sequence variation occurring commonly within a population at a single base pair position, where different sequence alternatives (alleles) exist in normal individuals in some population (Brookes, 1999). The number of one allele among two alleles at the single base pair position is called its allelic dosage. The frequency at which the least common allele occurs in a given population is named minor allele frequency (MAF). Two concepts in genetics will be mentioned in this thesis: phenotype and genotype. Phenotype is a description of an individual's actual physical characteristics or traits. For example, both the value of metabolites and BMI belong to phenotypes. In contrast, one's genotype refers to the individual's complete heritable genetic identity. Most phenotypes are influenced by both genotype and by environmental factors.

Nevertheless, GWAS has at least two weaknesses that need to be taken care of and improved. First of all, considering the complicated mechanisms between SNPs and traits, the signal to noise ratio is small. Thus, GWAS often collects large samples to obtain sufficient power. Secondly, multiple testing problem is severe in GWAS where a lot of statistical tests are performed simultaneously. Since the number of tests is the number of independent SNPs considered in GWAS, which can be hundreds of thousands or even millions, the significance threshold should be corrected, e.g. using Bonferroni correction, for multiplicity. Given the large sample size requirement and stringent significance threshold, the discovery power of GWAS is usually low in many cases.

Our study concerns an extra level of information, i.e. intermediate omics data besides genomics. Omics data refer to a field of study in biology ending with -omics, such as proteomics and metabolomics, which stand for protein and metabolite spectra generated by high-throughput biochemical techniques. Such modern omics techniques allow characterization of functional mechanisms, which can link commonly used genomic data to complex traits. It is clear that such intermediate phenotypes play an essential role as a natural mediator in the biology of complex traits, in the chain of "genomics - intermediate omics - complex traits". Although these high-throughput omics data are widely considered informative, combining information from multiple levels is rather challenging in statistical analysis (Ritchie et al., 2015).

In this paper, the trait we take as an example is body mass index (BMI), i.e. the ratio of one's weight to his or her height squared, which is of essential clinical relevance globally as a risk factor for a number of diseases, such as coronary heart disease (Freedman et al., 2001) and hyperlipidemia (Kawada, 2002). Since BMI is a very good measurement of one's level of body fat and also can be measured and calculated easily, clinicians deem it as an efficient screening tool. To date, using large GWAS meta-analysis, about a hundred independent genetic variants have been discovered to be associated with BMI, however, unfortunately, they still explain little of the variation of BMI (Locke et al., 2015b).

1.2 Previous studies of mediation analysis

The concept of mediation refers to the effect of one variable on another that is mediated through intermediate variables called mediators. Traditionally, mediation analysis has been developed, discussed and commonly used in the social sciences. Baron and Kenny (1986) firstly clarified the distinction between moderator and mediator, and introduced the standard regression approach to mediation. Later, in the causal inference literature, counterfactual notions were presented in Robins and Greenland (1992), so that the mediation effects can be generally defined without any specific statistical model. VanderWeele and Vansteelandt (2009) showed that the direct and indirect effects described in the counterfactual framework can be estimated using regression analysis under appropriate identification conditions. To deal with nonlinear and non-parametric models, Imai et al. (2010) developed an estimation method based on simulations,

which can also be used to achieve a set of sensitivity analyses mentioned in their study.

1.3 The aim of the study

Modern omics techniques are able to simultaneously measure many intermediate phenotypes between the genome and complex traits such as BMI. However, no clear and powerful method has been developed, using such omic information, to enhance our ability to identify more functional loci (i.e. genomic regions harboring the causal genes) and explain the underlying mechanisms of complex traits. Noticing the the large sample size requirement and stringent significance threshold of GWAS, the objective of this study is to develop more powerful methods based on mediation modeling for identifying associations between genetic variants and complex traits with the assistance of omic phenotypes.

The outline of the paper is as follows. Section 2 presents the datasets used in this study briefly. Section 3 demonstrates the methodological backgrounds of the mediation models and introduces two methods aiming to detect associations between SNPs and BMI potentially missed by standard GWAS. Section 4 contains empirical applications of the methods and discusses their pros and cons. Section 5 concludes.

2 Material

In this research, two datasets are analyzed. The first data ,which are used as discovery dataset, include 1,669 unrelated individuals from TWINGENE, a sub-study within the Swedish Twin Registry . The concentration of 157 metabolites were measured and annotated from circulating serum samples using ultra-performance liquid chromatography coupled with tandem mass spectrometry (UPLC-MS/MS). All the subjects were also genotyped using Illumina OmniExpress microarray, where 638,331 SNPs are available in our data.

In an independent cohort, PIVUS from Uppsala, Sweden, the same UPLC-MS/MS approach was used to profile the metabolome in 916 unrelated 70-years-old individuals, who were also genotyped using the Illumina OmniExpress array and imputed to the same reference panel

(Lind et al., 2005). This PIVUS dataset serves as a replication data in analysis below. For both TWINGENE and PIVUS, measurements of the BMI and all metabolites are inverse-Gaussian transformed to be standard normally distributed.

Besides these two datasets, the GWAS results of a very recent European-ancestry BMI meta-analysis by the GIANT consortium (Locke et al., 2015b) are referred to. The study examined associations between BMI and about 2.8 million SNPs in up to 339,224 individuals. Regression results including the effect size, standard error, p-value and allele frequency of each tested SNP are available.

In this thesis, both BMI and metabolites are continuous variables. The value of a SNP represents its allelic dosage. Let f denote the MAF of a SNP. We then assume the allelic dosage follows a binomial distribution $B(2, f)$, known as the Hardy-Weinberg equilibrium (HWE) (Lynch et al., 1998; Falconer et al., 1996). The genetic effect is assumed to be additive, which means the genetic effect of 1 versus 0 is half as that of 2 versus 0. In this way, we are able to use all the observations for estimating a single parameter, i.e. the additive genetic effect, which leads to the best power.

3 Methodology

3.1 Concepts of mediation models

In statistics, a mediating variable refers to the intermediate variable relating an independent variable to a dependent variable in a causal sequence (MacKinnon, 2008). More specifically, if X causes M and M causes Y in a causal chain relating X and Y, then the relationship is called mediation and M is called a mediator. In other words, mediating relationships occur when a variable transmits the effect of an independent variable on a dependent variable.

In our case, we are interested in the model shown in Figure 1. Assume a metabolite lies on a mediation pathway between a SNP and BMI, then the total genetic effect (TGE) is dissected to two parts: the mediated genetic effect (MGE) mediated by the metabolite and the direct genetic effect (DGE), which refers to the sum of all effects on other pathways between the SNP and

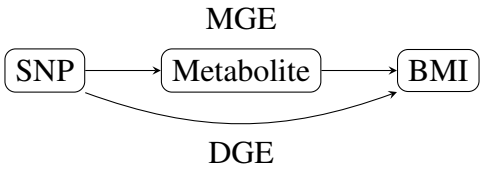


Figure 1: Example of a mediation model where a SNP can affect BMI mediated by a metabolite

BMI without the metabolite. Apart from TGE (= MGE + DGE), which GWAS focuses on, MGE is also informative for discovering mechanisms underlying SNPs and BMI. If multiple mediation pathways exist in parallel between the SNP and BMI, the signs of these parallel mediation effects may differ, so that these genetic effects cancel out each other, which may leads to a TGE close to zero. In this case, the existence of MGE indicates the association between a SNP and BMI but TGE does not. Hence, we are concerned with MGE instead of TGE and build a statistical mediation model in the form of Figure 1 to estimate MGE of the SNP on BMI mediated by the metabolite.

Although we assume the change in metabolite causes the change in BMI in the model above, the causality between a metabolite and BMI is always ambiguous in biology. Statistical tools such as mendelian randomization lack power when sample size is not large enough and the genetic effect size is tiny (Lawlor et al., 2008). As in most cases, metabolites are measured in a relatively small population, the statistical tools usually cannot help to determine the causality between metabolites and BMI. Therefore, we also come up with an alternative method using mediation model to select candidate SNPs instead of estimating their mediation effects.

3.2 Statistical mediation modeling

Since the number of SNPs in the datasets we deal with is huge, considering the computational time, for discovery purposes we propose a simple mediation modeling using linear regressions. The interaction term between the SNP and the metabolite is not included in the model due to lack of power to test it. Assume a complex trait $\mathbf{y}$ such as BMI is associated with a metabolite $\mathbf{m}$ and the allelic dosage of a SNP $\mathbf{s}$,

$$\mathbf{y} = \mathbf{1}\alpha_1 + \mathbf{m}\beta_1 + \mathbf{s}\beta_2 + \boldsymbol{\epsilon}_1 \tag{1}$$

and the metabolite is associated with the same SNP,

$$\mathbf{m} = \mathbf{1}\alpha_0 + \mathbf{s}\beta_0 + \boldsymbol{\epsilon}_0 \tag{2}$$

where $\boldsymbol{\epsilon}_0$ and $\boldsymbol{\epsilon}_1$ are i.i.d. normally distributed residuals, with variance σ_0^2 and σ_1^2, respectively. Both $\boldsymbol{\epsilon}_0$ and $\boldsymbol{\epsilon}_1$ are independent of $\mathbf{s}$. When covariates exist, we can replace the original $\mathbf{y}$ by the residuals of regression of $\mathbf{y}$ on the covariates, similar for $\mathbf{m}$. Therefore covariates such as age and sex are omitted in the equations for simplicity. The MGE in this case, which refers to the effect of the SNP on BMI mediated through the metabolite, is therefore $\beta_0\beta_1$. β_2 is the DGE. The TGE = MGE + DGE = $\beta_0\beta_1 + \beta_2$.

3.2.1 Test the omic-mediation effect

Despite that t-test should be used to test the significance of a regression coefficient, since t-statistic has an asymptotic standard normal distribution, we can use χ^2 test with one degree of freedom as a good substitute for t-test when sample size is large. Let $\hat{\beta}_m = \hat{\beta}_0\hat{\beta}_1$ be the estimate of MGE from the mediation model above. Although both $\hat{\beta}_0$ and $\hat{\beta}_1$ follow normal distribution, the distribution of the product $\hat{\beta}_m$ is symmetric but not normal under the null hypothesis $\beta_0\beta_1 = 0$. However, since the distribution of the product of two normal variables has shorter tail, meaning a smaller proportion of the population rests within its tail than would be under a normal distribution, the type I error of a test based on the product distribution will be less than a test based on normal distribution given the same rejection region. Therefore if the null hypothesis $\beta_m = 0$ is rejected using χ^2 test with one degree of freedom, it must be rejected using a test based on the product distribution. Since a conservative test does not generate more false positives, it is acceptable for a discovery step. So we can still use the χ^2 test, and the comparison of power can be presented clearly.

The variance of $\hat{\beta}_m$ can be obtained using $\hat{\beta}_0$, $\hat{\beta}_1$ and the corresponding standard errors. According to the delta method and noticing the independence between $\hat{\beta}_0$ and $\hat{\beta}_1$ (Bollen, 1987), we have

$$\text{Var}(\hat{\beta}_m) = \text{Var}(\hat{\beta}_0\hat{\beta}_1) \approx \hat{\beta}_0^2\text{Var}(\hat{\beta}_1) + \hat{\beta}_1^2\text{Var}(\hat{\beta}_0) \tag{3}$$

So the χ^2 statistic with one degree of freedom for $\hat{\beta}_m$ is

$$C = \frac{\hat{\beta}_m^2}{\text{Var}(\hat{\beta}_m)} = \frac{\hat{\beta}_0^2\hat{\beta}_1^2}{\beta_0^2\text{Var}(\hat{\beta}_1) + \beta_1^2\text{Var}(\hat{\beta}_0)} = \frac{C_0C_1}{C_0 + C_1}$$

where C_0 and C_1 are the chi-square statistics for $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. As the chi-square values are always positive, we have

$$C = \frac{C_0 C_1}{C_0 + C_1} = \frac{C_0}{C_0/C_1 + 1} < C_0 \tag{4}$$

and similarly $C < C_1$. Therefore, the significance of $\hat{\beta}_m$ is conditioned on that of both $\hat{\beta}_0$ and $\hat{\beta}_1$, which means both the SNP-metabolite and metabolite-BMI association have to be significant.

3.2.2 The power of the mediation analysis

Case 1: No direct genetic effect First, let us assume that $\beta_2 = 0$ and compare the power of the mediation test and that of genome-wide association analysis. Given the underlying mediation model (1) and (2), in an ordinary GWA analysis, TGE (= MGE when DGE = 0) is tested via the regression model

$$\mathbf{y} = \mathbf{1}(\alpha_0\beta_1 + \alpha_1) + \mathbf{s}\beta_0\beta_1 + (\boldsymbol{\epsilon}_0\beta_1 + \boldsymbol{\epsilon}_1) \tag{5}$$

Such a linear regression yields $\hat{\beta}_{\text{GWA}}$ that is also an unbiased estimate of $\beta_m = \beta_0\beta_1$, but its standard error is considerably bigger than that of $\hat{\beta}_m$. The allelic dosage $\mathbf{s}$ of a SNP with MAF f has a mean of $2f$ and variance $2f(1-f)$. Asymptotically, as the number of individuals n is sufficiently large, we have

$$\text{Var}(\hat{\beta}_{\text{GWA}}) = \frac{\text{Var}(\epsilon_0\beta_1 + \epsilon_1)}{\sum_{i=1}^{n}(s_i - \bar{s})^2} \approx \frac{\sigma_0^2\beta_1^2 + \sigma_1^2}{n\text{Var}(s_i)} = \frac{\sigma_0^2\beta_1^2}{2nf(1-f)} + \frac{\sigma_1^2}{2nf(1-f)}. \tag{6}$$

Similarly, since $\text{Var}(m_i) = \beta_0^2\text{Var}(s_i) + \text{Var}(\epsilon_0) = 2\beta_0^2 f(1-f) + \sigma_0^2$,

$$\text{Var}(\hat{\beta}_0) \approx \frac{\sigma_0^2}{n\text{Var}(s_i)} = \frac{\sigma_0^2}{2nf(1-f)}$$

and

$$\text{Var}(\hat{\beta}_1) \approx \frac{\sigma_1^2}{n\text{Var}(m_i)} = \frac{\sigma_1^2}{n(2\beta_0^2 f(1-f) + \sigma_0^2)}.$$

Therefore according to (3),

$$\begin{aligned}
\text{Var}(\hat{\beta}_m) &\approx \beta_1^2\text{Var}(\hat{\beta}_0) + \beta_0^2\text{Var}(\hat{\beta}_1) \\
&= \frac{\sigma_0^2\beta_1^2}{2nf(1-f)} + \frac{\sigma_1^2\beta_0^2}{n(2\beta_0^2 f(1-f) + \sigma_0^2)} \\
&= \frac{\sigma_0^2\beta_1^2}{2nf(1-f)} + \frac{\sigma_1^2}{2nf(1-f) + n\sigma_0^2/\beta_0^2}.
\end{aligned} \tag{7}$$

Comparing (6) and (7), it is obvious that $\text{Var}(\hat{\beta}_{\text{GWA}}) > \text{Var}(\hat{\beta}_m)$ due to $n\sigma_0^2/\beta_0^2 > 0$. So that the ordinary GWA has less power than mediation model. The power boost via the mediator can be expressed as the proportion of χ^2 statistic increased compared to ordinary GWA, i.e.

$$\frac{\text{Var}(\hat{\beta}_{\text{GWA}}) - \text{Var}(\hat{\beta}_m)}{\text{Var}(\hat{\beta}_m)} = \frac{\sigma_0^2\sigma_1^2}{2f(1-f)\beta_0^2(\sigma_1^2 + \sigma_0^2\beta_1^2) + \sigma_0^4\beta_1^2} \tag{8}$$

From (8), the reduction of the standard error of $\hat{\beta}$ by the mediation analysis increases when f decreases, thus the power boost is larger for rare variants. Additionally, the smaller β_0 and β_1 are, the larger the power boost is.

In particular, if the complex trait and the metabolite are both transformed to be standard normally distributed, we have $2f(1-f)\beta_0^2 + \sigma_0^2 = 1$ and $\beta_1^2 + \sigma_1^2 = 1$. Consequently, (8) can be simplified as

$$\frac{\text{Var}(\hat{\beta}_{\text{GWA}}) - \text{Var}(\hat{\beta}_m)}{\text{Var}(\hat{\beta}_m)} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_1^2} - 2}$$

where $0 < \sigma_0^2, \sigma_1^2 < 1$. In this case, our conclusions above still hold.

Case 2: Direct genetic effect exists To compare the discovery power of standard GWA analysis and that of our omic-mediation analysis (OMA) given $\beta_2 \neq 0$, simulations are drawn. The simulation results are summarized in Figure 2. A causal SNP with MAF of 0.30 is simulated for a population with 1,000 individuals. The SNP has a narrow sense heritability of 1% (i.e. the proportion of variance explained is 1%) on the omic mediator, and the mediator explains 5% of the variance of the complex trait. The simulated MGE is set to be 1. The bottom horizontal axis represents the simulated DGE on the trait. The top horizontal axis indicates the corresponding total narrow sense heritability (h^2) of the SNP on the trait. The vertical line marks the scenario that the TGE is zero, i.e. DGE = - MGE. Each point was calculated based on 10,000 simulations.

As shown in Figure 2, the GWA analysis may have better power if and only if DGE is sufficiently large compared to MGE. However, the DGE can be regarded as the sum of the mediated effects of the same SNP in all the other pathways except the one through the analyzed mediator; as these effects may have different directions and consequently cancel out each other, it is unlikely that the DGE can be large enough to generate a large TGE. In fact, we already learned from large GWAS that most functional loci have rather small TGE thus are very challenging to

identify using standard GWAS. According to the discussion above, we have shown, analytically and empirically, it is reasonable to expect that OMA is more powerful than the standard GWA when a good or complete mediator is available.

3.2.3 Procedure of mediation analysis

In our analysis, not all the combinations of SNPs and metabolites should be modeled and tested using the mediation analysis. Although more powerful than the direct test between each SNP and BMI, as we have shown in (4), the significance of the MGE has to be conditioned on the significance of both the SNP-metabolite and metabolite-BMI associations. Therefore, there is a requirement of the mediator to be associated with both the SNP and BMI. The steps we performed in our study are:

Step 1 Regress BMI on each metabolites, select the top five BMI-associated metabolites.

Step 2 For each of the selected metabolites, perform GWAS to find the SNP most correlated to the metabolite.

Step 3 Build the mediation model and estimate MGE, its variance and p-value by the method mentioned above for each of the five SNP - metabolite combinations.

It is worth noting that the MGE in our model depends on the association between SNP and metabolite and that between metabolite and BMI. Since GWAS is performed for each of the five selected metabolites, we actually run $5 \times 638,331$ tests for associations between SNPs and metabolites. For the associations between metabolites and BMI, 157 tests are performed. Hence the total number of tests is $5 \times 638,331 + 157 \approx 5 \times 1,000,000$. In accordance with a genome-wide significance threshold of 5×10^{-8} for a single GWA scan, the correspondent 5% Bonferroni-corrected significance threshold should be 10^{-8} instead of $0.05/5 = 0.01$ even though only five MGE are tested here.

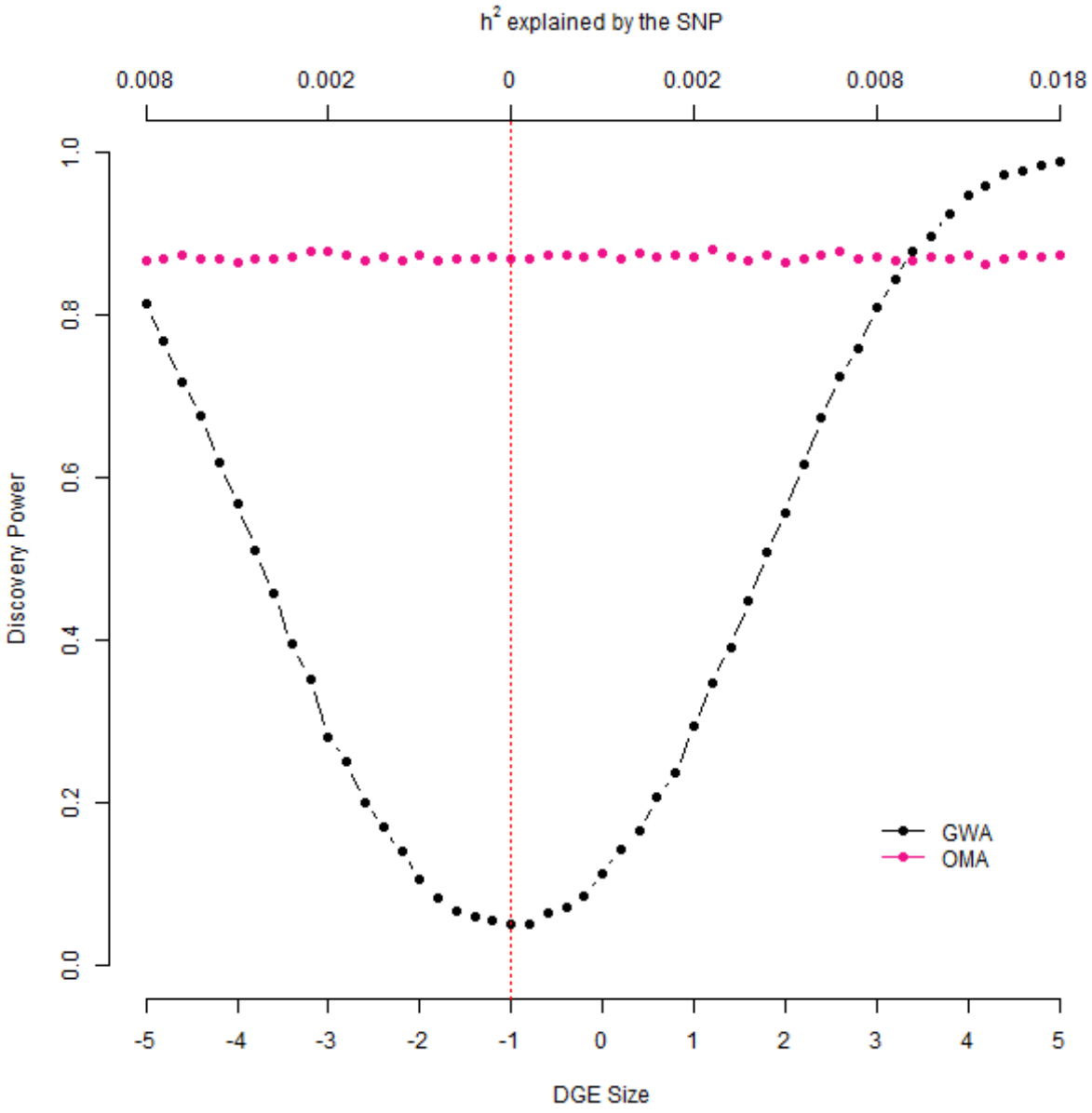
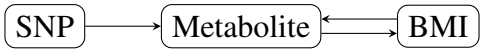


Figure 2: Comparison of the discovery power of standard GWA analysis and that of omic-mediation analysis (OMA).



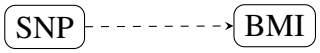


Figure 3: Steps to select candidate SNP. Solid arrow means association can be verified by regression or biology literature. Dashed arrow means association remaining to check in a GWAS result.

3.3 Selecting candidate SNPs using mediation model

When a metabolite is correlated to BMI without known biological causality, the metabolite can still be used as an indicator for selecting candidate SNPs. By decreasing the number of SNPs to be tested, the significance threshold becomes less stringent. If n independent SNPs are set as candidates, then the significance threshold turns to be $0.05/n$. In this research, we pick one SNP as candidate for each metabolite indicator to maximally lower the significance threshold. The procedure is stated below.

Step 1 According to the biology literature, choose a metabolite correlated to BMI. Test their correlation using data.

Step 2 Perform GWAS on the metabolite to find the SNP most correlated to it.

Step 3 Test the association between the SNP and BMI in a GWAS dataset.

The steps are presented in Figure 3. Unlike the first method, the association between SNP and metabolite is merely suggestive for that between SNP and BMI, which refers to TGE in this case. By doing this, since only one test is performed to verify the association between SNP and BMI, a GWAS level of significance is not necessary. Consequently some genetic effects on BMI, which are neglected because of mid-range error and stringent significance threshold, might be discovered.

Table 1: Top five BMI-associated metabolites

Metabolite	Discovery Data: TwinGene		Replication Data: PIVUS	
	Effect size(s.e.)	P	Effect size(s.e.)	P
Uric acid	$0.306(0.025)$	1.02×10^{-32}	$0.290(0.034)$	1.01×10^{-16}
L-Tyrosine	$0.259(0.024)$	2.04×10^{-25}	$0.182(0.033)$	5.88×10^{-08}
Glycerophospholipids PC(33:1)	$-0.212(0.025)$	1.99×10^{-17}	$-0.234(0.033)$	1.60×10^{-12}
L-proline	$0.208(0.025)$	5.29×10^{-16}	$0.169(0.034)$	8.21×10^{-07}
Arachidonic acid ethyl ester	$0.183(0.025)$	1.28×10^{-13}	$0.142(0.033)$	2.61×10^{-05}

4 Results and Discussion

4.1 Statistical mediation modeling

We build SNP - Metabolite - BMI mediation models for different combinations of SNPs and metabolites. Age and sex are included as covariates in all subsequent analyses. As we have shown in the Methodology section, the p-value of the mediation effect is always larger than that of the SNP effect on the metabolite and that of the metabolite effect on BMI. Hence we perform our step-by-step analysis following the procedure in Section 3.2.3.

4.1.1 Prioritizing candidate SNP - Metabolite associations

We regress BMI on each of the 157 metabolites and rank the metabolites according to their level of association (p-values) with BMI in TWINGENE. The top five BMI-associated metabolites are presented in Table 1. Then for each of the five chosen metabolites, the SNP most correlated to the metabolite is detected by GWAS (Table 2) using the R package GenABEL (Aulchenko et al., 2007).

4.1.2 Applying the omic-mediation analysis

After the five SNP-metabolite combinations are chosen, we can apply omic-mediation analysis to them. For the "two-stage" association of SNP - metabolite - BMI, MGE is calculated by

Table 2: The most correlated SNP for each metabolites

Metabolite	SNP	Discovery Data: TwinGene		Replication Data: PIVUS	
		Effect size(s.e.)	P	Effect size(s.e.)	P
Uric acid	rs1014290	$-0.274(0.041)$	2.40×10^{-11}	$-0.241(0.051)$	2.83×10^{-06}
L-Tyrosine	rs10431051	$-0.220(0.055)$	6.40×10^{-05}	$0.071(0.071)$	3.21×10^{-01}
Glycerophospholipids PC(33:1)	rs341093	$-0.546(0.127)$	1.86×10^{-05}	$0.064(0.135)$	6.35×10^{-01}
L-proline	rs2005883	$0.424(0.046)$	1.15×10^{-19}	$0.446(0.064)$	5.62×10^{-12}
Arachidonic acid ethyl ester	rs174538	$-0.229(0.038)$	1.65×10^{-09}	$-0.277(0.048)$	9.62×10^{-09}

Table 3: The omic-mediation analysis results for the five SNP-metabolite combinations

Metabolite	SNP	Discovery Data: TwinGene		Replication Data: PIVUS	
		MGE(s.e.)	P	MGE(s.e.)	P
Uric acid	rs1014290	$-0.086(0.014)$	3.66×10^{-09}	$-0.073(0.018)$	3.36×10^{-05}
L-Tyrosine	rs10431051	$-0.058(0.015)$	1.75×10^{-04}	$0.013(0.013)$	3.29×10^{-01}
Glycerophospholipids PC(33:1)	rs341093	$0.117(0.030)$	1.22×10^{-04}	$-0.015(0.032)$	6.35×10^{-01}
L-proline	rs2005883	$0.090(0.015)$	1.14×10^{-09}	$0.079(0.019)$	4.03×10^{-05}
Arachidonic acid ethyl ester	rs174538	$-0.043(0.009)$	1.99×10^{-06}	$-0.042(0.012)$	3.93×10^{-04}

multiplying the SNP effect on the metabolite and the effect of the metabolite on BMI. The corresponding standard error is calculated based on the delta method (see Methodology part). According to the results presented in Table 3, rs1014290 and rs2005883, whose p-values of MGE reach the corrected significance 10^{-8}, are significantly associated with BMI mediated by uric acid and L-proline separately. In the replication dataset PIVUS, for each of the two pathways, the MGE is similar to that in discovery data in terms of size and direction. Their p-values also reach the replication significance threshold of 0.05.

4.1.3 Discussion

Neither rs1014290 nor rs2005883 has been found to be associated with BMI in large meta-analyses of GWAS. The p-values of their TGEs from the meta-analysis by the GIANT consortium are 0.721 and 0.344, separately. As demonstrated in the Methodology section, the statistical power of testing the mediation effect is higher than that of testing the association between the SNP and BMI directly when DGE of the SNP on BMI is around zero or the DGE and MGE cancel out each other. For this reason, some SNPs which are undetectable in much larger GWA studies can be detected by OMA. We provide a straightforward workflow to bridge genomics and complex traits using omics measurements, which is essential to genetic studies in the omics era.

However, there are two assumptions we make might be violated when the omic-mediation analysis is performed. First, in order to build the mediation model, we assume the pathway is from a SNP to a metabolite to BMI. But the true causality between the metabolite and BMI might be mutual or even reversed as in Figure 4. In fact, if BMI affects the metabolite instead of that the metabolite regulates BMI, $\hat{\beta}_m$ and its p-value from our method do not infer the MGE at all. In our case, we report the association between rs1014290 and BMI mediated by uric acid (UA). Although the correlation between UA and BMI has been mentioned in Yue et al. (2012) and Ryu et al. (2012), the causality might be mutual (Sajja Srikanth and Sushma) or even reverse (Ishizaka et al., 2010). Therefore the association between rs1014290 and BMI is questionable in the situation that the change in BMI causes the change in UA.

Given the pathway is true, the residuals $\boldsymbol{\epsilon}_1$ in (1) and $\boldsymbol{\epsilon}_0$ in (2) are assumed to be independent,

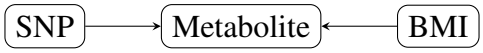


Figure 4: Example of the mediation model failure due to the reverse causality between metabolite and BMI.

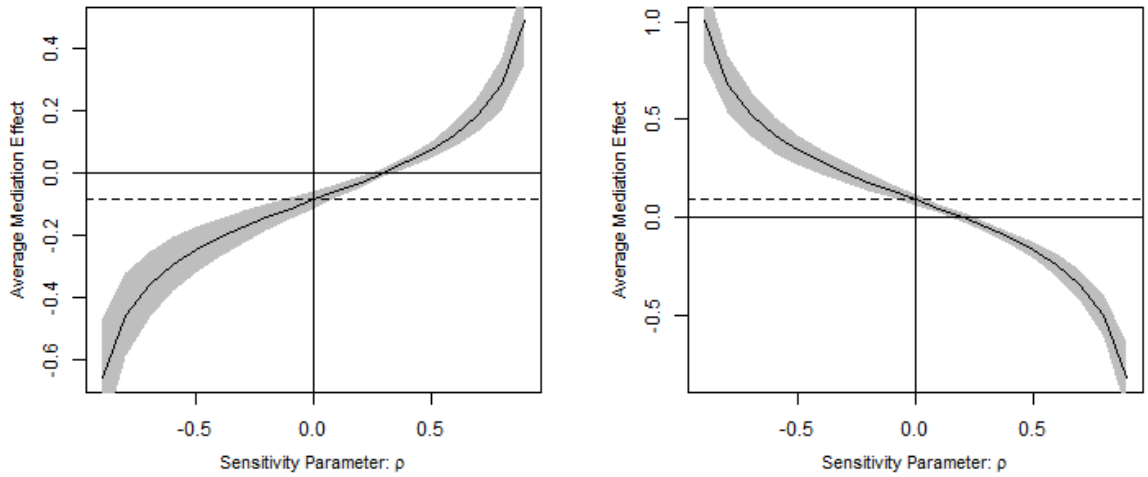


Figure 5: MGE of rs1014290 (left) and rs2005883 (right) as a function of degree of violation of no covariate assumption. Grey region is confidence interval.

which means there is no covariate except for sex and age affecting both the metabolite and BMI. This assumption is too strict to be met since a metabolite is almost surely affected by several SNPs and metabolites affecting BMI. When the assumption is violated, the least squares estimators are not unbiased even when sample size is large. Although this assumption is difficult to test from the data, a sensitivity analysis can be used. By expressing the MGE as a function of ρ, i.e. the correlation between ϵ_{i1} and ϵ_{i2}, a sensitivity analysis investigates how robust the results are to the violation of the no-covariate assumption. The results of sensitivity analysis for the two mediation models we built are shown in Figure 5. We can see that the MGE is around zero when ρ equals to 0.3 for rs1014290 and 0.2 for rs2005883. In this regard, the association between rs1014290 and BMI, where a more severe violation is allowed, is more robust than that between rs2005883 and BMI.

4.2 Selecting candidate SNPs using the mediation model

In this section, we regard the metabolite as an indicator for selecting candidate SNPs. As we have stated in Section 2.3, correlation rather than causality between a metabolite and BMI is sufficient to set the metabolite as an indicator and perform this method for discovery. According to Banfi and Del Fabbro (2006), there is a correlation between creatine and BMI in elite athletes. The correlation can be verified by regressing BMI on creatine (Table 4). It is noteworthy that the regression is just used for verifying rather than predicting. Then by regressing creatine on each SNP, rs4673546 is found as the most creatine-associated SNP. The association between rs4673546 and creatine is replicated in PIVUS. Consequently, according to the GWAS results of the meta-analysis by the GIANT consortium Locke et al. (2015a), the effect size of rs4673546 on BMI is -0.0141 (s.e. = 0.0047), with a p-value = 0.0027. As we have discussed in Section 2.3, since multiplicity is avoided, the significance threshold is 0.05 here. Hence the analysis concludes that rs4673546 is significantly associated with BMI.

4.2.1 Discussion

Generally speaking, this method is more biology-based due to the metabolite used as indicator is selected according to biology literature instead of statistical method. Additionally, since only correlation is required, this method can be performed more widely.

The method assumes SNPs are independent to each other. However, if n SNPs ($n > 1$) are chosen as candidates for each metabolite, there is a chance that some of them are correlated to each other. In such a case, the n tests concerning the associations between SNPs and BMI are correlated, which means the significance threshold $0.05/n$ is conservative. Nevertheless, a conservative significance threshold is acceptable for a discovery research as we have mentioned before.

Table 4: The correlations between creatine and BMI (the first line) and between rs4673546 and creatine (the second line)

Correlation Checked	Discovery Data: TwinGene		Replication Data: PIVUS	
	Coefficient(s.e.)	P	Coefficient(s.e.)	P
creatine - BMI	$0.110(0.028)$	0.0001	$0.119(0.038)$	0.0016
rs4673546 - creatine	$0.287(0.038)$	5.27×10^{-14}	$0.174(0.051)$	0.0007

4.3 Further research

In our research, OMA is based on a simple mediation model which might be worth generalizing in many aspects. For example, the causality between metabolite and BMI is assumed to be one way, while mutual effect is more common and plausible for the relationship between metabolite and BMI. Therefore the estimation and testing concerning a mediation model with a feedback loop between mediator and outcome will be practical. In addition, multiple levels of mediator deserves more attention.

5 Conclusion

As we have shown, if there is an underlying mediation pathway between a genetic variant and a complex trait, comparing to OMA, direct GWA analysis has less power and the power difference even scales up with the effect of the mediator on the trait. By the analysis of MGE with a correct mediator measured, OMA is able to discover the associations mediated by metabolites between SNPs and BMI, which can hardly be detected via GWAS even with large sample size. When the causality between a metabolite and BMI is unclear, omics data can also be helpful with respect to selecting SNPs for further association tests so that the stringent standard GWAS significance level is not required. In conclusion, omics data play an important role in detecting the missing genetic architectures underlying complex traits. We therefore emphasize that more efforts should be dedicated into measuring and analyzing intermediate omic phenotypes in genetic studies.

Acknowledgements

Foremost, I greatly appreciate my supervisor Dr. Xia Shen for his excellent guidance, constructive suggestions and great patience. Besides, I would like to express my deepest gratitude to Prof. Yudi Pawitan for giving fruitful comments and offering me the opportunity to do thesis in his team where I fell in love with genetics. My sincere appreciation also goes to Prof. Fan Yang Wallentin for her support of my thesis and my life as an international student. Thanks to my friend Dr. Shaobo Jin who gave me lots of help in terms of studying and living. Thanks to the Swedish Twin Registry and PIVUS study for the datasets used in this thesis. I also thank Prof. Erik Ingelsson for providing access to these two datasets for genotype and BMI data.

References

Mark D Adams, Jenny M Kelley, Jeannine D Gocayne, Mark Dubnick, Mihael H Polymeropoulos, Hong Xiao, Carl R Merril, Andrew Wu, Bjorn Olde, Ruben F Moreno, et al. Complementary dna sequencing: expressed sequence tags and human genome project. *Science*, 252 (5013):1651–1656, 1991.

Yurii S Aulchenko, Stephan Ripke, Aaron Isaacs, and Cornelia M Van Duijn. Genabel: an r library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–1296, 2007.

Giuseppe Banfi and Massimo Del Fabbro. Relation between serum creatinine and body mass index in elite athletes of different sport disciplines. *British journal of sports medicine*, 40(8): 675–678, 2006.

Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

Kenneth A Bollen. Total, direct, and indirect effects in structural equation models. *Sociological methodology*, 17(1):37–69, 1987.

Anthony J Brookes. The essence of snps. *Gene*, 234(2):177–186, 1999.

Douglas S Falconer, Trudy FC Mackay, and Richard Frankham. *Introduction to quantitative genetics (4th edn)*, volume 12. [Amsterdam, The Netherlands: Elsevier Science Publishers (Biomedical Division)], c1985-, 1996.

David S Freedman, Laura Kettel Khan, William H Dietz, Sathanur R Srinivasan, and Gerald S Berenson. Relationship of childhood obesity to coronary heart disease risk factors in adulthood: the bogalusa heart study. *Pediatrics*, 108(3):712–718, 2001.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.

Nobukazu Ishizaka, Yuko Ishizaka, Akiko Toda, Mizuki Tani, Kazuhiko Koike, Minoru Yamakado, and Ryozo Nagai. Changes in waist circumference and body mass index in relation to changes in serum uric acid in japanese individuals. *The Journal of rheumatology*, 37(2): 410–416, 2010.

T Kawada. Body mass index is a good predictor of hypertension and hyperlipidemia in a rural japanese population. *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity*, 26(5):725–729, 2002.

Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

Lars Lind, Nilla Fors, Jan Hall, Kerstin Marttala, and Anna Stenborg. A comparison of three different methods to evaluate endothelium-dependent vasodilation in the elderly the prospective investigation of the vasculature in uppsala seniors (pivus) study. *Arteriosclerosis, thrombosis, and vascular biology*, 25(11):2368–2375, 2005.

Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015a.

Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015b.

Michael Lynch, Bruce Walsh, et al. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, 1998.

David Peter MacKinnon. *Introduction to statistical mediation analysis*. Routledge, 2008.

Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, 2015.

James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155, 1992.

Hye Sook Ryu, Hyunjin Park, and Mieun Yun. Correlation between serum uric acid, bmi, fasting blood sugar, tg and hdl in korean health check examinees. *The FASEB Journal*, 26 (1_MeetingAbstracts):645–4, 2012.

Y Sajja Srikanth and K Sushma. Relationship between serum uric acid and bmi in pre diabetics and type ii diabetics in rural population–a pilot study.

Tyler VanderWeele and Stijn Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468, 2009. ISSN 1938-7989.

Ji-Rong Yue, Chang-Quan Huang, and Bi-Rong Dong. Association of serum uric acid with body mass index among long-lived chinese. *Experimental gerontology*, 47(8):595–600, 2012.