이학석사학위논문

HisCoM-PCA: Hierarchical structural Component Model for Pathway analysis of Common vAriants

계층적 구조 모형을 이용한 common variants 의 패스웨이 분석

2019 년 8 월

서울대학교 대학원

협동과정 생물정보학과

Nan Jiang

HisCoM-PCA: Hierarchical structural Component Model for Pathway analysis of Common vAriants

by

Nan Jiang

A thesis

submitted in fulfillment of the requirement for the degree of Master

in

Bioinformatics

Interdisciplinary Program in Bioinformatics College of Natural Sciences

Seoul National University

August, 2019

Abstract

HisCoM-PCA: Hierarchical structural Component Model for Pathway analysis of Common vAriants

Nan Jiang
Interdisciplinary Program in Bioinformatics
The Graduate School
Seoul National University

Genome-wide association studies (GWAS) have been widely used in identifying phenotype-related genetic variants by many statistical methods, such as logistic regression and linear regression. However, the identified SNPs with stringent statistical significance just explain a small portion of the overall estimated genetic heritability. To address this 'missing heritability' issue, gene-based and pathway-based analysis have been developed in many studies. The biological mechanisms and some related pathways have been reported using pathway-based methods in GWAS datasets. However, many of these methods often neglecting the correlation between genes and between pathways. Here, we construct a hierarchical component model with considering of the correlation existing both between genes and between pathways. Based on this model, we propose a novel pathway analysis method for GWAS datasets, named Hierarchical structural Component Model for Pathway analysis of Common vAriants (HisCoM-PCA). HisCoM-PCA first summaries the common variants in each gene into the gene-level statistics and then analyzes all pathways simultaneously by ridge-type penalization on both gene and pathway effects on the phenotype. The statistical significance of the gene and pathway coefficients can be examined by permutation tests. Through simulation study for both binary and continuous phenotypes using GAW17 simulation dataset, HisCoM-PCA controlled type I error well and showed a higher empirical power than several comparison methods. In addition, we applied our method to SNP chip dataset of KARE for four human physiologic traits: (1) type 2 diabetes; (2) hypertension; (3) systolic blood pressure; and (4) diastolic blood pressure. Those results showed that HisCoM-PCA could successfully identify signal pathways with superior statistical and biological significance. Our approach has an advantage of providing an intuitive biological interpretation for the association between common variants and phenotypes through the pathway information.

Keywords: common variants, Genome-wide association study, hierarchical components, pathway analysis.

Student Number: 2017-20507

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Genome−wide association studies (GWAS) have made great achievement for investigating the association between a set of genetic variants and a trait of interest. GWAS typically focus on associations between single−nucleotide polymorphisms (SNPs) and traits such as type 2 diabetes (T2D) [1]. To identify the common variants in GWAS, many statistical methods such as logistic regression and linear regression have been widely used. Since most of these methods are based on the single variant analysis, the statistically significant results sometimes may suffer from a lack of biological interpretation. In addition, it has been reported that only a small portion of the total heritability of traits can be explained by these identified SNPs [2]. To enhance the interpretation of the results from SNPs, many gene−based and pathway−based association analysis methods have been developed. Biological pathways, which have complex interaction with each other, always have more direct influence on the related biological behaviors rather than genes [3]. Thus, it is easier to interpret the pathway−based results than SNP−based results. The pathway−based association methods developed for GWAS often identify pathways based on results from the single analysis of SNPs. These methods often use only top SNPs according to the p-values obtained from single SNP analysis. However, such analysis process ignores genetic information in the SNPs which are not selected [4-6]. In addition, the high correlations always exist between pathways, potentially caused by many shared genes between pathways. The methods neglecting these correlations may mislead the association results [7].

Considering these deficiencies, a hierarchical component model has been constructed, which is named as PHARAOH (Pathway-based approach using HierArchical components of collapsed RAre variants Of High-throughput sequencing data). PHARAOH performs pathway analysis for rare variants using a single hierarchical model. PHARAOH includes a collapsing step for rare variants, since the rare variants data are usually sparse. The gene-level summary statistics are obtained by the special weight approach for rare variants. It analyzes entire genes and pathways by adding ridge-type penalties on both gene and pathway effects to traits [8]. PHARAOH is usually used to perform analysis for rare variants rather than common variants since the special collapsing step. Common variants data usually need dimension reduction instead of collapsing.

In this study, we utilized the main idea of PHARAOH and principal component analysis (PCA) to construct a hierarchical component model for common variants. Based on this model, we proposed a novel pathway analysis method for GWAS datasets, named Hierarchical structural Component Model for Pathway analysis of Common vAriants (HisCoM-PCA).

HisCoM-PCA has several distinctive features. First, HisCoM-PCA can identify associations between a trait and entire pathways using a single model. It can simultaneously quantify both the effects of pathways and genes to the phenotype. Second, HisCoM-PCA performs the pathway analysis using the gene-level summary statistics from SNPs within the same genes. Third, HisCoM-PCA allows the potential correlations between genes and between pathways by adding the ridge-type penalties on both genes and pathways effects. In addition, HisCoM-PCA is not only used for binary phenotypes but also continuous phenotypes. Overall, HisCoM-PCA can identify associated genes and pathways under controlling correlation between within them.

In this study, we applied HisCoM-PCA for binary phenotypes, type 2 diabetes (T2D) and hypertension (HT), and continuous phenotypes, systolic blood pressure (SBP) and diastolic blood pressure (DBP), using large-scale SNP data from a Korean population study (8,840 samples) [9] and KEGG pathway database (186 pathways) [10]. Furthermore, HisCoM-PCA was compared to three existing pathway-based approaches: GSA-SNP2 [4], sARTP [11], MAGMA [12]. To check the power and type I error of HisCoM-PCA, the simulation study was performed with Genetic Analysis Workshop (GAW) 17 generated dataset [13]. The empirical power of HisCoM-PCA was compared with other three existing methods.

The results of both simulation study and real data analysis demonstrated that HisCoM-pGWAS successfully identified the statistically associated and biologically plausible pathways for traits of interest.

Chapter 2

Materials

2.1. Real data

2.1.1. KARE cohort dataset

The KARE project is a large-scale cohort study for Korea population. It recruited almost 10,000 participants from Ansan and Ansung, which represent city population and countryside population [9]. More than hundreds of papers have been completed using this cohort data for genetic analysis study. The common variant genotype data of 8840 individuals were produced with the Affymetrix Genome-Wide Human SNP array 5.0. This chip consists of about 50 million autosomal SNPs and total 352,228 SNPs are available. In this study, we excluded SNPs for which the minor allele frequencies (MAF) were less than 0.05, the genotype calling rates were less than 95%, and Hardy-Weinberg equilibrium p-values were less than 10^{-6}. We only kept the subjects with gender consistencies and those whose calling rates were more than 90%. After such quality control process, missing values were imputed only for existing variants.

2.1.2. Definition of Type 2 diabetes

An individual is defined as T2D patient according to the following criteria: (1) under treatment for T2D, (2) fasting plasma glucose(FPG) ≥126 mg/Dl, 2-hour postprandial blood glucose (Glu120) ≥ 200mg/dL or glycated hemoglobin (HbA1c) ≥ 6.5%, and (3) age of disease onset ≥40 years. A total of 1,288 subjects are diagnosed as T2D patients among 8840 individuals. There are 3,687 individuals selected as normal subjects according to the inclusion criteria: (1) FPG <100 mg/dL, Glu120 <140 mg/dL and HbA1c < 5.7% (2) no history of diabetes [14]. Demographic variables of 4,975 selected subjects are summarized in Table1.

Table 1. Demographic variables for KARE cohort(T2D)

	T2D subjects	Normal subjects
Area(Ansan/Ansung)	673/615	1,607/2,080
Gender(Male/Female)	671/617	1,679/2,008
Age(Mean±SD)	55.92(±8.80)	49.88(±8.31)
BMI(Mean±SD)	25.54(±3.27)	24.10(±2.90)
Number of subjects	1,288	3,687

SD: standard deviation; BMI: body mass index

2.1.3. Definition of hypertension

A total of 2,008 individuals are defined as hypertensive cases according to the following criteria: (1) SBP ≥ 140 mm Hg and/or DBP ≥ 90 mm Hg, and (2) treatment with antihypertension medication. There are 4,569 individuals defined as normotensive controls according to the criteria: SBP < 120 mm Hg and DBP < 80 mm Hg. The subjects with pre-hypertensive status were excluded from the analysis. For quantitative trait analysis of SBP and DBP, 1,019 subjects are excluded due to hypertensive therapy or drug treatments, which are likely to influence blood pressure [15]. The basic characteristics and blood pressure of the subjects are listed in Table 2.

Table 2. Basic characteristics of study subjects (blood pressure)

(a) Basic characteristics of hypertensive cases and normotensive controls

	HT subjects	Normal subjects
Area(Ansan/Ansung)	1,204/804	1,756/2,813
Gender(Male/Female)	916/1,092	2,065/2,504
Age(Mean±SD)	56.74(±8.42)	49.43(±8.09)
BMI(Mean±SD)	25.62(±3.27)	24.03(±2.94)
Number of subjects	2,008	4,569

SD: standard deviation; BMI: body mass index;

(b) Basic characteristics of subjects for blood pressure analysis

	Subjects
Area(Ansan/Ansung)	3,589/4,222
Gender(Male/Female)	3,784/4,027
Age(Mean±SD)	51.45(±8.75)
BMI(Mean±SD)	24.40(±3.07)
SBP(Mean±SD)	115.54(±17.20)
DBP(Mean±SD)	74.10(±11.23)
Number of subjects	7,811

SD: standard deviation; BMI: body mass index;
SBP: systolic blood pressure; DBP: diastolic blood pressure

2.2. Simulation data

To check the power and type I error rate of HisCoM-PCA, a simulation study was performed using simulation data from the Genetic Analysis Workshop 17 (GAW17) [13]. In brief, a GAW17 simulation dataset was generated for 697 individuals from the 1000 Genomes Project [16], containing 24,487 SNVs and four phenotypes (Q1, Q2, Q4, and AFFECTED). Among the four simulated phenotypes, only Q1 was generated using pathway information, and was simulated to be affected by 9 genes from the vascular endothelial growth factor (VEGF) pathway, as defined by Ingenuity Pathway Analysis [17]. We next examined the power according to the proportion of identifying the VEGF pathway from the entirety of pathways in the KEGG database. Type I error of HisCoM-PCA was examined by the proportion of identifying null pathways which did not contain causal genes. Both type I error and power were calculated by analysis for Q1. To compare the power with other existing methods, we also analyzed the GAW17 dataset using sARTP, a self-contained version of MAGMA, a competitive version of MAGMA and GSA-SNP2.

Chapter 3

Methodology

3.1. HisCoM-PCA

3.1.1 Step1: SNPs dimension reduction by Principal Component Analysis (PCA)

At the first step, reduce dimension of the common variants located from the same genes by PCA. After performing PCA for each gene, choose a part of the PCs as gene-level summary statistics. PCs can be selected simply by number of the PCs for each gene. Different threshold of cumulative proportion of variances can also be defined to select the PCs for the corresponding genes. In this study, we used the following guidelines to choose the number of PCs in our analysis: (1) the first PC, (2) PCs whose cumulative proportion of variances are more than 30%. In this study, we use a R's function prcomp in the stats package to conduct PCA.

3.1.2. Step2: Pathway analysis with a hierarchical component model(HisCoM)

After reducing dimension of common variants for each gene, perform pathway analysis using the selected PCs with a hierarchical component model. The model has been used for pathway analysis of rare variants. Before statistical analysis, map the selected PCs with pathways using some well-known pathway databases, such as KEGG. The total scheme of HisCoM-PCA is showed in Figure 1.

In this model, pathways are defined as a weighted component of a set of PCs as showed in Figure 1. Let us define y_j as the phenotype of the j^{th} subject and assume that phenotype independently follow an exponential family distribution $(j = 1, \dots, N)$. Let K be the number of pathways, T_k be the number of genes in the k^{th} pathway and N_{kt} be the number of PCs for the t^{th} gene in the k^{th} pathway. Let g_{itk} denote the i^{th} PC of the t^{th} gene in k^{th} pathway $(k = 1, \dots, K; t = 1, \dots, T_k; i = 1, \dots, N_{tk})$. Let w_{itk} denote a weight assigned to g_{itk} and β_k denote the coefficient connecting the k^{th} pathway to the phenotype. For each individual, the relationships between PCs and binary phenotype are established in such a way that:

$$logit(\pi) = \beta_0 + \sum_{k=1}^{K} \left[\sum_{t=1}^{T_k} \sum_{i=1}^{N_{kt}} g_{itk} w_{itk} \right] \beta_k$$

The alternating regulated least-squares (ALS) algorithm is used to estimate model parameters in such component-based approach. According to the ALS algorithm, two steps are alternated until convergence [18].

$\boldsymbol{Step2-1}$: For fixed the weight coefficient estimates w_{kti}, update the pathway coefficient estimates β_k in the least−squares.

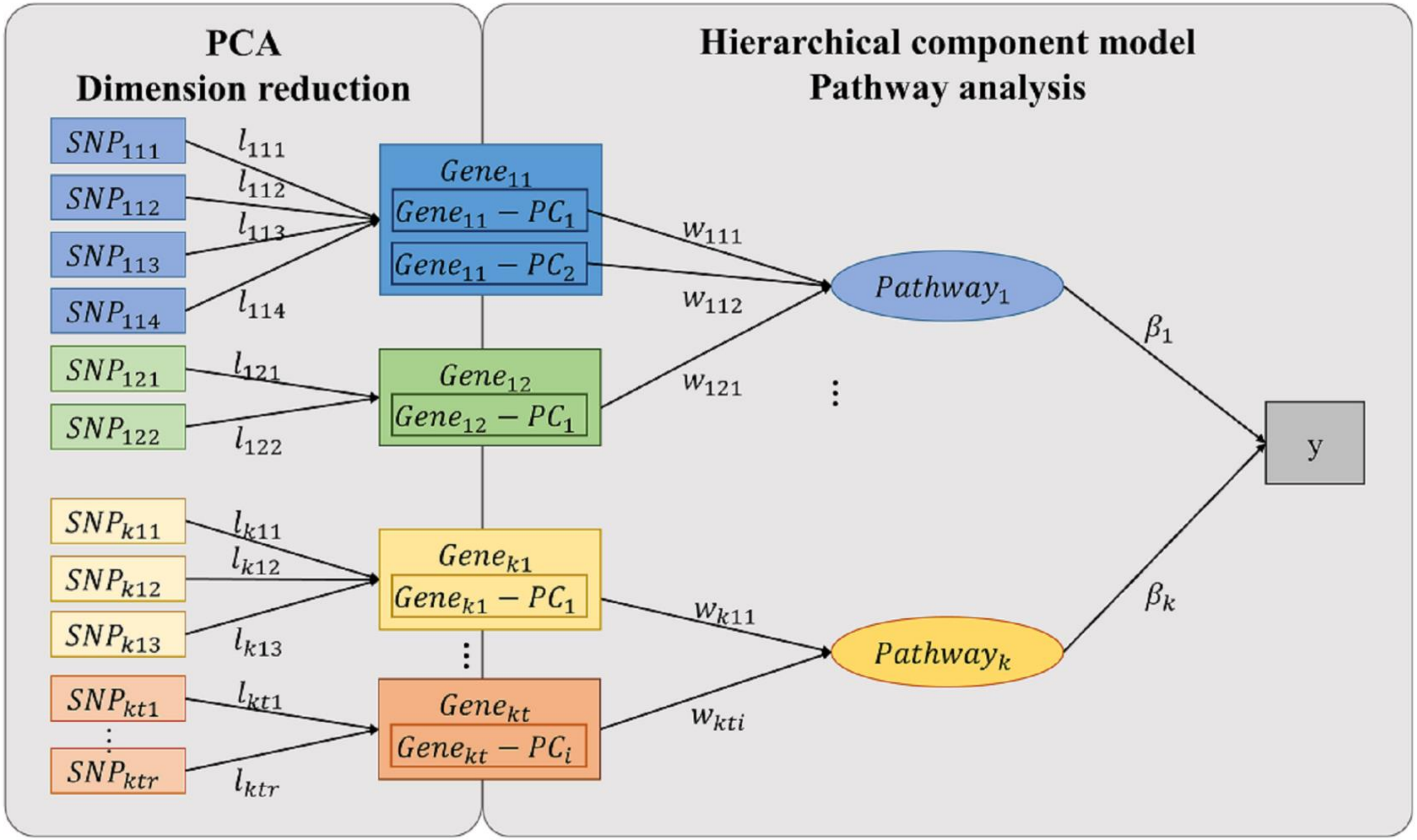
$\boldsymbol{Step2-2}$: For fixed the pathway coefficient estimates β_k, update the weight coefficient estimates w_{itk} in the least−squares.

To allow the potential correlations exist in the biological process, we utilize penalization approach on the effects of both genes and pathways. In this study, we adopt a ridge−type penalty to control multi−collinearity between genes and between pathways. Then, we seek to maximize the penalized log−likelihood function, which is given as follows:

$$\phi = \sum_{j=1}^{N} \log p(y_j; \beta_k, \delta) - \frac{1}{2}\lambda_g \sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{i=1}^{N_{kt}} w_{itk}^2 - \frac{1}{2}\lambda_p \sum_{k=1}^{K} \beta_k^2$$

where $p(y_j; \beta_k, \delta)$ is the probability distribution for the phenotype of the j^{th} individual. λ_g and λ_p are ridge parameters for genes and pathways. After estimation, we perform permutation test by resampling the phenotypes to test the significance of parameters.

Figure 1: A schematic diagram of HisCoM−PCA



Chapter 4

Result

4.1. Real data analysis of common variants from KARE

For KARE data, PLINK 1.90 [19] was used to perform the quality control analysis with the criteria described in Material section. The SNPs were mapped to the UCSC hg19 genomic coordination. Missing genotype data was imputed using the Beagle 5.0 [20] software program. Then the SNPs were annotated with genes using SnpEff v.4.3 [21]. After mapping these genes to KEGG pathway database, a total of 3,996 genes were matched with 186 KEGG pathways. We performed pathway analysis for four phenotypes: T2D, HT, SBP, and DBP. Following association tests conducted by other previous studies for KARE dataset, age, sex, body mass index (BMI), and area were included as covariates in the pathway analysis. In addition to HisCoM-PCA, other existing methods such as sARTP, MAGMA (self-contained version and competitive version) and GSA-SNP2 were used for a purpose of comparison. After performing PCA for each gene, we performed two HisCoM-PCA: one using the 3,996 first PCs and the second using 4,486 PCs whose cumulative proportion of variances were more than 30%. In addition to pathway analysis, HisCoM−PCA performed gene analysis at the same time. The tuning parameters λ_g and λ_p, were chosen based on five−fold cross−validation (CV). To test the significance of pathways, we performed permutation test by generating 1,000 permuted phenotypes.

HisCoM−PCA using the first PCs identified 14 pathways for T2D, 15 pathways for HT, 3 pathways for SBP, and 9 pathways for DBP, respectively at a 5% significance level. HisCoM−PCA using the 4,486 PCs identified 13 pathways for T2D, 20 pathways for HT, 6 pathways for SBP and 7 pathways for DBP, respectively at the same significance level. These different PC selection criteria provided very consistent results. Both identified 10 common pathways for T2D, 14 common pathways for HT, three common pathways for SBP, and five common pathways for DBP.

4.1.1. Real data analysis for T2D

For T2D analysis, HisCoM−PCA successfully identified the well−known pathways biologically related with T2D. For example, the pathways such as calcium signaling pathway, renin−angiotensin system pathway, and phosphatidylinositol signaling pathway were known to be related to insulin resistance or insulin sensitivity [22−25]. Calcium signaling is crucial for insulin secretion in pancreatic β −cells [22, 23]. In phosphatidylinositol signaling system, PI3K/PtdIns P3 signaling is known as an important role in the insulin stimulated glucose metabolism pathway which is associated with obesity and T2D [25]. Some diseases such as Alzheimer' s disease(AD), asthma, and dilated cardiovascular have been reported to share molecular pathways or risk factors with T2D [26-29]. For example, several studies have shown that insulin resistance is related to risk of AD as well as T2D [26]. Application of HisCoM-PCA with T2D successfully identified the pathways of these diseases. In addition, folate biosynthesis pathway and hedgehog signaling pathway are also reported to be potentially relate to T2D [30, 31]. These pathway results for T2D using HisCoM-PCA and other four methods are summarized in Table3.

Table 3: Significant pathways identified for T2D

Pathway	HisCoM−PCA (1^{st} PC)		HisCoM−PCA (30% PCs)		MAGMA (Competitive)	MAGMA (Self−contained)	GSA−SNP2	sARTP
	# PCs	P	# PCs	P	P values			
FOLATE BIOSYNTHESIS	8	0.0040	8	0.0020	0.0537	0.0540	0.0019	0.0052
HEDGEHOG SIGNALING PATHWAY	47	0.0060	50	0.0160	0.3073	0.1356	0.0280	0.1209
OLFACTORY TRANSDUCTION	238	0.0060	246	0.0020	0.0482	0.0062	0.0036	0.0836
BIOSYNTHESIS OF UNSATURATED FATTY ACIDS	19	0.0100	20	0.0120	0.0890	0.0355	0.0373	0.0453
ALZHEIMERS DISEASE	120	0.0140	142	0.0360	0.1808	0.0278	0.0446	0.2004
CALCIUM SIGNALING PATHWAY	141	0.0140	195	0.0260	0.0791	0.0248	0.0329	0.7193
ASTHMA	22	0.0160	22	0.0080	0.0517	0.0623	0.0737	0.0782
ACUTE MYELOID LEUKEMIA	45	0.0320	48	0.0420	0.0781	0.1895	0.5442	0.6037
MELANOGENESIS	79	0.0340	89	0.0120	0.0786	0.0133	0.0223	0.3000
LONG TERM POTENTIATION	55	0.0400	78	0.0280	0.1284	0.0715	0.1616	0.8167
PHOSPHATIDYLINOSITOL SIGNALING SYSTEM	64	0.1119	83	0.0060	0.7656	0.4586	0.1107	0.7707
DILATED CARDIOMYOPATHY	76	0.0300	106	0.0819	0.3994	0.0417	0.0408	0.7150
RENIN ANGIOTENSIN SYSTEM	13	0.0380	14	0.0539	0.3794	0.2651	0.0917	0.0651

4.1.2. Real data analysis for blood pressure

The pathways related to blood pressure (BP) were also identified by HisCoM−PCA using phenotypes HT, SBP, and DBP. Calcium signaling pathway and complement and coagulation cascades pathway were known to be related to BP regulation [32, 33]. BP regulation is influenced by regulators of vascular tone. The regulators of vascular tone are dependent on ion channels such as voltage−gated Ca2+ channels in calcium signaling pathway. Kallikrein−kinin system (KKS) is an important regulator of BP by influencing vascular tone and renal salt handling. It is well known that KKS is a large picture of complement and coagulation cascades pathway. There are also some disease pathways such as maturity onset diabetes of the young (MODY) and hypertrophic cardiomyopathy (HCM) identified by HisCoM−PCA. Some previous studies have shown that MODY and HCM may have association with HP [34−36]. These pathway results for HT, SBP, and DBP using HisCoM−PCA and other four methods are shown in Table 4.

Table 4: Significant pathways identified for blood pressure

(a) Significant pathways of hypertension

Pathway	HisCoM−PCA (1st PC)		HisCoM−PCA (30% PCs)		MAGMA (Competitive)	MAGMA (Self−contained)	GSA−SNP2	sARTP
	# PCs	P	# PCs	P	P values			
INOSITOL PHOSPHATE METABOLISM	45	0.0020	54	0.0040	0.1966	0.0003	0.0000	0.0001
PHOSPHATIDYLINOSITOL SIGNALING SYSTEM	64	0.0020	83	0.0040	0.2256	0.0027	0.0001	0.0001
UBIQUITIN MEDIATED PROTEOLYSIS	107	0.0020	113	0.0020	0.1076	0.0428	0.5963	0.0001
CALCIUM SIGNALING PATHWAY	141	0.0080	195	0.0060	0.0994	0.0268	0.0228	0.0001
NEUROTROPHIN SIGNALING PATHWAY	102	0.0100	113	0.0100	0.6801	0.1073	0.3157	0.0001
EPITHELIAL CELL SIGNALING IN HELICOBACTER PYLORI INFECTION	61	0.0120	66	0.0500	0.0826	0.0076	0.1245	0.0001
COMPLEMENT AND COAGULATION CASCADES	57	0.0140	61	0.0300	0.1428	0.0651	0.3405	0.0001
MATURITY ONSET DIABETES OF THE YOUNG	18	0.0180	19	0.0360	0.0670	0.0474	0.0239	0.0001
SNARE INTERACTIONS IN VESICULAR TRANSPORT	32	0.0220	33	0.0120	0.2071	0.0342	0.0653	0.0001

OTHER GLYCAN DEGRADATION	11	0.0260	12	0.0300	0.0535	0.0293	0.0767	0.0001
NITROGEN METABOLISM	20	0.0280	20	0.0420	0.5481	0.1157	0.3834	0.0001
PROTEASOME	39	0.0320	39	0.0260	0.1299	0.1896	0.3915	0.0001
GLYCOLYSIS GLUCONEOGENESIS	51	0.0340	54	0.0280	0.6343	0.3520	0.4836	0.0001
ARGININE AND PROLINE METABOLISM	36	0.0340	38	0.0320	0.5647	0.3794	0.0546	0.0001
HYPERTROPHIC CARDIOMYOPATHY HCM	70	0.1259	98	0.0300	0.0074	0.0027	0.0003	0.0001

(b) Significant pathways of SBP

Pathway	HisCoM−PCA (1st PC)		HisCoM−PCA (30% PCs)		MAGMA (Competitive)	MAGMA (Self−contained)	GSA−SNP2	sARTP
	# PCs	P	# PCs	P	P values			
VALINE LEUCINE AND ISOLEUCINE BIOSYNTHESIS	8	0.0160	8	0.0140	0.0281	0.0402	0.0782	0.1021
INTESTINAL IMMUNE NETWORK FOR IGA PRODUCTION	37	0.0300	37	0.0420	0.0995	0.1545	0.0268	0.3969
FOLATE BIOSYNTHESIS	8	0.0300	8	0.0340	0.1721	0.1939	0.3004	0.3650
AXON GUIDANCE	106	0.3197	156	0.0380	0.0149	0.0030	0.0649	0.0022

(c) Significant pathways of DBP

Pathway	HisCoM−PCA (1st PC)		HisCoM−PCA (30% PCs)		MAGMA (Competitive)	MAGMA (Self−contained)	GSA−SNP2	sARTP
	# PCs	P	# PCs	P	P values			
REGULATION OF AUTOPHAGY	23	0.0100	23	0.0360	0.1024	0.0597	0.2756	0.0268
VALINE LEUCINE AND ISOLEUCINE BIOSYNTHESIS	8	0.0120	8	0.0060	0.0477	0.0302	0.0280	0.0478
HOMOLOGOUS RECOMBINATION	24	0.0320	25	0.0240	0.3157	0.2229	0.0934	0.2029
INTESTINAL IMMUNE NETWORK FOR IGA PRODUCTION	37	0.0380	37	0.0440	0.0416	0.3198	0.1784	0.6592
BASAL TRANSCRIPTION FACTORS	27	0.0380	27	0.0400	0.2492	0.1362	0.3175	0.5487
CALCIUM SIGNALING PATHWAY	141	0.0440	195	0.0679	0.6141	0.2355	0.5279	0.3362
AXON GUIDANCE	106	0.1738	156	0.0160	0.4500	0.2196	0.1568	0.5813
UBIQUITIN MEDIATED PROTEOLYSIS	107	0.0699	113	0.0260	0.7291	0.2859	0.5946	0.5479
ALZHEIMERS DISEASE	120	0.0260	142	0.0519	0.7348	0.6050	0.5982	0.3647

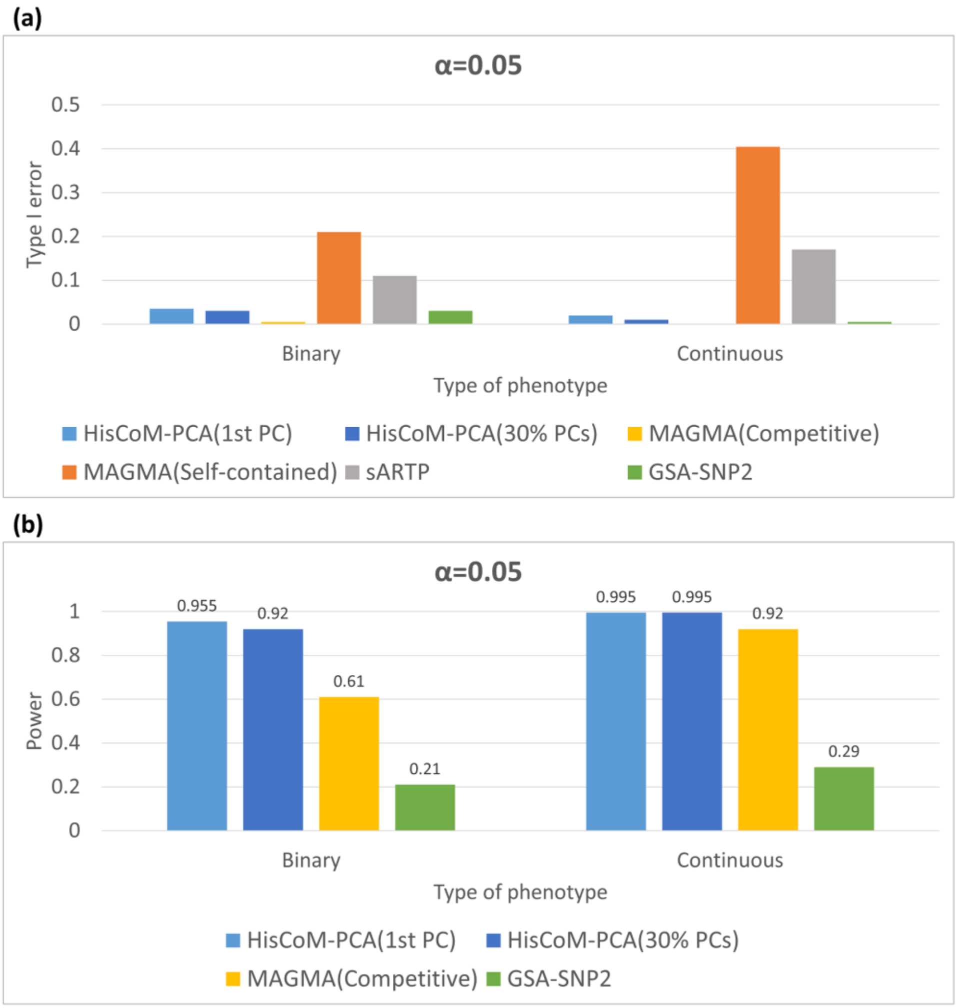
4.2. Simulation study using GAW17 dataset

To check the power and type I error of HisCoM-PCA, we performed a simulation study using the GAW17 dataset, for both binary and continuous types of a Q1 trait. For binary phenotypes, we transformed the continuous values of Q1 to binary values, using the median. Each SNP was then assigned to a gene, if its location was in, or within 20 kb of, the gene, and the KEGG database then used to map genes and pathways. In the simulation study, we chose the first PCs and PCs whose cumulative proportion of variances was more than 30%, after PCA of each gene. The tuning parameters of our method, λ_g and λ_p, were based on five-fold CV.

4.2.1. Type I error

To investigate where the type I error rate is controlled, we examined type I error by the proportion of identifying a null pathway whose number of genes was the same as the VEGF pathway. We checked the type I errors of HisCoM-PCA, sARTP, competitive version of MAGMA, self-contained version of MAGMA, and GSA-SNP2 (Figure 2).

Figure 2: Empirical type I errors and powers of HisCoM−PCA and other methods.



a) Empirical type I errors of HisCoM−PCA, sARTP, two versions of MAGMA, and GSA−SNP2. Empirical type I error indicates the times of identifying a null pathway among 200 replicates. (b)Empirical powers of HisCoM−PCA, competitive version of MAGMA and GSA−SNP2. Empirical power indicates the times of identifying VEGF pathway among 200 replicates.

To that end, HisCoM-PCA controlled type I error with PC selection criteria. GSA-SNP2, and the competitive version of MAGMA, also controlled type I error well. However, the type I errors of sARTP, and the self-contained version of MAGMA, were too inflated.

4.2.2. Power

According to the results of type I error, we only compared the power of HisCoM-PCA with GSA-SNP2, and the competitive version of MAGMA. To examine the power, we calculated proportion of identifying VEGF pathways from 168 KEGG pathways, with 200 replicates. The powers of the three methods are shown in Figure 2. For both continuous and binary phenotypes, HisCoM-PCA showed the highest power, compared to the other methods. The powers of HisCoM-PCA, with two types of phenotypes, and two criteria of PC selection, were all higher than 0.95. However, GSA-SNP2 only identified the VEGF pathway in 21% and 29% of cases for the binary and continuous phenotypes, respectively. While MAGMA showed higher power than GSA-SNP2, it showed only 0.6 power for the binary phenotype. However, HisCoM-PCA showed similar powers with either PC selection criteria, while all the methods showed higher power with continuous vs. binary phenotypes.

Chapter 5

Discussion

HisCoM−PCA is a novel method for pathway analysis of GWAS data. By applying HisCoM−PCA to a large population study dataset (KARE), we identified several biologically associated pathways for type−2 diabetes (T2D) and blood pressure (BP). For BP, we used three phenotypes: hypertension (HT), systolic blood pressure (SBP), and diastolic blood pressure (DBP). Whether the phenotype of interest is continuous or binary, HisCoM−PCA can successfully detect associated pathways with statistical significance. As self−directed validation, some pathways related to HT were also identified for SBP or DBP, simultaneously providing significant p−values. Beside pathway analysis, we performed gene analysis using HisCoM−PCA at the same time. The significant results of genes are shown in Additional file 1. To that end, HisCoM−PCA identified several genes well known to genetically influence T2D or BP, demonstrating that HisCoM−PCA can detect both pathways and genes having biological significance.

Other existing pathway methods revealed numbers of significant pathways. As shown in simulation studies, however, they have high chance of being false positives. On the other hand, some pathways identified by HisCoM−PCA were previously reported to be related to T2D or BP, while these pathways were not significant by other pathway identification methods we used for comparison. In addition, some pathways were jointly identified by other methods and HisCoM−PCA. Real data analysis showed that HisCoM−PCA can provide new candidates that other methods cannot successfully identify.

We also examined empirical power and type I error rate for both binary and continuous phenotypes, using the Genetic Analysis Workshop 17 (GAW17) simulation dataset for GWAS. Compared to several methods, HisCoM−PCA controlled type I error well and showed high statistical power. However, some methods, such as sARTP and the self−contained version of MAGMA, did not control type I error well. Moreover, the methods that well controlled type I error showed lower power than HisCoM−PCA. In the simulation study, HisCoM−PCA analysis of the first PC showed similar power to HisCoM−PCA with PCs whose cumulative proportion of variance was more than 30%. This may indicate that the power is similar, using multiple PCs, which can save a lot of computing time.

HisCoM−PCA performs both gene−based and pathway−based analysis directly from raw data. Most other existing methods use summary measures such as p−values or test statistics of univariate analysis. These are gene−level summary measures and are used as inputs to perform pathway analysis. However, since the values of these summaries do not directly represent the raw genetic data, this issue probably leads to false discoveries. In HisCoM−PCA, we can obtain gene−level summary statistics, by PCA, for each gene. These statistics are a linear combination of SNPs from the raw data. Using these values, subsequent analysis for genes and pathways may decrease the possibility of false discoveries.

HisCoM−PCA also considers correlations between pathways, an aspect usually neglected by other methods. Further, correlation between pathways may influence the combined effect of pathways on traits, similar to when correlations exist between genes in a specific pathway. To allow correlation between genes and between pathways, HisCoM−PCA applies a ridge−type penalization approach on coefficient estimation for both genes and pathways. Cross−validation is then used to detect the optimal tuning parameters of ridge−type penalties. During such consideration of correlations, HisCoM−PCA performs gene−based and pathway−based analyses simultaneously, using the entirety of genes and pathways. However, most existing methods perform these two analyses separately. In addition, other methods often are limited to performing single gene analysis and single pathway analysis.

In addition to the above advantages, HisCoM−PCA has high flexibility for users. First, PC selection criteria may be defined by the user. Second, users can perform both non−target and target pathway analysis. Since HisCoM−PCA controls the correlation between pathways, it is useful to detect associated pathways having similar molecular mechanisms. Thus, we strongly believe that our method, HisCoM-PCA, can be applied to any number of GWAS studies, resulting in the successful identification of genes and pathways associated with specific phenotypes.

Bibliography

1. Xue, A., et al., *Genome−wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes.* Nature communications, 2018. **9**(1): p. 2941.
2. Prasad, R. and L. Groop, *Genetics of type 2 diabetes—pitfalls and possibilities.* Genes, 2015. **6**(1): p. 87−123.
3. Costanzo, M., et al., *The genetic landscape of a cell.* science, 2010. **327**(5964): p. 425−431.
4. Yoon, S., et al., *Efficient pathway enrichment and network analysis of GWAS summary data using GSA−SNP2.* Nucleic acids research, 2018. **46**(10): p. e60−e60.
5. Zhang, K., et al., *i−GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome−wide association study.* Nucleic acids research, 2010. **38**(suppl_2): p. W90−W95.
6. Segrè, A.V., et al., *Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits.* PLoS genetics, 2010. **6**(8): p. e1001058.
7. Alexa, A., J. Rahnenführer, and T. Lengauer, *Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.* Bioinformatics, 2006. **22**(13): p. 1600−1607.
8. Lee, S., et al., *Pathway−based approach using hierarchical components of collapsed rare variants.* Bioinformatics, 2016. **32**(17): p. i586−i594.
9. Cho, Y.S., et al., *A large−scale genome−wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits.* Nature genetics, 2009. **41**(5): p. 527.
10. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic acids research, 2004. **32**(suppl_1): p. D277−D280.
11. Zhang, H., et al., *A powerful procedure for pathway−based meta−analysis using summary statistics identifies 43 pathways associated with type II diabetes in European populations.* PLoS genetics, 2016. **12**(6): p. e1006122.
12. de Leeuw, C.A., et al., *MAGMA: generalized gene−set analysis of GWAS data.* PLoS computational biology, 2015. **11**(4): p. e1004219.
13. Almasy, L., et al. *Genetic Analysis Workshop 17 mini−exome simulation.* in *BMC proceedings*. 2011. BioMed Central.
14. Lim, J., I. Koh, and Y.S. Cho, *Identification of genetic loci stratified by diabetic status and microRNA related SNPs influencing kidney function in Korean populations.* Genes & Genomics, 2016. **38**(7): p. 601−609.
15. Jin, H.−S., et al., *Replication of an African−American GWAS on blood pressure and hypertension in the Korean population.* Genes &

Genomics, 2011. **33**(2): p. 127.

16. Consortium, G.P., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061.
17. *Ingenuity Pathways Analysis software web link* Available from: http://www.ingenuity.com/.
18. Hwang, H. and Y. Takane, *Generalized structured component analysis.* Psychometrika, 2004. **69**(1): p. 81-99.
19. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**(1): p. 7.
20. Browning, B.L., Y. Zhou, and S.R. Browning, *A one-penny imputed genome from next-generation reference panels.* The American Journal of Human Genetics, 2018. **103**(3): p. 338-348.
21. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly, 2012. **6**(2): p. 80-92.
22. Levy, J., *Abnormal cell calcium homeostasis in type 2 diabetes mellitus.* Endocrine, 1999. **10**(1): p. 1-6.
23. Hodgkin, M., C. Hills, and P. Squires, *The calcium-sensing receptor and insulin secretion: a role outside systemic control 15 years on.* The Journal of endocrinology, 2008. **199**(1): p. 1-4.
24. Scheen, A.J., *Prevention of type 2 diabetes mellitus through inhibition of the Renin-Angiotensin system.* Drugs, 2004. **64**(22): p. 2537-2565.
25. Manna, P. and S.K. Jain, *Phosphatidylinositol-3, 4, 5-triphosphate and cellular signaling: implications for obesity and diabetes.* Cellular Physiology and Biochemistry, 2015. **35**(4): p. 1253-1275.
26. Han, W. and C. Li, *Linking type 2 diabetes and Alzheimer's disease.* Proceedings of the National Academy of Sciences, 2010. **107**(15): p. 6557-6558.
27. Luchsinger, J.A. and D.R. Gustafson, *Adiposity, type 2 diabetes, and Alzheimer's disease.* Journal of Alzheimer's Disease, 2009. **16**(4): p. 693-704.
28. Shore, S.A., *Obesity and asthma: possible mechanisms.* Journal of Allergy and Clinical Immunology, 2008. **121**(5): p. 1087-1093.
29. Chan, K.H.K., et al., *Shared molecular pathways and gene networks for cardiovascular disease and type 2 diabetes mellitus in women across diverse ethnicities.* Circulation: Cardiovascular Genetics, 2014. **7**(6): p. 911-919.
30. Al-Maskari, M.Y., et al., *Folate and vitamin B12 deficiency and hyperhomocysteinemia promote oxidative stress in adult type 2 diabetes.* Nutrition, 2012. **28**(7-8): p. e23-e26.
31. Thomas, M.K., et al., *Hedgehog signaling regulation of insulin production by pancreatic beta-cells.* Diabetes, 2000. **49**(12): p. 2039-2047.
32. Adragna, N.C. and P.K. Lauf, *K-Cl cotransport function and its*

potential contribution to cardiovascular disease. Pathophysiology, 2007. **14**(3-4): p. 135-146.

33. Kraja, A.T., et al., *Genetics of hypertension and cardiovascular disease and their interconnected pathways: lessons from large studies.* Current hypertension reports, 2011. **13**(1): p. 46-54.

34. Schober, E., et al., *Phenotypical aspects of maturity-onset diabetes of the young (MODY diabetes) in comparison with Type 2 diabetes mellitus (T2DM) in children and adolescents: Experience from a large multicentre database.* Diabetic Medicine, 2009. **26**(5): p. 466-473.

35. Misawa, K., et al., *Difference in coronary blood flow dynamics between patients with hypertension and those with hypertrophic cardiomyopathy.* Hypertension Research, 2002. **25**(5): p. 711-716.

36. Takeda, A. and N. Takeda, *Different pathophysiology of cardiac hypertrophy in hypertension and hypertrophic cardiomyopathy.* Journal of molecular and cellular cardiology, 1997. **29**(11): p. 2961-2965.

초 록

전장 유전체 상관성 분석 연구 (Genome-Wide Association Study, GWAS)에서 이미 많은 통계 방법을 이용하여 표현형과 관련된 대립유전자 빈도가 비교적 큰 변이(common variant)를 발굴 했다. 그러나 발굴된 통계적으로 유의미한 변이들로 추정된 유전력의 일부만 설명할 수 있다. 이러한 '유전적 결실' (missing heritability) 을 해결하기 위하여 유전자(gene) 기반 및 패스웨이(pathway) 기반한 연구가 많이 진행되고 있고 GWAS 데이터를 이용하여 생물학적 기작 및 관련된 패스웨이를 찾았다. 하지만 사용된 많은 방법들은 유정자 간 및 패스웨이 간의 상호 관계를 고려하지 않았다. 본 연구에서는 유전자 간 및 패스웨이 간의 상호 관계를 고려하는 계층적 구조 모형 기반으로 GWAS 데이터를 이용하는 새로운 패스웨이 기반 분석 방법을 개발 했음. 이 방법의 이름은 HisCoM-PCA(Hierarchical structural Component Model for Pathway analysis of Common vAriants)이다. HisCoM-PCA는 우선 동일한 유전자에 속하는 common variants를 한 통계량으로 요약하고, 계산된 통계량을 이용하여 유전자 기반 분석과 패스웨이 기반 분석을 릿지 회귀분석 방법을 통하여 동시에 진행한다. 그리고 순열검정법(permutation test)을 통해서 유전자와 패스웨이의 유의성 검정은 진행 한다. 본 연구에서 GAW17 시뮬레이션 데이터를 이용하여 이진형 표현형과 연속형 표현형에 대한 시뮬레이션을 통해 HisCoM-PCA는 제 I 형 오류를 잘 통제하고 여러 가지 방법보다 더 높은 검정력을 가지고 있는 것으로 확인 했다. 그리고 HisCoM-PCA를 한국인 유전체 분석사업(KARE) 자료에 적용하여 4가지 인체 표현형: (1) 2형 당뇨병; (2) 고혈압; (3) 수축기 혈압, (4) 이완기 혈압에

대하여 분석 했을 때, 분석 결과를 통하여 HisCoM-PCA는 통계적으로 유의미하고 생물적인 의미 있는 패스웨이를 발굴할 수 있는 것으로 확인됐다.

주요어: 전장 유전체 상관성 분석 연구, 단일염기다형성, 계층적 구조 모형, 패스웨이분석

학 번: 2017-20507