

# Using ASGSCA

Hela Romdhani, Stepan Grinek, Heungsun Hwang and Aurelie Labbe

October 17, 2016

## 1 Introduction

The **ASGSCA** (Association Study using GSCA) package provides tools to model and test the associations between multiple genotypes and multiple traits, taking into account prior biological knowledge. Functional genomic regions, e.g., genes and clinical pathways, are incorporated in the model as latent variables that are not directly observed. See ? for details. The method is based on Generalized Structured Component Analysis (GSCA) (?). GSCA is an approach to structural equation models (SEM) and thus constitutes two sub-models: measurement and structural models. The former specifies the relationships between observed variables (here genotypes and traits) and latent variables (here genes or more generally genomic regions and clinical pathways), whereas the structural model expresses the relationships between latent variables.

Assume we have data on  $J$  candidate SNPs ( $X_1, \dots, X_J$ ) and  $K$  traits ( $X_{J+1}, \dots, X_{J+K}$ ). Let  $I = J + K$  denote the total number of observed variables. Suppose that the  $J$  SNPs are mapped to  $G$  different genes or regions ( $\gamma_1, \dots, \gamma_G$ ) and the  $K$  traits are involved in  $T$  different clinical pathways ( $\gamma_{G+1}, \dots, \gamma_{G+T}$ ). The measurement model is given by

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

where  $S_\ell$  denotes the set of indices of the observed variables mapped to the  $\ell^{th}$  latent structure,  $w_{i\ell}$  denotes the weight associated with the observed variable  $X_i$  in the definition of the latent variable  $\ell$  and  $L = G + T$  is the total number of latent variables in the model. Let  $W$  denotes the  $I \times L$  matrix of weights. The structural model is given by

$$\gamma_\ell = \sum_{\ell'=1, \ell' \neq \ell}^L b_{\ell'\ell} \gamma_{\ell'} + \epsilon_\ell, \quad \ell = 1, \dots, L \quad (2)$$

where  $\epsilon_\ell$  represents the error term and  $b_{\ell'\ell}$  represents the path coefficient linking  $\gamma_{\ell'}$  to  $\gamma_\ell$ . Let  $B$  denote the  $L \times L$  matrix of path coefficients.

Figure ?? shows an example with four SNPs mapped to two different genes and four traits involved in two clinical

pathways. Gene 1 is only associated with clinical pathway 1 while gene 2 is related to both clinical pathways. One of the traits is involved in both clinical pathways.

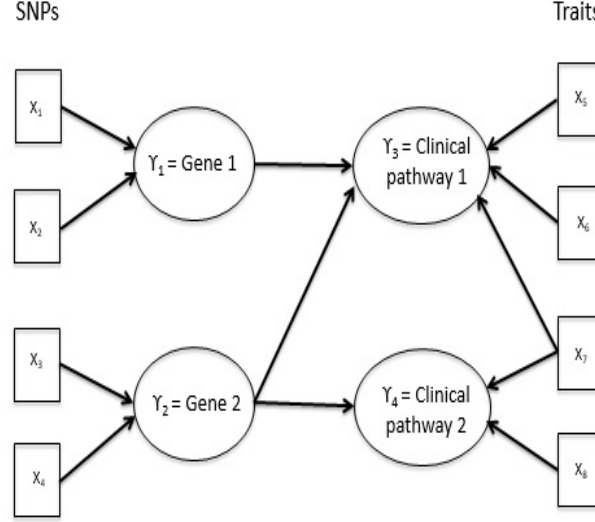


Figure 1: Example with 4 SNPs, 2 genes, 4 traits and 2 clinical pathways.

The **ASGSCA** package consists of one main function **GSCA** which allows one to estimate the model parameters and/or run tests for the null hypothesis  $H_0^{\ell', \ell} : b_{\ell' \ell} = 0$  of no effect of gene  $\gamma_{\ell'}$  on clinical pathway  $\gamma_{\ell}$ . Indeed,  $b_{\ell' \ell}$  quantifies the joint effect of the genotypes mapped to gene  $\gamma_{\ell'}$  on the traits involved in clinical pathway  $\gamma_{\ell}$  together.

## 2 Package use

First, the function **GSCA** can be used to estimate the weight and path coefficients by minimizing a global least square criterion using an Alternating Least-Squares (ALS) algorithm (??). This algorithm alternates between two main steps until convergence: 1) the weight coefficients  $w_{i\ell}$ ,  $i = 1, \dots, I$ ,  $\ell = 1, \dots, L$  are fixed, and the path coefficients  $b_{\ell' \ell}$ ,  $\ell, \ell' = 1, \dots, L$ ,  $\ell \neq \ell'$  are updated in the least-squares sense; 2) the weights  $w_{i\ell}$  are updated in the least-squares sense for fixed path coefficients  $b_{\ell' \ell}$ . Note here that if one runs the function **GSCA** twice for the same dataset, signs of the parameters estimates may change. Nevertheless, the meaning of the estimate remains the same. For example, for the same dataset, a first estimation could result in a positive path coefficient estimate between two latent variables as well as all positive weight estimates for the two latent variables, while a second run of the function could produce a negative path coefficient estimate between the same latent variables because it yields negative weight estimates for one latent variable whereas positive weight estimates for the other latent variable.

The function **GSCA** also allows one to test for the association between genes (multiple genotypes) and clinical pathways (multiple traits). It performs permutation test procedures for the significance of the path coefficients relating two latent variables. The user has the option to specify a subset of path coefficients to be tested, otherwise

the test is performed on all the gene-clinical pathway connections in the model.

The dataset should be given in a data frame object. To run the **GSCA** function, one should also provide two matrices  $W0$  and  $B0$  that indicate connections between the different components of the model. Concretely,  $W0$  is an  $I \times L$  matrix (the rows correspond to the genotypes and traits and the columns to genes and clinical pathways) with 0's and 1's, where a value of 1 indicates an arrow from the observed variable in the row to the latent variable in the column. Similarly,  $B0$  is an  $L \times L$  matrix with 0's and 1's, where a value of 1 indicates an arrow from the latent variable in the row to the latent variable in the column.

A dataset containing some variables of interest from the Quebec Child and Adolescent Health and Social Survey (QCAHS), observed on 1707 French Canadian participants (860 boys and 847 girls), is included in the package. Detailed descriptions of the QCAHS design and methods can be found in ?. The dataset contains z-score transformation (standardized for age and sex) of 8 traits grouped into three clinical pathways: lipid metabolism, energy metabolism and blood pressure control. Table ?? gives a list of the traits and the corresponding clinical pathways. The dataset also contains genotypic data on 35 variants within 25 genes listed in Table ?? along with biological pathways within which they fall. Among the considered genetic variants, 33 are SNPs that were coded according to the additive model. The other two are polymorphisms with more than two alleles. The first, with three alleles, belongs to the gene APOE. It has three alleles and then admits 6 different genotypes. It is coded using 5 indicator variables APOE1-APOE5. The second is a variant from gene PCSK9, also with three alleles but only 4 different genotypes are observed in our dataset. It was coded using 3 indicator variables PCSK9Leu1-PCSK9Leu3.

Table 1: Traits of interest in the QCAHS dataset.

Lipid pathway only	Both Lipid and Energy pathways	Energy pathway only	Blood pressure (BP) control pathway
High-density lipoprotein (HDL)	Low-density-lipoprotein (LDL)	Fasting glucose	Systolic blood pressure (SBP)
	Apolipoprotein B (APOB)	Fasting insulin	Diastolic blood pressure (DBP)
	Triglycerides (TG)		

## 2.1 Analysis of a subset of the data

We first focus on a subset of the data: the traits involved in the lipid and/or energy metabolisms and the genes CETP, LPL, PGC and TNFa. The corresponding path model is illustrated in Figure ?. It involves 12 observed variables (6 SNPs and 6 traits) and 6 latent variables (4 genes and 2 clinical pathways)

Table 2: Candidate genotypes available in QCAHS possibly related to the considered pathways.

Lipid pathway only		Both Lipid and Energy pathways		Energy pathway only		Blood pressure control pathway	
Gene	Variant	Gene	Variant	Gene	Variant	Gene	Variant
CETP	TaqIB	PGC	G1564A	TNFa	G308A	eNOS	T-786C
ApoC3	C-482T		G-1302A		G238A		Glu298Asp
ABCA1	Arg219Lys						
FABP-2	T54A		T45G			a23-AR	DelGlu301-303
ApoA1	G-75A		G276T			b1-AR	Gly389Arg
APOE	E1E2E3		-11391			b2-AR	Gly16Arg
HL	C-514T		-11377				Gln27Glu
LPL	Hind III	Adiponectin		PPARg2	Pro12Ala	b3-AR	Trp64Arg
MTP	G-493T					ACE	Ins/Del
PON1	A192G					AGT	Met235Thr
PON2	C311G					AGTR1	A1166C
PCSK9	R46L					LEPR	Lys656Asn
	PCSK9Leu						Gln223Arg
							Lys109Arg

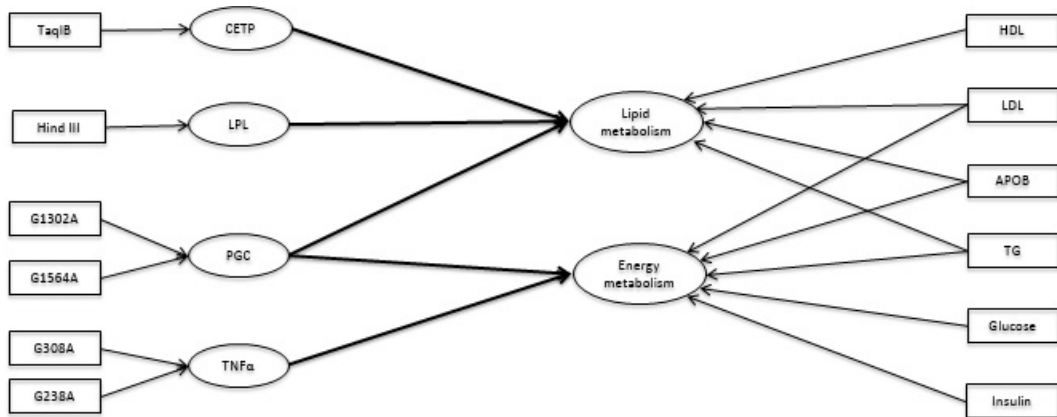


Figure 2: Path model for the considered variables.

We use the function **GSCA** to estimate the model and/or test the gene-clinical pathway connections as follows. Note that in the dataset QCAHS, the rows correspond to the individuals and the columns to the genotypes first then the traits.

```
> library(ASGSCA)
> data("QCAHS")
> #Names of all the observed variables: the SNPs then the traits
> colnames(QCAHS)

[1] "TaqIB"      "C482T"      "Arg219Lys"  "T54A"      "G75A"
[6] "APOE1"      "APOE2"      "APOE3"      "APOE4"      "APOE5"
[11] "C514T"      "HindIII"    "G493T"      "A192G"      "C311G"
[16] "R46L"      "PCSK9Leu1"  "PCSK9Leu2"  "PCSK9Leu3"  "G1302A"
[21] "G1564A"     "C11377G"    "C11391A"    "T45G"      "G276T"
[26] "Pro12Ala"   "G308A"      "G238A"      "T786C"      "Glu298Asp"
[31] "DelGlu301303" "Gly389Arg"  "Gly16Arg"   "Gln27Glu"   "Trp64Arg"
[36] "InsDel"     "Met235Thr"  "A1166C"     "Lys656Asn"  "Gln223Arg"
[41] "Lys109Arg"  "HDL"        "LDL"        "APOB"       "TG"
[46] "Glucose"    "Insulin"    "SBP"        "DBP"

> #Extract the variables of interest
> QCAHS1=data.frame(QCAHS$TaqIB,QCAHS$HindIII,QCAHS$G1302A,QCAHS$G1564A,QCAHS$G308A,QCAHS$G238A,
+ QCAHS$HDL,QCAHS$LDL,QCAHS$APOB,QCAHS$TG,QCAHS$Glucose,QCAHS$Insulin)
> #Names of the observed variables used in this example
> ObservedVar=c("TaqIB","HindIII","G1302A","G1564A","G308A","G238A","HDL","LDL","APOB",
+ "TG","Glucose","Insulin")
> colnames(QCAHS1)=ObservedVar
> #Define the vector of the latent variables names
> LatentVar=c("CETP","LPL","PGC","TNFa","Lipid metabolism","Energy metabolism")
> #Construction of the matrices W0 and B0 describing the model illustrated in Figure 2.
> W0=matrix(rep(0,12*6),nrow=12,ncol=6, dimnames=list(ObservedVar,LatentVar))
> W0[1,1]=W0[2,2]=W0[3:4,3]=W0[5:6,4]=W0[7:10,5]=W0[8:12,6]=1
> B0=matrix(rep(0,6*6),nrow=6,ncol=6, dimnames=list(LatentVar,LatentVar))
> B0[1:3,5]=B0[3:4,6]=1
> W0
```

CETP LPL PGC TNFa Lipid metabolism Energy metabolism

TaqIB	1	0	0	0	0	0
HindIII	0	1	0	0	0	0
G1302A	0	0	1	0	0	0
G1564A	0	0	1	0	0	0
G308A	0	0	0	1	0	0
G238A	0	0	0	1	0	0
HDL	0	0	0	0	1	0
LDL	0	0	0	0	1	1
APOB	0	0	0	0	1	1
TG	0	0	0	0	1	1
Glucose	0	0	0	0	0	1
Insulin	0	0	0	0	0	1

> B0

	CETP	LPL	PGC	TNFa	Lipid metabolism	Energy metabolism
CETP	0	0	0	0	1	0
LPL	0	0	0	0	1	0
PGC	0	0	0	0	1	1
TNFa	0	0	0	0	0	1
Lipid metabolism	0	0	0	0	0	0
Energy metabolism	0	0	0	0	0	0

The first row of  $W_0$  indicates an arrow directed from SNP TaqIB to gene CETP and no connection between this SNP and the other latent variables in the model. The first row of B0 indicates an arrow directed from gene CETP to the clinical pathway Lipid metabolism and no connection between this gene and the other latent variables in the model.

If one only wants to estimate the parameters of the model, the argument `estim` should be set to `TRUE` while `path.test` should be set to `FALSE`.

> `GSCA(QCAHS1,W0, B0,latent.names=LatentVar, estim=TRUE,path.test=FALSE,path=NULL,nperm=1000)`

\$Weight

	CETP	LPL	PGC	TNFa	Lipid metabolism	Energy metabolism
TaqIB	1	0	0.0000000	0.0000000	0.0000000	0.0000000
HindIII	0	1	0.0000000	0.0000000	0.0000000	0.0000000
G1302A	0	0	-0.2782234	0.0000000	0.0000000	0.0000000

G1564A	0	0	0.8716073	0.0000000	0.00000000	0.00000000
G308A	0	0	0.0000000	0.5915015	0.00000000	0.00000000
G238A	0	0	0.0000000	0.8456388	0.00000000	0.00000000
HDL	0	0	0.0000000	0.0000000	0.88870166	0.00000000
LDL	0	0	0.0000000	0.0000000	0.35988349	0.4983567
APOB	0	0	0.0000000	0.0000000	-0.09049176	0.9348191
TG	0	0	0.0000000	0.0000000	0.37018755	-0.4523388
Glucose	0	0	0.0000000	0.0000000	0.00000000	-0.2102506
Insulin	0	0	0.0000000	0.0000000	0.00000000	0.1480296

\$Path

	CETP	LPL	PGC	TNFa	Lipid metabolism	Energy metabolism
CETP	0	0	0	0	0.21070120	0.00000000
LPL	0	0	0	0	0.06648644	0.00000000
PGC	0	0	0	0	0.02204009	0.06243911
TNFa	0	0	0	0	0.00000000	0.10306018
Lipid metabolism	0	0	0	0	0.00000000	0.00000000
Energy metabolism	0	0	0	0	0.00000000	0.00000000

The output is a list of two matrices. The first one contains the estimates of the weights of the observed variables in the rows corresponding to the latent variables in the columns. The second matrix contains the estimates of the path coefficients relating the latent variables in the rows to those in the columns. The value  $\pm 0.278$  in the Weight output is the estimate of the weight coefficient relating SNP G1302A to gene PGC. The value  $\pm 0.210$  in the Path output corresponds to the estimate of the path coefficient between gene CETP and the clinical pathway Lipid metabolism. The obtained path model estimate is illustrated in Figure ??

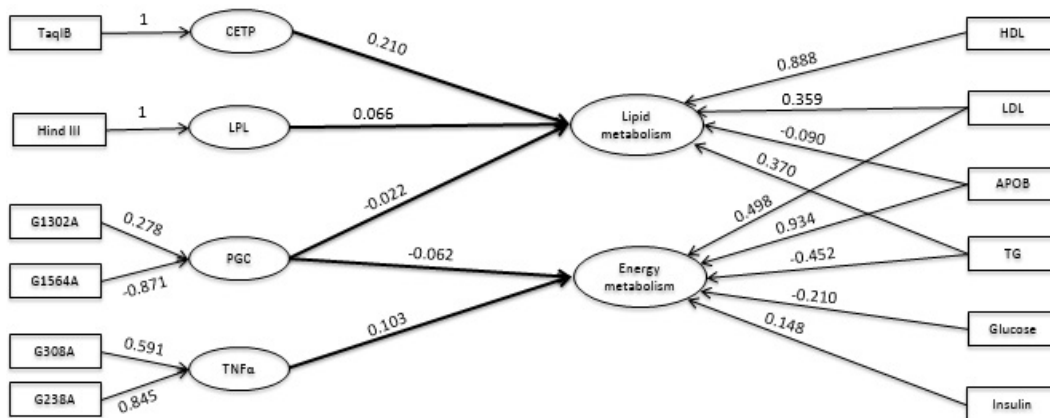


Figure 3: Path model estimated using GSCA.

To perform both estimation and test procedures, both arguments `estim` and `path.test` should be set to `TRUE`. The test procedure is based on random sampling of permutations, so we use the R function `set.seed` to insure the user will obtain the same results displayed here.

```
> set.seed(2)
> GSCA(QCAHS1,W0, B0,latent.names=LatentVar, estim=TRUE,path.test=TRUE,path=NULL,nperm=1000)
```

\$Weight

	CETP	LPL	PGC	TNFa	Lipid metabolism	Energy metabolism
TaqIB	1	0	0.0000000	0.0000000	0.00000000	0.00000000
HindIII	0	1	0.0000000	0.0000000	0.00000000	0.00000000
G1302A	0	0	-0.2782212	0.0000000	0.00000000	0.00000000
G1564A	0	0	0.8716085	0.0000000	0.00000000	0.00000000
G308A	0	0	0.0000000	0.5915036	0.00000000	0.00000000
G238A	0	0	0.0000000	0.8456374	0.00000000	0.00000000
HDL	0	0	0.0000000	0.0000000	0.88870172	0.00000000
LDL	0	0	0.0000000	0.0000000	0.35988337	0.4983573
APOB	0	0	0.0000000	0.0000000	-0.09049229	0.9348187
TG	0	0	0.0000000	0.0000000	0.37018786	-0.4523389
Glucose	0	0	0.0000000	0.0000000	0.00000000	-0.2102503
Insulin	0	0	0.0000000	0.0000000	0.00000000	0.1480297

\$Path

	CETP	LPL	PGC	TNFa	Lipid metabolism	Energy metabolism
CETP	0	0	0	0	0.21070124	0.00000000
LPL	0	0	0	0	0.06648647	0.00000000
PGC	0	0	0	0	0.02203967	0.06243912
TNFa	0	0	0	0	0.00000000	0.10306011
Lipid metabolism	0	0	0	0	0.00000000	0.00000000
Energy metabolism	0	0	0	0	0.00000000	0.00000000

\$pvalues

	CETP	LPL	PGC	TNFa	Lipid metabolism	Energy metabolism
CETP	NA	NA	NA	NA	0.000	NA
LPL	NA	NA	NA	NA	0.004	NA
PGC	NA	NA	NA	NA	0.566	0.113

TNFa	NA	NA	NA	NA	NA	0.010
Lipid metabolism	NA	NA	NA	NA	NA	NA
Energy metabolism	NA	NA	NA	NA	NA	NA

In this case, a matrix containing the *p-values* for all path coefficients is also given. NA is obtained where no connection is specified in  $B_0$ . For example, the obtained *p-value* for the path coefficient corresponding to the connection between gene LPL and the Lipid metabolism is equal to 0.004. This means that, under the studied model, the effect of gene LPL on the Lipid metabolism is significant (say for test level of 5%) which means that the effect of the SNP Hind III on the traits HDL, LDL, APOB and TG together is significant. The p-value 0.113 obtained for the connection between gene PGC and the clinical pathway Energy metabolism means that the joint effect of the SNPs G1302A and G1564A on the traits LDL, APOB, TG, Glucose and insulin together is not significant.

Now, if one only needs the results of the test procedure the argument `estim` could be set to `FALSE`.

```
> set.seed(2)
> GSCA(QCAHS1,W0, B0,latent.names=LatentVar,estim=FALSE,path.test=TRUE,path=NULL,nperm=1000)
```

	CETP	LPL	PGC	TNFa	Lipid metabolism	Energy metabolism
CETP	NA	NA	NA	NA	0.000	NA
LPL	NA	NA	NA	NA	0.004	NA
PGC	NA	NA	NA	NA	0.566	0.113
TNFa	NA	NA	NA	NA	NA	0.010
Lipid metabolism	NA	NA	NA	NA	NA	NA
Energy metabolism	NA	NA	NA	NA	NA	NA

It is also possible to perform the test for a subset of path coefficients of interest by assigning to the argument `path` a matrix of two columns. Each row of this matrix contains the indices of the two latent variables corresponding to an association to be tested. In the following example we test for the significance of the path coefficient relating the gene LPL (latent variable 2) and Lipid metabolism (latent variable 5) as well as the one relating the gene PGC (latent variable 3) and Energy metabolism (latent variable 6).

```
> set.seed(2)
> path0=matrix(c(2,3,5,6),ncol=2)
> path0
```

```
      [,1] [,2]
[1,]    2    5
[2,]    3    6
```

```
> GSCA(QCAHS1,W0, B0,latent.names=LatentVar, estim=FALSE,path.test=TRUE,path=path0,
+ nperm=1000)
```

	CETP	LPL	PGC	TNFa	Lipid metabolism	Energy metabolism
CETP	NA	NA	NA	NA	NA	NA
LPL	NA	NA	NA	NA	0.005	NA
PGC	NA	NA	NA	NA	NA	0.112
TNFa	NA	NA	NA	NA	NA	NA
Lipid metabolism	NA	NA	NA	NA	NA	NA
Energy metabolism	NA	NA	NA	NA	NA	NA

Only p-values for the specified path coefficients are computed, the others are set to NA.

## 2.2 Analysis of the complete data

Now, let's analyse all the available data. The corresponding GSCA model involves in total 49 observed variables (41 genotype variables and 8 phenotypes) and 28 latent variables (25 genes and 3 clinical pathways). Figure ?? shows a diagram illustrating the path model we want to fit to the data. The corresponding matrices W0 and B0 are included in the package.

```
> ObservedVar=colnames(QCAHS)
```

```
> ObservedVar
```

```
[1] "TaqIB"      "C482T"      "Arg219Lys"  "T54A"      "G75A"
[6] "APOE1"      "APOE2"      "APOE3"      "APOE4"      "APOE5"
[11] "C514T"      "HindIII"    "G493T"      "A192G"      "C311G"
[16] "R46L"       "PCSK9Leu1"  "PCSK9Leu2"  "PCSK9Leu3"  "G1302A"
[21] "G1564A"     "C11377G"    "C11391A"    "T45G"       "G276T"
[26] "Pro12Ala"   "G308A"      "G238A"      "T786C"      "Glu298Asp"
[31] "DelGlu301303" "Gly389Arg"  "Gly16Arg"   "Gln27Glu"   "Trp64Arg"
[36] "InsDel"     "Met235Thr"  "A1166C"     "Lys656Asn"  "Gln223Arg"
[41] "Lys109Arg"  "HDL"        "LDL"        "APOB"       "TG"
[46] "Glucose"    "Insulin"    "SBP"        "DBP"
```

```
> #Define the vector of the latent variables names
```

```
> LatentVar=c("CETP","APOC3","ABCA1","FABP-2","APOA1","APOE","HL","LPL","MTP","PON1","PON2","PCSK9",
+ "PGC","ADIPO","PPARG2","TNFa","eNOS","a23AR","b1AR","b2AR","b3AR","ACE","AGT","AGTR1","LEPR",
+ "Lipid metabolism", "Energy metabolism","BP control")
```

```
> #The matrices W0 and B0 describing the model illustrated in Figure 2.
```

```
> data(W0); data(B0)
```

```
> dim(W0)
```

```
[1] 49 28
```

```
> dim(B0)
```

```
[1] 28 28
```

We use the function GSCA to estimate the parameters of the model and test for the significance of the connections between genes and clinical pathways as shown in the comment below. The execution of this command takes around 2 hours on a current high-end laptop (26000 permutations of 1707 individuals). The user can run the function below or skip it and load the pre-computed results included in the package (ResQCAHS.rda). For convenience we do not display the returned matrices, however the obtained path model estimate is illustrated in Figure ?? and the significant gene-clinical pathway associations are extracted using the code below and reported in Table ??.

```
> #set.seed(4)
```

```
> #ResQCAHS=GSCA(QCAHS,W0, B0,latent.names=LatentVar, estim=TRUE,path.test=TRUE,path=NULL,nperm=1000)
```

```
> data("ResQCAHS")
```

```
> indices <- which(ResQCAHS$pvalues<0.05, arr.ind=TRUE)
```

```
> ind.row=indices[,1] ; ind.col=indices[,2]
```

```
> Significant<- matrix(rep(0,nrow(indices)*3),ncol=3);colnames(Significant)=c("Gene", "Pathway", "pval")
```

```
> Significant[,1] <- rownames(ResQCAHS$pvalues)[ind.row]
```

```
> Significant[,2] <- colnames(ResQCAHS$pvalues)[ind.col]
```

```
> Significant[,3]<-ResQCAHS$pvalues[indices]
```

```
> Significant
```

	Gene	Pathway	pval
[1,]	"CETP"	"Lipid metabolism"	"0.001"
[2,]	"APOE"	"Lipid metabolism"	"0"
[3,]	"LPL"	"Lipid metabolism"	"0.016"
[4,]	"PON2"	"Lipid metabolism"	"0.007"
[5,]	"PCSK9"	"Lipid metabolism"	"0"
[6,]	"ADIPO"	"Energy metabolism"	"0.043"
[7,]	"TNFa"	"Energy metabolism"	"0.002"
[8,]	"AGT"	"BP control"	"0.02"

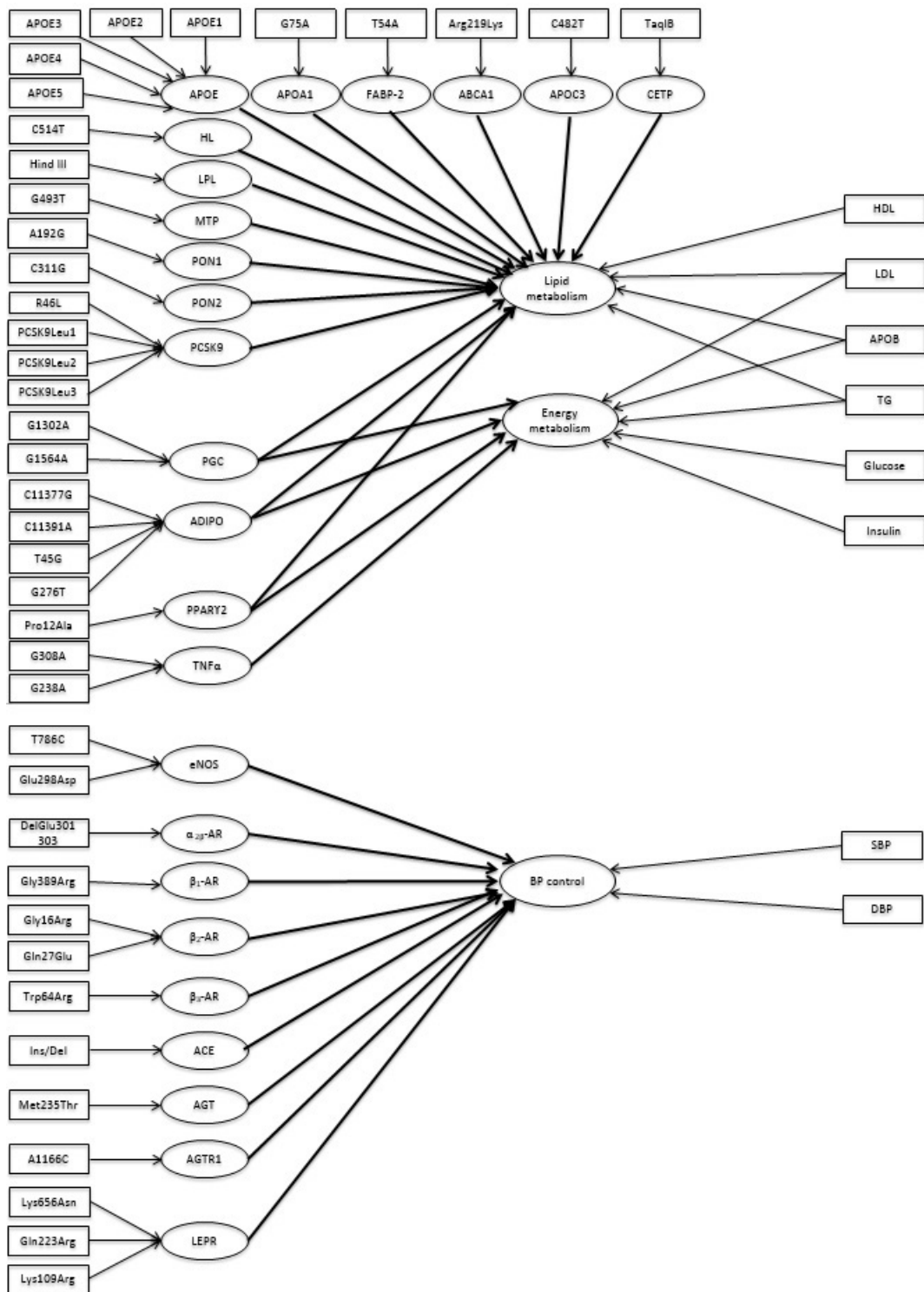


Figure 4: Path model for the QCAHS data.



Table 3: Results for QCAHS data analysis. A significant association (at level 5%) between a gene and a clinical pathway is indicated by "X".

Gene	Clinical pathways		
	Lipid	Energy	Blood pressure
CETP	X		
ApoE	X		
LPL	X		
PON2	X		
PCSK9	X		
Adiponectin		X	
TNF $\alpha$		X	
AGT			X

## References

- de Leeuw, J., Young, F. W., and Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, 471-503.
- Hwang, H. and Takane, Y. (2004). Generalized structured component analysis. *Psychometrika*, 69:81-99.
- Paradis, G., Lambert, M., O'Loughlin, J., Lavallee, C., Aubin, J., Berthiaume, P., Ledoux, M., Delvin, E., Levy, E., and Hanley, J. (2003). 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.
- Romdhani, H., Hwang, H., Paradis, G., Roy-Gagnon, M.-H. and Labbe, A. (2014). Pathway-based Association Study of Multiple Candidate Genes and Multiple Traits Using Structural Equation Models, submitted.