Stony Brook University

Statistical Methods for Association Analysis of Biological Data

A Dissertation Presented

by

Erya Huang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

In

Applied Mathematics and Statistics

Stony Brook University

May 2015

Stony Brook University

The Graduate School

Erya Huang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Wei Zhu – Dissertation Advisor
Professor, Deputy Chair, Department of Applied Mathematics and Statistics

Song Wu – Chairperson of Defense
Assistant Professor, Department of Applied Mathematics and Statistics

Xuefeng Wang, Ph.D. – Committee Member
Assistant Professor, Department of Applied Mathematics and Statistics

Wadie F. Bahou, MD – Committee Member
Professor, Department of Medicine

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Statistical Methods for Association Analysis of Biological Data

by

Erya Huang

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2015

Genome-wide association studies (GWA studies) are an important tool for identifying disease susceptibility variants for common and complex diseases. Traditional approaches to data analysis in GWA studies suffer with the multiple testing problem and also ignore any potential relationships between gene variants. We introduced here a novel two-stage framework with the combination of partial correlation network analysis (PCNA) and data mining techniques. This network-based technique, focusing on SNPs in joint modeling and their partial associations, alleviated the multiple testing problem and consequently increased the power to detect biologically relevant variants and their associations. Variable selection was achieved through penalized logistic regression using sparse-group lasso (SGL) penalty by grouping SNPs based on their: 1) pairwise canonical correlation measurement; or 2) biological information such as gene mapping. Network construction was based on pairwise partial correlation coefficients.

Simulation studies have indicated that this two-stage approach achieved high accuracy and a low false-positive rate in the identification of known individual and two-way association targets, which elucidated that it is possible to recover the true direct relationship even for high-dimensional situation. Subsequently, we illustrated the proposed approach in a search for potential significant SNP-SNP/gene-gene associations with nicotine dependence using a real data example from a GWA study conducted by the Washington University at St. Louis. The result would provide researchers potentially biologically relevant genetic networks for further investigation.

Another contribution of this thesis is the exploration of miRNA-mRNA regulatory set associated with essential thrombocytosis (ET) through the introduction of an application of penalized technique to canonical correlation analysis on microarray data sets. The identified variables were successfully tested by leave-one-out cross validation and a network exploration system.

Dedications

To my family, friends, and mentors who have supported me in my work. This work would not be possible without their constant encouragement.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to acknowledge the following people for their invaluable insight and commentary on this work, without which I would not have been able to clarify my thoughts and ideas.

Dr. Wei Zhu

Dr. Hongyan Chen

1. Introduction

1.1. Single Nucleotide Polymorphism (SNP)

Variations in the human genome can alter the risk of developing a disease. Among a variety of genetic polymorphisms, including copy number polymorphisms, inversions, and short tandem repeats, single nucleotide polymorphisms (SNPs, pronounced as 'SNiP') are the most common form of genetic variations, accounting for approximately 90% of human DNA polymorphisms (Collins, Brooks and Chakravarti 1998). An estimated 10 million SNPs are commonly occurring in the human genome (Norrgard 2008). By definition, a SNP represents a single base pair change in DNA sequence (*Figure 1*) (Onay, et al. 2006). Within a population, typically two different sequence alternatives correspond to the same SNP location. These alternatives are called alleles and such SNP is categorized as bi-allelic polymorphism. Scarcely, there are also tri- and tetra-allelic polymorphisms existing in human genome. The frequency of a SNP is defined as minor allele frequency (m.a.f.), referring to the frequency of the less common allele (Bush and Moore 2012). Strictly speaking, SNPs should be distinguished from rare variations, with the criterion that the m.a.f. of SNPs are 1% or larger. A combination of functionally relevant SNPs may additively or synergistically affect the intrinsic properties and the function of the proteins to a variable degree (Onay, et al. 2006). On the other hand, the

term mutation is referred to rare genetic variant (m.a.f. < 1%), which corresponds to obvious functional consequences on the protein level and ultimately leads to the disease state.

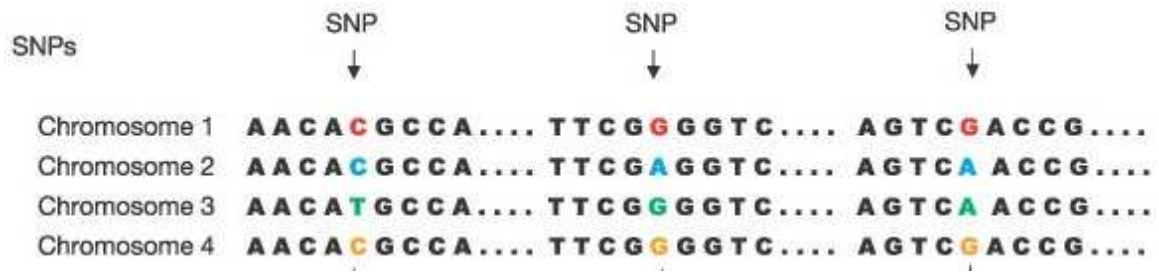


Figure 1. SNPs represent single base pair changes in DNA sequence (Gibbs, et al. 2003).

Mendelian disorders are typically regarded as diseases of largely genetic causation, in which disease phenotypes are mainly driven by rare genetic variants in a single gene locus (Lu, Latourelle, O'Connor, Dupuis and Kolaczyk 2013). Most Mendelian disorder diseases are rare, including Huntington's disease, Phenylketonuria, Cystic fibrosis, Sickle cell anemia, and Oculocutaneous albinism (MacDonald, et al. 1992, Chial 2008, Bush, et al. 2012). Because of its simple genetic architecture, Mendelian disorders follow an autosomal dominant or recessive inheritance pattern in families with the disease and the identification of the disease-causing mutation(s) in a single gene is relatively straightforward with the collection of sufficient family materials (Cho 2010). A typical strategy is called linkage analysis, in which a collection of genomic markers from the affected family are genotyped and the shared inheritance of genetic variants is linked to that of the disease phenotype. In 1989, linkage analysis has been successfully applied to the identification of missense multiple mutations within the *CFTR* (cystic fibrosis transmembrane regulator) gene as the main cause of Cystic fibrosis (Riordan, et al. 1989).

It is rapidly followed with subsequent successes in uncovering many disease-associated mutations for Mendelian disorders.

Apart from the rare Mendelian diseases, a majority of common diseases are complex diseases, including asthma, Autism, Alzheimer's, Type II Diabetes, and various types of common cancer (Stevenson 1992, Gatz, et al. 1997, Altshuler, et al. 2000, Easton, et al. 2007). Unlike Mendelian diseases, complex diseases arise as a result of various combinations of multiple factors, such as genetic, environmental and developmental factors (Freimer and Sabatti 2007, Stratton and Rahman 2007). Therefore, any individual SNP only accounts for a small to moderate contribution to the overall risk of a disease phenotype of interest (Bush, et al. 2012). Unsurprisingly, traditional analysis methods for Mendelian diseases, such as family-based linkage analysis, to identify individually important variant, may not be likely to succeed in genetic studies of complex diseases. This in turn, calls for population-based association studies or genome-wide association studies.

1.2. Genome-Wide Association Study (GWA Studies)

1.2.1. Association Study

The idea of association studies is to compare the allele frequency of an individual SNP, or a set of SNPs, in a cohort of unrelated cases to that of unrelated control subjects

conditioning on the confounding effects (gender, age, etc.). SNPs with higher frequency in the case cohorts are considered to be potentially associated with the disease phenotype of interest. Association study is claimed to be more powerful than linkage analysis in the capability of detecting lower penetrance alleles (Greenberg 1993, Hodge 1994, Risch and Merikangas 1996). Despite of some notable findings, early association studies (before 2005) have not achieved much success, limited by several crucial factors, such as the number of available polymorphisms. As a result of the following advances, interests in association studies have been renewed in the last few years.

1) The International HapMap Project

Collaborating participants from multiple countries including Japan, China, the United Kingdom, Canada, Nigera, and the United States, the International HapMap Project started in 2002, aiming to identify and localize genetic variants across the genome, to characterize correlations among variants, and to learn how the variants are distributed among people within or among populations from different parts of the world (Gibbs, et al. 2003). The project has since included eleven human populations of European, Asian, and African ancestry, and has genotyped 1.6 million common SNPs (Consortium 2010). The free HapMap information provided by the project for researchers from worldwide, greatly facilitates both the design and analysis of association studies, which will lead to the revolution of diagnosis, treatment, and prevention of diseases.

2) Development of genotyping technologies

Novel and improved genotyping technologies have been developed rapidly to efficiently and very accurately genotype genomic DNA from large numbers of individuals.

Chip-based microarray platforms primarily from Illumina (San Diego, CA) and Affymetrix (Santa Clara, CA) can assay one million or more SNPs (Bush, et al. 2012). The next-generation sequencing approaches are also available recently to provide all the DNA sequence variation in the genome. The development and mature of these technologies have and will continue substantially reducing the costs and increasing the rates for high-throughput SNP discovery, which makes genome wide association studies technically and financially feasible.

3) Sufficiently large number of participants

Appropriately large and well-characterized clinical samples have been assembled for many common diseases.

1.2.2. Genome-Wide Association Study

These significant accomplishments have made possible the extension of association studies to the whole genome level, which becomes genome-wide association study. Genome-wide association study (GWA study), also known as whole genome association study, is an important tool to systematically investigate the entire genome in large population (between case and control cohorts) in the effort to identify disease susceptibility polymorphisms for common and complex diseases (Burton, et al. 2007). GWA studies can be conducted without prior knowledge of position or function, and are thus one step beyond candidate gene studies, which study at most hundreds of variants selected based on limited biological pathway information (Pearson and Manolio 2008). In

the last decade, GWA studies have revolutionized in the field of genetics, enjoying plenty of successes in the search of genetic factors associating with common complex traits (Risch, et al. 1996, Arnaud-Lopez, et al. 2008, Easton and Eeles 2008, Lettre and Rioux 2008). According to the National Human Genome Institute Published Genome-wide Association Studies Catalog (www.genome.gov/gwastudies, accessed August 28, 2014), by the end of 2013, more than 12,000 SNPs associated with 17 trait categories have been identified by 1,778 published GWA studies. The ultimate goal of GWA studies is to identify SNPs for a better understanding of disease etiology, risk prediction, new leads for studying underlying biology of disease susceptibility for developing new prevention and treatment strategies (Kooperberg, LeBlanc and Obenchain 2010).

In most circumstances, the study design of GWA studies is the case-control design, which classifies individuals as a binary categorical variable – affected or unaffected. This design is relatively easier and less expensive than others. However, without well established, it can also lead to spurious association results, with the concern that the underlying population structure divergence (also referred to as population stratification), at genomic regions irrelevant to the disease of interest would result to the allele frequency difference (Chen 2011). The family-based trio design, which compares the frequencies of transmitted (to an affected offspring) and un-transmitted (from heterozygous parents) alleles, could address this concern (Cho 2010). Besides of qualitative traits, GWA studies have also gained successes in quantitative traits such as height (Weedon, et al. 2007) or high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol levels in heart disease (Teslovich, et al. 2010). This dissertation focuses exclusively on GWA studies in

case-control design. With this design, GWA studies typically share four common steps: (1) selecting sufficient large number of affected and unaffected subjects with the disease or trait of interest; (2) isolating and genotyping DNA for high genotyping quality data; (3) performing statistical analysis to test associations between SNPs and the disease or trait of interest; (4) verifying results by replicating the identified SNPs in an independent population or examining the functional implications via experiments (Pearson, et al. 2008).

The Common Disease/Common Variant (CD/CV) hypothesis was proposed based on the idea of complex diseases (Reich and Lander 2001) and is theoretically fundamental for GWA studies. As its name suggests, common diseases are caused by common genetic variations. According to this hypothesis, each related SNP would be merely slightly correlated to the prevalence of the disease (i.e. with small effect size or penetrance), and the overall genetic risk of the disease of interest would be spread across multiple genetic factors. Thus large sample size and a significant number of genetic markers are required for a significant finding. A conjugate of CD/CV is the Common Disease/Rare Variant (CD/RV) hypothesis, which postulates common disease is caused by multiple rare variants of moderate to large effect (*Figure 2*) (Zhou, Sehl, Sinsheimer and Lange 2010). Published in 2005, the first GWA study paper investigating on age-related macular degeneration (AMD; *Figure 3*) is cited as a supporting example of CD/RV (Klein, et al. 2005). With a relatively small sample size, this study has successfully identified SNPs within the complement factor *H* gene with exceptionally large effect size (odds ratio > 2) on risk for developing AMD. However in general, the CD/CV hypothesis

is true for most common diseases (typically with odds ratio range between 1.2 - 2) (Hindorff, et al. 2009).

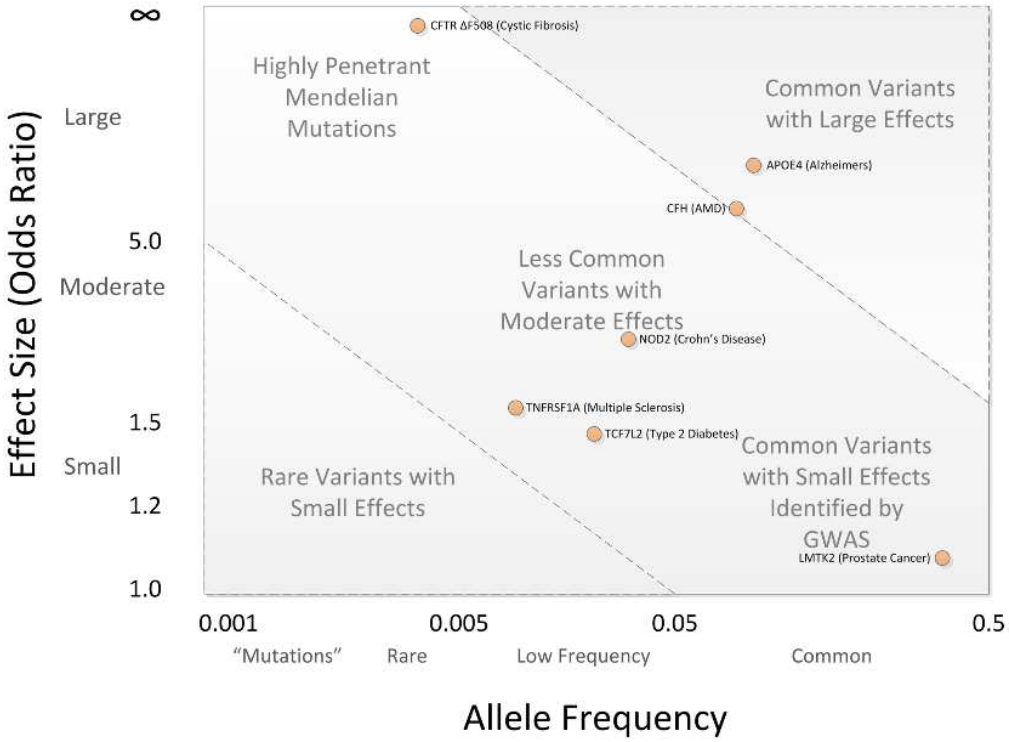


Figure 2. Spectrum of disease allele effects (Bush, et al. 2012)

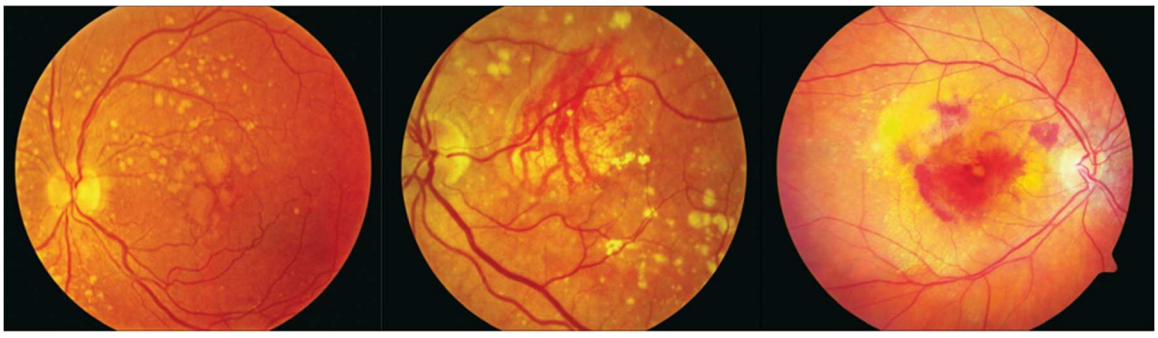


Figure 3. Three categories of AMD on the basis of the risk of developing vision loss. (Coleman, Chan, Ferris III and Chew 2008)

1.2.3. Association Tests

1.2.3.1. Single-locus Analysis

The typical data for GWA studies usually contains hundreds of thousands of SNPs and thousands of samples. While the number of SNPs has been increased greatly as a result of the development of genotyping technologies, it is impractical to expect the number of samples to be similarly expanded. As a result, one of the challenges of analyzing GWA study data is the “small n, large p” high-dimensional problem, which induces the naïve implementation (for example, incorporating all SNPs of interest in the standard methods) to be infeasible and undesirable (i.e. inversion of ill-definite sample covariance matrix or ill-posed least squares criterion in multivariate regressions).

Currently the standard approach to study the association between SNPs and the trait of interest in GWA studies is single-locus analysis, which examines each individual SNP independently at a time with a specified model, relating disease trait to the SNP and other potentially relevant covariates, and qualifies the significance of SNP via the p-value of an appropriate test. Case-control design GWA studies are usually analyzed using either contingency table method (with chi-square test of association, or Fisher’s exact test), or logistic regression model (with single SNP as the predictor). The latter one is usually more preferred by its ability to incorporate other covariates (Chen 2011). In either method, SNP can be coded as 0, 1, and 2, according to the number of minor alleles.

Analyses could be performed assuming a dominant, recessive or additive effect for each SNP. In the dominant model, it assumes that having one or more copies of the major alleles increases the risk compared to the minor allele. A recessive model combines variants with one or more minor alleles. An additive model assumes that there is a uniform, linear increase in the risk for each copy of major allele. (Onay, et al. 2006, Bush, et al. 2012))

Analyzing a typical GWA study using the single-locus analysis means that hundreds of thousands to millions of association tests will be conducted, and the conventional p-value criteria ($P < 0.05$) for statistical significance in a single test would be no longer suitable; otherwise, the cumulative likelihood of false positives (falsely detecting significant SNPs while in fact they are irrelevant) would be considerably high. This so-called multiple testing problem is counted as one of the challenging problems in GWA studies. Fortunately, this problem can be corrected to some degree. The simplest yet most conservative method is the Bonferroni correction (Bland and Altman 1995), which adjusts the level of significance α to be α/K (K is the total number of tests). However, the assumption of this correction -- each test to be mutually independent -- is generally unmet since linkage disequilibrium between SNPs can induce correlation between many of the tests. Moreover, the stringent threshold of significance level would result in the missing of many biologically relevant SNPs. Another approach is permutation testing. By randomly resampling the disease phenotypes with replacement from the original data and repeating a predefined large number of times, it generates an empirical distribution of test statistics under the null hypothesis, providing information on parameter

estimate and model selection (Onay, et al. 2006). Albeit straightforward, it is noticeably computationally intensive. False discovery rate (FDR) serves as another alternative. The idea was brought in by Benjamini and Hochberg in 1995 (Benjamini and Hochberg 1995) and is to estimate the proportion (usually 5%) of significant findings to be false positives (in another way, the proportion of errors among the rejected null hypotheses). These approaches have been widely applied to GWA studies (van den Oord 2008). The end result after the correction is a list of SNPs potentially significantly associated with the disease of interest, which in turn can be mapped to their closest genes (Lu, et al. 2013).

Despite of the success of single-locus analysis (Hindorff LA , Lu, et al. 2013), there are several potentially large problems. It is widely argued that this strategy often lacks the power to uncover the relatively small effect sizes of most genetic variants (Wang, Li and Hakonarson 2010). Neither does it adjust for correlation among SNPs (Lu, et al. 2013). Though corrections are applied, the multiple testing problem does not vanish. Additionally, this approach does not extend in a natural manner to search for interactions between variants and thus failed to explain entire underlying genetic variation in complex diseases (Kogelman and Kadarmideen 2014).

1.2.3.2. SNP-SNP Interactions

Realizing the limitations of traditional single-locus analysis, alternative approaches such as SNP-SNP interaction analyses have been developed recently. It has recently

been established that genes do not work alone; biological processes in the cell such as biochemical interactions and regulatory activities lead to complicated interaction patterns among genes and SNPs (Schäfer and Strimmer 2005). Therefore gene-gene/SNP-SNP interactions in molecular networks or pathways may have a great impact on unveiling the mechanism of complex diseases (Lin, et al. 2013). The motivation of SNP-SNP interaction analyses is to increase the power to detect disease-associated SNPs, and furthermore, to detect statistical interactions between loci that are informative about the biological and biochemical pathways that underpin the diseases (Cordell 2009).

The most common way to detect interactions is to fit a standard logistic regression model including all the main effects and relevant interaction terms on the log odds scale and test whether the interaction terms have significant effects (Cordell 2009). It is obvious that this model is limited by the high-dimensional problem. An alternative is to identify individual SNPs via single-locus analysis first and perform an exhaustive examine for all pair-wise SNP combinations between the chosen loci. Theoretically, methods such as chi-square test could be used in the exhaustive search to analyze all SNP-SNP interactions; however, the time required to perform such analysis increases exponentially with the number of variants analyzed (Li, et al. 2011). Another familiar exhaustive searching method for analyzing interaction effects is stepwise regression (Cordell and Clayton 2002). The forward stepwise selection starts with a model including all SNPs and covariates and search for the most significant interactions to enter into the model based on the score statistics; while backward elimination was done through the likelihood ratio test. Stepwise selection is popular yet suffers a number of failures too, i.e. overfitting or

instability due to sparsity of data. In short, exhaustive searching methods have successfully picked out potential disease-associated genes in some areas, e.g. breast cancer (Onay, et al. 2006); however, they are computationally intractable, even for highly efficient algorithms, and will encounter the multiple testing issue analogous to that in single-locus analysis.

Pathway-based approaches have also been recently developed on the grounds that, complex cellular pathways are often involved in disease susceptibility and functionally related genes can have coordinated gene expression patterns (Wang, et al. 2010). In this strategy, SNPs could be grouped together into SNP sets on the basis of certain biological criterion, i.e. gene mapping or pathway information. Then genome-wide test was performed on SNP sets, instead of individual SNPs. With the use of prior biological knowledge on gene function or pathways, the analyses of GWA studies would be facilitated to be more powerful and to have a better chance to identify genes and biological mechanisms. Moreover, by reducing the number of tests substantially compared to the single-locus analysis, the pathway-based approaches have kept the multiple testing problems in a benign form.

On top of the above methods, data mining techniques are considered as the mainstream in the current analysis approaches. They attempt to step through a particular sequence of regression models and to find the model that best fits the data (Cordell 2009). Allowing a large number of predictors, along with considering sparsity property of data,

the use of techniques such as penalized regression for identifying disease-associated SNPs and SNP-SNP interactions has been emerged. We will introduce and discuss these approaches in depth in the following chapters.

1.3. Goal

We introduced here a novel two-stage framework with the combination of partial correlation network analysis (PCNA) and data mining techniques. The proposed approach dealt with the multiple testing problem, along with the ignorance of potential relationships between gene variants by traditional methods to data analysis in GWA studies. The primary goal of the study was to achieve a reasonably sparse structure with the application of PCNA in GWA studies; and to thus extend PCNA for categorical/mixed variables to the high-dimensional arena. The approach would be tested through simulation studies, and be performed in a search for potential significant SNP-SNP/gene-gene associations with nicotine dependence in a GWA study.

1.4. Outline of the Dissertation

The rest of this dissertation was organized as follows. Chapter 2 introduced penalized regression models for categorical data, including penalties such as lasso, group lasso and sparse-group lasso. In Chapter 3, we presented our two-stage approach, and investigated its performance by means of simulation studies. Thereafter we illustrated the

method by analyzing COGEND data set, a GWA study data set in Chapter 4. Chapter 5 described the incorporation of sparse method in canonical correlation analysis and its application to microarray data for genetic regulatory network associated to essential thrombocytosis. We concluded the dissertation with a discussion of our methodology and future work.

2. Penalized Regression Methods

Analyzing SNPs together in a regression model enables us to consider the impact of SNPs among others. However, as introduced in the previous chapter, particularly in the situation of GWA studies, where there is the high-dimensional problem or where variables are highly correlated, traditional logistic regression model breaks down (Schäfer, et al. 2005). It is also believed that among the large number of SNPs, only a small proportion express under a certain set of conditions and affect the phenotype (Onay, et al. 2006). Thus we expect the coefficient vector to be sparse. Motivated by these challenges, penalized regression methods, such as the lasso (Tibshirani 1996) and the elastic net (Zou and Hastie 2005), were introduced and extended to offer an attractive and powerful alternative in GWA studies. They simultaneously carry out variable selection and provide coefficient estimates (Kooperberg, et al. 2010). There are different sparsity patterns of the penalty term. All of them shrink down the size of the coefficients of variables with little or none effects on the trait of interest with various degrees of constraints, some of which would coerce these coefficients to be zero. Penalized regression methods have prevailed conventional single-locus methods by the fact that it would yield fewer correlated variants (Ayers and Cordell 2010), be more powerful with a lower false discovery rate (He and Lin 2011), and could incorporate SNP-SNP interactions in a natural way (Lu, et al. 2013).

2.1. Penalized Linear Regression Methods

For a continuous trait, penalized regression methods minimize the sum of squared deviations of predicted values from observed ones with a penalty term. Suppose we have n samples and p predictors represented by an $n * p$ matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)^T$, where $\boldsymbol{X}_i = \left(x_{i1}, \dots, x_{ip}\right)^T$ as the predictor matrix and an $n * 1$ vector $\boldsymbol{Y} = (y_1, \dots, y_n)^T$ as the response variable. Assume the observations are independent and all x_{ij} are standardized (i.e. $\frac{\sum_i x_{ij}}{n} = 0,\ \sum_i \frac{x_{ij}^2}{n} = 1,\ i = 1,\ 2, \dots, n$), thus the interaction term β_0 is ignored. In ordinary linear regression model, estimates of coefficients are derived by

$$\hat{\beta} \ = \ arg\, \min_{\beta} (\frac{1}{2} \|y \ - \ X\beta\|_2^2 \ + \ \boldsymbol{P}(\boldsymbol{\beta}))$$

where β is the vector of coefficients, and $\|a\|_2^2 = \sum_j a_j^2$ is the L_2 norm. The first term in the above function ($\|y \ - \ X\beta\|_2^2$) represents the loss function minimized in ordinary least squares. The second term ($\boldsymbol{P}(\boldsymbol{\beta})$) is the penalty. The most well-known penalty is a L_1 penalty of the form

$$\boldsymbol{P}(\boldsymbol{\beta}) = \ \lambda \|\beta\|_1 = \ \lambda \sum\nolimits_j \left|\beta_j\right|$$

introduced by Tibshirani in 1996 and is called “least absolute shrinkage and selection operator”, or the lasso (Tibshirani 1996). Here $\lambda\ (> 0)$, the tuning constant, controls the strength of the penalty which constrains each β_j toward the origin and thus enforces sparse solutions. The regression based on the penalty with a squared L_2 penalty of the form

$$\boldsymbol{P}(\boldsymbol{\beta}) = \lambda \|\beta\|_2^2 = \lambda \sum\nolimits_j \beta_j^2$$

to the coefficients is called the ridge regression (Hoerl 1962, Hoerl and Kennard 1970). It has been claimed to be possible to distinguish causative from non-causative SNPs for quantitative traits (Malo, Libiger and Schork 2008) (*Figure 4*). A mixture of L_1 and L_2 penalties is called the elastic net (Zou, et al. 2005), which can be written as

$$\boldsymbol{P}(\boldsymbol{\beta}) = \lambda \left(\alpha * \|\beta\|_1 + (1-\alpha) * \|\beta\|_2^2\right) = \lambda \left(\alpha * \sum\nolimits_j \left|\beta_j\right| + (1-\alpha) * \sum\nolimits_j \beta_j^2\right)$$

It is obvious to tell that when α equals to one, the elastic net has reduced to the standard lasso and when α equals to zero, we have the ridge regression.

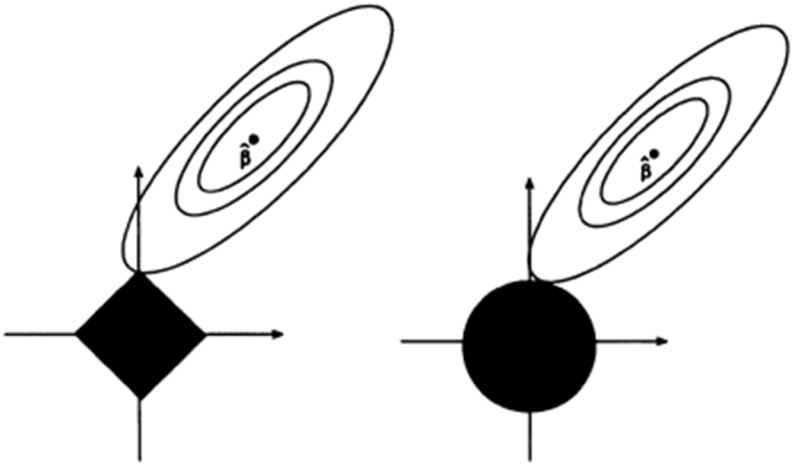


Figure 4. Geometric insight of the lasso (left) and the ridge (right) regression for two predictors (Tibshirani 1996).

To compare, the lasso encourages sparsity, selecting a subset of variables whose main effects best predict the response, and coercing the coefficients of other variables to be zero. It has been shown to outperform both subset selection and the ridge regression (Tibshirani 1996). However, its shortcomings include: a) will also impose heavy shrinkage

on large coefficients (Ayers, et al. 2010) and thus select unimportant variables to compensate this over-shrinkage (Huang, Breheny and Ma 2012), b) can only select at most n (sample size) nonzero variables (Ayers, et al. 2010), and c) when the variables are highly correlated or nearly linear dependent, the lasso tends to select only one of them at random (Bühlmann, Rütimann, van de Geer and Zhang 2013, Silver, et al. 2013). These drawbacks would be noteworthy especially in GWA studies, where SNPs are highly correlated or linearly dependent due to linkage disequilibrium (Balding 2006). For the ridge penalty, the estimates of coefficients have been only shrunk to small values, yet not been vanished (*Figure 4*). Thus it does not lead to sparse solutions. This difference is based on the fact that $|\beta_j|$ is much larger than β_j^2 for small β's (Wu, Chen, Hastie, Sobel and Lange 2009b). On the other hand, the elastic net is more stable than the lasso by the fact that it encourages groups of correlated variables to enter the model together, since the L_2 penalty form encourages similar coefficients for highly correlated variables (Ayers and Cordell 2013, Silver, et al. 2013). However, it is still unsatisfied of its ability of dealing with groups of variables with nearly linearly dependent, nor does it take linkage disequilibrium or biological information into account (Bühlmann, et al. 2013).

2.2. Penalized Logistic Regression Methods

Logistic regression model has been proven to be one of the most versatile techniques in the class of generalized linear models (Czepiel 2002). For a dichotomous response variable y_i (coded as 1 for cases and 0 for controls), the probability of being a case for subject i is

$$Prob(y_i = 1) = p_i = \frac{exp(X\beta)}{(1 + exp(X\beta))}$$

With the logit transform, that is, the natural logarithm of the odds ratio of p_i, the logistic regression function can be written as

$$log\left(\frac{p_i}{1 - p_i}\right) = \eta_i = X\beta$$

To estimate the coefficients, instead of minimizing the loss function to a sum of squares in the linear regression model, we maximize the likelihood function, which is

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

It is equivalent to maximize the log-likelihood function:

$$l(\beta) = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i)\log(1 - p_i) = \sum_{i=1}^{n} y_i \log\frac{p_i}{1 - p_i} + \log(1 - p_i)$$

$$= \sum_{i=1}^{n} y_i \eta_i - \log[1 + \exp(\eta_i)]$$

The idea of the lasso has also been applied to logistic regression model (Tibshirani 1996). The lasso penalized logistic regression model maximizes the above log-likelihood subject to an L_1 penalty which is dependent on the magnitude of the estimated parameters. Notice here that now we take the maximization, the penalty term is thus been subtracted. The derive of estimates of coefficients is written as

$$\hat{\beta}(\lambda) = \underset{\beta}{arg\,max}\big(l(\beta) - \boldsymbol{P}(\boldsymbol{\beta})\big) = \underset{\beta}{arg\,max}\left(\sum_{i=1}^{n} y_i\eta_i - \log[1+\exp(\eta_i)] - \lambda\|\beta\|_1\right) =$$

$$= \underset{\beta}{arg\,max}\,(\sum_{i=1}^{n} y_i\eta_i - \log[1+\exp(\eta_i)] - \lambda\sum\nolimits_j |\beta_j|)$$

The penalty constant λ can be tuned to give any desired number of predictors through methods such as cross validation. With a very large value of λ, there will be no variables selected in the model. As the value of λ decreases, number of variables entering the model would increase accordingly, with an order that is roughly determined by the impact of predictors on the response, except for correlated ones. Analogy to penalized linear regression, both the ridge and the elastic net penalties could be applied to logistic regression model by substituting the L_1 penalty term with an L_2 term and a mixture of L_1 and L_2 penalties, respectively.

An equivalent expression of the lasso penalized logistic regression model is to maximize

$$\underset{\beta}{max}\,(\sum_{i=1}^{n} y_i\eta_i - \log[1+\exp(\eta_i)])$$

under the constraint $\sum_j|\beta_j| \le s$. Like λ, s is also a user-specified parameter, which can be selected via model selection procedure, for example, cross-validation. Actually, there is a one-to-one correspondence between λ and s, i.e., if we have found $\hat{\beta}(\lambda)$, we can obtain s by $s = \sum_j|\widehat{\beta}_j(\lambda)|$ (Friedman, Hastie, Höfling and Tibshirani 2007).

2.3. The Group Lasso Penalty

Considering the fact that gene is a functional biological unit and variants within the same gene may affect the disease phenotype to a similar degree, it is not surprising that researchers would like to pursue a model fit for clustered or grouped variables. This shows the need of encouraging variables within the same group to enter a model together and encouraging sparsity between groups. Under the lasso and the ridge penalty, if a SNP is selected into the model, it does not strongly encourage or discourage another SNP within the same group -- located in the same gene or with high correlation -- from entering the model. It thus raises the discussion of how to select the penalty for a group of variables. Proposed by Yuan and Lin (Yuan and Lin 2006) for linear regression model and by Kim et al. (Kim, Kim and Kim 2006) and Meier et al. (Meier, Van De Geer and Bühlmann 2008) for logistic regression model, the group lasso penalty aims to address this problem by automatically including whole groups into the model if one variable amongst them is selected.

Suppose $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)^T$, where $\boldsymbol{X}_i = \left(x_{i1}, \dots, x_{ip}\right)^T$ is an $n * p$ genotype design matrix which can be divided into G groups and $\boldsymbol{Y} = (y_1, \dots, y_n)^T$, an $n * 1$ binary vector represents the response variable. Assume all observations are independent and standardized. If g indexes the G groups, the estimates of coefficients could be written as:

$$\hat{\beta}(\lambda) = arg\max_{\beta}(l(\beta) - \boldsymbol{P}(\boldsymbol{\beta}))$$

$$= arg\max_{\beta}\left(\sum_{i=1}^{n} y_i\eta_i - \log[1+\exp(\eta_i)] - \lambda\sum_{g=1}^{G}\sqrt{p_g}\left\|\beta^{(g)}\right\|_2\right)$$

where $\left\|\beta^{(g)}\right\|_2$ is the Euclidean (l_2) norm ($\left\|\beta^{(g)}\right\|_2 = \left(\sum_j {\beta_j^{(g)}}^2\right)^{1/2}$), applying in each group, which can be treated as an intermediate between the L_1 and L_2 penalty terms (*Figure 5*); p_g represents group size, that is, the number of variables in group g; and λ is the tuning parameter controlling the sparsity degree. The group lasso penalty term $\boldsymbol{P}(\boldsymbol{\beta})$ is a weighted sum of l_2 norms. The weight $\sqrt{p_g}$ could be replaced by other choices, allowing each group to be penalized to different extents; while itself penalizes large groups more heavily (Huang, et al. 2012). When $p_g = 1$ for all groups, the group lasso penalty will simplify to the standard lasso. It has been shown that under certain conditions, such as strong group sparsity, the performance of the group lasso penalty exceeds the standard lasso (Huang and Zhang 2010).

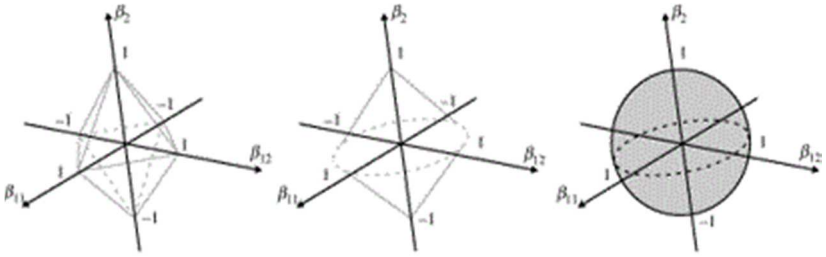


Figure 5. Illumination of penalty function of the standard lasso L_1 (left), the group lasso (center), and the ridge penalty L_2 (right) for two-variable continuous case (Yuan, et al. 2006).

There are many antecedents of incorporating penalized regression methods into GWA studies. To illustrate, the lasso penalized logistic regression model was used to pick

out significant SNP-SNP/gene-gene interactions in GWA study data (Wu, et al. 2009b, Lu, et al. 2013); the elastic net has been employed in GWA studies for the exploration of disease-causing SNPs (Cho, Kim, Oh, Kim and Park 2009); after being proposed by Yuan and Lin (Yuan, et al. 2006), the group lasso penalty has been applied for GWA study data to recognize pairwise interactions via hierarchical structure (Lim and Hastie 2013) and to identify pathways associated with quantitative traits of interest (Silver, Janousova, Hua, Thompson and Montana 2012).

2.4. The Sparse-group Lasso Penalty

Applying the group lasso means that once a variable is selected, all the other variables in the same group would also enter the model. If we are not interested in recognizing individual variables, group selection is a proper choice. However, in the field of genetics, if the predictors are biological molecules, such as genes or SNPs, researchers would like to identify not only particular gene groups or pathways that are closely related to the traits of interest, but also within the chosen groups, the particular "standouts" that play a more important role than their "group mates". For this reason, they would embrace the idea of selecting significant individual variables together with important groups, which leads to a mixture of the group and the lasso penalties.

In 2010, Friedman et al. (Friedman, Hastie and Tibshirani 2010) have briefly proposed a method combining the lasso with the group lasso penalty in an unpublished

note, which can be directly applied on non-orthonormal model matrices. Recently, Simon et al. (Simon, Friedman, Hastie and Tibshirani 2013) have continued this work, discussing this method, the so-called "sparse-group lasso" (SGL), in details and extended it to generalized regression models. The sparse-group lasso (SGL) penalty is one of the comparatively recent developments in sparse modelling. Simulation studies demonstrate that SGL accurately imposes the dual-level (between group and within group) sparsity pattern when comparing to both the group lasso penalty and the standard lasso (Friedman, et al. 2010, Simon, et al. 2013).

The sparse-group lasso method has integrated the lasso (L_1) and the group lasso (l_2) penalties to allow considerations of sparse effects both on the group-wise and within group levels. Consider the similar notation as the group lasso, that $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)^T$ is an $n * p$ genotype design matrix which can be divided into G groups, $\boldsymbol{Y} = (y_1, \dots, y_n)^T$, an $n * 1$ binary vector represents the response variable, and g indexes the G groups, sparse estimate for the coefficient vector is given by

$$\hat{\beta}(\lambda) = arg\min_{\beta}(-l(\beta) + \boldsymbol{P}(\boldsymbol{\beta}))$$

$$= arg\min_{\beta}(-l(\beta) + (1-\alpha)\lambda \sum_{g=1}^{G} \sqrt{p_g} \left\|\beta^{(g)}\right\|_2 + \alpha\lambda\|\beta\|_1)$$

where $l(\beta) = \sum_{i=1}^{n} y_i\eta_i - \log[1 + \exp(\eta_i)]$ is the log-likelihood function of logistic regression model; $\beta^{(g)}$ is the corresponding coefficient vector for $X^{(g)}$ in group g ($X^{(g)}$ is the sub-matrix of X containing only variables from group g); p_g is the number of variables in group g; both $\alpha \in [0, 1]$ and $\lambda > 0$ are parameters controlling sparsity. The group lasso

penalty (in Euclidean/l_2 form) enforces the sparsity at the group level; while the standard lasso (in L_1 form) enforces sparsity at individual level within selected groups (Silver, et al. 2013, Simon, et al. 2013).

Though looked similar to the elastic net penalty, the Euclidean form penalty in SGL does not differentiate at 0; however, within each non-zero group, it gives an elastic net fit (Simon, et al. 2013). SGL gains its popularity over other regular penalized regressions by the fact that, in SGL, both the group and the lasso penalties could improve the convergence rate in minimizing the objective function, and they are both compatible with coordinate descent for a fast optimization (Zhou, et al. 2010).

3. Two-stage Approach

3.1. Partial Correlation

Correlation is a widely used concept describing how two variables are related between mean values. This concept was first created by Francis Galton in 1888 to related measurements under different conditions (Stigler 1989). The degree of correlation is measured by correlation coefficient, which is denoted by ρ (pho) for a population and by r for a sample. One of the most familiar measurements is Pearson product-moment correlation coefficient (or Pearson's r). Defined in terms of moment, it was introduced by Karl Pearson to measure the linear dependence between two variables (Pearson 1920). The value of Pearson's r is always between -1 and 1. The geometrical interpretation of Pearson's r can be considered as the cosine of the angle between the two vectors in Euclidean space, each of which forms by pointing from the variable mean to the origin point (Fisher 1924). By Fisher transformation, correlation coefficients would approximately follow a normal distribution and can thus perform hypothesis tests and calculate confident intervals; otherwise, bootstrap resampling (Efron and Tibshirani 1994) could act as an alternative.

As it is well agreed, a high correlation coefficient between two variables in a network system may be indicative of three potential situations: (1) directly interacted; (2)

indirectly interacted; and (3) regulated by a common variable (*Figure 6*). However, to explore the interactions between biological molecules, the investigators are primarily interested in the direct interaction solely. Therefore in this section we would like to introduce the partial correlation coefficient (Yule 1907, Pearson 1915, Fisher 1924), which focuses specifically on the measurement of the strength of direct interaction.

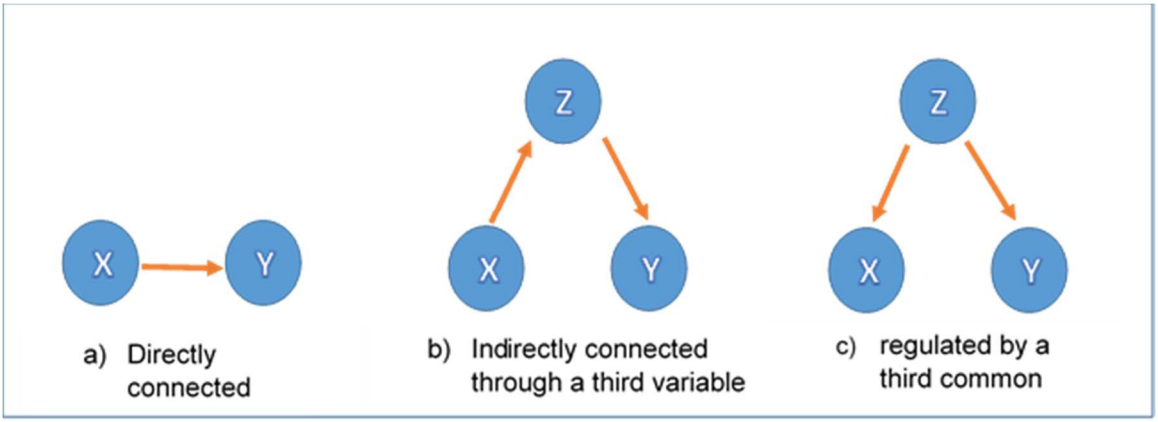


Figure 6. Potential relationships between variables X and Y with a high correlation coefficient. In the situations depicted in the center and right-hand side of the figure, the partial correlation coefficients of X and Y given Z would be zero (assuming ideal experimental conditions). Hence partial correlation coefficient prevents false positives due to indirect, rather than direct effects between two variables.

The partial correlation coefficient measures the degree of dependence between two random variables (e.g. gene activities) while controlling on the effects from one or several other variables. For example, the correlation $r_{ij.p}$ between variables X_i and X_j conditioning on X_p is the correlation between the parts of X_i and X_j that are uncorrelated with X_p. Under the assumption of normality, a partial correlation coefficient equals to zero if and only if the two variables are conditionally independent given the remaining variables. With a large set of variables, investigators may be of interest of the direct interaction relationships between all possible pairs. Proper statistic allows us to 1) measure the

strength of a relationship (i.e. the magnitude of a partial correlation coefficient); 2) determine whether a relationship is significant; and 3) compare the relationship of the same pair of variables between different groups. This is accomplished via different techniques considering the types of variables (e.g. continuous, categorical, or mixed data).

3.1.1. Partial Correlation with Continuous Data

Given p continuous random variables $\{X_i, i = 1, 2, \dots, p\}$, each of which has n samples, denote the set of variables as

$$X = (X_1, X_2, \dots \dots, X_P)^T \in \mathcal{R}^{n*p}$$

The rows of the matrix represent the samples and the columns represent the variables. Within each column (variable), the data are centered to the column mean. For any two random variables X_i and X_j, denote the set of all other variables as $X_{-(i,j)}$, i.e.

$$X_{-(i,j)} = X \setminus \{X_i, X_j\} = \{X_k, 1 \leq k \neq i, j \leq p\}$$

where X_i and $X_j \in \mathcal{R}^n$ are the ith and jth columns of X and $X_{-(i,j)} \in \mathcal{R}^{n*(p-2)}$ is the matrix obtained from X by deleting the ith and jth columns. Without loss of generality, we assume $i < j$. When the sample size (n) is larger than the number of variables (p), the standard estimate of partial correlation coefficient of X_i and X_j while controlling the effects of variables in the set of $X_{-(i,j)}$, can be calculated via three different methods.

The first method is achieved by "matrix inversion" (Schäfer, et al. 2005) and can be accomplished in the computation time of $\mathcal{O}(n^3)$. Denote the covariance matrix of X as $\Sigma = \left(\sigma_{ij}\right)_{p*p}$, which can be further decomposed into the variance components σ_i^2 and the Pearson correlation matrix $P = \left(r_{ij}\right)_{p*p}$. Since the data are column-centered, the estimate of covariance matrix Σ is obtained as

$$\widehat{\Sigma'}_{p*p} = (X - EX)^T(X - EX) \ = X^T X$$

where ^{T} denotes the transpose of a matrix and thus $X^T X$ is the inner product of X itself, that is, the sum of squares of all elements in X. The standard unbiased estimate of Σ is then given by

$$\hat{\Sigma}_{p*p} = \frac{1}{n-1}\widehat{\Sigma'}_{p*p} = \frac{1}{n-1}X^T X$$

In the setting of $n > p$, the above *p*-by-*p* matrix is symmetric and positive-semidefinite. If $\hat{\Sigma}$ is invertible, denote the precision (or concentration) matrix $\widehat{\Omega}$ as the inverse of $\hat{\Sigma}$ such that

$$\widehat{\Omega} = \left(\widehat{\omega}_{ij}\right)_{p*p} = \hat{\Sigma}^{-1}$$

Therefore, an unbiased estimate of partial correlation coefficient of X_i and X_j giving $X_{-(i,j)}$ is estimated as

$$\hat{\rho}_{ij} = -\frac{\widehat{\omega}_{ij}}{\sqrt{\widehat{\omega}_{ii}\widehat{\omega}_{jj}}}$$

Another simple way to compute partial correlation coefficient is by least square regression. Consider the two linear regression models

$$X_i = X_{-(i,j)}\beta^{(i)} + \varepsilon_i = \sum_{k \neq i,j} \beta_k^{(i)} X_k + \varepsilon_i$$

$$X_j = X_{-(i,j)}\beta^{(j)} + \varepsilon_j = \sum_{k \neq i,j} \beta_k^{(j)} X_k + \varepsilon_j$$

where ε_i and ε_j are the i.i.d. noises. The intercept term is not included in either model since all the variables are centered. The least square estimates of $\beta_k^{(i)}$ and $\beta_k^{(j)}$ can be obtained by solving the optimization problems of

$$\hat{\beta}^{(i)} = \left(\hat{\beta}_1^{(i)}, \hat{\beta}_2^{(i)}, \dots, \hat{\beta}_{i-1}^{(i)}, \hat{\beta}_{i+1}^{(i)}, \dots, \hat{\beta}_{j-1}^{(i)}, \hat{\beta}_{j+1}^{(i)}, \dots, \hat{\beta}_p^{(i)}\right)$$

$$= \arg \min_{\beta \in \mathcal{R}^{p-2}} \left\| X_i - X_{-(i,j)}\beta \right\|^2$$

$$\hat{\beta}^{(j)} = \left(\hat{\beta}_1^{(j)}, \hat{\beta}_2^{(j)}, \dots, \hat{\beta}_{i-1}^{(j)}, \hat{\beta}_{i+1}^{(j)}, \dots, \hat{\beta}_{j-1}^{(j)}, \hat{\beta}_{j+1}^{(j)}, \dots, \hat{\beta}_p^{(j)}\right)$$

$$= \arg \min_{\beta \in \mathcal{R}^{p-2}} \left\| X_j - X_{-(i,j)}\beta \right\|^2$$

$\|a\|_2^2 = \sum_j a_j^2$ is the L_2 norm, indicating the sum of squared elements of the matrix. The corresponding residuals are

$$\widehat{R_i} = X_i - X_{-(i,j)}\hat{\beta}^{(i)} = X_i - \sum_{k \neq i,j} \hat{\beta}_k^{(i)} X_k$$

$$\widehat{R_j} = X_j - X_{-(i,j)}\hat{\beta}^{(j)} = X_j - \sum_{k \neq i,j} \hat{\beta}_k^{(j)} X_k$$

The correlation between residuals is a measurement of the strength of the relationship between X_i and X_j when the effect of $X_{-(i,j)}$ has been removed and can thus represent the sample partial correlation coefficient

$$\hat{\rho}_{ij} = Corr\left(\widehat{R}_i, \widehat{R}_j\right).$$

A third version of the estimation of partial correlation coefficient also relates to the least square regression problem (Peng, Wang, Zhou and Zhu 2009). Construct p linear regression models

$$X_i = X_{-(i)}\beta^{(i)} + \varepsilon = \sum_{k \neq i} \beta_k^{(i)} X_k + \varepsilon, i = 1, 2, \dots, p$$

where ε are i.i.d. disturbance terms. The least square estimate of the regression coefficient vector is calculated as

$$\hat{\beta}^{(i)} = \left(\hat{\beta}_1^{(i)}, \hat{\beta}_2^{(i)}, \dots, \hat{\beta}_{i-1}^{(i)}, \hat{\beta}_{i+1}^{(i)}, \dots, \hat{\beta}_p^{(i)}\right) = \arg \min_{\beta \in \mathcal{R}^{p-1}} \left\|X_i - X_{-(i)}\beta\right\|^2$$

$$= \left(X_{-(i)}^T X_{-(i)}\right)^{-1} X_{-(i)}^T X_i, \quad for\ i = 1, 2, \dots, p$$

The sample partial correlation coefficient is then estimated as

$$\hat{\rho}_{ij} = sign\left(\hat{\beta}_j^{(i)}\right)\sqrt{\hat{\beta}_j^{(i)}\hat{\beta}_i^{(j)}}$$

Given $n > p$, the two coefficient vectors $\hat{\beta}_j^{(i)}$ and $\hat{\beta}_i^{(j)}$ always have the same sign and thus the term of square root in the above equation is well-defined. Based on the above formula, the process of searching for non-zero partial correlation coefficients is equivalence to the model selection problem under the regression setting.

The distribution of the sample partial correlation for continuous variables was described by Fisher (Fisher 1924). Assuming the original data of all variables come from a multivariate Gaussian distribution. It states that the random sampling distribution of a partial correlation coefficient controlling k variables, is exactly that of a total correlation coefficient with k fewer degrees of freedom. Thus, we can test the null hypothesis that the population partial correlation coefficient vanishes via an F-test:

$$F = \frac{\hat{\rho}_{ij}^2}{1 - \hat{\rho}_{ij}^2} * (N - k - 2) \sim F_{1,N-k-2}$$

where N is the sample size and k is the number of variables being controlled. Similarly, Fisher transformation can also be used:

$$Z = \frac{1}{2} \ln\left(\frac{1 + \hat{\rho}_{ij}}{1 - \hat{\rho}_{ij}}\right) \dot{\sim} N(0, 1)$$

There is no exact tests for the comparison of the equality between two partial correlation coefficients. For data with sufficiently large sample size, some methods of approximation are known. One of the most widely used is Fisher transformation (Fisher 1915). To compare two population partial correlation coefficients $\rho_{ij}^{(1)}$ and $\rho_{ij}^{(2)}$ conditioning on k variables, draw an independent sample from each of the population with sample size n_1 and n_2 respectively, and calculate the sample partial correlation coefficients $\hat{\rho}_{ij}^{(1)}$ and $\hat{\rho}_{ij}^{(2)}$. The test statistics of the null hypothesis that two population partial correlation coefficients $\rho_{ij}^{(1)}$ and $\rho_{ij}^{(2)}$ are equal is obtained as

$$Z = \frac{{}^{1}/_{2}\left(\ln\left(\frac{1+\hat{\rho}_{ij}^{(1)}}{1-\hat{\rho}_{ij}^{(1)}}\right) - \ln\left(\frac{1+\hat{\rho}_{ij}^{(2)}}{1-\hat{\rho}_{ij}^{(2)}}\right)\right)}{\sqrt{\frac{1}{n_1-k-3}+\frac{1}{n_2-k-3}}} \dot{\sim} N(0,1)$$

As has been noted previously, the bootstrap resampling method (Efron, et al. 1994) is a widely used non-parametric alternative for data failed to meet normality assumption. To determine whether two partial correlation coefficients are equal, one could perform either of the following: 1) to generate a bootstrap confidence interval for each sample partial correlation coefficient and see whether the two confidence intervals overlap; or 2) to bootstrap the difference between two partial correlation coefficients and see whether the bootstrap confidence interval of the difference contains zero.

Pradhan (Pradhan 2009) proposed a two-level regression method to convert the test to that of a regression coefficient. Denote the binary covariate as $G = \{0,1\}$, which effect on the pair of partial correlation coefficients is of interest. In the first step, two residual terms (prediction errors) $\widehat{R_\iota}$ and $\widehat{R_j}$ are obtained via linear regressions of X_i and X_j on $X_{-(i,j)}$ respectively.

$$X_i = X_{-(i,j)}\beta^{(i)} + \varepsilon_i$$

$$X_j = X_{-(i,j)}\beta^{(j)} + \varepsilon_j$$

$$\widehat{R_\iota} = X_i - X_{-(i,j)}\hat{\beta}^{(i)} = X_i - \sum_{k \neq i,j} \hat{\beta}_k^{(i)} X_k$$

$$\widehat{R_j} = X_j - X_{-(i,j)}\hat{\beta}^{(j)} = X_j - \sum_{k \neq i,j} \hat{\beta}_k^{(j)} X_k$$

The test of the Pearson correlation coefficient between two residuals gives the same significance to that of the slope coefficient in the linear regression model of $\widehat{R_\iota}$ and $\widehat{R_\jmath}$.

$$\widehat{R_\iota} = a_0 + a_1\widehat{R_\jmath} + \varepsilon$$

Or equivalently,

$$\widehat{R_\jmath} = c_0 + c_1\widehat{R_\iota} + \tau$$

Integrate covariate G into the above regressions brings the second stage of the model:

$$\widehat{R_\iota} = a_0 + a_1\widehat{R_\jmath} + \varepsilon = a_0 + (b_0 + b_1 G)\,\widehat{R_\jmath} + \varepsilon = a' + b_0\widehat{R_\jmath} + b_1 G\widehat{R_\jmath} + \varepsilon$$

$$\widehat{R_\jmath} = c_0 + c_1\widehat{R_\iota} + \tau = c_0 + (d_0 + d_1 G)\,\widehat{R_\iota} + \tau = c' + d_0\widehat{R_\iota} + d_1 G\widehat{R_\iota} + \tau$$

Therefore, the significance of coefficients b_1 and d_1, (i.e. the average p-value from tests of significance for b_1 and d_1) represents the significance of covariate effect G on the partial correlation coefficient between X_i and X_j controlling on $X_{-(i,j)}$.

3.1.2. Partial Correlation with Categorical and Mixed Data

While partial correlation and its corresponding properties and statistic are well defined for continuous variables, its application based upon categorical and mixed data has not yet been thoroughly investigated. Recently, a new estimate of partial correlation was proposed aiming at a solution to this problem (Chen 2011). Unlike other methods such as partial phi coefficient and partial polychoric correlation, this new estimate method is innovative and superior in the aspects of being capable to easily control for more than

one variables and also embracing both categorical and continuous variables simultaneously in the analysis (Chen 2011, Leong 2012).

Details of this new method are as follows. First consider the two-categorical case. Suppose we have p binary random variables $\{X_i, i = 1, 2, \dots, p\}$. To estimate the partial correlation coefficient between X_i and X_j when controlling on $X_{-(i,j)}$, two logistic regression models are performed with $X_{-(i,j)}$ as predictor variables and X_i and X_j as response variables respectively.

$$\pi_k = Prob\big(X_k = 1 | \, X_{-(i,j)}\big) = \frac{e^{X_{-(i,j)}\beta^{(k)}}}{1 + \, e^{X_{-(i,j)}\beta^{(k)}}} = 1 - Prob(X_k = 0 | \, X_{-(i,j)}), \quad k = i, j$$

$$logit\left(\frac{\pi_i}{1-\pi_i}\right) = \, X_{-(i,j)}\beta^{(i)}$$

$$logit\left(\frac{\pi_j}{1-\pi_j}\right) = \, X_{-(i,j)}\beta^{(j)}$$

Pearson residuals are obtained as

$$\hat{R}_i = \frac{X_i - \, \hat{\pi}_i}{\sqrt{\hat{\pi}_i \, (1 - \, \hat{\pi}_i)}}; \;\; \hat{R}_j = \frac{X_j - \, \hat{\pi}_j}{\sqrt{\hat{\pi}_j \, (1 - \, \hat{\pi}_j)}}$$

And the sample partial correlation coefficient is the conventional Pearson correlation coefficient between the two residuals

$$\hat{\rho}_{ij} = \, Corr\left(\hat{R}_i, \hat{R}_j\right).$$

The estimate method can be easily extended to a multi-categorical case. Consider variables with three classes as an example. Running two independent binary logistic models provides two sets of generalized residuals.

$$\begin{cases} logit\left(\frac{\pi_{i1}}{\pi_{i0}}\right) = X_{-(i,j)}\beta^{(i1)} \\ logit\left(\frac{\pi_{i2}}{\pi_{i0}}\right) = X_{-(i,j)}\beta^{(i2)} \end{cases}; \quad \begin{cases} logit\left(\frac{\pi_{j1}}{\pi_{j0}}\right) = X_{-(i,j)}\beta^{(j1)} \\ logit\left(\frac{\pi_{j2}}{\pi_{j0}}\right) = X_{-(i,j)}\beta^{(j2)} \end{cases}$$

For each observation x_{kz}, assign

$$x_{kz} = \begin{cases} 1, & if\ x_z\ belongs\ to\ category\ z \\ 0, & otherwise \end{cases}, z = 1, 2; k = 1, 2, \dots, p$$

Residuals are calculated as follows:

$$\begin{cases} \hat{R}_{i1} = \frac{X_{i1} - \widehat{\pi_{i1}}}{\sqrt{\widehat{\pi_{i1}}\,(1 - \widehat{\pi_{i1}})}} \\ \hat{R}_{i2} = \frac{X_{i2} - \widehat{\pi_{i2}}}{\sqrt{\widehat{\pi_{i2}}\,(1 - \widehat{\pi_{i2}})}} \end{cases}; \quad \begin{cases} \hat{R}_{j1} = \frac{X_{j1} - \widehat{\pi_{j1}}}{\sqrt{\widehat{\pi_{j1}}\,(1 - \widehat{\pi_{j1}})}} \\ \hat{R}_{j2} = \frac{X_{j2} - \widehat{\pi_{j2}}}{\sqrt{\widehat{\pi_{j2}}\,(1 - \widehat{\pi_{j2}})}} \end{cases}$$

The first canonical correlation coefficient between residual sets $\{\hat{R}_{i1},\ \hat{R}_{i2}\}$ and $\{\hat{R}_{j1},\ \hat{R}_{j2}\}$ can define the partial correlation coefficient between X_i and X_j. Canonical correlation test is readily available for significant partial correlation coefficient.

For mixed variables, the estimate could be determined similarly. Assume X_i is a three-class categorical variable, and also introduce continuous variable X_j into the model. For X_i, we have

$$\begin{cases} logit\left(\frac{\pi_{i1}}{\pi_{i0}}\right) = X_{-(i,j)}\beta^{(i1)} \\ logit\left(\frac{\pi_{i2}}{\pi_{i0}}\right) = X_{-(i,j)}\beta^{(i2)} \end{cases} \Rightarrow \begin{cases} \hat{R}_{i1} = \frac{X_{i1} - \widehat{\pi_{i1}}}{\sqrt{\widehat{\pi_{i1}}\,(1 - \widehat{\pi_{i1}})}} \\ \hat{R}_{i2} = \frac{X_{i2} - \widehat{\pi_{i2}}}{\sqrt{\widehat{\pi_{i2}}\,(1 - \widehat{\pi_{i2}})}} \end{cases}$$

where for each observation in X_i

$$x_{iz} = \begin{cases} 1, & if\ x_i\ belongs\ to\ category\ z \\ 0, & otherwise \end{cases}, \qquad z = 1, 2$$

and for X_j,

$$X_j = X_{-(i,j)}\beta^{(j)} \implies \widehat{R_j} = X_j - X_{-(i,j)}\hat{\beta}^{(j)}$$

The partial correlation coefficient between X_i and X_j is thus the first canonical correlation coefficient between residuals $\{\hat{R}_{i1},\ \hat{R}_{i2}\}$ and $\widehat{R_j}$; and an ANOVA F-test on the significance of regressing $\widehat{R_j}$ on $\hat{R}_{i1}$ and $\hat{R}_{i2}$ would be able to detect significant partial correlation coefficient (Chen 2011).

3.1.3. Partial Correlation Network Analysis (PCNA)

Given a group of multiple variables of interest, partial correlation coefficients distinguish direct from indirect interactions among all potential pairs. A corresponding partial correlation network analysis (PCNA) is a functional tool to represent graphically this direct interaction relationship. Formally, PCNA is an undirected graph, denoted by $G = \{V, E\}$, where the set V contains p nodes corresponding to p variables and the edge set E describes the conditional dependency relationship among $X_1, X_2, \ldots, X_p$ (Yuan and

Lin 2007) (*Figure* 7). An edge between two nodes X_i and X_j is present if and only if two corresponding variables are conditionally dependent on the other variables and such dependence does not have directional information (Chen 2011, Wang, Chao and Hsu 2011), i.e.

$$E = \{(i,j) | \, X_i \; and \; X_j \; are \; conditionally \; dependent; i,j \; \in V, 1 \leq i < j \leq p\}$$

$$X_i \perp X_j \, | \, X_{-(i,j)} \Leftrightarrow X_j \perp X_i \, | X_{-(i,j)}.$$

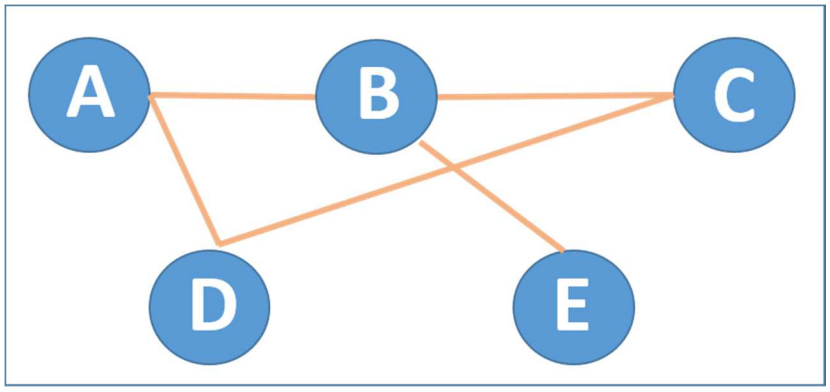


Figure 7. Partial correlation network analysis (PCNA). V = {A, B, C, D, E}. Two unconnected variables are not partially correlated and thus independent from each other when controlling the effects from all the other variables, i.e. B is independent to D given {A, C, E}.

A partial correlation network is popular for investigators by the fact that 1) the existence of an edge represents the significance of the corresponding partial correlation coefficient and 2) the strength of an direct interaction, if exists, can be measured by the magnitude of the partial correlation coefficient. It is informative compared to other methods (i.e., hierarchical clustering), which examine only the marginal pairwise correlations (Wang, et al. 2010).

Regular methods of PCNA such as exhaustive search do not account for the sparsity property and/or high-dimension problems, that is $n \ll p$, of the networks and thus would not work properly in these situations, i.e. the precision matrix is not unique in the high-dimensional case. Lately, a number of solutions are available focusing on PCNA with continuous variables, which application is computationally feasible and have successfully identified biologically meaningful genetic networks. Naïve improvements have been proposed by Fuente et al. (2004) to systematically testing all pair-wise correlations without conditioning first, then by conditioning on all other individual variables and subsequently on all possible pairs; in each step edges with non-significant coefficients are removed (De La Fuente, Bing, Hoeschele and Mendes 2004); and also by Schäfer and Strimmer (2005) for a shrinkage covariance estimation procedure (Schäfer, et al. 2005). However, as noted by Li and Gui (2006), neither of them accounted for sparsity during network estimation. Meinshausen and Bühlmann further developed a computationally attractive yet variable-by-variable approach, neighborhood selection with the lasso, which has been shown as an approximation to the exact problem by later studies (Meinshausen and Bühlmann 2006, Yuan, et al. 2006, Friedman, Hastie and Tibshirani 2008). Li and Gui (2006) have adopted a threshold gradient descent (TGD) regularization in the estimation of precision matrix (Li and Gui 2006); Yuan and Lin (2007) have introduced penalized maximum likelihood to the estimation of precision matrix with the constraint of positive definiteness, yet could not handle high-dimensional data (Yuan, et al. 2006); and Friedman et al. (2008) have presented the graphical lasso to use a coordinate descent procedure for the lasso with an impressive computational time

(Friedman, et al. 2008). In 2009, Peng et al. have further contributed to this problem by introducing an extension of algorithm for solving penalized optimization problems (Peng, et al. 2009).

It is noteworthy that so far sparse methods of PCNA mainly focus on continuous variables. It gives rise to needs of innovative extensions of PCNA when categorical variables are involved.

3.2. Two-stage Approach

When applying to the field of genetics, partial correlation estimates the degree of dependence between individual SNP and another SNP, or continuous/ categorical covariate while effects from other variables being controlled. PCNA performed graphically based on pairwise partial correlations and would be informative for pathway findings. However to date, this powerful tool, PCNA for categorical/mixed data proposed by Chen (Chen 2011), has not yet been directly applied to GWA studies for significant SNP-SNP/gene-gene associations. With the aim of achieving a reasonably sparse structure for SNP-SNP associations most related to disease of interest, and to thus extend PCNA for categorical/mixed variables to the high-dimensional arena, we described in this section a novel sequential analysis, combining variable selection process via SGL and connection identification stage through PCNA with categorical data. This approach enables us to

detect highly interconnected SNPs/genes that may work cooperatively in a pathway, and refines the multiple testing problem of traditional analysis methods in GWA studies.

Figure 8 shows the workflow of this two-stage approach. The main objective of this method is to firstly bring down dimension of variables and thus simultaneously identifies disease-susceptible SNPs with focus on variant association or biological information, and then to develop an association network in order to detect any potential genetic biomarkers, patterns and pathways for complex diseases and traits.

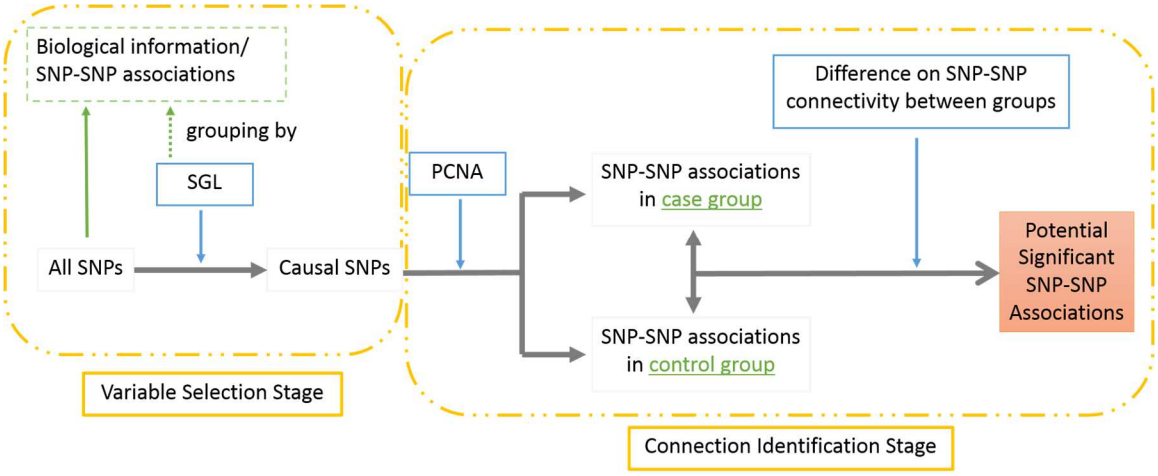


Figure 8. Workflow of two-stage approach.

The approach starts with variable selection accomplished through penalized regression using sparse estimation principles. Specifically, sparse-group lasso is performed on the data set as dimension reduction to detect any potential disease-susceptible SNPs. We also would like the estimation to allow for the use of biological knowledge or variant association in fitting the model. Details of the setting and processing are as follows.

3.2.1. Selection of Parameters

For the two-dimensional parameter (α, λ), α, as a convex combination of the lasso and the group lasso penalties, controls how the sparsity constraint is distributed between them ($\alpha = 0$ gives the group-lasso fit and $\alpha = 1$ gives the standard lasso fit) and λ regulates the degree of sparsity. When $\alpha \in [0, 1]$, as α approaches to 0, the model encourages greater sparsity at the group level than at the within group level. Thus, it is recommended to fix the value of α according to the reality of the problem (Simon, et al. 2013), i.e. $\alpha = 0.95$ would work for problems expected with strong overall sparsity and encouraged grouping; and $\alpha = 0.05$ would be preferred to problems with strong group-wise and medium or small within group sparsity. Considering the (biological) property of GWA study data, it is reasonable to set $\alpha = 0.95$ with the expectation of strong overall sparsity in our approach.

The value of λ can be tuned to select a user-predetermined number of SNP variants and/or other predictors. Reducing the value of λ would relax the penalty and hence encourage more variables entering the model. The number of non-zero variables selected by the model is generally a decreasing function of λ. Therefore once a variable been selected into the model, it will usually remain as λ decreases. In the situation of no preference in predetermined number of SNPs, the value of λ can be allowed to vary freely and be optimized by K-fold cross-validation process. To start with, λ was set to be large

enough (denoted as λ^{max}) that all estimates are zero. Decreasing the value of λ along a grid of values until some small proportion of λ^{max} results in a path of solutions, from which an optimal λ can be chosen by K-fold cross validation. In K-fold cross validation, we randomly divided the data into K equal-sized groups, leave out group k at a time, fit the model on data from the other $(K-1)$ groups (combined), and estimate parameters for the leave-out kth group. The testing errors would be averaged across all the K groups. Given different values of λ, the cross-validation curve $c(\lambda)$ represents the averaged testing error for each λ. The one with the smallest testing error would be the best value of λ. In our study, the default value for K is 10, i.e. the value of λ is selected via a 10-fold cross validation process.

3.2.2. Basis of Grouping

Grouping structure of variables could be decided according to different study goals and/or the availability of information. In GWA studies, to take advantage of prior biological/pathway information, SNPs mapped to the same gene can be clustered into one group (Huang, et al. 2012).

When biological information is unavailable or insufficient, clustering method is always a popular alternative. Variables with high associations are expected to be in the same group and to have low associations with those from different groups. Here we built the grouping structure of SNPs based on hierarchical clustering with a newly-proposed

pairwise association measurement – the canonical correlation measurement. Bühlmann et al. (Bühlmann, et al. 2013) have previously introduced canonical correlation for clustering variables in linear model and claimed that canonical correlation reflects the linear dependence among variables and thus addressed the identifiability problem. For categorical variables, Chen (Chen 2011) has also proposed using canonical correlation as pairwise association measurement.

Consider each SNP as a three-category variable – C/C, T/C, and T/T, with polynomial coding,

$$\text{C/C: } S = 0,\ S^2 = 0$$

$$\text{T/C: } S = 1,\ S^2 = 1$$

$$\text{T/T: } S = 2,\ S\text{^}2 = 4$$

The pairwise association between two SNPs X_i and X_j is defined as the first canonical correlation coefficient between $\left(\{S_{xi}, S_{xi}^2\},\ \{S_{xj},\ S_{xj}^2\}\right)$, which has been shown to be positively related to chi-square statistics and thus imply association strength.

$$Associaiton\left(X_i,\ X_j\right) = First\ CanCor\left(\{S_{xi},\ S_{xi}^2\},\ \{S_{xj},\ S_{xj}^2\}\right)$$

We thus applied the pairwise (first) canonical correlation coefficient as the dissimilarity matrix input in hierarchical clustering for SGL grouping basis. Notice that this measurement would not be affected by different coding schemes.

3.2.3. Algorithm

The traditional algorithm to fit the regular (unpenalized) logistic regression model is Newton's method (also known as the Newton-Raphson method). Recall that the log-likelihood function, along with its score function and Hessian matrix is expressed as follows:

$$l(\beta) = \sum_{i=1}^{n} (y_i \eta_i - \log[1 + \exp(\eta_i)]);$$

$$\nabla\, l(\beta) = \sum_{i=1}^{n} x_i\, (y_i - \eta_i);$$

$$\mathrm{d}^2\, l(\beta) = \sum_{i=1}^{n} x_i^2\, \eta_i\, (1 - \eta_i)$$

and the Newton's update of the estimates is

$$\beta^{(n+1)} = \beta^{(n)} - \frac{\nabla\, l\left(\beta^{(n)}\right)}{\mathrm{d}^2\, l(\beta^{(n)})}.$$

Fast and reliable, the Newton's method is a popular choice for low-dimensional problems welcomed by most statisticians (Wu, et al. 2009b); however, it is claimed to be computationally uncompetitive for high-dimensional problems (Zhou, et al. 2010).

The sparse-group lasso is fitted using blockwise descent: optimizes the penalized function with respect to a single group at a time, and cycles through disjoint groups until

convergence. For each group k, fix the coefficients of the other groups and estimate $\beta^{(k)}$ via an "accelerated generalized gradient descent algorithm with backtracking" to minimize

$$-l(\beta) + (1-\alpha)\lambda\left\|\beta^{(k)}\right\|_2 + \alpha\lambda\left\|\beta^{(k)}\right\|_1$$

The term $\sqrt{p_l}$ is suppressed and can be simply added back by replacing all future $(1-\alpha)\lambda$ by $\sqrt{p_l}\,(1-\alpha)\lambda$. The "accelerated generalized gradient descent algorithm" has introduced to the generalized gradient algorithm a momentum term proposed by Nesterov (Nesterov 2007) to improve the algorithm to $O(1/\sqrt{\epsilon})$, where ϵ is the convergence threshold, and a step size (t) optimization. The details of the algorithm idea are as follows (Simon, et al. 2013). First, denote the unpenalized negative log-likelihood function as

$l_k\left(\beta^{(-k)}, \beta^{(k)}\right) =$

$$-\frac{1}{n}\sum_{i=1}^{n} y_i\left(x_i^{(-k)^T}\beta^{(-k)} + x_i^{(k)^T}\beta^{(k)}\right) - \log\left(1+\exp\left(x_i^{(-k)^T}\beta^{(-k)} + x_i^{(k)^T}\beta^{(k)}\right)\right);$$

Apply the majorization minimization (MM) scheme to the function by

$$l_k\left(\beta^{(-k)}, \beta^{(k)}\right) \le l_k\left(\beta^{(-k)}, \beta_0\right) + (\beta - \beta_0)^T\nabla l_k\left(\beta^{(-k)}, \beta_0\right) + \frac{1}{2t}\|\beta - \beta_0\|_2^2$$

where β_0 is an initial point. Add the penalty terms and set the goal to find $\hat{\beta}$ optimizing the following function

$$M(\beta) = l_k\left(\beta^{(-k)}, \beta_0\right) + (\beta - \beta_0)^T\nabla l_k\left(\beta^{(-k)}, \beta_0\right)$$

$$+\frac{1}{2t}\|\beta - \beta_0\|_2^2 + (1-\alpha)\lambda\left\|\beta^{(k)}\right\|_2 + \alpha\lambda\left\|\beta^{(k)}\right\|_1$$

$$= \frac{1}{2t} \left\| \beta - (\beta_0 - t\nabla l_k(\beta^{(-k)}, \beta_0)) \right\|_2^2 + (1-\alpha)\lambda \left\| \beta^{(k)} \right\|_2 + \alpha\lambda \left\| \beta^{(k)} \right\|_1$$

Therefore we get that

a) $\hat{\beta} = 0$ if

$$\left\| S\left(\beta_0 - t\nabla l_k\left(\beta^{(-k)}, \beta_0\right), t\alpha\lambda\right) \right\|_2 \leq t(1-\alpha)\lambda$$

where $S(a,\ b) = sign(a)(|a| - a * b)_+$.

b) Otherwise $\hat{\beta}$ satisfies the updated formula for the generalized gradient step:

$$\hat{\beta} = \left(1 - \frac{t(1-\alpha)\lambda}{\|S(\beta_0 - t\nabla l_k(\beta^{(-k)}, \beta_0), t\alpha\lambda)\|_2}\right)_+ S\left(\beta_0 - t\nabla l_k\left(\beta^{(-k)}, \beta_0\right), t\alpha\lambda\right).$$

The right-hand side of the equation is denoted as $U\ (\beta_0, t)$ in the following description

$$U\ (\beta_0, t) = \left(1 - \frac{t(1-\alpha)\lambda}{\|S(\beta_0 - t\nabla l_k(\beta^{(-k)}, \beta_0), t\alpha\lambda)\|_2}\right)_+ S\left(\beta_0 - t\nabla l_k\left(\beta^{(-k)}, \beta_0\right), t\alpha\lambda\right).$$

An overview of the algorithm can thus be represented as:

1) (Group wise) Cycle through disjoint groups; at each group k, fix the coefficients of the other groups and execute the following steps;
2) Check if $\hat{\beta}^{(k)}$ satisfies the condition below. If yes, set $\hat{\beta}^{(k)} = 0$ for the whole group; if not, continue to the next step.

$$\left\| S\left(\frac{X^{(k)^T}\eta_{(-k)}}{n}, \alpha\lambda\right) \right\|_2 \leq (1-\alpha)\lambda;$$

where $\eta_{(-k)} = \frac{\exp(X^{(-k)^T}\beta^{(-k)})}{1+\exp(X^{(-k)^T}\beta^{(-k)})}$.

3) (Within group) Start with $\beta^{(k,l)} = \theta^{(k,l)} = \beta_0^{(k)}, t = 1$, and $l = 1$ (θ is the center; β_0 is the initial center point). Iterate the following steps until reach convergence:

   a. Update the score function by $g = \nabla l_k\left(\beta^{(-k)}, \beta_0^{(k)}\right)$;

   b. Optimize step size t by repeating $t = 0.8 * t$ until

$$l_k\left(U(\beta^{(k,l)}, t)\right) \leq l_k\left(\beta^{(k,l)}\right) + g^T\Delta_{(\mathrm{l,t})} + \frac{1}{2t}\left\|\Delta_{(\mathrm{l,t})}\right\|_2^2;$$

   c. Update $\theta^{(k,l)}$ by

$$\theta^{(k,l+1)} \leftarrow U\left(\beta^{(k,l)}, t\right);$$

   d. Update $\beta^{(k,l+1)}$ by

$$\beta^{(k,l+1)} \leftarrow \theta^{(k,l)} + \frac{l}{l+3}(\theta^{(k,l+1)} - \theta^{(k,l)});$$

   e. Set $l = l + 1$.

where $\Delta_{(\mathrm{l,t})} = U\left(\beta^{(k,l)}, t\right) - \beta^{(k,l)}$ is the difference between the old and new estimates.

Following the sparse-group lasso, PCNA is performed on the selected SNPs from SGL. An edge connecting two nodes (variables) represents significant pairwise partial correlation estimate. Two networks are created in case and control groups respectively. Within each group, multiple testing problem has been taken into full consideration in the view of the fact that false discovery rate (FDR) was assessed for the p-values from

significant tests and has been controlled at 0.05 (α). The process is as follows: with $n = \binom{14}{2}$ null hypotheses, rank the n corresponding p-values from smallest to largest to have $\{p_{(1)}, \ldots, p_{(n)}\}$; for a given level of significance $\alpha \geq 0$, compared $p_{(i)}$ with $\left(\frac{i}{n}\right) * \alpha$ starting from $p_{(1)}$ and stop whenever it comes to an integer j satisfied with $p_{(j)} \geq \left(\frac{i}{n}\right)\alpha, 1 \leq j \leq n$; we then can reject the hypotheses $H_{(j)}, \ldots, H_{(n)}$ at an FDR of α or better. Through the comparison of the existence of edge between the same pair of variables in two networks, difference of connectivity is considered as potential significant SNP-SNP association related to the trait of interest. When biological information is available, the PCNA result could further indicate information on potential gene-gene associations or biological pathways/patterns.

Traditional approaches in GWA studies usually consider SNPs in isolation, testing them one by one at a time and thus would be likely to fail to capture the inherent relationship among variants. Many approaches also ignore any potential functional relationships between variants. Our two-stage sequential analysis approach analyzes the variables together in the model to allow the consideration of the effect of one variable to another, which has been discussed to improve power in certain situations (Ayers, et al. 2010). Moreover, it takes account of factors specific to genome-wide data sets, such as variant association and/or biological information. In the following section, we further tested the approach through simulation studies.

3.3. Simulation

The performance of our proposed method is assessed in simulation studies. Simulation data were generated using statistical programming language R 2.15.3 (http://www.r-project.org/) with package – *scrime* (http://cran.r-project.org/web/packages/scrime/index.html). A data matrix containing $N\ (=1000)$ observations and $P\ (=50\ or\ 1000)$ SNPs was simulated given the selected minor allele frequencies (m.a.f.) of SNPs. All SNPs are firstly simulated mutually independently and thus unlinked. In the next step, the binary response variable Y is determined. For each observation, the probability of $Y=1$ is calculated via a logistic regression model, with the pre-specified significant individual SNPs and SNP-SNP interactions as predictors. The value of the response is then determined by a random draw from a Bernoulli distribution using this probability (Schwender and Fritsch 2008).

The simulation studies address two scenarios (*Table 1*). Both settings involve an $N*P$ SNP data matrix and an $N*1$ case/control status indicator y (y = 1 for case group and $y=0$ for control group). For both settings, the sample size is fixed at $N=1000$. The total numbers of SNPs are set to be $P=50$ and $P=1000$, representing low- and high-dimensional situations respectively. For each SNP, the minor allele frequency (m.a.f.) is drawn from a uniform distribution $Uni(0.2, 0.4)$. Significant individual SNPs are set to be SNP 2, 3, 7, and 8; as well as two pairwise SNP-SNP interactions: SNP2*SNP3 and SNP7*SNP8. Replications $(s=50\ or\ 1000)$ in each scenario are performed to average

out variability due to random sampling as recommended by Krämer et al. (Krämer, Schäfer and Boulesteix 2009).

Table 1. Simulation Settings

	Scenario 1	**Scenario 2**
Number of Data sets (s)	1000	50
Total Number of SNPs (P)	50	1000
Sample Size (N)	1000	1000
Minor Allele Frequency (m.a.f.) of Each SNP	drawn from a uniform distribution of [0.2, 0.4]	
Significant Items	Individual SNPs: SNP2, SNP3, SNP7, SNP8; SNP-SNP interactions: SNP2*SNP3; SNP7*SNP8.	

Our two-stage approach is performed on both scenarios, while controlling false positive rate to be 5%. Since gene mapping information is not available, we only consider grouping variables with dissimilar matrix based on pairwise canonical correlations (*Figure 9*). The performance is determined and compared by the difference between the detected and true targets, which is measured in terms of true positive rate (TPR) and positive predictive rate (PPR). We here defined the true positive rate (TPR) as the ratio of the number of times the true targets been identified out of the total number of simulations. In each simulation, we also take the record on whether all detected individual targets are true ones and define positive predictive rate (PPR) as the ratio of the number of times when all detected individual targets are true ones over the total number of simulations.

$$True\ positive\ rate = \frac{\#\ of\ times\ \{true\ targets\ are\ detected\}}{\#\ \{replications\ of\ simulations\}};$$

$$Positive\ predictive\ rate = \frac{\#\ of\ times\ \{all\ detected\ targets\ are\ true\ ones\ \}}{\#\ \{replications\ of\ simulations\}}$$

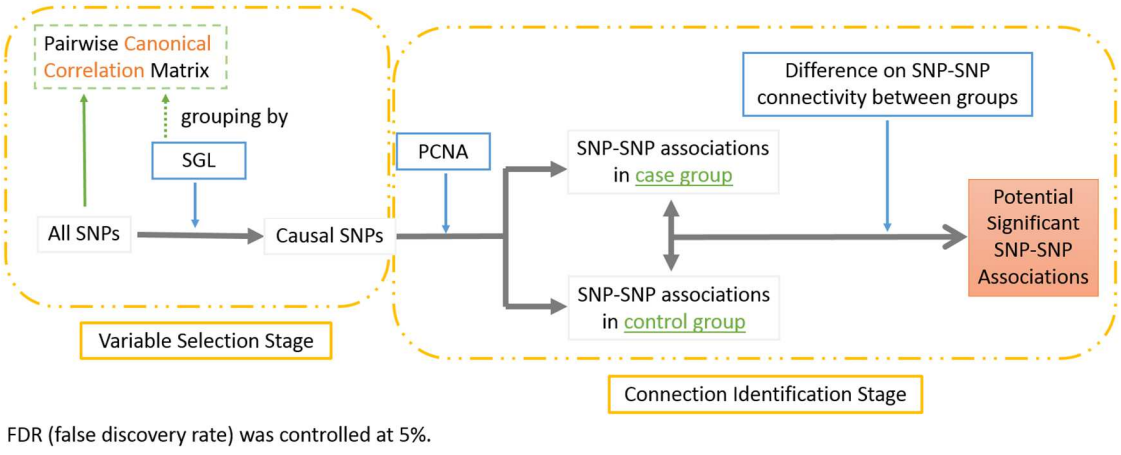


Figure 9. Workflow of two-stage approach applied on simulation data.

For comparison, a penalized forward-stepwise logistic regression model was also fit on the same simulation data using R package – *stepPLR* (Park, Hastie and Park 2009). In the model, a L_2 penalized term is added to the log-likelihood function:

$$\max_{\beta}(l(\beta) - \lambda\|\beta\|_2^2)$$

Bayesian information criterion (BIC) is used to guide the forward growing and λ is set to a default value as $\lambda = 1e-4$. Maximum number of terms to be added in the forward selection procedure is set as 10. The model allows interactions to enter if at least one of the main effects are present. The performance of penalized stepwise logistic regression model was compared to that of the two-stage approach (*Table 3-4*).

Table 2 summarized the performance of our two-stage approach in simulation data with both low- and high-dimensional settings. In the low-dimensional scenario, the two-stage approach has a high performance in identifying the true targets. Each of the four true individual SNPs has been successfully detected more than 80% of times, with the highest being 92.2%, out of total 1000 simulations. For SNP-SNP interactions, given the fact that PCNA can only functioned when there are more than two variables, we only focus on the simulations which more than two SNPs are identified by SGL. There are 821 out of 1000 satisfying this criteria. Out of these 821 simulations, 92.4% have identified the true SNP-SNP interaction between SNP 2 and 3, and ~90% (89.3%) have identified interaction between SNP 7 and 8. The comparatively high TPR comes at the prize of rather low positive predictive rate. 25.8% out of 1000 simulations have detected only individual targets that are true ones.

In high-dimensional scenario, the two-stage approach exhibits a little differently comparing to in low-dimensional setting. The TPRs of true individual targets are relatively low, with the highest one being only 52%. However, among the 14 (out of 50) simulations that more than two variables are spotted by SGL, the two-stage approach has successfully recognized interaction SNP2*SNP3 with an 85.7% TPR and SNP7*SNP8 with a 71.4% TPR. Though slightly lower than those in low-dimensional setting, the TPRs of SNP-SNP interactions for high-dimensional scenario are still satisfying. Surprisingly, 96% of the simulations have pinpointed only true individual targets. We thus claimed that in the high-dimensional situation, although the two-stage approach performed

conservatively in the identification of true individual variables, yet it preserved the similarly and impressively high performance in detecting true SNP-SNP interactions and a notably low false discovery rate when compared to that in low-dimensional setting.

The exciting performance of the two-stage approach was confirmed by the comparison with the stepwise penalized logistic regression model. Considering the TPRs in low-dimensional scenario, the two-stage approach performed as well as, if not better than, the stepwise penalized logistic regression model for individual variables. While the latter model recognized the true interactions with merely 23.1% and 11.6% TPRs, the proposed method has greatly surpassed it by TPRs three to six times higher concerning total 1,000 replications (*Table 3*).

Table 2. Result summary of the two-stage approach in simulation data

		Scenario 1	Scenario 2	Summary
Number of Simulations (s)		1,000	50	
Total Number of SNPs (P)		50	1,000	↑
Sample Size (N)		1,000	1,000	
True Positive Rate	**SNP2**	92.2% (out of 1,000)	52% (out of 50)	↓
	SNP3	81.7% (out of 1,000)	26% (out of 50)	
	SNP7	81.5% (out of 1,000)	32% (out of 50)	
	SNP8	80.3% (out of 1,000)	30% (out of 50)	
	SNP2_vs_SNP3	92.4% (out of 821)	85.7% (out of 14)	≈
	SNP7_vs_SNP8	89.3% (out of 821)	71.4% (out of 14)	≈
Positive Predictive Rate		25.8% (out of 1,000)	96% (out of 50)	↑

Table 3. Result summary of two methods in Scenario 1

Scenario 1		Two-stage	StepPlr
Number of Simulations (s)		1,000	
Total Number of SNPs (P)		50	
Sample Size (N)		1,000	
True Positive Rate	**SNP2**	92.2% (out of 1,000)	100% (out of 1,000)
	SNP3	81.7% (out of 1,000)	74.4% (out of 1,000)
	SNP7	81.5% (out of 1,000)	79.9% (out of 1,000)
	SNP8	80.3% (out of 1,000)	79.0% (out of 1,000)
	SNP2_vs_SNP3	92.4% (out of 821) 75.9% (out of 1,000)	23.1% (out of 1,000)
	SNP7_vs_SNP8	89.3% (out of 821) 73.3% (out of 1,000)	11.6% (out of 1,000)

* Two-stage: two-stage approach; StepPlr: penalized stepwise logistic regression.

With regards to the high-dimensional setting, the stepwise penalized model has recognized the individual true SNP with more than 50% of times (even reach 100% for SNP2); nevertheless, the TPRs of SNP-SNP interactions were considerately low, at 26% and 6% for SNP2*SNP3 and SNP7*SNP8, respectively. On the other hand, as noted previously, the two-stage approach did not have high TPRs for true individual targets, but has successfully detected SNP-SNP interactions in most of the simulations with more than two significant individuals identified (*Table 4*). Even considering the full situation with all 50 replications, these TPRs were still comparable to those of stepwise penalized model. We did not consider PPR here since for stepwise penalized model, the number of selected variables was user-specified, rendering the comparison of PPR less meaningful.

Table 4. Result summary of two methods in Scenario 2

Scenario 2		**Two-stage**	**StepPlr**
Number of Simulations (s)		50	
Total Number of SNPs (P)		1,000	
Sample Size (N)		1,000	
True Positive Rate	**SNP2**	52% (out of 50)	100% (out of 50)
	SNP3	26% (out of 50)	56% (out of 50)
	SNP7	32% (out of 50)	64% (out of 50)
	SNP8	30% (out of 50)	70% (out of 50)
	SNP2_vs_SNP3	85.7% (out of 14) 24.0% (out of 50)	26% (out of 50)
	SNP7_vs_SNP8	71.4% (out of 14) 20.0% (out of 50)	6% (out of 50)

* Two-stage: two-stage approach; StepPlr: penalized stepwise logistic regression.

To summarize, analysis results in simulation data indicated that, our two-stage approach performed well in finding true main effects and interactions in low-dimensional scenario. The fact that this approach had an impressively high (96%) positive predictive rate of true individual variables and satisfying true positive rates of SNP-SNP interactions particularly in high-dimensional situation is very encouraging to its application in the field of genetics for the exploration of biologically-relevant genomic associations. When compared with the stepwise penalized logistic regression model, the two-stage approach has notably outperformed it, especially with respect to SNP-SNP interactions.

4. Application to GWA Studies

Smoking has produced negative health and economic burdens. According to a report from US Department of Health and Human Services published in 2014 (Health and Services 2014), smoking is responsible for more than 480,00 deaths and $289 billion cost. Though a majority of smokers would like to quit, yet few are successful. Addiction to nicotine, a naturally occurring alkaloid found in tobacco, is considered as the main reason. The GWA study COGEND (Collaborative Genetic Study of Nicotine Dependence) was initiated in 2001, aiming to detect biological mechanisms, genes and environmental factors associated with heavy tobacco consumption, nicotine dependence, and related phenotypes (NCBI , COGEND 2013). Subjects in COGEND study age from 25 to 44. Case subjects are smokers defined as nicotine dependence (with a FTND (Fagerström Test for Nicotine Dependence) score of 4 or greater) and control subjects are smokers (smoked at least 100 cigarettes lifetime) who never had any symptoms of dependence (with lifetime FTND = 0) (COGEND 2013). The COGEND data used in this dissertation is from Dr. Laura Bierut's group, which is a subset of COGEND data set. It contains 2022 subjects (1114 cases and 908 controls). Totally 215 SNPs are included in the data set, locating in eight chromosomes (*Table 5*). SNPs are coded as (0, 1, 2) according to the number of minor alleles, whichever is identified with lower frequency in population.

Table 5. Chromosome information of SNPs

Chromosome	1	2	4	8	11	15	17	20
# of SNPs	4	17	10	37	4	112	20	11

The data were analyzed by our two-stage approach, which results were compared to that from the penalized stepwise logistic regression model. Since for real data the ground truth (i.e. the true underlying associations and networks) is unknown, the performance of different methods could not be directly compared. We thus focused on the similarities and dissimilarities of the results from the investigated methods.

4.1. Two-stage Approach

Our approach can be applied to the COGEND data set by grouping SNPs based on their: 1) pairwise canonical correlation measurement; or 2) biological information such as gene mapping (*Figure 10*). We denoted them as "Method A" and "Method B" respectively in the following description.

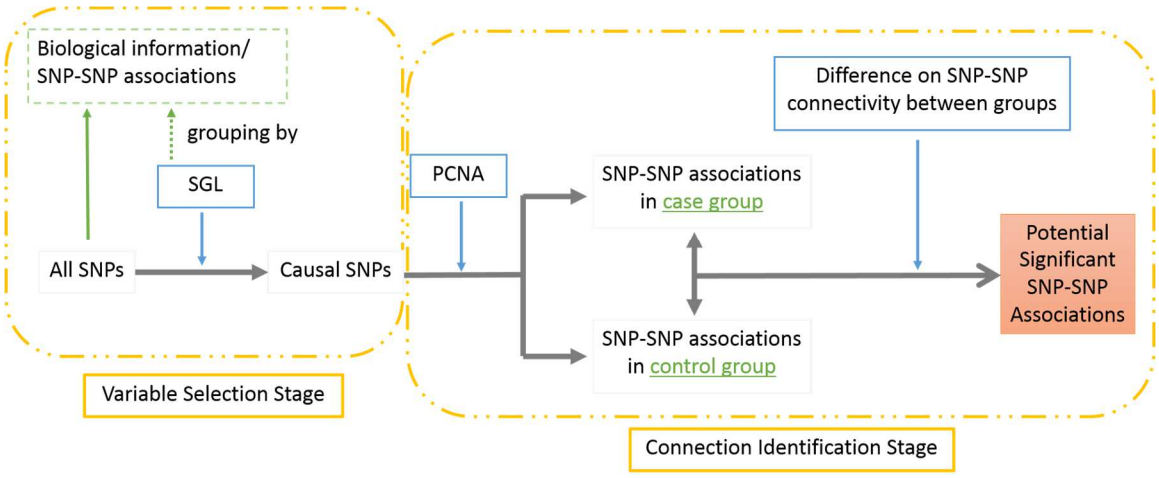


Figure 10. Workflow of two-stage approach in COGEND data analysis.

4.1.1. Method A: Grouping by Pairwise Canonical Correlation Measurements

Clustering was performed according to the dissimilarity matrix input based on pairwise canonical correlation measurements. R package – *DynamicTreeCut* (Langfelder, Zhang and Horvath 2008) was implemented to determine the clusters with criteria of threshold to be at least 10 SNPs per group. This procedure resulted in 11 clusters plus Cluster0 for outliers (containing 7 SNP that did not meet the clustering criteria; *Table 6*). As discussed in (Chen 2011), the clustering result based on canonical correlation measurement is extremely similar to those based on linkage disequilibrium and Cramér's V and is superior compared to those of Kendall's τ and Pearson's r. Based on this grouping structure, sparse-group lasso was applied, with α being set as 0.95 (corresponding to the expectation of strong overall sparseity), and the value of λ being optimized via 10-fold cross validation (*Figure 11*).

Table 6. Clustering results of Method A (minimum cluster size = 10 SNPs)

Cluster	**0**	1	2	3	4	5
# of SNPs	**7**	41	32	24	24	15
Cluster	6	7	8	9	10	11
# of SNPs	15	14	12	11	10	10

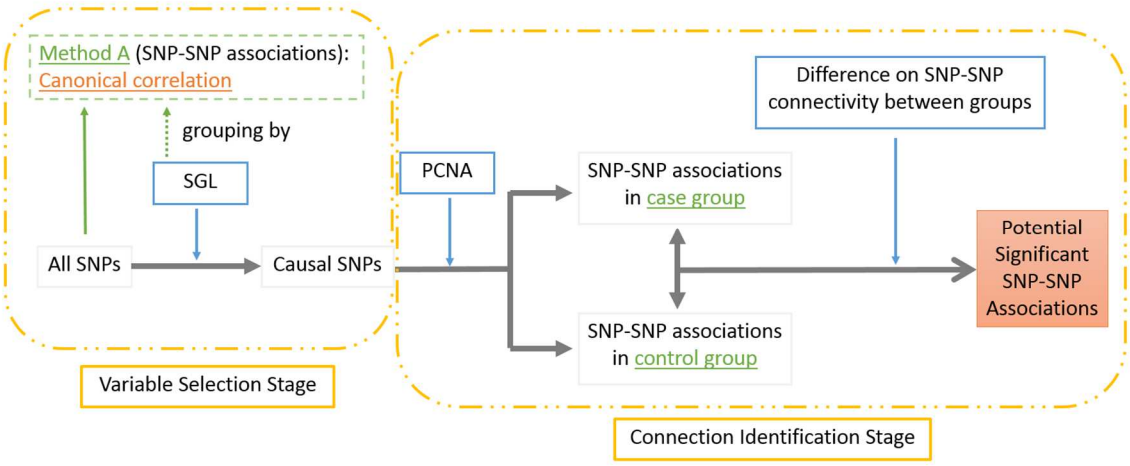


Figure 11. Workflow of Method A

The analysis of data with sparse-group lasso resulted in 14 SNPs with non-zero coefficients from two chromosomes (*Chr15* and *Chr20*) and four clusters (*Table 7*). Following this, we further explored the pairwise conditional relationship among these 14 SNP candidates that would be potentially associated to nicotine dependence. Data of these 14 SNPs were divided into case and control groups. Within each group, pairwise partial correlations were calculated, followed by tests of significance. Multiple testing problem has been taken into full consideration here in the view of the fact that false discovery rate (FDR) was assessed for the p-values from significant tests and has been controlled at 0.05 (α). Through the variable selection process via sparse-group lasso, we have brought down the number of testing dramatically and have thus kept the multiple testing problem in a more benign form. After carefully comparing the two partial correlation networks between groups, we have identified 15 pairs of SNP-SNP associations in interests, the significance of partial correlation of which was detected in one group yet not in the other (*Table 8; Figure 12-15)*.

Table 7. SNPs with non-zero coefficients recognized by sparse-group lasso in Method A

SNP	CHROMOSOME	CLUSTER
rs2036534	15	2
rs3813570	15	2
rs905739	15	2
rs667282	15	2
rs6495309	15	2
rs12440014	15	2
rs3813567	15	2
rs12914008	15	3
rs2036527	15	5
rs17486278	15	5
rs2236196	20	9
rs3787137	20	9
rs3787138	20	9
rs2229959	20	9

Table 8. Partial correlations of 15 SNP pairs in case and control groups

SNP1	SNP2	Partial Correlation Coefficients		Group (1 -- case; 0 -- control)
		Smoking	Non-smoking	
rs2036527	*rs2036534*	0.120	0.096	1
rs17486278	*rs2036534*	0.115	0.103	1
rs12914008	*rs2036534*	0.125	0.035	1
rs2036527	*rs3813570*	0.126	0.086	1
rs12914008	*rs3813570*	0.140	0.038	1
rs2036527	*rs905739*	0.115	0.089	1

rs17486278	*rs905739*	0.128	0.104	1
rs6495309	*rs2036527*	0.107	0.103	1
rs2236196	*rs2036527*	0.103	0.058	1
rs2229959	*rs2036527*	0.164	0.055	1
rs6495309	*rs17486278*	0.118	0.103	1
rs3813567	*rs17486278*	0.118	0.079	1
rs3787137	*rs6495309*	0.094	0.057	1
rs17486278	*rs3813570*	0.099	0.133	0
rs3787138	*rs12914008*	0.035	0.183	0

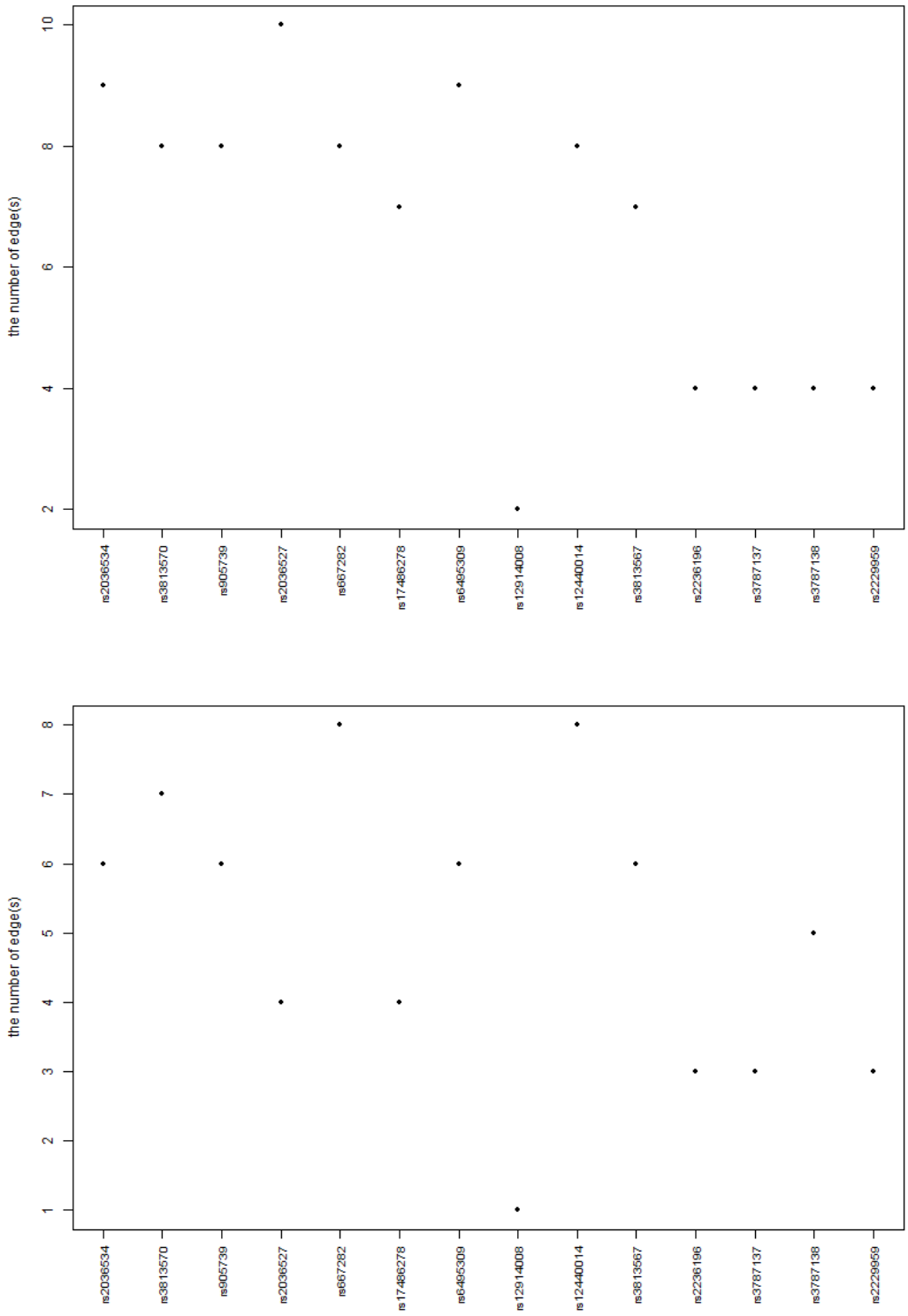


Figure 12. Numbers of edge(s) for each SNP in PCNA networks of case (above) and control (below) groups.

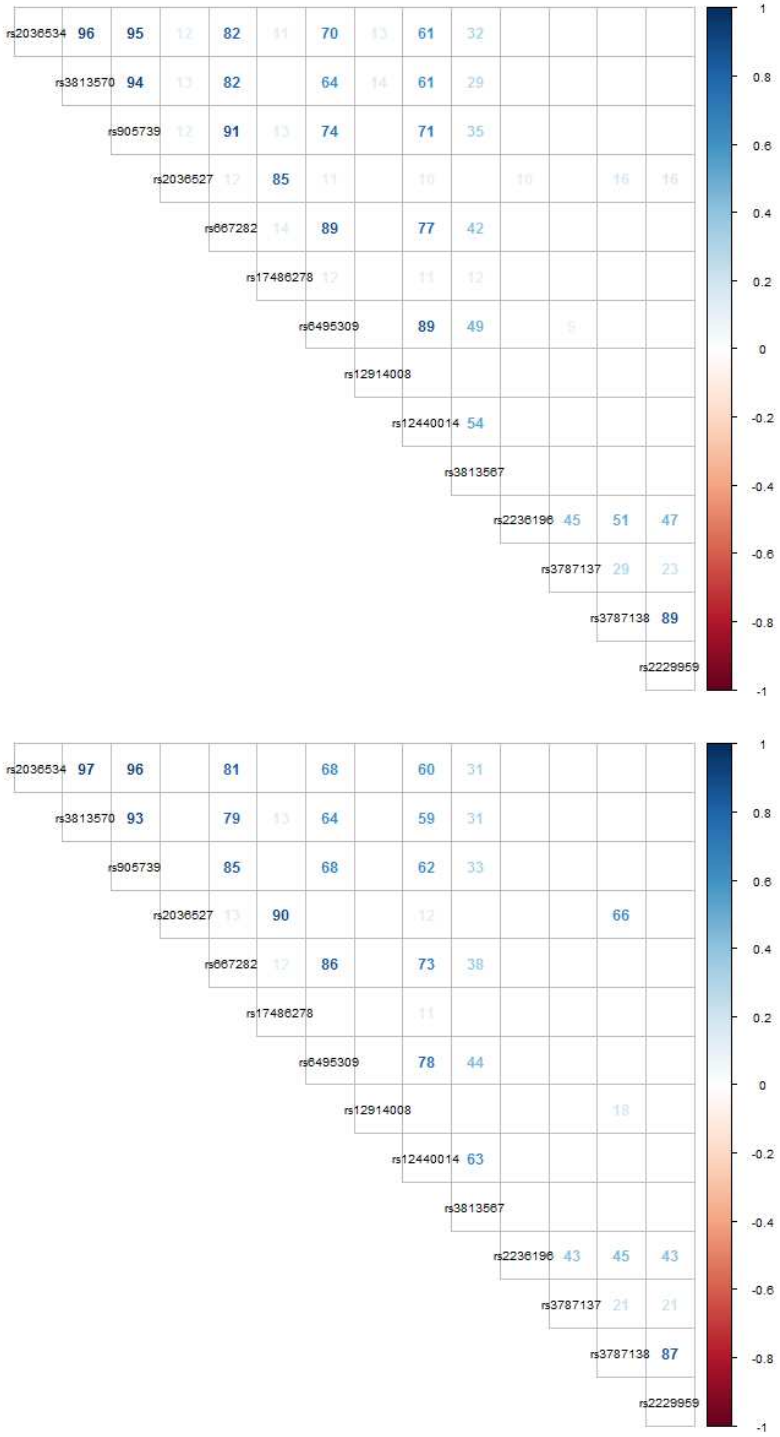


Figure 13. Correlation plots representing pairwise partial correlation coefficients among 14 selected SNP targets in case (above) and control (below) groups. Numbers indicated significant partial correlations translated into percentage and controlled with FDR at 5%. Insignificant partial correlations were suppressed to be expressed. Degree of

transparency represented the magnitude of partial correlation coefficients. Numbers in blue indicated positive partial correlations and red for negative partial correlations.

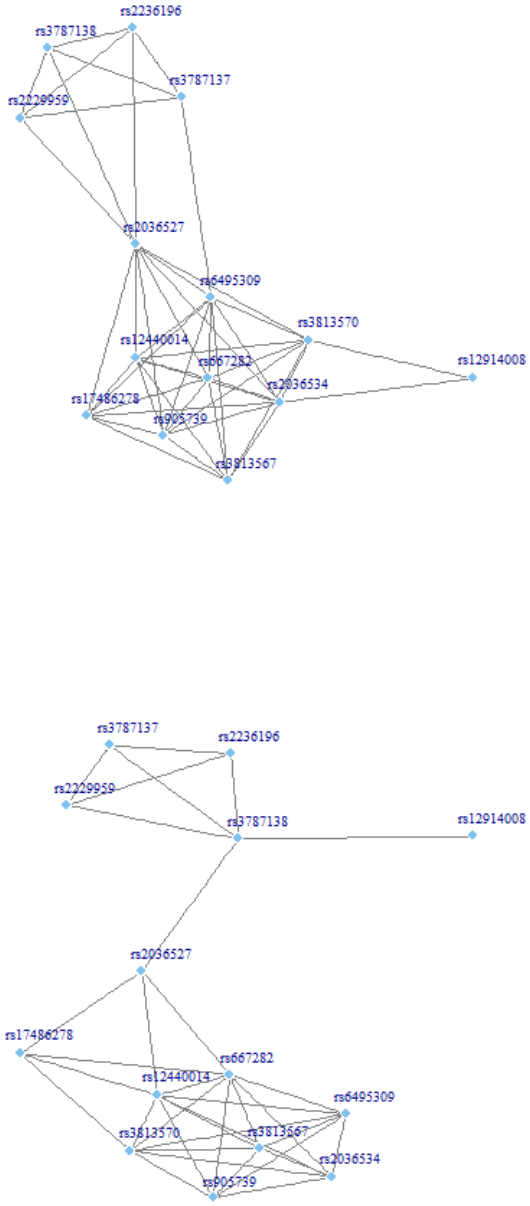


Figure 14. Partial correlation networks in case (above) and control (below) groups.

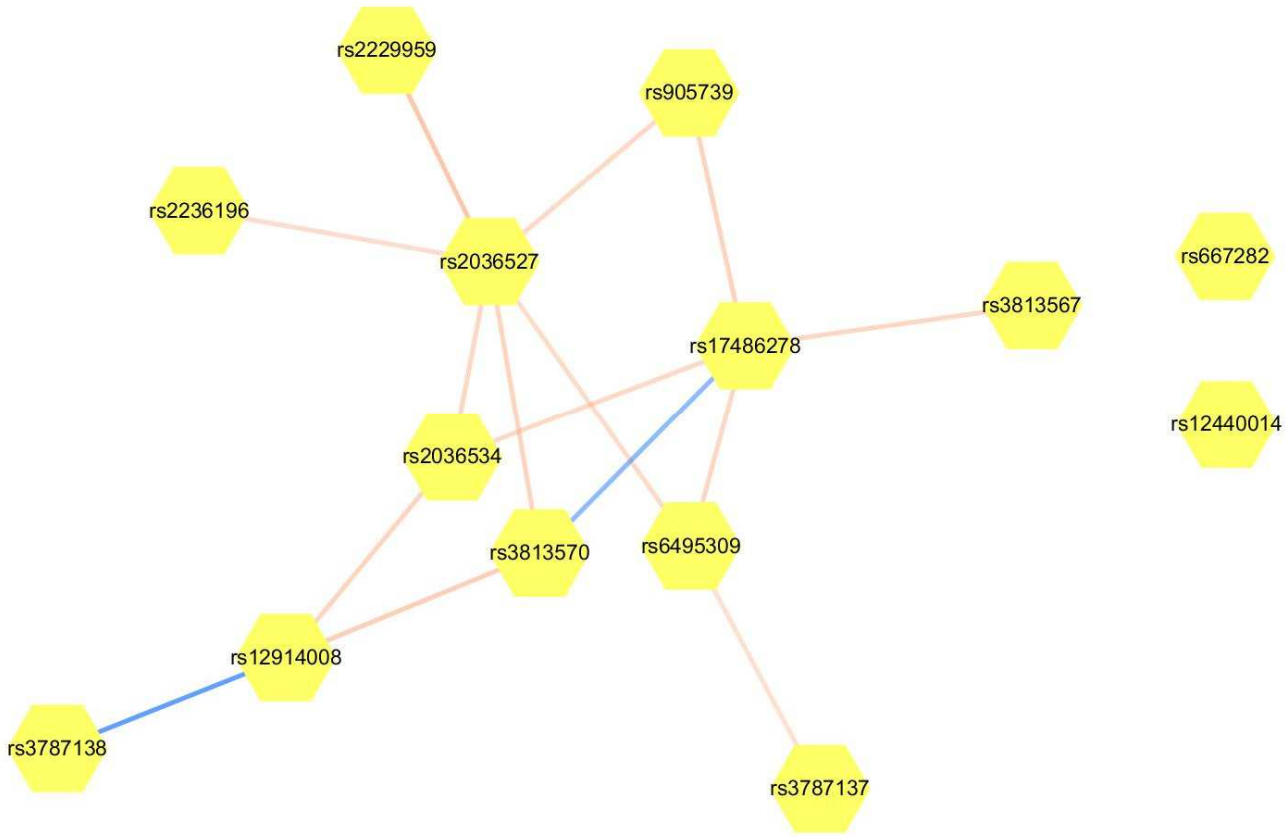


Figure 15. Partial correlation network with 14 selected SNP targets. Nodes represent SNP targets; while edges are recognized by pairwise partial correlations significant in one group yet not in the other (signified by different colors: Orange – Case; Blue – Control). Degree of transparency of edges represents the magnitude of partial correlations between the two connected nodes.

4.1.2. Method B: Grouping by Gene Mapping Information

The 215 SNPs can also be clustered according to their gene mapping information. In the database provided by National Institutes of Health (NIH; http://snpinfo.niehs.nih.gov/snpinfo/snpfunc.htm), we have mapped the SNPs to their corresponding genes, which formed 29 groups in total. The number of SNPs mapped per gene ranged from 1 to 40 *(Table 9-10)*. While majority of SNPs are within gene boundaries, there are 49 (~22.8%) SNPs identified in inter-genic regions. It has been suggested that these variants may still be potentially located in functionally significant regions that are outside gene boundaries (Silver, et al. 2013). Thus, these 49 SNPs were mapped to whichever nearby genes with shorter distance in base pairs (bp; *Table 13)*. Based on this grouping structure, sparse-group lasso was performed, with α being set as 0.95 and the value of λ being optimized via 10-fold cross validation (*Figure 16*).

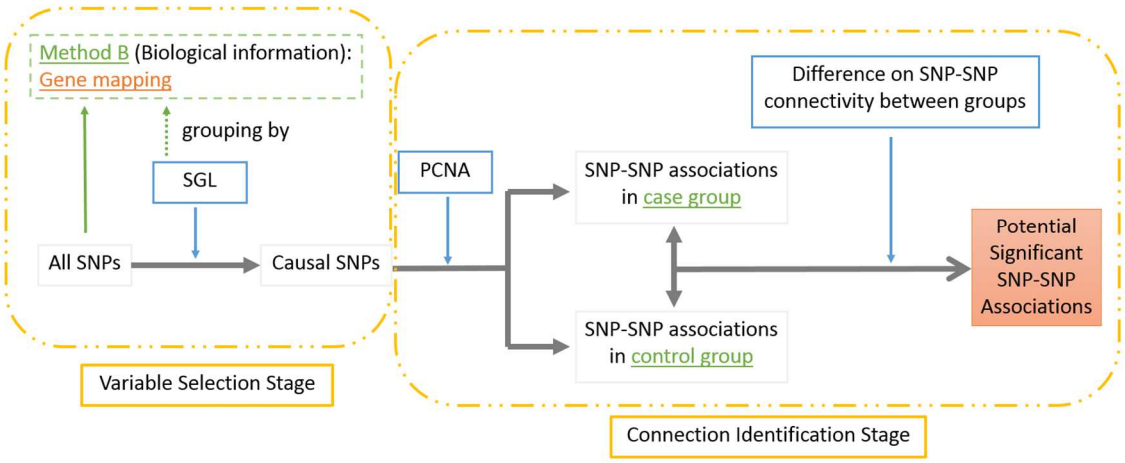


Figure 16. Workflow of Method B

Table 9. Summary of gene mapping information of SNPs

CHRNA7	***CHRNA3***	***CHRNA5***	***CHRNB3***	***IREB2***	***CHRNA2***
40	20	18	17	13	11
CHRNB4	***CHRNA9***	***CHRNA4***	***CHRNA6***	***CHRNB1***	***CHRNG***
11	10	9	9	9	6
CHRNA1	***LOC100130311***	***LOC123688***	***PSMA4***	***CHRNB2***	***CHRND***
5	5	5	5	4	4
CHRNA10	***EIF4E2***	***FGF11***	***ART1***	***LOC100130587***	***LOC100133187***
2	2	2	1	1	1
MINK1	***NUP98***	***POLR2A***	***TMEM102***	***ZBTB4***	
1	1	1	1	1	

Table 10. Summary of gene mapping and chromosome information

Gene	**Chromosome**							
	1	**2**	**4**	**8**	**11**	**15**	**17**	**20**
ART1	0	0	0	1	0	0	0	0
CHRNA1	0	0	0	1	1	2	1	0
CHRNA10	0	0	0	0	0	2	0	0
CHRNA2	0	0	0	1	0	8	1	1
CHRNA3	0	2	0	2	0	11	4	1
CHRNA4	0	1	0	0	0	8	0	0
CHRNA5	0	1	0	3	0	13	0	1
CHRNA6	1	2	0	0	1	1	2	2
CHRNA7	0	3	5	6	2	16	7	1
CHRNA9	3	0	0	0	0	6	0	1
CHRNB1	0	0	0	1	0	8	0	0
CHRNB2	0	0	0	3	0	1	0	0
CHRNB3	0	1	0	7	0	8	1	0
CHRNB4	0	3	2	2	0	3	0	1
CHRND	0	0	1	0	0	3	0	0
CHRNG	0	1	0	2	0	3	0	0
EIF4E2	0	0	0	0	0	2	0	0
FGF11	0	0	0	0	0	2	0	0
IREB2	0	0	0	5	0	7	0	1
LOC100130311	0	0	0	1	0	3	1	0
LOC100130587	0	0	0	0	0	0	1	0
LOC100133187	0	1	0	0	0	0	0	0
LOC123688	0	1	0	1	0	2	1	0
MINK1	0	0	1	0	0	0	0	0

NUP98	0	0	0	0	0	1	0	0
POLR2A	0	0	0	0	0	0	0	1
PSMA4	0	1	1	0	0	1	1	1
TMEM102	0	0	0	0	0	1	0	0
ZBTB4	0	0	0	1	0	0	0	0

The sparse-group lasso picked out 14 SNPs with non-zero coefficients from three chromosomes (*Chr8*, *Chr15* and *Chr20*) and seven genes (*Table 11*). Subsequently, pairwise partial correlation coefficients of these chosen SNPs were calculated in case and control groups respectively. The significance of coefficients was tested and controlled at FDR = 0.05. By comparing the two partial correlation networks between groups, 13 SNP-SNP associations of interests were identified, the partial correlations of which were significant in one group yet not in the other (*Table 12; Figure 17-20*).

Table 11. SNPs with non-zero coefficients recognized by sparse-group lasso in Method B

SNP	CHROMOSOME	GENE MAPPING
rs2292974	8	*CHRNA2*
rs12440014	15	*CHRNB4*
rs12914008	15	*CHRNB4*
rs1317286	15	*CHRNA3*
rs17486278	15	*CHRNA5*
rs2036527	15	*CHRNA5*
rs2036534	15	*LOC123688*
rs3813567	15	*CHRNB4*
rs3813570	15	*PSMA4*
rs667282	15	*CHRNA5*
rs2229959	20	*CHRNA4*

rs2236196	20	*CHRNA4*
rs3787137	20	*CHRNA4*
rs3787138	20	*CHRNA4*

Table 12. Partial correlations of 13 SNP pairs in case and control groups

SNP1	**SNP2**	**Partial Correlation Coefficients**		**Group (1 -- case; 0 -- control)**
		Smoking	**Non-smoking**	
rs2292974	*rs12440014*	0.113	0.055	1
rs2036534	*rs12914008*	0.222	0.034	1
rs3813570	*rs12914008*	0.136	0.035	1
rs2036534	*rs17486278*	0.128	0.103	1
rs3813567	*rs17486278*	0.120	0.096	1
rs2036534	*rs2036527*	0.120	0.099	1
rs2229959	*rs2036527*	0.244	0.062	1
rs3813570	*rs2036527*	0.129	0.089	1
rs3787137	*rs2036534*	0.121	0.023	1
rs3787138	*rs12914008*	0.049	0.182	0
rs2229959	*rs1317286*	0.061	0.122	0
rs3813570	*rs17486278*	0.101	0.129	0
rs3813570	*rs2229959*	0.044	0.098	0

Table 13. Summary of 49 SNPs located in inter-gene regions

SNP	Position	Nearby Genes	Distance (bp)	Gene Mapping
rs4796305	7276779	*C17orf74*||*TMEM102*	-5168||-2707	*TMEM102*
rs10958726	42655066	*C8orf40*||*CHRNB3*	-127786||-16653	*CHRNB3*
rs13277254	42669139	*C8orf40*||*CHRNB3*	-141859||-2580	*CHRNB3*
rs13277524	42669214	*C8orf40*||*CHRNB3*	-141934||-2505	*CHRNB3*
rs13280301	42669174	*C8orf40*||*CHRNB3*	-141894||-2545	*CHRNB3*
rs1530847	42667396	*C8orf40*||*CHRNB3*	-140116||-4323	*CHRNB3*
rs1955185	42668804	*C8orf40*||*CHRNB3*	-141524||-2915	*CHRNB3*
rs1955186	42668648	*C8orf40*||*CHRNB3*	-141368||-3071	*CHRNB3*
rs5005909	42647824	*C8orf40*||*CHRNB3*	-120544||-23895	*CHRNB3*
rs6474412	42669655	*C8orf40*||*CHRNB3*	-142375||-2064	*CHRNB3*
rs6474413	42670221	*C8orf40*||*CHRNB3*	-142941||-1498	*CHRNB3*
rs2231529	3649829	*CHRNA10*||*NUP98*	-639||-2987	*CHRNA10*
rs2231532	3649696	*CHRNA10*||*NUP98*	-506||-3120	*CHRNA10*
rs6578411	3652548	*CHRNA10*||*NUP98*	-3358||-268	*NUP98*
rs2565055	27393297	*CHRNA2*||*EPHX2*	-567||-11265	*CHRNA2*
rs6495309	76702300	*CHRNA3*||*CHRNB4*	-1923||-1391	*CHRNB4*
rs10107450	42749052	*CHRNA6*||*THAP1*	-6276||-61922	*CHRNA6*
rs7828365	42748471	*CHRNA6*||*THAP1*	-5695||-62503	*CHRNA6*
rs17732878	7303083	*CHRNB1*||*ZBTB4*	-1427||-338	*ZBTB4*
rs9298628	42725148	*CHRNB3*||*CHRNA6*	-13782||-1772	*CHRNA6*
rs9298629	42725343	*CHRNB3*||*CHRNA6*	-13977||-1577	*CHRNA6*
rs6987323	42716389	*CHRNB3*||*CHRNA6*	-5023||-10531	*CHRNB3*
rs7012713	42711460	*CHRNB3*||*CHRNA6*	-94||-15460	*CHRNB3*
rs3813567	76721606	*CHRNB4*||*LOC390612*	-964||-18832	*CHRNB4*
rs3971872	76729090	*CHRNB4*||*LOC390612*	-8448||-11348	*CHRNB4*
rs4790235	4746831	*CHRNE/LOC100130311*	4991/3339||317/175	*LOC100130311*
rs2276560	233159163	*EIF4E2*||*EFHD1*	-16999||-47405	*EIF4E2*

rs6749955	233146161	*EIF4E2||EFHD1*	-3997||-60407	*EIF4E2*
rs17483548	76517368	*LOC100129388||IREB2*	-48810||-205	*IREB2*
rs16954243	4753179	*LOC100130311||GP1BA*	-6173||-23193	*LOC100130311*
rs3760490	4748705	*LOC100130311||GP1BA*	-1699||-27667	*LOC100130311*
rs7214776	4752393	*LOC100130311||GP1BA*	-5387||-23979	*LOC100130311*
rs8080668	4758494	*LOC100130311||GP1BA*	-11488||-17878	*LOC100130311*
rs12442690	30088689	*LOC100130857||CHRNA7*	-32697||-21329	*CHRNA7*
rs1514246	30106782	*LOC100130857||CHRNA7*	-50790||-3236	*CHRNA7*
rs3087454	30108259	*LOC100130857||CHRNA7*	-52267||-1759	*CHRNA7*
rs3826029	30108777	*LOC100130857||CHRNA7*	-52785||-1241	*CHRNA7*
rs6494165	30108578	*LOC100130857||CHRNA7*	-52586||-1440	*CHRNA7*
rs4603829	61439336	*LOC100131010||CHRNA4*	-9679||-5773	*CHRNA4*
rs4469116	40032005	*LOC100132141||CHRNA9*	-4812||-221	*CHRNA9*
rs4602530	40031981	*LOC100132141||CHRNA9*	-4788||-245	*CHRNA9*
rs6823439	40031357	*LOC100132141||CHRNA9*	-4164||-869	*CHRNA9*
rs1107953	61471990	*LOC100133187||LOC100130152*	-1900||-22377	*LOC100133187*
rs12916483	76619452	*LOC123688||PSMA4*	-2682||-350	*PSMA4*
rs9901643	7360548	*POLR2A||TNFSF12*	-1895||-32551	*POLR2A*
rs2036527	76638670	*PSMA4||CHRNA5*	-10053||-6291	*CHRNA5*
rs503464	76644951	*PSMA4||CHRNA5*	-16334||-10	*CHRNA5*
rs880395	76631411	*PSMA4||CHRNA5*	-2794||-13550	*PSMA4*
rs905739	76632165	*PSMA4||CHRNA5*	-3548||-12796	*PSMA4*

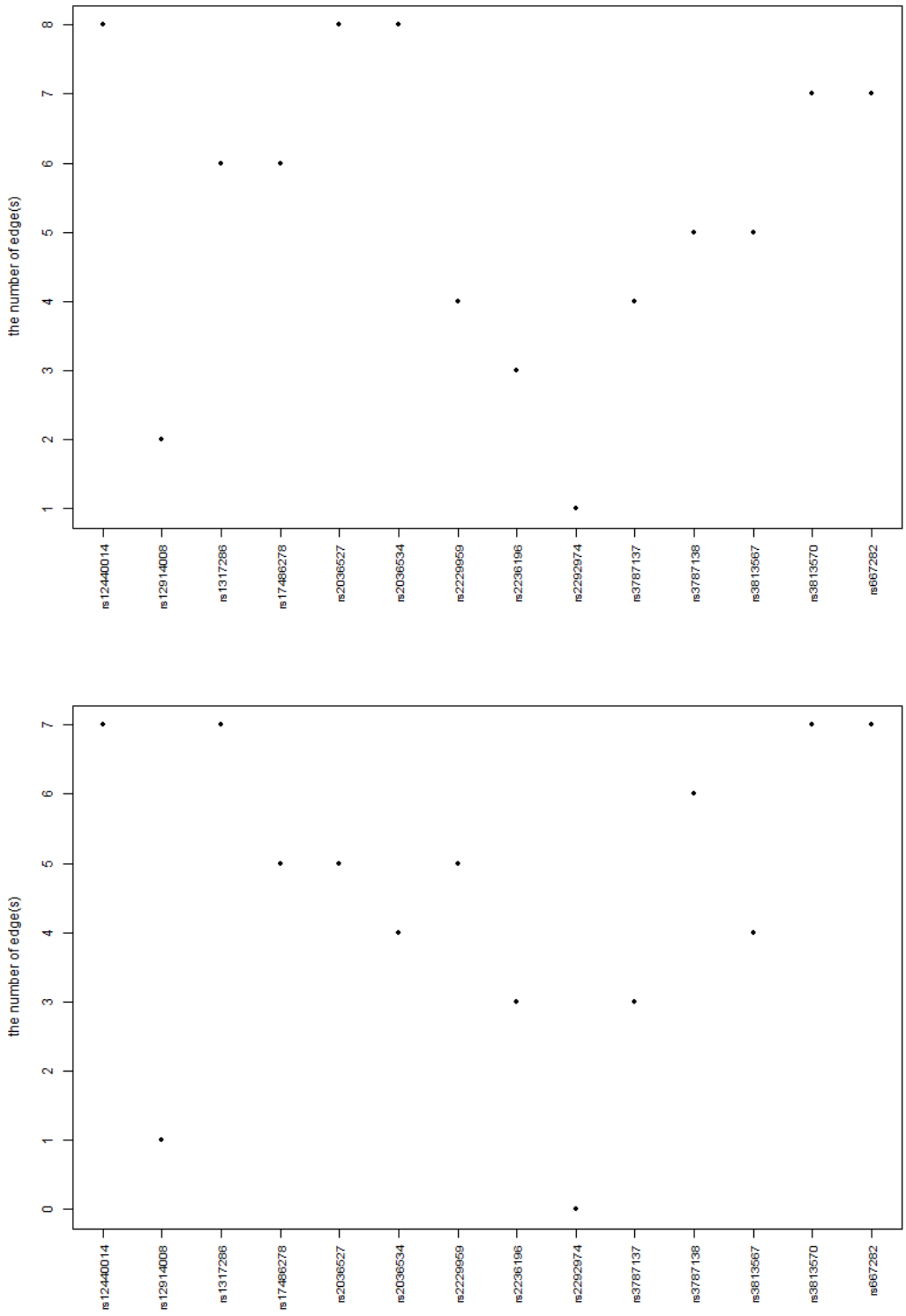


Figure 17. Numbers of edge(s) for each SNP in PCNA networks of case (above) and control (below) groups.

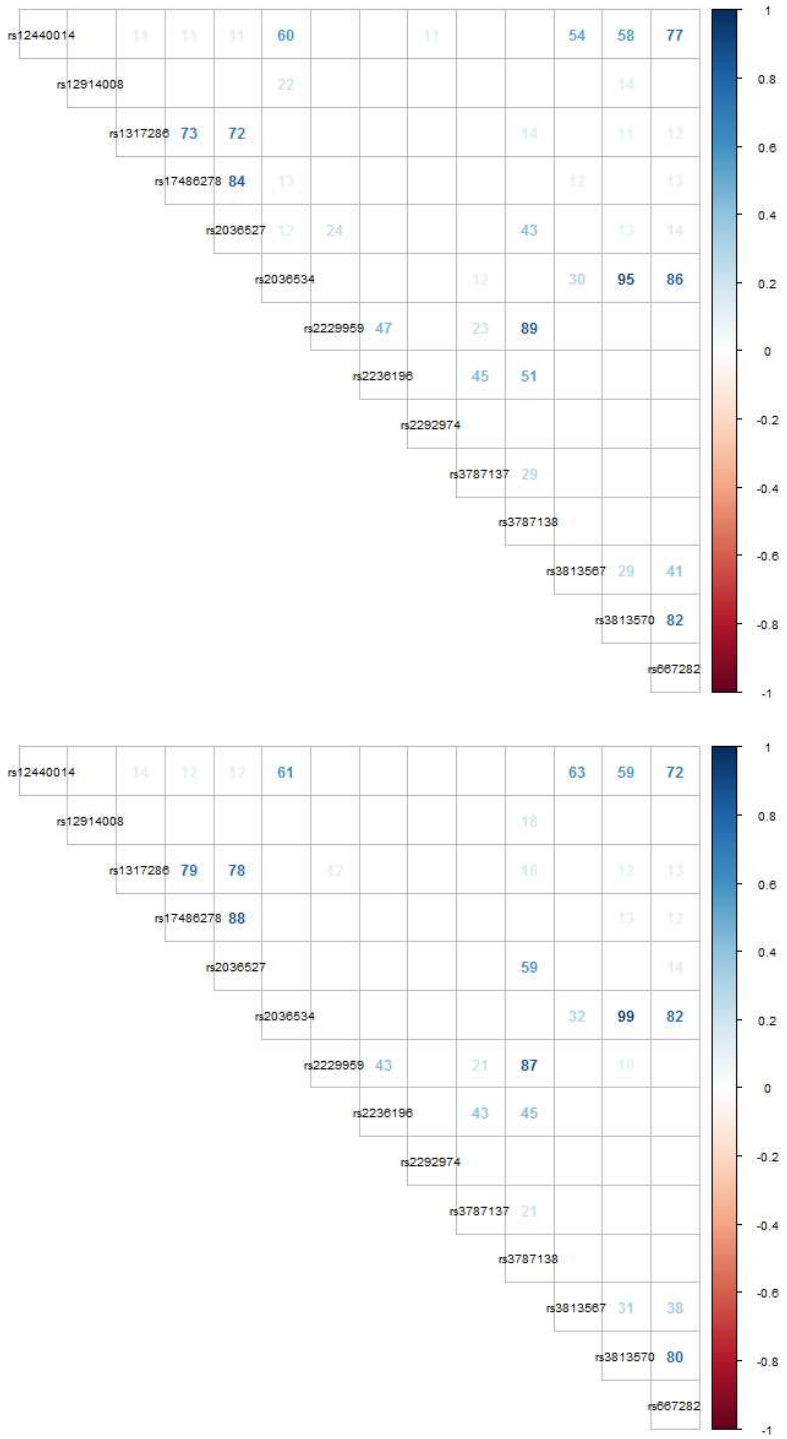


Figure 18. Correlation plots representing pairwise partial correlation coefficients among 14 selected SNP targets in case (above) and control (below) groups. Numbers indicated significant partial correlations translated into percentage and controlled with FDR at 5%. Insignificant partial correlations were suppressed to be expressed. Degree of

transparency represented the magnitude of partial correlation coefficients. Numbers in blue indicated positive partial correlations and red for negative partial correlations.

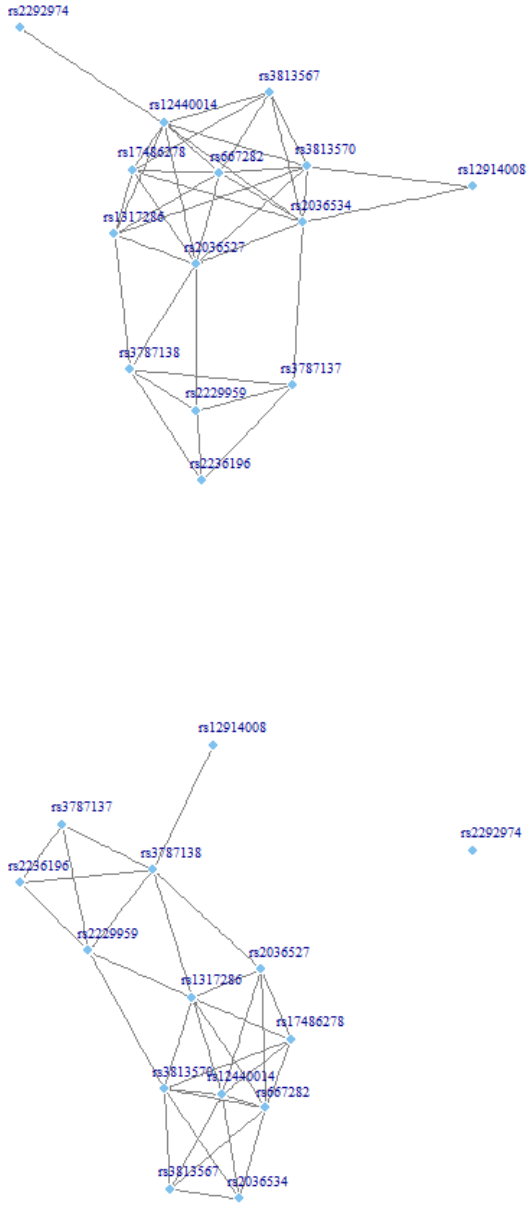


Figure 19. Partial correlation networks in case (above) and control (below) groups.

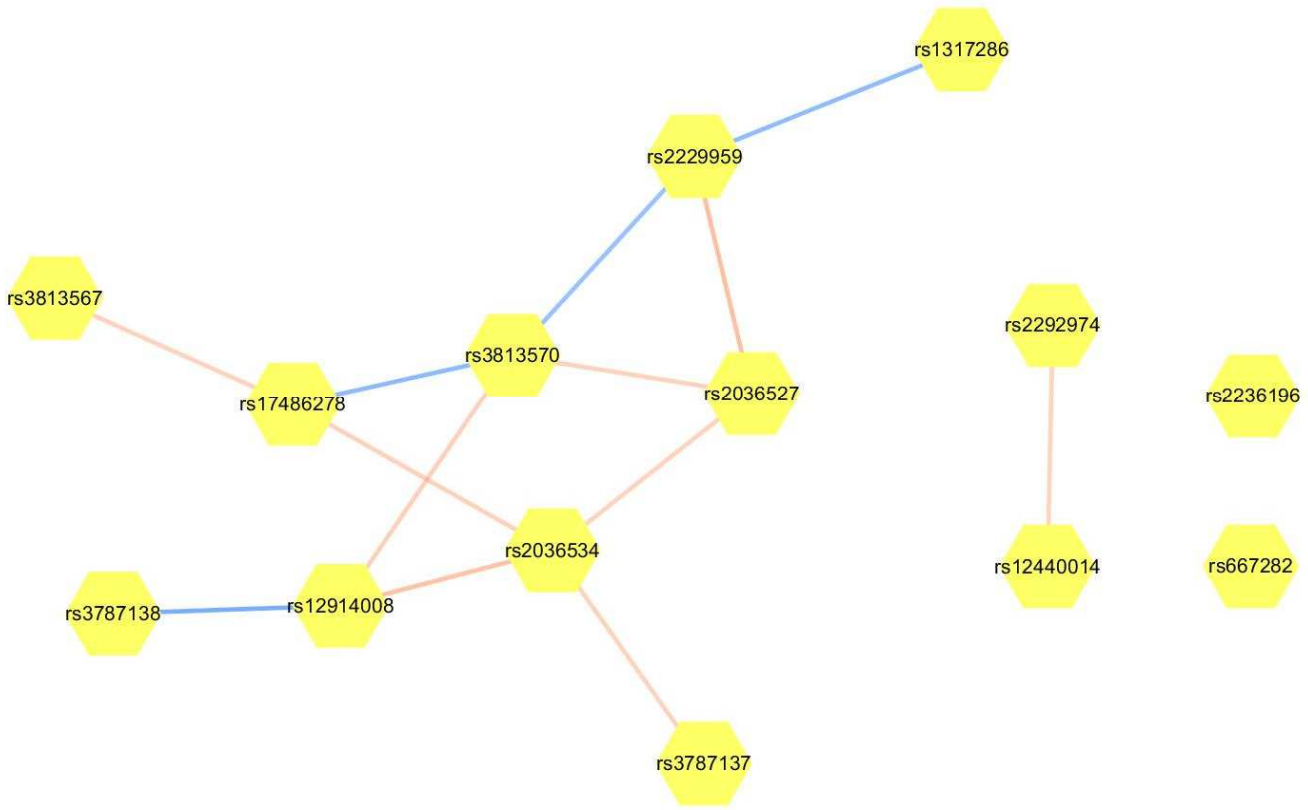


Figure 20. Partial correlation network with 14 selected SNP targets. Nodes represent SNP targets; while edges are recognized by pairwise partial correlations significant in one group yet not in the other (signified by different colors: Orange – Case; Blue – Control). Degree of transparency of edges represents the magnitude of partial correlations between the two connected nodes.

4.1.3. Summary

Comparing the results from Method A and Method B, 12 individual SNPs *(Table 14)* and 9 pairs of SNP-SNP associations *(Table 15)* were identical. The fact that most detected targets from Method A and Method B were the same demonstrates that the pairwise association measurement based on canonical correlation may well represent the relationship of SNPs concerning their gene mapping, and can thus be considered as an alternative when biological information of SNPs is insufficient or unavailable.

Table 14. Identical individual SNPs recognized by Method A and Method B

SNP	CHROMO-SOME	GENE MAPPING	SNP	CHROMO-SOME	GENE MAPPING
rs3813567	15	*CHRNB4*	*rs2036534*	15	*LOC123688*
rs17486278	15	*CHRNA5*	*rs3813570*	15	*PSMA4*
rs2036527	15	*CHRNA5*	*rs2229959*	20	*CHRNA4*
rs667282	15	*CHRNA5*	*rs2236196*	20	*CHRNA4*
rs12440014	15	*CHRNB4*	*rs3787137*	20	*CHRNA4*
rs12914008	15	*CHRNB4*	*rs3787138*	20	*CHRNA4*

Table 15. Identical SNP-SNP associations detected by Method A and Method B

SNP1	SNP2	SNP1	SNP2
rs2036534	*rs12914008*	*rs2229959*	*rs2036527*
rs3813570	*rs12914008*	*rs3813570*	*rs2036527*
rs2036534	*rs17486278*	*rs3787138*	*rs12914008*
rs3813567	*rs17486278*	*rs3813570*	*rs17486278*
rs2036534	*rs2036527*		

4.2. Method C: Penalized Stepwise Logistic Regression Model

The penalized stepwise logistic regression model was once again employed on the COGEND data set for comparison (denoted as "Method C") using R package "*stepPLR*". The value of λ was determined through a 5-fold cross-validation process and set as $\lambda = 1$. Items were picked via forward selection, followed by a backward deletion. Maximum number of items to be added in the selection procedure was set to be 30. The final model included 28 items, including individual SNPs and up to 7-way SNP-SNP interactions (*Table 16*).

Table 16. Individual targets and interactions (up to 7-way) identified in Method C

Individual SNPs *rs12440014* *rs2236196* *rs12914008* *rs16925377*	**Two-way Interactions** *rs2292974:rs12440014* *rs6494212:rs2236196* *rs1376866:rs12440014*
Three-way Interactions *rs8192479:rs2292974:rs12440014* *rs3787138:rs2292974:rs12440014* *rs2767:rs2292974:rs12440014* *rs6474412:rs1376866:rs12440014* *rs2289080:rs6494212:rs2236196*	

Four-way Interactions
rs11636605:rs8192479:rs2292974:rs12440014
rs4950:rs3787138:rs2292974:rs12440014
rs12442690:rs3787138:rs2292974:rs12440014
rs8192475:rs2289080:rs6494212:rs2236196
rs3971872:rs3787138:rs2292974:rs12440014
Five-way Interactions
rs16956223:rs4950:rs3787138:rs2292974:rs12440014
rs1051730:rs4950:rs3787138:rs2292974:rs12440014
rs950776:rs12442690:rs3787138:rs2292974:rs12440014
rs2036527:rs11636605:rs8192479:rs2292974:rs12440014
rs4796305:rs3971872:rs3787138:rs2292974:rs12440014
Six-way Interactions
rs1376866:rs16956223:rs4950:rs3787138:rs2292974:rs12440014
rs6495309:rs950776:rs12442690:rs3787138:rs2292974:rs12440014
rs10009228:rs1051730:rs4950:rs3787138:rs2292974:rs12440014
rs1500948:rs4796305:rs3971872:rs3787138:rs2292974:rs12440014
Seven-way Interactions
rs578776:rs1376866:rs16956223:rs4950:rs3787138:rs2292974:rs12440014
rs1827294:rs1500948:rs4796305:rs3971872:rs3787138:rs2292974:rs12440014

The individual targets recognized by Method C were similar to those from Method A and B (3 out of 4 are the same). One of the two-way interactions -- *rs2292974:rs12440014* was also detected by Method B (grouping by gene mapping information). Comparing to the two-stage approach (Method A and B), the penalized stepwise model (Method C) has identified relatively small number of either individual or two-way interaction targets. On the other hand, this model recognized interactions as long as at least one of the main effects is in the model; furthermore, interactions more than two-way were taken into consideration.

4.3. Discussion

Information of individual SNPs picked out by the three methods and corresponding phenotype-associated reference were summarized in *Table 19*. All selected single SNPs from three methods come from 8 genes (*CHRNA3, CHRNB4, CHRNA5, PSMA4, LOC123688, CHRNA4, CHRNA2* and *ART1; Table 17*). Among them, the first five genes are from the so-called “Chromosome 15q25.1 region” (*Table 18*). This region has been identified to be associated with smoking behavior and would increase the risks of nicotine dependence, and smoking-related diseases, such as lung cancer (Saccone, et al. 2009, VanderWeele, et al. 2012). The first three genes (*CHRNA3, CHRNB4,* and *CHRNA5*) lie very close to each other and are in strong linkage disequilibrium (LD) with each other. They are treated as the “*CHRNA5-CHRNA3-CHRNB4*” cluster in most studies. This nicotinic acetylcholine receptor (nAChR) subunit gene cluster located on Chromosome 15q24-25 has been claimed by GWA study findings, to be associating with nicotine

dependence (Hung, et al. 2008), smoking behavior (Berrettini, et al. 2008, Thorgeirsson, et al. 2008, Caporaso, et al. 2009, David, et al. 2012), and lung cancer (Amos, et al. 2008, Hung, et al. 2008). Totally 8 SNPs identified by at least one of the three methods located in this cluster. *PSMA4* (proteasome subunit, alpha type 4), a 20S proteasome structural protein gene, has been claimed as a strong candidate mediator of lung cancer cell proliferation and apoptosis (Liu, et al. 2009, Hansen, et al. 2010). Down-regulation of *PSMA4* expression decreases proteasome activity and induces apoptosis. In this gene, we have identified SNPs *rs3813570* and *rs905739*, which have been shown to relate to smoking behavior and nicotine dependence, respectively (Wang, et al. 2009, David, et al. 2012, Meyers, et al. 2013). *LOC123688* is a hypothetical gene and we have identified *rs2036534* from it.

Table 17. Corresponding genes mapped with SNPs identified by the three methods

Gene	SNPs Selected by Method A	SNPs Selected by Method B	SNPs Selected by Method C
CHRNA3		*rs1317286*	
CHRNB4	*rs12440014*	*rs12440014*	*rs12440014*
	rs12914008	*rs12914008*	*rs12914008*
	rs3813567	*rs3813567*	
	rs6495309		
CHRNA5	*rs17486278*	*rs17486278*	
	rs2036527	*rs2036527*	
	rs667282	*rs667282*	
PSMA4	*rs3813570*	*rs3813570*	
	rs905739		

LOC123688	*rs2036534*	*rs2036534*	
CHRNA4	*rs2229959*	*rs2229959*	
	rs2236196	*rs2236196*	*rs2236196*
	rs3787137	*rs3787137*	
	rs3787138	*rs3787138*	
CHRNA2		*rs2292974*	
ART1			*rs16925377*

Table 18. Chromosome 15q25.1 region and the included genes

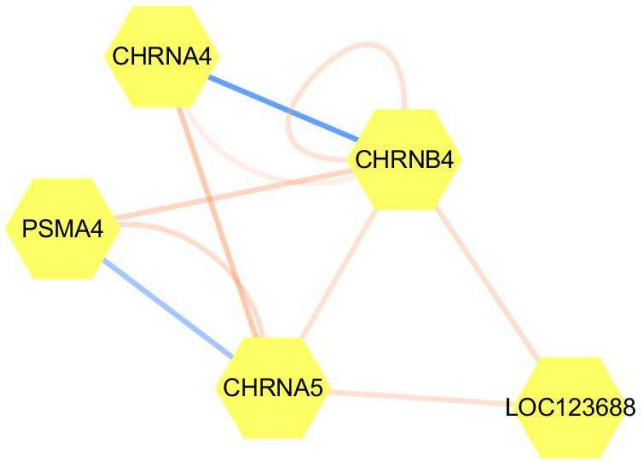
	Gene	**Description**
Chromosome 15q25.1 Region	*CHRNA5*	nicotinic cholinergic receptor subunit genes
	CHRNA3	
	CHRNB4	
	PSMA4	a proteasome subunit encoding gene
	LOC123688	a hypothetical gene
	IREB2	an iron responsive element-binding protein

Both *CHRNA4* and *CHRNA2* are nicotinic acetylcholine receptor (nAChR) subunits. There are in total 11 nAChR subunit-encoding genes – *CHRNA2, 3, 4, 5, 6, 7, 9, 10*; *CHRNB2, 3, 4* –located on 6 chromosomes (*Chr1, 4, 8, 11, 15*, and *20*). They code for

proteins that form receptors present in neuronal and other tissues and are strong candidate genes for smoking-related disease (Hung, et al. 2008). In the mouse models, the cholinergic receptor, nicotinic, alpha 4 (*CHRNA4*), encodes the $\alpha 4$ subunit of nAChRs, which, together with *CHRNB2*, form the most prevalent nAChRs in brain (Lou, et al. 2007) and has been reported to be involved in nicotine-induced reward, tolerance and sensitization (Tapper, et al. 2004). Both *rs2236196* (located in 3' UTR) and *rs3787137* (located in the intron) were reported to associate with smoking behavior (Li, et al. 2005, Hutchison, et al. 2007). *CHRNA2* has been linked to tobacco dependence and smoking intensity (Faraone, et al. 2004, Swan, et al. 2006); yet *rs2292974* has been reported to potentially associated with nicotine dependence (Philibert, et al. 2009). Recognized by stepwise penalized model, *rs16925377* located in *Chr11*, referring to gene *ART1* (ADP-Ribosyltransferase 1) that does not belong to neuronal nicotinic receptor family, although it may have modest effect on nicotine dependence (Saccone, et al. 2010).

In a genome-wide meta-analysis of smoking behaviors in African-Americans, both *rs667282* and *rs3813570* were reported to be weakly correlated with *rs2036527* in the study of CPD (cigarettes per day; r^2=0.2 in CEU (Northern and Western Europe) and 0.12 in YRI (Yoruba in Ibadan, Nigeria)); while they were also correlated with each other (r^2=0.60 in CEU and 0.32 in YRI) (David, et al. 2012). In our study, we have identified the SNP-SNP association, *rs2036527:rs3813570* in both Method A (partial correlation: 0.126 in case group, 0.086 in control group) and Method B (partial correlation: 0.129 in case group, 0.089 in control group) (*Table 8, Table 12*).

To gain further insight, we transferred the SNP partial correlation network to the corresponding gene network. For gene pairs associated with more than one SNP-SNP interactions in the same group, we simplified them to one; yet we keep both edges for the same gene pair if they are from different groups



(

Figure 21-*22*). Both methods have recognized the close relationship among genes in the "Chromosome 15q25.1 region". Besides, they also pinpointed the association of this region with *CHRNA4*. *CHRNA4* has been reported to be up-regulated under chronic nicotine exposure (Marks, et al. 1992) and its activation is sufficient for nicotine-induced reward, tolerance, and sensitization (Tapper 2004). It has been demonstrated that *CHRNA4* interacts with *CHRNB2* experimentally, and acts jointly with *BDNF* (Brain-Derived Neurotrophic Factor) or *NTRK2* (Neurotrophic Tyrosine Kinase, Receptor, Type 2) contributing to nicotine dependence in a yet unknown indirect manner (Li, Lou, Chen, Ma and Elston 2008). Our result proposed another potential jointly direct relationship of *CHRNA4* and genes in the "Chromosome 15q25.1 region" to associate with nicotine dependence.

Table 19. Summary of phenotype-associated reference of SNPs

SNP	Chr	Position (bp)	Gene	SNP Function	Phenotype	Reference
rs1317286	15	78603787	*CHRNA3*	Intron	Smoking behavior	(Berrettini, et al. 2008)
					Nicotine dependence	(Saccone, et al. 2009, Li, et al. 2010)
					Lung cancer	(Hung, et al. 2008, Amos, et al. 2010)
rs6495309	15	76702300	*CHRNB4*	Locus	Lung cancer	(Amos, et al. 2010, Yang, et al. 2012)
rs12914008	15	76710560	*CHRNB4*	Nonsynon	Nicotine dependence	(Zhang, Summah, Zhu and Qu 2011)
rs12440014	15	76713781	*CHRNB4*	Intron	Smoking behavior	(Stevens, et al. 2008)
					Lung cancer	(Amos, et al. 2010)
rs3813567	15	76721606	*CHRNB4*	(5' near gene)	Nicotine dependence	(Saccone, et al. 2007)
rs2036527	15	76638670	Upstream of *CHRNA5*		Smoking behavior	(Stevens, et al. 2008, Caporaso, et al. 2009, Broms, et al. 2012, David, et al. 2012, Zhu, et al. 2014)
					Nicotine dependence	(Greenbaum and Lerer 2009)
					Lung cancer	(Amos, et al. 2010, Scherf, et al. 2013)
rs667282	15	76650527	*CHRNA5*	Intron	Lung cancer	(Amos, et al. 2010)
rs17486278	15	76654537	*CHRNA5*	Intron	Smoking behavior	(Bierut, et al. 2008, Stevens, et al. 2008, Greenbaum, et al. 2009)
rs17486278	15	76654537	*CHRNA5*	Intron	Nicotine dependence	(Weiss, et al. 2008, Saccone, et al. 2009)
					Lung cancer	(Hansen, et al. 2010)

rs3813570	15	76619887	*PSMA4*	mRNA-UTR	Smoking behavior	(David, et al. 2012, Meyers, et al. 2013)
rs905739	15	76632165	*PSMA4*		Nicotine dependence	(Wang, et al. 2009)
rs2036534	15	76614003	*LOC123688*		Nicotine dependence	(Saccone, et al. 2009)
					Lung cancer	(Wu, et al. 2009a, Bae, et al. 2012)
rs2236196	20	61448000	*CHRNA4*	mRNA-UTR	Nicotine dependence	(Saccone, et al. 2007, Li, et al. 2008, Portugal and Gould 2008, Li and Burmeister 2009)
					Smoking behavior	(Li, et al. 2005, Hutchison, et al. 2007, Han, et al. 2011)
rs3787138	20	61449668	*CHRNA4*	Intron	Nicotine dependence	(Saccone, et al. 2010)
rs3787137	20	61449544	*CHRNA4*	Intron	Smoking behavior	(Li, et al. 2005, Portugal, et al. 2008)
					Nicotine dependence	(Lou, et al. 2007, Li, et al. 2008)
rs2229959	20	61451998	*CHRNA4*	Synon	Nicotine dependence	(Lou, et al. 2007, Breitling, et al. 2009, Greenbaum, et al. 2009)
rs2292974	8	27374308	*CHRNA2*	mRNA-UTR	Nicotine dependence	(Philibert, et al. 2009)
rs16925377	11	3677075	*ART1*		Nicotine dependence	(Saccone, et al. 2010)

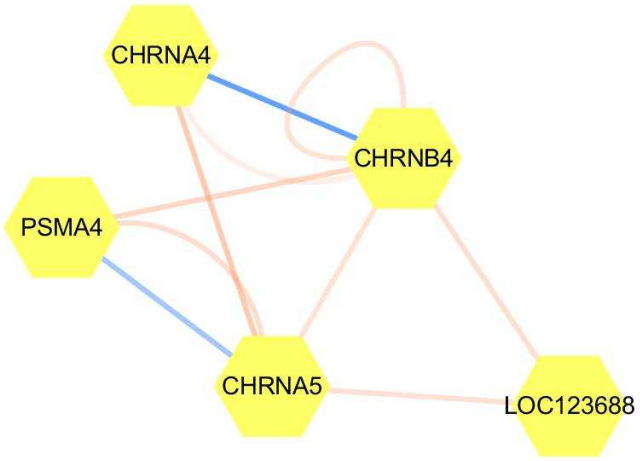


Figure 21. (Method A) Gene partial correlation network by mapping SNPs to their corresponding/nearest genes for potential gene-gene associations. Partial correlation coefficients of gene pairs with multiple SNP pair mappings would be the largest one amongst. Colors of edges represent different groups: Orange – Case; Blue – Control. Degree of transparency of edges represents the magnitude of partial correlations between the two nodes connected.

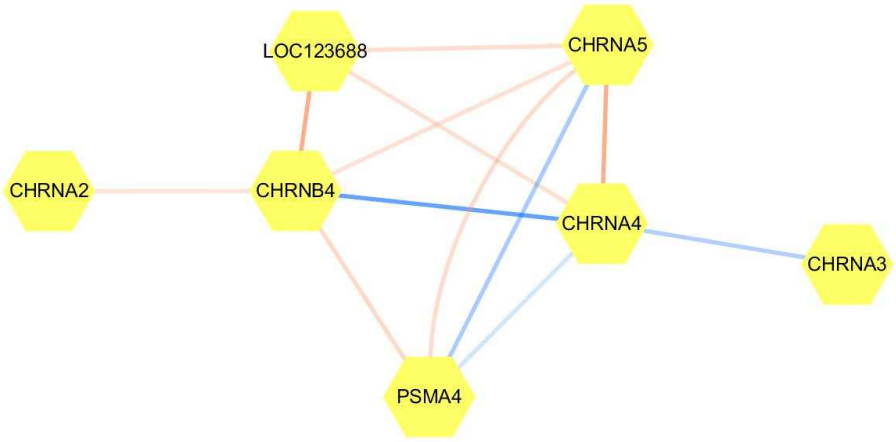


Figure 22. (Method B) Gene partial correlation network by mapping SNPs to their corresponding/nearest genes for potential gene-gene associations. Partial correlation coefficients of gene pairs with multiple SNP pair mappings would be the largest one amongst. Colors of edges represent different groups: Orange – Case; Blue – Control. Degree of transparency of edges represents the magnitude of partial correlations between the two nodes connected.

5. Exploring MicroRNA/Messenger RNA Regulatory Network on Essential Thrombocytosis

Platelets are anucleate blood cells which play an important key in haemostasis and thrombosis. Thrombocytosis is a disorder of platelet overproduction in the blood. It is classified as essential/primary thrombocytosis (ET) or reactive/secondary thrombocytosis (RT) due to the causes. Essential thrombocytosis is caused by a chronic myeloproliferative disorder with an unregulated surplus of platelets attributed to a malfunction in the body's feedback system. Complications of ET include stroke, heart attack, and formation of blood clots. To date, the genetic basis of ET is still under full investigation and no direct diagnostic tests are available (Gnatenko, et al. 2005).

Messenger RNA (mRNA) is an RNA molecule that is transcribed from a DNA template. It brings the genetic information and acts as the template in the process of protein synthesis (Kozak Mar. 1983). MicroRNA (miRNA) is single-stranded 21 to 23 nucleotide RNA molecule, which targets mRNAs through complementary pairing to the 3'-untranslated region (UTR) of mRNAs (Edelstein and Bray 2011) and regulates mRNA translation or stability (Filipowicz, Bhattacharyya and Sonenberg 2008). miRNAs have effects on protein synthesis through regulating mRNA destabilization or translational repression (Filipowicz, et al. 2008).

In this study, we have explored the potential miRNA/mRNA regulatory networks associated to essential thrombocytosis based on a 43-member cohort, through a combination of data-driven and knowledge-based analyses.

5.1. Analysis Method Introduction

5.1.1. Canonical Correlation Analysis (CCA)

Introduced by Hotelling in 1936 (Hotelling 1936), canonical correlation between two variable sets looks for the weighted combination of all variables within each variable set such that the correlation of the two combinations is maximized. The weighted combinations are called canonical variables or components. Canonical correlation is considered as a general model since it can be used when both the dependent and independent variables are either continuous or categorical data.

Consider an $n * p$ matrix X and an $n * q$ matrix Y. Without loss of generality, we assume $p < q$. Canonical correlation analysis (CCA) (Hotelling 1936) seeks for coefficient vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, such that the correlation between the linear combinations $\omega = \boldsymbol{u}'X$ and $\xi = \boldsymbol{v}'Y$ is maximized, i.e.

$$\max_{\boldsymbol{u},\boldsymbol{v}} Corr(\omega, \xi) = \max_{\boldsymbol{u},\boldsymbol{v}} \frac{\boldsymbol{u}'\Sigma_{XY}\boldsymbol{v}}{\sqrt{\boldsymbol{u}'\Sigma_{XX}\boldsymbol{u}}\sqrt{\boldsymbol{v}'\Sigma_{YY}\boldsymbol{v}}}$$

where Σ_{XX}, Σ_{YY}, and Σ_{XY} are the variance for X, Y, and the covariance for X and Y, respectively. It is attained by the canonical variate pairs

$$\omega = \boldsymbol{u}'X = e'\Sigma_{XX}{}^{-\frac{1}{2}}X; \qquad \xi = \boldsymbol{v}'Y = f'\Sigma_{YY}{}^{-\frac{1}{2}}Y$$

with e and f from the singular value decomposition (SVD) of a matrix K given by

$K = \Sigma_{XX}{}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}{}^{-\frac{1}{2}} = eDf'$ (Parkhomenko, Tritchler and Beyene 2007).

5.1.2. Sparse Supervised Canonical Correlation Analysis (Sparse sCCA)

In canonical correlation analysis, all variables are included in the linear combinations, yet for genetic data obtained via microarray studies or other high throughput methods, the number of variables usually surpasses tens of thousands, exceeding the number of study subjects. Thus the fitted linear combinations may not be easily interpreted and the application of standard algorithms may fail. These problems may be solved by introducing sparse loadings in the canonical components. Motivated by this idea, sparse canonical correlation analysis (SCCA) has been firstly proposed in 2007 (Parkhomenko, et al. 2007) and has been extended and widely applied in the genetic area. The idea of SCCA in the field of genetics is consistent with the belief that only a small section of genes are expressed under a certain conditions.

Sparse canonical correlation analysis can be accomplished via different methods, all of which has gained successes in studies of high-dimensional genomic data.

Parkhomenko et al. have firstly proposed an iterative algorithm to approximate singular vectors through soft-thresholding and applied it for the exploration of relationships between correlated sets of genome-wide SNP data and gene expression phenotypes (Parkhomenko, et al. 2007). Waaijenborg et al. have adapted the elastic net penalty to the estimates of canonical vectors and have successfully performed it to the examination of associations between gene expression and DNA markers data (Waaijenborg, Verselewel de Witt Hamer and Zwinderman 2008). Witten et al. in 2009 have introduced a regularized version of singular value decomposition (SVD) with the use of L_1 and/or fused lasso penalties and have then investigated genetic data of the same set of subjects obtained from multiple assays (Witten, Tibshirani and Hastie 2009a). In the same year, Waaijenborg et al. took a step further to incorporate ridge and elastic net penalties in SCCA for the identification of pathway genes (Waaijenborg and Zwinderman 2009). SCCA has also been extended to include more than two sets of variables to address the need of high-throughput data by Lee et al. (Lee, Lee, Lee and Pawitan 2011).

Based on the foundation of SCCA, Witten and Tibshirani (Witten and Tibshirani 2009b) have further presented “sparse supervised canonical correlation analysis (sparse sCCA)”, targeting on finding the sparse linear combinations of the two variable sets that are correlated with each other and also associated with the trait of interest.

Still consider an $n * p$ matrix X and an $n * q$ matrix Y, and assume that the columns of X and Y have been standardized with mean 0 and standard deviation 1.

Suppose in addition a categorical outcome vector $z \in \mathbb{R}^n$. The estimates of canonical vectors are defined as

$$\max_{u,v} u^T X^T Y v, \text{ subject to}$$

$$\|u\|^2 \leq 1, \|v\|^2 \leq 1, P_1(u) = \|u\|_1 \leq c_u, P_2(v) = \|v\|_1 \leq c_v,$$

$$u_j = 0 \; \forall \, j \notin Q_u, v_j = 0 \; \forall \, j \notin Q_v,$$

where P_1 and P_2 are convex penalty functions; c_u and c_v are assumed to be $1 \leq c_u \leq \sqrt{p}$ and $1 \leq c_v \leq \sqrt{q}$; Q_u and Q_v are the sets of variables with highest univariate association with the outcome z in X and Y, respectively; the threshold for variables to be included in Q_u and Q_v can either be fixed or be defined as a tuning parameter. u and v are obtained using an iterative algorithm with soft-thresholding. We have performed this sparse sCCA method on our genetic data set to investigate whether the expression of miRNA would have a significant effect on that of genes and vice versa.

5.2. Data Analysis

5.2.1. Data Structure and Processing

Our data included two data sets: 354 platelet-specific mRNA data from custom array and 939 miRNA data from Agilent microarray (Santa Clara, CA), which were paired with each other from 13 patients with essential thrombocytosis (ET) disease and 30 control subjects (*Table 20*).

Table 20. Data structure

	ET	Control	Total
# of Paired Subjects	13	30	43
	miRNA	**mRNA**	**Subject**
Total # of Items	939	354	43

Data were preprocessed and analyzed using R 2.15.3 with Bioconductor packages (http://www.bioconductor.org/). The original miRNA data set was filtered by two steps: The first step was performed using spot-flagging information provided by Agilent Feature Extraction Software (Glenda Delenstarr and Nair), keeping only miRNAs with more than 70% non-absent cells in any group. Next, miRNAs with more than 40% missing values in the sample sets were also been singled out. For mRNA data, proportion of missing expression data in the sample set for each mRNA was calculated and those with 50% or more absent data have been excluded. In addition, potential outlier was checked and filtered with a criteria of 3 standard deviations from the mean expression value. In both data sets, quantile normalization was applied to correct between-array variation (Pradervand, et al. 2009), followed by K-nearest neighbors algorithm for imputing missing expression data.

After data filtering and processing, there were totally 354 platelet-specific mRNAs and 392 miRNAs left. We further applied significance analysis of microarrays (SAM) (Chu, Li, Narasimhan, Tibshirani and Tusher 2001) on miRNA data to identify differentially expressed miRNAs between two groups. Established in 2001, SAM is a powerful

statistical technique aiming at determining significant changes in a set of microarray experiments. For each gene, specific empirical t-test was performed, followed by permutations of repeated measurements to identify the false discovery rate (FDR). By setting criteria with FDR equal to 0.05 and threshold of fold change (FC; ET versus control) to be 2 (or ½), SAM has pinpointed 50 significant miRNAs (*Figure 23; Table 21-21*).

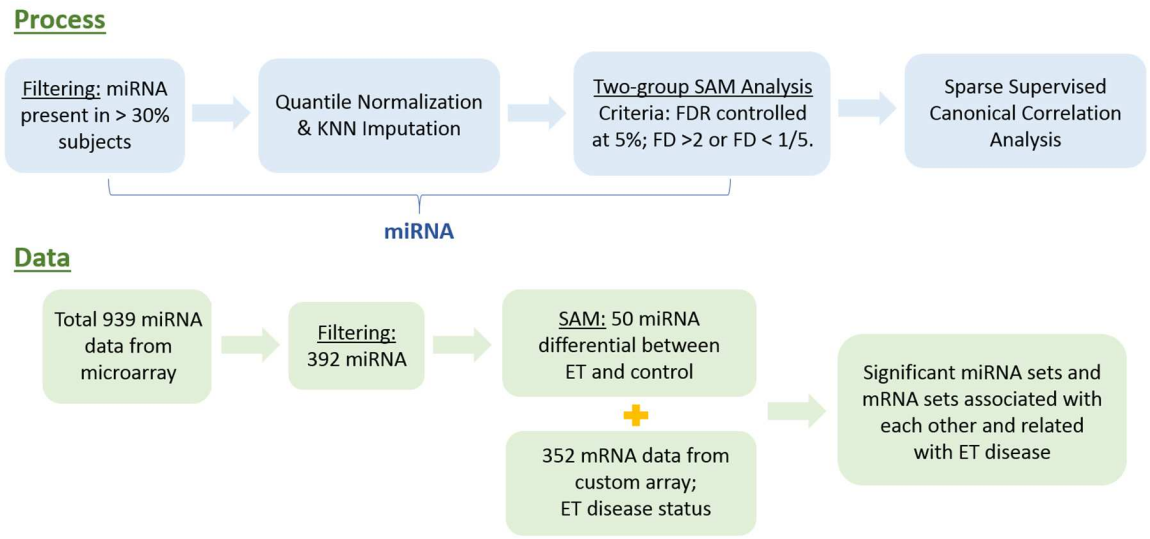


Figure 23. Workflow of data processing and analysis. (ET: Essential Thrombocytosis; SAM: Significance analysis of microarrays; FDR: False discovery rate; FD: Fold change (ET versus control))

Table 21. miRNAs identified by SAM up-regulated in ET group [Fold change (ET versus control) > 2] (28 miRNAs in total)

miRNA	Fold Change	miRNA	Fold Change
hsa-miR-490-5p	9.729	**hsa-miR-1274b**	2.391
hsa-miR-490-3p	6.085	**hsa-miR-1914***	2.356
hsa-miR-34a	4.707	**hsa-miR-29b-1***	2.348
hsa-miR-34b*	3.990	**hsa-miR-299-5p**	2.330
hsa-miR-9	3.773	**hsa-miR-424***	2.166
hsa-miR-424	3.183	**hsa-miR-493***	2.166
hsa-miR-1274a	2.982	**hsa-miR-487b**	2.155
hsa-miR-9*	2.912	**hsa-miR-449a**	2.143
hsa-miR-148a	2.791	**hsa-miR-656**	2.143

hsa-miR-1308	2.742	**hsa-miR-127-3p**	2.124
hsa-miR-1260	2.723	**hsa-miR-625**	2.119
hsa-miR-380	2.575	**hsa-miR-379***	2.073
hsa-miR-148a*	2.570	**hsa-miR-548c-5p**	2.059
hsa-miR-550	2.427	**hsa-miR-1287**	2.038

Table 22. miRNAs identified by SAM down-regulated in ET group [Fold change (ET versus control) < ½] (22 miRNAs in total)

miRNA	Fold Change	miRNA	Fold Change
hsa-miR-219-5p	0.499	**hsa-miR-181a**	0.417
hsa-miR-181a*	0.498	**hsa-miR-181c**	0.394
hsa-miR-196b	0.496	**hsa-miR-144***	0.383
hsa-miR-342-5p	0.495	**hsa-miR-33a**	0.383
hsa-miR-28-3p	0.492	**hsa-let-7d***	0.376
hsa-miR-10a	0.492	**hsa-miR-1301**	0.361
hsa-miR-328	0.489	**hsa-miR-182**	0.345
hsa-miR-106b*	0.487	**hsa-miR-150**	0.315
hsa-miR-423-5p	0.474	**hsa-miR-181c***	0.315
hsa-miR-101*	0.470	**hsa-miR-144**	0.262
hsa-miR-330-3p	0.440	**hsa-miR-551b**	0.218

5.2.2. Result of Data Analysis

With the 50 selected miRNAs as one variable set, all 354 mRNAs as the other, and the vector of subject disease status as a binary outcome vector, sparse sCCA was performed using R package "*PMA*" (Witten, et al. 2009b), aiming to identify significant miRNA sets whose expression may be associated with genomic gain or loss (changes of mRNA expression) that were also associated with essential thrombocytosis disease. In the result, one miRNA (*hsa-miR-34a*) was stood out with 10 corresponding mRNAs (*Table 23*). The canonical correlation coefficient between the two sets was 0.790.

Table 23. miRNA and mRNAs with non-zero weights by sparse sCCA

Name	Weight
hsa-miR-34a	1.000
HSD17B12	0.551
GLA	0.519
MMP1	0.448
PKIG	0.324
SERPINI1	0.241
CAV2	0.182
WASF1	0.161
NME4	0.058
TIMP1	0.030
TGFB1I1	0.014

Boxplot showing the canonical variables of both miRNA and mRNAs, stratified by disease types was given (*Figure 24*). It was clear that the values of both canonical variables were different by disease types. Scatter plot in *Figure 25* shown the relationship between two canonical variables by groups. The two variable sets were highly correlated in both ET and control group, especially in the former; and was also well separated between groups. Correlation plot (*Figure 26*)***Error! Reference source not found.*** indicated most variables were highly positively correlated pair-wisely.

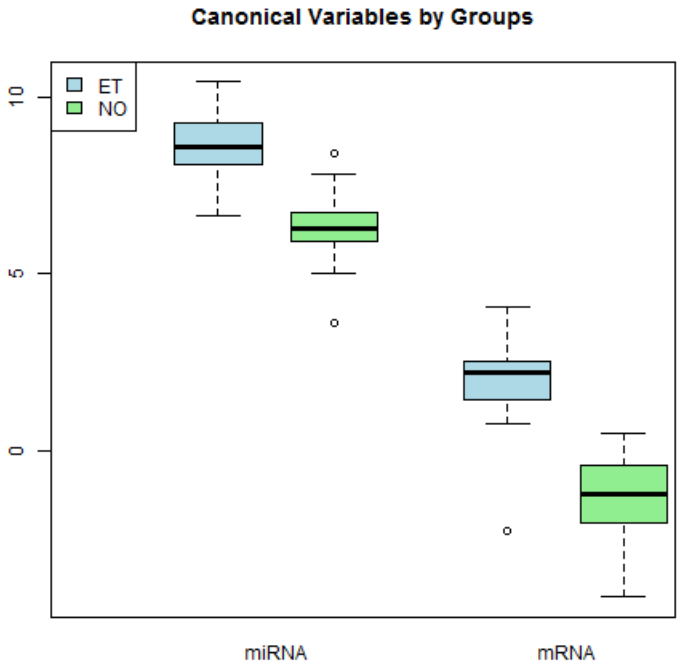


Figure 24. Boxplot of canonical variables stratified by groups. Boxes in blue represent ET subjects and boxes in green for control subjects.

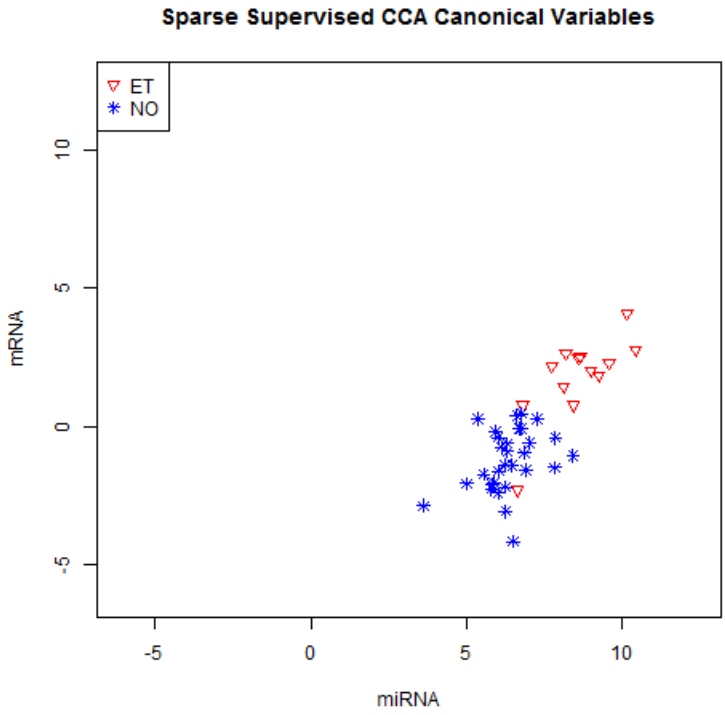


Figure 25. Scatter plot of canonical variables by groups. Points in blue represent control subjects and inverted triangles in red represent ET subjects.

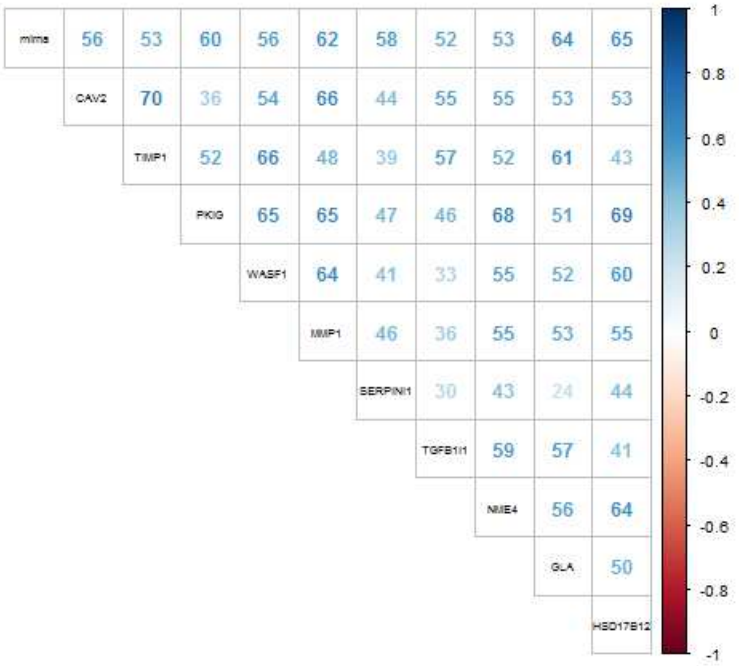


Figure 26. Upper triangle correlation plot showing pairwise correlation coefficients among variables. Numbers indicated correlation coefficients translated into percentage. The degree of transparency represented the magnitude of correlations. Numbers in blue indicated positive correlations and red for negative correlations.

To focus only on direct interaction between variables, we calculated pairwise partial correlation coefficients for each group (*Figure 27*). In ET group, when controlling other variables, almost every pair of variables were highly correlated, except for *NME4*. On the other hand, in control group, only 4 pairs of variables had significant correlations. Moreover, most of them had opposite directions comparing to themselves in ET group.

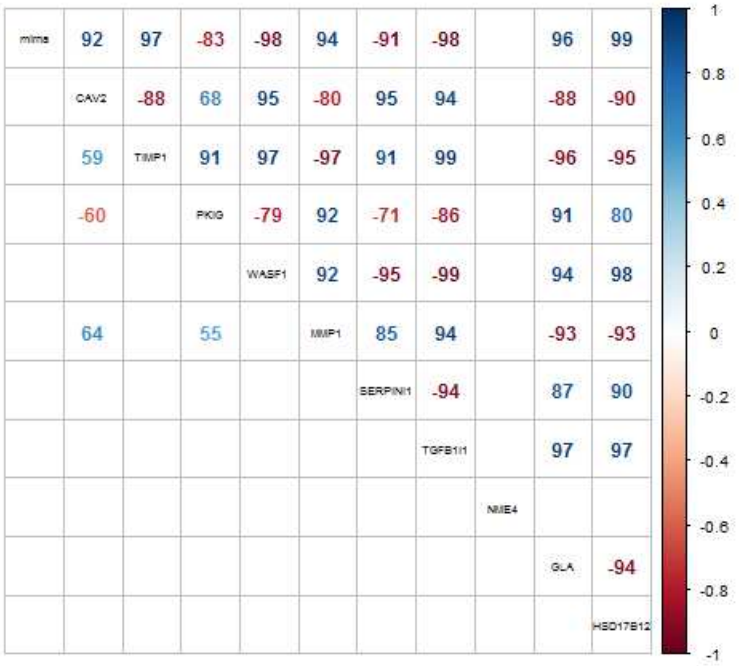


Figure 27. Correlation plot showing pairwise partial correlation coefficients among variables by groups. Upper triangle represents ET group and lower triangle for control group. Numbers indicated partial correlations translated into percentage. Multiple testing correction was applied and FDR was controlled at level of significance at 5%. Insignificant correlations were suppressed to be expressed. The degree of transparency represented the magnitude of partial correlation. Numbers in blue indicated positive partial correlations and red for negative partial correlations.

What's more, the 10 identified mRNAs along with *hsa-miR-34a* can serve as features in a multinomial logistic regression model to predict the disease type. This was confirmed by the leave-one-out cross validation result. The algorithm was as follows:

1. For $i \in 1, \ldots, N$, where N is the sample size:
    a. Split the data set into training and test data. The training data, denoted as $(X_{miRNA}^{train}, X_{mRNA}^{train}, y_i^{train})$, includes all the data except for the i^{th} subject, while

the test data, denoted as $(x_{miRNA}^{test}, x_{mRNA}^{test}, y_i^{test})$, includes only the data from the i^{th} subject.

b. Perform a generalized linear model (GLM) with the training data $(X_{miRNA}^{train}, X_{mRNA}^{train}, y_i^{train})$ to obtain coefficients B^{train}.

c. Use $(x_{miRNA}^{test}, x_{mRNA}^{test})^T \cdot \mathrm{B}^{train}$ as features in the GLM model to predict disease type y_i^{pred}.

2. Calculate the predicted rate:

$$r = \frac{1}{N} \sum_{i=1}^{N} I_{y_i^{pred} = y_i^{test}}$$

Result shown that the 11 variables (10 mRNAs and 1 miRNA) have correctly predicted the disease type of 35 samples, with a positive predicted rate equals to 81.40%. Predictions using only the 10 mRNAs produced a similar result, with 34 samples (79.07%) being correctly predicted (*Table 24*).

Table 24. Prediction results of leave-one-out cross validation

Variable Sets	**# of Subjects with Correct Prediction**	**Positive Predicted Rate (%)**
11 (*hsa-miR-34a* plus 10 mRNAs)	35	81.4
10 (mRNAs only)	34	79.07

5.3. Discussion

As in *Table 21*, the only miRNA *hsa-miR-34a* identified by sparse sCCA expressed around 4.7 times higher in ET group comparing to that in control group. It has been previously shown to express aberrantly in polycythemia vera (PV) granulocytes (Bruchova, Merkerova and Prchal 2008) and to be one of the miRNA members that expressed most differentially among ET, RT, and control groups (Xu, et al. 2012).

The platelet-expressed gene *HSD17B12*, standing for "Hydroxysteroid (17-β) Dehydrogenase 12", has been claimed previously to be associated with the distinction of ET platelets from normal ones (Gnatenko, et al. 2005). Two encoded protease/protease inhibitors – *MMP1* (Matrix Metallopeptidase 1) and *SERPINI1* (Serpin Peptidase Inhibitor, Clade I (Neuroserpin), Member 1) are a class of proteins well-associated in tumor invasiveness and cancer metastases and have both been detected over-expressed in ET group comparing to normal (Gnatenko, et al. 2005, Saito and Bunnett 2005). *MMP1* has been demonstrated to be related to inflammation in several studies (Brassart, et al. 2001, Herouy, et al. 2001, Zhang, et al. 2003, Andonovska, Dimova and Panov 2008). Moreover, members in matrix Metallopeptidase family have be indicated to be involved in the migration and invasion of leukemia cell (*MMP-2*) (He, et al. 2009); and to mediate megakaryocyte transendothelial migration and proplatelet formation (*MMP-9*) (Lane, et al. 2000).

CAV2 (Caveolin 2), *TGFB1I1* (Transforming Growth Factor Beta 1 Induced Transcript 1), and *NME4* (NME/NM23 Nucleoside Diphosphate Kinase 4) have been inferred to associate with tumors, metastasis, and multiple types of cancer. *TIMP1* (TIMP Metallopeptidase Inhibitor 1) has been discussed to be highly related to tumors, cancer metastasis and inflammation. *WASF1* (WAS Protein Family, Member 1) has been predicted as a potential target of hsa-miR-34a by a popular miRNA target prediction tools, TargetScan (http://www.targetscan.org/), which predicts regulatory targets using conserved complementary (Lewis, Burge and Bartel 2005). Moreover, the WAS protein family has been shown to be related to nucleosome and chromatin assembly, performing an important role in gene transcription that may regulate megakaryocytopoiesis and/or proplatelet formation (Schulze and Shivdasani 2004). A recent study in class prediction models of ET included a member from this family, *WASF3*, as one of the biomarkers segregating ET, RT and normal groups (Gnatenko, et al. 2010). Although instead of *WASF1*, the study pointed to *WASF3*, our result would implicate a specific role *WASF1* plays in the classification and prediction models for ET and normal cohorts.

Considering gene regulatory network, we referred to the Ingenuity Pathways Analysis (IPA) system (http://www.ingenuity.com/products/ipa), which helps build and explore transcriptional networks for researchers to gain insight into molecular interactions and disease processes. The IPA system has identified two networks with the 10 mRNA variables spotted by sparse sCCA. The network with the higher score is associated with functions of “Cardiac Hypertrophy, Cardio vascular Disease, and Developmental Disorder”, involving 9 selected mRNAs *GLA*, *HSD17B12*, *CAV2*, *MMP1*, *TIMP1*,

SERPINI1, *WASF1*, *PKIG*, and *TGFB1I1* (*Figure 28*). The other network with *NME4* is related to “Cellular Assembly and Organization, Cellular Function and Maintenance, and Nucleic Acid Metabolism” (*Figure 29*). This prediction result is remarkably consistent with the partial correlation result (*Figure 27*) that *NME4* is conditionally independent to all the other variables.

In conclusion, our data analysis on miRNA and mRNA data has predicted a close relationship of miRNA *hsa-miR-34a* and an mRNA set (including *HSD17B12*, *GLA*, *MMP1*, *PKIG*, *SERPINI1*, *CAV2*, *WASF1*, *NME4, TIMP1* and *TGFB1I1*). A majority of the identified variables have been linked to hematologic function by a sizable number of studies. Additionally, all 10 mRNAs are involved in two transcriptional networks corresponding to several essential functions. Altogether it alluded that the identified mRNA set might be considered as a contributor in the regulatory mechanism of ET disease; and the expression of miRNA *hsa-miR-34a* might had an effect on that of the mRNA set. Experiments focusing on this regulatory relationship are in demand for further confirmation.

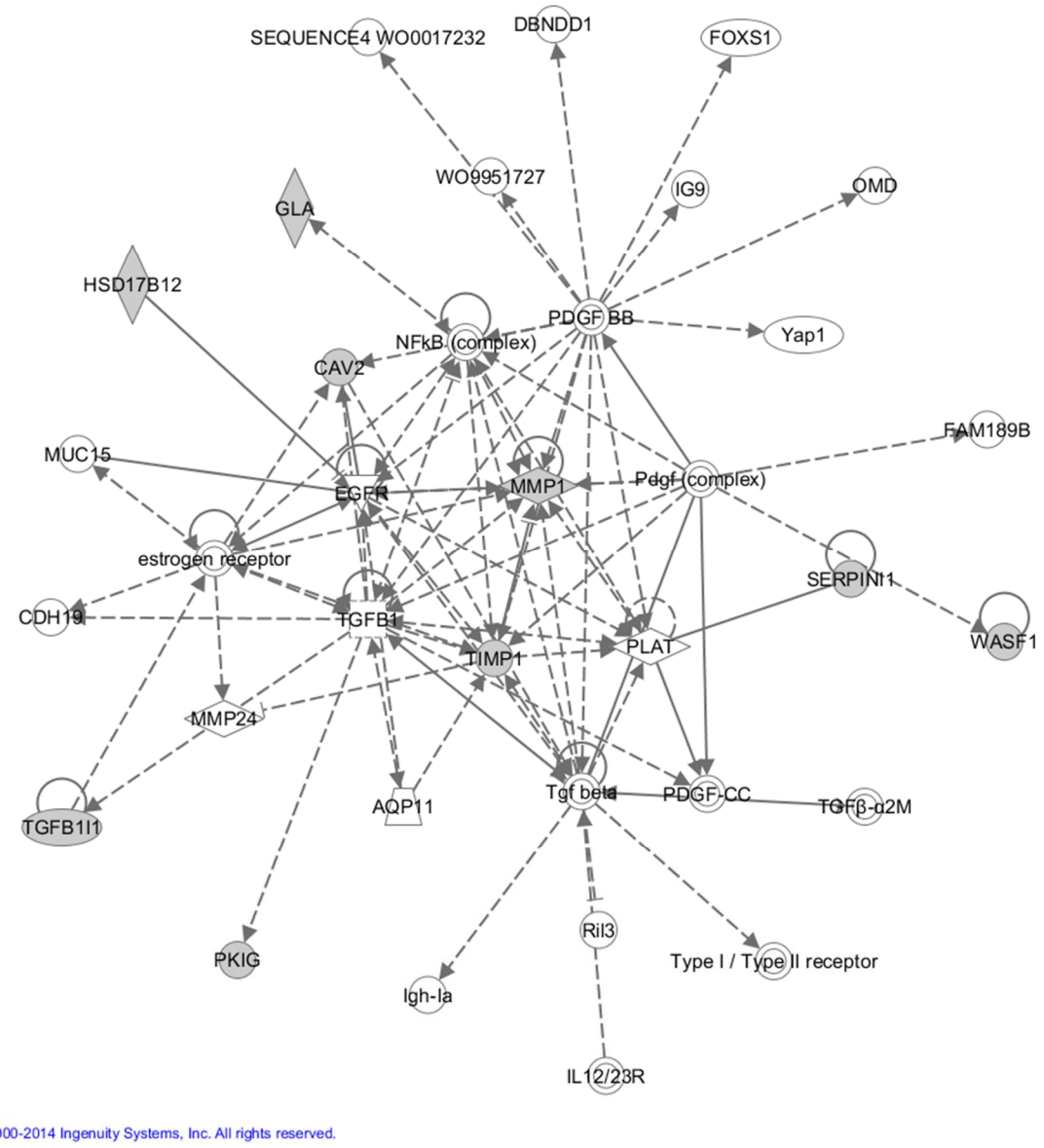


Figure 28. Structure of the first genetic network predicted by IPA system (*http://www.ingenuity.com/products/ipa*). mRNAs recognized by sparse sCCA are highlighted in shades.

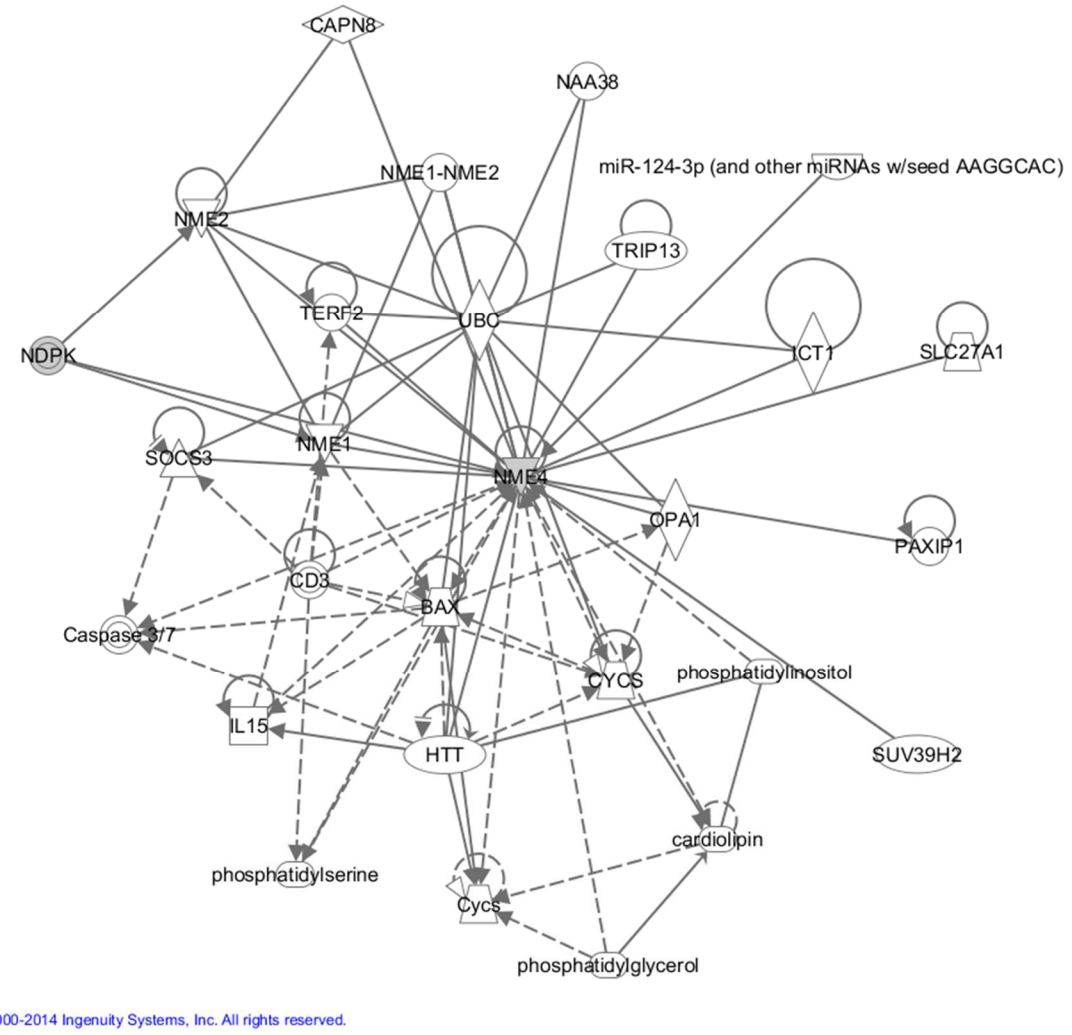


Figure 29. Structure of the second genetic network predicted by IPA system (*http://www.ingenuity.com/products/ipa*). mRNA *NME4* recognized by sparse sCCA is centered.

6. Discussion and Future Work

In 2011, Chen (Chen 2011) has proposed a new partial correlation coefficient estimating method for categorical/mixed variables. Firstly logistic regression models are performed, then the residuals are correlated via canonical correlation. The Pearson residuals are used for their asymptotic properties. It has been discussed later by a simulation study that this method and the partial phi coefficient converge in estimate and inference in a limiting case (Leong 2012); yet the former one outperforms the latter by its well-defined in the multi-categorical case and its readily capability in controlling for more than one variable. Moreover, this new estimating method can in a natural manner be extended to embrace mixed variables, measuring the relationship between continuous and categorical variables, which raises its potential as a powerful tool in the analysis of GWA studies.

In spite of the exciting fact, currently this method is not designed for the high-dimensional situation ($n \ll p$), and did not obtain reasonably sparse structure with GWA study data set in Chen's study. Accordingly, we introduced here a novel two-stage sequential analysis framework to approach this problem. The approach is a combination of penalized logistic regression model based on grouping SNPs and partial correlation network analysis. In this thesis, we have shown that it can naturally incorporate information of either variable association defined by pairwise canonical correlation

measurement, or of biological information, such as gene mapping, both of which have satisfying performance in simulation studies and/or the real data example. Other information, such as linkage disequilibrium, rare/common variants, pathway, etc. can also be easily embedded according to the desires of researchers. In numerical simulations of both low- and high-dimensional settings, this two-stage approach is, in general, found to be more powerful and at the same time less conservative than the traditional stepwise penalized logistic regression model. We illustrated the approach to a GWA study data set COGEND. The approach has identified the close relationship relating to nicotine dependence among genes in the so-called "Chromosome 15q25.1 region", after controlling all other SNPs/genes. The effects of genes in this region on nicotine dependence have been described and well discussed by many previous studies. We also inferred potential disease-susceptible interactions of gene *CHRNA4* and genes in the "Chromosome 15q25.1 region".

A sequential analysis was also applied in Chen's study in GWA study data set by combining clustering and network techniques. First, hierarchical clustering was conducted according to pairwise canonical correlation measurement. Then within each cluster, SNP with overall highest canonical correlation (similarities) to all the others was selected. Following this, PCNA was performed considering these "representative SNPs". Our approach has polished this stage-wise method by embedding the sparse-group lasso penalty and thus taking account of both group-wise and within group sparsity of genetic data, as well as integrating variable association and biological information into the analysis.

SGL has been used to data with high-dimensional predictors for the identification of genomic regions (Peng, et al. 2010), and has also been extended to multinomial classifier (Vincent and Hansen 2012) and network-based Cox regression model (Zhang, et al. 2013). Zhou et al. (Zhou, et al. 2010) have firstly applied the idea of "generalized linear models with mixed group and lasso penalties" on the arena of GWA studies, identifying rare variants by grouping SNPs into genes. Later on, Ayers and Cordell (Ayers, et al. 2013) have performed the sparse-group lasso to identify grouped common and rare variants in GWA studies, incorporating weighting methods to allow different contributions of variants. Focusing on the potential functional relationships between gene variants, Silver et al. (Silver, et al. 2013) conducted SGL to simultaneously identify pathways and genes that are related to the trait of interest. SNPs were grouped according to the prior mapping information to gene pathways. Since one SNP may be associated to more than one pathway, they also discussed a modification to deal with overlapping groups. However impressive their work has been, in our study, SGL was the first time introduced to PCNA for the detection of significant SNP-SNP associations in the GWA study arena, with the ability of incorporating with not only biological information but also pairwise association measurements among SNPs.

Partial correlation network analysis is a data-driven method and no assumptions about the network structure are required for the initiation of analysis. The goal of our approach is not to infer the network correctly; but instead to develop new hypotheses of associations between variants with confidence. This approach could be seen as a post-

GWA study method to explain more underlying genetic variation and unravel genome-wide genetic associations or biologically relevant pathways.

In all, this novel two-stage approach has several advantages:

a. Analyze all variables at once to consider the impact of one variable on another and to improve restricted power of single-locus analysis for detecting SNPs with small or moderate effects.

b. Reduce data dimension via sparse regression to keep the multiple testing problem in a more benign form and to identify potential target variables simultaneously.

c. Allow for the incorporation of biological information and/or variable association into the analysis by forming SNP clusters.

d. Focus on conditional dependence between variables and graphically present the differential connections between case and control groups.

e. Provide easy and direct results for a meaningful biological interpretation indicating potential gene-gene associations.

We have also introduced the application of sparse techniques to canonical correlation analysis for the establishment of regulatory network among genetic data. One miRNA together with ten mRNAs was pinpointed to be associated with essential thrombocytosis (ET), which has been verified via leave-one-out cross validation. Two

networks have been predicted to be potentially related to the genetic regularization basis of ET with the help of a network exploration system.

The following aspects of PCNA deserve main attentions in our future study, concerning especially its application in the arena of GWA studies. First is the test of significance for partial correlation pair. In most circumstances, GWA studies are in the case-control design. Tests discovering which edges are significantly different between groups would address the question whether the trait of interest is affected by the strength of variant pair associations. Equally important, PCNA accounting for nonlinear relationships should also be brought into focus. Various nonlinear regression models could be introduced into the analysis and new terminology is of necessity since we would be no longer dealing with partial correlations as traditionally defined.

Bibliography

Altshuler, D., et al. (2000), "The Common Pparγ Pro12ala Polymorphism Is Associated with Decreased Risk of Type 2 Diabetes," *Nature genetics*, 26, 76-80.

Amos, C. I., et al. (2010), "Nicotinic Acetylcholine Receptor Region on Chromosome 15q25 and Lung Cancer Risk among African Americans: A Case–Control Study," *J Natl Cancer Inst*, 102, 1199-1205.

Amos, C. I., et al. (2008), "Genome-Wide Association Scan of Tag Snps Identifies a Susceptibility Locus for Lung Cancer at 15q25. 1," *Nature genetics*, 40, 616-622.

Andonovska, B., Dimova, C., and Panov, S. (2008), "Matrix Metalloproteinases (Mmp-1, -8, -13) in Chronic Periapical Lesions," *Vojnosanit Pregl*, 65, 882-886.

Arnaud-Lopez, L., et al. (2008), "< I> Phosphodiesterase 8b</I> Gene Variants Are Associated with Serum Tsh Levels and Thyroid Function," *The American Journal of Human Genetics*, 82, 1270-1280.

Ayers, K. L., and Cordell, H. J. (2010), "Snp Selection in Genome-Wide and Candidate Gene Studies Via Penalized Logistic Regression," *Genetic epidemiology*, 34, 879-891.

Ayers, K. L., and Cordell, H. J. (2013), "Identification of Grouped Rare and Common Variants Via Penalized Logistic Regression," *Genetic epidemiology*, 37, 592-602.

Bae, E. Y., et al. (2012), "Replication of Results of Genome-Wide Association Studies on Lung Cancer Susceptibility Loci in a Korean Population," *Respirology*, 17, 699-706.

Balding, D. J. (2006), "A Tutorial on Statistical Methods for Population Association Studies," *Nature Reviews Genetics*, 7, 781-791.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Berrettini, W., et al. (2008), "A-5/A-3 Nicotinic Receptor Subunit Alleles Increase Risk for Heavy Smoking," *Molecular psychiatry*, 13, 368-373.

Bierut, L., et al. (2008), "Variants in Nicotinic Receptors and Risk for Nicotine Dependence," *American Journal of Psychiatry*, 165, 1163-1171.

Bland, J. M., and Altman, D. G. (1995), "Multiple Significance Tests: The Bonferroni Method," *Bmj*, 310, 170.

Brassart, B., et al. (2001), "Conformational Dependence of Collagenase (Matrix Metalloproteinase-1) up-Regulation by Elastin Peptides in Cultured Fibroblasts," *Journal of Biological Chemistry*, 276, 5222-5227.

Breitling, L., et al. (2009), "Association of Nicotinic Acetylcholine Receptor Subunit A4 Polymorphisms with Nicotine Dependence in 5500 Germans," *The pharmacogenomics journal*, 9, 219-224.

Broms, U., et al. (2012), "Analysis of Detailed Phenotype Profiles Reveals Chrna5-Chrna3-Chrnb4 Gene Cluster Association with Several Nicotine Dependence Traits," *Nicotine & Tobacco Research*, 14, 720-733.

Bruchova, H., Merkerova, M., and Prchal, J. T. (2008), "Aberrant Expression of Microrna in Polycythemia Vera," *Haematologica*, 93, 1009-1016.

Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C.-H. (2013), "Correlated Variables in Regression: Clustering and Sparse Estimation," *Journal of Statistical Planning and Inference*, 143, 1835-1858.

Burton, P. R., et al. (2007), "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls," *Nature*, 447, 661-678.

Bush, W. S., and Moore, J. H. (2012), "Genome-Wide Association Studies," *PLoS computational biology*, 8, e1002822.

Caporaso, N., et al. (2009), "Genome-Wide and Candidate Gene Association Study of Cigarette Smoking Behaviors," *PLoS One*, 4, e4653.

Chen, H. (2011), "*Clustering and Network Analysis with Single Nucleotide Polymorphism (Snp)*," Stony Brook University, Applied Mathematics and Statistics.

Chial, H. (2008), "Mendelian Genetics: Patterns of Inheritance and Single-Gene Disorders," *Nature Education*, 1, 3775-3781.

Cho, J. H. (2010), "Genome-Wide Association Studies: Present Status and Future Directions," *Gastroenterology*, 138, 1668-1672. e1661.

Cho, S., Kim, H., Oh, S., Kim, K., and Park, T. (2009), "Elastic-Net Regularization Approaches for Genome-Wide Association Studies of Rheumatoid Arthritis," in *BMC proceedings*, BioMed Central Ltd, p. S25.

Chu, G., Li, J., Narasimhan, B., Tibshirani, R., and Tusher, V. (2001), "Significance Analysis of Microarrays Users Guide and Technical Document."

COGEND. (2013), "Cogend Collaborative Genetic Study of Nicotine Dependence ".

Coleman, H. R., Chan, C.-C., Ferris III, F. L., and Chew, E. Y. (2008), "Age-Related Macular Degeneration," *The Lancet*, 372, 1835-1845.

Collins, F. S., Brooks, L. D., and Chakravarti, A. (1998), "A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation," *Genome Research*, 8, 1229-1231.

Consortium, I. H. (2010), "Integrating Common and Rare Genetic Variation in Diverse Human Populations," *Nature*, 467, 52-58.

Cordell, H. J. (2009), "Detecting Gene–Gene Interactions That Underlie Human Diseases," *Nature Reviews Genetics*, 10, 392-404.

Cordell, H. J., and Clayton, D. G. (2002), "A Unified Stepwise Regression Procedure for Evaluating the Relative Effects of Polymorphisms within a Gene Using Case/Control or Family Data: Application to< I> Hla</I> in Type 1 Diabetes," *The American Journal of Human Genetics*, 70, 124-141.

Czepiel, S. A. (2002), "Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation," *Available at czep. net/stat/mlelr. pdf*.

David, S., et al. (2012), "Genome-Wide Meta-Analyses of Smoking Behaviors in African Americans," *Translational psychiatry*, 2, e119.

De La Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004), "Discovery of Meaningful Associations in Genomic Data Using Partial Correlation Coefficients," *Bioinformatics*, 20, 3565-3574.

Easton, D. F., et al. (2007), "A Systematic Genetic Assessment of 1,433 Sequence Variants of Unknown Clinical Significance in the< I> Brca1</I> and< I> Brca2</I> Breast Cancer–Predisposition Genes," *The American Journal of Human Genetics*, 81, 873-883.

Easton, D. F., and Eeles, R. A. (2008), "Genome-Wide Association Studies in Cancer," *Human Molecular Genetics*, 17, R109-R115.

Edelstein, L. C., and Bray, P. F. (2011), "Micrornas in Platelet Production and Activation," *Blood*, 117, 5289-5296.

Efron, B., and Tibshirani, R. J. (1994), *An Introduction to the Bootstrap* (Vol. 57), CRC press.

Faraone, S. V., et al. (2004), "A Novel Permutation Testing Method Implicates Sixteen Nicotinic Acetylcholine Receptor Genes as Risk Factors for Smoking in Schizophrenia Families," *Human Heredity*, 57, 59-68.

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008), "Mechanisms of Post-Transcriptional Regulation by Micrornas: Are the Answers in Sight?," *Nat Rev Genet*, 9, 102-114.

Fisher, R. A. (1915), "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika*, 10, 507-521.

Fisher, R. A. (1924), "The Distribution of the Partial Correlation Coefficient," *Metron*, 3, 329-332.

Freimer, N. B., and Sabatti, C. (2007), "Human Genetics: Variants in Common Diseases," *Nature*, 445, 828-830.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The annals of applied statistics*, 1, 302-332.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 9, 432-441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "A Note on the Group Lasso and a Sparse Group Lasso," *arXiv preprint arXiv:1001.0736*.

Gatz, M., et al. (1997), "Heritability for Alzheimer's Disease: The Study of Dementia in Swedish Twins," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 52, M117-M125.

Gibbs, R. A., et al. (2003), "The International Hapmap Project," *Nature*, 426, 789-796.

Glenda Delenstarr, S. V., Mark Hartnett, Condie Carmack, William Love, and, and Nair, M., "Evaluating the Reproducibility of

Microarray Technology," *Agilent Technologies, Inc.*

Gnatenko, D. V., et al. (2005), "Platelets Express Steroidogenic 17beta-Hydroxysteroid Dehydrogenases. Distinct Profiles Predict the Essential Thrombocythemic Phenotype," *THROMBOSIS AND HAEMOSTASIS-STUTTGART-*, 94, 412.

Gnatenko, D. V., et al. (2010), "Class Prediction Models of Thrombocytosis Using Genetic Biomarkers," *Blood*, 115, 7-14.

Greenbaum, L., and Lerer, B. (2009), "Differential Contribution of Genetic Variation in Multiple Brain Nicotinic Cholinergic Receptors to Nicotine Dependence: Recent Progress and Emerging Open Questions," *Molecular psychiatry*, 14, 912-945.

Greenberg, D. A. (1993), "Linkage Analysis of" Necessary" Disease Loci Versus" Susceptibility" Loci," *Am J Hum Genet*, 52, 135.

Han, S., et al. (2011), "Association of Chrna4 Polymorphisms with Smoking Behavior in Two Populations," *Am J Med Genet B Neuropsychiatr Genet*, 156B, 421-429.

Hansen, H. M., et al. (2010), "Fine Mapping of Chromosome 15q25. 1 Lung Cancer Susceptibility in African-Americans," *Human Molecular Genetics*, 19, 3652-3661.

He, Q., and Lin, D.-Y. (2011), "A Variable Selection Method for Genome-Wide Association Studies," *Bioinformatics*, 27, 1-8.

He, Y., et al. (2009), "[Role of Wave1 in K562 Leukemia Cells Invasion and Its Mechanism]," *Zhonghua xue ye xue za zhi= Zhonghua xueyexue zazhi*, 30, 237-241.

Health, U. D. o., and Services, H. (2014), "The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General," *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health*, 17.

Herouy, Y., et al. (2001), "Inflammation in Stasis Dermatitis Upregulates Mmp-1, Mmp-2 and Mmp-13 Expression," *Journal of dermatological science*, 25, 198-205.

Hindorff LA, M. J., Morales J, Junkins HA, Hall PN, Klemm AK, and Manolio TA, "A Catalog of Published Genome-Wide Association Studies," 2014.

Hindorff, L. A., et al. (2009), "Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits," *Proceedings of the National Academy of Sciences*, 106, 9362-9367.

Hodge, S. E. (1994), "What Association Analysis Can and Cannot Tell Us About the Genetics of Complex Disease," *American journal of medical genetics*, 54, 318-323.

Hoerl, A. E. (1962), "Application of Ridge Analysis to Regression Problems," *Chemical Engineering Progress*, 58, 54-59.

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55-67.

Hotelling, H. (1936), "Relations between Two Sets of Variates," *Biometrika*, 28, 321-377.

Huang, J., Breheny, P., and Ma, S. (2012), "A Selective Review of Group Selection in High-Dimensional Models," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27.

Huang, J., and Zhang, T. (2010), "The Benefit of Group Sparsity," *The Annals of Statistics*, 38, 1978-2004.

Hung, R. J., et al. (2008), "A Susceptibility Locus for Lung Cancer Maps to Nicotinic Acetylcholine Receptor Subunit Genes on 15q25," *Nature*, 452, 633-637.

Hutchison, K. E., et al. (2007), "Chrna4 and Tobacco Dependence: From Gene Regulation to Treatment Outcome," *Archives of general psychiatry*, 64, 1078-1086.

Kim, Y., Kim, J., and Kim, Y. (2006), "Blockwise Sparse Regression," *Statistica Sinica*, 16, 375.

Klein, R. J., et al. (2005), "Complement Factor H Polymorphism in Age-Related Macular Degeneration," *Science*, 308, 385-389.

Kogelman, L. J., and Kadarmideen, H. N. (2014), "Weighted Interaction Snp Hub (Wish) Network Method for Building Genetic Networks for Complex Diseases and Traits Using Whole Genome Genotype Data," *BMC Systems Biology*, 8, 1-12.

Kooperberg, C., LeBlanc, M., and Obenchain, V. (2010), "Risk Prediction Using Genome-Wide Association Studies," *Genetic epidemiology*, 34, 643-652.

Kozak, M. (Mar. 1983), "Comparison of Initiation of Protein Synthesis in Procaryotes, Eucaryotes, and Organelles.," *MICROBIOLOGICAL REVIEWS*, 47, 1-45.

Krämer, N., Schäfer, J., and Boulesteix, A.-L. (2009), "Regularized Estimation of Large-Scale Gene Association Networks Using Graphical Gaussian Models," *BMC Bioinformatics*, 10, 384.

Lane, W. J., et al. (2000), "Stromal-Derived Factor 1–Induced Megakaryocyte Migration and Platelet Production Is Dependent on Matrix Metalloproteinases," *Blood*, 96, 4152-4159.

Langfelder, P., Zhang, B., and Horvath, S. (2008), "Defining Clusters from a Hierarchical Cluster Tree: The Dynamic Tree Cut Package for R," *Bioinformatics*, 24, 719-720.

Lee, W., Lee, D., Lee, Y., and Pawitan, Y. (2011), "Sparse Canonical Covariance Analysis for High-Throughput Data," *Statistical applications in genetics and molecular biology*, 10, 1-24.

Leong, S. H. Y. (2012), "*Partial Correlation Network Analysis for Mixed Data*," Stony Brook University, Applied Mathematics and Statistics.

Lettre, G., and Rioux, J. D. (2008), "Autoimmune Diseases: Insights from Genome-Wide Association Studies," *Human Molecular Genetics*, 17, R116-R121.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005), "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are Microrna Targets," *Cell*, 120, 15-20.

Li, C., et al. (2011), "Disease-Driven Detection of Differential Inherited Snp Modules from Snp Network," *Gene*, 489, 119-129.

Li, H., and Gui, J. (2006), "Gradient Directed Regularization for Sparse Gaussian Concentration Graphs, with Applications to Inference of Genetic Networks," *Biostatistics*, 7, 302-317.

Li, M. D., et al. (2005), "Ethnic-and Gender-Specific Association of the Nicotinic Acetylcholine Receptor A4 Subunit Gene (Chrna4) with Nicotine Dependence," *Human Molecular Genetics*, 14, 1211-1219.

Li, M. D., and Burmeister, M. (2009), "New Insights into the Genetics of Addiction," *Nat Rev Genet*, 10, 225-231.

Li, M. D., Lou, X. Y., Chen, G., Ma, J. Z., and Elston, R. C. (2008), "Gene-Gene Interactions among Chrna4, Chrnb2, Bdnf, and Ntrk2 in Nicotine Dependence," *Biol Psychiatry*, 64, 951-957.

Li, M. D., et al. (2010), "Association and Interaction Analysis of Variants in Chrna5/Chrna3/Chrnb4 Gene Cluster with Nicotine Dependence in African and European Americans," *Am J Med Genet B Neuropsychiatr Genet*, 153B, 745-756.

Lim, M., and Hastie, T. (2013), "Learning Interactions through Hierarchical Group-Lasso Regularization," *arXiv preprint arXiv:1308.2719*.

Lin, H.-Y., et al. (2013), "Snp-Snp Interaction Network in Angiogenesis Genes Associated with Prostate Cancer Aggressiveness," *PLoS One*, 8, e59688.

Liu, Y., et al. (2009), "Haplotype and Cell Proliferation Analyses of Candidate Lung Cancer Susceptibility Genes on Chromosome 15q24-25.1," *Cancer Res*, 69, 7844-7850.

Lou, X.-Y., et al. (2007), "A Generalized Combinatorial Approach for Detecting Gene-by-Gene and Gene-by-Environment Interactions with Application to Nicotine Dependence," *The American Journal of Human Genetics*, 80, 1125-1137.

Lu, C., Latourelle, J., O'Connor, G. T., Dupuis, J., and Kolaczyk, E. D. (2013), "Network-Guided Sparse Regression Modeling for Detection of Gene-by-Gene Interactions," *Bioinformatics*, 29, 1241-1249.

MacDonald, M. E., et al. (1992), "The Huntington's Disease Candidate Region Exhibits Many Different Haplotypes," *Nature genetics*, 1, 99-103.

Malo, N., Libiger, O., and Schork, N. J. (2008), "Accommodating Linkage Disequilibrium in Genetic-Association Analyses Via Ridge Regression," *The American Journal of Human Genetics*, 82, 375-385.

Marks, M. J., et al. (1992), "Nicotine Binding and Nicotinic Receptor Subunit Rna after Chronic Nicotine Treatment," *The Journal of neuroscience*, 12, 2765-2784.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008), "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 53-71.

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the Lasso," *The Annals of Statistics*, 1436-1462.

Meyers, J., et al. (2013), "Interaction between Polygenic Risk for Cigarette Use and Environmental Exposures in the Detroit Neighborhood Health Study," *Translational psychiatry*, 3, e290.

NCBI, "The Genetic Architecture of Smoking and Smoking Cessation," *The Database of Genotypes and Phenotypes*.

Nesterov, Y. (2007), "Gradient Methods for Minimizing Composite Objective Function."

Norrgard, K. (2008), "Genetic Variation and Disease: Gwas," *Nature Education*, 1.

Onay, V. Ü., et al. (2006), "Snp-Snp Interactions in Breast Cancer Susceptibility," *BMC cancer*, 6, 114.

Park, M. Y., Hastie, T., and Park, M. M. Y. (2009), "Package 'Stepplr'."

Parkhomenko, E., Tritchler, D., and Beyene, J. (2007), "Genome-Wide Sparse Canonical Correlation of Gene Expression with Genotypes," in *BMC proceedings*, BioMed Central Ltd, p. S119.

Pearson, K. (1915), "On the Partial Correlation Ratio," *Proceedings of the Royal Society of London. Series A*, 91, 492-498.

Pearson, K. (1920), "Notes on the History of Correlation," *Biometrika*, 13, 25-45.

Pearson, T. A., and Manolio, T. A. (2008), "How to Interpret a Genome-Wide Association Study," *JAMA*, 299, 1335-1344.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Models," *Journal of the American Statistical Association*, 104.

Peng, J., et al. (2010), "Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer," *The annals of applied statistics*, 4, 53.

Philibert, R. A., et al. (2009), "Examination of the Nicotine Dependence (Nicsnp) Consortium Findings in the Iowa Adoption Studies Population," *Nicotine & Tobacco Research*, ntn034.

Portugal, G. S., and Gould, T. J. (2008), "Genetic Variability in Nicotinic Acetylcholine Receptors and Nicotine Addiction: Converging Evidence from Human and Animal Research," *Behav Brain Res*, 193, 1-16.

Pradervand, S., et al. (2009), "Impact of Normalization on Mirna Microarray Expression Profiling," *RNA*, 15, 493-501.

Pradhan, K. (2009), "*Partial Correlation Analysis in Functional Brain Imaging Studies*," Stony Brook University, Applied Mathematics and Statistics

Reich, D. E., and Lander, E. S. (2001), "On the Allelic Spectrum of Human Disease," *TRENDS in Genetics*, 17, 502-510.

Riordan, J. R., et al. (1989), "Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA," *Science*, 245, 1066-1073.

Risch, N., and Merikangas, K. (1996), "The Future of Genetic Studies of Complex Human Diseases," *Science*, 273, 1516-1517.

Saccone, N. L., et al. (2010), "Multiple Cholinergic Nicotinic Receptor Genes Affect Nicotine Dependence Risk in African and European Americans," *Genes, Brain and Behavior*, 9, 741-750.

Saccone, N. L., et al. (2009), "The Chrna5-Chrna3-Chrnb4 Nicotinic Receptor Subunit Gene Cluster Affects Risk for Nicotine Dependence in African-Americans and in European-Americans," *Cancer Res*, 69, 6848-6856.

Saccone, S. F., et al. (2007), "Cholinergic Nicotinic Receptor Genes Implicated in a Nicotine Dependence Association Study Targeting 348 Candidate Genes with 3713 Snps," *Human Molecular Genetics*, 16, 36-49.

Saito, T., and Bunnett, N. W. (2005), "Protease-Activated Receptors," *Neuromolecular medicine*, 7, 79-99.

Schäfer, J., and Strimmer, K. (2005), "An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks," *Bioinformatics*, 21, 754-764.

Scherf, D. B., et al. (2013), "Epigenetic Screen Identifies Genotype-Specific Promoter DNA Methylation and Oncogenic Potential of Chrnb4," *Oncogene*, 32, 3329-3338.

Schulze, H., and Shivdasani, R. A. (2004), "Molecular Mechanisms of Megakaryocyte Differentiation," in *Semin Thromb Hemost*, Copyright© 2004 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA., pp. 389-398.

Schwender, H., and Fritsch, A. (2008), "Scrime: Analysis of High-Dimensional Categorical Data Such as Snp Data," *R package version*, 1.

Silver, M., et al. (2013), "Pathways-Driven Sparse Regression Identifies Pathways and Genes Associated with High-Density Lipoprotein Cholesterol in Two Asian Cohorts," *PLoS Genet*, 9, e1003939.

Silver, M., Janousova, E., Hua, X., Thompson, P. M., and Montana, G. (2012), "Identification of Gene Pathways Implicated in Alzheimer's Disease Using Longitudinal Imaging Phenotypes with Sparse Regression," *NeuroImage*, 63, 1681-1694.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, 22, 231-245.

Stevens, V. L., et al. (2008), "Nicotinic Receptor Gene Variants Influence Susceptibility to Heavy Smoking," *Cancer Epidemiology Biomarkers & Prevention*, 17, 3517-3525.

Stevenson, J. (1992), "Evidence for a Genetic Etiology in Hyperactivity in Children," *Behavior genetics*, 22, 337-344.

Stigler, S. M. (1989), "Francis Galton's Account of the Invention of Correlation," *Statistical Science*, 4, 73-79.

Stratton, M. R., and Rahman, N. (2007), "The Emerging Landscape of Breast Cancer Susceptibility," *Nature genetics*, 40, 17-22.

Swan, G. E., et al. (2006), "A Genome-Wide Screen for Nicotine Dependence Susceptibility Loci," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 141, 354-360.

Tapper, A. R. (2004), "4* Receptors: Sufficient for Reward, A Nicotine Activation Of," *Science*, 1099420, 306.

Tapper, A. R., et al. (2004), "Nicotine Activation of A4* Receptors: Sufficient for Reward, Tolerance, and Sensitization," *Science*, 306, 1029-1032.

Teslovich, T. M., et al. (2010), "Biological, Clinical and Population Relevance of 95 Loci for Blood Lipids," *Nature*, 466, 707-713.

Thorgeirsson, T. E., et al. (2008), "A Variant Associated with Nicotine Dependence, Lung Cancer and Peripheral Arterial Disease," *Nature*, 452, 638-642.

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

van den Oord, E. J. (2008), "Controlling False Discoveries in Genetic Studies," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 147, 637-644.

VanderWeele, T. J., et al. (2012), "Genetic Variants on 15q25. 1, Smoking, and Lung Cancer: An Assessment of Mediation and Interaction," *American journal of epidemiology*, 175, 1013-1020.

Vincent, M., and Hansen, N. R. (2012), "Sparse Group Lasso and High Dimensional Multinomial Classification," *arXiv preprint arXiv:1205.1245*.

Waaijenborg, S., and Zwinderman, A. (2009), "Sparse Canonical Correlation Analysis for Identifying, Connecting and Completing Gene-Expression Networks."

Waaijenborg, y., Verselewel de Witt Hamer, P. C., and Zwinderman, A. H. (2008), "Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis," *Statistical Applications in Genetics & Molecular Biology*, 7, 1-27.

Wang, J. C., et al. (2009), "Risk for Nicotine Dependence and Lung Cancer Is Conferred by Mrna Expression Levels and Amino Acid Change in Chrna5," *Human Molecular Genetics*, 18, 3125-3135.

Wang, K., Li, M., and Hakonarson, H. (2010), "Analysing Biological Pathways in Genome-Wide Association Studies," *Nature Reviews Genetics*, 11, 843-854.

Wang, P., Chao, D. L., and Hsu, L. (2011), "Learning Oncogenic Pathways from Binary Genomic Instability Data," *Biometrics*, 67, 164-173.

Weedon, M. N., et al. (2007), "A Common Variant of Hmga2 Is Associated with Adult and Childhood Height in the General Population," *Nature genetics*, 39, 1245-1250.

Weiss, R. B., et al. (2008), "A Candidate Gene Approach Identifies the Chrna5-A3-B4 Region as a Risk Factor for Age-Dependent Nicotine Addiction," *PLoS Genet*, 4, e1000125.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009a), "A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis," *Biostatistics*, 10, 515-534.

Witten, D. M., and Tibshirani, R. J. (2009b), "Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data," *Statistical applications in genetics and molecular biology*, 8, 1-27.

Wu, C., et al. (2009a), "Genetic Variants on Chromosome 15q25 Associated with Lung Cancer Risk in Chinese Populations," *Cancer Res*, 69, 5065-5072.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009b), "Genome-Wide Association Analysis by Lasso Penalized Logistic Regression," *Bioinformatics*, 25, 714-721.

Yang, L., et al. (2012), "Functional Polymorphisms of Chrna3 Predict Risks of Chronic Obstructive Pulmonary Disease and Lung Cancer in Chinese," *PLoS One*, 7, e46071.

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49-67.

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19-35.

Yule, G. U. (1907), "On the Theory of Correlation for Any Number of Variables, Treated by a New System of Notation," *Proceedings of the Royal Society of London. Series A*, 79, 182-193.

Zhang, B. B., et al. (2003), "Diagnostic Value of Platelet Derived Growth Factor-Bb, Transforming Growth Factor-Beta1, Matrix Metalloproteinase-1, and Tissue Inhibitor of Matrix Metalloproteinase-1 in Serum and Peripheral Blood Mononuclear Cells for Hepatic Fibrosis," *World J Gastroenterol*, 9, 2490-2496.

Zhang, J., Summah, H., Zhu, Y.-g., and Qu, J.-M. (2011), "Nicotinic Acetylcholine Receptor Variants Associated with Susceptibility to Chronic Obstructive Pulmonary Disease: A Meta-Analysis," *Respiratory research*, 12, 158.

Zhang, W., et al. (2013), "Network-Based Survival Analysis Reveals Subnetwork Signatures for Predicting Outcomes of Ovarian Cancer Treatment," *PLoS computational biology*, 9, e1002975.

Zhou, H., Sehl, M. E., Sinsheimer, J. S., and Lange, K. (2010), "Association Screening of Common and Rare Genetic Variants by Penalized Regression," *Bioinformatics*, 26, 2375.

Zhu, A. Z., et al. (2014), "Association of Chrna5-A3-B4 Snp Rs2036527 with Smoking Cessation Therapy Response in African American Smokers," *Clinical Pharmacology & Therapeutics*.

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301-320.