Stony Brook University

Clustering and Network Analysis with Single Nucleotide Polymorphism (SNP)

A Dissertation Presented

by

Hongyan Chen

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

December 2011

Stony Brook University

The Graduate School

Hongyan Chen

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

Hongshik Ahn – Dissertation Co-Advisor
Professor, Department of Applied Mathematics and Statistics

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

Ellen Li – Outside Member
Professor, Department of Medicine

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Clustering and Network Analysis with Single Nucleotide Polymorphism (SNP)

by

Hongyan Chen

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2011

The goal of the genome-wide association studies (GWAS) is to investigate the relationships between disease phenotypes and genotypes, which are usually determined by a large number of single nucleotide polymorphisms (SNPs). Currently GWAS are often underpowered to identify SNPs with small to moderate effect sizes. In order to overcome this difficulty, two major approaches, (1) meta-analysis by increasing sample size and (2) SNP pre-selection by dimension reduction, are often adopted. Dimension reduction for SNP data has been arduous due to the categorical nature of SNP that renders most association measures such as the Pearson correlation or the Euclidean distance inappropriate. In this thesis, we propose a novel (partial) canonical correlation association measure for categorical data that can be implemented to major dimension reduction approaches including: *cluster analysis* (CA) and *partial correlation network analysis* (PCNA) towards the analysis of GWAS data. Its performance is examined and comparison is made to other existing association measures.

Network analysis methods such as PCNA and the Bayesian network serve as not only dimension reduction approaches but also data driven pathway discovery tools. A key objective in modern genetic studies is to discover the regulatory causal relationships between genetic mutations measured by SNPs and the resulting functional changes often gauged by gene expression levels. With the former being categorical and the latter continuous numerical data, we now face the problem of mixed data types. Our novel partial canonical correlation measure developed for categorical data can be readily extended to PCNA with mixed variables. This new approach is illustrated by using a real data example from a study on inflammatory bowel diseases conducted at Stony brook University Medical Center and the Washington University at St. Louis. Comparison is also made to Bayesian network analysis for mixed data and guidelines provided on the pros and cons of each method.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Single Nucleotide Polymorphism (SNP)

Single nucleotide polymorphism or SNP (pronounced ‘SNiP’), is a major form of genome variation, accounts for approximately 90% of human DNA polymorphisms (Collins et al., 1998). By the widely accepted definition, a SNP represents a single nucleotide genetic variation along the genome sequence that exists in individuals from some population (Brookes, 1999). These different sequence alternatives within a population are also called alleles. The number of alleles corresponds to the same SNP could be two, three or four, and such SNP is categorized as bi-, tri-, and tetra-allelic polymorphism, respectively. However, tri-allelic and tetra-allelic SNPs hardly exist in human genome, hence bi-allelic SNPs are the major target of the Human Genome Project (Sachidanandam et al., 2001). Strictly speaking, SNP should be distinguished from rare variations, with the criterion that the least frequent allele has an abundance of 1% or greater. Nevertheless, this criterion is not always implemented in practice (Brookes, 1999). Some other extensions to the above SNP definition have been also seen in the past decade, for instance, the single nucleotide variation map on cDNA for human chromosome 21 was constructed in 2001 (Deutsch et al., 2001), focusing on the final protein product differences caused by SNPs, although such a cSNPs study ignores potential effects from RNA editing. Finally, it should be noted that single nucleotide polymorphisms do not include insertion/deletion variants or multiple-base alternation, which are other important biological sources of disease. In this dissertation, we will focus our interest on the bi-allelic SNPs along the genome DNA sequence but without the 1% least allele frequency restriction.

To understand the significance of SNPs studies, we first examine the frequency of SNPs across individuals. Several research groups have reported independently that the occurrence of single nucleotide variation in genomic DNA is at the level of 1/1000bp

(Li and Sadler, 1991; Wang et al., 1998; Lai et al., 1998; Nickerson et al., 1998). Moreover, such variation occurs with significantly different frequencies in different genome regions. Usually more polymorphisms are observed in non-coding sequences while fewer ones are found in coding exons - with a reported 100-fold difference between these two regions (Nachman et al., 1998; Guillaudeux et al., 1998; Horton et al., 1998). Considering that there are over 3 billion base pairs for the entire human genome, we could make a reasonable estimate that between any two individuals millions of SNPs can be identified, among which there are hundreds of thousands of amino acid variations. Combining this with the fact that it is 1/10 of the protein variations between human and primates, a great promise thereby lie in the investigation of the relationship between SNPs to phenotypic differences within a population (Brookes, 1999).

From the biological point of view, phenotypic differences should result from a combination of genetic and environmental factors. And many diseases have been found to be heavily influenced by genetic factors, for example the Alzheimer's disease (Gatz et al., 1997) and Autism (Stevenson, 1992). In addition, we can assume influences from genetic factors depend primarily upon the SNP patterns. The SNP genotype – disease phenotype relations can be classified into two broad categories: (1) "simple gene disorders", where genetic variations directly affect genes that cause disease, and (2), "complex disease", where any single related SNP is not sufficient to cause significant phenotypic difference, but instead, only modifies the disease risk. The complex disease scenario, where a collection of SNPs from various genes contribute to the same disease, is more common (Brookes, 1999; Wang et al., 2005). In this scenario, environmental factors would also be essential for disease formation – for example, the effect of smoking on lung cancer (Vial, 1986). A concrete example for complex disease is stature or body height, a quantitative trait. One research group claims that they have identified 20 selected SNPs that will explain the height variation, but only to a degree of 3% (Weedon et al., 2008). Assuming these 20 SNPs are the most influential ones, a set of thousands of SNPs is required to explain the entire

height variation (Goldstein, 2009). On the other hand, treating disease as a categorical variable, several SNPs have been identified for Type 2 diabetes in 2008 by Manolio et al. According to their family-genetic study, the SNP with the strongest effect would have a sibling relative risk only about 1.02. Given that no single SNP may have enough contribution to a certain phenotypic variation and SNPs can be located on numerous loci, traditional genetic-disease analysis methods such as the family linkage analysis may fail to identify any significant SNP candidate. This in turn, calls for association studies or genome-wide association studies.

1.2 Genome Wide Association Study (GWAS)

The idea of association analysis is rather straightforward: if a SNP has an effect on disease occurrence (case-control studies), one allele of this SNP (also called risk allele) should have higher frequency in disease group, compared to non-diseased controls. Therefore, association studies mainly test the allelic frequency differences between control and disease groups, with appropriate control for the confounding effects (gender, age, race, etc.). This method has already shown to be more powerful to detect disease-associated alleles than the traditional family linkage analysis (Greenberg, 1993; Hodge, 1994; Risch and Merikangas, 1996).

When such an analysis is conducted at the whole genome level where usually 100,000 SNPs are genotyped and tested, it becomes the genome-wide association study, or GWAS. It is widely believed that the first GWAS paper was published in 2005, an investigation on age-related macular degeneration (AMD) (Klein et al., 2005) (Figure 1.1). There are also several other papers performing similar studies published in the same year, aiming to identify SNPs associated with myocardial infarction (MI) (Ozaki et al., 2002) or Crohn's Disease (CD) (Yamazaki et al., 2005). The paper from Klein's group is regarded as the landmark study because of two main reasons. First, SNP genotyping was conducted with new microarray technology not the traditional PCR-Invader method, which has much lower success rate of genotyping (~70%).

Second, SNPs were selected "randomly" along the entire human genome instead of gene-based regions. Furthermore, with a rather small sample size (50 controls and 96 cases), they successfully identified common risk alleles with exceptionally large effect size, e.g., odds ratio at 4.6 for heterozygous and odds ratio at 7.4 for homozygous risk alleles. These results brought great excitement to the human genetics community. Subsequently, nearly 400 GWAS research articles were published in 2007 and that year was named the year of GWAS by the journal *Science* (Ku et al., 2010).

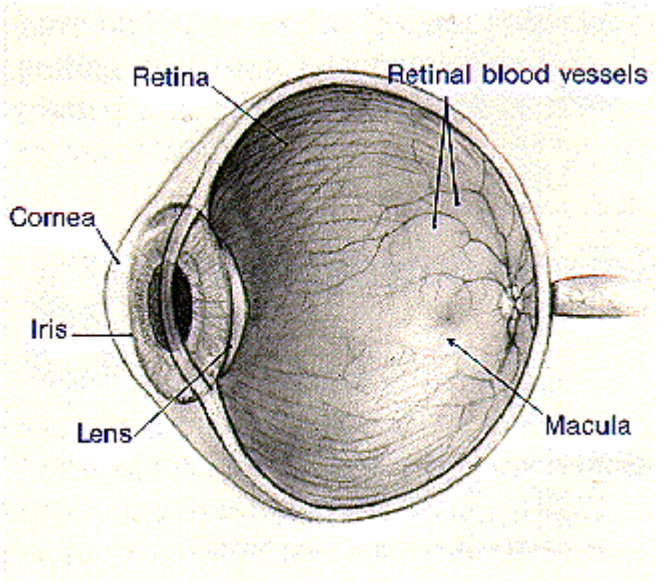


Figure 1.1 Macular retinopathy. Age-related macular degeneration (AMD) is a medical condition which usually affects older adults and results in a loss of vision in the center of the visual field (macula) (http://www.medicinenet.com).

Although GWAS can be customized in many ways, they still share several common steps at the beginning: quality control, bias correction, statistical testing and visualization (Corvin, et al., 2010).

1) Quality control is the first and the most time-consuming step of GWAS. The control has two-fold meaning. First, unqualified SNPs are excluded if they exhibit imprecise mapping to the genome, excessive missing values, low minor allele frequency, etc. Afterwards, subjects are examined by checking agreement between chrX/chrY genotypes and excessive missing values.
2) As GWAS aims to test the allelic frequency difference between cases and controls

for qualitative traits, it is a great concern to ensure such frequency difference does not come from the underlying population structure divergence between cases and controls, also referred to as population stratification. For instance, genetic drift of allele frequencies could result from different ancestry of two groups (e.g., African and European), producing false positive SNPs. In practice, individuals with mixed ancestry should be excluded or a statistical method should be used to control this bias

3) In case-control studies, logistic regression with a single SNP as the predictor is applied or an independence test on a contingency table serves as an alternative. Both are suitable for a study on association between individual SNP and disease, whereas logistic regression is able to incorporate other covariates (Corvin, et al., 2010). SNP is sometimes coded as 0, 1, 2 in terms of the number of one allele, usually the minor or risk allele – however, such practice has severe limitations in analysis, especially regression analysis. In addition, covariates such as gender, race can be included in the statistical modeling process, such as logistic regression, to control for potential confounding effects. Since such analysis is repeated for each SNP, a GWAS typically generates over 100,000 tests. These tests are typically not mutually independent as SNPs could be located physically close to each other, causing linkage disequilibrium. Previous experience suggests that to guarantee a familywise Type I error rate of 0.05 for the simultaneous testing of 1,000,000 SNPs, each individual test has to be conducted at the significance level of 5×10^{-8} (Pe'er et al., 2008). In addition, for controlling FDR at 0.05, individual tests are required at the same significant level as that for familywise error rate of 0.05 (Xing et al., 2010).
4) To present massive statistical test results appropriately, the Manhattan plot is widely used to transform p-values into logarithmic form and then to depict SNPs with significantly small p-values by genomic position (Corvin, et al., 2010).

It is noteworthy that steps described above are without prior knowledge. If other information on SNPs is available or downstream statistical analyses are necessary, we

can prioritize GWAS results in pathway analysis or other bioinformatics approaches (Cantor et al., 2010).

Despite the promising prospect of GWAS, several challenging problems already emerged in recent years. The most important one is the common-disease common-variant (CD/CV) hypothesis, which is theoretically fundamental for GWAS (Manolio, 2010). According to this hypothesis, disease susceptibility is a result of a number of common genetic variants, which present in more than 1% of the population. Hence, affected individuals would share a significant number of risk alleles, not a just single one. In short, as its name suggests, common disease is caused by common variants. A conjugate of CD/CV is the disease heterogeneity hypothesis, where distinct genetic variants are assumed to occur in different affected individuals, and thus each risk allele is a rare event (<1%) within a population. Therefore, the first GWAS paper is a "lucky" exception because the CD/CV model would suggest that each SNP should provide relatively small effect size (odds ratio < 1.5). In practice, after the GWAS on AMD, researchers have not successfully found common risk alleles with considerably large effect size (odds ratio > 2.0): most of the ensuing findings fall into an odds ratio range 1.1 – 1.3 (Cantor et al., 2010). Subsequently, GWAS is questioned as whether it is a powerful approach to detect common variants with small effect size. Figure 1.2 below (Wang et al., 2005) demonstrates this practical issue. Considering an odds ratio of 1.2 with allelic frequency of 0.2, we may need a sample size as unrealistically large as 11,000 at a relatively loose significance level of 10^{-6}. Besides, the CD/CV model would exclude rare SNPs and even they are included in GWAS, the chance of finding them is extremely low. One way to increase the power of GWAS is to directly increase the sample size, most commonly with meta-analysis by combing independent GWAS datasets. Alternatively, we can perform a more thorough pre-selection of the SNPs for testing (Brookes, 1999). In this dissertation, we propose that this pre-selection can be achieved with traditional dimensional reduction techniques, such as hierarchical clustering analysis and partial correlation network analysis (PCNA) – enabled by our newly developed

correlation/partial correlation measures for SNPs as categorical data.

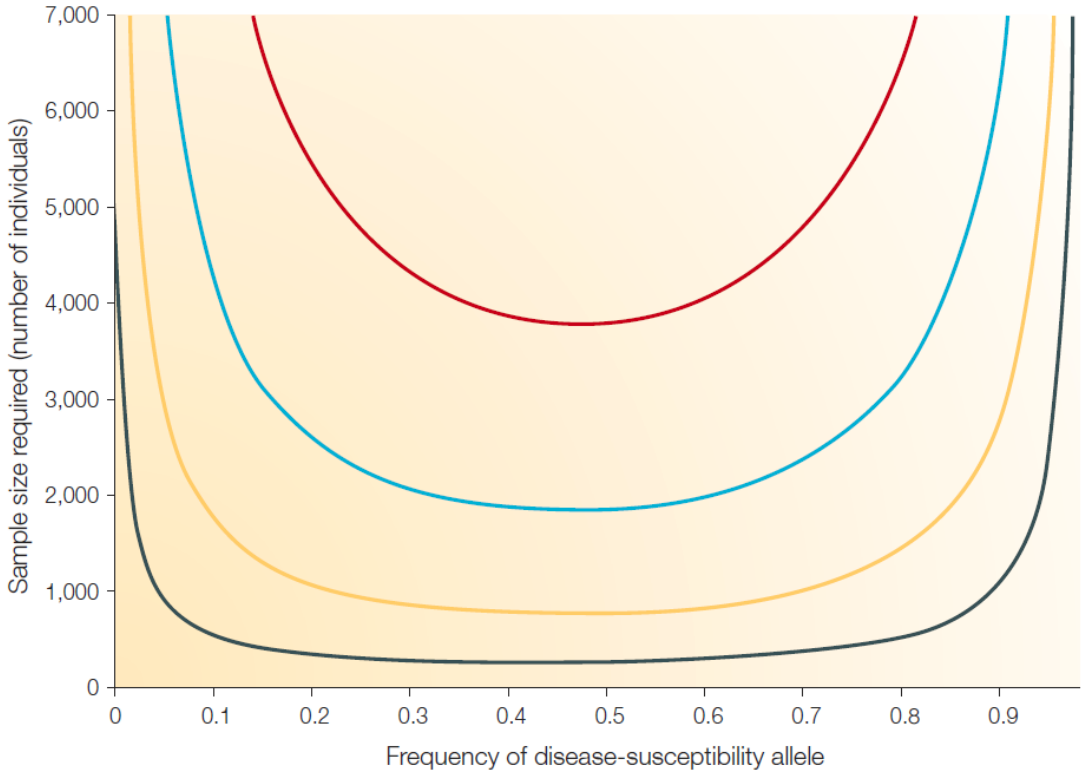


Figure 1.2 Effects of allele frequency and effect size on sample-size requirements. Sample size required for a balanced-design case-control study with allelic odds ratios of 1.2 (red), 1.3 (blue), 1.5 (yellow) and 2 (black) are shown, assuming a multiplicative model with a statistical power of 80% and a Type 1 error rate of 10^{-6} via logistic regression analysis (Wang et al., 2005).

In addition, other concerns on the future of GWAS includes inconsistent findings (Herbert et al., 2006; Maraganore et al., 2005), indirect study designs, etc. Indirect design would interfere with biological interpretation of GWAS results. SNPs identified by GWAS are most likely to be surrogate markers. If SNPs are synonymous (no amino acid alternation), the biological functions of these SNPs remain ambiguous. The association between SNPs and disease phenotype could be due to either SNPs regulate transcription level and/or their physical proximity (i.e. nearly perfect linkage disequilibrium) rather than the true functional variants.

Chapter 2 Clustering Analysis with SNP

2.1 Categorical Properties and Coding Schemes of SNP

Typically, a SNP is treated as a nominal categorical variable. Within a population, SNP can be assigned a minor allele frequency, which is the lowest allele frequency at an observable locus in a particular population. This is simply the lesser of the two allele frequencies for bi-allelic SNP. Therefore, a SNP can be denoted as *AA*, *Aa* and *aa*, where A/a are normal/minor alleles respectively. Under certain circumstances, we can assign a value to the SNP that records the frequency of either allele *A* or *a*: 0, 1 or 2. And such an order obviously has a biological meaning. Therefore, SNPs might also be treated as an ordinal categorical variable with assumption that the heterozygous *Aa* lies between two homozygous *AA* and *aa*. Unfortunately, it is still inappropriate to treat a SNP as a single numeric variable because the 0, 1, 2 coding represents a monotone linear relationship between the three SNP categories and corresponds to severe modeling assumptions in regression analysis. Dummy variable coding should be considered instead.

A categorical variable is binary when it has only two categories. In this case, coding strategy has little effect on the subsequent analysis. For a variable with more than two categories, different coding would yield different interpretations of each individual coding variable. For instance, in multiple regression analysis, the coefficients would represent different comparisons under different coding schemes. Nevertheless, overall model fit is always the same regardless of coding scheme. In more recent times, quite a few widely accepted coding schemes have been developed, including treatment/dummy coding, effects/sum coding, contrast coding and polynomial coding (Cohen and Cohen, 1983; Kaufman and Sweet, 1974; Serlin and Levin, 1985; Wendorf, 2004).

In general, a categorical variable C with g groups needs to be coded by (g – 1) dummy variables: D_1, D_2… $D_{(g-1)}$. In treatment/dummy coding, groups are well defined and a reference group usually exists. If this reference group is assigned as D_i = 0, i = 1, 2 … (g - 1), then the intercept term β_0 in a multiple regression analysis: $Y = \beta_0 + \beta_1 D_1 + \ldots + \beta_{(g-1)} D_{(g-1)}$ represents the reference group mean. If the goal is to compare a single group to the grand mean, a base group should be coded as D_i = -1, i = 1, 2 … (g - 1), and the sum of assigned values of each coding variable should be equal to zero. Sometimes the base group can have a different value other than -1: its advantage is to make coefficient terms in regression more straightforward for interpretation.

Based on effects/sum coding, an additional constraint can be implemented: all the pairwise inner products of coding variables D_1, D_2… $D_{(g-1)}$ must be zero (orthogonal constraint). Under such an orthogonal contrast coding scheme, we are able to catch unique portions of the variance and test specific (i.e., theory-guided) hypotheses (Cohen and Cohen 1983). Finally, if the categorical variable is ordinal, polynomial coding can be derived from the coding strategies mentioned above so as to perform a trend analysis to capture all the (g - 1) – order trends. This coding strategy also overcomes the collinearity problem from natural polynomial coding (Muller and Fetterman, 2002).

2.2 Clustering Categorical Data

Clustering is a popular data mining tool for discovering the underlying structures in a dataset of interest. By classifying observations/variables into different subsets, one can easily perform further studies more efficiently since high associations are expected within the same subset and low associations among members from different subsets. These associations must be quantified by a distance/dissimilarity function, based on which clustering algorithms can be developed. The existing clustering algorithms are mainly divided into two groups: partition clustering and hierarchical

clustering. However, most of them generally deal with continuous data only (Kaufman and Rousseeuw, 1990). The major challenge of applying those algorithms to categorical variables is that datasets with categorical variables have a greatly different data structure: the distance functions that are suitable for continuous data may not be applicable to categorical one. As an example, the Euclidean distance function cannot be used directly for measuring distance with nominal categorical variables since arbitrary coding schemes would assign different values and thus alter the measurement. Hence, in order to overcome those challenges, researchers have developed several modified algorithms specifically for data with categorical attributes in partition clustering.

K-means algorithm proposed by MacQueen (MacQueen, 1967) is a widely used example of the partition algorithm. It starts with a pre-defined cluster number *K*, and subsequentially initializes *K* cluster centers. Afterwards, it performs iterations until a convergence criterion is reached. Obviously it is not reasonable to apply it to categorical variables, because they do not have measurable coordinates for calculation. Instead, K-modes algorithm was introduced (Huang, 1997), in which modes replace means to represent cluster centers of categorical data and mismatch counting is used as distance function:

$$d_{ij} = \sum_{k=1}^{p} \varphi(x_{ik}, x_{jk})$$

$$\varphi(x_{ik}, x_{jk}) = \begin{cases} 0, x_{ik} = x_{jk} \\ 1, x_{ik} \neq x_{jk} \end{cases}$$

where *i*, *j* are individual subjects with *p* categorical attributes. Similar to K-means, it uses iterations to update memberships of every mode and data point. However, it still relies on coding schemes of the categorical variable. More importantly, it might be sensitive to the initial parameters that start the algorithm. Different input orders would generate different outcomes from the same dataset.

Different from the K-means idea based on distance measurement, Cheeseman and

Stutz in 1995 reported a model-based algorithm AutoClass, which is formulated on the basis of mixture models not on the distance function. Instead of using sample space, AutoClass introduces model space that constitutes all possible probability density functions from different numbers of mixture components or modes. AutoClass algorithm determines the most probable set of partitions for a given dataset through a Bayesian model selection procedure. Thus it is a Bayesian unsupervised clustering algorithm. It selects the desired model as the one whose probability density function form has the "best" posterior probability. A fundamental flaw of this algorithm is the computational burden: the computational complexity is $O(nlogn)$. Moreover, EM algorithm is implemented in AutoClass clustering method to determine the global maxima. EM algorithm is believed to have a rather slow convergence rate and sensitive to initial values, which keep AutoClass from being applied to high-dimensional datasets.

In recent years, a new partition algorithm originated from Hamming Distance (HD) vector was introduced (Zhang et al., 2006). Hamming Distance (metric) has been used in clustering categorical data with K-modes. As proposed by their group, the center of the dataset can be identified by minimizing the sum of the HDs over all data points. Afterwards, the categorical sample space is projected into the one-dimensional space as a histogram representing HD frequencies from every data point to that center, which is also defined as the Categorical Distance (CD) vector. Unlike K-means or AutoClass, this algorithm does not require any convergence criteria or parameters for specified models. At each iteration step, it would identify one new cluster from the remaining dataset based on the CD vector pattern until no more significant clusters exist. Therefore, this CD algorithm will determine the number of clusters K automatically. The main drawback of this algorithm is that Hamming Distance is only reasonable for nominal data. When categorical variables are ordered, the CD algorithm will cause a serious loss of ordering information.

On the other hand, dealing with categorical attributes in hierarchical clustering

remains relatively unexplored. It might be due to the fact that usually only pairwise dissimilarities are calculated (for average, complete and single linkage agglomeration methods). Hence, iterative updating of categorical cluster centers would be not necessary. Therefore, the key in hierarchical clustering with categorical data is to define certain appropriate dissimilarity measurements between categorical attributes; several such measurements already exist, either for the general situation or specifically for SNPs studies.

2.3 Pairwise Association Measurement between Categorical Variables

2.3.1 Traditional Measurements

The primary technique to analyze pairwise categorical attributes is with an *r* by *c* contingency table:

		Variable 2				
		Group 1	Group 2	…	Group *c*	Total
Variable 1	Group 1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
	Group 2	n_{21}	n_{21}	...	n_{2c}	$n_{2.}$
	…	...	...	...	...	
	Group *r*	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
	Total	$n_{.1}$	$n_{.2}$	...	$n_{.c}$	$n_{..}$

Table 2.1 A typical *r* by *c* contingency table for two categorical variables.

The most common statistic derived from a contingency table is Pearson's chi-square statistic for independence assessment (Pearson, 1900; Plackett, 1983),

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{[n_{ij}-E(n_{ij})]^2}{E(n_{ij})}, E(n_{ij})=\frac{n_{i.}n_{.j}}{n..}$$

When sample size is large (n_{ij} > 5), under the null hypothesis where independence

holds, this statistic asymptotically follows a chi-square distribution with degrees of freedom equals to $(r - 1)(c - 1)$ (Cochran, 1954). An alternative approach is Fisher's exact test, which performs better especially for the small sample size scenario (Fisher, 1922).

Cramér's V is a popular extension to the chi-square statistic by normalizing it with both sample size and degree of freedom (Cramér, 1946), we obtain

$$V = \sqrt{\frac{\chi 2}{N(k-1)}}$$

where N is the total number of observations and k is the lesser of r and c. Two major advantages may explain why Cramér's V is a popular measurement. First, the value is normalized in the range between 0 and 1, and such value is readily interpreted: V may be viewed as the ratio of the actual association between two variables to their maximum possible one. Second, it is well adapted to studies involving a large contingency table, e.g., in multiallelic loci research (Brynedal et al, 2007).

When categorical variables have rank order, the standard chi-square test would not be powerful enough as it does not take into account category ordering information. Agresti has introduced several other strategies as alternatives using Mann-Whitney Test and weighted sum of differences (Agresti, 1983). Moreover, other powerful methods to detect association between ordered categorical variables have been reported, such like Kendall's τ, Goodman and Kruskal's γ, etc. (Kendall, 1938; Goodman and Kruskal, 1954). Both methods rely on calculation of concordant and discordant pairs, which are only applicable for ordered categorical variables. Consider a pair of bivariate observation data-set $\{X_1, Y_1\}$ and $\{X_2, Y_2\}$. If $\text{sgn}(X_2 - X_1) = \text{sgn}(Y_2 - Y_1)$, the pair is termed concordant; on the other hand, if $\text{sgn}(X_2 - X_1) = -\text{sgn}(Y_2 - Y_1)$ the pair is discordant. Letting C and D be the total number of concordant and discordant pairs respectively, we have

$$\tau = \frac{C - D}{\sqrt{(n^2 - \sum_{i=1}^{r} n_{i.}{}^2)(n^2 - \sum_{j=1}^{c} n_{.j}{}^2)}}$$

$$\gamma = \frac{C - D}{C + D}$$

Similar to properties of Pearson's correlation coefficient *r*, both measurements lie between -1 and 1, where 0 suggests independence holds between two variables. When the sample size is sufficiently large, corresponding test statistics can be also derived for hypothesis testing.

Another important aspect of measurement between categorical variables is inter-rater agreement assessment. Among techniques in this field, Cohen's kappa receives continuing popularity (Cohen, 1960):

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where Pr(*a*) is the relative observed agreement among raters, and Pr(*e*) is the hypothetical probability of agreement by chance or the probability of each observer randomly saying each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$. Based on this method, several derivatives have been developed for the purposes of taking disagreement differently (Cohen, 1968) or relating categorical data to continuous data (King and Chinchilli, 2001). A similar statistic, called Scott's *pi*, was used more specifically in communications studies (Scott, 1955). The difference between those two methods is how Pr(*e*) is calculated. In addition, Fleiss' κ, an extension of Scott's *pi*, generalizes the classic two-rater agreement problem to multi-rater case (Fleiss, 1971).

If a contingency table is 2 by 2 or can be collapsed into a 2 by 2 one with matched pairs of subjects, McNemar's test will be useful to test marginal homogeneity (McNemar, 1947). It is worthy to point out that an application of this test, termed transmission disequilibrium test (TDT), has been reported as a

family-based association test and it is robust to the presence of population structure (Spielman et al., 1993).

2.3.2 Linkage Disequilibrium (LD)

The concept of linkage disequilibrium (LD) was introduced to population genetics much earlier than the onset of SNPs research (Lewontin and Kojima, 1960). It is mainly used to describe the non-random association of alleles at two or more loci. The existence of a significant LD among loci suggests a strong association, which could be a result of many possible factors, including physical proximate location, the recombination rate, selection, etc. Consider an example with two loci *A/a* and *B/b* and assume we already know the four possible haplotype frequencies:

Haplotype	Frequency
AB	h_{11}
Ab	h_{12}
aB	h_{21}
ab	h_{22}

We are then able to obtain corresponding individual allele frequencies:

Allele	Frequency
A	$p_1 = h_{11}+h_{12}$
a	$p_2 = h_{21}+h_{22}$
B	$q_1 = h_{11}+h_{21}$
b	$q_2 = h_{12}+h_{22}$

If there is no linkage between two loci haplotypes are completely random combinations of *A/a* and *B/b*, and the following four equations hold:

Haplotype	Expected frequency
AB	$h_{11} = p_1q_1$
Ab	$h_{12} = p_1q_2$
aB	$h_{21} = p_2q_1$

ab	$h_{22} = p_2q_2$

However, if two loci are not independent to each other, the allelic variation in one locus would influence the other one, and subsequently we would expect to observe deviation from the above four hypothesized equations. The degree of deviation is commonly denoted by *D*:

	A	*a*	Total
B	$h_{11} = p_1q_1 + D$	$h_{21} = p_2q_1 - D$	q_1
b	$h_{12} = p_1q_2 - D$	$h_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	1

When $D \neq 0$, we conclude linkage disequilibrium is observed between these two loci. A larger value of *D* indicates stronger association between *A/a* and *B/b*. Nevertheless, the direct use of *D* is not desirable: the sign of *D* is arbitrary since either allele can be assigned as the capital allele *A/B*; besides, the possible value range of *D* is bounded by calculated allelic frequencies, which may vary from one study to another. Instead, two other metrics have been devised: *D'* and r^2 (or *r*).

$$D' = \begin{cases} \dfrac{D}{D_{\max}} = \dfrac{D}{\min(p_1q_2, p_2q_1)}, D \geq 0 \\ \dfrac{D}{D_{\min}} = \dfrac{D}{\min(p_1q_1, p_2q_2)}, D < 0 \end{cases}$$

$$r^2 = \frac{D^2}{p_1p_2q_1q_2}, r = \sqrt{r^2}$$

Both can be viewed as standardized *D*, whose values are comparable across LD measurements of different loci. Particularly, r^2 is preferable in population genetic studies since its expected value is a function of required sample size for association mapping, given a fixed genetic effect (Sham et al., 2000; Pritchard and Przeworski, 2001).

LD measurements have recently become extremely useful in GWAS mainly because of two reasons. By measuring pairwise LDs of a set of SNPs, “tag-SNPs” could be selected so as to work as surrogates for analyses. Such tag-SNPs usually

have strong LDs with all the other SNPs in a given set. In other words, the use of tag-SNPs greatly reduces the number of SNPs that are required for either commercial array genotyping (cost reduction) or statistical tests (dimension reduction). Secondly, as mentioned in the first chapter, GWAS is an indirect approach to find out pathogenic genes. In practice, a relatively small set of SNPs will be identified that shows strong associations with disease phenotype directly or implicitly, with the underlying genes. In addition, the associations between SNP candidates and underlying genes are always modeled with LD measurements (Wang et al., 2005).

Due to the fact that LD is based on haplotype but not genotype, which is the data form in GWAS, heterozygous ambiguity exists (Figure 2.1). The non-one-to-one corresponding pattern between genotype and haplotype leads us to estimate haplotype frequencies. One widely accepted approach is to solve the maximum likelihood function with EM algorithm (Becker and Knapp, 2004).

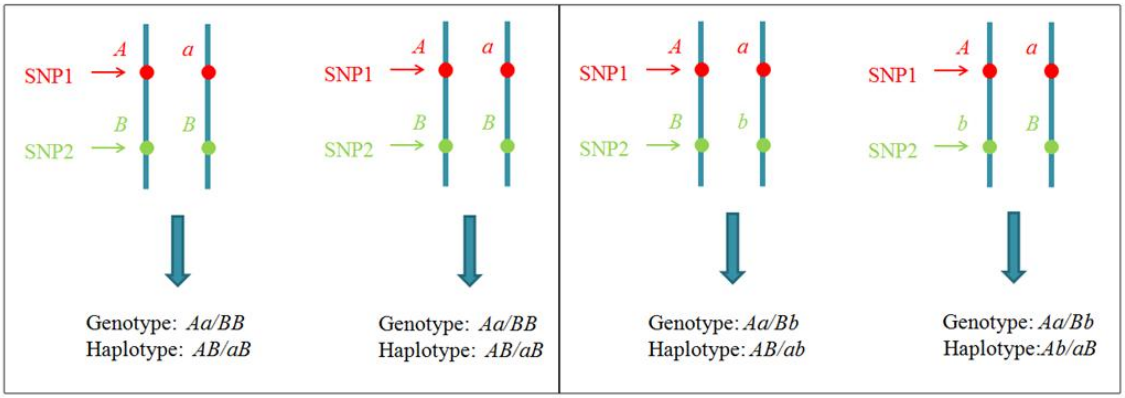


Figure 2.1 Heterozygous ambiguity from genotype to haplotype. The left panel shows a clear transform from genotype to haplotype; the right panel depicts why one-to-one corresponding fails when heterozygote exists in both loci, where one genotype is related to two possible haplotype combinations.

Suppose the following contingency table has been constructed, where $n_1 \ldots n_9$ are numbers of observations and the four haplotype frequencies required for LD maresure are $P_{AB} = P_{11}$, $P_{Ab} = P_{10}$, $P_{aB} = P_{01}$, $P_{ab} = P_{00}$.

	BB	*Bb*	*bb*
AA	*AB|AB* (n_1)	*AB|Ab* (n_2)	*Ab|Ab* (n_3)

Aa	*AB*|*aB* (n_4)	*AB*|*ab* or *Ab*|*aB* (n_5)	*Ab*|*ab* (n_6)
aa	*aB*|*aB* (n_7)	*aB*|*ab* (n_8)	*ab*|*ab* (n_9)

Table 2.2 A sample table of pairwise bi-allelic SNPs. Note that n_5 is a mixture of two possible haplotype combinations.

Assuming $n_1 \ldots n_9$ follow a multinomial distribution, we can write down the likelihood function in terms of P_{11}, P_{10}, P_{01} and P_{00}, where $N = n_1 + n_2 + \ldots + n_9$:

$$L = \frac{N!}{n_1!n_2!\ldots n_9!}(P_{11}^{\ 2})^{n1}(2P_{11}P_{10})^{n2}\ldots(P_{00}^{\ 2})^{n9}$$

$$L = \frac{N!}{n_1!n_2!\ldots n_9!}2^{n2+n4+n6+n8}P_{11}^{\ 2n1+n2+n4}P_{10}^{\ 2n3+n2+n6}P_{01}^{\ 2n7+n8+n4}P_{00}^{\ 2n9+n8+n6}(2P_{11}P_{00}+2P_{10}P_{01})^{n5}$$

$(2P_{11}P_{00} + 2P_{10}P_{01})$ indicates a mixture of two possible haplotype combinations in n_5. By introducing a new parameter θ, we let the proportion of $P_{11}P_{00}$ (*AB*|*ab*) be θ and the proportion of $P_{10}P_{01}$ (*Ab*|*aB*) be $(1 - \theta)$ and reorganize the likelihood function

$$L = \frac{N!}{n_1!\ldots(\theta n_5)!((1-\theta)n_5)!\ldots n_9!}2^{n2+n4+n5+n6+n8}P_{11}^{\ 2n1+n2+n4+\theta n5}P_{10}^{\ 2n3+n2+n6+(1-\theta)n5}P_{01}^{\ 2n7+n8+n4+(1-\theta)n5}P_{00}^{\ 2n9+n8+n6+\theta n5}$$

Afterwards, EM algorithm can be implemented. In the expectation step, given certain initial guess of P_{11}, P_{10}, P_{01} and P_{00},

$$\theta n_5 \sim \text{binomial}\left(n_5, \frac{P_{11}P_{00}}{P_{11}P_{00}+P_{10}P_{01}}\right)$$

$$E[\theta] = \frac{P_{11}P_{00}}{P_{11}P_{00}+P_{10}P_{01}}$$

Then P_{11}, P_{10}, P_{01} and P_{00} are updated with this expected value of θ,

$$P_{11}^{new} = \frac{2n_1 + n_2 + n_4 + n_5\dfrac{P_{11}P_{00}}{P_{11}P_{00}+P_{10}P_{01}}}{N}$$

$$P_{10}^{new} = \frac{n_2 + 2n_3 + n_6 + n_5\dfrac{P_{10}P_{01}}{P_{11}P_{00}+P_{10}P_{01}}}{N}$$

$$P_{01}^{new} = \frac{n_4 + 2n_7 + n_8 + n_5\dfrac{P_{10}P_{01}}{P_{11}P_{00}+P_{10}P_{01}}}{N}$$

$$P_{00}^{new} = \frac{n_6 + n_8 + 2n_9 + n_5\dfrac{P_{11}P_{00}}{P_{11}P_{00}+P_{10}P_{01}}}{N}$$

Iterations are terminated when convergence occurs. In practice, we set the convergence criterion as the l^2-norm of the difference vector between two rounds of estimated P_{11}, P_{10}, P_{01} and P_{00} less than $1e^{-10}$.

2.3.3 Canonical Correlation Measurement

In this section, we propose a novel pairwise association measurement for categorical variables. Recall SNP is a three-category variable – *AA*, *Aa* and *aa*. With the simple dummy coding,

$$AA: X_1 = 0, X_2 = 0$$
$$Aa: X_1 = 0, X_2 = 1$$
$$aa: X_1 = 1, X_2 = 0$$

or with polynomial coding,

$$AA: S = 0, S^2 = 0$$
$$Aa: S = 1, S^2 = 1$$
$$aa: S = 2, S^2 = 4$$

The corresponding coding matrices are

$$A_1 = \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 4 \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Note that they are linearly transformable to each other

$$A_2 \begin{bmatrix} -0.5 & 2 \\ 0.5 & -1 \end{bmatrix} = A_1$$

Now define association between two SNPs X and Y is computed as the (first) canonical correlation between ($\{S_X, S_X^2\}$, $\{S_Y, S_Y^2\}$). Such measurement will remain constant even if simple dummy coding is used because the two-dimensional dummy variable spaces created under those two coding schemes are linear. Canonical correlation seeks to find out the maximized correlation from the one-dimensional space projected from dummy variables; therefore, its value will be invariant to coding schemes listed above.

2.3.4 Link Canonical Correlation Analysis to Chi-square Test

Using canonical correlation in categorical variable studies is not a completely new idea. As early as 1973, sociologists gave the first report on how chi-square tests on contingency tables can be treated as a special form of canonical correlation analysis (Darlington et al., 1973). Years later, *t*-test for either correlated or independent samples, general linear model (ANOVAs), discriminate analysis were also successfully incorporated into the general canonical correlation testing system (Knapp, 1978).

As demonstrated in most multivariate textbooks (Anderson, 1958; Morrison, 1976), a typical canonical correlation analysis involves two sets of variables: $X = [X_1 ... X_p]$ and $Y = [Y_1 \dots Y_q]$. Without loss of generality, let q be no greater than p. Then canonical correlations can be obtained as the square root of eigenvalues - λ_i's of the matrix $M = R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY}$, where R represents sample correlation matrices (standardized variance-covariance matrices). Its significance can be verified with an F-statistic as follows:

$$F(pq, (ms - pq/2 + 1)) = \frac{(1 - \Lambda^{1/s}) / pq}{(\Lambda^{1/s}) / (ms - pq/2 + 1)}$$

$$\Lambda = \prod_{i=1}^{q} (1 - \lambda_i)$$

$m = N - 3/2 - (p + q)/2$, N is the sample size

$$s = \sqrt{\frac{p^2 q^2 - 4}{p^2 + q^2 - 5}}$$

Now suppose these two sets of variables X and Y are coded dummy variables, corresponding to two categorical variables A and B, with $(p + 1)$ and $(q + 1)$ groups, respectively. The ordinary chi-square independence test would use the $(p + 1) \times (q + 1)$ contingency table and the obtained statistic follows chi-square distribution with degree of freedom of pq. Since $\chi^2(pq) = pq \times F(pq, \infty)$, the two tests are exchangeable when sample size is sufficiently large (Knapp, 1978).

Moreover, the quantity of canonical correlation is closely related to Cramér's *V* because the squared Cramér's *V* is the mean squared canonical correlations between coded dummy variables (Cramér's, 1946). And Cramér's *V* can be expressed in terms of a modified LD *r* with Hardy-Weinberg principle and certain assumption on heterozygous ambiguity. Table 2.3 provides a simple example to illustrate how these three measures can be linked to each other. Cramér's *V* can be obtained from the Chi-square statistic:

$$V = \sqrt{\frac{\chi 2}{2N}} = 0.154 \Rightarrow V^2 = \frac{\chi 2}{2N} = 0.024$$

We can also have two canonical correlations from dummy variable coding:

$$r_1 = 0.217, r_2 = 0.005 \Rightarrow \frac{r_1^2 + r_2^2}{2} = 0.024$$

Additionally, we estimate the LD between two SNPs: *phi* = 0.65, r^2 = 0.023, *r* = 0.15

Now assuming Hardy-Weinberg principle holds in our data and *phi* = 0.5

$$\hat{P}_A = \sqrt{P_{AA}}, \hat{P}_B = \sqrt{P_{BB}}, \phi = 0.5$$
$$\Rightarrow D = P_{AB} - \hat{P}_A \hat{P}_B$$
$$\Rightarrow (r')^2 = \frac{4D^2}{[P_A(1-P_A) + \Delta_A][P_B(1-P_B) + \Delta_B)]} = 0.024; \Delta_A = P_{AA} - (P_A)^2, \Delta_B = P_{BB} - (P_B)^2$$
$$\Rightarrow r' = 0.154$$

Therefore, Cramér's *V* takes equal weights of canonical correlations: $\sqrt{(r_1^2 + r_2^2)/2}$, while our measure focuses on the first one r_1. A large difference between r_1 and r_2 will indicate a significant difference between the two measures. Another possible measure would be giving different (optimal) weights of r_1 and r_2. Note that Cramér's *V* is also close to the LD measure *r* in our example, which is expected since Cramér's *V* already serves as an alternative global LD measure (more than two SNPs) (Brynedal et al, 2007).

	BB	*Bb*	*bb*

AA	3	5	2
Aa	3	5	3
aa	2	3	4

Table 2.3 An example to show the squared Cramér's *V* is the mean squared canonical correlations.

Hence previous studies on canonical correlation analysis with categorical variables provide us a theoretical support of using it to measure association strength between SNPs: the first canonical correlation is positively related to the chi-square statistic, whose value implies association strength. Additionally, in Knapp's paper, simple dummy variable coding was suggested. We have already shown that both simple dummy variable coding and the linear/quadratic minor allele frequency coding will generate the same canonical correlation measures. On the other hand, these published papers focused on the testing point of view, which is significance testing of canonical correlations and its equivalence to the independence test on contingency table. In particular, they did not investigate the first canonical correlation quantity and its potential applications, such as the dissimilarity matrix input in hierarchical clustering analysis that we have performed with real datasets.

Chapter 3 Clustering Application to GWAS Data

In this chapter, we examined the performance of the canonical correlation measure when it serves as the input dissimilarities for hierarchical clustering with SNPs. Clustering outputs were compared to outputs using with other measures, including linkage disequilibrium (r^2 and r), Cramér's V, Kendall's τ and Pearson' r. These four other measures were selected for distinct reasons: LD measure is the conventional one in GWAS; Cramér's V is derived from chi-square test, which is explicitly related to canonical correlation measure; Kendall's τ is most suitable for ordinal variables, which might be true when SNP is codes as 0, 1 and 2, the number of risk alleles; Pearson's r is to check whether it is applicable to treat SNP as a single numeric variable (0, 1 and 2 coding).

To make a feasible comparison, we need to control all the other factors in clustering, most importantly, the linkage method and the cluster number determination. In the following applications to real datasets, we used the traditional average linkage method for all clustering analyses. However, we did not calculate common statistics for cluster number determination, such as R^2, pseudo-F and pseudo-t^2 since they all rely on Euclidean distance and cluster centers (mean). Instead, we implemented a package in *R*: *DynamicTreeCut* (Langfelder et al., 2008). With this method, clusters are defined by cutting branches off a dendrogram, but not with some arbitrary height value. The cutting procedure would be specific/dynamic to different clusters, controlled by certain parameters, e.g., minimum number of variables in a cluster, maximum joining tree height, etc. Therefore, these parameters can be set to be the same levels in different SNP association measures comparison. This package is also able to detect outliers in a dendrogram, which can be viewed as SNPs that are distant from all defined clusters.

Since clustering analysis is rather an exploratory approach, chromosome

locations could serve as a general standard (prior knowledge) to evaluate clustering results. SNPs with closer physical distance are more likely to be genetically linked. Thus a reasonable clustering would separate SNPs completely or partially by their chromosome locations: good performance will generate clusters whose SNP members come from the same chromosome. In other words, it will group SNPs from the same chromosome into one cluster. Furthermore, such evaluation can be more thoroughly conducted if SNP positions along the same chromosome are also available.

3.1 Collaborative Genetic Study of Nicotine Dependence (COGEND)

We first applied the canonical correlation concept and other association measurements to the COGEND data from our collaborator, Laura Bierut's group. Strictly speaking, it is a subset from a GWAS since it contains 2022 subjects (1114 cases and 908 controls) and only 215 SNPs located in eight different chromosomes, but the number of SNPs should be sufficient for clustering. As explained earlier, SNPs are coded with the following rule: heterozygous *Aa* is always set to be 1; homozygous reference genotype *AA* is identified with higher frequency in samples and then set to be 0; the minor allele homozygous is *aa* and codes as 2. For instance, if a SNP has three genotypes: CG/CC (24%), CC/CC (35%) and CG/CG (41%), the corresponding coding is 1 (*Aa*), 2 (*aa*) and 0 (*AA*), respectively. Such coding can be directly used for Pearson's r and Kendall's τ. For canonical correlation measurement, polynomial dummy variables are required: CG/CC {1, 1}, CC/CC {2, 4} and CG/CG {0, 0}. Additionally, we ensured all three possible genotypes exist in samples for every SNP. However, we did not exclude those 15 SNPs with rare minor allele frequencies < 5% (Figure 3. 1) since it has been found that less common SNPs could be also associated with nicotine dependence (Saccone et al., 2009).

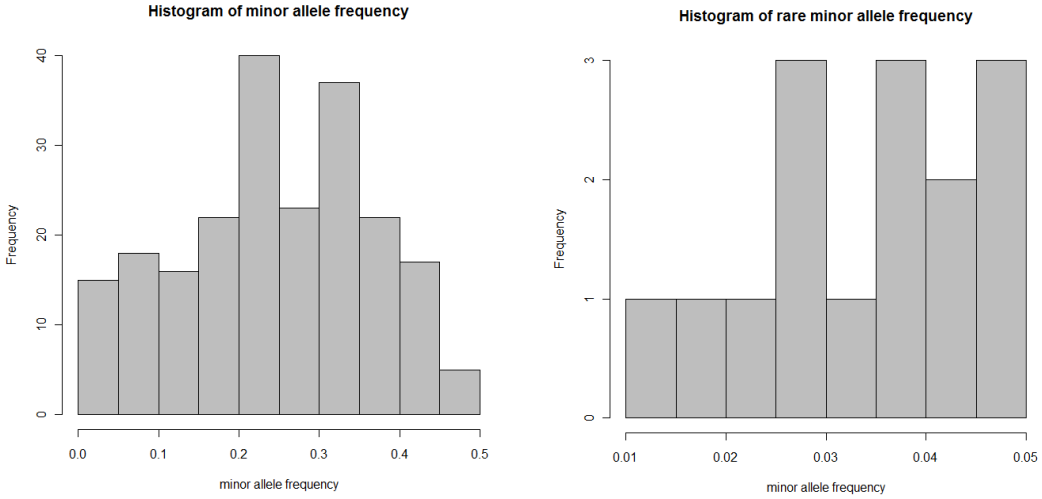


Figure 3.1 Minor allele frequency distribution of COGEND data. Left panel is the overall distribution of all 215 SNPs while right panel displays more details of rare minor alleles (frequency < 5%).

3.1.1 Clustering Evaluation

Clustering outputs are summarized in tables below (Table 3.1 – Table 3.6). Note that "Cluster0" represents outliers detected by *DynamicTreeCut* package (minimum cluster size = 10 SNPs).

Canonical correlation

Cluster	Chromosome							
	1	2	4	8	11	15	17	20
Cluster0	**0**	**0**	**0**	**1**	**1**	**2**	**0**	**0**
Cluster1	**4**	**5**	**0**	**26**	**1**	**0**	**6**	**0**
Cluster2	0	0	0	0	0	31	0	0
Cluster3	0	0	0	0	0	24	0	0
Cluster4	0	0	0	0	0	24	0	0
Cluster5	0	0	0	0	2	15	0	0
Cluster6	0	0	0	0	0	16	0	0
Cluster7	0	0	0	0	0	0	14	0

Cluster8	0	12	0	0	0	0	0	0
Cluster9	0	0	0	0	0	0	0	11
Cluster10	0	0	10	0	0	0	0	0
Cluster11	0	0	0	10	0	0	0	0

Table 3.1 Clustering pattern of 215 SNPs based on canonical correlation. Highlighted (red) rows are clusters not achieving separation goal. Cluster0 is a collection of outliers.

Linkage disequilibrium (r^2)

Cluster	Chromosome							
	1	2	4	8	11	15	17	20
Cluster0	**4**	**12**	**10**	**12**	**4**	**10**	**6**	**1**
Cluster1	**0**	**5**	**0**	**25**	**0**	**3**	**0**	**0**
Cluster2	0	0	0	0	0	32	0	0
Cluster3	0	0	0	0	0	29	0	0
Cluster4	0	0	0	0	0	24	0	0
Cluster5	0	0	0	0	2	14	0	0
Cluster6	0	0	0	0	0	0	14	0
Cluster7	0	0	0	0	0	0	0	10

Table 3.2 Clustering pattern of 215 SNPs based on linkage disequilibrium r^2. Highlighted (red) rows are clusters not achieving separation goal. Cluster0 is a collection of outliers.

Linkage disequilibrium (r)

Cluster	Chromosome							
	1	2	4	8	11	15	17	20
Cluster0	**0**	**0**	**0**	**1**	**4**	**2**	**0**	**0**
Cluster1	**4**	**5**	**0**	**26**	**0**	**0**	**6**	**0**
Cluster2	0	0	0	0	0	32	0	0
Cluster3	0	0	0	0	0	24	0	0

Cluster4	0	0	0	0	0	24	0	0
Cluster5	0	0	0	0	0	15	0	0
Cluster6	0	0	0	0	0	15	0	0
Cluster7	0	0	0	0	0	0	14	0
Cluster8	0	12	0	0	0	0	0	0
Cluster9	0	0	0	0	0	0	0	11
Cluster10	0	0	10	0	0	0	0	0
Cluster11	0	0	0	10	0	0	0	0

Table 3.3 Clustering pattern of 215 SNPs based on linkage disequilibrium *r*. Highlighted (red) rows are clusters not achieving separation goal. Cluster0 is a collection of outliers.

Cramér's *V*

Cluster	Chromosome							
	1	2	4	8	11	15	17	20
Cluster0	**4**	**0**	**0**	**2**	**4**	**2**	**1**	**1**
Cluster1	**0**	**5**	**0**	**25**	**0**	**0**	**4**	**0**
Cluster2	0	0	0	0	0	31	0	0
Cluster3	0	0	0	0	0	24	1	0
Cluster4	0	0	0	0	0	24	0	0
Cluster5	0	0	0	0	0	16	0	0
Cluster6	0	0	0	0	0	15	0	0
Cluster7	0	0	0	0	0	0	14	0
Cluster8	0	12	0	0	0	0	0	0
Cluster9	0	0	10	0	0	0	0	0
Cluster10	0	0	0	10	0	0	0	0
Cluster11	0	0	0	0	0	0	0	10

Table 3.4 Clustering pattern of 215 SNPs based on Cramér's *V*. Highlighted (red) rows are clusters not achieving separation goal. Cluster0 is a collection of outliers.

Kendall's τ

Cluster	Chromosome							
	1	2	4	8	11	15	17	20
Cluster0	0	0	0	0	0	0	0	0
Cluster1	**4**	**6**	**1**	**28**	**2**	**1**	**7**	**7**
Cluster2	**0**	**2**	**0**	**1**	**2**	**33**	**5**	**0**
Cluster3	**0**	**7**	**3**	**4**	**0**	**24**	**1**	**3**
Cluster4	**0**	**1**	**3**	**4**	**0**	**30**	**0**	**1**
Cluster5	**0**	**1**	**3**	**0**	**0**	**18**	**0**	**0**
Cluster6	**0**	**0**	**0**	**0**	**0**	**6**	**7**	**0**

Table 3.5 Clustering pattern of 215 SNPs based on Kendall's τ. Highlighted (red) rows are clusters not achieving separation goal. Cluster0 is a collection of outliers.

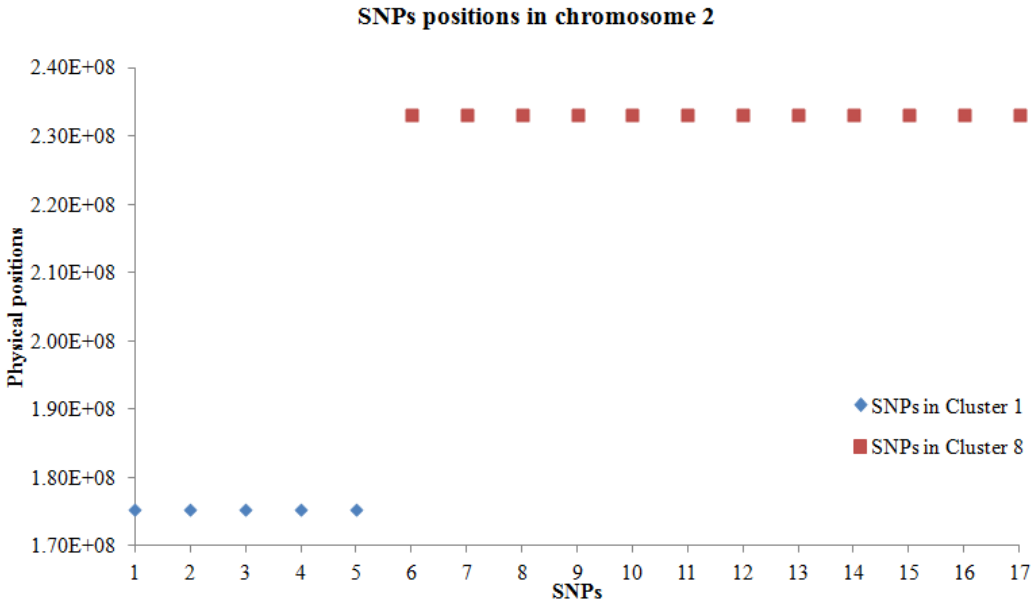
Pearson' r

Cluster	Chromosome							
	1	2	4	8	11	15	17	20
Cluster0	0	0	0	0	0	0	0	0
Cluster1	**4**	**6**	**1**	**28**	**0**	**1**	**7**	**7**
Cluster2	**0**	**7**	**3**	**5**	**0**	**25**	**1**	**3**
Cluster3	**0**	**2**	**0**	**0**	**4**	**32**	**4**	**0**
Cluster4	**0**	**1**	**3**	**4**	**0**	**30**	**0**	**1**
Cluster5	**0**	**1**	**3**	**0**	**0**	**18**	**1**	**0**
Cluster6	**0**	**0**	**0**	**0**	**0**	**6**	**7**	**0**

Table 3.6 Clustering pattern of 215 SNPs based on Pearson' r. Highlighted (red) rows are clusters not achieving separation goal. Cluster0 is a collection of outliers.

As shown in tables above, each row represents chromosomal separation information within each cluster. And ideally each cluster should contain SNPs from only one chromosome. Highlighted red rows indicate clusters not achieving separation goal. Therefore, according to the chromosome separation criterion, we

conclude that canonical correlation, LD (r) and Cramér's V produce extremely similar clustering patterns and they perform much better than Kendall's τ and Pearson' r. Regarding comparison between r and r^2 for LD measures, r is a better choice mainly because there are a great number of SNPs classified as outliers with r^2 measure. Furthermore, if we examine these tables by columns (chromosomes) instead of by rows (clusters), we will find that SNPs from the same chromosome could be grouped into two or more clusters, in a quite consistent manner (canonical correlation, LD r and Cramér's V). For example, 17 SNPs in chromosome 2 were grouped into two clusters (the same 5 and 12 SNPs, respectively) across all the three measures; 20 SNPs from chromosome 17 were grouped into two clusters (the same 6 and 14 SNPs, respectively) with canonical correlation and LD r. Such separation might suggest a poor clustering but it is reasonable based on detailed chromosomal positions (Figure 3.2).



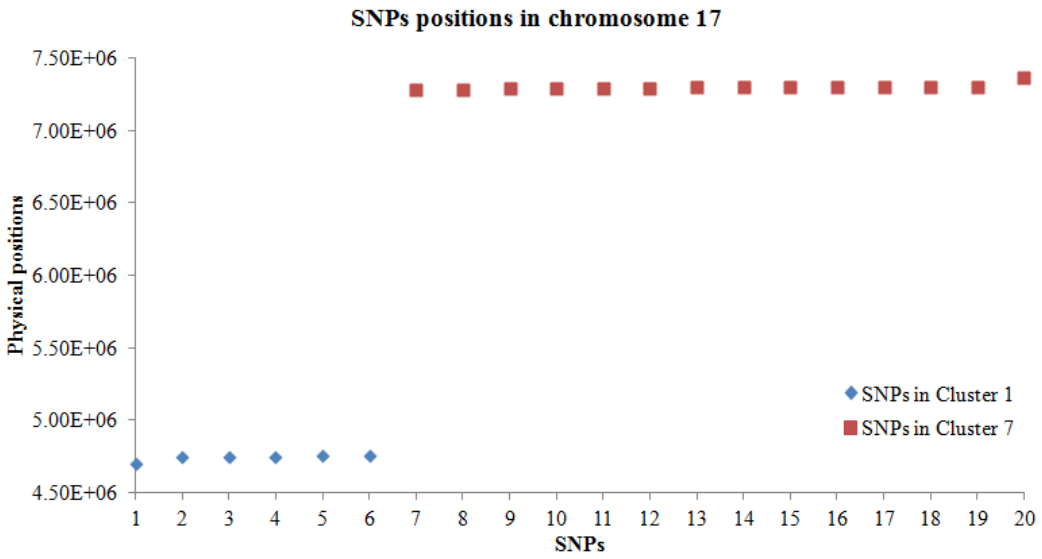


Figure 3.2 SNPs positions in chromosome 2 and 17. Clustering information is based on canonical correlation measure and LD (*r*).

A more extreme case is in chromosome 15, where there are more than 100 SNPs mainly located in two regions, one around 3e7 (region I) while the other around 8e7 (region II) (Figure 3.3). Thus it is worthy to conduct clustering analysis specifically with SNPs from chromosome 15. Results are shown in Table 3.7 – Table 3.9.

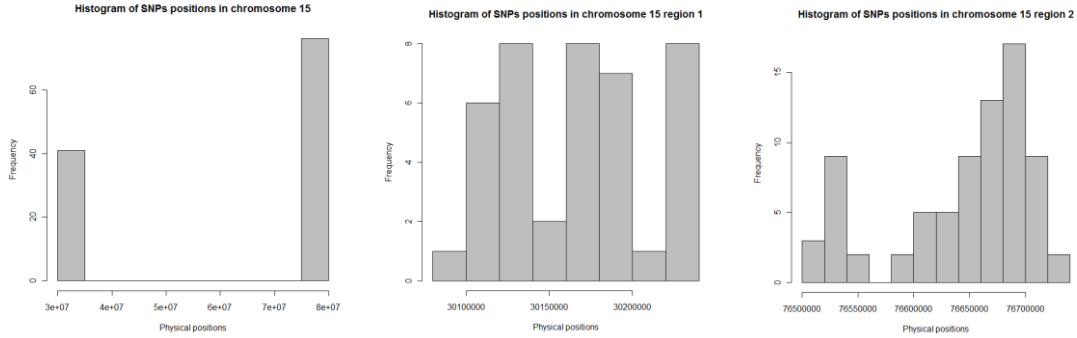


Figure 3.3 SNPs positions in chromosome 15. The most left panel shows the overall information, clearly SNPs are located in two distant regions along chromosome 15; the right two panels show more detailed histograms of each region.

Canonical correlation

	Number of SNPs	Positional range	
		min	max

Cluster0	2	30228925	76729090
Cluster1	39	30088689	30232287
Cluster2	31	76518863	76721606
Cluster3	24	76619452	76713073
Cluster4	16	76517368	76711042

Table 3.7 Clustering pattern of 112 SNPs from chromosome 15 based on canonical correlation. Cluster0 is a collection of outliers.

Linkage disequilibrium (*r*)

	Number of SNPs	Positional range	
		min	max
Cluster0	2	30228925	76729090
Cluster1	24	30088689	30232287
Cluster2	15	30172411	30228226
Cluster3	32	76518863	76721606
Cluster4	24	76619452	76713073
Cluster5	15	76517368	76711042

Table 3.8 Clustering pattern of 112 SNPs from chromosome 15 based on Linkage disequilibrium *r*. Cluster0 is a collection of outliers.

Cramér's *V*

	Number of SNPs	Positional range	
		min	max
Cluster0	1	30228925	30228925
Cluster1	24	30088689	30232287
Cluster2	15	30172411	30228226
Cluster3	32	76518863	76729090
Cluster4	24	76619452	76713073
Cluster5	16	76517368	76711042

Table 3.9 Clustering pattern of 112 SNPs from chromosome 15 based on Cramér's *V*. Cluster0 is a collection of outliers.

Similar to the previous tables, clustering patterns across all three measures are consistent. SNPs from region II were grouped into three clusters and one or two SNPs were detected as outliers in Cluster0. Nevertheless, only the canonical correlation measure successfully identified all the SNPs from region I as a single cluster. However, two clusters from region I generated by LD and Cramér's *V* overlapped, which is not desirable according to the physical location criterion (Figure 3.4). Hence, in this case, it appears that the canonical correlation measure achieved a slightly better performance than the other two.

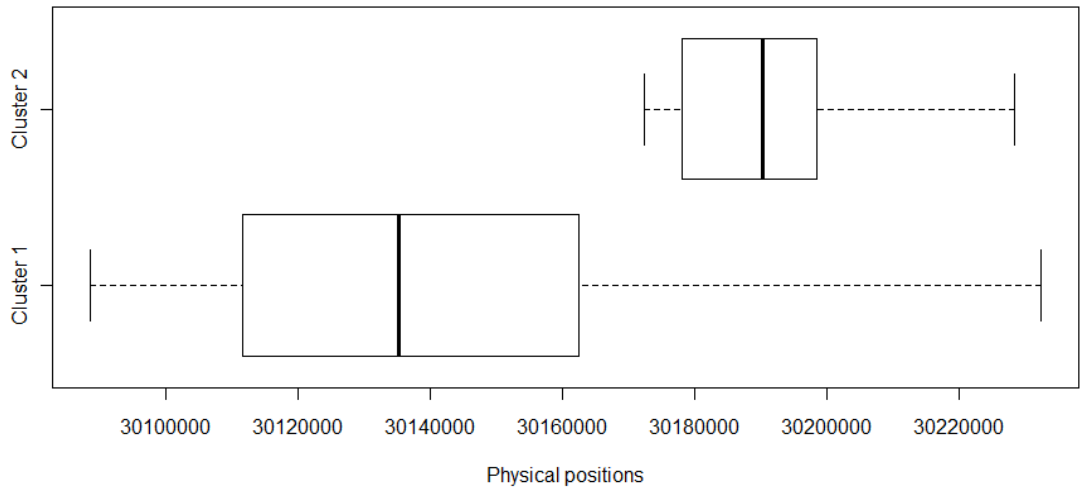


Figure 3.4 Box plots of two clusters from chromosome 15 region I, with linkage disequilibrium *r* or Cramér's *V*. Box plots illustrate two clusters are overlapped in terms of SNPs locations.

3.1.2 Biological Interpretation of Clustering Results

We first revisit clustering tables with all the 215 SNPs, by looking at outliers in Cluster0. Canonical correlation, LD (*r*) and Cramér's *V* share four common outliers: *rs17621256* in chromosome 8, *rs16925377* in chromosome 11, *rs2611605* and *rs3971872* in chromosome 15. *rs17621256* is located in the intron5 region of a neuronal nicotinic receptor gene *CHRNB3*, but did not show significant association

with “dizziness” to tobacco; on the other hand, the most significant SNPs identified in previous study, including *rs10958726*, *rs13277254* and *rs13277524*, are located in the upstream region (promoter) of *CHRNB3* (Enringer et al., 2009), and all three were grouped into the same cluster (Cluster1) in our analyses. *rs16925377* refers to a gene that does not belong to neuronal nicotinic receptors family, although it may have modest effect on nicotine dependence (Saccone et al., 2010). In our dataset, the other three SNPs from chromosome 11 are *rs2231532*, *rs2231529* and *rs6578411*. *rs2231532* is in the proximate region around *CHRNA10*, another neuronal nicotinic receptor while the other two have not been reported yet, regarding nicotinic phenotype association. More convincing evidence to support our clustering analyses comes from *rs2611605* and *rs3971872*, two *CHRN* related SNPs. Although quite a few *SNPs* in chromosome 15 have been identified to influence nicotine-addiction trait through *CHRN* genes cluster (Saccone et al., 2009), there is no paper thus far to claim similar findings for *rs2611605* and *rs3971872*.

Another way to verify our clustering results is to examine specifically SNPs that are already reported in nicotine-addiction studies. Complying with these previous findings, we would expect SNPs that biologically work together to be in the same cluster while SNPs from different functional proteins should be separated. Two recently identified SNPs on nicotine-addiction field are *rs16969968*, a non-synonymous *CHRNA5* SNP and *rs578776*, a *CHRNA3* SNP. It is reported that their effect’s directions are opposite (odds ratio equals to 1.4, 0.7, respectively) and their correlation is low (r^2 = 0.2). Furthermore, researchers performed joint logistic regression and found both of them remained significant (Saccone et al., 2009). Putting them together, a consistent clustering would put these two SNPs into different clusters. Taking canonical correlation measure for an example, in clustering analysis with all the 215 SNPs, *rs16969968* is in cluster 6 and *rs578776* is in cluster 2; in clustering analysis restricted to SNPs in chromosome 15, *rs16969968* is in cluster 4 and *rs578776* is in cluster 2. On the other hand, considering the four SNPs that refer to *IREB2* (*rs17483548*, *rs17405217*, *rs2656052*, and *rs17484235*), a neighboring gene

of *CHRN* gene cluster, we found all of them will be grouped into the same cluster, regardless of whether clustering is conducted with all the 215 SNPs or with only SNPs from chromosome 15.

3.2 Crohn's Disease Location Study

Subsequently, similar analyses were conducted on another dataset on Crohn's diseases. Crohn's diseases (CD) are chronic relapsing inflammatory intestinal disorder that can affect any segment of the intestine often in a discontinuous manner (Goyette et al., 2007; Abraham et al., 2009). One great advantage of GWAS on CD is that it has been intensively explored and more than 30 susceptibility loci have now been identified through genome-wide association studies (Barrett et al., 2008).

As a further exploratory step, CD patients are phenotypically heterogeneous. Efforts have been made to subphenotype the patients in order to facilitate genoytpe-phenotype correlations. Both the Vienna and Montreal classifications have classified the patients on the basis of three major parameters: age of diagnosis, disease location and disease behavior (Satsangi et al., 2006; Louis et al., 2001). While disease behavior changes over time (Unkart et al., 2008), disease location remains fairly stable. Based on both the Vienna and Montreal classifications, there are four major patterns of disease location: L1, ileal disease with or without cecal disease (ileal CD); L2, colonic disease only (Crohn's colitis); L3, ileal disease with colonic disease beyond the cecum; L4, proximal intestinal disease. Most CD patients have ileal and/or colonic disease (L1, L2, and L3). Only a small number of patients have disease restricted to the proximal gut (L4). Two identified CD-related genes: *NOD2* and *ATG16L1* have been previously associated with the subset of Crohn's disease patients with ileal disease location compared to control patients without inflammatory bowel diseases. These studies incorporated a relatively limited set of susceptibility loci (Cuthbert et al., 2002; Lesage et al., 2002; Prescott et al., 2007; Fowler et al., 2008; Van Limbergen et al., 2008; Márquez et al., 2009).

3.2.1 Clustering with 29 SNPs

Our dataset contains 628 CD patients within the Washington University Digestive Diseases Research Core Center Tissue Procurement Facility database (recruited between April 2005 - February 2010) that have complete genotype information on 31 established CD risk alleles (Barrett et al., 2008) (Table 3.10) and complete clinical information on disease location (L1-L4), smoking, gender, race and age of diagnosis (Table 3.11). From a case-control study point of view, we intended to carry out comparison between (L1 + L3) vs. L2, which is ileal (case) vs. non-ileal (control). With chi-square test, univariate results on such association are also attached in Table 3.12 and 3.13. Furthermore, in practice, we excluded those six L4 observations and combined three SNPs information on *NOD2*. Therefore it is a dataset including 622 samples and 29 SNPs in total. Afterwards, we are able to verify whether clustering analyses would be consistent with our published findings (Chen et al., 2011).

Gene	SNP	L1 *n = 288*	L2 *n = 131*	L3 *n = 203*	L4 *n = 6*	L1 + L3 vs. L2 *P-value*
NOD2	(composite)					**<0.0001**
R/R	rs2066847	36 (12%)	3 (2%)	13 (7%)	0 (0%)	
R/NR	rs2066844	87 (30%)	18 (14%)	47 (23%)	2 (33%)	
NR/NR	rs2066845	165 (57%)	110 (84%)	143 (70%)	4 (67%)	
ATG16L1	rs2241880					0.1227
R/R		101 (35%)	39 (30%)	74 (36%)	4 (67%)	
R/NR		130 (45%)	56 (43%)	91 (45%)	0 (0%)	
NR/NR		57 (20%)	36 (27%)	38 (19%)	2 (33%)	
IL23R	rs11209026					0.6729
R/R		276 (96%)	125 (95%)	186 (92%)	6 (100%)	
R/NR		12 (4%)	6 (5%)	17 (8%)	0 (0%)	
NR/NR		0 (0%)	0 (0%)	0 (0%)	0 (0%)	
IRGM	rs13361189					0.1971
R/R		11 (4%)	7 (5%)	4 (2%)	0 (0%)	
R/NR		77 (27%)	24 (18%)	41 (20%)	0 (0%)	
NR/NR		200 (69%)	100 (76%)	158 (78%)	6 (100%)	
STAT3	rs744166					0.5200
R/R		95 (33%)	48 (37%)	64 (32%)	1 (17%)	
R/NR		143 (50%)	60 (46%)	109 (54%)	2 (33%)	

NR/NR		50 (17%)	23 (18%)	30 (15%)	3 (50%)	
ICOSLG	rs762421					0.7857
R/R		48 (17%)	18 (14%)	25 (12%)	0 (0%)	
R/NR		140 (49%)	63 (48%)	108 (53%)	4 (67%)	
NR/NR		100 (35%)	50 (38%)	70 (34%)	2 (33%)	
X21q21	rs1736135					0.1709
R/R		110 (38%)	58 (44%)	98 (48%)	2 (33%)	
R/NR		143 (50%)	49 (37%)	80 (39%)	4 (67%)	
NR/NR		35 (12%)	23 (18%)	25 (12%)	0 (0%)	
7p12	rs1456893					0.9087
R/R		149 (52%)	65 (50%)	101 (50%)	2 (33%)	
R/NR		119 (41%)	55 (42%)	85 (42%)	3 (50%)	
NR/NR		20 (7%)	11 (8%)	17 (8%)	1 (17%)	
LOC4411108	rs2188962					0.0562
R/R		57 (20%)	17 (13%)	49 (24%)	0 (0%)	
R/NR		131 (45%)	62 (47%)	96 (47%)	4 (67%)	
NR/NR		100 (35%)	52 (40%)	58 (29%)	2 (33%)	
ITLN1	rs2274910					0.9385
R/R		122 (42%)	60 (46%)	94 (46%)	4 (67%)	
R/NR		135 (47%)	58 (44%)	88 (43%)	2 (33%)	
NR/NR		31 (11)	13 (10%)	21 (10%)	0 (0%)	
CCR6	rs2301436					0.8455
R/R		64 (22%)	29 (22%)	51 (25%)	2 (33%)	
R/NR		150 (52%)	71 (54%)	101 (50%)	4 (67%)	
NR/NR		74 (26%)	31 (24%)	51 (25%)	0 (0%)	
PTPN2	rs2542151					0.0751
R/R		21 (7%)	2 (2%)	9 (4%)	0 (0%)	
R/NR		79 (27%)	45 (34%)	69 (34%)	4 (67%)	
NR/NR		188 (65%)	84 (64%)	125 (62%)	2 (33%)	
PTPN22	rs2476601					0.3271
R/R		251 (87%)	112 (85%)	175 (86%)	2 (33%)	
R/NR		36 (13%)	17 (13%)	27 (13%)	4 (67%)	
NR/NR		1 (0%)	2 (2%)	1 (0%)	0 (0%)	
TNFSF15	rs4263839					**0.0241**
R/R		159 (55%)	72 (55%)	94 (46%)	3 (50%)	
R/NR		108 (38%)	56 (43%)	86 (42%)	3 (50%)	
NR/NR		21 (7%)	3 (2%)	23 (11%)	0 (0%)	
ORMDL3	rs2872507					0.1031
R/R		58 (20%)	35 (27%)	35 (17%)	1 (17%)	
R/NR		133 (46%)	62 (47%)	103 (51%)	2 (33%)	
NR/NR		97 (34%)	34 (26%)	65 (32%)	3 (50%)	
MST1	rs3197999					0.7062
R/R		31 (11%)	16 (12%)	19 (9%)	1 (17%)	
R/NR		121 (42%)	56 (43%)	83 (41%)	2 (33%)	

NR/NR		136 (47%)	59 (45)	101 (50%)	3 (50%)	
C11orf30	rs7927894					0.4907
R/R		54 (19%)	22 (17%)	41 (20%)	1 (17%)	
R/NR		130 (45%)	67 (51%)	92 (45%)	4 (67%)	
NR/NR		104 (36%)	42 (32%)	70 (34%)	1 (17%)	
C13orf31	rs3764147					0.3284
R/R		20 (7%)	14 (11%)	14 (69%)	1 (17%)	
R/NR		103 (36%)	47 (36%)	90 (44%)	1 (17%)	
NR/NR		165 (57%)	70 (53%)	99 (49%)	4 (67%)	
PTGER4	rs4613763					0.2659
R/R		16 (6%)	3 (2%)	11 (5%)	0 (0%)	
R/NR		84 (30%)	36 (27%)	61 (30%)	3 (50%)	
NR/NR		188 (65%)	92 (70%)	131 (65%)	3 (50%)	
CDKAL1	rs6908425					0.5650
R/R		181 (63%)	80 (61%)	122 (60%)	4 (67%)	
R/NR		96 (33%)	44 (34%)	75 (37%)	2 (33%)	
NR/NR		11 (4%)	7 (5%)	6 (3%)	0 (0%)	
6q21	rs7746082					0.4046
R/R		24 (8%)	8 (6%)	23 (11%)	0 (0%)	
R/NR		105 (36%)	55 (43%)	80 (39%)	4 (67%)	
NR/NR		159 (55%)	68 (52%)	100 (49%)	2 (33%)	
1q24	rs9286879					**0.0493**
R/R		25 (9%)	7 (5%)	20 (10%)	0 (0%)	
R/NR		112 (39%)	65 (50%)	76 (37%)	3 (50%)	
NR/NR		151 (52%)	59 (45%)	107 (53%)	3 (50%)	
IL12B	rs10045431					0.0760
R/R		161 (56%)	76 (58%)	96 (47%)	4 (67%)	
R/NR		112 (39%)	43 (33%)	95 (47%)	2 (33%)	
NR/NR		15 (5%)	12 (9%)	12 (6%)	0 ()	
JAK2	rs10758669					0.7569
R/R		127 (44%)	59 (45%)	75 (37%)	2 (33%)	
R/NR		123 (43%)	57 (44%)	103 (51%)	2 (33%)	
NR/NR		37 (13%)	15 (11%)	25 (12%)	2 (33%)	
10p11	rs17582416					0.3411
R/R		40 (14%)	12 (9%)	29 (14%)	1 (17%)	
R/NR		128 (44%)	63 (48%)	94 (46%)	1 (17%)	
NR/NR		120 (42%)	56 (43%)	80 (39%)	4 (67%)	
NKX2-3	rs11190140					0.5257
R/R		76 (26%)	32 (24%)	48 (24%)	1 (17%)	
R/NR		141 (49%)	71 (54%)	100 (49%)	3 (50%)	
NR/NR		71 (25%)	28 (21%)	55 (27%)	2 (33%)	

ZNF365	rs10995271					0.2432
R/R		61 (21%)	18 (14%)	34 (17%)	0 (0%)	
R/NR		126 (44%)	70 (53%)	102 (50%)	3 (50%)	
NR/NR		101 (35%)	43 (33%)	67 (33%)	3 (50%)	
LOC651731	rs11584383					0.6580
R/R		172 (60%)	71 (54%)	109 (54%)	5 (83%)	
R/NR		90 (31%)	52 (40%)	84 (41%)	1 (17%)	
NR/NR		26 (9%)	8 (6%)	10 (5%)	0 (0%)	
MUC19	rs11175593					0.5385
R/R		0 (0%)	0 (0%)	0 (0%)	0 (0%)	
R/NR		21 (7%)	6 (5%)	10 (5%)	0 (0%)	
NR/NR		267 (93%)	125 (95%)	193 (95%)	6 (100%)	

Table 3.10 Joint distribution of CD genotypes with the four major disease locations. L1, L2, L3 include subjects with L4 as a modifier. L4 refers to patients with only L4 disease location. P-values result from chi-square test on individual SNP with L1 + L3 vs. L2. SNPs with significant P-values (<.05) are highlighted in red.

	L1 *n = 288*	L2 *n = 131*	L3 *n = 203*	L4 *n = 6*	L1 + L3 vs. L2 *P-value*
Gender (male)	128 (44%)	65 (50%)	92 (45%)	5 (92%)	0.3244
Race					0.0852
White	262 (91%)	110 (84%)	182 (90%)	5 (92%)	
Black	23 (8%)	18 (14%)	18 (9%)	0 (0%)	
Other	3 (1%)	3 (2%)	3 (1%)	1 (8%)	
Smoking habit					**0.0047**
Smoker	113 (39%)	33 (25%)	73 (36%)	2 (33%)	
Ex-smoker	17 (6%)	17 (13%)	15 (7%)	1 (17%)	
Non-smoker	158 (55%)	81(62%)	115 (57%)	3 (50%)	
Age of Diagnosis					0.3553
A1 (<17 y)	32 (11%)	21 (13%)	43 (21%)	0 (0%)	
A2 (14-40y)	202 (70%)	79 (60%)	123 (61%)	5(92%)	
A3 (>40y)	54 (19%)	31 (27%)	37 (18%)	1 (8%)	
Surgery	222 (77%)	54 (41%)	132 (65%)	4 (67%)	

Table 3.11 Patient clinical characteristics. L1, L2, L3 include subjects with L4 as a modifier. L4 refers to patients with only L4 disease location. P-values result from chi-square test on individual variable with L1 + L3 vs. L2. Smoking is highlighted in red with significant p-value (<0.05).

However, it could be inappropriate for us to evaluate clustering outputs according to chromosomal separation, because of the sparsity of these 29 SNPs. The 29 selected

SNPs are located in fifteen different chromosomes. Five of them are in chromosome 1, which is the most "concentrated" one in terms of the number of SNPs. Alternatively, since it is a rather small set of variables, the entire dendrogram would be readable and comparable across different measures. In addition, we actually can not apply R package *DynamicTreeCut* here since we may not be able to define a reasonable minimum cluster size in this case. Dendrograms based on canonical correlation, linkage disequilibrium *r*, Cramér's *V* and Pearson's *r* are displayed below (Figure 3.5 – Figure 3.8).

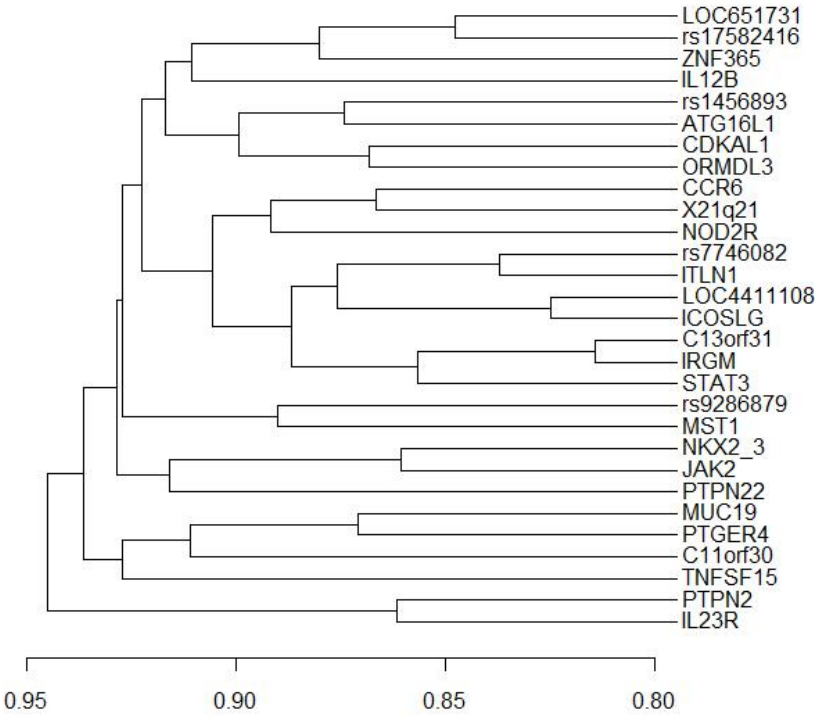


Figure 3.5 Clustering dendrogram of 29 SNPs with canonical correlation measure.

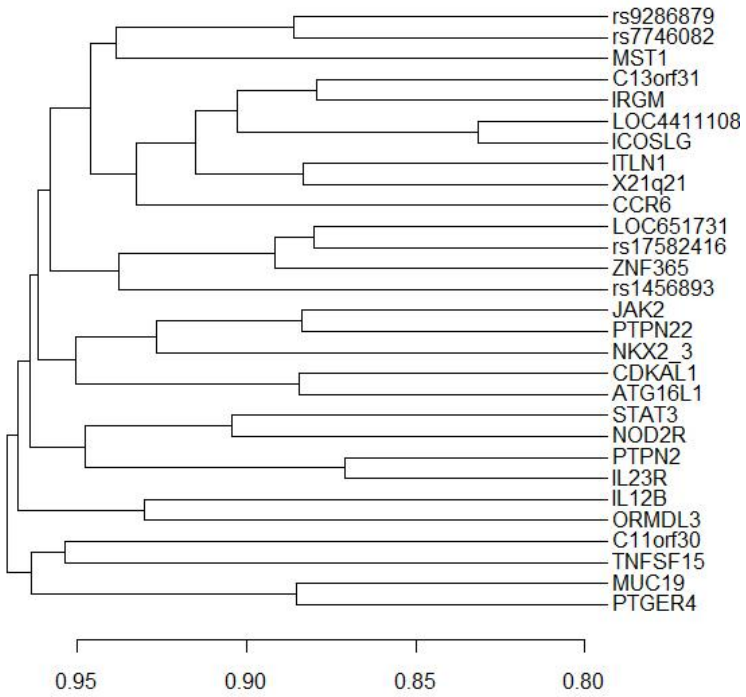


Figure 3.6 Clustering dendrogram of 29 SNPs with linkage disequilibrium *r*.

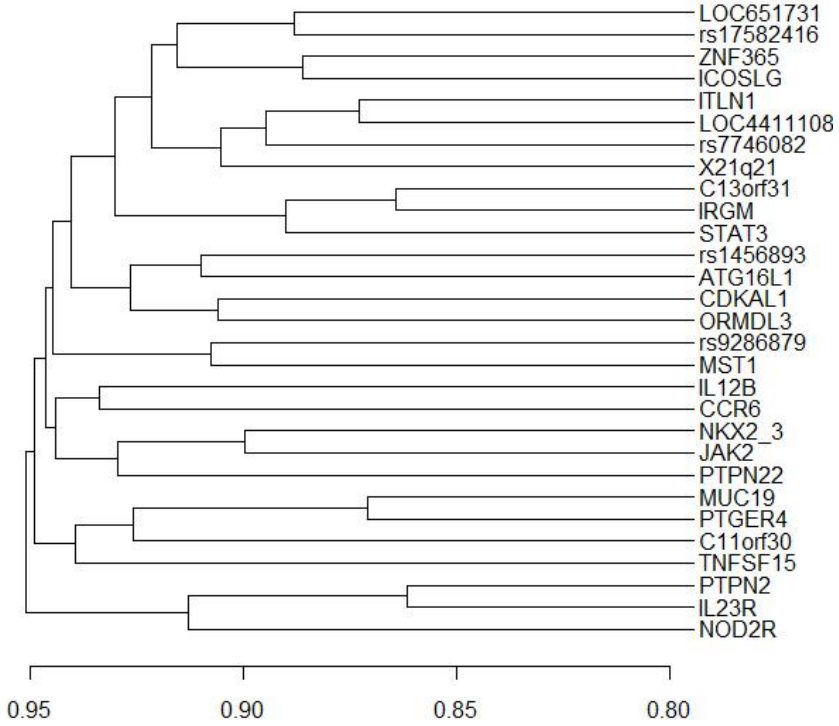


Figure 3.7 Clustering dendrogram of 29 SNPs with Cramér's *V*.

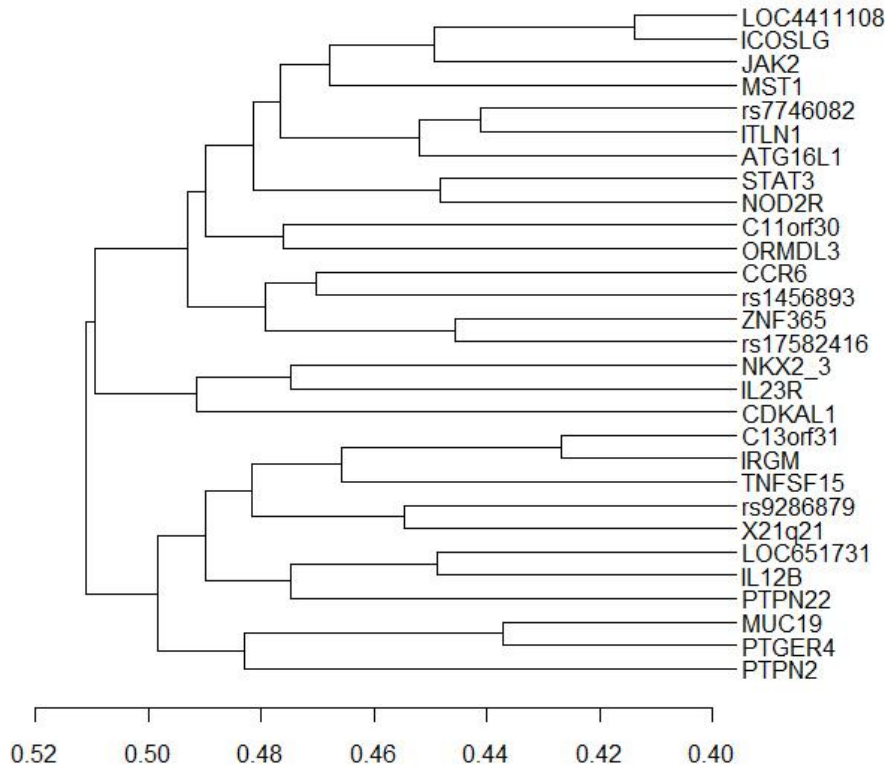


Figure 3.8 Clustering dendrogram of 29 SNPs with Pearson's *r*.

As expected, dendrograms based on canonical correlation, linkage disequilibrium *r* and Cramér's *V* share several consistent and biologically meaningful patterns. For instance, two immunity – related genes *IL23R* and *PTPN2* are always clustered at the very early step with all the three measures, these two genes have also been associated with other autoimmune disorders (Xavier and Podolsky, 2007). However, they appear to be distant to each other with Pearson' *r*, suggesting that SNPs should not be treated as numeric variables. More importantly, it is also possible for us to connect clustering output to our published variable selection result (Chen et al., 2011), where we identified *TNFSF15* and *NOD2* after logistic regression with stepwise selection (Table 3.12). According to outputs above, *TNFSF15* and *NOD2* are located in distant branches: their mergence occurred in later stage of clustering.

	P-value	R/R vs. NR/NR	R/NR vs. NR/NR	R/R vs. R/NR
NOD2	<0.0001	6.102	2.552	2.391

smoking	0.0082	1.689	0.647	2.611
TNSF15	0.0479	0.227	0.218	1.045

Table 3.12 L1 +L3 (ileal CD and ileocolonic) vs. L2 (nonileal) logistic regression analysis with stepwise variable selection.

3.2.2 Including Smoking in Analyses

It is noteworthy that we included smoking and other covariates in Table 3.11 when running variable selection in logistic regression model. And smoking was found to be significant in the final result. In practice, we also treated smoking as a categorical variable: non-smoker, ex-smoker and current smoker. Thus it might be acceptable to consider smoking as a "hypothesized" SNP: non-smoker, ex-smoker and current smoker corresponds to NR/NR, NR/R and R/R genotypes, respectively. Updated clustering results (Figure 3.9 – Figure 3.12), however, vary regarding the position of smoking in dendrogram. With canonical correlation and Cramér's *V* measures, *NOD2* and smoking are close to each other while linkage disequilibrium *r* successfully recognized smoking as a distinct branch. Such discrepancy might have two-fold meanings. First, canonical correlation measure would be more consistent with Cramér's *V* index because of their underlying theoretical connection. Second, simply treating smoking as another SNP variable remains questionable. The LD measure worked well with smoking status but does not guarantee the appropriateness of such application, especially when we notice that LD is derived from the allelic combination, a biological fact only meaningful with SNP data. Therefore, we should be cautious when extending clustering analysis with canonical correlation measure (or other measures) to a set of variables including other covariates. Finally, the clustering fact that smoking and NOD2 are close to each other does not necessarily contradict to our findings (Table 3.12). Although two highly correlated variables are less likely to be retained into a logistic regression model together, both of them can be still selected if they account for distinct effects on the response variable.

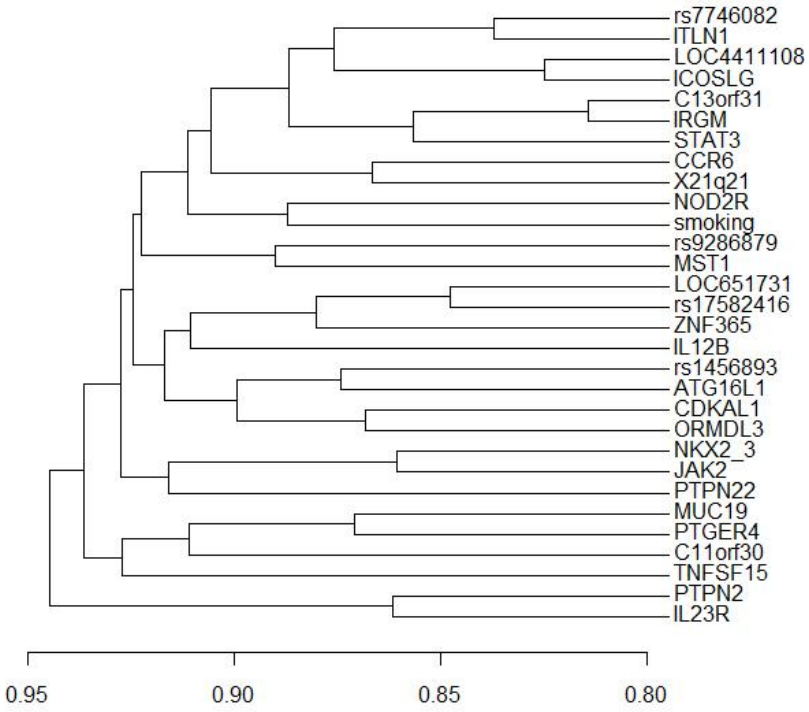


Figure 3.9 Clustering dendrogram of 29 SNPs and smoking status with canonical correlation measure.

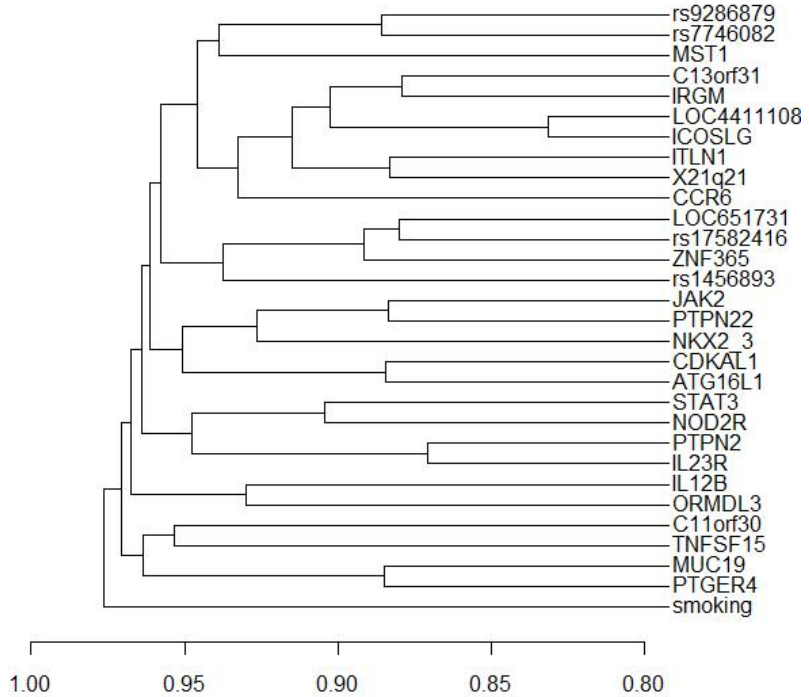


Figure 3.10 Clustering dendrogram of 29 SNPs and smoking status with linkage disequilibrium *r*.

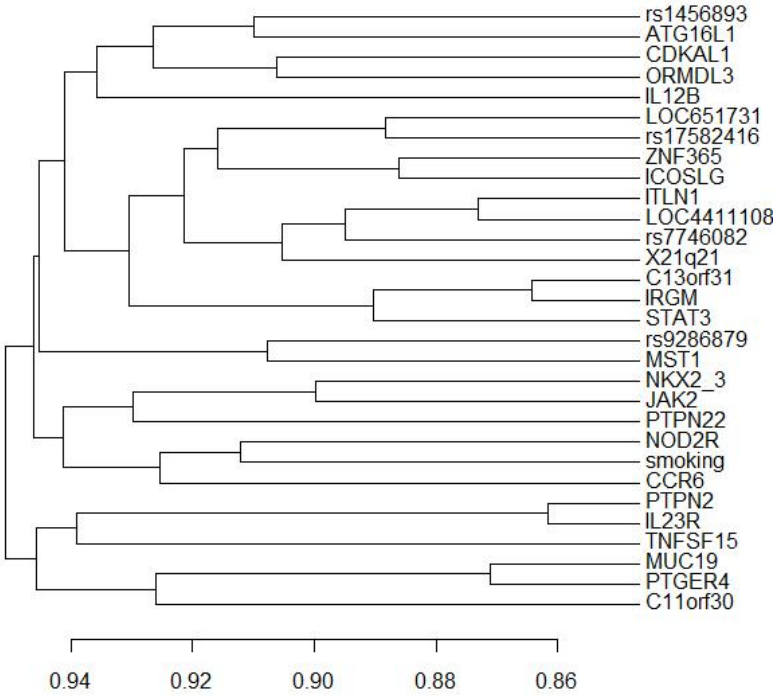


Figure 3.11 Clustering dendrogram of 29 SNPs and smoking status with Cramér's *V*.

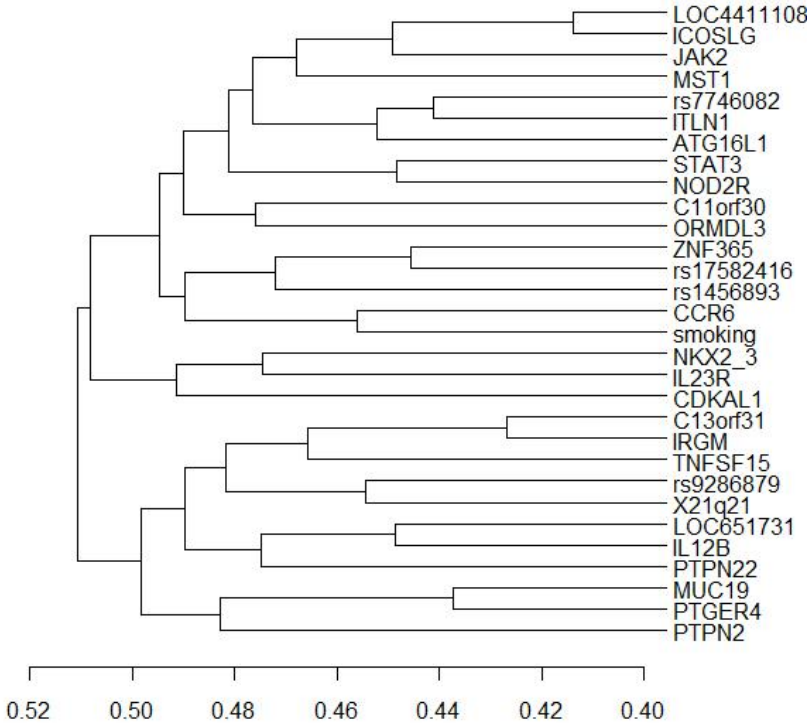


Figure 3.12 Clustering dendrogram of 29 SNPs and smoking status with Pearson's *r*.

Chapter 4 Network Analysis with SNP

Strictly speaking, motivation of performing network analysis with SNP is two-fold. First, similar to clustering analysis, network analysis may serve as another dimension reduction technique. For instance, if the underlying network represents a hub-structure, hub-SNPs can be identified through network analysis and be used for the follow-up analysis. Second, it would be informative for pathway findings. A typical example will be in cancer research. It is well known that loss of heterozygosity (LOH) is a major source in cancerogenic pathways. As a result, SNPs of interest can be coded as 1 (LOH occurs) and 0 (LOH does not occur) and then put into a categorical network analysis. In this way, edges linking SNPs together may indicate potential cancerogenic network.

Furthermore, we will have other experimental information from the sample biological sample, such as microarray, proteomics, covariates, etc. These other variables are likely to be continuous. Therefore, network analysis with SNP should be able to handle such mixed variable scenarios. In this chapter, we first review the existing methods for pure categorical and mixed scenarios, respectively and then propose our novel partial canonical correlation measure, which is readily adapted to both situations.

4.1 Partial Correlation Network Analysis (PCNA)

The studies on partial correlation analysis can be traced back to the early 20th century by Pearson, Fisher and others (Isserlis, 1914; Pearson, 1915, Fisher, 1924; Goodman and Kruskal, 1979). It is the correlation between two variables while effects from additional variables are controlled at the same time. One of its usages is in the causal analysis through graphic modeling. As a typical example, the partial correlation between two variables is compared to the conventional correlation without

excluding other variable effects: no difference between the two correlations suggests that the controlled variables have no effect on relation between these two variables. On the other hand, if the partial correlation approaches 0, it indicates that the original correlation is spurious and caused by controlled variables. In other words, partial correlation excludes the confounding effects.

Furthermore, a corresponding partial correlation network analysis (PCNA) can be performed graphically based on pairwise partial correlations of variables of interest. A statistically non-zero partial correlation between two variables is denoted with an edge to link them. Like most parametric statistical analyses, traditional PCNA requires the sample size (n) to be larger than the number of variables (p). Nevertheless, methods derived from the sparse property of partial correlation matrix have been introduced in recent studies to estimate partial correlation under the insufficient sample size condition or the high dimensional scenario ($p > n$): Schafer and Strimmer (2005) proposed a shrinkage covariance estimation procedure to overcome the ill-conditioned problem of sample covariance matrix when $p > n$; Li and Gui (2006) introduced a threshold gradient descent regularization procedure; Meinshausen and Buhlmann (2006) reported a variable-by-variable approach for neighborhood selection via the lasso regression; Yuan and Lin (2007) proposed a penalized maximum likelihood approach which performs model selection and estimation simultaneously and ensures the positive definiteness of the estimated concentration matrix; Friedman et al. (2008) proposed an improved algorithm so as to address problems with high dimensions; Bickel and Levina (2008) proposed to regularize the covariance matrix by hard thresholding for families of covariance matrices satisfying suitable sparsity assumptions; Peng et al (2009) developed a new algorithm based on the joint sparse regression model (JSRM). The simulation study shows an improvement in performance for $p >> n$ data, and efficiency in identifying network hubs. There are also quite a few applications of PCNA to exploratory gene microarray analysis, which generates many “hub-genes” as potential key regulators within the whole regulatory network (Barabasi and Oltvai, 2004). Finally, it should be noted that

so far PCNA mainly focuses on continuous (numeric) variable. When categorical variables are involved in PCNA, it is often referred to a categorical Markov Random Field problem.

4.1.1 PCNA with Categorical Data

In graph theory, PCNA generates an undirected graph: $G = \{V, E\}$, where node set V corresponds to variables $\{X_i\}$, the edge set E indicates pairwise relationship. If nodes are not linked by an edge, two corresponding variables are conditionally independent and such independence does not have directional information:

$$X_r \perp X_s \mid X_{-(r,s)} \Leftrightarrow X_s \perp X_r \mid X_{-(r,s)}$$

When V includes only continuous variables, joint Gaussian distribution is often assumed, so as to estimate the precision matrix (inversed variance-covariance matrix Σ^{-1}). On the other hand, when V is a set of only categorical variables, it is regarded as a categorical Markov Random Field, which possesses properties similar to a partial correlation network under the continuous scenario (Figure 4.1). Strictly speaking, continuous Markov Random Field is also a common tool for continuous PCNA. Nevertheless, estimation on a categorical Markov Random Field is notably harder, due to the complexity of the partition function.

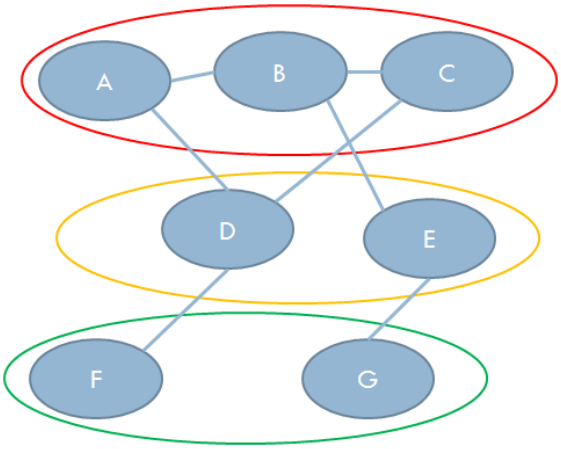


Figure 4.1 An illustration of Markov Random Field. $V = \{A, B, C, D, E, F, G\}$. Three major properties are: 1) two non-connected variables are conditionally independent given all the others, e.g., A is independent to C given $\{B, D, E, F, G\}$; 2)

a variable is conditionally independent of all the others given its connected neighbors, e.g., *E* is independent of {*A*, *C*, *D*}; 3) any two variable sets are conditionally independent given a separate one, e.g., {*A*, *B* ,*C*} is independent to {*F*, *G*} given {*D*, *E*}.

Consider a binary Markov random field. Under certain assumptions from graph theory (maximum cliques in graph ≤ 2), the probability distribution can be factorized as the following quadratic exponential model (Wang et al., 2011):

$$f(X_1,...,X_p)=\frac{1}{Z(\Theta)}\exp\left(\sum_{i=1}^{p}\theta_{i,i}X_i+\sum_{1\le i<i'\le p}\theta_{i,i'}X_iX_{i'}\right)$$

It has been shown that with this expression of the joint distribution, $\theta_{i,i'}$ will directly represent conditionally independence:

$$X_r \perp X_s \mid X_{-(r,s)} \Leftrightarrow \theta_{i,i'}=0$$

Therefore, identifying the edge set *E* in graph leads to estimating non-zero coefficients $\theta_{i,i'}$. As mentioned earlier, one major challenge is from the partition function $Z(\Theta)$, which is composed of 2^p terms. As a result, it is infeasible to directly maximize the exact joint distribution function. Instead, several groups have published their efforts on bypassing it. For example, one could use the log-determinant relaxation to approximate the log-partition function (D'Aspermont et al., 2008; Kolar and Xing, 2008). Recently, two independent pieces of research work successfully decomposed it into a joint logistic regression problem (Guo et al., 2010; Wang et al., 2011). In the former paper, the joint distribution function was approximated with pseudo-likelihoods (Besag, 1975):

$$f(X_1,...,X_p)\approx\prod_{i=1}^{p}f(X_i\mid X_1,...,X_{i-1},X_{i+1},...,X_p)$$

Afterwards, each conditional probability density function was assumed to follow Bernoulli distribution, or equivalently modeled with *p* logistic regressions:

$$X_i \mid X_1,...,X_{i-1},X_{i+1},...,X_p \sim Bernoulli(\mu_i)$$

$$\mu_i = \frac{\exp(\theta_{i,i} + \sum_{k\neq i}\theta_{i,k}X_k)}{1+\exp(\theta_{i,i} + \sum_{k\neq i}\theta_{i,k}X_k)}$$

$$f(X_1,...,X_p) = \prod_{i=1}^{p} f(X_i \mid X_1,...,X_{i-1},X_{i+1},...,X_p) = \prod_{i=1}^{p} \mu_i^{X_i}(1-\mu_i)^{1-X_i}$$

Alternatively, the partition function $Z(\Theta)$ can be canceled out with the likelihood ratio:

$$f(X_1,...,X_p) = \frac{1}{Z(\Theta)}\exp\left(\sum_{i=1}^{p}\theta_{i,i}X_i + \sum_{1\le i<i'\le p}\theta_{i,i'}X_iX_{i'}\right)$$

$$\Rightarrow \begin{cases} f(X_1=0, X_2=x_2,...,X_p=x_p) = \frac{1}{Z(\Theta)}\exp\left(\sum_{i=2}^{p}\theta_{i,i}x_i + \sum_{2\le i<i'\le p}\theta_{i,i'}x_ix_{i'}\right) \\ f(X_1=1, X_2=x_2,...,X_p=x_p) = \frac{1}{Z(\Theta)}\exp\left(\sum_{i=1}^{p}\theta_{i,i}x_i + \sum_{1\le i<i'\le p}\theta_{i,i'}x_ix_{i'}\right) \end{cases}$$

$$\Rightarrow \frac{f(X_1=1, X_2=x_2,...,X_p=x_p)}{f(X_1=0, X_2=x_2,...,X_p=x_p)} = \frac{f(X_1=1 \mid X_2=x_2,...,X_p=x_p)}{f(X_1=0 \mid X_2=x_2,...,X_p=x_p)}$$

$$= \exp\left(\theta_{1,1} + \sum_{2\le i<i'\le p}\theta_{i,i'}x_ix_{i'}\right)$$

$$\Rightarrow \text{logit}\left\{f(X_1=1 \mid X_2=x_2,...,X_p=x_p)\right\} = \theta_{1,1} + \sum_{2\le i<i'\le p}\theta_{i,i'}x_ix_{i'}$$

$$\vdots$$

$$\text{logit}\left\{f(X_p=1 \mid X_1=x_1,...,X_{p-1}=x_{p-1})\right\} = \theta_{p,p} + \sum_{1\le i<i'\le p-1}\theta_{i,i'}x_ix_{i'}$$

The early work on estimating coefficients within p logistic regressions focused on separate model fitting with l_1-regularization (Ravikumar et al., 2010):

$$\max_{\{\theta_{j,k}\}_{k=1}^{p}} \sum_{i=1}^{n}\left[x_{i,j}(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}) - \log\left\{1+\exp\left(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}\right)\right\}\right] - \lambda_j\sum_{k\neq j}|\theta_{j,k}|$$

Criticisms on this approach mainly address the symmetry issue that separate fitting will not ensure $\theta_{i,i'} = \theta_{i',i}$. Since it is an undirected graph, symmetry restriction must be imposed. Consequently, certain post hoc rules are necessary when inequality occurs, such like maximum aggregation, minimum aggregation, etc. Furthermore, λ_j's are usually set to be the same value in practice, which is obviously not suitable for a

non-homogeneous network, i.e., existence of hubs. Hence joint structure estimation method was later proposed:

$$\max_{\Theta} \sum_{j=1}^{p} \sum_{i=1}^{n} \left[x_{i,j}(\theta_{j,j} + \sum_{k \neq j} \theta_{j,k} x_{i,k}) - \log \left\{ 1 + \exp\left(\theta_{j,j} + \sum_{k \neq j} \theta_{j,k} x_{i,k} \right) \right\} \right] - \lambda \sum_{j<j'} | \theta_{j,j'} |$$

subject to $\theta_{j,j'} = \theta_{j',j}$

The maximization problem is solvable with various algorithms (Friedman et al., 2007 and 2010; Hofling and Tibshirani, 2009; Peng et al., 2009). In other words, we are able to construct a SNP network with the control of complexity of network by using joint sparse logistic regression model.

4.1.2 Partial Canonical Correlation Measurement

We hereby present another aspect of PCNA with SNPs. It can be viewed as an effort on making ordinary partial correlation measurable for categorical variables. In a general definition, partial correlation is the Pearson's correlation *r* between residuals from two regressions on controlled variables:

$$X = Z\beta + \varepsilon_x, Y = Z\beta' + \varepsilon_y$$

$$r_{xy|Z} = cor(\hat{\varepsilon}_x, \hat{\varepsilon}_y)$$

If *X* and *Y* are categorical variables here, we can obtain Pearson residuals from logistic regressions. Considering a binary case:

$$\text{logit}\left(\frac{\pi_x}{1-\pi_x} \right) = Z\beta_x, \text{logit}\left(\frac{\pi_y}{1-\pi_y} \right) = Z\beta_y$$

For each observation x_i and y_i, Pearson residuals are calculated as the following

$$\hat{r}_{x_i} = \frac{x_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}, \hat{r}_{y_i} = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$$

Pearson residuals are mainly used in diagnostics for logistic regression. Since they follow normal distribution asymptotically (Homster and Lemeshow, 2000), the sum of squared residuals forms a chi-square statistics, and may serve as goodness fit index. On the other hand, Pearson residual was also suggested for a second-stage analysis,

by only including confounding effects from logistic regression in the first step (other graduate students work from our group). Thus it would be acceptable to apply conventional correlation measure/test here to detect conditional dependency between *X* and *Y* given *Z* (*Z* refers to controlled or cofounding effect). Additionally, there are other types of residuals that can be derived from logistic regression model, for instance, deviance residuals based on likelihood ratio concept. But other types of residuals would not have asymptotic normality property and subsequently are not appropriate in a second-state analysis.

More importantly, this method can be readily extended to a multi-category scenario, where multinomial logistic regression will be used. Taking SNP for an example, a reference group is needed to fit a multinomial model, which is applicable for the SNP. As explained in our canonical correlation measure, we can always treat non-risk homozygous (*AA*) as the baseline. Subsequently, we can have generalized residual form:

$$\begin{cases} \text{logit}\left(\frac{\pi_{x1}}{\pi_{x0}}\right) = Z\beta_{x1} \\ \text{logit}\left(\frac{\pi_{x2}}{\pi_{x0}}\right) = Z\beta_{x2} \end{cases}, \begin{cases} \text{logit}\left(\frac{\pi_{y1}}{\pi_{y0}}\right) = Z\beta_{y1} \\ \text{logit}\left(\frac{\pi_{y2}}{\pi_{y0}}\right) = Z\beta_{y2} \end{cases}$$

$$\Rightarrow \begin{cases} \hat{r}_{x_{i1}} = \frac{x_i - \hat{\pi}_{i1}}{\sqrt{\hat{\pi}_{i1}(1-\hat{\pi}_{i1})}}, x_i = 1 \text{ when } x_i \text{ belongs to category 1; } x_i = 0 \text{ otherwise} \\ \hat{r}_{x_{i2}} = \frac{x_i - \hat{\pi}_{i2}}{\sqrt{\hat{\pi}_{i2}(1-\hat{\pi}_{i2})}}, x_i = 1 \text{ when } x_i \text{ belongs to category 2; } x_i = 0 \text{ otherwise} \end{cases}$$

$$\begin{cases} \hat{r}_{y_{i1}} = \frac{y_i - \hat{\pi}_{i1}}{\sqrt{\hat{\pi}_{i1}(1-\hat{\pi}_{i1})}}, y_i = 1 \text{ when } y_i \text{ belongs to category 1; } y_i = 0 \text{ otherwise} \\ \hat{r}_{y_{i2}} = \frac{y_i - \hat{\pi}_{i2}}{\sqrt{\hat{\pi}_{i2}(1-\hat{\pi}_{i2})}}, y_i = 1 \text{ when } x_i \text{ belongs to category 2; } y_i = 0 \text{ otherwise} \end{cases}$$

It can also be shown that residuals $\{r_{x1}, r_{x2}\}$ and $\{r_{y1}, r_{y2}\}$ will asymptotically follow multivariate normal distribution (Seber and Nyangoma, 2000). Hence, the partial correlation between *X* and *Y* given *Z* can be defined as the first canonical correlation between $\{r_{x1}, r_{x2}\}$ and $\{r_{y1}, r_{y2}\}$. And using canonical correlation test on Pearson

residuals to detect significant edges in a network is a novel application.

To verify our new measure/test, simulations are performed, addressing four different scenarios. In a simple case involving three categorical variables – *X*, *Y* and *Z*, we treat *Y* as response variable that is affected by both *X* and *Z*, which are simulated independently (Figure 4.2). We expect to see significant partial correlation between *X* and *Y* while controlling *Z* in our new measure. In other words, the first canonical correlation test between $\{r_{x1}, r_{x2}\}$ and $\{r_{y1}, r_{y2}\}$ is expected to be significant. Assuming they are all SNPs, *Y* is simulated as the following

$$\begin{cases} \text{logit}\left(\frac{\pi_{y1}}{\pi_{y0}}\right) = \beta_1 + X_1\beta_{X11} + X_2\beta_{X12} + Z_1\beta_{z11} + Z_2\beta_{z12} \\ \text{logit}\left(\frac{\pi_{y2}}{\pi_{y0}}\right) = \beta_2 + X_1\beta_{X21} + X_2\beta_{X22} + Z_1\beta_{z21} + Z_2\beta_{z22} \end{cases}$$

Note that X_1, X_2, Z_1 and Z_2 refer to coded dummy variables (linear and quadratic minor allele frequencies *S* and S^2) in regression. Set $[\beta_1, \beta_{X11}, \beta_{X12}, \beta_{Z11}, \beta_{Z12}]^T$= [-2, 1, 0.2, 1, 0.2]^T; $[\beta_2, \beta_{X21}, \beta_{X22}, \beta_{Z21}, \beta_{Z22}]^T$= [-3, 1.2, 0.2, 1.2, 0.2]^T. The intercept term is negative since we expect the reference group (*AA*) to be the majority; the coefficient for S^2 is set to be lower than that for *S*. Effects for *X* and *Z* are set to be equal. *X* and *Z* are generated independently with a multinomial distribution (0.5, 0.3, 0.2), corresponding to (*AA*, *Aa*, *aa*) with *N* (sample size) = 100. For one simulated dataset, *Y* will have the following distribution: *AA* – 58, *Aa* - 27 and *aa* – 15, which resembles the distributions of *X* and *Z*. We also calculated the *LDs* between SNPs: $r^2_{x,y}$ = 0.50 and $r^2_{y,z}$ = 0.34, which are not large effect sizes (r^2 < 0.8). Thus the coefficients set for the logistic regressions are appropriate since we do not arbitrarily create a dataset with large effect sizes that might be easier to be detected with canonical correlation measure. After 500 simulations, we record the number of times the first canonical correlation test between $\{r_{x1}, r_{x2}\}$ and $\{r_{y1}, r_{y2}\}$ is significant (Figure 4.3): 461/500 = 92.2%. We also record the number of times the first canonical correlation test between $\{r_{x1}, r_{x2}\}$ and $\{r_{z1}, r_{z2}\}$ is not significant (Figure 4.4): 433/500 = 86.6%

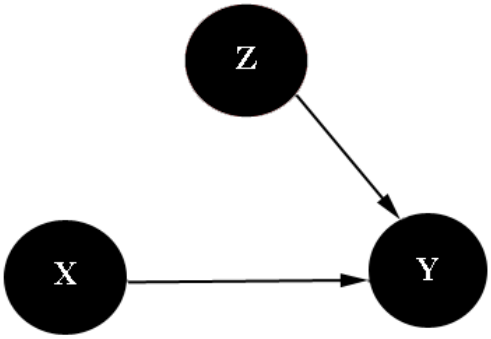


Figure 4.2 Simulation scenario I for pure categorical network.

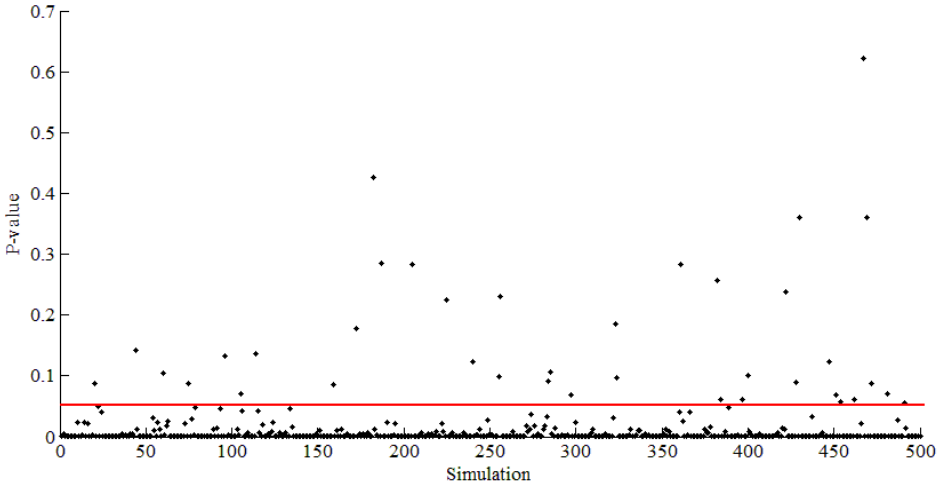


Figure 4.3 p-values plot from the canonical correlation test between {r_{x1}, r_{x2}} and {r_{y1}, r_{y2}}.

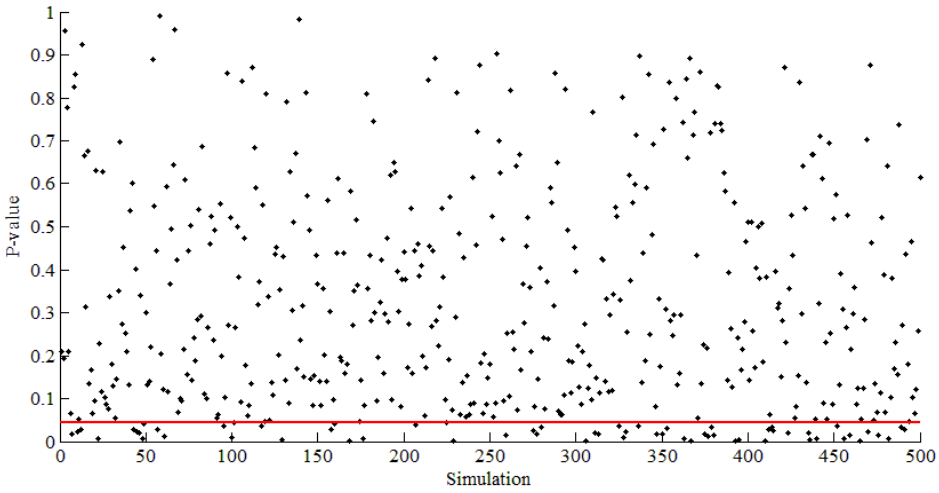


Figure 4.4 p-values plot from the canonical correlation test between {r_{x1}, r_{x2}} and {r_{z1}, r_{z2}}.

In a second scenario, Z does not affect Y, ($\beta_{Z21}=\beta_{Z22}=\beta_{Z11}=\beta_{Z12}=0$), while the same values for all the other parameters remain the same (Figure 4.5). We record the number of times the first canonical correlation test between $\{r_{x1}, r_{x2}\}$ and $\{r_{y1}, r_{y2}\}$ is significant: 402/500 = 80.4%

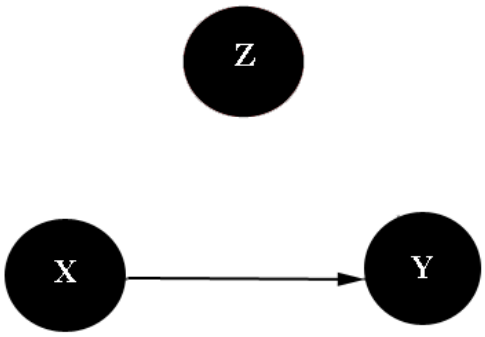


Figure 4.5 Simulation scenario II for pure categorical network.

Now considering the situation where X and Y has no conditional relation (Figure 4.6), then we have

$$\begin{cases} \text{logit}\left(\frac{\pi_{x1}}{\pi_{x0}}\right) = \beta_1 + Z_1\beta_{z11} + Z_2\beta_{z12} \\ \text{logit}\left(\frac{\pi_{x2}}{\pi_{x0}}\right) = \beta_2 + Z_1\beta_{z21} + Z_2\beta_{z22} \end{cases}$$

$$\begin{cases} \text{logit}\left(\frac{\pi_{y1}}{\pi_{y0}}\right) = \beta_1 + Z_1\beta_{z11} + Z_2\beta_{z12} \\ \text{logit}\left(\frac{\pi_{y2}}{\pi_{y0}}\right) = \beta_2 + Z_1\beta_{z21} + Z_2\beta_{z22} \end{cases}$$

For X, set $[\beta_1, \beta_{Z11}, \beta_{Z12}, \beta_2, \beta_{Z21}, \beta_{Z22}]^T$ = [-2, 1, 0.2, -3, 1.2, 0.2]^T; for Y, set $[\beta_1, \beta_{Z11}, \beta_{Z12}, \beta_2, \beta_{Z21}, \beta_{Z22}]^T$ = [-2, 1, 0.2, -3, 1.2, 0.2]^T. Thus effects from Z to X and Y are equal. Z and the other parameters are generated as mentioned above. We record the number of times the first canonical correlation test between $\{r_{x1}, r_{x2}\}$ and $\{r_{y1}, r_{y2}\}$ is not significant: 469/500 = 93.8%

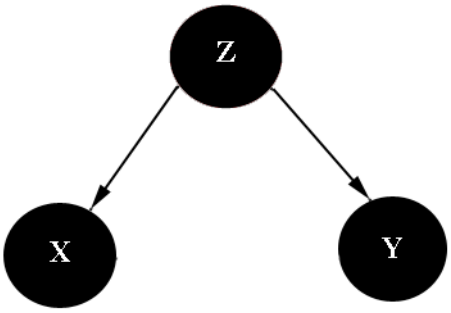


Figure 4.6 Simulation scenario III for pure categorical network.

Figure 4.7 displays another very typical partial correlation scenario for continuous variables, where *X* and *Y* would be correlated but not partially correlated given *Z*. Data are simulated sequentially:

$$X \sim \text{multinomial}(0.5, 0.3, 0.2)$$

$$\Rightarrow \begin{cases} \text{logit}\left(\frac{\pi_{z1}}{\pi_{z0}}\right) = \beta_1 + X_1\beta_{x11} + X_2\beta_{x12} \\ \text{logit}\left(\frac{\pi_{z2}}{\pi_{z0}}\right) = \beta_2 + X_1\beta_{x21} + X_2\beta_{x22} \end{cases}$$

$$\Rightarrow \begin{cases} \text{logit}\left(\frac{\pi_{y1}}{\pi_{y0}}\right) = \beta_1 + Z_1\beta_{z11} + Z_2\beta_{z12} \\ \text{logit}\left(\frac{\pi_{y2}}{\pi_{y0}}\right) = \beta_2 + Z_1\beta_{z21} + Z_2\beta_{z22} \end{cases}$$

From *X* to *Z*, $[\beta_1, \beta_{X11}, \beta_{X12}, \beta_2, \beta_{X21}, \beta_{X22}]^T$ = [-2, 1, 0.2, -3, 1.2, 0.2]^T; from *Z* to *Y*, $[\beta_1, \beta_{Z11}, \beta_{Z12}, \beta_2, \beta_{Z21}, \beta_{Z22}]^T$ = [-2, 1, 0.2, -3, 1.2, 0.2]^T. Hence we expect our canonical correlation test will not reject the null hypothesis regarding *X* and *Y*, the number of times of such output: 471/500 = 94.2%; on the other hand, we expect significant canonical correlation from the test between *X* and *Z* (490/500 = 98%) and the test between *Y* and *Z* (440/500 = 88%)

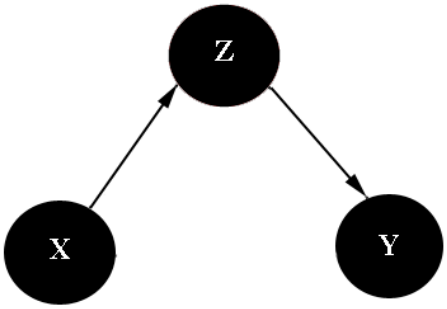


Figure 4.7 Simulation scenario IV for pure categorical network.

Finally, it is noteworthy to point out that there are other conditional association tests available in our simple simulation scenarios, such as generalized Cochran–Mantel–Haenszel (CMH) test on a $M \times N \times K$ table, which can be viewed as an association test on a contingency table while controlling the third stratification variable (Agresti 2002). But this test would be not applicable with more than three variables, and more importantly, can be not generalized to a mixed scenario discussed later. Hence, we did not compare our method with CMH test in our simulation examples as CMH test can be used in real SNP datasets (COGEND and Crohn's Disease). The main purpose of simulations illustrated above is to verify that our method is powerful enough to detect the underlying true conditional dependencies in model.

4.2 Network Analysis with SNP and Other Variables

In real data analysis, it is quite likely for us to have other types of variable than SNP. Subsequently, different approaches would be applicable, depending on the properties of other variables and specific research aims, for example, categorical covariate in network and mixed Bayesian network.

4.2.1 Covariate in network

Suppose the number of covariates is manageable and covariates are categorical, we can generate networks one by one, according to each assignment to covariates. Then the task turns to be evaluating whether those networks are significantly different from each other. One previous graduate student from our research group, Dr. Kith Pradhan, studied comprehensively on this topic. Generally speaking, there are four ways to conduct such partial correlation analysis: Fisher's transformation, bootstrapping, two-level regression and likelihood ratio test. The former two are existing methods and the others are novel developed methods. However, not all of them can be directly applied to our research problems since they are all under ordinary continuous variables scenario. The most applicable extension from Dr. Ktih Pradhan's work would be bootstrapping, where we compare partial correlations in a non-parametric way, without the continuous case assumption.

4.2.2 Mixed Bayesian network

A Bayesian network is a directed probabilistic graph model, presenting a joint distribution of variables in this network. The first part of a Bayesian network is similar to an undirected graph, which contains $G = \{V, E\}$, except E in a Bayesian network has arrows, the directional information. In practice, such information might indicate causality, but not always. The second part is local distribution for each variable, given its parents in the first part. The reason why local distributions are sufficient to describe the joint one is brought by the local Markov property: each variable is conditionally independent of its non-descendants given its parent variables. Therefore, by chain rule

$$f(X_1,...,X_n) = \prod_{i=1}^{n} f(X_i \mid Pa(X_i))$$

An example is demonstrated in Figure 4.8. In order to fully represent information on this Bayesian network, five local probability functions need to be specified: P(A), P(B|E, A), P(C|B), P(D|A) and P(E). Generally speaking, there are three scenarios to

represent local distributions in Bayesian network, depending on the variable types in it.

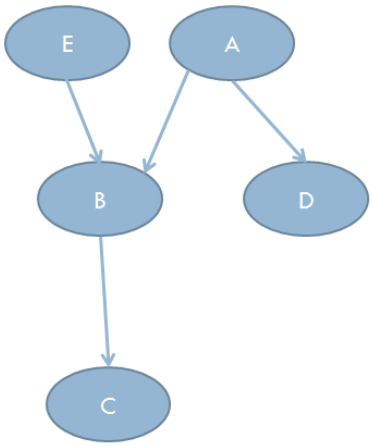


Figure 4.8 An example on a Bayesian network can be represented with local distributions. The graph itself implies conditional independences. By chain rule P(*A*, *B*, *C*, *D*, *E*) = P(*A*)P(*B*, *C*, *D*, *E*|*A*) = P(*A*)P(*D*|*A*)P(*B*, *C*, *E*|*A*) = P(*A*)P(*D*|*A*)P(*B*, *C*|*E*, *A*)P(*E*|*A*) = P(*A*)P(*D*|*A*)P(*E*)P(*C*|*B*, *E*, *A*)P(*B*|*E*, *A*) = P(*A*)P(*B*|*E*, *A*)P(*C*|*B*)P(*D*|*A*)P(*E*).

When all variables in the network are discrete (case I), denoting $\{U_1,\ldots,U_k\}$ as parents for a certain variable X, $P(X|U_1,\ldots,U_k)$ can be represented as a categorical distribution after specifying each assignment to $U_1,\ldots,U_k$, in the form of tables (Table 4.1) or the logistic regression model. When all variables are continuous (case II), people commonly assume Gaussian conditional densities:

$$P(X \mid u_1,...,u_k) \sim N(a_0 + \sum_i a_i u_i, \sigma^2)$$

It has already been shown if all variables in Bayesian network follow the distribution above, the joint distribution is a multivariate Gaussian. For a more general case with both types of variables, studies are restricted to the scenario where there are both discrete and continuous parents to a continuous descendent (case III). Hence under different value assignments to discrete parents, the descendent variable follows a Gaussian distribution conditional on the other continuous parent, similar to case II. One major reason continuous parents to discrete variable are not allowed is to ensure exact computation methods (Friedman et al., 2000; Bottcher and Dethlefsen, 2003). Nevertheless, if such a situation where continuous parents point to a discrete variable

is encountered in real data analyses, some researchers compromise by simply treating a discrete descendent like a continuous one (Bottcher and Dethlefsen, 2003).

$P(X=x_1\|U_1,\ldots U_k),\ldots,P(X=x_n\|U_1,\ldots U_k)$	U_1	…	U_k
$\theta_1, \ldots, \theta_n, \theta_1+\ldots+\theta_n = 1$	u_1	…	u_k
…	…	…	…
$\theta_{1'}, \ldots, \theta_{n'}, \theta_{1'}+\ldots+\theta_{n'} = 1$	$u_{1'}$	…	$u_{k'}$

Table 4.1 Specification of local distribution in a categorical Bayesian network. $\theta_1, \ldots, \theta_n$ are parameters required to describe local probability mass function of X.

4.2.3 Mixed Network with Partial Canonical Correlation Measure

There are three possible scenarios in a mixed case: between categorical variables, between continuous variables and between categorical and continuous variables. The former two can be solved with our partial canonical correlation measure and conventional partial correlation measure, respectively. In the last scenario, continuing to use SNP as an example, where there are three categories:

$$\begin{cases} \text{logit}\left(\frac{\pi_{x1}}{\pi_{x0}}\right) = Z\beta_{x1} \\ \text{logit}\left(\frac{\pi_{x2}}{\pi_{x0}}\right) = Z\beta_{x2} \end{cases}$$

$$\Rightarrow \begin{cases} \hat{r}_{x_{i1}} = \frac{x_i - \hat{\pi}_{i1}}{\sqrt{\hat{\pi}_{i1}(1-\hat{\pi}_{i1})}}, x_i = 1 \text{ when } x_i \text{ belongs to category 1; } x_i = 0 \text{ otherwise} \\ \hat{r}_{x_{i2}} = \frac{x_i - \hat{\pi}_{i2}}{\sqrt{\hat{\pi}_{i2}(1-\hat{\pi}_{i2})}}, x_i = 2 \text{ when } x_i \text{ belongs to category 2; } x_i = 0 \text{ otherwise} \end{cases}$$

$$Y = Z\beta_y + \varepsilon_y$$

$$\Rightarrow \hat{\varepsilon}_y = Y - Z\hat{\beta}_y$$

Consequently, the partial correlation between X and Y can be defined as the first canonical correlation between $\{r_{x1}, r_{x2}\}$ and ε_y. Furthermore, a test on the significance of the first canonical correlation will be equivalent to the corresponding ANOVA F-test on the regression model significance with ε_y on r_{x1} and r_{x2} (Knapp, 1978).

Two simulation scenarios are represented below to verify this partial correlation concept in a mixed network. Simulations are similar to what we conducted in pure categorical network, except we let Y be continuous (Figure 4.9). As a result, Y is simulated in a different manner, while the others are the same

$$Y = \beta_1 + X_1\beta_{X1} + X_2\beta_{X2} + Z_1\beta_{Z1} + Z_2\beta_{Z2} + \varepsilon$$

Set $[\beta_1, \beta_{X1}, \beta_{X2}, \beta_{Z1}, \beta_{Z2}]^T = [-0.5, 1, -0.2, 1, -0.2]^T$, $\varepsilon \sim N(0,1)$. It is expected that X and Y are correlated given Z while X is conditionally independent from Z. We still record the number of times true patterns are recognized through canonical correlation testing: 498/500 = 99.6% for X, $Y|Z$ and 366/500 = 73.2% for X, $Z|Y$.

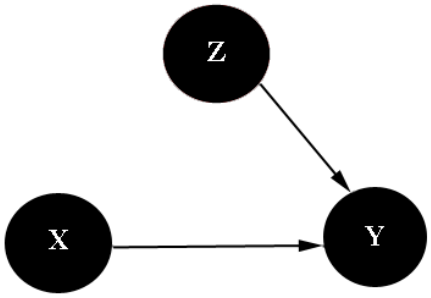


Figure 4.9 Simulation scenario I for mixed network. X and Z are categorical while Y is continuous.

Moreover, we also consider the following scenario (Figure 4.10),

$$X \sim \text{multinomial}(0.5, 0.3, 0.2)$$

$$\Rightarrow \begin{cases} \text{logit}\left(\frac{\pi_{z1}}{\pi_{z0}}\right) = \beta_1 + X_1\beta_{x11} + X_2\beta_{x12} \\ \text{logit}\left(\frac{\pi_{z2}}{\pi_{z0}}\right) = \beta_2 + X_1\beta_{x21} + X_2\beta_{x22} \end{cases}$$

$$Y = \beta_1 + Z_1\beta_{z1} + Z_2\beta_{z2} + \varepsilon$$

From X to Z, the parameters are the same as those previously used $[\beta_1, \beta_{X11}, \beta_{X12}, \beta_2, \beta_{X21}, \beta_{X22}]^T = [-2, 1, 0.2, -3, 1.2, 0.2]^T$; From Z to Y, we set $[\beta_1, \beta_{Z1}, \beta_{Z2}]^T = [-0.5, 1, -0.2]^T$, $\varepsilon \sim N(0,1)$. Examining the partial correlation between X and Y, the number of times our partial canonical test to successfully conclude a non-significant correlation

given *Z*: 457/500 = 91.4%

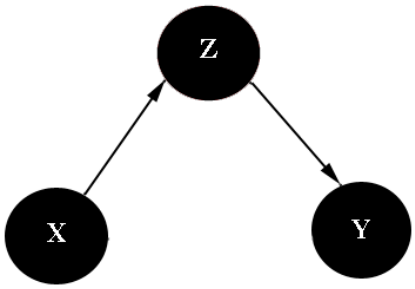


Figure 4.10 Simulation scenario II for mixed network. *X* and *Z* are categorical while *Y* is continuous.

Note that this approach will generate an undirected mixed network, without directional information as Bayesian network has. On the other hand, by excluding such information in the network, it does not have limitations caused by it. As explained earlier, in practice, mixed Bayesian network does not allow discrete variables to have continuous parents, which will not be an issue in PCNA with our partial canonical correlation measure. In addition, PCNA does not aim to find out the overall optimized network, but to focus on pairwise partial correlations. Hence, we might have a more intensive structure in comparison with Bayesian network, implying more information on pathway/pattern discovery. Summing them together, PCNA and Bayesian network require distinct interpretation.

Chapter 5 Network application to GWAS data

We used the same two datasets as we did in clustering analysis: COGEND - 2022 subjects (1114 cases and 908 controls) with 215 SNPs; Crohn's Disease - 491 patients having ileal disease location (considered as cases) and 131 patients having non-ileal disease location (considered as controls) with 29 SNPs. SNPs networks were then constructed with partial canonical correlation introduced earlier. For comparison purpose, we also established another two networks either by treating SNPs as single numeric variables (Pearson's partial correlation) or by using joint sparse logistic regression model (Wang et al., 2011) in an R package *LogitNet*.

Potential covariates in network were also examined when we applied these methods to Crohn's Disease data. If the covariate is categorical (phenotype), SNPs networks under different phenotypic groups were estimated and compared to find distinct relations; if the covariate is continuous (mRNA), a Bayesian mixed network was built up so as to identify causal pathways from SNP to downstream gene.

5.1 Collaborative Genetic Study of Nicotine Dependence (COGEND)

Edges (grey lines) in SNPs network were either defined as significantly non-zero partial (canonical) correlations through statistical testing or numerically non-zero values under certain threshold ($> 1e^{-6}$). For partial (canonical) correlations, since pairwise tests are conducted, multiple test correction must be implemented. In our study, false discovery rate is controlled at 0.05. Regarding "hub-SNPs" identification, we first plotted the numbers of edges from every SNP, producing a general picture to verify whether the underlying network possesses a hub-structure or not. If the hub-structure truly exists, it will be easy for us to set a threshold for selecting hub SNPs. Taking Figure 5.1 for an example, if two SNP variables pass the test on the fist canonical correlation at FDR = 0.05, they are connected with grey edge. SNPs with

notably large number of edges will be then considered as hubs. However, it appears most SNPs are intensively connected to each other. Plotting shows no evident gap between hypothesized hub-SNPs and the others.

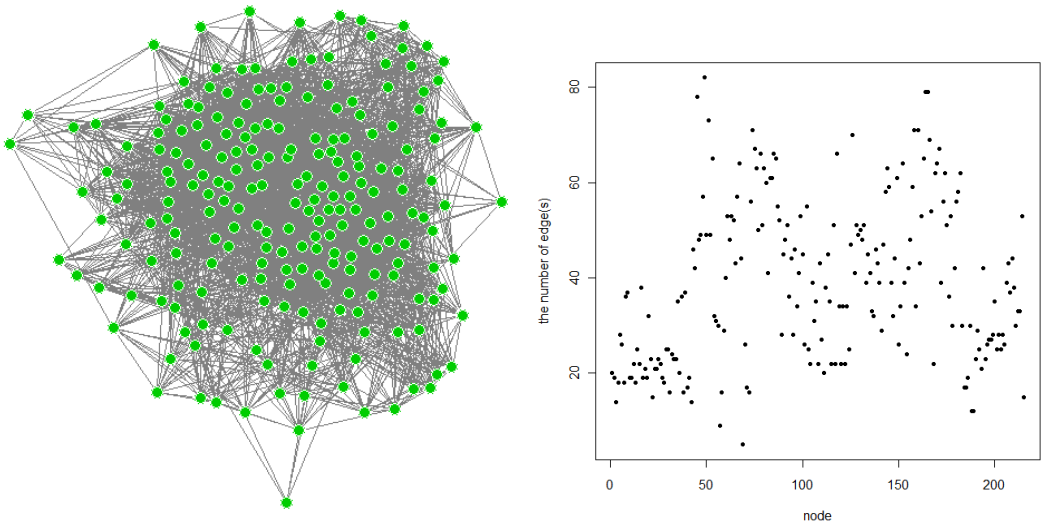


Figure 5.1 COGEND network analysis based on partial canonical correlation (FDR = 0.05). Edges are recognized by significance test on the first canonical correlation. According to edge numbers plotting, there is no convincing evidence to support a hub-SNP structure in network.

Network based on Pearson's *r* hardly exhibits a hub-structure (Figure 5.2) and there are fifteen outliers considerably apart from the main network. Furthermore, a main advantage of joint sparse model estimation is that we can control the overall network complexity with the l_1 penalty term, whereas this approach also generated an intensive network pattern with COGEND data (Figure 5.3). *rs2600685* will be recognized as the only reasonable hub, which is different from partial canonical correlation measure. Given the complexity of all three networks, it is hard for us to present a thorough evaluation regarding their performances. Hence we further combined clustering and network techniques to perform a sequential analysis: hierarchical clustering with all 215 SNPs was conducted first and then SNP representative was selected within each cluster. The concept is similar to that in LD studies – SNP that has overall highest similarities (canonical correlations) to all the other SNPs within each cluster will be chosen. Subsequently, 11 SNPs from 11

clusters and 4 SNPs from cluster 0 (outliers) (Table 3.1) were included in the following network analysis. Lastly, we compare networks based on partial canonical correlation measure and sparse joint logistic regression (Figure 5.4 & Figure 5.5). According to the final outputs, *rs1316971* is recognized by both methods while *rs1107953* and *rs2337980* are recognized by sparse joint logistic regression only. All the three SNPs belong to CHRN gene family – neuronal nicotinic receptor genes. Moreover, *rs1316971* is a marker of *CHRNB4*, and was reported to affect early alcohol and tobacco initiation in young adults (Schlaepfer et al., 2007). On the other hand, *rs2337980* is located in *CHRNA7*, a gene related to schizophrenia (Peng et al., 2008) and *rs1107953* has not associated with any neurological disease so far. Therefore, our partial canonical correlation measure successfully identified a "hub-SNP" that is truly associated with nicotine-addition.

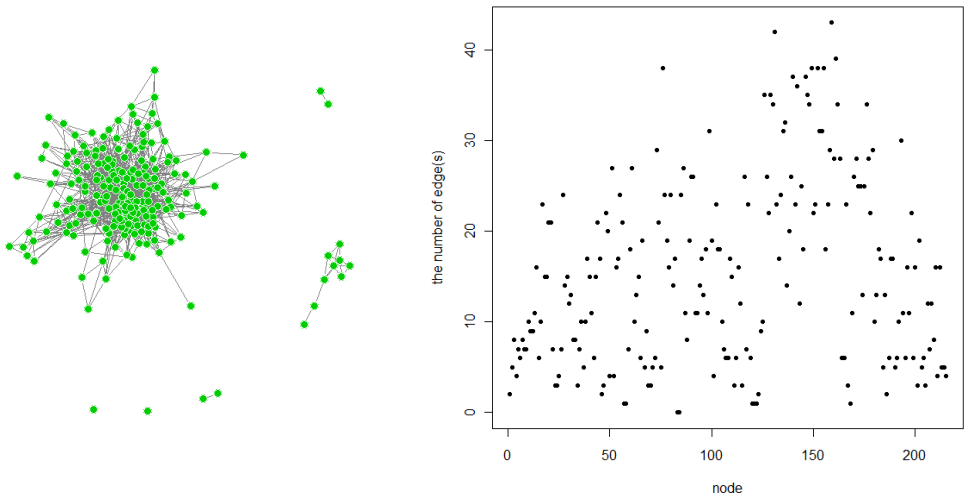


Figure 5.2 COGEND network analysis based on partial correlation measure (FDR = 0.05). Edges are recognized by significance test on Pearson's partial correlation. According to edge numbers plotting, there is no convincing evidence to support a hub-SNP structure in network, whereas fifteen SNPs are apart from the main framework.

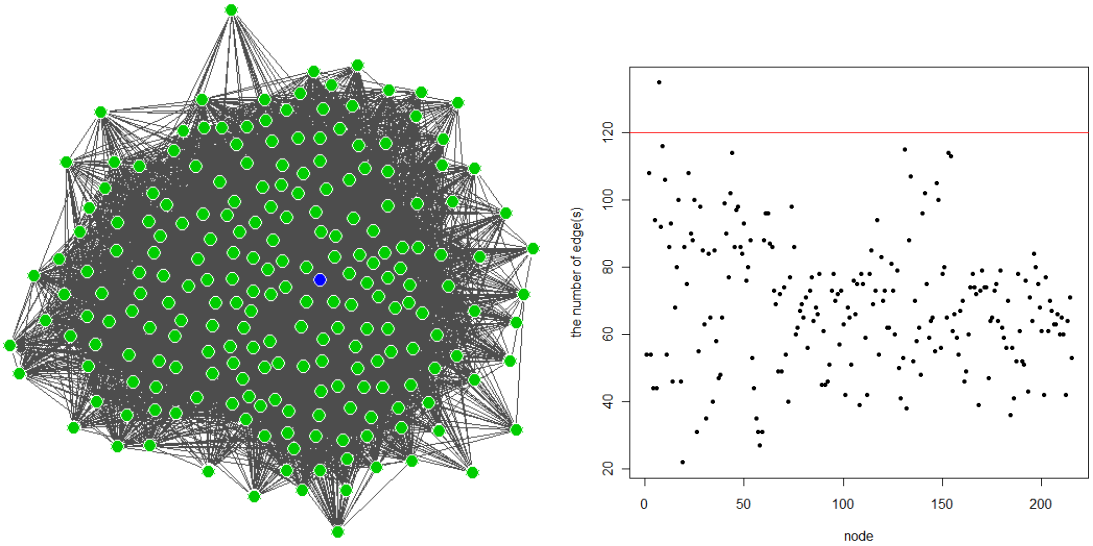


Figure 5.3 COGEND network analysis based on joint sparse logistic regression modeling. Edges are recognized by non-zero values (>1e^{-6}). According to edge numbers plotting, a threshold = 120 (red line) is set to distinguish hub-SNP (blue) from the others (green) in the left panel.

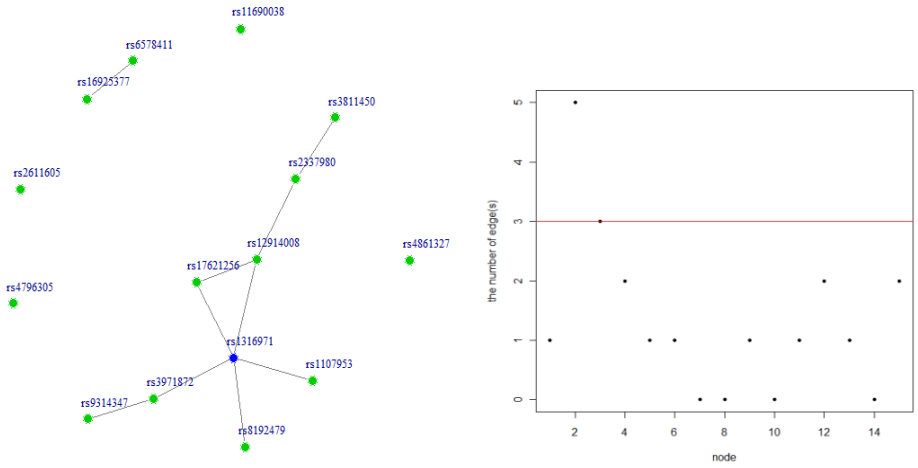


Figure 5.4 COGEND network analysis based on partial canonical correlation after clustering with canonical correlation measure (FDR = 0.05). Edges are recognized by significance test on the first canonical correlation. A threshold = 3 (red line) is set to distinguish hub-SNP (blue) from the others (green).

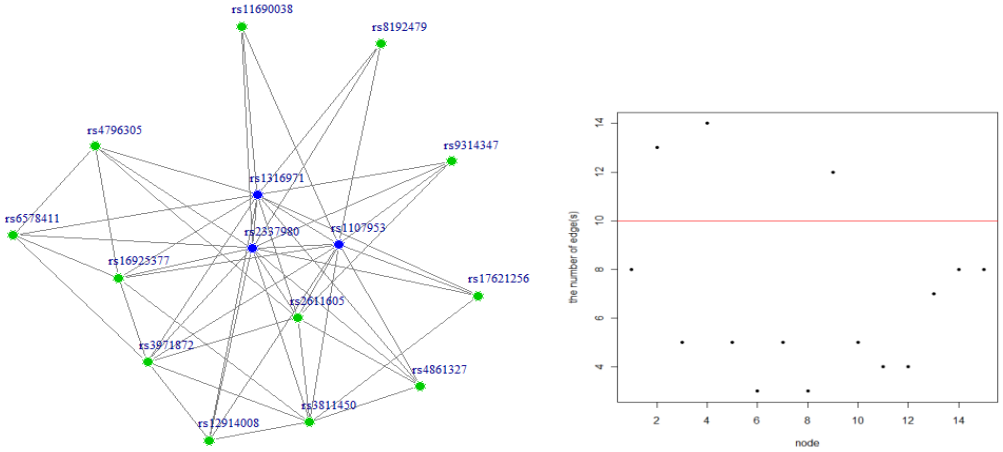


Figure 5.5 COGEND network analysis based on joint sparse logistic regression modeling after clustering with canonical correlation measure. Edges are recognized by non-zero values (>1e^{-6}). A threshold = 10 (red line) is set to distinguish hub-SNP (blue) from the others (green).

5.2 Crohn's Disease Study

5.2.1 Network Analysis with 29 SNPs

Figure 5.6 compiles network analysis outputs based on partial canonical correlation measure. Figure 5.7 and 5.8 show results based on partial correlation and sparse logistic regression respectively. It appears that network would still exhibit a fairly hub-structure with sparse logistic regression modeling while it failed to do so when we conducted partial (canonical) correlation tests. In sparse logistic regression modeling regarding hub-structure, two SNPs, *IL23R* and *STAT3* were selected as hub-SNPs.

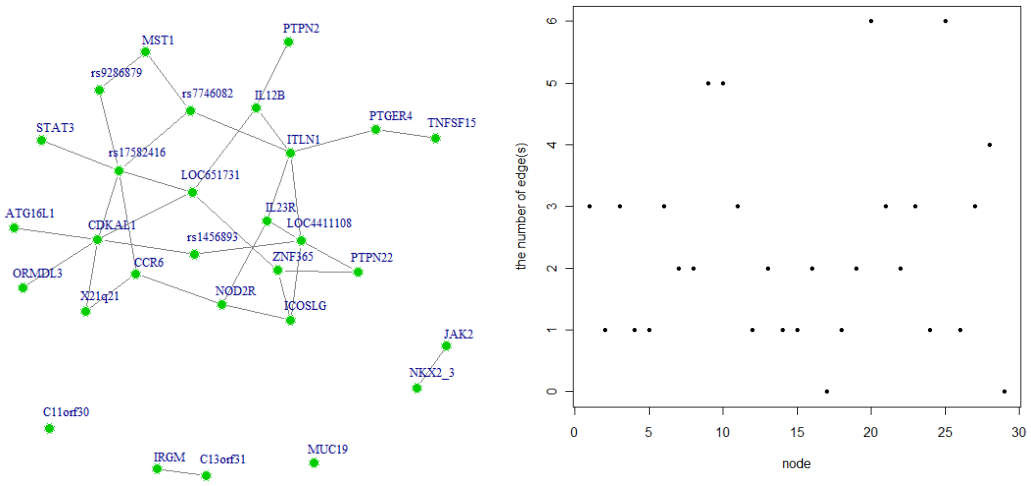


Figure 5.6 Network analysis of Crohn's disease study based on partial canonical correlation measure (FDR = 0.05). Edges are recognized by significance test on the first canonical correlation. There is no convincing evidence to support a hub-SNP structure in network.

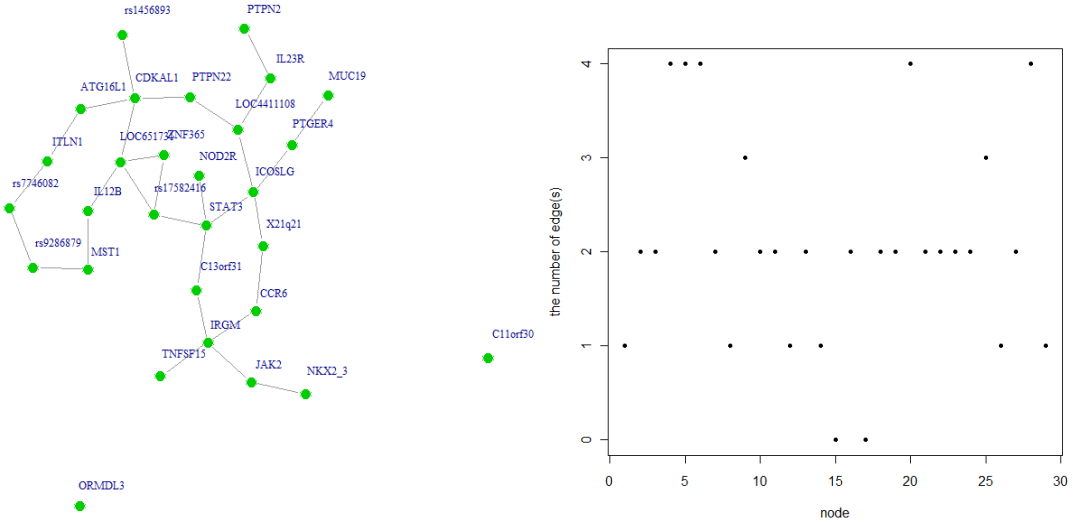


Figure 5.7 Network analysis of Crohn's disease study based on partial correlation measure (no multiple test correction). Edges are recognized by significance test on Pearson's partial correlation. According to edge numbers plotting, there is no convincing evidence to support a hub-SNP structure in network. After correction with FDR = 0.05, there is no significant partial correlations in network.

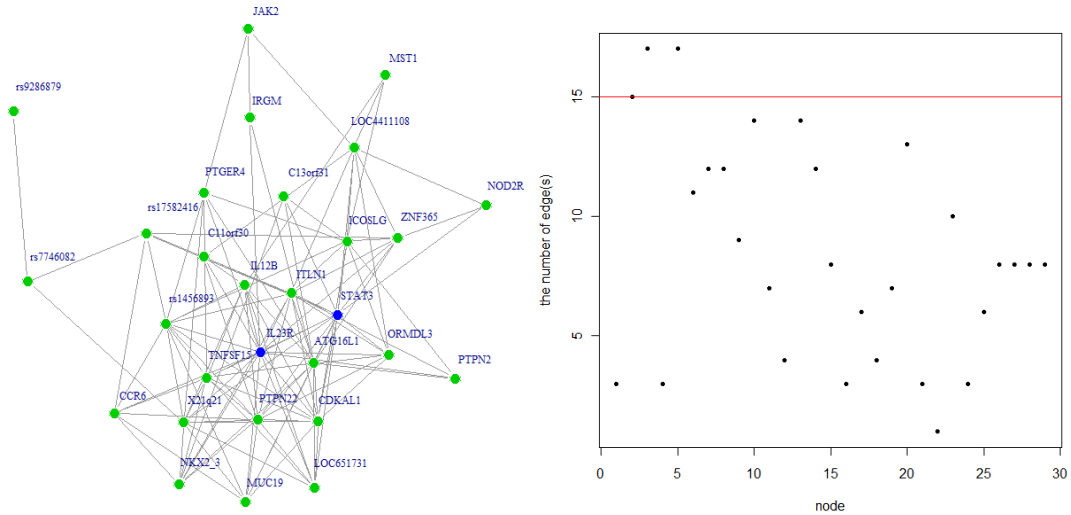


Figure 5.8 Network analysis of Crohn's disease based on sparse logistic regression modeling. Edges are recognized by non-zero values (>1e^{-6}). According to edge numbers plotting, a threshold = 15 (red line) is set to distinguish hub-SNPs (blue) from the others (green) in the left panel.

5.2.2 Phenotype as Covariate in Network

Similar to the "tag-SNPs", hub-SNPs can be viewed to include most of the information from all SNPs. Hence it seems that classification performance with those hub-SNPs should be performed, by including them as independent variables in any classification model. However, disease location information was out of our consideration when we constructed SNPs network. As a result, there is no natural connection from SNPs network to additional information on disease location. In fact, neither of *IL23R* and *STAT3* can be associated significantly with phenotype. It can be seen through univariate chi-square tests between individual SNP and phenotype (Chen et al., 2011). Besides, as demonstrated with variable selection, *NOD2* and *TNFSF15* were selected in the final model, whereas, both of them own relatively smaller number of edges in Figure 5.6 - 5.8. Such discrepancy implies that most of SNPs will have little effect size on disease and can be regarded as the majority "background", while the truly disease-associated SNPs would be far away from the majority and show weaker relations to these background SNPs. In order to connect SNPs network

to disease phenotype study, we might need other approaches, e.g., treating disease phenotype as a covariate in network analysis.

As explained in the last chapter, within case and control groups, we can construct SNPs network separately and then compare the underlying structural difference. For every possible edge in network (regardless of non-zero significance), a confidence interval may be obtained with bootstrapping method. If the corresponding confidence intervals from case and control groups do not overlap, we conclude that this partial relation between SNPs is different. Table 5.1 summarizes 44 different conditional relations with sparse logistic regression modeling. After incorporating phenotype as a covariate, we found previously identified hub-SNPs *IL23R* and *STAT3* did not account for the major difference between two networks, while *ATG16L1* possesses quite a few distinct conditional dependencies in case and control groups, further suggesting that hub-SNPs should not be assumed to have greater effect size on phenotype. Because results from partial canonical correlation measure are not readily interpretable, subsequently, a simple alternative is simply to focus on non-zero partial correlations (conditional dependencies) that are shown in one network but not the other (Table 5.2).

Different conditional relations		
ATG16L1-STAT3	ICOSLG-MST1	CCR6-JAK2
ATG16L1-ICOSLG	ICOSLG-JAK2	PTPN2-C11orf30
ATG16L1-ITLN1	21q21-rs1456893	PTPN22-TNFSF15
ATG16L1-TNFSF15	21q21-TNFSF15	PTPN22-IL12B
ATG16L1-rs7746082	21q21-LOC651731	PTPN22-NKX2_3
ATG16L1-IL12B	rs1456893-ITLN1	PTPN22-MUC19
ATG16L1-NKX2_3	rs1456893-PTPN22	TNFSF15-IL12B
ATG16L1-LOC651731	rs1456893-TNFSF15	TNFSF15-NKX2_3
IL23R-LOC651731	rs1456893-MUC19	TNFSF15-LOC651731
STAT3-LOC4411108	LOC4411108-ITLN1	C11orf30-IL12B

STAT3-ITLN1	ITLN1-PTPN22	C13orf31-NKX2_3
STAT3-TNFSF15	ITLN1-CDKAL1	CDKAL1-IL12B
STAT3-LOC651731	ITLN1-NKX2_3	IL12B-LOC651731
ICOSLG-LOC4411108	ITLN1-ZNF365	rs17582416-ZNF365
ICOSLG-ITLN1	CCR6-PTPN22	

Table 5.1 Different conditional relations of SNPs from L1 + L3 vs. L2, according to sparse logistic regression model.

Different conditional relations		
NOD2-CCR6	STAT3-PTPN22	rs1456893-rs17582416
ATG16L1-IRGM	STAT3-TNFSF15	LOC4411108-ITLN1
ATG16L1-STAT3	STAT3-ORMDL3	LOC4411108-ORMDL3
ATG16L1-ITLN1	STAT3-C13orf31	LOC4411108-C13orf31
ATG16L1-PTPN2	STAT3-CDKAL1	LOC4411108-JAK2
ATG16L1-PTPN22	STAT3-ZNF365	LOC4411108-ZNF365
ATG16L1-TNFSF15	STAT3-LOC651731	ITLN1-C11orf30
ATG16L1-PTGER4	ICOSLG-rs1456893	ITLN1-CDKAL1
ATG16L1-CDKAL1	ICOSLG-LOC4411108	ITLN1-rs17582416
ATG16L1-IL12B	ICOSLG-ITLN1	ITLN1-NKX2_3
ATG16L1-rs17582416	ICOSLG-PTPN22	ITLN1-ZNF365
ATG16L1-ZNF365	ICOSLG-ORMDL3	CCR6-TNFSF15
ATG16L1-MUC19	ICOSLG-PTGER4	CCR6-rs17582416
IL23R-STAT3	ICOSLG-CDKAL1	CCR6-NKX2_3
IL23R-21q21	ICOSLG-rs17582416	PTPN22-C11orf30
IL23R-rs1456893	ICOSLG-ZNF365	PTPN22-NKX2_3
IL23R-ITLN1	21q21-rs1456893	PTPN22-MUC19
IL23R-CCR6	21q21-PTPN22	TNFSF15-PTGER4
IL23R-PTPN2	21q21-TNFSF15	TNFSF15-NKX2_3
IL23R-TNFSF15	21q21-MST1	TNFSF15-LOC651731

IL23R-PTGER4	21q21-PTGER4	ORMDL3-CDKAL1
IL23R-CDKAL1	21q21-CDKAL1	MST1-rs17582416
IL23R-IL12B	21q21-IL12B	C11orf30-ZNF365
IL23R-NKX2_3	21q21-MUC19	CDKAL1-IL12B
IL23R-ZNF365	rs1456893-LOC4411108	CDKAL1-NKX2_3
IL23R-MUC19	rs1456893-ITLN1	CDKAL1-MUC19
STAT3-ICOSLG	rs1456893-CCR6	rs7746082-rs9286879
STAT3-21q21	rs1456893-PTPN22	IL12B-ZNF365
STAT3-LOC4411108	rs1456893-PTGER4	IL12B-LOC651731
STAT3-ITLN1	rs1456893-CDKAL1	rs17582416-ZNF365
STAT3-CCR6	rs1456893-IL12B	NKX2_3-ZNF365

Table 5.2 Different non-zero conditional relations of SNPs from L1 + L3 vs. L2, according to sparse logistic regression model.

5.2.3 Continuous Variables in Network

We also investigated relations between SNPs and continuous variables in a mixed network. To ensure an interpretable network analysis, we limited SNPs of interest to *NOD2* and *ATG16L1*, the two major SNPs associated with Crohn's Disease. Continuous variables are mRNA expression profiles, which were chosen by classification feature selection (CD vs. control) according to microarray data. We then compared the network outputs from Bayesian mixed network and PCNA with our partial canonical correlation measure. It is noteworthy that the dataset used here is slightly different: we have to ensure every observation for this study has both SNPs and microarray information, thus the sample size is much smaller (n = 98). Finally, the R package *deal* was used to search the optimal Bayesian network pattern, complying with maximum likelihood principle (Bottcher and Dethlefsen, 2003).

We first included the nine most important mRNAs selected by *boosting* in R package *CMA*. Figure 5.9 shows the result from Bayesian mixed network. Because of

the local Markov property of Bayesian network, arrows in the network may suggest casual relationships and therefore have biological interpretations. For instance, notice that *FOLH1* is the only parent variable of *BAX*, in other words, expression level of *BAX* is independent to all the other information in this network, given *FOLH1*'s expression. *FOLH1* encodes a folate hydrolase, a type II transmembrane glycoprotein belonging to the M28 peptidases family. It works as a glutamate carboxypeptidase on folate and other substrates. On the other hand, *BAX* encodes Bcl-2-associated X protein, which facilitate apoptosis. It has been found that low Bax expression in mucosal tissue may prevent T cells from activation and apoptosis, which is necessary for immune homeostasis. And subsequently, the decay of immune system triggers Crohn's disease formation (Itoh et al., 2001). Although there is no published research yet reporting *FOLH1* and *BAX* could act in the same physiological pathway, folate deficiency syndrome has been associated with decreased Bax expression in colon cancer cells (Novakovic et al., 2006). Regarding direct effects from SNP to mRNA, there are three conditional dependencies: *ATG16L1* to *PACSIN1*, *NOD2* to *SORD* and *NOD2* to *CYP26B1*. It is well recognized that *ATG16L1* plays a key role in the autophage pathway, affecting intestinal Paneth cells (Cadwell et al., 2008). Consistently, *PACSIN1* belongs to a kinase family that regulates macroautophage process, aiming to maintain cellular homeostasis (Szyniarowski et al., 2011); *NOD2* may be the most important Crohn's disease associated phenotype. The underlying *NOD2* protein mainly recognizes bacterial pepetidoglycans and regulates downstream immune reactions. Correspondingly, it has been claimed that T cell activation involves *SORD* and *CYP26B1* (Wang et al., 2008; Takeuchi et al., 2011). To sum them together, putting both genotype and mRNA in a mixed network provides us another opportunity to perform data-driven pathway analysis. It may bridge the gap between statistical modeling and biological bench work. The conditional relationships in network, or more specifically causalities, give potential directions in terms of finding out novel pathogenic pathways of interest.

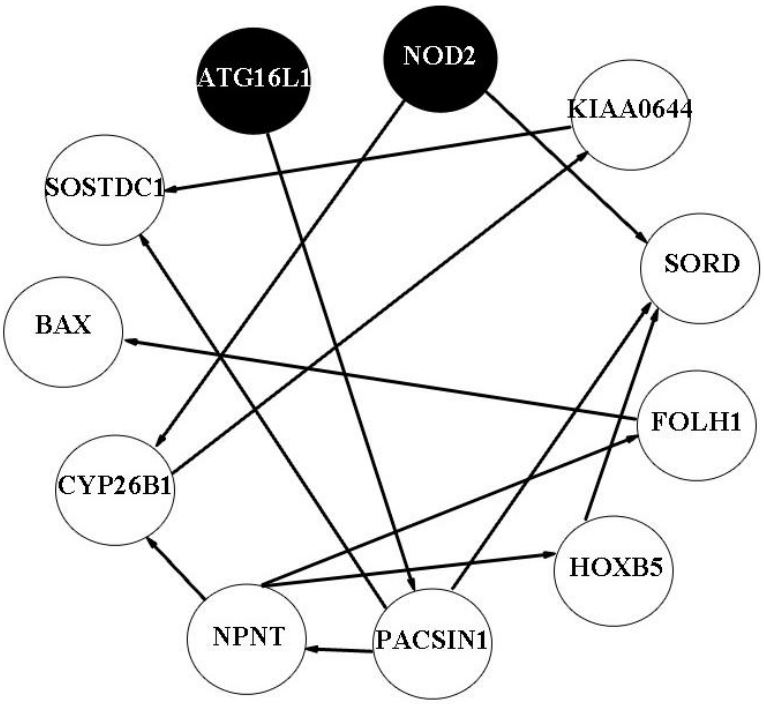


Figure 5.9 Bayesian mixed network with genotype and mRNAs. Black nodes indicate categorical SNPs - *NOD2* and *ATG16L1* while white nodes are continuous variables representing mRNA profile of nine Crohn's Disease related genes. Arrows distinguish parents and descendents.

We applied our partial canonical correlation measure to this mixed variables situation. It is worthy to point out again that such PCNA is an undirected graph modeling, which means it needs different interpretation on outputs. Figure 5.10 depicts a similar network structure. Regardless of direction/causal relationships, arrows in Figure 5.9 are partially overlapped with the edges in Figure 5.10. For instance, *KIAA0644* is connected with *SOSTDC1* and *CYP26B1* in a Bayesian network while it is associated with *SOSTDC1* and *FOLH1* in terms of partial canonical correlation measure. Nonetheless, the biological meaningful pathways in Figure 5.9 (*ATG16L1* to *PACSIN1*, *FOLH1* to *BAX*, etc.) are still kept in PCNA; it may provide some new relations that are worthy to explore in further studies. For example, *ATG16L1* and *BAX* are linked together in Figure 5.10, suggesting a possible crosstalk between autophage and apoptosis pathways, which can be trigger by common signals.

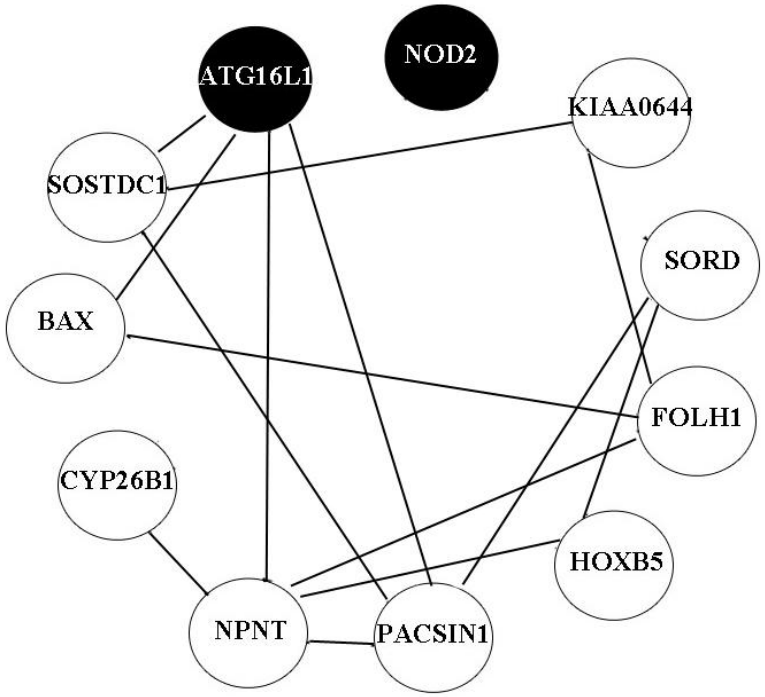


Figure 5.10 Mixed network analysis with genotype and mRNAs based on partial canonical correlation (FDR = 0.05). Note that edges have no direction information.

In practice, there are several other ways to perform feature selection among genes (probes) in microarray, including Prediction Analysis for Microarrays (PAM), RandomForest and least absolute shrinkage and selection operator (lasso). More importantly, each method would generate distinct feature list (results not shown). Therefore, we should conduct a batch of mixed network analyses with every mRNA list, so as to achieve a thorough investigation on relationship between SNPs and mRNAs. Alternatively, we can choose features (mRNAs) that are consistently picked across different platforms as input. According to work from other students in our group, there are six mRNAs – *FOLH1*, *KIAA0644*, *AK130891*, *NPC1L1*, *C4orf7*, and *DB340110* selected more than three times from the four methods mentioned above. Subsequently, these six mRNA expression profiles were included in a new round of mixed network analyses, along with *NOD2* and *ATG16L1*. Figure 5.11 – 5.12 are outputs based on Bayesian network and partial canonical correlation measure, respectively. It appears that no effect from *NOD2* or *ATG16L1* to mRNA expressions will be recognized in an optimized Bayesian network. Regardless, the identified

network structure in Figure 5.11 would be not unique in terms of class equivalence. The sub-structure with *AK130891*, *NPC1L1*, *C4orf7*, and *DB340110* can be replaced with alternative relationships (red lines). Hence causal directions within this structure cannot be confirmed unless there is additional biological evidence. On the other hand, partial canonical correlation measure generates a highly similar network, but also provides additional clues from genotypes to gene expressions - *ATG16L1* to *DB340110*. However, *DB340110* currently refers to an unexplored mRNA. There is no published study on its function so far. Hence, more evidence from future biological work would be necessary to support our finding in a mixed network according to partial canonical correlation measure.

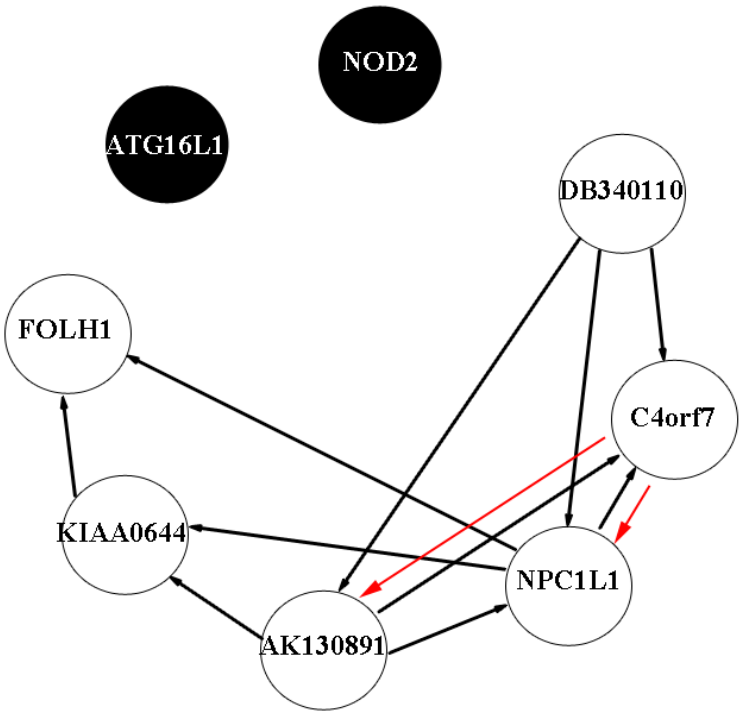


Figure 5.11 Bayesian mixed network with genotype and mRNAs selected by all the four methods. Black nodes indicate categorical SNPs - *NOD2* and *ATG16L1* while white nodes are continuous variables representing mRNA profiles of six Crohn's Disease related probes. Red lines indicate an equivalent alternative.

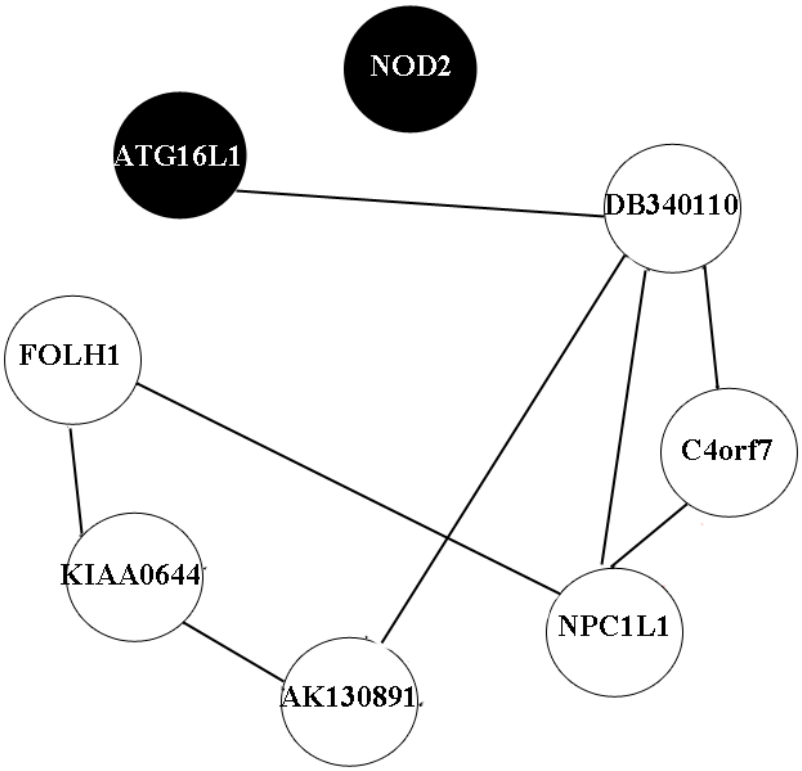


Figure 5.12 Mixed network analysis with genotype and mRNAs selected by all the four methods, based on partial canonical correlation (FDR = 0.05). Black nodes indicate categorical SNPs - *NOD2* and *ATG16L1* while white nodes are continuous variables representing mRNA profiles of six Crohn's Disease related probes. Note that *ATG16L1* is linked to *DB340110*.

Chapter 6 Discussion and Future Work

6.1 Canonical Correlation Measure in Hierarchical Clustering Analysis

Applications to both GWAS datasets support that our novel canonical correlation measures as an appropriate way to quantify pairwise association strength of SNPs. According to our clustering results, canonical correlation performs at least as well as the popular linkage disequilibrium measure *r*. And Cramér's *V*, another measure closely related to canonical correlation analysis, also works quite well. The other two measures included for comparison purpose, Pearson' *r* and Kendall's *τ*, did not achieve satisfying outputs. It implies that we must be cautious when treating SNPs as continuous or ordinal variables, even though the 0, 1 and 2 coding has biological meaning in terms of the risk allele number.

Besides the performance outcomes, canonical correlation measure may have additional advantages in practical sense. For instance, it would be feasible for us to define categorical cluster centers with this measurement. With polynomial coding {*S*, S^2} in our studies, we can determine cluster mean in the form of { $\bar{S}, \overline{S^2}$ }, assuming the two coded dummy variables are in a continuous manner. Consequently, this new definition could be helpful for cluster number determination with the ordinary R^2 computation or for obtaining representative SNPs after clustering. Figure 6.1 shows a comparison of R^2 plot from COGEND data between polynomial coding and single numeric coding (0, 1 and 2); the difference is immaterial. This is expected to some degree since polynomial coding eventually added one more calculation for sum of squares, which is from the coded variable S^2. The other dummy variable *S* includes equivalent information from single numeric coding. Given the fact that S^2 is highly correlated with *S*, results having information from both *S* and S^2 should be similar to results having information merely from *S*. Said in another way, it also explains why single numeric coding with SNP may also work well in certain GWAS articles

(Parkhomenko et al., 2009).

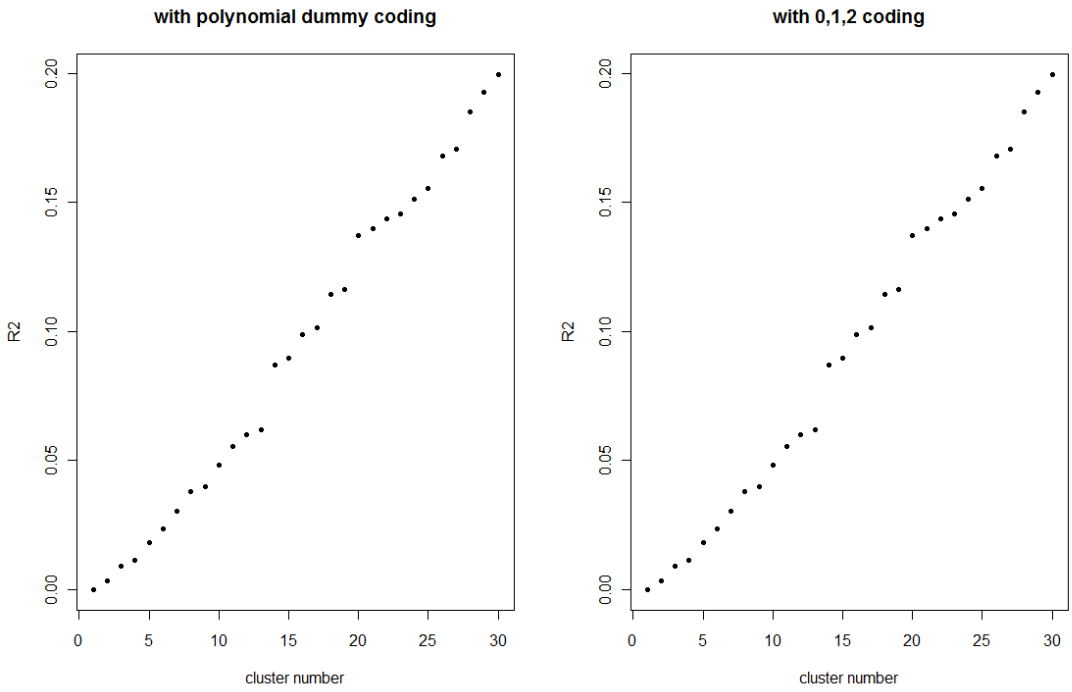


Figure 6.1 Clustering number determination with R^2. Clustering is based on COGEND dataset with 215 SNPs, sum of squares were calculated through either polynomial dummy coding (left) or single numeric coding – 0, 1 and 2 (right).

Because the ultimate goal in GWAS is to find out disease-related SNPs, clustering is merely a dimensional reduction approach to achieve this. With the linkage disequilibrium measure, tag-SNP concept is commonly applied – select a SNP representative that shows highest LDs to all the others in each cluster/region; with the canonical correlation measure, we may be able to use cluster means $\{ \bar{S}, \overline{S^2} \}$ directly. Either way, we then include these representatives in a logistic regression model to predict disease phenotype. Table 6.1 - 6.2 summarize their classification performances based on COGEND data. We discovered the canonical correlation measure has lower specificity (28.6%) than LD (36.9%) and slightly higher sensitivity (76.8%) than LD (76.3%). A more comprehensive comparison is also made across three measures: canonical correlation, linkage disequilibrium and Cramér's *V* with the conventional tag-SNP concept (Table 6.3). SNPs that have the highest (median) similarities to all

the other SNPs within each cluster are selected into the logistic regression model to examine the prediction power from selected SNPs. SNPs identified as outliers (cluster 0) were also included. As shown in Table 6.3, none of them achieves a satisfying overall performance. The major reason for such poor classification would be confounding effects, most likely coming from the population structure. Previous studies on nicotine-addiction have pointed out that SNPs may be in different association patterns with phenotype within African-Americans and European–Americans subgroups (Saccone et al., 2009). Unfortunately, we do not have such covariate information on our data. Regardless, there must be other confounding factors in our study. For instance, *rs16969968*, a SNP described earlier shows a medium effect size (odds ratio = 1.40) on a dataset without population ancestry discrimination, but has a much smaller effect size (odds ratio = 1.18) in ours. Nevertheless, instead of focusing on particular classification performance, our canonical correlation measure implies another general way to select the most relevant SNPs information from high-dimensional data.

	Truly non-addicted	Truly addicted	Total
Predicted as non-addicted	260	258	518
Predicted as addicted	648	856	1504
Total	908	1114	2022

Table 6.1 Classification performance with cluster means from canonical correlation measure. Leave-one-out cross validation was applied to estimate prediction accuracies and 0.5 was set to be the threshold since our dataset exhibits a fairly balanced design (1114 cases and 908 controls).

	Truly non-addicted	Truly addicted	Total
Predicted as non-addicted	335	264	252
Predicted as addicted	573	850	1770
Total	908	1114	2022

Table 6.2 Classification performance with tag-SNPs from LD *r*. Leave-one-out cross validation was applied to estimate prediction accuracies and 0.5 was set to be

the threshold since our dataset exhibits a fairly balanced design (1114 cases and 908 controls).

Measure	Specificity (%)	Sensitivity (%)	Overall accuracy (%)
Canonical correlation	35.4	74.2	56.8
Linkage disequilibrium	36.9	76.3	58.6
Cramér's *V*	36.2	73.8	56.9

Table 6.3 Classification performance with tag-SNP concept from three measures. Leave-one-out cross validation was applied to estimate prediction accuracies and 0.5 was set to be the threshold since our dataset exhibits a fairly balanced design (1114 cases and 908 controls).

Finally, there have been more risk loci identified from GWAS on Crohn's disease since our published study (Chen et al., 2011). After Barrett's paper in 2008, the number of SNPs associated with Crohn's disease is doubled (Khor et al., 2011). Besides, much more subjects are available now for statistical analysis. The latest GWAS study even includes >15,000 subjects in both disease and control groups (Rivas et al., 2011). Therefore, more thorough clustering analysis with increasing SNP candidates and sample size could be conducted in future, so as to further verify our stepwise variable selection findings.

6.2 Partial Canonical Correlation Measure in Network analysis

6.2.1 Categorical Network

Upon results from two GWAS datasets, our partial canonical correlation measure appears to be largely different from the joint sparse logistic modeling, in terms of the hub-SNPs recognition (COGEND) and the overall network structure (Crohn's

disease), while ordinary partial correlation measure would generate greater discrepancies. In particular, we did not obtain reasonably sparse (or hub) structure with either dataset. Further work on other real data or simulation will be necessary to thoroughly evaluate their performances. Moreover, we found there are certain practical problems in R package *LogitNet*. It is suppose to be able to handle multi-category variables, such like SNP with NR/NR, NR/R and R/R genotypes (Wang et al., 2011), but it is only applicable to binary data in practice. As a result, we had to combine R/R and NR/R together.

Another potential application of partial canonical correlation between Pearson residuals would be in clustering analysis. Since its value is bound by 0 and 1, it may be also suitable to be used as dissimilarity input. In fact, partial canonical correlation measure might be more relevant to fundamental GWAS research purpose. For instance, in our Crohn's Disease subphenotype dataset, we performed variable selection in a logistic regression model and successfully identified two associated SNPs - NOD2 and TNFSF15. Variable selection can be viewed as a detection of conditional effect from a variable given all the others that are already selected in the model. Therefore, we should expect that partial canonical correlation measure gives consistent clustering output. Most importantly, NOD2 and TNFSF15 are located in distance braches (Figure 6.2).

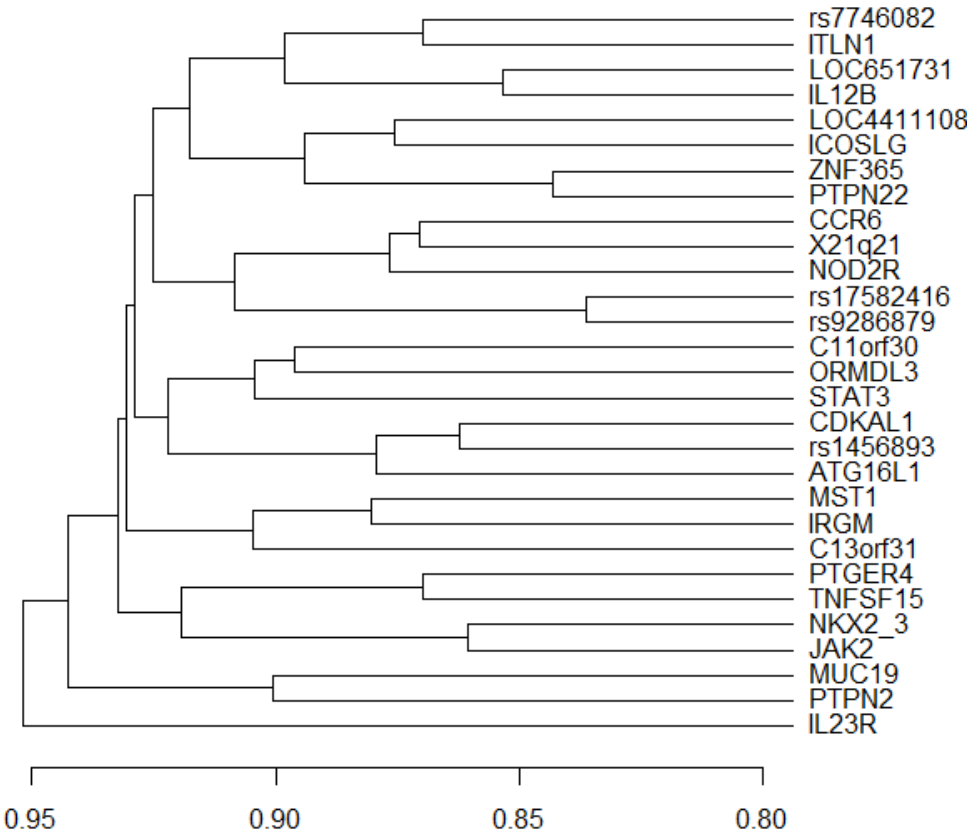


Figure 6.2 Clustering dendrogram of 29 SNPs with partial canonical correlation measure.

A future direction of network analysis is to incorporate genotype information when establishing a SNPs network. In reality, we pay more attention to disease-associated SNPs than hub-SNPs within an intra-SNP network. Applications to both COGEND and Crohn's disease data also reveal such a problem, considering the poor classification performance with hub-SNPs as prediction variables. Inclusion of genotype information may be achieved through network structure comparison between case and control groups (Table 5.1 and Table 5.2). Alternatively, there has been a biological solution for this problem: code SNPs in terms of disease-associated information other than three-category genotype. For example, loss of heterozygosity (LOH) is a commonplace phenomenon in cancer studies. As result, SNPs can be coded in binary form, where 1 indicates the occurrence of LOH after comparing cancer tissue with normal tissue (Wang et al., 2011). However, this solution may not suitable for a general scenario, because it is not always feasible for us to incorporate

disease information in this way. In addition, a recent paper discussed a joint allelic – linkage disequilibrium test with a specific coding scheme (Kim et al., 2010). Let *X*, *Y* denote any two SNPs: {*AA*, *Aa*, *aa*} and {*BB*, *Bb*, *bb*} and be coded as -1, 0 and 1 along with the order of genotypes listed in brackets. It can be shown that a logistic regression on disease phenotype relates the regression coefficients test to the corresponding joint allelic – LD test:

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_X X + \beta_Y Y + \beta_{XY} XY$$

$$(\beta_X, \beta_Y, \beta_{XY}) = 0 \Leftrightarrow \text{Joint Allelic - LD test}$$

It is essential to realize that interpretation from this test is completely different from sparse joint logistic regression or partial canonical correlation: non-zero edge from this test refers to a rejection on allelic – LD test: (β_X, β_Y, β_{XY}) $\neq$ 0, not conditional independency between *X* and *Y*.

Finally, we list the comparison between Markov random field and our partial canonical correlation measure in the following:

1) Both are able to detect conditional dependency. Edges identified in a network have the same interpretation.
2) Markov random field is conducted with numerical method while partial canonical correlation is based on testing (multiple test correction is required). Hence, threshold is set differently: non-zero values in MRF and a certain significance test level in partial canonical correlation.
3) Current numerical methods in Markov random field mainly deals with binary data while our method is designed for multi-category case.
4) According to our real data application, our method would not generate an intensive network structure, even without an explicit sparsity control in MRF. However, sparsity control implementation could be a future direction.

5) Our current method is not designed to the $n<<p$ scenario, which can be handled with Markov random field method. But this difficulty can be remedied through a combination with clustering analysis (COGEND) (Figure 5.4).

6.2.2 Mixed Network

In our further studies on the Crohn's disease data, mixed network was implemented for pathway discovery purpose. The existing Bayesian network and the extension of our partial canonical correlation measure were applied to find out relationships between SNPs and mRNAs. According to our results, they would provide similar outputs. Pros and cons of each method are summarized below:

1) Bayesian mixed network is able to distinguish parents variables between descendents, thus provides directional (causal) relationship. On the other hand, our partial canonical correlation measure aims to make ordinary PCNA applicable to mixed data, generating an undirected graph. Although it might also be possible to have causality information from a simple partial correlations structure, such as partial correlations within three or four continues variables, the interpretations on edges from the two methods are different.
2) Current Bayesian mixed network analysis (*deal*) does not allow continuous variables pointing to discrete descendents, which may become a practical issue when discrete variables other than SNPs are included. Unlike genotypes that are expected to be parents of mRNA expression levels, other covariates such as smoking status and disease phenotype would not have such clear prior information. In other words, we should not exclude potential effect from continuous mRNAs to these discrete covariates. Partial canonical correlation measure does not require such restriction in analysis. It is suitable for a more general mixed scenario.
3) Equivalent class is an essential concept in Bayesian network, which implies that the optimized network structure is not unique – there are several equivalent ways to represent the entire network with different local distributions. Programming is

not able to distinguish them since they all possess the same score. Under this condition, casual pathways cannot be verified statistically (Figure 5.9).

4) For a large number of controlled categorical variables, multicollinearity could exist in regression/logistic regression with the partial canonical correlation measure, causing the obtained Pearson's residuals to be unreliable. However, this practical issue can be solved by principle component analysis on the controlled categorical variables.
5) With the Crohn's disease data, we also tried including all the 17 mRNAs from the *boosting* feature selection on microarray data, but then the entire network will contain excessive edges for both methods (results not shown). We only included 2 major SNPs – *NOD2* and *ATG16L1*, given the fact that there are 29 SNP candidates. As explained earlier, such restriction on the variable number is mainly to ensure the overall network interpretability. Thus, sparsity control would be a future development of our method. In regards to the Bayesian network, a combination of focusing on mixed variables scenario (Friedman et al., 2000; Bottcher and Dethlefsen, 2003) and obtaining sparse solutions (Tipping, 2001) will be necessary for a more thorough investigation in the future.

References

Abraham, C. and Cho, J. (2009). Mechanisms of disease: inflammatory bowel disease. *New England Journal of Medicine* **361**, 2066 - 2078.

Agresti, A. (1983). Testing Marginal Homogeneity for Ordinal Categorical Variables. *Biometrics* **39**, 505-510.

Agresti, A. (2002). Categorical data analysis (2nd edition). New York: Wiley.

Ahmad, T., Armuzzi, A. and Bunce, M. (2002). The molecular classification of the clinical manifestations of Crohn's disease. *Gastroenterology* **122**, 854 - 866.

Anderson, T. W. (1958). An introduction to multivariate statistical analysis.

Barabasi, A. L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101 -115.

Barrett, J. C. , Hansoul, S. and Nicolae, D. L. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics* **40**, 955 - 962.

Becker, T. and Knapp, M. (2004). Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genetic Epidemiology* **27**, 21 – 32.

Besag. J. (1975). Statistical Analysis of Non-Lattice Data. *The Statistician* **24**, 179-195.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics* **36**, 199 - 227.

Bottcher, S. G. and Dethlefsen, C. (2003). deal: a package for learning Bayesian networks. *Journal of Statistical Software* **8**, 1 – 40.

Brynedal, B., Duvefelt, K., Jonasdottir, G., Roos, I. M., and Åkesson, E. (2007). HLA-A Confers an HLA-DRB1 Independent Influence on the Risk of Multiple Sclerosis. *PLoS One* **2**, e664.

Brookes, A. J. (1999). The essence of SNPs. *Gene* **234**, 177 – 186.

Cadwell, K., Liu, J. Y., Brown, S. L., Miyoshi, H., Loh, J., K.Lennerz, J., Kishi, C., Kc, W., Carrero, J. A., Hunt, S., Stone, C. D., Brunt, E. M., Xavier, R. J., Sleckman, B.

P., Li, E., Mizushima, N., Stappenbeck, T. S., and Virgin IV H. W. (2008). *Nature* **456**, 259 – 263.

Cantor, R. M., Kenneth, L., and Sinsheimer, J. S. (2010). Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *American Journal of Human Genetics* **86**, 6 – 22.

Cheeseman, P. and Stutz, J. (1995). Bayesian Classification (AUTOCLASS): Theory and Results. *Advances in Knowledge Discovery and Data Mining*, eds. AAI Press, 153–180.

Chen, H., Lee, A., Bowcock, A. Zhu, W., Li, E., Ciorba, M., and Hunt, S. (2011). Influence of Crohn's Disease Risk Alleles and Smoking on Disease Location. *Diseases of the Colon & Rectum* **54**, 1020 - 1025.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37–46.

Cohen, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**, 213–220.

Cohen, J. and Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences.

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.

Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10**, 417-451.

Corvin, A., Craddock, N., and Sullivan, P. F. (2010). Genome-wide association studies: a primer. *Psychological Medicine* **40**, 1063 – 1077.

Cramér, H. (1946). Mathematical Methods of Statistics.

Cuthbert, A., Fisher, S., and Croucher, P. J. (2002). The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease. *Gastroenterology* **122**, 867 - 874.

D'asermont, A., Banerjee, O., and Ghaoui, L. (2008). First-order Methods for Sparse Covariance selection. *Journal on* Matrix *Analysis and Applications* **30**, 56 – 66.

Darlington, R. B., Weinberg, S. L., Walberg, H. J. (1973). Canonical variate analysis

and related techniques. *Review of Educational Research* **43**, 433 – 454.

Deutsch, S., Iseli, C., Bucher, P., Antonarakis, S. E., and Scott, H. S. (2001). A cSNP map and database for human chromosome 21. *Genome Research* **11**, 300 -3 07.

Ehringer, M. A., McQueen, M. B., Hoft, N. R., Saccone, N. L., Stitzel, J. A., Wang, J. C., and Bierut, L. J. (2009). Association of CHRN genes with "dizziness" to tobacco. *American Journal of Medical Genetics* Part B **153**, 600 – 609.

Fisher, R.A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* **85**, 87–94.

Fisher, R.A. (1924). The distribution of the partial correlation coefficient. *Metron* **3**, 329 - 332.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378–382.

Fowler, E. V., Doecke, J., Simms, L. A. (2008). ATG16L1 T300A shows strong associations with disease subgroups in a large Australian IBD population: further support for significant disease heterogeneity. *American Journal of Gastroenterology* **103**, 2519 - 2526.

Freidman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayseian Network to Analyze Expression Data. *Journal of Computational Biology* **7**, 601 - 620.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *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., Hofling, H., and Tibshirani, R. (2010). Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33.

Gatz, M., Pedersen, N. L., Berg, S., Johansson, B., Johansson, K., Mortimer, J. A., Posner, S. F., Viitanen, M., Winblad, B., and Ahlbom, A. (1997). Heritability for Alzheimer's disease: the study of dementia in Swedish twins. *Journals of Gerontology* **52**, M117 – M125.

Goldstein, D. B. (2009). Common genetic variation and human traits. *New England Journal of Medicine* **360**, 1696 – 1698.

Goodman, L. A., Kruskal, W. H. (1954). Measures of association for cross

classifications. *Journal of the American Statistical Association* **49**, 732-764.

Goodman, L. A. and Kruskal, W. H. (1979). Measures of Association for Cross-classifications. New York: Springer-Verlag.

Goyette, P., Labbé, C. and Trinh, T. T. (2007). Molecular pathogenesis of inflammatory bowel disease: genotypes, phenotypes and personalized medicine. *Annals of Medicine* **39**, 177 - 99.

Greenberg, D. A. (1993). Linkage analysis of 'necessary' disease loci versus 'susceptibility' loci. *American Journal of Human Genetics* **52**, 135 – 143.

Guillaudeux, T., Janer, M., Wong, G. K., Spies, T., and Geraghty, D. E. (1998). The complete genomic sequence of 424015 bp at the centromeric end of the HLA class I region: gene content and polymorphism. *Proceedings of the National Academy of Sciences USA* **95**, 9494 – 9499.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Joint Structure Estimation for Categorical Markov Networks. To appear.

Herbert, A., Gerry, N. P., McQueen, M. B., Heid, I. M., Pfeufer, A. and LLLig, T. (2006). A common genetic variant is associated with adult and childhood obesity. *Science* **312**, 279 – 283.

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.

Hofling, H. and Tibshirani, R. (2009). Estimation of Sparse Binary Pairwise Markov Networks using Pseudo-likelihoods. *Journal of Machine Learning Research* **10**, 883 - 906.

Homster, D. W. and Lemeshow, S. (2000). Applied logistic regression. Wiley-Interscience.

Horton, R., Niblett, D., Milne, S., Palmer, S., Tubby, B., Trowsdale, J., and Beck, S. (1998). Large-scale sequence comparisons reveal unusually high levels of variations in the HAL-DQB1 locus in the class II region of the human MHC. *Journal of Molecular Biology* **282**, 71 – 97.

Huang, Z. X. (1997). Extensions to the *K*-Means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery* **2**, 283–304.

Isserlis, L. (1914). On the partial correlation ratio - Part I Theoretical. *Biometrika* **10**, 391 - 411.

Itoh, J. de la Motte, C., Strong, S., Levine, A. and Fiocchi, C. (2001). Decrease Bax expression by mucosal T cells favours resistance to apoptosis in Crohn's disease. *Gut* **49**, 35 – 41.

Kaufman, L. and Rousseeuw, P. J. (1990). Finding Groups in Data: An introduction to Cluster Analysis.

Kaufman, D. and Sweet, R. (1974). Contrast coding in least squares regression analysis. *American Educational Research Journal* **11**, 359-377.

Kendall, M. (1938). A New Measure of Rank Correlation. *Biometrika* **30**, 81–89.

Khor, B., Gardet, A., Xavier, R. J. (2011). Genetics and pathogenesis of inflammatory bowel disease. *Nature* **474**: 307 – 317.

Kim, S., Morris, N. J., Won, S. and Elston, R. C. (2010). Single-Marker and Two-Marker Association Tests for Unphased Case-Control Genotype Data, With a Power Comparison. *Genetic Epidemiology* **34**, 67 -77.

King T. S. and Chinchilli V. M. (2001). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine* **20**, 2131–2147.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., Sangiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385 – 389.

Knapp, T. R. (1978). Canonical correlation analysis: a general parametric significance-testing system. *Psychological Bulletin* **85**, 410 – 416.

Kolar, M. and Xing, E. (2008). Improved estimation of high-dimensional Ising models. Eprint arXiv:0811.1239.

Ku, C. S., Loy, E. Y., Pawitan, Y., and Chia, K. S. (2010). The pursuit of genome-wide association studies: where are we now? *Journal of Human Genetics* **55**, 195 – 206.

Lai, E., Riley, J., Purvis, I., and Roses, A. (1998). A 4 Mb high-density single nucleotide polymorphism-based map around human APOE. *Genomics* **54**, 31 – 38.

Langfelder, P., Zhang, B. and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut library for R. *Bioinformatics* **24**, 719 – 720.

Lesage, S., Zouali, H., and Cezard, J. P. (2002). CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *American Journal of Human Genetics* **70**, 845 - 857.

Lewontin, R. C. and Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution* **14**, 458–472

Li, W. and Sadler, L. A. (1991). Low nucleotide diversity in man. *Genetics* **129**, 513 – 523.

Li, H. Z. and Gui, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics* **7**, 302 - 317.

Louis, E., Collard, A., and Oger, A. F. (2001). Behavior of Crohn's disease according to the Vienna classification: changing pattern over the course of the disease. *Gut* **49**, 777 - 882.

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Symposium at Mathematical Statistics and Probability*, 281–297

Manolio, T. A., Brooks, L. D., and Collins, F.S. (2008). A HapMap harvest of insights into the genetics of common disease. *Journal of Clinical Investigation* **118**, 1590 – 1605.

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine* **363**, 166 – 176.

Maraganore, D. M., de Andrade, M., Lesnick, T. G., Strain, K. J., Farrer, M. J., Rocca, W. A. (2005). High resolution whole-genome association study of Parkinson disease. *American Journal of Human Genetics* **77**, 685 – 693.

Márquez, A., Núñez, C. and Martínez, A., Role of ATG16L1 Thr300Ala polymorphism in inflammatory bowel disease: a study in the Spanish population and a meta-analysis. *Inflammatory Bowel Diseases* **15**, 1697 - 1704.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 153–157.

Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34**, 1436 - 1462.

Morrison, D. F. (1976). Multivariate statistical methods.

Muller, K. E. and Fetterman, B. A. (2002). Regression and ANOVA: an integrated approach using SAS software.

Nachman, M. W., Bauer, V. L., Crowell, S. L. and Aquadro, C. F. (1998). DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**, 1133 – 1141.

Nickerson, D. A., Taylor, S. L., Weiss, K. M., Clark, A. G., Hutchinson, R. G., Stengard, J., Salomaa, V., Vartiainen, E., Boerwinkle, E. and Sing, C. F. (1998). DNA sequence diversity in a 9.7kb region of the human lipoprotein lipase gene. *Nature Genetics* **19**, 233 – 240.

Novakovic, P., M., Stempak, J., Sohn, K-J., and Kim, Y-I. (2006). Effects of folate deficiency on gene expression in the apoptosis and cancer pathways in colon cancer cells. *Carcinogenesis* **27**, 916 – 924.

Ozaki, K., Ohnishi, Y., Lida, A., Sekine, A., Yamada, R., and Tsunoda, T. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics* **32**, 650 – 654.

Pearson, K. (1900). On the Criterion that a given System of Deviations from the probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from random Sampling. *Philosophical Magazine Series* **5**, 157-175.

Pearson, K. (1915). On the partial correlation ratio. *Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character* **91**, 492 - 498.

Pe'er. I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genome-wide association studies of nearly all common variants. *Genetic Epidemiology* **32**, 381 – 385.

Peng, J., Wang, P., Zhou, N. F., and Zhu, J. (2009). Partial Correlation Estimation by Joint Sparse Regression Models. *Journal of the American Statistical Association* **104**, 735 - 746.

Peng, Z., Wan, X., and Jiang, T. (2008). The transmission disequilibrium analysis between neuronal nicotinic acetylcholine receptor alpha 7 subunit gene polymorphisms and schizophrenia. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* **25**: 154 – 158.

Plackett, R.L. (1983). Karl Pearson and the chi-squared test. *International Statistical*

Review **51**, 59–72.

Pitchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* **69**, 1 – 14.

Prescott, N. J., Fisher, S. A.. and Franke, A. (2007). A nonsynonymous SNP in ATG16L1 predisposes to ileal Crohn's disease and is independent of CARD15 and IBD5. *Gastroenterology* **132**, 1665 - 1671.

Ravikumar, P., Wainwright, M., Raskutti, G. and Yu. B. (2010). High-dimensional Ising model selection using l1-regularized logistic regression. *Annals of Statistics* **38**, 1287 - 1319.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516 – 1517.

Rivas, M. A. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics* **43**: 1066 – 1073.

Saccone, N. L., Wang, J. C., Breslau, N., Johnson, E. O., Hatsukami, D., Saccone, S. F., Grucza, R. A., Sun, L., Duan, W. and Budde, J. (2009). The CHRNA5-CHRNA3-CHRNB4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans. *Cancer Research* **69**, 6848 – 6856.

Saccone, N. L., Schwantes-An, T.-H., Wang, J. C., Grucza, R. A., Breslau, N., Hatsukami, D., Johnson, E. O., Rice, J. P., Goate, A. M. and Bierut, L. J. (2010). Multiple cholinergic nicotinic receptor genes affect nicotine dependence risk in African and European Americans. *Genes, Brain and Behavior* **9**, 741 – 750.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., and Mullikin, J. C. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–33.

Satsangi, J., Silverberg, M. S. and Vermeire, S. (2006). The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. *Gut* **55**, 749 - 753.

Schafer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21**, 754 - 764.

Schlaepfer, I. R., Hoft, N. R., Collins, A. C., Corley, R. P., Hewitt, J. K., Hopfer, C. J., Lessem, J. M., McQueens, M. B., Rhee, S. H., and Ehringer, M. A. (2007). The

CHRNA5/A3/B4 gene cluster variability as an important determinant of early alcohol and tobacco initiation in young adults. *Biological Psychiatry* **63**: 1039 – 1046.

Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* **17**, 321-325.

Sczyniarowski, P., Corcelle-Termeau, E., Farkas, T., Hoyer-Hansen, M., Nylandsted, J., Kallunki, T., and Jaattela, M. (2011). A comprehensive siRNA screen for kinases that suppresee macroqutophage in optimal growth conditions. *Autophagy* **7**, 892 – 903.

Seber, G. A. F. and Nyangoma, S. O. (2000). Residuals for multinomial models. *Biometrika* **87**, 183 – 191.

Serlin, R. C. and Levin, J. R. (1985). Teaching how to derive directly interpretable coding schemes for multiple regression analysis. *Journal of Educational Genetics* **10**, 223-238.

Sham, R. C., Cherny, S. S., Purcell, S., and Hewitt, J. K. (20000. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *American Journal of Human Genetics* **66**, 1616 – 1630

Spielman, R. S., McGinnis, R. E., and Ewens, W.J., 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**, 506–16.

Stevenson, J. (1992). Evidence for a genetic etiology in hyperactivity in children. *Behavior Genetics* **22**, 337 – 344.

Takeuchi, H., Yokota, A., Ohoka, Y., and Iwata, M. (2011). Cyp26b1 regulates retinoic acid-dependent signals in T cells and its expression is inhibited by transforming growth factor-β. *PLoS ONE* **6**, e16089.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **1**, 211 – 244.

Unkart, J. T., Anderson, L., and Li, E., 2008. Risk factors for surgical recurrence after ileocolic resection of Crohn's disease. *Disease of the Colon & Rectum* **51**, 1211 – 1216.

Vial, W. C. (1986). Cigarette smoking and lung disease. *American Journal of the Medical Sciences* **291**, 130 – 142.

Van Limbergen, J., Russell, R. K., and Nimmo, E. R. (2008). Autophagy gene

ATG16L1 influences susceptibility and disease location but not childhood-onset in Crohn's disease in Northern Europe. *Inflammatory Bowel Diseases* **14**, 338 - 346.

Wang, D. G. , Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M., and Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077 – 1082.

Wang, M., Windgassen, D., and Papoutsakis, E. T. (2008). Comparative analysis of transcriptional profiling of CD3+, CD4+ and CD8+ T cells identifies novel immune response players in T-cell activation. *BMC Genomics* 9:225.

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

Wang, W. Y. S., Barratt, B. J., Clayton, D. G., and Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics* **6**, 109 – 118.

Weedon, M. N., Lango, H., and Lindgren, C. M. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics* **40**, 575 – 583.

Wendorf, C. A. (2004). Primer on multiple regression coding: Common forms and the additional case of repeated contrasts. *Understanding Statistics* **3**, 47-57

Xavier, R. J. and Podolsky, D. K. (2007). Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448**, 427 – 434.

Xing, C., Cohen, J. C. and Boerwinkle, E. (2010). A weighted false discovery rate control procedure reveals alleles at FOXA2 that influence fasting glucose levels. *American Journal of Human Genetics* **86**, 440 – 446.

Yamazaki, K., McGovern, D., Ragoussis, J., Paolucci, M., Butler, H., and Jewell, D. (2005). Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn’s disease. *Human Molecular Genetics* **14**, 3499 – 3506.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19 - 35.

Zhang, P., Wang, X., and Song, P. X. K. (2006). Clustering Categorical Data Based on Distance Vectors. *Journal of the American Statistical Association* **101**, 355-367.