5-2016

Integrated analysis of miRNA/mRNA expression and gene methylation using sparse canonical correlation analysis.

Dake Yang
University of Louisville

INTEGRATED ANALYSIS OF MIRNA/MRNA EXPRESSION AND GENE METHYLATION USING SPARSE CANONICAL CORRELATION ANALYSIS

By

Dake Yang

B.A., BIT, 2008

M.A., University of Louisville, 2011

A Dissertation Submitted to the Faculty of the

School of Public Health & Information Sciences of the University of Louisville

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

In Biostatistics: Decision Science

Department of Bioinformatics and Biostatistics

University of Louisville

Louisville, Kentucky

May 2016

INTEGRATED ANALYSIS OF MIRNA/MRNA EXPRESSION AND GENE METHYLATION USING SPARSE CANONICAL CORRELATION ANALYSIS

By

Dake Yang

B.A., BIT, 2008

M.A., University of Louisville, 2011

April 1st, 2016

by the following Dissertation Committee:

_____
Dissertation Director: Dr. Guy Brock

_____
Dissertation Co-Director: Dr. Douglas Lorenz

_____
Second Committee Member: Dr. Maiying Kong

_____
Third Committee Member: Dr. KB Kulasekera

_____
Fourth Committee Member: Dr. Partha Mukhopadhyay

_____
Fifth Committee Member: Dr. Dongfeng Wu

DEDICATION

This dissertation is dedicated to my loving parents

Mr. Xiaoli Yang

and

Mrs. Qiong Hu

who have given me invaluable educational opportunities, and their words of encouragement give me motivation of moving ahead

ACKNOWLEDGEMENTS

I wish to thank my committee members who were more than generous with their expertise and precious time. A special thanks to Dr. Guy Brock, my committee chairman for his countless hours of reflecting, reading, encouraging, and most of all patience throughout the entire process. Thank you Dr. Douglas Lorenz, Dr. Maiying Kong, Dr. KB Kulasekera, Dr. Partha Mukhopadhyay and Dr. Dongfeng Wu for agreeing to serve on my committee.

I would like to acknowledge and thank my school division for allowing me to conduct my research and providing any assistance requested. Special thanks goes to the members of staff development and human resources department for their continued support.

Finally I would like to thank the beginning teachers, mentor-teachers and administrators in our school division that assisted me with this project. Their excitement and willingness to provide feedback made the completion of this research an enjoyable experience.

ABSTRACT

INTEGRATED ANALYSIS OF MIRNA/MRNA EXPRESSION AND GENE METHYLATION USING SPARSE CANONICAL CORRELATION ANALYSIS

Dake Yang

April 1st, 2016

MicroRNAs (miRNAs) are a large number of small endogenous non-coding RNA molecules (18-25 nucleotides in length) which regulate expression of genes post-transcriptionally. While a variety of algorithms exist for determining the targets of miRNAs, they are generally based on sequence information and frequently produce lists consisting of thousands of genes. Canonical correlation analysis (CCA) is a multivariate statistical method that can be used to find linear relationships between two data sets, and here we apply CCA to find the linear combination of differentially expressed miRNAs and their corresponding target genes having maximal negative correlation. Due to the high dimensionality, sparse CCA is used to constrain the problem and obtain a solution. A novel gene set enrichment analysis statistic is proposed based on the sparse CCA results for estimating the significance of predefined gene sets. The methods are illustrated with both a simulation study and real miRNA-mRNA expression data.

DNA methylation is a process of adding a methyl group to DNA by a group of enzymes collectively known as DNA methyltransferases which is an epigenetic modification critical to normal genome regulation and development. In order to understand the role of DNA methylation in gene differentiation, we analyze genome-scale DNA methylation patterns and gene expression data using sparse CCA to find linear combinations between the two data sets which have maximal negative correlation. In a similar spirit to the miRNA-mRNA study, we create a GSEA statistic with weight vectors from the sparse CCA method and assess the significance of predefined gene sets. The method is exemplified with real gene expression / DNA methylation data regarding the development of the embryonic murine palate.

TABLE OF CONTENTS

PAGE

CHAPTER I

BACKGROUND

MicroRNAs (miRNAs) are a large number of small endogenous non-coding RNA molecules (18-25 nucleotides in length) processed from 70–100 nucleotide hairpin pre-miRNAs. The miRNAs are transcribed by RNA polymerase II from independent genes or represent introns of messenger RNA transcripts. The miRNAs have been discovered and found to execute key functions in a ribonucleoprotein complex called RNA-induced silencing complex (RISC) and guide the RISC to the target mRNA in both plant and animal systems (Nelson and Weiss, 2008). The miRNAs bind to their target mRNA 5'UTR and can down regulate gene expression through directly inhibiting their translation and/or resulting in the destabilization of their target mRNAs at the posttranscriptional level. Currently, thousands of these small regulators have been identified in various species. It is believed that each miRNA potentially targets between 100 and 200 mRNAs, and miRNAs regulate between 20%- 30% of all human genes (Flynt and Lai, 2008; Nilsen, 2007). So, the potential relationships between miRNAs and mRNAs are extremely complex. Therefore, miRNAs play a major role in multiple essential biological processes including development, differentiation, apoptosis and cellular proliferation. There is also strong evidence that miRNAs are involved in pathological processes and contribute to the occurrence and development of some cancers. Specifically, abnormally expressed miRNAs have been shown to be crucial contributors and may serve as biomarkers in many human diseases, as found by comparing distinct miRNA expression for human cancers with their normal counterparts.

The development of microarray technology has equipped scientific researchers with the ability to simultaneously study, in a single experiment, the expression patterns of thousands of genes within the cells of a biological sample. This technology has been successfully extended to the arena of miRNAs to generate "microRNA gene expression profiles" of the cell cycle (Corney, et al., 2007), cell differentiation (Zhan, et al., 2007), cell death (Kren, et al., 2009), embryonic development, stem cell differentiation (Lakshmipathy, et al., 2007), different types of cancers (Gottardo, et al., 2007), the diseased heart (Tatsuguchi, et al., 2007)

and normal as well as diseased neural tissue (Ferretti, et al., 2009). The typical first step in determining the important miRNAs for regulation of gene expression is identifying differentially expressed miRNAs. That is, miRNAs that are differentially expressed between normal and diseased tissue types, or exhibit changes in expression over time. Then, these miRNAs are evaluated to determine which biochemical and molecular systems they target. Of critical importance is to discover how the miRNAs are targeting the biological pathways, i.e. what specific genes/transcripts (within those pathways) are being regulated by the differentially expressed miRNAs. Identifying the putative target transcripts based on sequence complementarity between the 3'-UTR of the mRNAs and the 'seed region' of the miRNA (nucleotides 2–7) is an important step. There are several databases which include lists of miRNA targets which are computationally predicted, including miRBase (Kozomara and Griffiths-Jones, 2014), miRanda (John, et al., 2004) and TargetScan (Lewis, et al., 2003). However, the issue is that computational methods predict hundreds to thousands of target mRNAs for each miRNA. Furthermore, the information concerning which of the potential miRNA targets are regulated during the biological process of interest is not included. To help solve this problem, we can identify the predicted target mRNAs which are inversely correlated with miRNA expression values as the potential mRNA-miRNA associations. The motivation behind this approach is that the main regulatory mechanism of miRNAs is to bind complementary regions within the 3'-untranslated region of mRNA transcripts which results in degradation of the mRNA target transcript. So, a more definitive determination of miRNA-mRNA interactions involves integrated analysis from both miRNA and mRNA expression values. These potential targets can then be further analyzed for enrichment in certain biological functions or pathways.

Regulation of gene expression by miRNA binding of mRNA transcripts can be considered one mode of epigenetic regulation. Another mode of epigenetic regulation is DNA methylation of cytosine nucleotides, which is an epigenetic mechanism that occurs throughout the human genome. This covalent modification is a genomic DNA mark that commonly happens at a 5-carbon position of cytosine, generally within a 5'-CpG-3' dinucleotide. Approximately 1.5% of human genomic DNA contains this dinucleotide (Lister, et al., 2009) which usually forms as clusters of un-methylated cytosine guanosine dinucleotides (CpGs) called CpG islands. These islands are generally present in gene promoter regions and do not methylate. DNA methylation occurs at the 5' carbon of the cytosine ring by adding a methyl group (Bird, 2002) and forming 5-methylcytosine. These methyl groups modify the function of DNA and effectively suppress transcription.

In general, DNA methylation effects biological processes in two ways. First, DNA methylation can steadily change the expression of genes in cells from embryonic cell division and differentiation of stem cells into a particular organization. The resulting change is usually one-way and permanent, stopping a cell from turning back into a stem cell or converted into different cell types. Second, via deleting hydroxyl methyl groups rather than completely removing methyl groups by dilution as cells divide or in a faster active process, DNA methylation can be passively deleted (Wossidlo, et al., 2011). DNA methylation is usually deleted and re-established through continuous cell division in the process of development.

DNA methylation at the 5' position of cytosine has been found in each examined vertebrate and generally, reduces gene expression with a specific effect. It usually occurs in the CpG dinucleotide context in adult somatic cells. However, non- CpG methylation is common in embryonic stem cells (Haines, et al., 2001) and also plays a role in neural development (Lister, et al., 2013). In mammalian DNA, 60%- 90% of CpGs are methylated, and 5-methylcytosine is primarily found in CpGs (Tucker, 2001). These CpGs play a key role in maintenance of cellular functions and the regulation of gene expression (Jones and Takai, 2001). In the human genome, these CpG sites exist in less than expected frequencies for the majority of the genome but are found more frequently among CpG islands. These CpG islands are generally found in or near promoter regions of genes (Herman and Baylin, 2003) and act as potential regulators of gene expression. In the promoter region, hypermethylation usually occurs in the CpG island area and is related to gene inactivation. Recently, a genome-wide high-resolution DNA methylation analysis of a primary human fibroblast cell line demonstrated that in genomic DNA, 4.25% of total cytosines are methylated, 67.7% of CpGs are methylated, and 99.98% of DNA methylation occurs in CpG dinucleotides (Lister, et al., 2009). In many disease processes, such as cancer (Baylin, 2005), gene promoter CpG islands have abnormal hypermethylation, causing transcriptional silencing which can be inherited in the daughter cells after cell division. These changes in DNA methylation are considered to be an important part of the development of cancer. Hypomethylation generally appears with chromosome instability and loss of imprinting; on the other hand hypermethylation is linked with the promoter methylation and possible secondary gene silencing (cancer suppressor genes), but may be an epigenetic therapy target. DNA methylation is essential during embryonic development, and DNA methylation patterns usually keep high fidelity to the daughter cells. Hypermethylation and hypomethylation have been associated with a large number of human malignant tumors compared to normal tissue. Generally

speaking, for carcinogenic methylation changes there is an increase in DNA methylation associated tumor suppressor genes and a decrease related to oncogenes (Gonzalo, 2010).

Many studies have explored the association of gene expression and DNA methylation, but only a few reported the combination of two features using a gene set enrichment analysis to enhance biological pathway analysis. In our research, we develop a statistic to combine gene expression and DNA methylation and performed a gene set enrichment analysis to detect relevant pathways to the phenotype of interest. The purpose of the integration of two features is to increase the power of detecting significant pathways related to the gene expression.

In this dissertation, we develop methods for integrating miRNA and mRNA expression data, as well as mRNA expression and DNA methylation data, based on sparse canonical correlation analysis (SCCA). This approach has an advantage relative to pairwise comparisons by reducing the dimension of the data to potentially increase both statistical power using this data and biological interpretability. Further, we develop a novel gene set enrichment analysis (GSEA) approach based on the integrated analysis using SCCA. GSEA allows for testing the potential enrichment of pathways and biological terms of genes that are significantly associated with the phenotype of interest. The rest of this dissertation is organized as follows. In Chapter 2, we develop an approach for integrated analysis of miRNA and mRNA expression data using SCCA and motivate the derivation of our novel test statistic for GSEA based on this analysis. In Chapter 3, we evaluate our SCCA-GSEA statistic for miRNA / mRNA expression data using simulated data, and compare it with a similar statistic based on pairwise correlation analysis of miRNA / mRNA data. In Chapter 4 we evaluate the same SCCA-GSEA statistic for detecting GO terms and KEGG pathways using several real data sets of miRNA / mRNA expression data in cancer vs. normal tissue and in tissue related to embryonic development of the murine neural tube. In Chapter 5, we develop an analog of the SCCA-GSEA statistic for use with integrated analysis of mRNA expression and DNA methylation data. In Chapter 6 we evaluate this statistic for detecting regions of epigenetic regulation associated with GO terms in data relating to murine embryonic palate development. Finally, in Chapter 7 we finish with some concluding remarks and potential areas for future research.

CHAPTER II

METHOD OF INTEGRATED ANALYSIS OF MIRNA AND MRNA EXPRESSION

2.1. Sparse canonical correlation analysis

In this research we focus on identifying the predicted target genes which have maximum negative correlation with miRNAs of interest, e.g. miRNAs that are down- or up-regulated between two sets of biological samples. So, the objective is to find an analytic method which establishes the relationships between sets of measurements from the same group of subjects and further reduces the dimensionality of the data. An often used method is principal component analysis (PCA). PCA is used for reducing the dimensions of the data sets, modeling the potential structure in the data and then aggregating the original variables into composite latent variables. As a final step, PCA is often used to model the relationships between the latent variables and additional outcome variables, an approach called principal components regression (PCR). But, there are two main disadvantages of this approach. One is that in our research we want to measure the correlation between two or among more sets of variables from populations, but PCA is mainly to maximize the variance within only one set of variables by creating composite measures. Another disadvantage is that with large scale data sets, these composite measures are based on thousands of variables. Since PCA creates latent variables (principal components) which are linear combinations of the entire sets of variables, the resulting components may lack interpretability and be difficult to visualize. The first disadvantage can be solved using the canonical correlation analysis (CCA) method. CCA is a classical technique due to Hotelling (1936) that identifies relationships among sets of variables on the same set of subjects. Specifically, CCA seeks linear combinations of the variables in two populations which have maximal correlation. Suppose there are two data $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with the same number of observations n. The first data matrix $\boldsymbol{X}_1$ is a $n \times p_1$ matrix corresponding to p_1 variables with n observations and the second matrix $\boldsymbol{X}_2$ is a $n \times p_2$ matrix corresponding to p_2 variables on the same set of observations. We assume that all columns of the matrices are standardized. The objective of CCA is to identify linear combinations of variables in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ which have maximum positive correlation.

We define that sample covariance for these standard data in matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is $\frac{1}{n-1}\boldsymbol{X}_1^T\boldsymbol{X}_2$. Then, we define the linear combinations of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ as $\boldsymbol{U} = \boldsymbol{X}_1\boldsymbol{u}$ and $\boldsymbol{V} = \boldsymbol{X}_2\boldsymbol{v}$, where vectors $\boldsymbol{u} \in \mathbb{R}^{p_1}$ and $\boldsymbol{v} \in \mathbb{R}^{p_2}$ are the weights used to determine the linear combination of measurements in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ which are maximally correlated with each other. The linear combinations $\boldsymbol{U} = \boldsymbol{X}_1\boldsymbol{u}$ and $\boldsymbol{V} = \boldsymbol{X}_2\boldsymbol{v}$ are termed the sample canonical variables. So, CCA aims to find $\boldsymbol{u}$ and $\boldsymbol{v}$ in order to maximize $\boldsymbol{u}^T\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{v}$, subject to $\boldsymbol{u}^T\boldsymbol{X}_1^T\boldsymbol{X}_1\boldsymbol{u} = \boldsymbol{v}^T\boldsymbol{X}_2^T\boldsymbol{X}_2\boldsymbol{v} = 1$.

The traditional CCA (Hotelling, 1936) approach fits the linear combinations or canonical vectors by including all variables from both data sets. In our research, the number of genomic regions / features of interest under consideration often reaches tens of thousands, while the number of samples is typically limited (in the tens to hundreds). In this case, linear combinations of the whole set of features lack biological interpretability because there are too many variables under consideration. Furthermore, insufficient sample size and high dimensional data result in inaccurately estimated parameters and many computational problems (Parkhomenko, et al., 2009).

Sparse CCA (SCCA) is an extension to classical CCA which solves the aforementioned problems concerning high dimensional data, and aids in biological interpretability by identifying sparse groups of associated variables. Instead of including all the variables in both data sets for finding correlation between two sets of variables as with traditional CCA, SCCA uses a penalty term to reduce the dimensionality of the problem. Thus, the results of SCCA are expected to be more robust compared to CCA in the high dimensional setting.

The SCCA method introduced by (Witten and Tibshirani, 2009) maximized $\boldsymbol{u}^T\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{v}$ subject to constraints on the norms of the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. Specifically, the SCCA criterion proposed in (Witten and Tibshirani, 2009) is

$$\text{maximize}_{u,v}\ \boldsymbol{u}^T\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{v}$$

$$\text{subject to } \|\boldsymbol{u}\|^2 \le 1, \|\boldsymbol{v}\|^2 \le 1, P_1(\boldsymbol{u}) \le c_1, P_2(\boldsymbol{v}) \le c_2$$

Here, the penalty functions P_1 and P_2 are either lasso (with $P_1(\boldsymbol{u}) = \|\boldsymbol{u}\|_1$) or fused lasso (with $P_1(\boldsymbol{u}) = \sum_j|\boldsymbol{u}_j| + \sum_j|\boldsymbol{u}_j - \boldsymbol{u}_{(j-1)}|$) penalties. The lasso penalty results in sparse $\boldsymbol{u}$ and/or $\boldsymbol{v}$ for appropriately

chosen c_1 and c_2 (where $1 \leq c_1 \leq \sqrt{p_1}$ and $1 \leq c_2 \leq \sqrt{p_2}$), while the fused lasso penalty results in u and/or v which are both sparse and smooth. Witten and Tibshirani (2009) introduce algorithms for estimating unconstrained u and v and for when u and v are constrained to be non-negative (or non-positive).

Witten and Tibshirani (2009) also introduced the concept of sparse multiple CCA (sparse mCCA) which generalizes sparse CCA to the setting of multiple data sets $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ where $K > 2$. Here, the goal is to find $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K$ which maximizes $\sum_{i<j} \boldsymbol{u}_i^T \boldsymbol{X}_i^T \boldsymbol{X}_j \boldsymbol{u}_j$ subject to $\|\boldsymbol{u}_i\|^2 \leq 1, P_i(\boldsymbol{u}_i) \leq c_i$, where the P_is are again convex penalty functions.

The algorithm proposed by Witten and Tibshirani (2009) for calculating the canonical covariate of the SCCA is as follows:

1. Initialize $\boldsymbol{w_2}$ to have L_1 norm 1.

2. Iterate the following two steps until convergence:

(a) $\boldsymbol{w_1} \leftarrow \arg max_{\boldsymbol{w_1}} \boldsymbol{w_1^T X_1^T X_2 w_2}$ subject to $\|\boldsymbol{w_1}\|^2 \leq 1, P_1(\boldsymbol{w_1}) \leq c_1$.

(b) $\boldsymbol{w_2} \leftarrow \arg max_{\boldsymbol{w_2}} \boldsymbol{w_1^T X_1^T X_2 w_2}$ subject to $\|\boldsymbol{w_2}\|^2 \leq 1, P_2(\boldsymbol{w_2}) \leq c_2$.

If P_1 is an L_1 penalty then the update has the form

$$\boldsymbol{w_1} \leftarrow \frac{S(\boldsymbol{w_1^T X_2 w_2}, \Delta_1)}{\|S(\boldsymbol{w_1^T X_2 w_2}, \Delta_1)\|^2},$$

where $\Delta_1 = 0$, $\|\boldsymbol{w_1}\|_1 \leq c_1$; otherwise, $\Delta_1 \geq 0$ chosen so that $\|\boldsymbol{w_1}\|_1 = c_1$. Here $S(.)$ is a soft-threshold operator; that is $S(\boldsymbol{w_1}, c) = sign(\boldsymbol{w_1})(|\boldsymbol{w_1}| - c)_+$.

In our application $\boldsymbol{X}_1$ consists of the gene expression measurements of the predicted miRNA target transcripts, and $\boldsymbol{X}_2$ consists of the miRNA expression measurements at corresponding time points. In each case, the columns represent the different miRNAs / mRNAs and the rows are the values from different subjects. Since our goal is to find combinations of miRNA and mRNA measurements which are maximally negatively correlated, the variables in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ will be standardized to have mean zero and standard deviation one and then either $\boldsymbol{X}_1$ or $\boldsymbol{X}_2$ will be multiplied by negative one prior to application of SCCA. Application of SCCA to these transformed matrices will identify linear combinations of the original $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ which have

maximum negative correlation. As stated previously, the goal of SCCA is to find unit vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ such that $\boldsymbol{u}^T \boldsymbol{X}_1^T \boldsymbol{X}_2 \boldsymbol{v}$ is maximized subject to constraints on $\boldsymbol{u}$ and $\boldsymbol{v}$. In our problem, since neither matrix of miRNA or mRNA expression measurements is ordered the lasso penalty will be used for both weight vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. One exception is that, when the number of miRNAs is small (e.g., ten or fewer), we may use no penalty for the miRNA data and retain all the miRNAs in the analysis. Further, for interpretation purposes and for construction of our GSEA statistic described below the weights are constrained to be non-negative. A permutation procedure will be used for both selecting the optimal set of tuning parameters (c_1 and c_2) and determining the significance of the correlation between the canonical variables $\boldsymbol{X}_1 \boldsymbol{u}$ and $\boldsymbol{X}_2 \boldsymbol{v}$. This procedure is advantageous for small samples, since it does not require cross-validation or splitting the sample into training and test sets. Subsequent canonical variables can be obtained by applying the procedure to the components of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ which are orthogonal to the previously obtained canonical variables. The SCCA was performed using the R package *PMA* provided by

(https://cran.r-project.org/web/packages/PMA/PMA.pdf).

2.2. Gene set enrichment analysis (GSEA)

Gene Set Analysis (GSA) is a method for estimating the significance of predefined gene sets, rather than individual genes. The gene sets could be derived from different sources, for example the sets of genes representing various biological pathways of cells (e.g., Gene Ontology , KEGG (Kanehisa, et al., 2004), Biocarta (http://www.biocarta.com), Reactome (http://www.reactome.org), MSigDB (Subramanian, et al., 2005), Pathway Interaction Database (Schaefer, et al., 2009)). The motivation behind GSA is that these genes inside the gene sets are closely related and will have similar expression patterns. Hence, there is potential for increased statistical power as well as biological interpretability because of the strong relationships between genes within the same gene set.

The GSA method works roughly as follows. Suppose we have *N* genes in the data. The initial step is to calculate a test statistic for each of the genes, e.g. for studies concerning two sets of samples (diseased and control samples) the two sample t-statistic t_i (or some variant thereof) is appropriate. The next step is to identify the predefined gene sets for all *N* genes which are denoted $GS_K = (gs_1, gs_2, \dots, gs_K)$. In GSA we use a cut-off for the t_i (e.g, a threshold for the *p*-value of the test statistics) to obtain a gene list, and then test for association between this gene list and each of the pre-defined gene sets $s_1, \dots, s_k$ using Fisher's exact test. An example 2x2 table for GSA is given in **Table 2.1**, where *N* = a + b + c + d. Results can be ordered based on the p-value from Fisher's exact test or on the fold-enrichment $a/(a+b) \Big/ (a+c)/(b+d)$ of the statistically significant gene list for genes from the given gene set. Note that statistically significant results can also be found for gene sets having an under- representation within the list of statistically significant genes, though this is usually of less practical interest.

T-test	Gene set	
	In Gene set	**Not in Gene set**
Significant	a	b
Not significant	c	d

Table 2.1: Example 2x2 table for GSA based on Fisher's test. Letters represent counts in each cell

An extension to GSA, termed gene set enrichment analysis (GSEA), has an advantage over GSA in that the user does not have to specify a significance threshold for inclusion of genes within the gene list but instead uses the entire range of information in the collective set of test statistics (Subramanian, et al., 2005). We begin with a predefined collection of gene sets $GS_K = (gs_1, gs_2, \ldots, gs_K)$ and compute a test statistic (e.g., a *t*-statistic) t_j for all N genes in our data. Let $\boldsymbol{T}_k = (t_1, t_2, \ldots, t_{n_k})$ be the gene scores for the n_k genes in gene set gs_k. Then, a gene set score (statistic) $gs_k(\boldsymbol{T}_k)$ is computed for each gene set gs_k. For Subramanian's original GSEA this was equal to a signed and weighted version of the Kolmogorov-Smirnov statistic, but later authors proposed simpler alternative statistics for $gs_k(\boldsymbol{T}_k)$ including the mean of $\boldsymbol{T}_k$ (Jiang and Gentleman, 2007; Tian, et al., 2005). The idea of characterizing the significance of a gene set is that if some or all of the gene set scores within gs_k are higher (or lower) than expected, their sum of scores $\boldsymbol{T}_k$ will be higher (or lower) than expected. The statistical significance of the gene set scores $gs_k(\boldsymbol{T}_k)$ can be determined by either permuting the gene scores t_j or by permuting the phenotypes across the samples and re-calculating the gene scores. The former addresses the null hypothesis that the scores in a given gene set do not differ from the scores outside of the gene set (the so-called competitive test), while the latter addresses the null hypothesis that the gene set does not contain genes whose expression levels are associated with the phenotype (the self-contained test) (Tian, et al., 2005). In general, the GSEA method has three specific steps: (1) rank all genes with some kind of score for each gene, (2) define a specific overall test statistic for each pre-specified gene set, (3) conduct a permutation testing procedure to assess the significance of the statistics.

2.3. GSEA score based on integrating miRNA / mRNA expression data

A modified version of GSEA which combines both mRNA and miRNA gene expression measurements is constructed based on the SCCA. The GSEA is based on a novel statistic constructed from the two sets of weight vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ obtained from SCCA, with constraints that both $\boldsymbol{u}$ and $\boldsymbol{v}$ are non-negative. We first use function *CCA.permute* in R package *PMA* provided by

(https://cran.r-project.org/web/packages/PMA/PMA.pdf) to obtain the best penalties for both mRNA and miRNA data sets. This function automatically selects best penalties for sparse CCA using the penalized matrix decompostion. The penalties are selected using a permutation procedure for each predetermined penalty value. After the permutation process, the function give z-statistic and p-value for each pair of canonical variables resulting from a given predetermined penalty value. The best penalties should have both a significant p-value (< 0.05) and best z-statistic (larger z-statistic correspond to better tuning parameter values). When the data sets are highly correlated, the *CCA.permute* function may return a very small penalty (a smaller penalty means fewer non-negative values in the $\boldsymbol{u}$ and $\boldsymbol{v}$ weight vectors are obtained). In this case, we can use the one standard error rule which is generally used within cross-validation. That is, we select the largest penalty value that has a z-statistic within one standard deviation of the optimal penalty z-statistic (that is, within the optimal z-statistic minus 1). Then, we apply the new penalty to SCCA to obtain weight vectors $\boldsymbol{u}$ and $\boldsymbol{v}$**.** After we get the first pair of canonical vectors, a second pair of canonical vectors could be obtained from the residual matrices of mRNA and miRNA. $\boldsymbol{X_1}^{\boldsymbol{residual}} = \boldsymbol{X_1} - \widehat{\boldsymbol{X}}_{\boldsymbol{1}}, \widehat{\boldsymbol{X}}_{\boldsymbol{1}} = \boldsymbol{X_1 u}((\boldsymbol{X_1 u})'\boldsymbol{X_1 u})^{-\boldsymbol{1}}(\boldsymbol{X_1 u})'$ and $\boldsymbol{X_2}^{\boldsymbol{residual}} = \boldsymbol{X_2} - \widehat{\boldsymbol{X}}_{\boldsymbol{2}}$, $\widehat{\boldsymbol{X}}_{\boldsymbol{2}} = \boldsymbol{X_2 v}((\boldsymbol{X_2 v})'\boldsymbol{X_2 v})^{-\boldsymbol{1}}(\boldsymbol{X_2 v})'$. Where $\boldsymbol{X}_1$ indicates mRNA data, $\boldsymbol{X}_2$ indicates miRNA data, $\boldsymbol{u}$ and $\boldsymbol{v}$ are weight vectors obtained from SCCA. Then, we apply $\boldsymbol{X_1}^{\boldsymbol{residual}}$ *and* $\boldsymbol{X_2}^{\boldsymbol{residual}}$ with the same procedure of the original mRNA and miRNA data sets to obtain the second pair of canonical vectors. We repeat the procedure above until no significant tuning parameter is found by the *CCA.permute* function. Specifically, after application of SCCA we obtain multiple weight vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k$ and $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$where $\boldsymbol{u_i} = (u_1, u_2, \dots, u_{p_1})$ for the weights associated with mRNAs and non-negative weight vector $\boldsymbol{v_i} = \left(v_1, v_2, \dots, v_{p_2}\right)$ for the weights associated with miRNAs, $i = 1, 2 \dots k$. Then, the $\boldsymbol{u}$ and $\boldsymbol{v}$ vectors are the summation of multiple weight vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k$ and $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ separately. The test statistic is constructed from two parts. The first part consists of the normalized $\boldsymbol{u}$ vector $\boldsymbol{u}_{\text{norm}}$, such

that the mean of $\boldsymbol{u}_{\text{norm}}$ is zero and the variance is one. This component simply indicates the degree to which each putative target gene is represented in the weight vector $\boldsymbol{u}$. The second part consists of the weight vector v multiplied with the putative target matrix $\boldsymbol{P}_{p_2 \times p_1}$, where P_{ij} = 1 if miRNA i putatively targets gene j and is zero otherwise. This part is calculated as $\boldsymbol{v}^* = \boldsymbol{v}^T \boldsymbol{P}$, where the dimension of $\boldsymbol{v}^*$ is $1 \times p_1$. This component incorporates the weights associated with each miRNA into the per-gene scores, and also accounts for the degree of targeting associated with each miRNA / mRNA. The $\boldsymbol{v}^*$ scores are also normalized to have mean zero and standard deviation one ($\boldsymbol{v}^*_{\text{norm}}$). Under the null hypothesis miRNA and mRNA are not correlated, which means $\boldsymbol{u}_{\text{norm}}$ scores are independent to $\boldsymbol{v}^*_{\text{norm}}$. Both $\boldsymbol{u}_{\text{norm}}$ and $\boldsymbol{v}^*_{\text{norm}}$ scores have mean 0 and standard deviation 1. Although we do not know the distribution of these scores, we could also normalize the summation of $\boldsymbol{u}_{\text{norm}}$ and $\boldsymbol{v}^*_{\text{norm}}$ scores to have mean 0 and standard deviation one. So we set the final statistic associated with each gene by summing the normalized $\boldsymbol{u}$ and $\boldsymbol{v}^*$ scores and then dividing by $\sqrt{2}$, which is denoted as $\boldsymbol{Z} = \frac{1}{\sqrt{2}}(\boldsymbol{u}_{\text{norm}} + \boldsymbol{v}^*_{\text{norm}})$.

Then, we calculate an aggregate gene enrichment score for each gene set, where the gene sets are pre-determined from e.g. the KEGG and GO databases. Specifically, suppose there are *K* pre-determined gene sets with $n_1, n_2, \ldots, n_K$ genes in each set. In our terminology, the vector $\boldsymbol{Z}_j$consists of the components of $\boldsymbol{Z}$ corresponding to the genes in gene set j. Then, the GSEA statistic $gs_1(\boldsymbol{Z}_1),\ gs_2(\boldsymbol{Z}_2), \ldots, gs_K(\boldsymbol{Z}_K)$ for each gene set is calculated by the sum of the per gene statistic included in each gene set then divided by the square root of the number of genes in each gene set:

$$gs_k(\boldsymbol{Z}_k) = \frac{1}{\sqrt{n_k}}\left(\sum_{j=1}^{n_k} z_j\right),$$

where $k = 1,2, \ldots, K$ and $\boldsymbol{Z}_k = (z_1,\ z_2, \ldots,\ z_{n_k})$ are the gene statistics for gene set gs_k.

The motivation for using this statistic is that under the null hypothesis of no association between $\boldsymbol{X}_1$and $\boldsymbol{X}_2$, each component of $\boldsymbol{Z}_K$ is expected to have mean 0 and standard deviation 1. Presuming that the components of $\boldsymbol{Z}_K$ are also independent under the null, then the statistic $gs_k(\boldsymbol{Z}_k)$ is also expected to have mean 0 and standard deviation 1. While the distribution of $gs_k(\boldsymbol{Z}_k)$ is unknown, the gene sets scores defined in this manner then indicate how many standard deviations away from the null expectation the gene set statistic is.

When $\sqrt{n_k}$ gets large enough, the distribution of $gs_k(\boldsymbol{Z}_k)$ approximately converges to asymptotic normal $N(0,1)$.

Finally, we calculate the permutation *p*-value of the GSEA statistic for each gene set using both the competitive test and the self-contained test. For the competitive test, we first resample the per-gene statistics $\boldsymbol{Z} = (z_1, z_2, \dots, z_{p_1})$ without replacement to obtain permuted statistics $\boldsymbol{Z}^m$ for $m = 1, \dots, M$ permutations. Next, permuted gene set statistics $gs_1^m(\boldsymbol{Z}_1^m), gs_2^m(\boldsymbol{Z}_2^m), \dots, gs_K^m(\boldsymbol{Z}_K^m)$ are calculated for each of the original gene sets, where $\boldsymbol{Z}_k^m = (z_1^m, z_2^m, \dots, z_{n_k}^m)$ are the permuted gene statistics for each gene set gs_k. The permutation *p*-value $p_{\text{perm},k}$ for each gene-set k is then calculated as the proportion of the permuted GSEA statistics that are larger than the original GSEA statistic:

$$p_{\text{perm},k} = \frac{1}{M}\sum_{m=1}^{M} I\left(gs_k^m(\boldsymbol{Z}_k^m) > gs_k(\boldsymbol{Z}_k)\right),$$

where $I(\cdot)$ is the indicator function.

For the self-contained test, we resample the samples of the mRNA data set without replacement to obtain permuted mRNA data. Then, we apply SCCA on the permuted mRNA data and original miRNA data to get permuted statistics vector $\boldsymbol{u}^m$ and vector $\boldsymbol{v}^m$ for $m = 1, \dots, M$ permutations. Then, a permuted per gene statistic $\boldsymbol{Z}^{ms}$ is calculated by $\boldsymbol{u}^m$ and $\boldsymbol{v}^m$ using the same procedure as we used to calculate the $\boldsymbol{Z}$ score in the competitive-test. Next, the self-contained permuted gene set statistics $gs_1^{ms}(\boldsymbol{Z}_1^{ms}), gs_2^{ms}(\boldsymbol{Z}_2^{ms}), \dots, gs_K^{ms}(\boldsymbol{Z}_K^{ms})$ are calculated for original gene sets, where $\boldsymbol{Z}_k^{ms} = (z_1^{ms}, z_2^{ms}, \dots, z_{n_k}^{ms})$ are the self- contained permuted gene statistics for gene set gs_k. The permutation *p*-value $p_{\text{self-perm},k}$ for each gene-set k is then calculated as the proportion of the permuted GSEA statistics that are larger than the original GSEA statistic:

$$p_{\text{self-perm},k} = \frac{1}{M}\sum_{m=1}^{M} I\left(gs_k^{ms}(\boldsymbol{Z}_k^{ms}) > gs_k(\boldsymbol{Z}_k)\right),$$

where $I(\cdot)$ is the indicator function.

2.4. Integrated analysis of miRNA and mRNA expression data based on pairwise correlation analysis

For comparison purposes, we construct a pairwise correlation (PWC) GSEA statistic based on the pairwise Pearson product-moment correlation coefficients between the mRNA (matrix $\boldsymbol{X}_1$) and miRNA (matrix $\boldsymbol{X}_2$) expression measurements. Let $\mathbf{Q}_{X_2X_1}$ denotes the sample correlation matrix between $\boldsymbol{X}_2$ and $\boldsymbol{X}_1$ (which we abbreviate as $\mathbf{Q}$ in what follows). To construct the per-gene statistics used for the PWC approach, we filter $\mathbf{Q}$ in the following manner. First, all the non-negative correlations are set to be zero. Second, all correlation coefficients with adjusted p-values (based on the Benjamini- Hochberg method for controlling the false discovery rate) above a predetermined level α are set to be zero as well. Then, we denote the filtered $\mathbf{Q}$ matrix as $\mathbf{Q}^*$. In essence, $\mathbf{Q}^*$ contains only significant (after controlling for multiple comparisons) pairwise negative correlations between miRNAs and mRNAs (all other elements are zero). This matrix is then multiplied element-wise with the putative target matrix $\boldsymbol{P}_{p_2\times p_1}$ which we create in SCCA section and the final PWC per-gene statistics obtained as $\mathbf{Z}_{\text{pwc}}^{\text{T}} = {p_1}^{-1}\mathbf{1}^{\text{T}}\mathbf{Q}^* \circ \boldsymbol{P}$, where $\circ$ denotes the Hadamard product and $\mathbf{1}^{\text{T}}$ is a vector of ones of length p_1. The GSEA statistic is calculated as the mean of the z_{pwc} values for each gene set, as in the SCCA method. The permutation *p*-value was calculated by the same process as competitive test for the SCCA method. The function *rcorr* in R package *Hmisc* (Harrell, 2014) was used to calculate the sample correlation matrix.

CHAPTER III

SIMULATION OF INTEGRATED ANALYSIS OF MIRNA AND MRNA EXPRESSION

3.1 Simulation Strategy

The goal of the simulation study was to compare the ability of the SCCA and PWC methods to detect statistically significant gene sets (e.g., biological pathways) between two sets of variables from mRNA and miRNA population data sets. We use the same approach as (Witten and Tibshirani, 2009) to simulate correlated miRNA and mRNA sample data, while the approach to simulating gene sets was adopted from (Efron and Tibshirani, 2007) and (Abatangelo, et al., 2009). Here $\boldsymbol{X}_1$ is a $n \times p_1$ matrix consisting of the mRNA expression measurements and $\boldsymbol{X}_2$ is a $n \times p_2$ matrix of miRNA expression measurements on the same set of subjects. We presume that only the first r_1 variables in $\boldsymbol{X}_1$ are highly correlated with the first r_2 variables in $\boldsymbol{X}_2$, while the rest of the variables in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are uncorrelated. The data sets are simulated as follows. First, we generate a latent random vector for both data sets $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_n]^T$ from $N(0, \sigma_\gamma^2 \boldsymbol{I}_n)$, where $\boldsymbol{I}_n$ is the $n \times n$ identify matrix. Then, we generate vectors $\boldsymbol{u} \in \mathbb{R}^{p_1}$ and $\boldsymbol{v} \in \mathbb{R}^{p_2}$ where $u_1, u_2, \ldots, u_{r_1}$ are independent and identically distributed (iid) $N(\mu_u, \sigma_u^2)$, $v_1, v_2, \ldots, v_{r_2}$ are iid $N(\mu_v, \sigma_v^2)$, and the remaining elements in $\boldsymbol{u}$ and $\boldsymbol{v}$ are set to zero. The vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are the weights used to determine the linear combination of measurements in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ which are maximally inversely correlated with each other. So, the values in each data set are generated as follows:

$$x_{1ij} = u_j \gamma_i + e_{1ij} \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, p_1 ,$$

$$x_{2ij} = -v_j \gamma_i + e_{2ij} \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, p_2 ,$$

where e_{1ij} and e_{2ij} are both $N(0, \sigma_e^2)$. The negative sign for the product $-v_j \gamma_i$ ensures negative correlation between the r_1 variables in $\boldsymbol{X}_1$ and the r_2 variables in $\boldsymbol{X}_2$, as long as the parameters for the distributions of $\boldsymbol{u}$ and $\boldsymbol{v}$ are chosen appropriately.

Lastly, we simulated the putative target matrix $\boldsymbol{P}_{p_2 \times p_1}$ which determines whether miRNA i putatively targets gene j (P_{ij} = 1) or not (P_{ij} = 0). First, for each of the i miRNAs the number of targeted genes is simulated by an integer uniform random number $n_{\text{target,i}}$ between $n_{\text{target,min}}$ and $n_{\text{target,max}}$. Second, which genes are targeted by each miRNA is simulated. For the $p_2 - r_2$ unassociated miRNAs these are generated as a random sample without replacement of $n_{\text{target,i}}$ genes from the p_1 total number of genes. For the r_2 associated miRNAs a fraction p_{related} of the targets are randomly selected from the r_1 associated genes, and the remaining are randomly selected from the $p_1 - r_1$ other genes.

3.2 Simulation study results – single gene set

Our initial simulation study consisted of a single significant gene set having inversely correlated mRNA and corresponding targeting miRNA expression measurements. The total number of genes (mRNAs) was set to 1000, the total number of miRNAs to 50, and we assumed 50 gene sets each consisting of 20 genes within each set. All other parameters were fixed. Details of the simulation study parameters are given in **Table 3.1**. P-values for the significance of each gene set based on the SCCA and PWC approaches were calculated based on the competitive test permutation procedures outlined in **Sections 2.1** and **2.2**, respectively, with adjustment for multiple comparisons based on the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) for controlling the false discovery rate. Power for each method was calculated as the proportion of times out of 100 replicates that the associated gene set was declared significant (adjusted p-value < 0.05). The type I error rate for each method was calculated as the proportion of times out of 100 replicates that the remaining gene sets were declared significant. The simulation studies were performed based on 100 replications and the averaged results are presented in figures (3.1- 3.6).

Table 3.1: Summary of Parameters in Simulation Studies

Parameters	Description	Value
n	Sample Size	10, 20, 30, 40, 50
p_1	Total number of mRNAs	1000
p_2	Total number of miRNAs	50
r_1	Number of related mRNAs	5, 10 and 20
r_2	Number of related miRNAs	1, 5 and 20
n_{gs}	Number of gene sets	50
n_g	Number of genes in each gene set	20
μ_u, μ_v	Mean for weight vector for related mRNAs and miRNAs	$\mu_u = \mu_v = 1$
σ_γ	Standard deviation for latent vector relating miRNAs and mRNAs	1
σ_u	Standard deviation for weight vector for related mRNAs	0.1
σ_v	Standard deviation for weight vector for related miRNAs	0.2
σ_e	Standard deviation for error of expression measurements (both miRNA and mRNA)	0.5, 1, 1.5
$n_{\text{target,min}}$	Minimum number of putative targets for each miRNA	25
$n_{\text{target,max}}$	Maximum number of putative targets for each miRNA	40
p_{related}	From the r_2 associated miRNAs the fraction of targets selected from the r_1 associated genes	0.5
α	Threshold for PWC method	0.05

Figure 3.1- 3.2. show the results for the simulated data sets where the parameters of the simulated data sets were number of related mRNAs $P_1 = 5$ for all nine data sets, and number of related miRNAs $P_2 = 1, 5, 20$ respectively for three of the nine data sets, each row of the plot represents the power of data sets with same number of related miRNAs (e.g. first row of plots indicated the three data sets with related miRNAs equaled to 20). The columns of the figure represent the power of data sets with different error rates for both miRNA and mRNA data sets. In the figure, the x-axis represents the number of samples in each data set (i.e. n= 10, 20, 25, etc.), and the y-axis indicates the power or the error rate of detecting the gene sets for each method. The solid lines with triangles represent the power or error rate under each sample size for the SCCA method and the dotted lines with circles indicate the power or error rate under each sample size for the PWC method.

Figures 3.1- 3.2 indicate the power and error rate of the data sets with correlated miRNAs equal to 1, 5 and 20 respectively shown in each row. For figure 3.1, we can see that under the conditions related miRNA equaled to 1, $\sigma_e = 0.5$ and $\sigma_e = 1$, except the sample size of the data equal to 10 for $\sigma_e = 1$, the power of the pair wise correlation method is larger than the SCCA method. Also, the power of PWC is larger than SCCA when related miRNAs are equal to 5, $\sigma_e = 0.5$ and the sample size equal to 25. However, under other conditions the power of SCCA method is larger than power of PWC. Second, we can also see that the power increased with sample size for both methods and that the power increases when the number of targeting and correlated miRNAs increases. Opposite, the power decreases with increasing standard deviation of expression measurements for both mRNA and miRNA under same sample size. Figure 3.2 is the error rate for both SCCA and PWC methods. We can see that the error rate of the PWC method equals zero under every condition. And the error rate for the SCCA method is also small. The largest value of the error rate is 0.002 which is considerably less than 0.05. We can conclude that while the nominal error rate is not exceeded by either method, the methods may be lacking power due to being overly conservative.

Figures 3.3- 3.4 show the results for the simulated data sets where the parameters of the simulated data sets were number of related mRNAs $P_1 = 10$ for all the nine data sets. And other settings were the same as the simulated data sets corresponding to Figure 3.1.

Figures 3.3- 3.4 indicate the data sets with correlated mRNAs equal to 10. For Figure 3.3, we can see that the power of PWC is larger than SCCA only when related number of miRNA= 1 and $\sigma_e = 0.5$ and $\sigma_e = 1$. For other conditions power of pair wise correlation method is less than or equal to the SCCA method. Other results are similar to Figure 3.1. Figure 3.4 is the error rate for both SCCA and PWC methods. The results are similar to Figure 3.2. The largest value of error rate is 0.0012 which is less than 0.05.

Figures 3.5 and 3.6 show the results for the simulated data sets where the parameters of the simulated data sets were number of related mRNAs $P_1 = 20$ for all the nine data sets. And other settings were the same as the simulated data sets corresponding to Figure 3.1.

Figures 3.5 and 3.6 indicate the data sets with correlated mRNAs equal to 20. For figure 3.5, the result is similar to figure 3.3. The figure 3.6 is the error rate for both SCCA and PWC methods. We can see that the

error rate of PWC method equals to 0 under every condition. And the error rate for SCCA method is also small. The largest value of error rate is 0.0016 which is less than 0.05.

In comparing the two methods, the PWC method has larger power than the SCCA when the number of miRNAs is small. But it changes when the number of associated miRNAs increases. This separation is greatest when the subset size is small and the standard deviation is large, with the power of the methods rapidly converging to each other as the sample size increases. Also, we can conclude that the power increases as the number of related mRNAs and miRNAs increases. The error rates of both methods are small for all conditions, possibly indicating a lack of power due to being overly conservative.

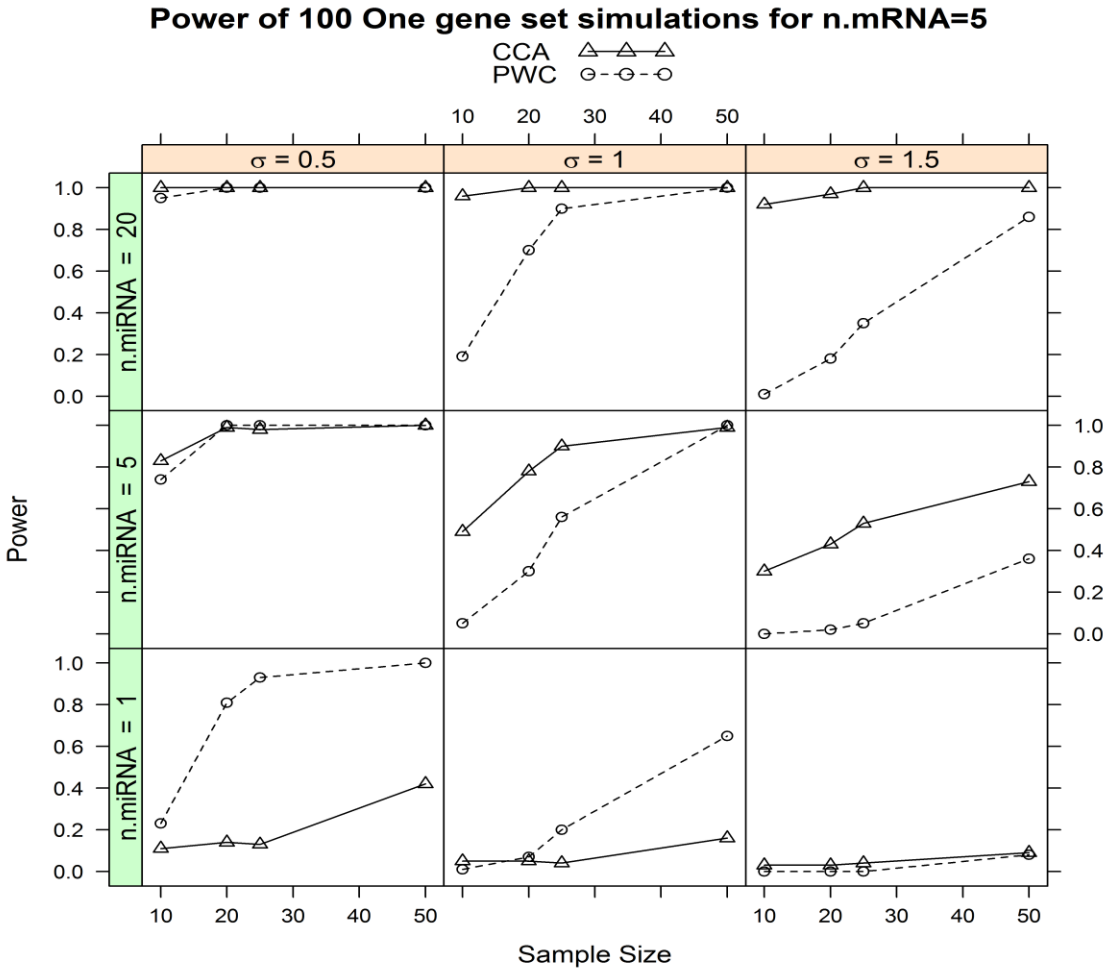


Figure 3.1. The average power of both pair-wise correlation and SCCA methods over 100 simulated data sets. In each data set, there were five correlated mRNAs, where $P_1 = 5, r_2 = 995$. There are 9 plots in the figure, where $\sigma_e = 0.5,\ \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set (i.e. n =5, 15, 25, etc.), and y- axis indicates the power of detecting the correlated genes by two methods. The dotted lines indicate the power under each sample size for PWC. The solid line indicated power for SCCA method.

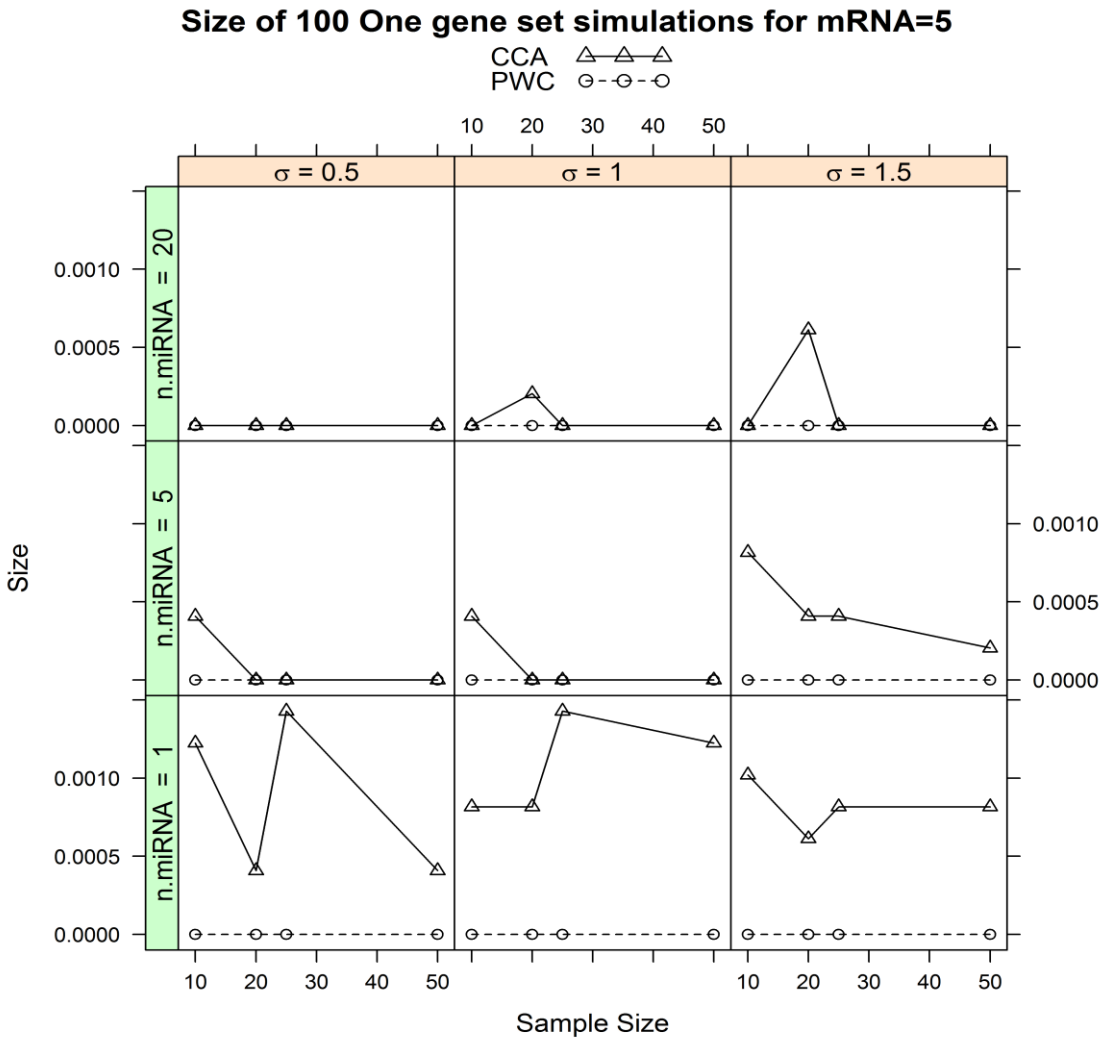


Figure 3.2. The average error rate of both pair-wise correlation and SCCA methods over 100 simulated data sets. In each data set, there were five correlated mRNAs, where $P_1 = 5, r_2 = 995$. There are 9 plots in the figure, where $\sigma_e = 0.5, \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set (i.e. $n = 5$, 15, 25, etc.), and y-axis indicates the error rate of detecting the correlated genes by two methods. The dotted lines indicate the error rate under each sample size for PWC. The solid line indicated error rate for SCCA method.

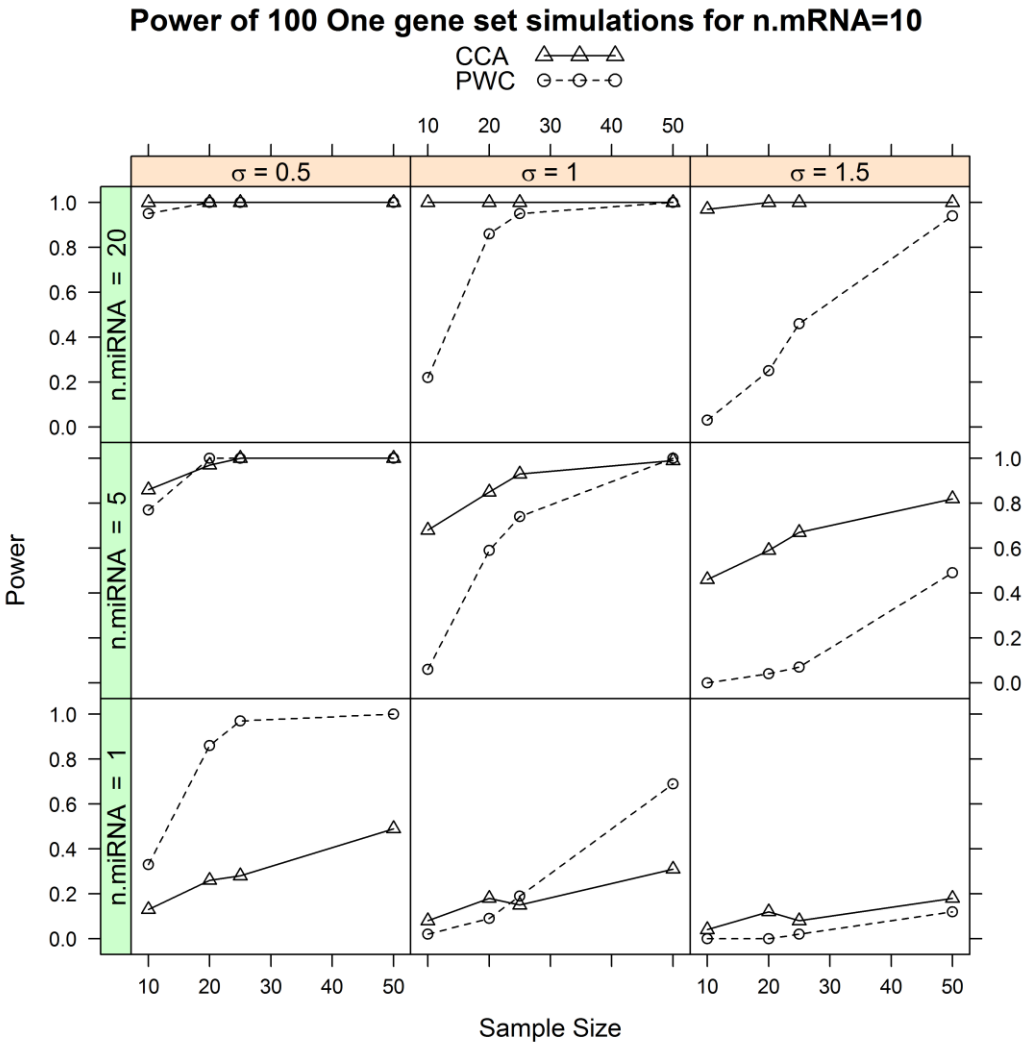


Figure 3.3. The average power of both pair-wise correlation and SCCA methods over 100 simulated data sets. In each data set, there were ten correlated mRNAs, where $P_1 = 10, r_2 = 990$. There are 9 plots in the figure, where $\sigma_e = 0.5, \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set, and y-axis indicates the power of detecting the correlated genes by two methods. The dotted lines indicate the power under each sample size for PWC. The solid line indicated power for SCCA method.

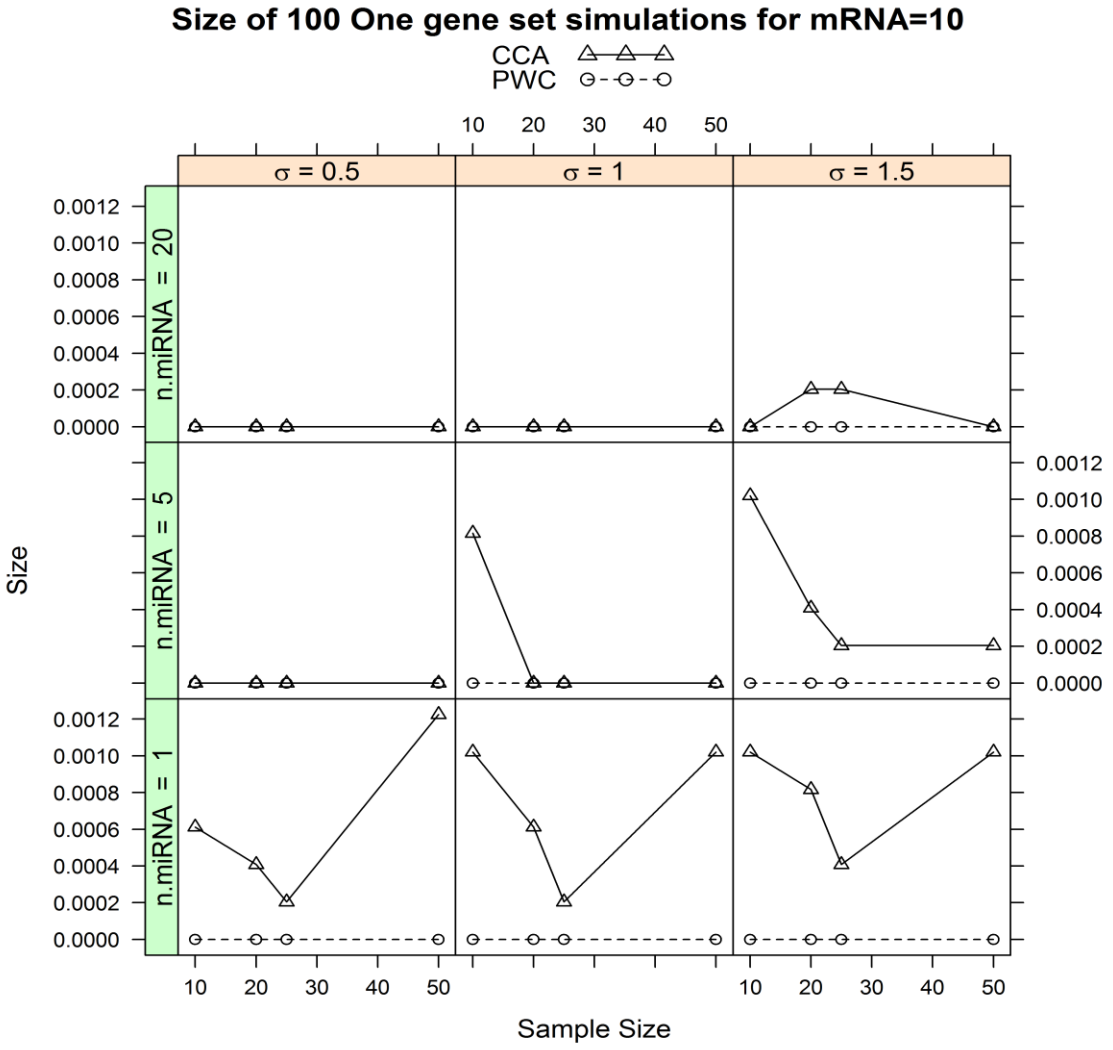


Figure 3.4. The average error rate of both pair-wise correlation and SCCA methods over 100 simulated data sets. In each data set, there were ten correlated mRNAs, where $P_1 = 10, r_2 = 990$. There are 9 plots in the figure, where $\sigma_e = 0.5, \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set and y-axis indicates the error rate of detecting the correlated genes by two methods. The dotted lines indicate the error rate under each sample size for PWC. The solid line indicated error rate for SCCA method.

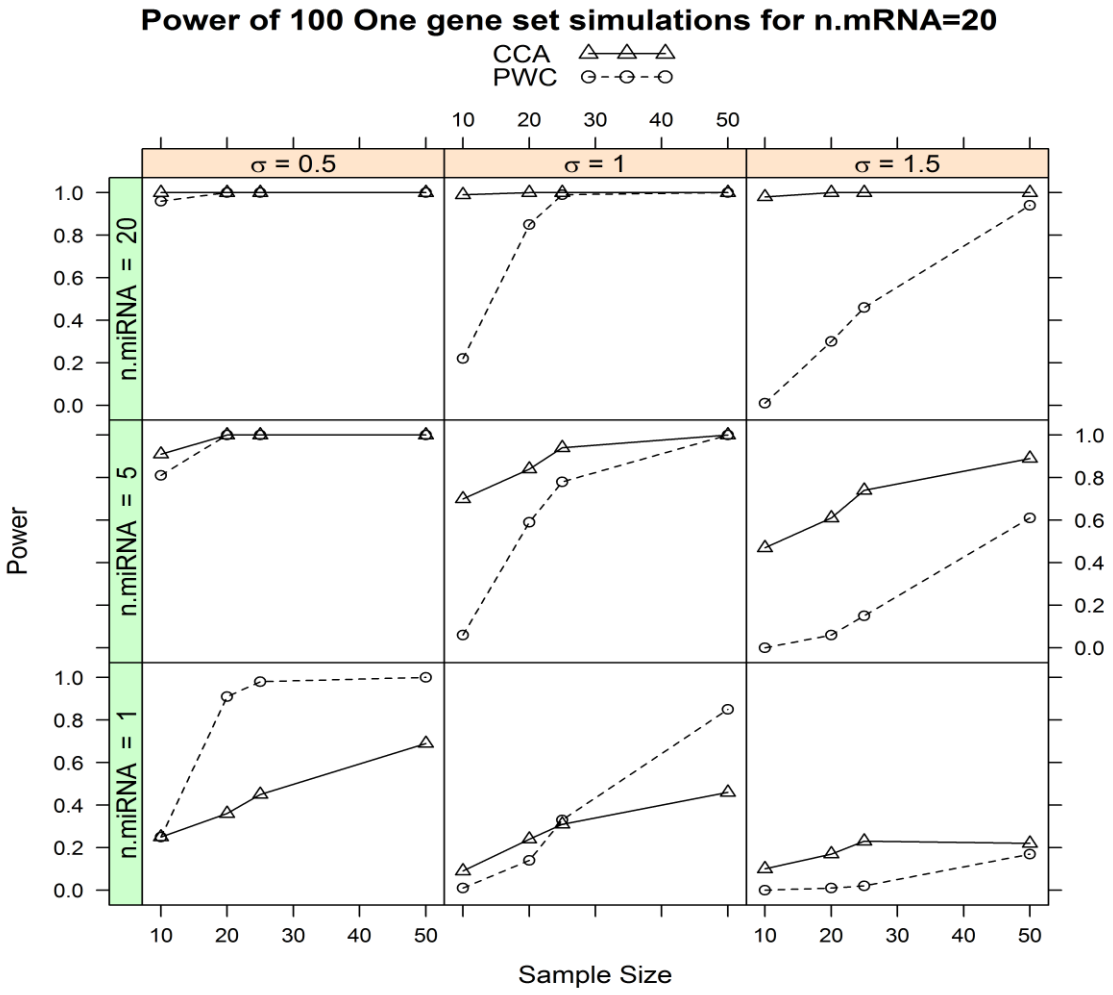


Figure 3.5. The average power of both pair-wise correlation and SCCA methods over 100 simulated data sets. In each data set, there were twenty correlated mRNAs, where $P_1 = 20, r_2 = 980$. There are 9 plots in the figure, where $\sigma_e = 0.5, \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set, and y-axis indicates the power of detecting the correlated genes by two methods. The dotted lines indicate the power under each sample size for PWC. The solid line indicated power for SCCA method.

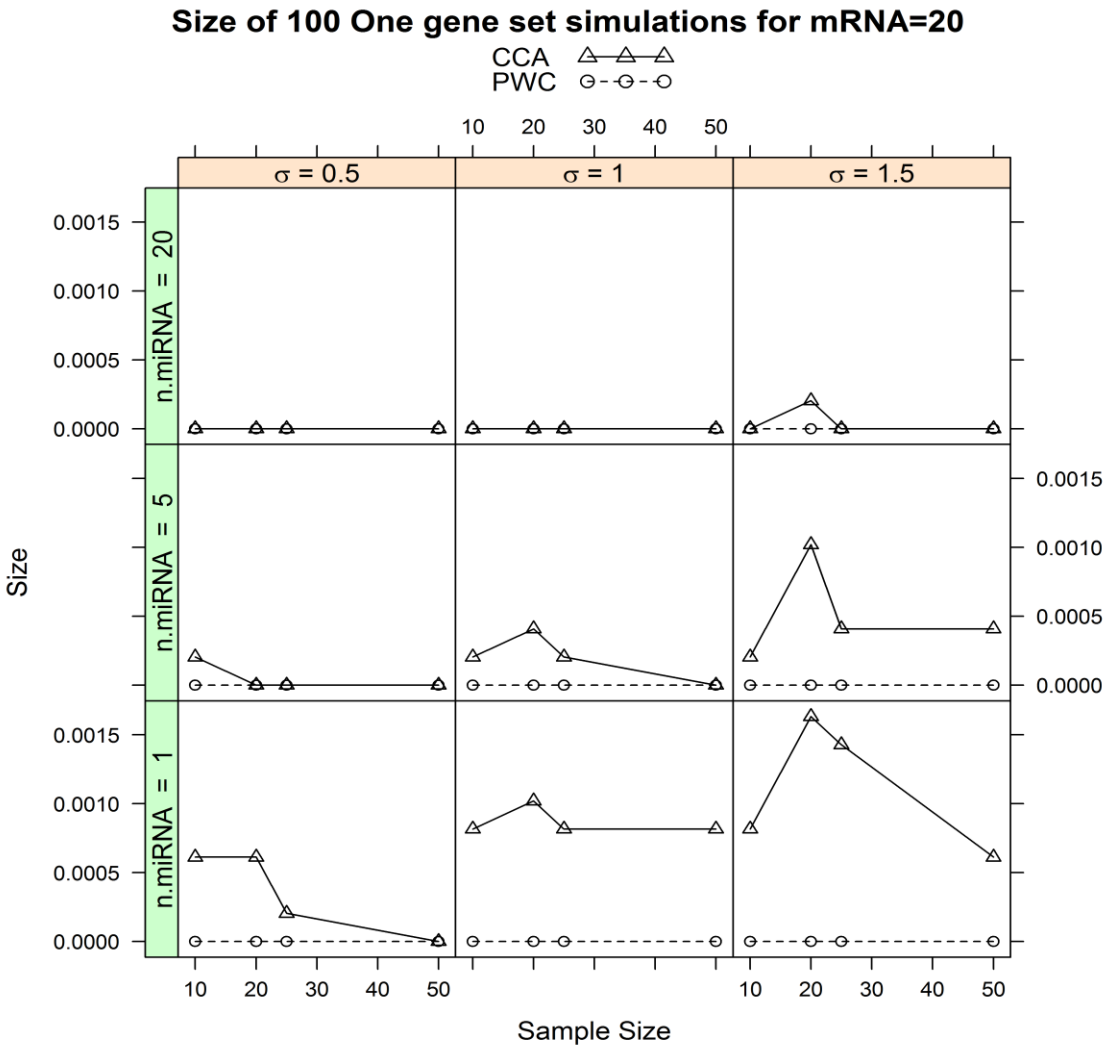


Figure 3.6. The average error rate of both pair-wise correlation and SCCA methods over 100 simulated data sets. In each data set, there were twenty correlated mRNAs, where $P_1 = 20, r_2 = 980$. There are 9 plots in the figure, where $\sigma_e = 0.5, \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set and y-axis indicates the error rate of detecting the correlated genes by two methods. The dotted lines indicate the error rate under each sample size for PWC. The solid line indicates error rate for SCCA method.

3.3. Simulation study results – two gene sets

The goal of the two data sets simulation study is similar to the one gene set simulation. This simulation study consisted of two significant gene sets each having inversely correlated mRNA and corresponding targeting miRNA expression measurements. In two gene sets simulation, we generated data sets as follows. We first separately created two data sets each with the same parameters and a single significant gene set, then bound the two data sets by columns. For each data set, the number of mRNAs was set to 500, the total number of miRNAs to 25, and we assumed 25 gene sets each consisting of 20 genes within each set. All other parameters were fixed. Details of the simulation study parameters are given in **Table 3.2**. P-values for the significance of each gene set based on the SCCA and PWC approaches were calculated based on the competitive test permutation procedures outlined in **Sections 2.1** and **2.2**, respectively, with adjustment for multiple comparisons based on the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) for controlling the false discovery rate. Power for each method was calculated as the proportion of times out of 100 replicates that the associated gene set was declared significant (adjusted p-value < 0.05). The type I error rate for each method was calculated as the proportion of times out of 100 replicates that the remaining gene sets were declared significant. The simulation studies were performed based on 100 replications and the averaged results are presented in Figures 3.7 to 3.12.

Table 3.2: Summary of Parameter in Simulation Studies for each data pair

Parameters	Description	Value
n	Sample Size	10, 20, 30, 40, 50
p_1	Total number of mRNAs	500
p_2	Total number of miRNAs	25
r_1	Number of related mRNAs	5, 10 and 20
r_2	Number of related miRNAs	1, 5 and 20
n_{gs}	Number of gene sets	25
n_g	Number of genes in each gene set	20
μ_u, μ_v	Mean for weight vector for related mRNAs and miRNAs	$\mu_u = \mu_v = 1$
σ_γ	Standard deviation for latent vector relating miRNAs and mRNAs	1
σ_u	Standard deviation for weight vector for related mRNAs	0.1
σ_v	Standard deviation for weight vector for related miRNAs	0.2
σ_e	Standard deviation for error of expression measurements (both miRNA and mRNA)	0.5, 1, 1.5
$n_{\text{target,min}}$	Minimum number of putative targets for each miRNA	25
$n_{\text{target,max}}$	Maximum number of putative targets for each miRNA	40
p_{related}	From the r_2 associated miRNAs the fraction of targets selected from the r_1 associated genes	0.5
α	Threshold for PWC method	0.05

Figures 3.7 and 3.8 show the results for two data set simulations, where the parameters for both data sets were number of related mRNAs $P_1 = 5$ for all nine data sets, and number of related miRNAs $P_2 = 1, 5, 20$ respectively for each corresponding to a single row in the figure. The parameters indicated that there were 5 correlated mRNAs included in each data set and 1, 5 and 20 correlated miRNAs respectively in each data. In each figure, the x-axis represents the sample size in each data set (i.e. n= 10, 20, 30, etc.), and the y-axis indicates the power or the error rate of detecting the gene sets by the two methods. The solid lines with triangles represented the power or error rate under each sample size for the SCCA method and the dotted lines with circles mean the power or error rate under each sample size for the PWC method.

Figures 3.7- 3.8 indicate the power and error rate of the data set with correlated mRNAs equal to 5 and correlated miRNAs equal to 1, 5 and 20 which respectively correspond to rows of Figure 3.7. We can see

that under the conditions when the related miRNAs is one and $\sigma_e = 0.5$, the power of pair wise correlation is larger than that for SCCA for all the sample sizes. And for $\sigma_e = 1$, under the condition that sample sizes of the data sets equal to 10 and 20, the average power of the PWC method is less than the SCCA method. But when the sample size is larger than 20, the average power of PWC is larger than the average power of the SCCA method. However, under other conditions the power of PWC is smaller than SCCA for all the sample sizes. We can also see that the power increases as the sample size increases and the number of related miRNAs increase for both methods. Opposite, the power decreases with increasing standard deviation of expression measurements for both mRNA and miRNA under the same sample size. Figure 3.8 is the error rate for both SCCA and PWC methods. We can see that the error rate of the PWC method is zero under every condition and the error rate for the SCCA method is also small. The largest value of error rate is 0.018 which is less than 0.05. Hence similar conclusions hold as with the single gene set simulation study.

Figure 3.9 and 3.10 show the results for two data sets simulation, where the parameters for both data sets were $P_1 = 10$ for all nine data sets, and $P_2 = 1, 5, 20$ respectively. The parameters indicate that there were 10 correlated mRNAs included in each data set and 1, 5 and 20 correlated miRNAs respectively in each data. And other settings were the same as the simulated data sets corresponding to Figure 3.7.

Figures 3.9- 3.10 indicate the power and error rate of the data set with correlated mRNAs equal to 10 and correlated miRNAs equal to 1, 5 and 20 which respectively corresponds to rows of Figure 3.9. The result of power of Figure 3.9 is similar to that for Figure 3.7. The Figure 3.10 is the error rate for both SCCA and PWC methods. We can see that the result is similar to Figure 3.8, error rate of PWC method equals to 0 under every condition. And the largest value of error rates is 0.0015 which are less than 0.05.

Figure 3.11- 3.12 show the results for two data sets simulation, where the parameters for both data sets were $P_1 = 20$ for all nine data sets, and other settings were the same to the simulated data sets corresponding to figure 3.7. Looking at Figures 3.11- 3.12, the result of Figure 3.11 is similar to Figures 3.7 and 3.9. The Figure 3.12 is the error rate for both SCCA and PWC methods. The result of Figure 3.12 is similar to Figures 3.8 and 3.10, error rate of PWC method equals to 0 under every condition. And the largest value of error rates is 0.0015 which are less than 0.05.

Over all conditions, several general observations can be concluded by looking at the figures. First, both power of SCCA and PWC increase with increasing sample size and number of miRNAs and mRNAs, as expected. Opposite, the power decreases and as the error rate increases with increasing standard deviation of expression measurements for both mRNAs and miRNAs.

In comparing the two methods, the PWC method has larger power than SCCA only when the number of miRNAs is equal to $1, \sigma_e = 0.5,$ and $\sigma_e = 1$ with a small sample size. But it changes as the number of miRNAs associated with the mRNAs in the gene set increases. This separation is greatest when the subset size is small and the standard deviation is large, with the power of both methods rapidly converging to each other as the subset size increases. The error rates of both methods are small for all conditions and possibly indicate that the methods are overly conservative.

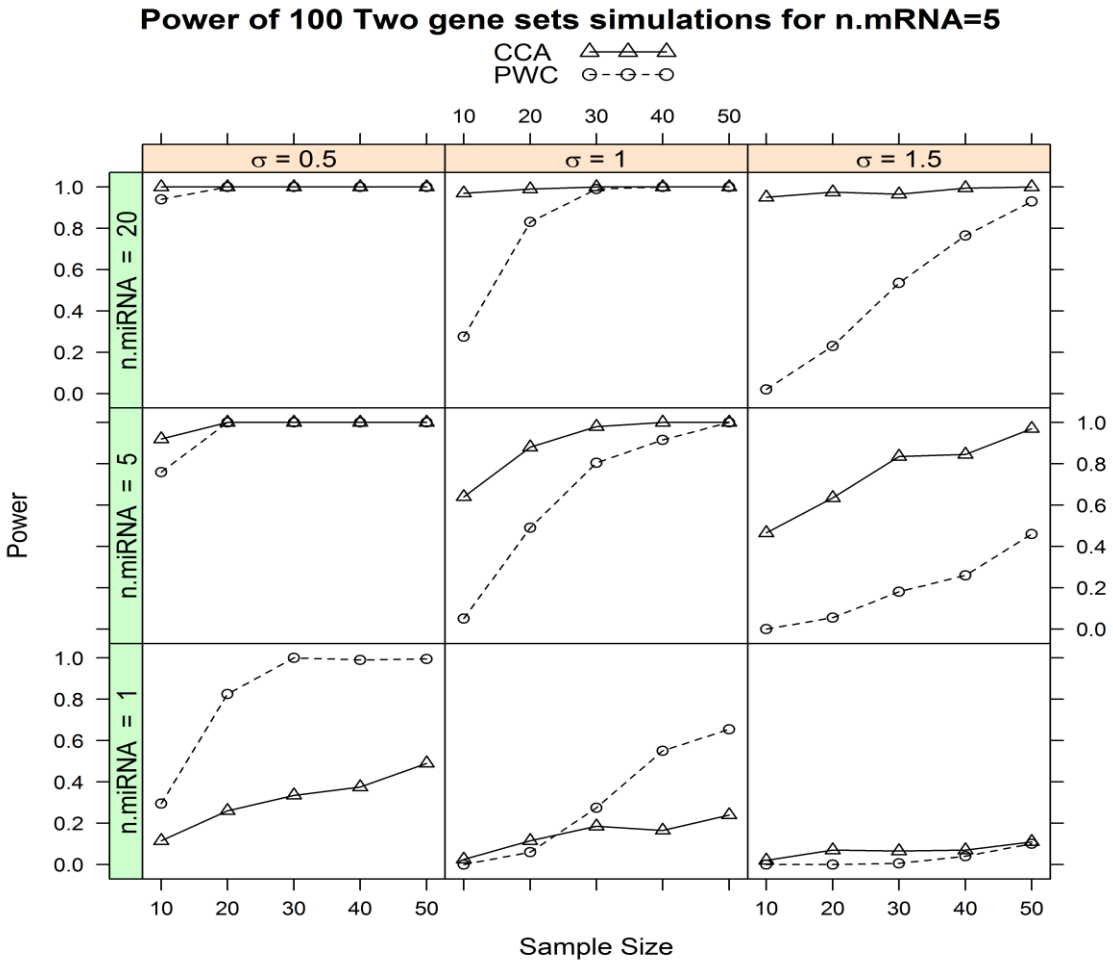


Figure 3.7. The average power of both pair wise correlation and SCCA methods over 100 two data sets simulation. In each data set, there were five correlated mRNAs, where $P_1 = 5, r_2 = 495$. There are 9 plots in the figure, where $\sigma_e = 0.5, \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set (i.e. n =10, 20, 30, etc.), and y-axis indicates the power of detecting the correlated genes by two methods. The dotted lines indicate the power under each sample size for PWC. The solid line indicated power for SCCA method.

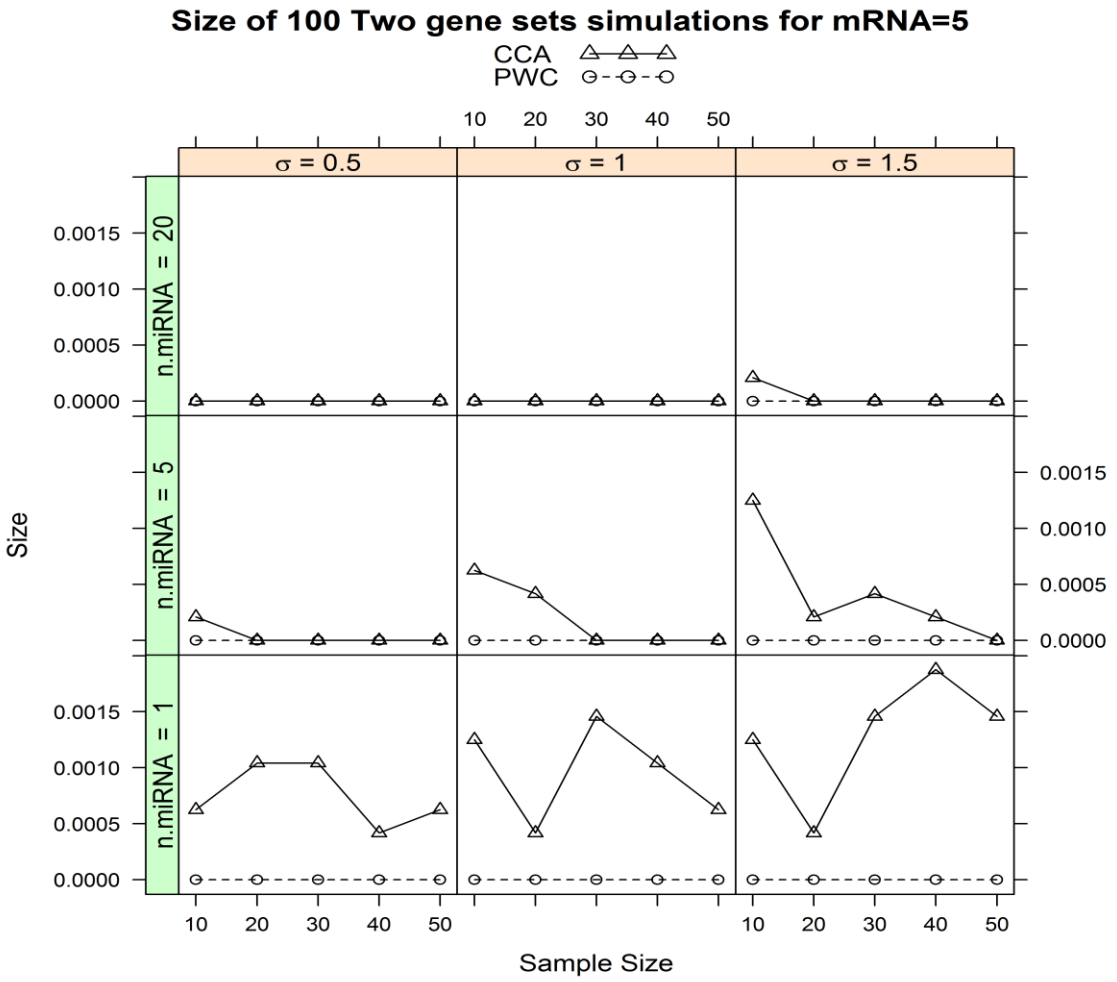


Figure 3.8. The average error rate of both pair wise correlation and SCCA methods over 100 two data sets simulation. In each pair of data sets, there were 5 correlated mRNAs, where $P_1 = 5, r_2 = 495$. There are 9 plots in the figure, where $\sigma_e = 0.5, \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set (i.e. n =5, 15, 25, etc.), and y-axis indicates the error rate of detecting the correlated genes by two methods. The dotted lines indicate the error rate under each sample size for PWC. The solid line indicated error rate for SCCA method.

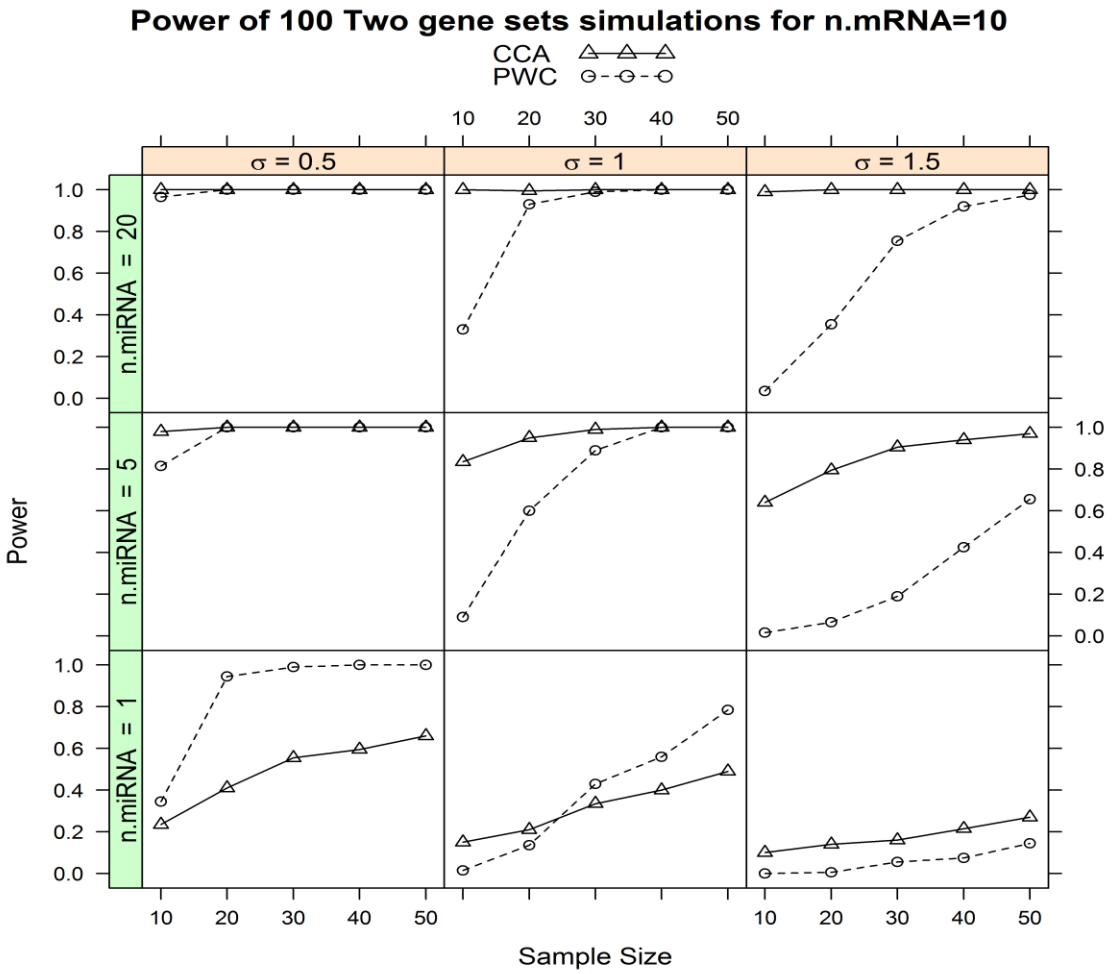


Figure 3.9. The average power of both pair wise correlation and SCCA methods over 100 two data sets simulation. In each data set, there were ten correlated mRNAs, where $P_1 = 10, r_2 = 490$. There are 9 plots in the figure, where $\sigma_e = 0.5, \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set, and y-axis indicates the power of detecting the correlated genes by two methods. The dotted lines indicate the power under each sample size for PWC. The solid line indicated power for SCCA method.

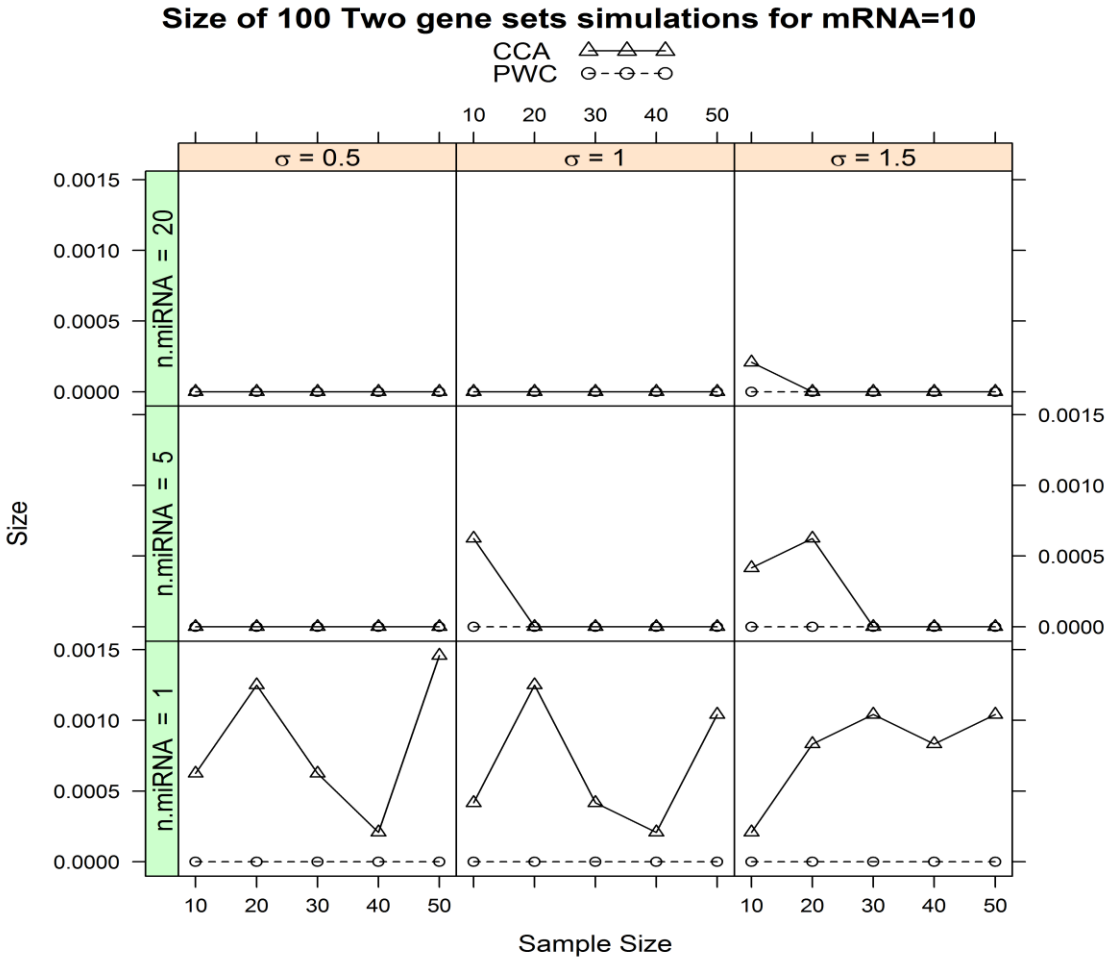


Figure 3.10. The average error rate of both pair wise correlation and SCCA methods over 100 two data sets simulation. In each data set, there were ten correlated mRNAs, where $P_1 = 20, r_2 = 480$. There are 9 plots in the figure, where $\sigma_e = 0.5, \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set, and y-axis indicates the power of detecting the correlated genes by two methods. The dotted lines indicate the power under each sample size for PWC. The solid line indicated power for SCCA method.

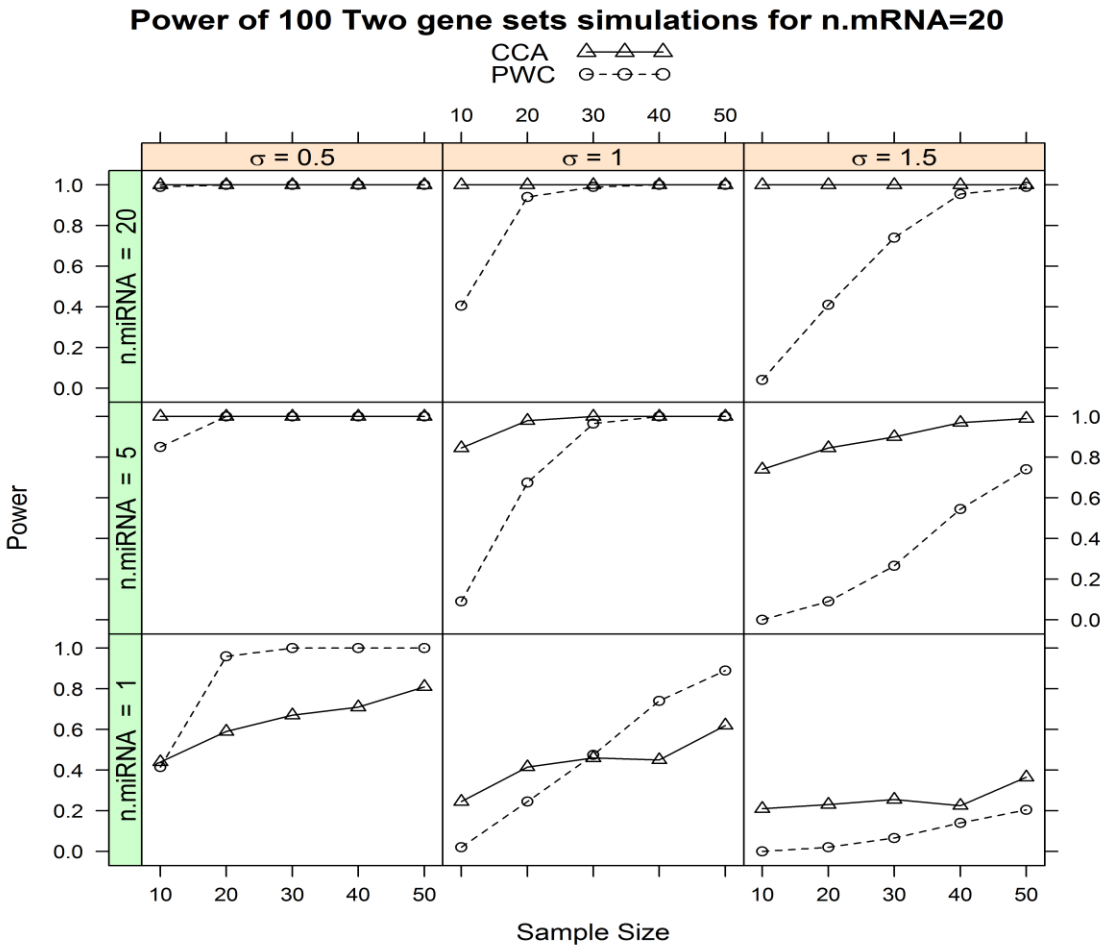


Figure 3.11. The average power of both pair wise correlation and SCCA methods over 100 two data sets simulation. In each data set, there were twenty correlated mRNAs, where $P_1 = 20, r_2 = 480$. There are 9 plots in the figure, where $\sigma_e = 0.5, \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set, and y-axis indicates the power of detecting the correlated genes by two methods. The dotted lines indicate the power under each sample size for PWC. The solid line indicated power for SCCA method.

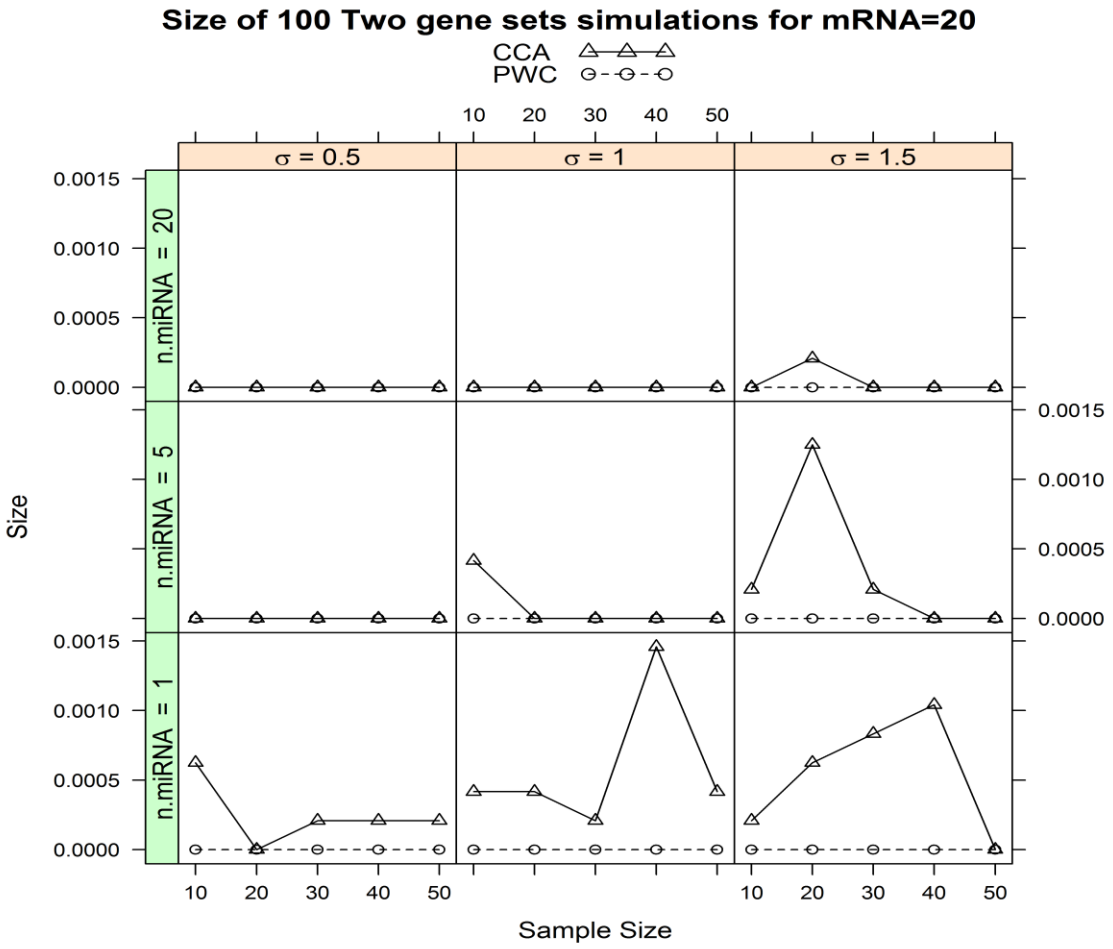


Figure 3.12. The average error rate of both pair wise correlation and SCCA methods over 100 two data sets simulation. In each data set, there were twenty correlated mRNAs, where $P_1 = 20, r_2 = 480$. There are 9 plots in the figure, where $\sigma_e = 0.5, \sigma_e = 1\ and\ \sigma_e = 1.5$ respectively for each column and number of related miRNAs equal to 1, 5 and 20 for each row. In each plot of the figure, x-axis indicates the number of observations in each data set, and y-axis indicates the power of detecting the correlated genes by two methods. The dotted lines indicate the power under each sample size for PWC. The solid line indicated power for SCCA method.

CHAPTER IV

REAL DATA ANALYSIS

4.1. Prostate Cancer

The miRNA and mRNA microarray data sets of human prostate cancer and normal cell we used were obtained from the Broad Institute and downloaded from the database:

(http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi) (Lu, et al., 2005). The original data set contained six prostate cancer tumor and six normal human tissues with both miRNA and mRNA expression. The miRNA data was filtered by the minimum value of 32 and $\log_2$ transformed, so the minimum value was $log_2(32) = 5$. The mRNA data were obtained using Affymetrix Genechips. There were two chips used in the data, Hu35KsubA and Hu6800. The data totally contained 16,063 probes which respectively 8,934 and 7,129 probes in Hu35KsubA and Hu6800 microarrays. Here, we first chose the 8,934 probes from Hu35KsubA chip to do the initial analysis.

In the research, we first filtered the data. For original miRNA data, there were 217 miRNAs and 12 samples (6 tumor and 6 normal). We filtered out the 9 miRNAs which had a variance 0 across the samples. In mRNA expression data, we chose 8,934 mRNAs probes from Hu35KsubA chips. We used the function *nsFilter* within R package *genefilter* to filter the mRNA data by removing duplicate probes mapping to the same Entrez Gene ID (the probe with the highest variance across the samples was retained) and probes with a variance below the 50^{th} percentile. After filtering 2,917 probes remained.

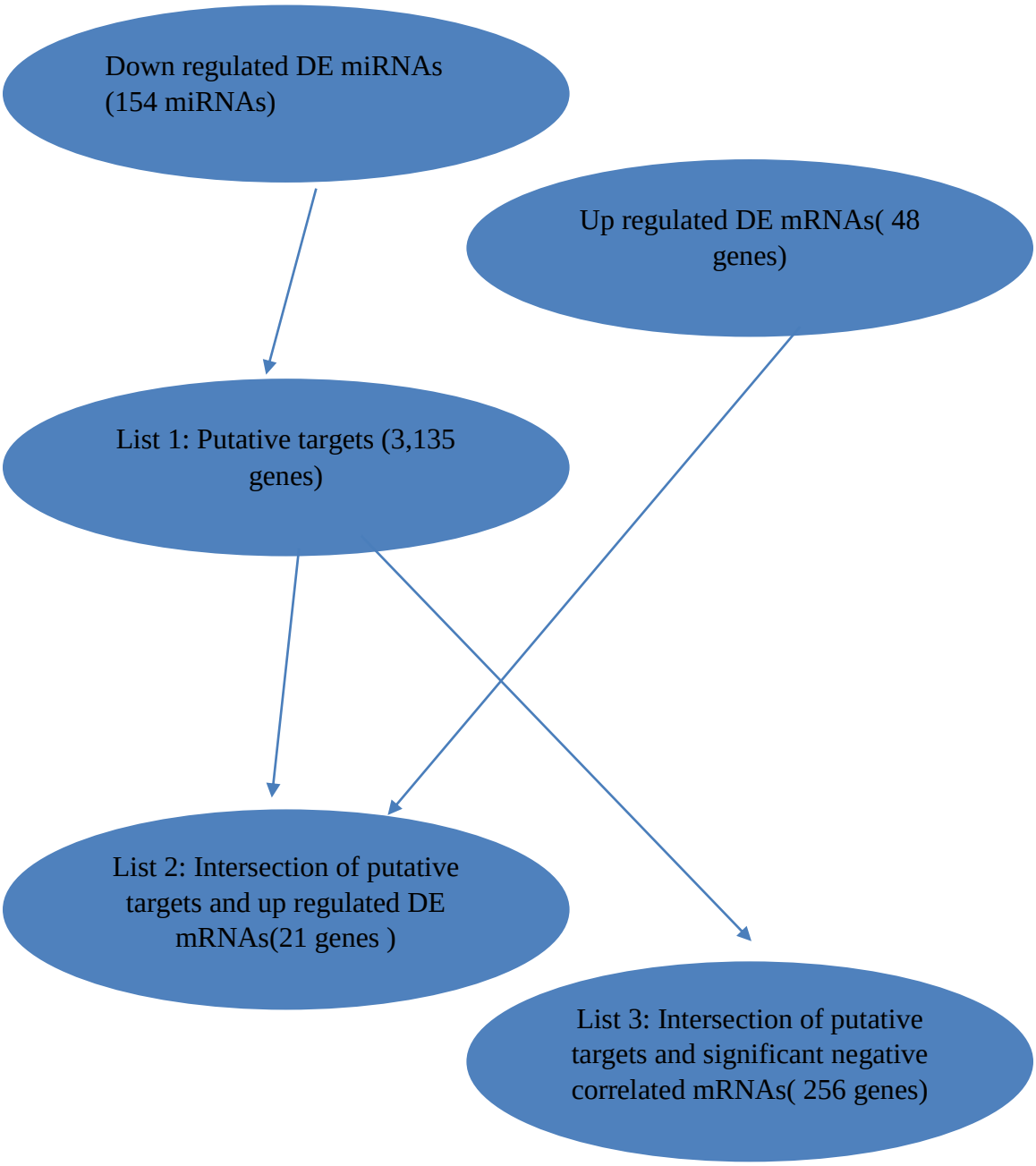
Second, we determined down regulated differentially expressed miRNAs between normal and tumor samples using the empirical Bayes method in R package *limma* (Ritchie, et al., 2015) and (Smyth, 2004). We identified 154 significant miRNAs with adjusted *p*-values (based on Benjamini-Hochberg correction) ≤ 0.05. All of the DE miRNAs were down regulated in the tumor tissue.

Third, we determined three lists (see Table 4.1 and Figure 4.1) of target mRNAs which we would use in the later gene-set enrichment analysis with DAVID. The first list consisted of putative target genes of the DE miRNAs, based on the intersection of targets in the miRBase (Kozomara and Griffiths-Jones, 2014) and TargetScan (Lewis, et al., 2003) databases. This resulted in 3,135 putative target genes of the down regulated miRNAs. The second list consisted of intersecting this putative target list with the up regulated DE genes. We determined up regulated DE genes using the same procedures as for miRNAs. Here we identified 48 up-regulated and DE (adjusted p-value < 0.05) mRNAs. The intersection of this list with the list of putative target genes resulted in 21 total genes. The third list was the overlapped genes between putative target mRNAs of miRNAs and significant genes obtained by pair wise correlation method. Here we identified 256 mRNAs with significant correlation (adjusted p-value < 0.05) between miRNAs and mRNA. The intersection of this list with the list of putative target genes resulted in 256 total genes.

Table 4.1. Description of method of obtaining gene lists

Method	Brief Description
List 1: Putative target mRNAs for down regulated DE miRNAs	Obtained list of putative targets from down regulated DE miRNAs by intersecting putative targets from miRBase and TargetScan. Uploaded this gene list to DAVID for gene set analysis
List 2: Intersection of putative targets (List 1) and up regulated DE mRNAs	Obtained gene list by intersecting putative target mRNAs (List1) and up regulated DE mRNAs of the data set. Uploaded this gene list to DAVID for gene set analysis
List 3: Intersection of putative targets (List 1) and significant mRNAs by PWC method	Obtained gene list by intersecting putative target mRNAs (List1) and significant negative correlated mRNAs detected by PWC method with adjust p-value less than 0.05. Uploaded this gene list to DAVID for gene set analysis

Figure 4.1. Flow chart for prostate cancer:



After getting the three lists, we used the SCCA method based on the 154 significant down-regulated miRNAs data and the 2,963 filtered mRNAs expressed data from Hu35KsubA chip. After normalizing each of the matrices so that expression measurements for each miRNA / mRNA had mean zero and standard deviation one, the miRNA data was multiplied by -1. The *CCA.permute* function in package *PMA* (Witten, et al., 2009) was used to determine optimal penalty parameters for SCCA with multiple sets of canonical variables. But in the result, only the first set of canonical variables had significant permuted p-value. So, the first set of canonical variables were used and there were 46 non-zero elements in the u vector, which meant that 31 mRNAs were selected by the SCCA function. And there were 3 non-zero elements in v vector, which indicated 3 miRNAs were selected.

The next step was KEGG pathway analysis with the SCCA GSEA method. We first used the *GeneSetCollection* function within the Bioconductor package *GSEABase* to construct a collection of gene sets of pathways from the KEGG database. There were 217 pathways collected from KEGG. FDR adjusted p-values from these pathways are given in **Table 4.2**, where the permuted p-value was calculated by the self-contained method. For comparison purposes, the results from DAVID analysis of the KEGG database with default parameters based on all 3,315 putative targets, the intersection of these targets with the 48 DE up-regulated genes, and the intersection of these targets with the genes significantly correlated with the miRNAs are given in **Tables 4.3-4.5**. The former is based on a huge number of genes and identifying germane pathways based on this large set is a daunting task. The middle is based on only 21 genes, and returns only a single pathway. The latter contains six KEGG pathways.

Table 4.2. KEGG pathway analysis by First canonical vector and Self-contained test				
KEGG ID	**Pathway**	**Statisitc**	**P-value**	**Adj.P-value**
4974	Protein digestion and absorption	2.724	0.005	0.293
51	Fructose and mannose metabolism	1.908	0.008	0.293
534	Glycosaminoglycan biosynthesis - heparan sulfate	1.731	0.008	0.293
520	Amino sugar and nucleotide sugar metabolism	1.387	0.009	0.293
532	Glycosaminoglycan biosynthesis - chondroitin sulfate	1.168	0.016	0.293
4672	Intestinal immune network for IgA production	0.330	0.017	0.293
4973	Carbohydrate digestion and absorption	1.369	0.018	0.293
100	Steroid biosynthesis	1.085	0.022	0.293
4970	Salivary secretion	1.433	0.022	0.293
5332	Graft-versus-host disease	0.230	0.026	0.293
5221	Acute myeloid leukemia	0.895	0.034	0.293
4210	Apoptosis	0.844	0.035	0.293
5020	Prion diseases	0.165	0.036	0.293
5218	Melanoma	0.821	0.038	0.293
4664	Fc epsilon RI signaling pathway	0.798	0.043	0.293
5213	Endometrial cancer	0.722	0.044	0.293
5214	Glioma	0.757	0.044	0.293
5223	Non-small cell lung cancer	0.722	0.045	0.293
5142	Chagas disease (American trypanosomiasis)	0.610	0.048	0.293
5210	Colorectal cancer	0.691	0.048	0.293

Table 4.2. KEGG pathway found by SCCA and GSEA method in Prostate Cancer data. We list first 20 pathways with p-value less than 0.05.

Table 4.3. KEGG Pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
hsa04144:Endocytosis	2.10	p<0.001	0.001
hsa04722:Neurotrophin signaling pathway	2.33	p<0.001	0.001
hsa04310:Wnt signaling pathway	2.17	p<0.001	0.001
hsa04330:Notch signaling pathway	2.87	0.001	0.028
hsa04910:Insulin signaling pathway	1.93	0.001	0.039
hsa04120:Ubiquitin mediated proteolysis	1.90	0.001	0.040
hsa04930:Type II diabetes mellitus	2.46	0.007	0.16
hsa05200:Pathways in cancer	1.44	0.007	0.14
hsa05211:Renal cell carcinoma	2.07	0.011	0.19
hsa04010:MAPK signaling pathway	1.45	0.015	0.23
hsa04012:ErbB signaling pathway	1.89	0.015	0.21
hsa00510:N-Glycan biosynthesis	2.31	0.017	0.21
hsa05210:Colorectal cancer	1.84	0.024	0.27
hsa04916:Melanogenesis	1.75	0.024	0.25
hsa04520:Adherens junction	1.88	0.025	0.24
hsa00730:Thiamine metabolism	4.82	0.041	0.36
hsa04530:Tight junction	1.51	0.06	0.44
hsa04140:Regulation of autophagy	2.21	0.06	0.46
hsa05215:Prostate cancer	1.63	0.07	0.48
hsa04070:Phosphatidylinositol signaling system	1.70	0.08	0.48

Table 4.3. KEGG pathways found by DAVID online software in Prostate Cancer data. We analyzed predicted target gene list of 154 significant down regulated miRNAs by DAVID. The putative target list contained 3,135 genes. We list first 20 pathways.

Table 4.4. KEGG pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
hsa04150:mTOR signaling pathway	27.94	0.06	0.90

Table 4.4. KEGG pathway found by DAVID online software in Prostate Cancer data. We analyzed intersection gene list of predicted target gene list and up-regulated mRNAs by DAVID. The intersection list contained 21 genes. There was only one pathway found.

Table 4.5. KEGG pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
hsa04330:Notch signaling pathway	8.07	0.003	0.27
hsa04150:mTOR signaling pathway	7.30	0.004	0.21
hsa04910:Insulin signaling pathway	3.94	0.008	0.24
hsa05200:Pathways in cancer	2.31	0.025	0.48
hsa05220:Chronic myeloid leukemia	4.05	0.07	0.80
hsa05210:Colorectal cancer	3.61	0.09	0.83

Table 4.5. KEGG pathway found by DAVID online software in Prostate Cancer data. We analyzed intersection gene list of predicted target gene list and mRNAs found by pair wise correlation method of 154 significant down regulated miRNAs by DAVID. The intersection list contained 256 genes and there were six pathways been found.

Then, we produced a similar analysis based on the GO database. FDR adjusted p-values for these pathways corresponding to the GSEA statistic resulting from the SCCA method are given in **Table 4.6**. For comparison purposes, the results from DAVID analysis of the GO database with default parameters based on all 3,135 putative targets, the intersection of these targets with the 48DE up-regulated genes and the intersection of putative target genes with significant pair-wise correlations are given in **Tables 4.7- 4.9**.

Table 4.6. GO pathway analysis by First canonical vector and Self-contained test

GO ID	GO Term	Statisitc	P-value	Adj.P-value
GO:0016874	ligase activity	4.706	p<0.001	0.327
GO:0018024	histone-lysine N-methyltransferase activity	6.489	p<0.001	0.372
GO:0004514	nicotinate-nucleotide diphosphorylase (carboxylating) activity	7.708	0.001	0.372
GO:0006469	negative regulation of protein kinase activity	4.412	0.001	0.372
GO:0030278	regulation of ossification	6.947	0.001	0.372
GO:0007422	peripheral nervous system development	5.632	0.002	0.372
GO:0008045	motor neuron axon guidance	5.632	0.002	0.372
GO:0019674	NAD metabolic process	5.416	0.002	0.372
GO:0035284	brain segmentation	6.995	0.002	0.372
GO:0071320	cellular response to cAMP	5.672	0.002	0.372
GO:0071371	cellular response to gonadotropin stimulus	6.947	0.002	0.372
GO:0009435	NAD biosynthetic process	5.381	0.003	0.372
GO:0043679	axon terminus	5.302	0.003	0.372
GO:0048168	regulation of neuronal synaptic plasticity	4.912	0.003	0.372
GO:0001102	RNA polymerase II activating transcription factor binding	3.982	0.004	0.372
GO:0006611	protein export from nucleus	4.843	0.004	0.372
GO:0031105	septin complex	4.558	0.004	0.372
GO:0033147	negative regulation of intracellular estrogen receptor signaling pathway	5.834	0.004	0.372
GO:0007611	learning or memory	4.332	0.005	0.372
GO:0031625	ubiquitin protein ligase binding	2.748	0.005	0.372

Table 4.6. Gene Ontology (GO) terms found by SCCA and GSEA method in Prostate Cancer data. We list first 20 GO terms with p-value less than 0.05.

Table 4.7. GO analysis by DAVID			
GO Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0046907~intracellular transport	1.79	p<0.001	1.61E-08
GO:0044451~nucleoplasm part	1.86	p<0.001	1.61E-08
GO:0031981~nuclear lumen	1.48	p<0.001	1.09E-08
GO:0016568~chromatin modification	2.16	p<0.001	9.94E-07
GO:0005654~nucleoplasm	1.59	p<0.001	2.59E-07
GO:0008104~protein localization	1.57	p<0.001	2.19E-06
GO:0030163~protein catabolic process	1.70	p<0.001	1.69E-06
GO:0044265~cellular macromolecule catabolic process	1.63	p<0.001	2.02E-06
GO:0051603~proteolysis involved in cellular protein catabolic process	1.70	p<0.001	1.72E-06
GO:0015031~protein transport	1.61	p<0.001	1.99E-06
GO:0044257~cellular protein catabolic process	1.70	p<0.001	1.74E-06
GO:0045184~establishment of protein localization	1.60	p<0.001	1.57E-06
GO:0043632~modification-dependent macromolecule catabolic process	1.69	p<0.001	3.89E-06
GO:0019941~modification-dependent protein catabolic process	1.69	p<0.001	3.89E-06
GO:0005794~Golgi apparatus	1.56	p<0.001	1.58E-06
GO:0009057~macromolecule catabolic process	1.58	p<0.001	3.72E-06
GO:0070013~intracellular organelle lumen	1.36	p<0.001	1.45E-06
GO:0070727~cellular macromolecule localization	1.82	p<0.001	6.64E-06
GO:0000123~histone acetyltransferase complex	3.91	p<0.001	2.70E-06
GO:0034613~cellular protein localization	1.81	p<0.001	9.07E-06

Table 4.7. Gene Ontology (GO) found by DAVID online software in Prostate Cancer data. We analyzed predicted target gene list of 154 significant down regulated miRNAs by DAVID. The putative target list contained 3,135 genes. We list first 20 GO with p-value less than 0.05.

Table 4.8. GO analysis By DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0004672~protein kinase activity	4.51	0.049	0.99
GO:0006468~protein amino acid phosphorylation	4.27	0.06	1.00
GO:0005624~membrane fraction	3.72	0.08	0.99
GO:0043066~negative regulation of apoptosis	6.03	0.08	1.00
GO:0043069~negative regulation of programmed cell death	5.95	0.08	1.00
GO:0060548~negative regulation of cell death	5.93	0.08	1.00
GO:0005626~insoluble fraction	3.58	0.08	0.94
GO:0016310~phosphorylation	3.56	0.09	1.00
GO:0005643~nuclear pore	19.03	0.09	0.88
GO:0032553~ribonucleotide binding	2.23	0.10	0.99
GO:0032555~purine ribonucleotide binding	2.23	0.10	0.99

Table 4.8. Gene Ontology (GO) found by DAVID online software in Prostate Cancer data. We analyzed intersection gene list of predicted target gene list and up-regulated mRNAs by DAVID. The intersection list contained 21 genes. There were 11 GO pathways found.

Table 4.9. GO analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0008104~protein localization	2.58	p<0.001	0.002
GO:0015031~protein transport	2.72	p<0.001	0.001
GO:0045184~establishment of protein localization	2.69	p<0.001	0.001
GO:0034613~cellular protein localization	3.02	p<0.001	0.036
GO:0070727~cellular macromolecule localization	3.00	p<0.001	0.032
GO:0046907~intracellular transport	2.42	p<0.001	0.05
GO:0006886~intracellular protein transport	2.95	p<0.001	0.07
GO:0046320~regulation of fatty acid oxidation	12.33	p<0.001	0.12
GO:0005794~Golgi apparatus	2.06	p<0.001	0.22
GO:0006605~protein targeting	3.53	0.001	0.18
GO:0051028~mRNA transport	5.55	0.002	0.22
GO:0051236~establishment of RNA localization	4.98	0.003	0.33
GO:0050658~RNA transport	4.98	0.003	0.33
GO:0050657~nucleic acid transport	4.98	0.003	0.33
GO:0006403~RNA localization	4.83	0.003	0.34
GO:0050872~white fat cell differentiation	29.58	0.004	0.39
GO:0031981~nuclear lumen	1.63	0.005	0.49
GO:0019217~regulation of fatty acid metabolic process	7.04	0.005	0.45
GO:0015931~nucleobase, nucleoside, nucleotide and nucleic acid transport	4.28	0.006	0.46
GO:0000398~nuclear mRNA splicing, via spliceosome	3.61	0.007	0.48

Table 4.9. Gene Ontology (GO) found by DAVID online software in Prostate Cancer data. We analysed intersection gene list of predicted target gene list and mRNAs found by pair wise correlation method of 154 significant down regulated miRNAs by DAVID. The intersection list contained 256 genes and we list first 20 GO pathways with p-value less than 0.05.

After we analyzed the 8,934 probes from the Hu35KsubA chip, we did a similar analysis on the 7,129 probes in the Hu6800 microarrays.

After filtering 7,129 probes in Hu6800 microarrays with function *nsFilter* within R package *genefilter*, we had 2,643 probes left. Then, we used same procedures with Hu35KsubA chip analysis to determined three lists of targets mRNAs. The first list consisted of 2,234 putative target genes of the DE miRNAs. The second list consisted of 23 intersected genes between putative targets and the 68 up regulated DE genes. The third list was based on 233 overlapped genes between putative target mRNAs of miRNAs and mRNAs that had significant inverse correlation with the differentially expressed miRNAs.

After getting the three lists, we applied the SCCA method based on the 154 significant down-regulated miRNAs data and all the 2,643 filtered mRNAs from the Hu6800 chip. In the result, only the first set of canonical variables had significant p-value, so, with the first set canonical variables there were 49 non-zero elements in the $\boldsymbol{u}$ vector, which meant that 49 mRNAs were selected by the SCCA function. And there were 2 non-zero elements in $\boldsymbol{v}$ vector, which indicated 2 miRNAs were selected.

The next step was KEGG pathway analysis with the SCCA- GSEA method. The pathways are given in **Table 4.10**. The results from DAVID analysis of the KEGG database are given in **Tables 4.11- 4.13**.

Table 4.10. KEGG pathway analysis by First canonical vector and Self-contained test				
KEGG ID	**Pathway**	**Statisitc**	**P-value**	**Adj.P-value**
565	Ether lipid metabolism	4.359	0.01	0.581
592	alpha-Linolenic acid metabolism	2.757	0.01	0.581
3010	Ribosome	2.774	0.01	0.581
3008	Ribosome biogenesis in eukaryotes	3.35	0.01	0.709
4950	Maturity onset diabetes of the young	3.289	0.02	0.763
3013	RNA transport	2.883	0.02	0.763
3022	Basal transcription factors	3.121	0.03	0.790
4150	mTOR signaling pathway	1.613	0.03	0.790

Table 4.10. KEGG pathway found by SCCA and GSEA method in the Prostate Cancer data with Hu68000 chip. We list first 8 pathways with p-value less than 0.05.

Table 4.11. KEGG pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
hsa05200:Pathways in cancer	1.84	p<0.001	p<0.001
hsa04510:Focal adhesion	2.07	p<0.001	p<0.001
hsa04020:Calcium signaling pathway	2.07	p<0.001	p<0.001
hsa04010:MAPK signaling pathway	1.84	p<0.001	p<0.001
hsa04722:Neurotrophin signaling pathway	2.11	p<0.001	p<0.001
hsa04012:ErbB signaling pathway	2.23	p<0.001	p<0.001
hsa04910:Insulin signaling pathway	1.90	p<0.001	p<0.001
hsa04350:TGF-beta signaling pathway	2.16	p<0.001	p<0.001
hsa04810:Regulation of actin cytoskeleton	1.67	p<0.001	p<0.001
hsa05215:Prostate cancer	2.11	p<0.001	p<0.001
hsa05211:Renal cell carcinoma	2.28	p<0.001	p<0.001
hsa04620:Toll-like receptor signaling pathway	2.03	p<0.001	p<0.001
hsa04720:Long-term potentiation	2.26	p<0.001	p<0.001
hsa04512:ECM-receptor interaction	2.10	p<0.001	p<0.001
hsa04070:Phosphatidylinositol signaling system	2.16	p<0.001	p<0.001
hsa05214:Glioma	2.26	p<0.001	p<0.001
hsa04660:T cell receptor signaling pathway	1.90	p<0.001	p<0.001
hsa05216:Thyroid cancer	2.95	p<0.001	p<0.001
hsa04730:Long-term depression	2.15	p<0.001	p<0.001
hsa05221:Acute myeloid leukemia	2.26	p<0.001	p<0.001

Table 4.11. KEGG pathway found by DAVID online software in Prostate Cancer data with Hu68000 chip. We analyzed predicted target gene list of 154 significant down regulated miRNAs by DAVID. The putative target list contained 2,234 genes. We list first 20 pathways.

Table 4.12. KEGG pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-Value**	**Benjamini**
hsa04020:Calcium signaling pathway	27.94	0.06	0.74

Table 4.12. KEGG pathway found by DAVID online software in Prostate Cancer data with Hu68000 chip. We analyzed intersection gene list of predicted target gene list and up-regulated mRNAs by DAVID. The intersection list contained 23 genes. There was only one pathway discovered.

Table 4.13. KEGG pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-Value**	**Benjamini**
hsa04020:Calcium signaling pathway	2.74	0.016	0.84
hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)	3.52	0.05	0.95
hsa04142:Lysosome	2.74	0.06	0.93
hsa04722:Neurotrophin signaling pathway	2.59	0.08	0.91
hsa05010:Alzheimer's disease	2.30	0.08	0.86
hsa05215:Prostate cancer	3.01	0.08	0.81

Table 4.13. KEGG pathway found by DAVID online software in Prostate Cancer data with Hu68000 chip. We analyzed intersection gene list of predicted target gene list and mRNAs found by pair wise correlation method of 154 down regulated miRNAs by DAVID. The intersection list contained 233 genes and there were six pathways been found.

After KEGG pathway analysis, we did a similar analysis on GO terms based on the Hu6800 microarrays. FDR adjusted p-values from these pathways are given in **Table 4.14**. For comparison purposes, the results from DAVID analysis of the GO database with default parameters are given in **Tables 4.15**- **4.17** for all putative targets, all putative targets intersected with significantly up-regulated mRNAs, and all putative targets that were significantly inversely correlated with the down-regulated miRNAs.

Table 4.14. GO pathway analysis by First canonical vector and Self-contained test				
GO ID	**GO Term**	**Statisitc**	**P-value**	**Adj.P-value**
GO:0002039	p53 binding	5.729	p<0.001	p<0.001
GO:0006413	translational initiation	4.969	p<0.001	p<0.001
GO:0019003	GDP binding	6.519	p<0.001	0.156
GO:0019068	virion assembly	9.036	p<0.001	0.156
GO:0019082	viral protein processing	9.036	p<0.001	0.156
GO:0044267	cellular protein metabolic process	4.183	p<0.001	0.156
GO:0044822	poly(A) RNA binding	4.151	p<0.001	0.175
GO:0071236	cellular response to antibiotic	8.104	p<0.001	0.175
GO:0075733	intracellular transport of virus	8.322	p<0.001	0.260
GO:0006184	GTP catabolic process	4.825	0.001	0.421
GO:0015031	protein transport	4.764	0.001	0.484
GO:0031901	early endosome membrane	5.105	0.001	0.484
GO:0016197	endosomal transport	5.341	0.002	0.484
GO:0019058	viral life cycle	4.623	0.002	0.484
GO:0030914	STAGA complex	8.003	0.002	0.484
GO:0047497	mitochondrion transport along microtubule	7.709	0.002	0.484
GO:0060675	ureteric bud morphogenesis	9.23	0.002	0.484
GO:0060760	positive regulation of response to cytokine stimulus	8.131	0.002	0.484
GO:1900103	positive regulation of endoplasmic reticulum unfolded protein response	8.551	0.002	0.484
GO:0000139	Golgi membrane	3.636	0.003	0.484

Table 4.14. Gene Ontology (GO) found by SCCA and GSEA method in Prostate Cancer data with Hu68000 chip. We list first 20 of 278 significant GO pathways with p-values less than 0.05.

Table 4.15. GO analysis by DAVID			
GO Term	**Fold Enrichment**	**P-Value**	**Benjamini**
GO:0006357~regulation of transcription from RNA polymerase II promoter	2.14	p<0.001	p<0.001
GO:0044459~plasma membrane part	1.59	p<0.001	p<0.001
GO:0010033~response to organic substance	2.08	p<0.001	p<0.001
GO:0010604~positive regulation of macromolecule metabolic process	1.95	p<0.001	p<0.001
GO:0051173~positive regulation of nitrogen compound metabolic process	2.06	p<0.001	p<0.001
GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.08	p<0.001	p<0.001
GO:0010557~positive regulation of macromolecule biosynthetic process	2.04	p<0.001	p<0.001
GO:0042127~regulation of cell proliferation	1.93	p<0.001	p<0.001
GO:0006793~phosphorus metabolic process	1.82	p<0.001	p<0.001
GO:0006796~phosphate metabolic process	1.82	p<0.001	p<0.001
GO:0009891~positive regulation of biosynthetic process	1.99	p<0.001	p<0.001
GO:0031328~positive regulation of cellular biosynthetic process	1.99	p<0.001	p<0.001
GO:0051254~positive regulation of RNA metabolic process	2.19	p<0.001	p<0.001
GO:0007242~intracellular signaling cascade	1.67	p<0.001	p<0.001
GO:0045941~positive regulation of transcription	2.08	p<0.001	p<0.001
GO:0010628~positive regulation of gene expression	2.06	p<0.001	p<0.001
GO:0045893~positive regulation of transcription, DNA-dependent	2.18	p<0.001	p<0.001
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	2.36	p<0.001	p<0.001
GO:0007167~enzyme linked receptor protein signaling pathway	2.40	p<0.001	p<0.001
GO:0030528~transcription regulator activity	1.57	p<0.001	p<0.001

Table 4.15. Gene Ontology (GO) found by DAVID online software in Prostate Cancer data with Hu68000 chip. We analyzed predicted target gene list of 154 significant down regulated miRNAs by DAVID. The putative target list contained 2,234 genes. We list first 20 GO pathways with p-value less than 0.05.

Table 4.16. GO analysis by DAVID			
Term	**Fold Enrichment**	**P-Value**	**Benjamini**
GO:0030554~adenyl nucleotide binding	3.29	0.005	0.47
GO:0001883~purine nucleoside binding	3.24	0.006	0.29
GO:0001882~nucleoside binding	3.22	0.006	0.21
GO:0000166~nucleotide binding	2.60	0.010	0.26
GO:0055117~regulation of cardiac muscle contraction	147.04	0.013	1.00
GO:0017076~purine nucleotide binding	2.71	0.015	0.31
GO:0005524~ATP binding	3.08	0.016	0.28
GO:0032559~adenyl ribonucleotide binding	3.04	0.017	0.26
GO:0006793~phosphorus metabolic process	3.63	0.018	0.99
GO:0006796~phosphate metabolic process	3.63	0.018	0.99
GO:0051173~positive regulation of nitrogen compound metabolic process	4.57	0.019	0.95
GO:0006809~nitric oxide biosynthetic process	98.03	0.019	0.90
GO:0046209~nitric oxide metabolic process	90.49	0.021	0.87
GO:0031328~positive regulation of cellular biosynthetic process	4.29	0.023	0.84
GO:0009891~positive regulation of biosynthetic process	4.23	0.024	0.81
GO:0045429~positive regulation of nitric oxide biosynthetic process	56.02	0.034	0.87
GO:0006942~regulation of striated muscle contraction	53.47	0.035	0.85
GO:0016310~phosphorylation	3.68	0.038	0.84
GO:0032555~purine ribonucleotide binding	2.47	0.042	0.48
GO:0032553~ribonucleotide binding	2.47	0.042	0.48

Table 4.16. Gene Ontology (GO) found by DAVID online software in Prostate Cancer data with Hu68000 chip. We analyzed intersection gene list of predicted target gene list and up-regulated mRNAs by DAVID. The intersection list contained 23 genes. There were 20 GO pathways were listed here.

Table 4.17. GO analysis by DAVID			
Term	**Fold Enrichment**	**P-Value**	**Benjamini**
GO:0010033~response to organic substance	2.78	p<0.001	p<0.001
GO:0043566~structure-specific DNA binding	6.18	p<0.001	p<0.001
GO:0005829~cytosol	2.30	p<0.001	p<0.001
GO:0006357~regulation of transcription from RNA polymerase II promoter	2.67	p<0.001	0.001
GO:0003690~double-stranded DNA binding	7.25	p<0.001	0.001
GO:0051173~positive regulation of nitrogen compound metabolic process	2.72	p<0.001	0.003
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	3.38	p<0.001	0.004
GO:0042493~response to drug	4.35	p<0.001	0.004
GO:0031328~positive regulation of cellular biosynthetic process	2.56	p<0.001	0.004
GO:0009891~positive regulation of biosynthetic process	2.52	p<0.001	0.004
GO:0042802~identical protein binding	2.60	p<0.001	0.003
GO:0045893~positive regulation of transcription, DNA-dependent	2.89	p<0.001	0.006
GO:0051254~positive regulation of RNA metabolic process	2.86	p<0.001	0.006
GO:0010604~positive regulation of macromolecule metabolic process	2.27	p<0.001	0.007
GO:0010557~positive regulation of macromolecule biosynthetic process	2.49	p<0.001	0.008
GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.51	p<0.001	0.009
GO:0048732~gland development	5.10	p<0.001	0.009
GO:0045941~positive regulation of transcription	2.55	p<0.001	0.013
GO:0016564~transcription repressor activity	3.24	p<0.001	0.015
GO:0010628~positive regulation of gene expression	2.48	p<0.001	0.019

Table 4.17. Gene Ontology (GO) found by DAVID online software in Prostate Cancer data with Hu68000 chip. We analyzed intersection gene list of predicted target gene list and mRNAs found by pair wise correlation method of 154 significant down regulated miRNAs by DAVID. The intersection list contained 233 genes and we list first 20 GO pathways with p-value less than 0.05.

4.2. Colon Cancer

The miRNA and mRNA microarray data sets of human colon cancer and normal cell we used were obtained from the Broad Institute and downloaded from the database:

(http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi) (Ting, et al., 2005)

The original data set contained seven colon cancer tumor and four normal human tissues for both miRNA and mRNA expression data. The miRNA data was filtered by minimum value 32 and log2 transformed, so, the minimum value was $log_2(32) = 5$. The mRNA data was obtained by using Affymetrix GENECHIP analysis software. There were 2 chips Hu35KsubA and Hu6800 in the data. The mRNA contained 16,063 probes which was 8,934 and 7,129 probes respectively in each chip.

The first part of the analysis was on the 8,934 Hu35KsubA probes. In the paper, we first filtered out the 18 miRNAs which variances were 0, and there were 199 miRNAs left. Then, we applied the function *nsFilter* within R package *genefilter* on the mRNA data. After filtering, 2,917 features were left.

Second, we determined significant down regulated differentially expressed miRNAs between tumor and normal samples. We first made a contrast comparing tumor to normal samples, second fitted the linear model to estimate the contrast by the *lmFit* function under the *limma* package, then used the empirical Bayes method to compute moderated t-statistics and the log-odds of differential expression. We identified 26 significant DE miRNAs with adjusted p-values of moderated t-statistics less than or equal to 0.05 with a log fold change less than zero. The 26 DE miRNAs were down regulated in the tumor tissue.

Third, we determined three lists (similarly to the prostate cancer data in **Section 4.1**) of target mRNAs which we would use in the gene set enrichment analysis with DAVID. The first list was consisted of target mRNAs of differentially expressed miRNAs. At last, we got 1,354 overlapped putative targeting genes in the list.

The second list consisted of the intersection of putative target mRNAs and up regulated differentially expressed genes in the tumor tissues. We determined up regulated differentially expressed mRNA using the similar procedures when we determined down regulated differentially expressed miRNAs. We identified 200

up regulated differentially expressed mRNAs. Then we discovered 36 overlapped genes between up regulated DE mRNAs and the list of putative gene targets.

The third list was the overlapped genes between putative target mRNAs of miRNAs and the target mRNAs which had significant inverse correlation with the down-regulated miRNAs. We got 100 mRNAs with FDR adjusted p-value of pair wise correlation between miRNAs and mRNAs less or equal to 0.05. And all of these 100 genes were contained in the putative target list.

Third, we applied the SCCA method based on the 26 significantly down regulated miRNAs in tumor tissues. and the 2,963 filtered mRNA expression data from the Hu35KsubA chip. We first normalized both miRNA and mRNA data sets so that each miRNA/ mRNA expression vector had mean 0 and standard deviation 1, then multiplied the miRNA data by -1. In the SCCA procedure, we first used the function *CCA.permute* in package *PMA(Witten, et al., 2009)* to determine the optimal penalties to be used in the *CCA* function for obtaining multiple sets of canonical variables. In the result, only first set of canonical variables had significant permuted p-values, so, penalties of 0.1 would be used for both mRNA and miRNA expression data. Then, we applied the *CCA* function with the mRNA and miRNA expression data. In the result, there were 390 non-zero elements in $\boldsymbol{u}$ vector, which meant that 390 mRNAs were selected by the SCCA function. And there were 3 non-zero elements in the $\boldsymbol{v}$ vector, which indicated that 3 miRNAs were identified.

The next step was KEGG pathway analysis with the SCCA- GSEA method. We first used the function *GeneSetCollection* within the R package *GSEABase* to construct a collection of gene sets from pathways in the KEGG database. There were 212 gene sets of pathways collected from KEGG. FDR adjusted p-values from these pathways are given in **Table 4.18**. For comparison, DAVID analysis of the KEGG database with default parameters based on the three gene lists we created are given in **Tables 4.19- 4.21**. The former is based on a 1,354 genes and identifies a large number of pathways. The middle is based on only 36 genes, and returns only two pathways. The latter contains only one KEGG pathway. Then, we produced a similar analysis based on GO terms. The result is in the **Table 4.22** for the SCCA-GSEA method. The results of the three gene lists analyzed by DAVID are given in **Tables 4.23- 4.25**.

Table 4.18. KEGG pathway analysis by First canonical vector and Self-contained test				
ENTREZID	**Pathway**	**Statisitc**	**P-value**	**Adj.P-value**
512	Mucin type O-Glycan biosynthesis	3.96	p<0.001	p<0.001
4614	Renin-angiotensin system	0.90	p<0.001	p<0.001
5214	Glioma	3.02	0.00	0.03
4310	Wnt signaling pathway	3.12	0.00	0.06
4720	Long-term potentiation	4.13	0.00	0.09
4740	Olfactory transduction	4.80	0.00	0.09
4971	Gastric acid secretion	3.49	0.01	0.10
4114	Oocyte meiosis	2.59	0.01	0.16
4912	GnRH signaling pathway	2.53	0.01	0.18
4012	ErbB signaling pathway	2.45	0.01	0.18
4916	Melanogenesis	2.37	0.01	0.18
4950	Maturity onset diabetes of the young	1.83	0.02	0.22
3040	Spliceosome	2.07	0.02	0.24
4722	Neurotrophin signaling pathway	2.00	0.02	0.28
982	Drug metabolism - cytochrome P450	0.99	0.03	0.34
4020	Calcium signaling pathway	2.02	0.03	0.34
830	Retinol metabolism	0.87	0.04	0.35

Table 4.18. KEGG pathway found by SCCA and GSEA method in Colon Cancer data. There were 18 pathways with p-value less than 0.05.

Table 4.19. KEGG pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
hsa04144:Endocytosis	2.55	p<0.001	0.015
hsa04910:Insulin signaling pathway	2.69	p<0.001	0.032
hsa04330:Notch signaling pathway	4.09	p<0.001	0.06
hsa04310:Wnt signaling pathway	2.41	p<0.001	0.05
hsa04360:Axon guidance	2.48	p<0.001	0.07
hsa04722:Neurotrophin signaling pathway	2.41	p<0.001	0.10
hsa05210:Colorectal cancer	2.80	p<0.001	0.10
hsa04930:Type II diabetes mellitus	3.64	p<0.001	0.09
hsa04012:ErbB signaling pathway	2.70	p<0.001	0.10
hsa04370:VEGF signaling pathway	2.85	p<0.001	0.10
hsa05211:Renal cell carcinoma	2.75	p<0.001	0.17
hsa04120:Ubiquitin mediated proteolysis	2.03	p<0.001	0.25
hsa04520:Adherens junction	2.50	p<0.001	0.24
hsa04660:T cell receptor signaling pathway	2.18	p<0.001	0.24
hsa04666:Fc gamma R-mediated phagocytosis	2.25	p<0.001	0.26
hsa05213:Endometrial cancer	2.88	p<0.001	0.25
hsa04150:mTOR signaling pathway	2.88	p<0.001	0.25
hsa04920:Adipocytokine signaling pathway	2.55	p<0.001	0.25
hsa04720:Long-term potentiation	2.51	p<0.001	0.25
hsa04115:p53 signaling pathway	2.51	p<0.001	0.25

Table 4.19. KEGG pathway found by DAVID online software in Colon Cancer data. We analyzed predicted target gene list of 28 down regulated miRNAs by DAVID. The putative target list contained 1,354 genes. We list first 20 pathways with p-value less than 0.05.

Table 4.20. KEGG pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
hsa04120:Ubiquitin mediated proteolysis	22.27	0.00	0.02
hsa04330:Notch signaling pathway	43.28	0.04	0.07

Table 4.20. KEGG pathway found by DAVID online software in Colon Cancer data. We analyzed intersection gene list of predicted target gene list and up-regulated mRNAs by DAVID. The intersection list contained 36 genes. There was only two pathway discovered.

Table 4.21. KEGG pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
hsa04120:Ubiquitin mediated proteolysis	5.57	0.09	0.89

Table 4.21. KEGG pathway found by DAVID online software in Colon Cancer data. We analyzed intersection gene list of predicted target gene list and mRNAs found by pair wise correlation method of 26 down regulated miRNAs by DAVID. The intersection list contained 100 genes and there was only one pathway had been found.

Table 4.22. GO pathway analysis by First canonical vector and Self-contained test				
GO ID	**GO Term**	**Statisitc**	**P-Value**	**Adj.P-value**
GO:0014733	regulation of skeletal muscle adaptation	6.79	p<0.001	0.33
GO:0051924	regulation of calcium ion transport	8.31	p<0.001	0.33
GO:0090129	positive regulation of synapse maturation	8.31	p<0.001	0.33
GO:0004683	calmodulin-dependent protein kinase activity	5.88	0.001	0.36
GO:0007268	synaptic transmission	3.20	0.001	0.36
GO:0007369	gastrulation	7.16	0.001	0.36
GO:0010976	positive regulation of neuron projection development	5.57	0.001	0.36
GO:0033017	sarcoplasmic reticulum membrane	5.72	0.001	0.36
GO:0048169	regulation of long-term neuronal synaptic plasticity	6.67	0.001	0.36
GO:0060333	interferon-gamma-mediated signaling pathway	4.44	0.001	0.36
GO:0060998	regulation of dendritic spine development	8.08	0.001	0.36
GO:2001235	positive regulation of apoptotic signaling pathway	4.80	0.001	0.36
GO:0050885	neuromuscular process controlling balance	3.91	0.002	0.38
GO:0051233	spindle midzone	4.54	0.002	0.41
GO:0030666	endocytic vesicle membrane	3.93	0.003	0.41
GO:0046686	response to cadmium ion	4.40	0.003	0.41
GO:0072383	plus-end-directed vesicle transport along microtubule	5.96	0.003	0.41
GO:0005815	microtubule organizing center	2.94	0.005	0.41
GO:0006606	protein import into nucleus	4.00	0.005	0.41
GO:0008333	endosome to lysosome transport	4.05	0.005	0.41

Table 4.22. Gene Ontology (GO) found by SCCA and GSEA method in Colon Cancer data. We list first 20 of 181 GO pathways with p-value less than 0.05.

Table 4.23. GO analysis by DAVID			
GO Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0031981~nuclear lumen	1.56	p<0.001	p<0.001
GO:0016564~transcription repressor activity	2.27	p<0.001	0.004
GO:0044451~nucleoplasm part	1.94	p<0.001	0.001
GO:0019898~extrinsic to membrane	1.97	p<0.001	0.002
GO:0016568~chromatin modification	2.28	p<0.001	0.053
GO:0005654~nucleoplasm	1.66	p<0.001	0.003
GO:0005768~endosome	2.19	p<0.001	0.003
GO:0046907~intracellular transport	1.73	p<0.001	0.052
GO:0006325~chromatin organization	2.00	p<0.001	0.041
GO:0010629~negative regulation of gene expression	1.84	p<0.001	0.034
GO:0012505~endomembrane system	1.66	p<0.001	0.005
GO:0000123~histone acetyltransferase complex	4.47	p<0.001	0.009
GO:0043005~neuron projection	2.02	p<0.001	0.010
GO:0019941~modification-dependent protein catabolic process	1.71	p<0.001	0.09
GO:0043632~modification-dependent macromolecule catabolic process	1.71	p<0.001	0.09
GO:0016481~negative regulation of transcription	1.81	p<0.001	0.08
GO:0045184~establishment of protein localization	1.60	p<0.001	0.07
GO:0008104~protein localization	1.54	p<0.001	0.07
GO:0016570~histone modification	2.79	p<0.001	0.06
GO:0008536~Ran GTPase binding	9.45	p<0.001	0.09

Table 4.23. Gene Ontology (GO) found by DAVID online software in Colon Cancer data. We analyzed predicted target gene list of 28 down regulated miRNAs by DAVID. The putative target list contained 1,354 genes. We list first 20 GO with p-value less than 0.05.

Table 4.24. GO analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0008270~zinc ion binding	2.289	0.010	0.581
GO:0016570~histone modification	15.841	0.014	0.990
GO:0016569~covalent chromatin modification	15.338	0.015	0.912
GO:0030900~forebrain development	12.714	0.021	0.902
GO:0030901~midbrain development	80.524	0.023	0.858
GO:0046914~transition metal ion binding	1.899	0.036	0.785
GO:0045665~negative regulation of neuron differentiation	39.042	0.048	0.960
GO:0019941~modification-dependent protein catabolic process	4.489	0.051	0.943
GO:0043632~modification-dependent macromolecule catabolic process	4.489	0.051	0.943
GO:0017015~regulation of transforming growth factor beta receptor signaling pathway	33.035	0.056	0.934
GO:0051603~proteolysis involved in cellular protein catabolic process	4.295	0.056	0.909
GO:0044257~cellular protein catabolic process	4.273	0.057	0.884
GO:0000122~negative regulation of transcription from RNA polymerase II promoter	7.265	0.058	0.861
GO:0048663~neuron fate commitment	30.676	0.060	0.845
GO:0016568~chromatin modification	7.053	0.061	0.824
GO:0030163~protein catabolic process	4.143	0.062	0.801
GO:0000123~histone acetyltransferase complex	26.909	0.068	0.999
GO:0016573~histone acetylation	26.841	0.069	0.813
GO:0031625~ubiquitin protein ligase binding	26.714	0.070	0.865
GO:0046907~intracellular transport	3.922	0.070	0.799

Table 4.24. Gene Ontology (GO) found by DAVID online software in Colon Cancer data. We analyzed intersection gene list of predicted target gene list and up-regulated mRNAs by DAVID. We list first 20 of 27 GO pathways with p-value less than 0.05

Table 4.25. GO analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0000123~histone acetyltransferase complex	16.493	0.002	0.252
GO:0016573~histone acetylation	15.657	0.002	0.806
GO:0006473~protein amino acid acetylation	14.453	0.003	0.644
GO:0008134~transcription factor binding	3.330	0.003	0.427
GO:0043543~protein amino acid acylation	12.526	0.004	0.644
GO:0016570~histone modification	7.700	0.004	0.541
GO:0016569~covalent chromatin modification	7.456	0.004	0.503
GO:0043966~histone H3 acetylation	21.679	0.008	0.668
GO:0016568~chromatin modification	4.114	0.014	0.814
GO:0008015~blood circulation	5.051	0.016	0.813
GO:0003013~circulatory system process	5.051	0.016	0.813
GO:0003712~transcription cofactor activity	3.294	0.018	0.851
GO:0004468~lysine N-acetyltransferase activity	13.486	0.020	0.754
GO:0004402~histone acetyltransferase activity	13.486	0.020	0.754
GO:0015672~monovalent inorganic cation transport	3.545	0.026	0.903
GO:0060177~regulation of angiotensin metabolic process	75.156	0.026	0.881
GO:0002002~regulation of angiotensin levels in blood	75.156	0.026	0.881
GO:0007507~heart development	4.370	0.026	0.860
GO:0006605~protein targeting	4.370	0.026	0.860
GO:0030695~GTPase regulator activity	2.960	0.029	0.781

Table 4.25. Gene Ontology (GO) found by DAVID online software in Colon Cancer data. We analyzed intersection gene list of predicted target gene list and mRNAs found by pair wise correlation method of 26 down regulated miRNAs by DAVID. The intersection list contained 100 genes and we list first 20 of 65 GO with p-value less than 0.05.

The next step was the analysis of the 7,071 probes in the Hu6800 microarrays, the process being identical to the analysis of the Hu35KsubA chip data. First, we identified 26 significant miRNAs with adjusted p-values of moderated t-statistics less than or equal to 0.05 and with a log 2 fold change less than 0, which meant all the DE miRNAs were down regulated in colon tumor tissue relative to normal tissue.

Then, we determined three gene lists (as with the Hu35KsubA chip) of target mRNAs which we would use in the later analysis with DAVID. The first list contained 974 genes, the second list consisted of 36 genes and there were 53 genes in the third list.

After getting the three lists, we used the SCCA method based on the data from the 26 significantly down regulated miRNAs and all the 2,643 filtered mRNA expression data from the Hu6800 chip. We first ran the *CCA* function with gene expressed data. We obtained the ***u*** and ***v*** vectors from the CCA function by the SCCA method, where the ***u*** vector with length 2643 was the canonical vector for mRNA expressed data and ***v*** with length 23 was the canonical vector of miRNA data. In the result, there were 950 non-zero elements in ***u*** vector, which meant that 950 mRNAs were selected by the SCCA function. And 13 non-zero elements in ***v*** vector, which indicated 13 miRNAs were selected.

The next step was KEGG pathway analysis with the gene set enrichment analysis method. The last step of GSEA was to calculate the permutation p-value of overall statistic for each pathway. The result is in **Table 4.26**. The results of the DAVID analysis of the three gene lists are in **Tables 4.27- 4.29**. Then, we conducted a similar analysis based on GO terms. The result is in **Table 4.30** for the SCCA-GSEA method. The results of the DAVID analysis of the three gene lists are in **Tables 4.31- 4.33**.

Table 4.26. KEGGpathway analysis by First canonical vector and Self-contained test				
ENTREZID	**Pathway**	**Statisitc**	**P-value**	**Adj.P-value**
30	Pentose phosphate pathway	2.327	p< 0.001	p< 0.001
270	Cysteine and methionine metabolism	2.274	p< 0.002	p< 0.002
450	Selenocompound metabolism	0.000	p< 0.003	p< 0.003
512	Mucin type O-Glycan biosynthesis	0.000	p< 0.004	p< 0.004
760	Nicotinate and nicotinamide metabolism	1.400	p< 0.005	p< 0.005
3018	RNA degradation	1.697	p< 0.006	p< 0.006
3040	Spliceosome	4.017	p< 0.007	p< 0.007
3050	Proteasome	1.965	p< 0.008	p< 0.008
3450	Non-homologous end-joining	0.631	p< 0.009	p< 0.009
970	Aminoacyl-tRNA biosynthesis	2.637	0.001	0.020
3010	Ribosome	2.738	0.001	0.020
310	Lysine degradation	1.168	0.003	0.050
1040	Biosynthesis of unsaturated fatty acids	1.501	0.003	0.050
480	Glutathione metabolism	1.531	0.009	0.139
740	Riboflavin metabolism	1.108	0.011	0.158
250	Alanine, aspartate and glutamate metabolism	1.987	0.015	0.203
4512	ECM-receptor interaction	1.437	0.016	0.203
10	Glycolysis / Gluconeogenesis	1.228	0.025	0.300
1100	Metabolic pathways	0.962	0.034	0.387
3013	RNA transport	2.317	0.039	0.413

Table 4.26. KEGG pathway found by SCCA and GSEA method in Colon Cancer data with Hu6800 chip. There were first 20 of 24 KEGG pathways with p-value less than 0.05.

Table 4.27. KEGG pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
hsa04010:MAPK signaling pathway	2.16	p<0.001	p<0.001
hsa05014:Amyotrophic lateral sclerosis (ALS)	4.15	p<0.001	p<0.001
hsa04510:Focal adhesion	2.32	p<0.001	p<0.001
hsa04012:ErbB signaling pathway	3.16	p<0.001	p<0.001
hsa04720:Long-term potentiation	3.44	p<0.001	p<0.001
hsa05410:Hypertrophic cardiomyopathy (HCM)	3.07	p<0.001	p<0.001
hsa05200:Pathways in cancer	1.89	p<0.001	p<0.001
hsa04350:TGF-beta signaling pathway	2.84	p<0.001	0.003
hsa05216:Thyroid cancer	4.74	p<0.001	0.003
hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)	2.89	p<0.001	0.005
hsa04360:Axon guidance	2.34	p<0.001	0.005
hsa04930:Type II diabetes mellitus	3.51	p<0.001	0.006
hsa04310:Wnt signaling pathway	2.18	p<0.001	0.006
hsa04910:Insulin signaling pathway	2.24	p<0.001	0.007
hsa05414:Dilated cardiomyopathy	2.54	p<0.001	0.009
hsa04114:Oocyte meiosis	2.37	p<0.001	0.008
hsa04512:ECM-receptor interaction	2.62	p<0.001	0.009
hsa05210:Colorectal cancer	2.62	p<0.001	0.009
hsa05213:Endometrial cancer	3.17	p<0.001	0.009
hsa04722:Neurotrophin signaling pathway	2.22	p<0.001	0.012

Table 4.27. KEGG pathway found by DAVID online software in Colon Cancer data with Hu6800 chip. We analyzed predicted target gene list of 28 down regulated miRNAs by DAVID. The putative target list contained 974 genes. We list the first 20 of 95 KEGG pathways with p-value less than 0.05.

Table 4.28. KEGG pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
hsa04510:Focal adhesion	9.90	p<0.001	p<0.001
hsa04512:ECM-receptor interaction	15.79	p<0.001	p<0.001
hsa04810:Regulation of actin cytoskeleton	4.11	0.06	0.61

Table 4.28. KEGG pathway found by DAVID online software in Colon Cancer data with Hu6800 chip. We analyzed list 2 by DAVID. The gene list contained 36 genes, and three KEGG pathway were found.

Table 4.29. KEGG pathway analysis by DAVID			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
hsa04510:Focal adhesion	7.85	p< 0.001	p< 0.001
hsa04512:ECM-receptor interaction	12.52	p< 0.001	0.003
hsa04810:Regulation of actin cytoskeleton	4.08	0.029	0.49
hsa05211:Renal cell carcinoma	7.51	0.06	0.64

Table 4.29. KEGG pathway found by DAVID online software in Colon Cancer data with Hu6800 chip. We analyzed intersection gene list of predicted target gene list and mRNAs found by pair wise correlation method of 28 down regulated miRNAs by DAVID. The intersection list contained 53 genes and there was only one pathway had been found.

Table 4.30. GO pathway analysis by First canonical vector and Self-contained test				
GO ID	**GO Term**	**Statisitc**	**P-Value**	**Adj.Pvalue**
GO:0000398	mRNA splicing, via spliceosome	4.154	p< 0.001	p< 0.001
GO:0002199	zona pellucida receptor complex	4.119	p< 0.001	p< 0.001
GO:0003697	single-stranded DNA binding	4.158	p< 0.001	p< 0.001
GO:0003723	RNA binding	3.774	p< 0.001	p< 0.001
GO:0005515	protein bindin	1.919	p< 0.001	p< 0.001
GO:0005829	cytosol	2.690	p< 0.001	p< 0.001
GO:0005832	chaperonin-containing T-complex	4.404	p< 0.001	p< 0.001
GO:0006457	protein folding	3.148	p< 0.001	p< 0.001
GO:0007339	binding of sperm to zona pellucida	3.884	p< 0.001	p< 0.001
GO:0007599	hemostasis	4.536	p< 0.001	p< 0.001
GO:0008380	RNA splicing	4.492	p< 0.001	0.029
GO:0009566	fertilization	4.121	p< 0.001	0.029
GO:0010467	gene expression	4.211	p< 0.001	0.029
GO:0010899	regulation of phosphatidylcholine catabolic process	4.385	p< 0.001	0.029
GO:0016032	viral process	3.575	p< 0.001	0.029
GO:0016070	RNA metabolic process	3.625	p< 0.001	0.029
GO:0016192	vesicle-mediated transport	3.804	p< 0.001	0.049
GO:0030529	ribonucleoprotein complex	3.537	p< 0.001	0.049
GO:0044822	poly(A) RNA binding	6.577	p< 0.001	0.049
GO:0051082	unfolded protein binding	4.811	p< 0.001	0.070

Table 4.30. Gene Ontology (GO) found by SCCA and GSEA method in Colon Cancer data with Hu6800 chip. We list first 20 of 292 GO pathways with p-value less than 0.05.

Table 4.31. GO analysis by DAVID			
GO Term	**Fold Enrichment**	**P-Value**	**Benjamini**
GO:0044459~plasma membrane part	1.64	p< 0.001	p< 0.001
GO:0006357~regulation of transcription from RNA polymerase II promoter	2.17	p< 0.001	p< 0.001
GO:0030528~transcription regulator activity	1.68	p< 0.001	p< 0.001
GO:0019226~transmission of nerve impulse	2.61	p< 0.001	p< 0.001
GO:0006796~phosphate metabolic process	1.86	p< 0.001	p< 0.001
GO:0006793~phosphorus metabolic process	1.86	p< 0.001	p< 0.001
GO:0005829~cytosol	1.74	p< 0.001	p< 0.001
GO:0007517~muscle organ development	3.15	p< 0.001	p< 0.001
GO:0007267~cell-cell signaling	2.12	p< 0.001	p< 0.001
GO:0007507~heart development	3.09	p< 0.001	p< 0.001
GO:0007167~enzyme linked receptor protein signaling pathway	2.57	p< 0.001	p< 0.001
GO:0046907~intracellular transport	2.05	p< 0.001	p< 0.001
GO:0045202~synapse	2.60	p< 0.001	p< 0.001
GO:0007268~synaptic transmission	2.65	p< 0.001	p< 0.001
GO:0044057~regulation of system process	2.61	p< 0.001	p< 0.001
GO:0051173~positive regulation of nitrogen compound metabolic process	2.03	p< 0.001	p< 0.001
GO:0005516~calmodulin binding	3.63	p< 0.001	p< 0.001
GO:0048878~chemical homeostasis	2.17	p< 0.001	p< 0.001
GO:0060537~muscle tissue development	3.73	p< 0.001	p< 0.001
GO:0005667~transcription factor complex	3.08	p< 0.001	p< 0.001

Table 4.31. Gene Ontology (GO) found by DAVID online software in Colon Cancer data with Hu6800. We analyzed predicted target gene list of 26 down regulated miRNAs by DAVID. The putative target list contained 974 genes. We list first 20 GO terms with p-value less than 0.05.

Table 4.32. GO analysis by DAVID			
Term	**Fold Enrichment**	**P-Value**	**Benjamini**
GO:0044420~extracellular matrix part	21.14	p< 0.001	0.001
GO:0005578~proteinaceous extracellular matrix	9.02	p< 0.001	0.005
GO:0031012~extracellular matrix	8.37	p< 0.001	0.005
GO:0005201~extracellular matrix structural constituent	18.30	0.001	0.14
GO:0007155~cell adhesion	4.55	0.001	0.49
GO:0022610~biological adhesion	4.54	0.001	0.29
GO:0030198~extracellular matrix organization	15.30	0.002	0.31
GO:0030199~collagen fibril organization	41.16	0.002	0.26
GO:0005581~collagen	35.34	0.003	0.09
GO:0044421~extracellular region part	3.44	0.006	0.14
GO:0043062~extracellular structure organization	9.76	0.007	0.54
GO:0032964~collagen biosynthetic process	159.15	0.012	0.67
GO:0007015~actin filament organization	16.58	0.013	0.64
GO:0046164~alcohol catabolic process	14.74	0.017	0.68
GO:0030036~actin cytoskeleton organization	7.04	0.017	0.65
GO:0005198~structural molecule activity	3.72	0.018	0.67
GO:0007160~cell-matrix adhesion	13.41	0.020	0.66
GO:0030029~actin filament-based process	6.60	0.021	0.64
GO:0007010~cytoskeleton organization	4.56	0.021	0.61
GO:0030674~protein binding, bridging	12.56	0.022	0.60

Table 4.32. Gene Ontology (GO) found by DAVID online software in Colon Cancer data with Hu6800 chip. We analyzed intersection gene list of predicted target gene list and up-regulated mRNAs of 26 down regulated miRNAs by DAVID. The intersection list contained 36 genes. We list 20 GO with p-value less than 0.05.

Table 4.33. GO analysis by DAVID			
Term	**Fold Enrichment**	**P-Value**	**Benjamini**
GO:0044420~extracellular matrix part	15.24	p< 0.001	0.006
GO:0005581~collagen	33.97	p< 0.001	0.015
GO:0005578~proteinaceous extracellular matrix	6.50	0.001	0.029
GO:0031012~extracellular matrix	6.03	0.001	0.032
GO:0007155~cell adhesion	3.55	0.003	0.87
GO:0022610~biological adhesion	3.54	0.003	0.65
GO:0005201~extracellular matrix structural constituent	13.42	0.003	0.35
GO:0007264~small GTPase mediated signal transduction	5.43	0.004	0.64
GO:0030199~collagen fibril organization	28.56	0.005	0.56
GO:0030198~extracellular matrix organization	10.62	0.006	0.56
GO:0016071~mRNA metabolic process	4.48	0.010	0.68
GO:0003697~single-stranded DNA binding	15.74	0.015	0.66
GO:0005178~integrin binding	14.67	0.017	0.56
GO:0032964~collagen biosynthetic process	110.43	0.018	0.83
GO:0043062~extracellular structure organization	6.78	0.020	0.83
GO:0016564~transcription repressor activity	4.57	0.022	0.54
GO:0005829~cytosol	2.24	0.026	0.56
GO:0006397~mRNA processing	4.30	0.027	0.88
GO:0007015~actin filament organization	11.50	0.027	0.85
GO:0008544~epidermis development	6.00	0.027	0.83

Table 4.33. Gene Ontology (GO) found by DAVID online software in Colon Cancer data with Hu6800. We analyzed intersection gene list of predicted target gene list and mRNAs found by pair wise correlation method of 26 down regulated miRNAs by DAVID. The intersection list contained 53 genes and we list first 20 GO with p-value less than 0.05.

4.3 Birth Defects Center, Dental school neural tube data

The miRNA and mRNA microarray data sets from the murine embryonic neural tube (NT) development study contained four 8.5- NT- arrays, four 9.0-NT-arrays and four 9.5-NT- arrays for both miRNA and mRNA expression data. The data were collected at three gestational days (GD), 8.5, 9.0, and 9.5. The miRNA expression data was obtained by using AffyBatch analysis software and the annotation of the data was *mirna20*. The number of samples in the miRNA data was 12 and the number of miRNA genes was 20,706. The mRNA data totally contained 12 samples and 45,101 features. The mRNA data was obtained by using Affymetrix GENECHIP analysis software. The annotation of the data was *mouse4302*.

In the analysis, we first filtered the data set. For original miRNA data, there were 20,706 miRNAs and 12 samples. We kept the miRNAs for which the name of genes started with "mmu-". As the result of filtering, there were 1,412 genes left. In the mRNA expression data, we used the function *affy::rma* within R package *pd.mirna.2.0* to filter the mRNA data by removing duplicate probes mapping to the same Entrez Gene ID (the probe with the highest variance across the samples was retained) and probes with a variance below the 50^{th} percentile. After filtering there were 10,336 probes remaining.

Second, we determined differentially expressed miRNAs between 9.5- NT and 8.5- NT arrays using the empirical Bayes method in the R package *limma* (Ritchie, et al., 2015) (Smyth, 2004). We identified 183 significant miRNAs with adjusted *p*-values (based on Benjamini-Hochberg correction) ≤ 0.05. 52 out of 183 miRNAs with positive log fold change were up regulated on GD 9.5 relative to GD 8.5, and other 131 miRNAs with negative log fold changes are down regulated.

Third, we determined three lists of target mRNAs which we would use in the later analysis with DAVID. The first part was the result of 52 up regulated miRNAs. The first list consisted of putative target genes of the DE miRNAs, based on the intersection of targets in the miRBase (Kozomara and Griffiths-Jones, 2014) and TargetScan (Lewis, et al., 2003) databases. This resulted in 5,154 putative target genes of the up regulated miRNAs. The second list consisted of intersecting this putative target list with the up regulated DE genes. We determined up regulated DE genes using the same procedures as for the miRNAs. Here we identified 3,037 down regulated and DE (adjusted p-value < 0.05) mRNAs. The intersection of this list with the list of putative target genes resulted in 530 total genes. The third list was the overlapped genes between putative

target mRNAs of miRNAs and significant genes obtained by pair wise correlation with differentially expressed miRNAs. Here we identified 456 mRNAs with significant negative correlation (adjusted p-value < 0.05) between miRNAs and mRNA. The intersection of this list with the list of putative target genes resulted in 424 total genes.

After getting the three lists, we used the SCCA method based on the expression data of the 52 up-regulated miRNAs and all the 10,336 filtered mRNA expression data from the mouse4302 chip. After normalizing each of the matrices so that expression measurements for each miRNA / mRNA had mean zero and standard deviation one, the miRNA data was multiplied by -1. The *CCA.permute* function in package *PMA* (Witten, et al., 2009) was used to determine optimal penalty parameters for SCCA with multiple sets of canonical variables. In the result, there were 8,603 non-zero elements in the $\boldsymbol{u}$ vector, which meant that 8,603 mRNAs were selected by the SCCA function. And there were 43 non-zero elements in $\boldsymbol{v}$ vector, which indicated 43 miRNAs were selected.

The next step was KEGG pathway analysis with the SCCA GSEA method. We first used the *GeneSetCollection* function within the Bioconductor package *GSEABase* to construct a collection of gene sets of pathways from the KEGG database. There were 224 pathways collected from KEGG. FDR adjusted p-values from these pathways are given in **Table 4.34**. For comparison purposes, the results from DAVID analysis of the KEGG database with default parameters based on all 5,154 putative targets, the 416 intersection of these targets with the 2,838 DE down regulated genes and the 424 intersection of putative target genes and significant pair-wise correlation are given in **Tables 4.35- 4.37**.

Table 4.34. KEGG pathway analysis by Multiple canonical vectors and Self-contained test			
pathway	**Statistic**	**P-value**	**Adj.P-value**
Oxidative phosphorylation	2.065	p< 0.001	p< 0.001
Glycine, serine and threonine metabolism	1.416	p< 0.001	p< 0.001
Protein export	1.613	p< 0.001	p< 0.001
Chemokine signaling pathway	1.825	p< 0.001	p< 0.001
Lysosome	1.559	p< 0.001	p< 0.001
Cardiac muscle contraction	2.406	p< 0.001	p< 0.001
Natural killer cell mediated cytotoxicity	1.751	p< 0.001	p< 0.001
Fc gamma R-mediated phagocytosis	2.148	p< 0.001	p< 0.001
Leukocyte transendothelial migration	2.071	p< 0.001	p< 0.001
Regulation of actin cytoskeleton	1.096	p< 0.001	p< 0.001
Vitamin digestion and absorption	2.190	p< 0.001	p< 0.001
Small cell lung cancer	1.975	p< 0.001	p< 0.001
Rheumatoid arthritis	2.316	p< 0.001	p< 0.001
Hypertrophic cardiomyopathy (HCM)	1.712	p< 0.001	p< 0.001
Dilated cardiomyopathy	1.635	p< 0.001	p< 0.001
Phagosome	2.601	0.001	0.012
NOD-like receptor signaling pathway	1.053	0.001	0.012
Bladder cancer	1.442	0.001	0.012
Fc epsilon RI signaling pathway	1.369	0.002	0.023
Collecting duct acid secretion	2.265	0.003	0.031

Table 4.34. KEGG pathway found by SCCA and GSEA method in Neural Tube data. We list 20 of 48 KEGG pathways with FDR adjusted p-value less than 0.05.

Table 4.35 KEGG Pathway analysis of predicted targets of up regulated miRNAs			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
mmu04360:Axon guidance	3.56	p<0.001	p<0.001
mmu04010:MAPK signaling pathway	2.54	p<0.001	p<0.001
mmu05200:Pathways in cancer	2.35	p<0.001	p<0.001
mmu05210:Colorectal cancer	3.42	p<0.001	0.001
mmu04910:Insulin signaling pathway	2.76	p<0.001	0.001
mmu04810:Regulation of actin cytoskeleton	2.23	p<0.001	0.003
mmu05221:Acute myeloid leukemia	3.64	p<0.001	0.008
mmu04720:Long-term potentiation	3.21	0.001	0.011
mmu04310:Wnt signaling pathway	2.32	0.001	0.015
mmu04150:mTOR signaling pathway	3.52	0.001	0.014
mmu04510:Focal adhesion	2.09	0.001	0.013
mmu05211:Renal cell carcinoma	2.96	0.002	0.026
mmu04660:T cell receptor signaling pathway	2.34	0.003	0.036
mmu04722:Neurotrophin signaling pathway	2.26	0.003	0.034
mmu04914:Progesterone-mediated oocyte maturation	2.64	0.003	0.034
mmu05220:Chronic myeloid leukemia	2.73	0.004	0.038
mmu04012:ErbB signaling pathway	2.58	0.004	0.036
mmu04916:Melanogenesis	2.42	0.005	0.040
mmu05215:Prostate cancer	2.50	0.005	0.042
mmu04730:Long-term depression	2.64	0.008	0.060

Table 4.35. KEGG pathway found by DAVID online software in Neural Tube data. We analyzed predicted target gene list of 52 up regulated miRNAs by DAVID. The putative target list contained 5,154 genes. We list first 20 KEGG pathways.

Table 4.36 KEGG Pathway analysis of mRNAs from intersection of different data base of up regulated miRNAs			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
mmu00600:Sphingolipid metabolism	6.83	p<0.001	0.06
mmu04144:Endocytosis	2.64	0.003	0.19
mmu05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)	3.83	0.009	0.32
mmu05200:Pathways in cancer	2.03	0.011	0.29
mmu04960:Aldosterone-regulated sodium reabsorption	4.88	0.018	0.36
mmu04142:Lysosome	2.76	0.025	0.41
mmu04666:Fc gamma R-mediated phagocytosis	2.51	0.09	0.81
mmu00071:Fatty acid metabolism	3.64	0.09	0.79

Table 4.36. KEGG pathway found by DAVID online software in Neural Tube data. We analyzed intersection gene list of predicted target gene list and down-regulated mRNAs by DAVID. The intersection list contained 530 genes.

Table 4.37 KEGG analysis of mRNAs have significant correlation with up regulated miRNAs			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
mmu04144:Endocytosis	2.86	0.001	0.12
mmu05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)	3.85	0.009	0.42
mmu05200:Pathways in cancer	2.04	0.010	0.34
mmu04960:Aldosterone-regulated sodium reabsorption	4.91	0.018	0.42
mmu00600:Sphingolipid metabolism	4.91	0.018	0.42
mmu04666:Fc gamma R-mediated phagocytosis	2.95	0.030	0.52
mmu04510:Focal adhesion	2.08	0.048	0.63
mmu04142:Lysosome	2.43	0.07	0.69
mmu04810:Regulation of actin cytoskeleton	1.90	0.08	0.70
mmu00071:Fatty acid metabolism	3.67	0.09	0.73

Table 4.37. KEGG pathway found by DAVID online software in Neural Tube data. We analyzed intersection gene list of predicted target gene list and mRNAs with significant correlation by DAVID. The intersection list contained 456 genes.

Then, we produced the similar process of KEGG pathway analysis on GO analysis. FDR adjusted p-values from these pathways are given in **Table 4.38**. For comparison purposes, the results from DAVID analysis of the GO database with default parameters based on all 5,154 putative targets, the 530 intersection of these targets with the 3,037 DE down regulated genes and the 456 intersection of putative target genes and those with significant negative pair-wise correlation are given in **Tables 4.39- 4.41**.

Table 4.38. GO pathway analysis by Multiple canonical vectors and Self-contained test				
GO	**GO Term**	**Statistic**	**P-Value**	**Adjust P-value**
GO:0000086	G2/M transition of mitotic cell cycle	0.52	p< 0.001	p< 0.001
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	1.74	p< 0.001	p< 0.001
GO:0000165	MAPK cascade	0.82	p< 0.001	p< 0.001
GO:0000187	activation of MAPK activity	0.394	p< 0.001	p< 0.001
GO:0000266	mitochondrial fission	0.363	p< 0.001	p< 0.001
GO:0000904	cell morphogenesis involved in differentiation	1.417	p< 0.001	p< 0.001
GO:0001501	skeletal system development	0.432	p< 0.001	p< 0.001
GO:0001503	ossification	1.125	p< 0.001	p< 0.001
GO:0001525	angiogenesis	0.873	p< 0.001	p< 0.001
GO:0001568	blood vessel development	0.986	p< 0.001	p< 0.001
GO:0001569	patterning of blood vessels	0.467	p< 0.001	p< 0.001
GO:0001570	vasculogenesis	2.483	p< 0.001	p< 0.001
GO:0001649	osteoblast differentiation	1.626	p< 0.001	p< 0.001
GO:0001656	metanephros development	0.452	p< 0.001	p< 0.001
GO:0001657	ureteric bud development	1.815	p< 0.001	p< 0.001
GO:0001658	branching involved in ureteric bud morphogenesis	1.928	p< 0.001	p< 0.001
GO:0001666	response to hypoxia	2.633	p< 0.001	p< 0.001
GO:0001701	in utero embryonic development	1.708	p< 0.001	p< 0.001
GO:0001708	cell fate specification	0.728	p< 0.001	p< 0.001
GO:0001709	cell fate determination	0.931	p< 0.001	p< 0.001

Table 4.38. GO pathway found by SCCA and GSEA method in Neural Tube data. We list first 20 pathways with p-value less than 0.05.

Table 4.39. GO analysis of predicted targets mRNAs of up regulated miRNAs			
GO Term	**Fold Enrichment**	**P-Value**	**Benjamini**
GO:0006357~regulation of transcription from RNA polymerase II promoter	2.17	p<0.001	p<0.001
GO:0045449~regulation of transcription	1.53	p<0.001	p<0.001
GO:0051252~regulation of RNA metabolic process	1.60	p<0.001	p<0.001
GO:0006355~regulation of transcription, DNA-dependent	1.59	p<0.001	p<0.001
GO:0030528~transcription regulator activity	1.64	p<0.001	p<0.001
GO:0010604~positive regulation of macromolecule metabolic process	1.87	p<0.001	p<0.001
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	2.20	p<0.001	p<0.001
GO:0016568~chromatin modification	2.54	p<0.001	p<0.001
GO:0006350~transcription	1.46	p<0.001	p<0.001
GO:0045893~positive regulation of transcription, DNA-dependent	2.08	p<0.001	p<0.001
GO:0051254~positive regulation of RNA metabolic process	2.07	p<0.001	p<0.001
GO:0045941~positive regulation of transcription	1.99	p<0.001	p<0.001
GO:0010628~positive regulation of gene expression	1.97	p<0.001	p<0.001
GO:0000267~cell fraction	1.88	p<0.001	p<0.001
GO:0009792~embryonic development ending in birth or egg hatching	2.04	p<0.001	p<0.001
GO:0048514~blood vessel morphogenesis	2.62	p<0.001	p<0.001
GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.91	p<0.001	p<0.001
GO:0043009~chordate embryonic development	2.02	p<0.001	p<0.001
GO:0003677~DNA binding	1.42	p<0.001	p<0.001
GO:0046872~metal ion binding	1.25	p<0.001	p<0.001

Table 4.39. GO pathway found by DAVID online software in Neural Tube data. We analyzed predicted target gene list of 52 up regulated miRNAs by DAVID. The putative target list contained 5,154 genes. We list first 20 pathways.

Table. 4.40. GO analysis of mRNAs from intersection of different data base of up regulated miRNAs			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0003700~transcription factor activity	2.69	p<0.001	p<0.001
GO:0009792~embryonic development ending in birth or egg hatching	3.22	p<0.001	p<0.001
GO:0030528~transcription regulator activity	2.10	p<0.001	p<0.001
GO:0043565~sequence-specific DNA binding	2.80	p<0.001	p<0.001
GO:0043009~chordate embryonic development	3.04	p<0.001	p<0.001
GO:0007389~pattern specification process	3.57	p<0.001	p<0.001
GO:0007507~heart development	3.96	p<0.001	p<0.001
GO:0001944~vasculature development	3.71	0.001	p<0.001
GO:0035239~tube morphogenesis	4.39	0.001	p<0.001
GO:0003677~DNA binding	1.75	0.001	p<0.001
GO:0001568~blood vessel development	3.62	0.001	p<0.001
GO:0048729~tissue morphogenesis	3.52	0.002	0.002
GO:0045893~positive regulation of transcription, DNA-dependent	2.76	0.003	0.002
GO:0035295~tube development	3.34	0.003	0.001
GO:0051254~positive regulation of RNA metabolic process	2.74	0.003	0.001
GO:0006357~regulation of transcription from RNA polymerase II promoter	2.36	0.004	0.001
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	2.83	0.004	0.003
GO:0045941~positive regulation of transcription	2.51	0.005	0.003
GO:0048568~embryonic organ development	3.30	0.005	0.004
GO:0048705~skeletal system morphogenesis	4.41	0.008	0.004

Table 4.40. GO pathway found by DAVID online software in Neural Tube data. We analyzed intersection gene list of predicted target gene list and down-regulated mRNAs by DAVID. The intersection list contained 530 genes.

Table. 4.41. GO analysis of mRNAs have significant correlation with up regulated miRNAs			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0010033~response to organic substance	2.80	p<0.001	p<0.001
GO:0043566~structure-specific DNA binding	2.16	p<0.001	p<0.001
GO:0005829~cytosol	3.20	p<0.001	p<0.001
GO:0006357~regulation of transcription from RNA polymerase II promoter	2.73	p<0.001	p<0.001
GO:0003690~double-stranded DNA binding	3.03	p<0.001	p<0.001
GO:0051173~positive regulation of nitrogen compound metabolic process	1.80	p<0.001	p<0.001
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	3.44	p<0.001	0.001
GO:0042493~response to drug	3.81	p<0.001	0.001
GO:0031328~positive regulation of cellular biosynthetic process	3.57	p<0.001	0.001
GO:0009891~positive regulation of biosynthetic process	4.22	p<0.001	0.001
GO:0042802~identical protein binding	3.48	p<0.001	0.001
GO:0045893~positive regulation of transcription, DNA-dependent	3.39	p<0.001	0.003
GO:0051254~positive regulation of RNA metabolic process	3.22	p<0.001	0.003
GO:0010604~positive regulation of macromolecule metabolic process	2.65	p<0.001	0.003
GO:0010557~positive regulation of macromolecule biosynthetic process	2.63	p<0.001	0.003
GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.27	p<0.001	0.003
GO:0048732~gland development	4.86	p<0.001	0.010
GO:0045941~positive regulation of transcription	2.73	p<0.001	0.005
GO:0016564~transcription repressor activity	1.63	p<0.001	0.005
GO:0010628~positive regulation of gene expression	4.25	p<0.001	0.006

Table 4.41. GO pathway found by DAVID online software in Neural Tube data. We analyzed intersection gene list of predicted target gene list and mRNAs with significant correlation by DAVID. The intersection list contained 456 genes.

The second part was the results of 131 down regulated miRNAs, the process of analyzing was similar to the analysis of the 52 up regulated miRNAs. First, we determined three lists of target mRNAs which we would use in the later analysis with DAVID. The first list contained 8,804 genes, the second list consisted of 671 genes and there were 1,114 genes in the third list.

After getting the three lists, we used the SCCA method based on the expression data from the 131 down-regulated miRNAs on GD 9.5 relative to GD 8.5and all the 10,336 filtered mRNA expression data from the mouse4302 chip. After normalizing each of the matrices so that expression measurements for each miRNA / mRNA had mean zero and standard deviation one, the miRNA data was multiplied by -1. The *CCA.permute* function in package *PMA* (Witten, et al., 2009) was used to determine optimal penalty parameters for SCCA with a single set of canonical variables. In the result, there were 5,479 non-zero elements in the $\boldsymbol{u}$ vector, which meant that 5,479 mRNAs were selected by the SCCA function. And there were 43 non-zero elements in $\boldsymbol{v}$ vector, which indicated 43 miRNAs were selected.

The next step was KEGG pathway analysis with the SCCA GSEA method. We first used the *GeneSetCollection* function within the Bioconductor package *GSEABase* to construct a collection of gene sets of pathways from the KEGG database. There were 224 pathways collected from KEGG. FDR adjusted p-values from these pathways are given in **Table 4.42**. For comparison purposes, the results from DAVID analysis of the KEGG database with default parameters based on all 8,804 putative targets, the 671 intersection of these targets with the 2,838 DE up-regulated genes and the intersection of putative target genes and 1,114 genes with significant negative pair-wise correlation with the differentially expressed miRNAs are given in **Tables 4.43- 4.45**.

Table 4.42. KEGG pathway analysis by Multiple canonical vectors and Self-contained test			
Pathway	**Statistic**	**P-value**	**Adj.P-value**
Fatty acid elongation in mitochondria	3.527	p< 0.001	p< 0.001
Steroid biosynthesis	1.627	p< 0.001	p< 0.001
Pyrimidine metabolism	3.092	p< 0.001	p< 0.001
One carbon pool by folate	4.865	p< 0.001	p< 0.001
DNA replication	3.348	p< 0.001	p< 0.001
Nucleotide excision repair	1.509	p< 0.001	p< 0.001
Mismatch repair	2.181	p< 0.001	p< 0.001
Parkinson's disease	3.070	p< 0.001	p< 0.001
Oxidative phosphorylation	2.096	0.003	0.065
Protein export	2.222	0.003	0.065
Aminoacyl-tRNA biosynthesis	1.330	0.005	0.091
Ribosome biogenesis in eukaryotes	2.825	0.005	0.091
Ribosome	2.310	0.010	0.168
RNA polymerase	1.042	0.014	0.218
Selenocompound metabolism	0.867	0.030	0.436
Huntington's disease	1.104	0.033	0.436
Terpenoid backbone biosynthesis	1.175	0.035	0.436
Valine, leucine and isoleucine degradation	1.131	0.036	0.436
RNA transport	2.386	0.039	0.447

Table 4.42. KEGG pathway found by SCCA and GSEA method in Neural Tube data. We list first 20 KEGG pathways with p-value less than 0.05.

Table 4.43 KEGG Pathway analysis of preticted targets of down regulated miRNAs			
Pathway	**Fold Enrichment**	**P-value**	**Benjamini**
mmu04360:Axon guidance	3.00	p< 0.001	p< 0.001
mmu05200:Pathways in cancer	2.15	p< 0.001	p< 0.001
mmu04510:Focal adhesion	2.47	p< 0.001	p< 0.001
mmu04310:Wnt signaling pathway	2.48	p< 0.001	p< 0.001
mmu04810:Regulation of actin cytoskeleton	2.18	p< 0.001	p< 0.001
mmu04010:MAPK signaling pathway	2.02	p< 0.001	p< 0.001
mmu05210:Colorectal cancer	2.84	p< 0.001	p< 0.001
mmu05211:Renal cell carcinoma	2.84	p< 0.001	p< 0.001
mmu04722:Neurotrophin signaling pathway	2.19	p< 0.001	p< 0.001
mmu04520:Adherens junction	2.62	p< 0.001	p< 0.001
mmu04350:TGF-beta signaling pathway	2.48	p< 0.001	p< 0.001
mmu05218:Melanoma	2.64	p< 0.001	p< 0.001
mmu05220:Chronic myeloid leukemia	2.54	p< 0.001	p< 0.001
mmu04530:Tight junction	2.11	p< 0.001	p< 0.001
mmu04012:ErbB signaling pathway	2.42	p< 0.001	p< 0.001
mmu05221:Acute myeloid leukemia	2.79	p< 0.001	p< 0.001
mmu04144:Endocytosis	1.86	p< 0.001	p< 0.001
mmu04910:Insulin signaling pathway	2.06	p< 0.001	p< 0.001
P00034:Integrin signalling pathway	1.60	p< 0.001	p< 0.001
mmu04666:Fc gamma R-mediated phagocytosis	2.26	p< 0.001	p< 0.001

Table 4.43. KEGG pathway found by DAVID online software in Neural Tube data. We analyzed predicted target gene list of 131 down regulated miRNAs by DAVID. The putative target list contained 8,804 genes. We list first 20 KEGG pathways.

Table. 4.44. KEGG Pathway analysis of mRNAs from intersection of different data base of down regulated miRNAs			
Pathway	**Fold Enrichment**	**P-value**	**Benjamini**
mmu04360:Axon guidance	4.76	p< 0.001	p< 0.001
mmu04310:Wnt signaling pathway	3.66	p< 0.001	p< 0.001
mmu04510:Focal adhesion	3.15	p< 0.001	p< 0.001
mmu05200:Pathways in cancer	2.41	p< 0.001	p< 0.001
mmu05210:Colorectal cancer	4.23	p< 0.001	p< 0.001
mmu04010:MAPK signaling pathway	2.55	p< 0.001	p< 0.001
mmu04666:Fc gamma R-mediated phagocytosis	3.71	p< 0.001	p< 0.001
mmu04330:Notch signaling pathway	5.19	p< 0.001	p< 0.001
mmu05220:Chronic myeloid leukemia	4.10	p< 0.001	p< 0.001
mmu05214:Glioma	4.46	p< 0.001	p< 0.001
mmu05221:Acute myeloid leukemia	4.56	p< 0.001	p< 0.001
mmu05213:Endometrial cancer	4.49	p< 0.001	p< 0.001
mmu05223:Non-small cell lung cancer	4.33	p< 0.001	p< 0.001
mmu04664:Fc epsilon RI signaling pathway	3.48	p< 0.001	p< 0.001
mmu04912:GnRH signaling pathway	3.21	p< 0.001	p< 0.001
mmu04914:Progesterone-mediated oocyte maturation	3.36	p< 0.001	p< 0.001
mmu04810:Regulation of actin cytoskeleton	2.27	p< 0.001	p< 0.001
mmu05212:Pancreatic cancer	3.61	p< 0.001	p< 0.001
mmu04660:T cell receptor signaling pathway	2.86	p< 0.001	p< 0.001
mmu05215:Prostate cancer	3.17	p< 0.001	p< 0.001

Table 4.44. KEGG pathway found by DAVID online software in Neural Tube data. We analyzed intersection gene list of predicted target gene list and down-regulated mRNAs of 131 down regulated miRNAs by DAVID. The intersection list contained 671 genes.

Table. 4.45. KEGG analysis of mRNAs have significant correlation with down regulated miRNAs			
Pathway	**Fold Enrichment**	**P-Value**	**Benjamini**
mmu04360:Axon guidance	4.62	p< 0.001	p< 0.001
mmu04310:Wnt signaling pathway	3.76	p< 0.001	p< 0.001
mmu04010:MAPK signaling pathway	2.79	p< 0.001	p< 0.001
mmu04510:Focal adhesion	3.06	p< 0.001	p< 0.001
mmu05210:Colorectal cancer	4.17	p< 0.001	p< 0.001
mmu05200:Pathways in cancer	2.36	p< 0.001	p< 0.001
2.7.11.1	2.35	p< 0.001	0.005
mmu04810:Regulation of actin cytoskeleton	2.48	p< 0.001	0.001
mmu04666:Fc gamma R-mediated phagocytosis	3.43	p< 0.001	0.001
mmu05214:Glioma	4.20	p< 0.001	0.001
mmu05220:Chronic myeloid leukemia	3.83	p< 0.001	0.001
P00057:Wnt signaling pathway	1.76	p< 0.001	0.016
mmu04660:T cell receptor signaling pathway	3.04	p< 0.001	0.002
mmu04330:Notch signaling pathway	4.48	p< 0.001	0.003
mmu04910:Insulin signaling pathway	2.76	p< 0.001	0.004
mmu05213:Endometrial cancer	4.31	p< 0.001	0.004
mmu04012:ErbB signaling pathway	3.35	p< 0.001	0.004
mmu05223:Non-small cell lung cancer	4.15	p< 0.001	0.004
P00005:Angiogenesis	1.87	p< 0.001	0.029
mmu05221:Acute myeloid leukemia	3.93	p< 0.001	0.006

Table 4.45. KEGG pathway found by DAVID online software in Neural Tube data. We analyzed intersection gene list of predicted target gene list and mRNAs with significant correlation by DAVID. The intersection list contained 1,114 genes.

Then, we produced a similar analysis on based on the GO database. FDR adjusted p-values from these terms for the SCCA-GSEA method are given in **Table 4.46**. For comparison purposes, the results from DAVID analysis of the GO database with default parameters based on all 8,804 putative targets, the 671 intersection of these targets with the 2,838 DE up-regulated genes and the 1,114 intersection of putative target genes and significant pair-wise correlation are given in **Tables 4.47**- **4.49**.

Table 4.46. GO pathway analysis by Multiple canonical vectors and Self-contained test				
GO ID	**GO Term**	**Statistic**	**P-Value**	**Adjust P-value**
GO:0000028	ribosomal small subunit assembly	0.438	p< 0.001	p< 0.001
GO:0000038	very long-chain fatty acid metabolic process	2.59	p< 0.001	p< 0.001
GO:0000056	ribosomal small subunit export from nucleus	1.324	p< 0.001	p< 0.001
GO:0000076	DNA replication checkpoint	0.482	p< 0.001	p< 0.001
GO:0000245	spliceosomal complex assembly	3.839	p< 0.001	p< 0.001
GO:0000506	glycosylphosphatidylinositol-N-acetylglucosaminyltransferase	1.805	p< 0.001	p< 0.001
GO:0000778	condensed nuclear chromosome kinetochore	0.432	p< 0.001	p< 0.001
GO:0000785	chromatin	2.181	p< 0.001	p< 0.001
GO:0000801	central element	3.039	p< 0.001	p< 0.001
GO:0001940	male pronucleu	4.261	p< 0.001	p< 0.001
GO:0003406	retinal pigment epithelium development	0.602	p< 0.001	p< 0.001
GO:0003723	RNA binding	3.199	p< 0.001	p< 0.001
GO:0003729	mRNA binding	2.598	p< 0.001	p< 0.001
GO:0003735	structural constituent of ribosome	3.731	p< 0.001	p< 0.001
GO:0003777	microtubule motor activity	4.139	p< 0.001	p< 0.001
GO:0003796	lysozyme activity	1.708	p< 0.001	p< 0.001
GO:0003857	3-hydroxyacyl-CoA dehydrogenase activity	2.991	p< 0.001	p< 0.001
GO:0004111	creatine kinase activity	2.848	p< 0.001	p< 0.001
GO:0004322	ferroxidase activity	2.041	p< 0.001	p< 0.001
GO:0004402	histone acetyltransferase activity	2.504	p< 0.001	p< 0.001

Table 4.46. GO pathway found by SCCA and GSEA method in Neural Tube data. We list first 20 GO pathways with p-value less than 0.05.

Table. 4.47. GO analysis of preticted targets mRNAs of down regulated miRNAs			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0045449~regulation of transcription	1.53	p< 0.001	p< 0.001
GO:0003677~DNA binding	1.56	p< 0.001	p< 0.001
GO:0006350~transcription	1.56	p< 0.001	p< 0.001
GO:0030528~transcription regulator activity	1.71	p< 0.001	p< 0.001
GO:0006357~regulation of transcription from RNA polymerase II promoter	2.01	p< 0.001	p< 0.001
GO:0045941~positive regulation of transcription	2.00	p< 0.001	p< 0.001
GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.95	p< 0.001	p< 0.001
GO:0010628~positive regulation of gene expression	1.97	p< 0.001	p< 0.001
GO:0045893~positive regulation of transcription, DNA-dependent	2.06	p< 0.001	p< 0.001
GO:0051254~positive regulation of RNA metabolic process	2.04	p< 0.001	p< 0.001
GO:0003700~transcription factor activity	1.72	p< 0.001	p< 0.001
GO:0051173~positive regulation of nitrogen compound metabolic process	1.90	p< 0.001	p< 0.001
GO:0010604~positive regulation of macromolecule metabolic process	1.81	p< 0.001	p< 0.001
GO:0043009~chordate embryonic development	2.01	p< 0.001	p< 0.001
GO:0009792~embryonic development ending in birth or egg hatching	2.00	p< 0.001	p< 0.001
GO:0010557~positive regulation of macromolecule biosynthetic process	1.88	p< 0.001	p< 0.001
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	2.10	p< 0.001	p< 0.001
GO:0031328~positive regulation of cellular biosynthetic process	1.84	p< 0.001	p< 0.001
GO:0009891~positive regulation of biosynthetic process	1.84	p< 0.001	p< 0.001
GO:0051252~regulation of RNA metabolic process	1.44	p< 0.001	p< 0.001

Table 4.47. GO pathway found by DAVID online software in Neural Tube data. We analyzed predicted target gene list of 131 down regulated miRNAs by DAVID. The putative target list contained 8,804 genes. We list first 20 GO pathways.

Table. 4.48. GO analysis of mRNAs from intersection of different data base of down regulated miRNAs			
GO Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0045449~regulation of transcription	1.81	p< 0.001	p< 0.001
GO:0003677~DNA binding	1.88	p< 0.001	p< 0.001
GO:0006350~transcription	1.81	p< 0.001	p< 0.001
GO:0030528~transcription regulator activity	2.01	p< 0.001	p< 0.001
GO:0006357~regulation of transcription from RNA polymerase II promoter	2.33	p< 0.001	p< 0.001
GO:0016568~chromatin modification	3.35	p< 0.001	p< 0.001
GO:0030182~neuron differentiation	2.68	p< 0.001	p< 0.001
GO:0010629~negative regulation of gene expression	2.61	p< 0.001	p< 0.001
GO:0005856~cytoskeleton	1.92	p< 0.001	p< 0.001
GO:0010558~negative regulation of macromolecule biosynthetic process	2.56	p< 0.001	p< 0.001
GO:0032990~cell part morphogenesis	3.33	p< 0.001	p< 0.001
GO:0016481~negative regulation of transcription	2.65	p< 0.001	p< 0.001
GO:0044451~nucleoplasm part	2.46	p< 0.001	p< 0.001
GO:0048667~cell morphogenesis involved in neuron differentiation	3.53	p< 0.001	p< 0.001
GO:0031327~negative regulation of cellular biosynthetic process	2.49	p< 0.001	p< 0.001
GO:0005654~nucleoplasm	2.31	p< 0.001	p< 0.001
GO:0009890~negative regulation of biosynthetic process	2.47	p< 0.001	p< 0.001
GO:0007409~axonogenesis	3.68	p< 0.001	p< 0.001
GO:0048812~neuron projection morphogenesis	3.53	p< 0.001	p< 0.001
GO:0051252~regulation of RNA metabolic process	1.68	p< 0.001	p< 0.001

Table 4.48. GO pathway found by DAVID online software in Neural Tube data. We analyzed intersection gene list of predicted target gene list and down-regulated mRNAs of 131 down regulated miRNAs by DAVID. The intersection list contained 671 genes.

Table. 4.49. GO analysis of mRNAs have significant correlation with down regulated miRNAs			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0045449~regulation of transcription	1.76	p< 0.001	p< 0.001
GO:0003677~DNA binding	1.83	p< 0.001	p< 0.001
GO:0030528~transcription regulator activity	2.00	p< 0.001	p< 0.001
GO:0006350~transcription	1.77	p< 0.001	p< 0.001
GO:0030182~neuron differentiation	2.71	p< 0.001	p< 0.001
GO:0044451~nucleoplasm part	2.51	p< 0.001	p< 0.001
GO:0006357~regulation of transcription from RNA polymerase II promoter	2.25	p< 0.001	p< 0.001
GO:0005654~nucleoplasm	2.33	p< 0.001	p< 0.001
GO:0032990~cell part morphogenesis	3.22	p< 0.001	p< 0.001
GO:0048667~cell morphogenesis involved in neuron differentiation	3.44	p< 0.001	p< 0.001
GO:0007409~axonogenesis	3.61	p< 0.001	p< 0.001
GO:0016568~chromatin modification	3.06	p< 0.001	p< 0.001
GO:0048812~neuron projection morphogenesis	3.45	p< 0.001	p< 0.001
GO:0010629~negative regulation of gene expression	2.45	p< 0.001	p< 0.001
GO:0031981~nuclear lumen	1.97	p< 0.001	p< 0.001
GO:0048858~cell projection morphogenesis	3.19	p< 0.001	p< 0.001
GO:0048666~neuron development	2.73	p< 0.001	p< 0.001
GO:0031175~neuron projection development	3.05	p< 0.001	p< 0.001
GO:0051252~regulation of RNA metabolic process	1.63	p< 0.001	p< 0.001
GO:0010558~negative regulation of macromolecule biosynthetic process	2.36	p< 0.001	p< 0.001

Table 4.49. GO pathway found by DAVID online software in Neural Tube data. We analyzed intersection gene list of predicted target gene list and mRNAs with significant correlation by DAVID. The intersection list contained 1,114 genes.

CHAPTER V

METHOD OF INTEGRATED ANALYSIS OF MRNA AND METHYLATION

5.1. Application of sparse mCCA to Murine Palate Methylome data

In our previous chapters, we showed that sparse CCA can perform an integrative analysis on two data sets with the same number of samples but different variables. But now, we need to analyze more than two data sets. (Gifi, 1990) introduced a number of methods that generalized CCA to more than two data sets, and Witten and Tibshirani (Witten and Tibshirani, 2009) extended their sparse CCA approach to sparse multiple CCA (mCCA). Here, we briefly review their methodology with the context of applying it to our genomewide methylation data. In our application we use methylated DNA probes from 21 chromosomes collected from the murine embryonic palate during gestational days (GDs) 12 to 14 (three arrays per GD) (Seelan, et al., 2013). We have M = 21 data sets $\boldsymbol{X_1}, \boldsymbol{X_2} \ldots \boldsymbol{X_{20}}$, , and there are $\boldsymbol{p_m}$ variables (here, probes in methylated regions) and n samples for data set $\boldsymbol{X_m}$ where m= 1,2…21, n= 9 in our research. We normalized each variable in the data sets to have mean 0 and standard deviation 1. Then, the criterion of multiple CCA for obtaining weight vectors $\boldsymbol{w_1}, \boldsymbol{w_2} \ldots \boldsymbol{w_{20}}$ is to maximize

$\sum_{i<j} \boldsymbol{w_i^T X_i^T X_j w_j}$ subject to $\boldsymbol{w_m^T X_m^T X_m w_m} = 1, \forall m,$ where $w_m \in \mathbb{R}^{p_m}$.

As can be seen, when M = 2 multiple CCA reduces to traditional CCA. Following this logical spirit, Witten and Tibshirani (2009) extended the criterion for SCCA with two data sets to sparse multiple CCA. Again, we suppose the samples within each data set are independent so that $\boldsymbol{X_m^T X_m} = \boldsymbol{I}$ for any m. Then the criterion for sparse mCCA is: $maximize_{\boldsymbol{w_1, w_2 \ldots w_M}} \sum_{i<j} \boldsymbol{w_i^T X_i^T X_j w_j}$ subject to $\|\boldsymbol{w_i}\|^2 \leq 1, P_i(\boldsymbol{w_i}) \leq c_i, \forall i$,

where P_i is a lasso or fused lasso penalty. When $c_i$is set appropriately, the canonical vector $\boldsymbol{w_i}$ which related to data set $\boldsymbol{X_i}$ will be sparse and smooth.

The algorithm of Witten and Tibshirani (2009) for calculating the canonical weight vectors of the sparse mCCA is:1. Set initial value for each.$\boldsymbol{w_i} \in \mathbb{R}^{p_m}$2. For every data set, we repeat iteration until the canonical vector $\boldsymbol{w_i}$converges:

$\boldsymbol{w_i} \leftarrow argmax_{w_i} \boldsymbol{w_i^T X_i^T} (\sum_{j \neq i} \boldsymbol{X_j w_j})$ subject to $\|\boldsymbol{w_i}\|^2 \leq 1, P_i(\boldsymbol{w_i}) \leq c_i$ For the example of $\boldsymbol{L_1}$ penalty of P_i the update of $\boldsymbol{w_i}$ follows the form as:

$$\boldsymbol{w_i} \leftarrow \frac{S(\boldsymbol{X_i^T}(\sum_{j \neq i} \boldsymbol{X_j w_j}), \Delta_i)}{\left\| S(\boldsymbol{X_i^T}(\sum_{j \neq i} \boldsymbol{X_j w_j}), \Delta_i) \right\|_2}$$

When $\|\boldsymbol{w_i}\|_1 = c_i$ we choose $\Delta_i > 0$, in addition`$\Delta_i = 0$ when $\|\boldsymbol{w_i}\|_1 < c_i$.

Witten and Tibshirani (2009) used the sparse mCCA approach to investigate genome wide correlation in copy number patterns. In our research, we posed a similar question concerning whether the methylated probes on separate chromosomes have similar changes in pattern. Hence, we apply sparse mCCA on data sets X_i where $i = 1,2 \ldots 21$, each contains methylated probes on chromosome *i*. Because the methylated probes are ordered along the chromosome, a fused lasso penalty is used on all data sets.

5.2. Integrated analysis of methylated regions of interest (MRIs) measurements and mRNA expression using SCCA

In this part we first identify the DE mRNAs (up and down-regulated between GD 12 and 14) which have maximum negative correlation with MRIs. We determine the correlation between MRIs and mRNAs by the SCCA method which we described in the previous section. The procedure is similar to the integrated analysis of miRNA and mRNA expression using SCCA. Suppose there are two types of data sets $\boldsymbol{X_{i1}}$ and $\boldsymbol{X_{i2}}$ with the same number of observations *n*, where $i = 1,2 \ldots, 21$ (the total number of chromosomes). In our application, the first type of data matrix $\boldsymbol{X_{i1}}$has dimension $n \times p_{i1}$ (p_{i1} variables with n observations, where $i = 1,2 \ldots 21$) indicates DE mRNAs in each chromosome. The second type of data matrix $\boldsymbol{X_{i2}}$ has dimension $n \times p_{i2}$ (p_{i2} variables with n observations, where $i = 1,2 \ldots 21$) represents MRIs in each chromosome. For any pair of these two data sets under same chromosome, we first identify DE genes (up and down-regulated between GD 12 and 14) and MRIs. Second, for each chromosome the entire set of DE gene expression

measurements and the entire set of MRIs will be analyzed using SCCA to determine how the global changes in methylation patterns along the chromosome impact gene expression patterns on the chromosome. Since methylation measurements are ordered along the chromosome, we should follow (Witten and Tibshirani, 2009) and use the lasso penalty (Tibshirani, 1996) on the matrix of mRNA expression measurements and the fused lasso penalty (Tibshirani, et al., 2005) on the matrix of methylation measurements. However, for GSEA we need to create the gene set scores from the weight vectors obtained from SCCA and hence we need the weight vectors to be positive. Since the SCCA software requires using the same penalty for each data set to obtain positive weight vectors, we also use the lasso penalty on the MRIs matrix. After the SCCA process, we again create a GSEA statistic on the basis of the SCCA analysis, similar in spirit to the statistic for integrating miRNA and mRNA expression data.

Since we apply integrated analysis of MRIs and mRNA on separate chromosomes, we create GSEA statistics for each chromosome. After the SCCA procedure, we obtain the weight vectors $\boldsymbol{u_i}$ and $\boldsymbol{v_i}$ on chromosome *i*, where *i*= 1, 2…21, $\boldsymbol{u_i}$ is the weight vector of mRNA expression matrix of chromosome *i* with dimensional $1 \times p_{i1}$and $\boldsymbol{v_i}$ is the weight vector of MRI probe intensity matrix of chromosome *i* with dimensional $1 \times p_{i2}$. The test statistic is constructed from two parts. The first part consists of the normalized $\boldsymbol{u_i}$ vector $\boldsymbol{u_{i_{\text{norm}}}}$, such that the mean of $\boldsymbol{u_{i_{\text{norm}}}}$ is zero and the variance is one. This component simply indicates the genes that are represented in the weight vector $\boldsymbol{u_i}$. The second part consists of the averaged MRI probe scores which correlate with specific genes selected by SCCA. This component incorporates the weights associated with MRI probe scores into the per-gene scores. We first map the MRI probes to the nearest gene. Second, we calculated the averaged score of MRI probes for each gene and called the new vector $\boldsymbol{v_i}^*$with dimension $1 \times p_{i1}$; same with $\boldsymbol{u_i}$ The $\boldsymbol{v_i}^*$ scores are also normalized to have mean zero and standard deviation one ($\boldsymbol{v_{i_{\text{norm}}}^*}$). Lastly, we set the final statistic associated with each gene by summing the normalized $\boldsymbol{u_i}$ and $\boldsymbol{v_i}^*$ scores and then dividing by $\sqrt{2}$, which is denoted as $\boldsymbol{Z_i} = \frac{1}{\sqrt{2}}(\boldsymbol{u_{i\ \text{norm}}} + \boldsymbol{v_{i_{norm}}^*})$.

Then, we calculate an aggregate gene enrichment score for each pre-determined gene set. Suppose there are *K* pre-determined gene sets with $n_1, n_2, \ldots, n_K$ genes in each set. In our terminology, the vector $\boldsymbol{Z}_{ij}$consists of the components of $\boldsymbol{Z_i}$ corresponding to the genes in gene set *j* . Then, the GSEA statistic

$gs_1(\boldsymbol{Z}_{i1}),\ gs_2(\boldsymbol{Z}_{i2}), \ldots, gs_K(\boldsymbol{Z}_{iK})$ for each gene set is calculated by the sum of the per gene statistic included in each gene set then divided by the square root of the number of genes in each gene set:

$$gs_k(\boldsymbol{Z}_{ik}) = \frac{1}{\sqrt{n_k}}\left(\sum_{j=1}^{n_k} z_{ij}\right),$$

where $k = 1,2, \ldots, K$ and $\boldsymbol{Z}_{ik} = (z_{i1},\ z_{i2}, \ldots,\ z_{in_k})$ are the gene statistics for gene set gs_k.

Finally, we calculate the permutation *p*-value of the GSEA statistic for each gene set using competitive test. We first resample the per-gene statistics $\boldsymbol{Z_i} = (z_{i1},\ z_{i2}, \ldots,\ z_{ip_1})$ without replacement to obtain permuted statistics $\boldsymbol{Z_i}^p$ for $p = 1, \ldots, P$ permutations. Next, permuted gene set statistics $gs_1^p(\boldsymbol{Z}_{i1}^p),\ gs_2^p(\boldsymbol{Z}_{i2}^p), \ldots, gs_K^p(\boldsymbol{Z}_{iK}^p)$ are calculated for each of the original gene sets, where $\boldsymbol{Z}_{ik}^p = (z_{i1}^p,\ z_{i2}^p, \ldots, z_{in_k}^p)$ are the permuted gene statistics for each gene set gs_k. The permutation *p*-value $p_{\text{perm},k}$ for each gene-set k is then calculated as the proportion of the permuted GSEA statistics that are larger than the original GSEA statistic: $p_{\text{perm},k} = \frac{1}{P}\sum_{p=1}^{P} I(\,gs_k^p(\boldsymbol{Z}_{ik}^p) > gs_k(\boldsymbol{Z}_{ik}))$,

where $I(\cdot)$ is the indicator function.

CHAPTER VI

REAL DATA ANALYSIS OF MURINE PALATAL METHYLOME DATA

The original methylation data set obtained from the developing secondary palate of mouse embryos was collected using NimbleGen 2.1M mouse promoter arrays (Seelan, et al., 2013). Data were analyzed using the Bioconductor package Ringo(Toedling, et al., 2007). It contained three arrays per each gestational day (GD), GD 12, 13, and 14. The total number of samples in the original methylation data was 9 and the number of probes was 2,064,266. The mRNA data was collected in prior studies using Affymetrix GENECHIP analysis software and the annotation of the data was mgu74av2. The number of samples in the mRNA data was 9 (three per GD) and the number of genes was 21,706. Since the gene expression data was collected in prior studies and the samples were different for the methylation data, we averaged the data by GD when we did integrated gene expression and methylation data analysis.

In the analysis, we first filtered the data set. For the original methylation data, there were 2,064,266 probes and 9 samples. We kept the 70,072 methylation probes which were in methylated regions (Seelan et al, 2013) and averaged the data by GD day. In the mRNA expression data, we used the function *nsFilter* within R package *genefilter* to filter the mRNA data by removing duplicate probes mapping to the same Entrez Gene ID (the probe with the highest variance across the samples was retained) and probes with a variance below the 50^{th} percentile. After filtering 12,488 probes remained and we again averaged the data by GD day.

Second, we used sparse multiple CCA to calculate the canonical vectors at whole methylated regions to identify which methylated regions are correlated with each other (to identify methylated regions which have correlated changes). The result is in **Figure 6.1**

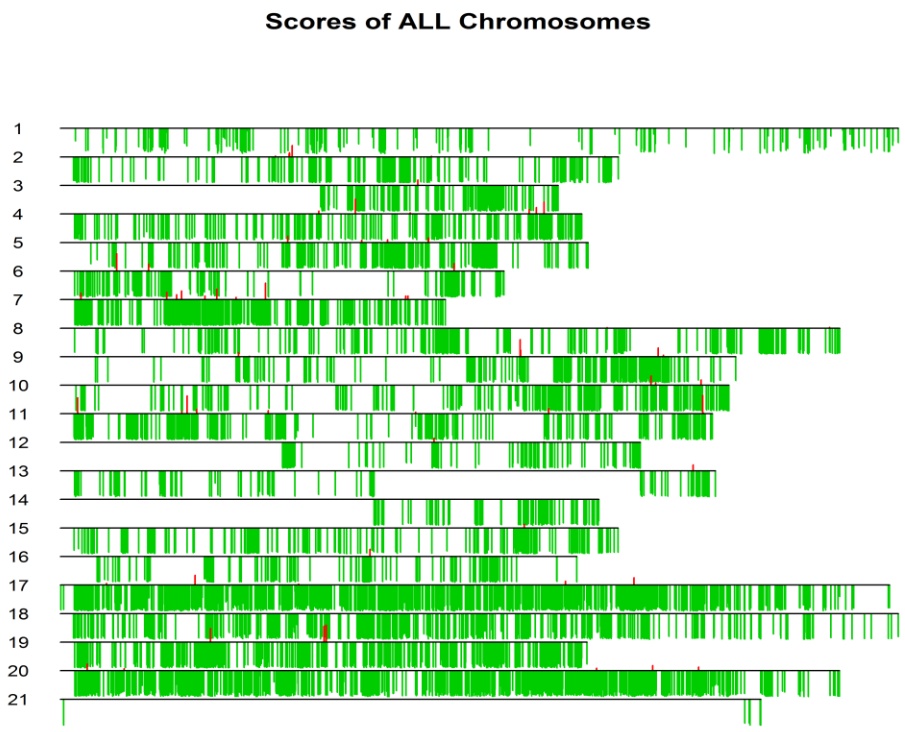


Figure 6.1. Sparse mCCA treating each chromosome as a separate data set, in order to identify genomic regions that have correlated methylation patterns. The canonical vectors $w_{1,} \ldots, w_{21}$ are shown. Positive values of the canonical vectors are shown in red, and negative values are in green.

Third, we determined differentially expressed mRNAs between GD-14 and GD-12 arrays on chromosome 1 using the empirical Bayes method in R package *limma* (Ritchie, et al., 2015) (Smyth, 2004). We identified 42 significant mRNAs with adjusted *p*-values (based on the Benjamini- Hochberg correction) ≤ 0.05. Nine out of the 42 mRNAs had a negative log fold change and were down regulated in GD-14 compared to GD-12, and the other 33 mRNAs had a positive log fold changes and were up regulated.

Fourth, we used the SCCA method based on the 9 down-regulated mRNAs data and the 3,157 methylated probes on chromosome 1. After normalizing each of the matrices so that expression measurements for each mRNA / Methylation had mean zero and standard deviation one, the mRNA data was multiplied by -1. The *CCA.permute* function in package *PMA* (Witten, et al., 2009) was used to determine the optimal penalty parameters for SCCA with a single set of canonical variables. In the result, there were 9 non-zero elements in the $\boldsymbol{u}$ vector, which meant that 9 mRNAs were selected by the SCCA function. And there were 210 non-zero elements in $\boldsymbol{v}$ vector, which indicated 210 methylated probes were selected.

The next step was GO pathway analysis with the SCCA GSEA method. We first used the *GeneSetCollection* function within the Bioconductor package *GSEABase* to construct a collection of gene sets of pathways from the GO database. There were 380 pathways collected from GO. FDR adjusted p-values from these pathways are given in **Table 6.1**. For comparison purposes, the results from DAVID analysis of the GO database with default parameters based on the 80 genes which mapped to the methylated probes with positive loadings in the SCCA method is given in **Table 6.2.**

Table. 6.1. GO pathway analysis				
GO ID	GO Term	Statistic	P-Value	Adj.pvalue
GO:0016311	dephosphorylation	0.274	0.007	0.820
GO:0031225	anchored component of membrane	0.276	0.008	0.820
GO:0005886	plasma membrane	0.666	0.010	0.820
GO:0030016	myofibril	0.273	0.010	0.820
GO:0005515	protein binding	1.242	0.012	0.820
GO:0009986	cell surface	0.351	0.021	0.820
GO:0016020	membrane	0.940	0.029	0.820
GO:0005622	intracellular	0.491	0.042	0.820
GO:0004035	alkaline phosphatase activity	0.183	0.049	0.820
GO:0016791	phosphatase activity	0.183	0.049	0.820

Table 6.1. GO pathway found by SCCA and GSEA method in chromosome 1 of the murine palate data.

Table 6.2. GO analysis with DAVID				
Term	**Genes**	**Fold Enrichment**	**P-value**	**Benja mini**
GO:0006928~cell motion	enabled homolog (Drosophila); similar to SH2/SH3 adaptor protein; neuron navigator 1; Fc receptor, IgE, high affinity I, gamma polypeptide; GLI-Kruppel family member GLI2	3.778	0.040	1.000
GO:0009084~glutamine family amino acid biosynthetic process	predicted gene 4949, glutamate-ammonia ligase (glutamine synthetase); pyrroline-5-carboxylate reductase family, member 2	36.974	0.052	1.000
GO:0019842~vitamin binding	LMBR1 domain containing 1; solute carrier family 19 (sodium/hydrogen exchanger), member 3; selenocysteine lyase	6.460	0.076	1.000
GO:0005212~structural constituent of eye lens	crystallin, gamma D, crystallin, gamma A	24.814	0.076	0.999
GO:0070013~intracellular organelle lumen	RNA binding motif protein 8a; UDP-glucose ceramide glucosyltransferase-like 1; isoleucine-tRNA synthetase 2, mitochondrial; inhibitor of growth family, member 5; staufen (RNA binding protein) homolog 2 (Drosophila); death effector domain-containing; calsequestrin 1; RIKEN cDNA 6430706D22 gene	2.007	0.090	1.000
GO:0043233~organelle lumen 60 genes from our list are not in the output.	RNA binding motif protein 8a; UDP-glucose ceramide glucosyltransferase-like 1; isoleucine-tRNA synthetase 2, mitochondrial; inhibitor of growth family, member 5; staufen (RNA binding protein) homolog 2 (Drosophila); death effector domain-containing; calsequestrin 1; RIKEN cDNA 6430706D22 gene	2.001	0.091	0.998

Table 6.2. GO pathway found by DAVID software in the murine palate data. We analyzed the list of genes on chromosome 1 which mapped to the methylated probes with positive loadings in SCCA with DAVID. The list contained 80 genes.

Then, we did a similar analysis of chromosome 1 on chromosome 2. First, we determined differentially expressed mRNAs between GD-14 and GD-12 arrays on chromosome 2 using the empirical Bayes method in R package *limma* (Ritchie, et al., 2015) (Smyth, 2004). We identified 42 significant mRNAs with adjusted *p*-values (based on Benjamini-Hochberg correction) ≤ 0.05. Twelve out of 40 mRNAs with negative log fold change were down regulated on GD-14 versus GD-12, and the other 28 mRNAs had a positive log fold change and were up regulated.

Fourth, we used the SCCA method based on the 12 down-regulated mRNAs and the 3,157 methylated probes on chromosome 2. After normalizing each of the matrices so that expression measurements for each mRNA / methylation probe had mean zero and standard deviation one, the mRNA data was multiplied by -1. The *CCA.permute* function in package *PMA* (Witten, et al., 2009) was used to determine optimal penalty parameters for SCCA with a single set of canonical variables. In the result, there were 12 non-zero elements in the $\boldsymbol{u}$ vector, which meant that 12 mRNAs were selected by the SCCA function. And there were 733 non-zero elements in $\boldsymbol{v}$ vector, which indicated 733 methylated probes were selected.

The next step was GO pathway analysis with the SCCA GSEA method. We first used the *GeneSetCollection* function within the Bioconductor package *GSEABase* to construct a collection of gene sets of pathways from the GO database. There were 380 pathways collected from GO. FDR adjusted p-values from these pathways are given in **Table 6.3**. For comparison purposes, the results from DAVID analysis of the GO database with default parameters based on 76 genes which mapped to the methylated probes with positive loadings in the SCCA method is given in **Table 6.4.**

Table 6.3. GO Pathway analysis				
GO ID	**GO Term**	**Statistic**	**P-Value**	**Adj.pvalue**
GO:0009791	post-embryonic development	0.434	0	0.000
GO:0035116	embryonic hindlimb morphogenesis	0.434	0	0.000
GO:0042995	cell projection	0.547	0	0.000
GO:0001894	tissue homeostasis	0.334	0.001	0.012
GO:0001958	endochondral ossification	0.334	0.001	0.012
GO:0003924	TPase activity	0.334	0.001	0.012
GO:0004871	signal transducer activity	0.334	0.001	0.012
GO:0005834	heterotrimeric G-protein complex	0.334	0.001	0.012
GO:0006112	energy reserve metabolic process	0.334	0.001	0.012
GO:0006306	DNA methylation	0.334	0.001	0.012
GO:0007186	G-protein coupled receptor signaling pathway	0.334	0.001	0.012
GO:0007189	adenylate cyclase-activating G-protein coupled receptor signaling pathway	0.334	0.001	0.012
GO:0007191	adenylate cyclase-activating dopamine receptor signaling pathway	0.334	0.001	0.012
GO:0007606	sensory perception of chemical stimulus	0.334	0.001	0.012
GO:0019001	guanyl nucleotide binding	0.334	0.001	0.012
GO:0030425	dendrite	0.334	0.001	0.012
GO:0031234	extrinsic component of cytoplasmic side of plasma membrane	0.334	0.001	0.012
GO:0031683	G-protein beta/gamma-subunit complex binding	0.334	0.001	0.012
GO:0031852	mu-type opioid receptor binding	0.334	0.001	0.012
GO:0035255	ionotropic glutamate receptor binding	0.334	0.001	0.012

Table 6.3. GO pathway found by SCCA and GSEA method in chromosome 2 of the murine palate data.

Table 6.4 GO pathway analysis			
Term	**Fold Enrichment**	**P-value**	**Benjamini**
GO:0030326~embryonic limb morphogenesis	11.674	0.005	0.908
GO:0035113~embryonic appendage morphogenesis	11.674	0.005	0.908
GO:0035137~hindlimb morphogenesis	24.978	0.006	0.799
GO:0035108~limb morphogenesis	9.846	0.007	0.721
GO:0035107~appendage morphogenesis	9.846	0.007	0.721
GO:0060173~limb development	9.515	0.008	0.651
GO:0048736~appendage development	9.515	0.008	0.651
GO:0008219~cell death	3.350	0.030	0.958
GO:0016265~death	3.273	0.033	0.945
GO:0009791~post-embryonic development	9.761	0.036	0.937
GO:0042981~regulation of apoptosis	3.071	0.041	0.936
GO:0043067~regulation of programmed cell death	3.033	0.043	0.923
GO:0010941~regulation of cell death	3.017	0.044	0.905
GO:0009886~post-embryonic morphogenesis	40.440	0.047	0.900

Table 6.4. GO pathway found by DAVID software in the murine palate data. We analyzed list of genes on chromosome 2 which mapped back to the methylated probes with positive loadings in SCCA with DAVID. The list contained 76 genes.

CHAPTER VII

DISCUSSION

In this research, we developed a novel GSEA approach for integrated analysis of miRNA / mRNA expression data and miRNA / methylation data. Our methodology uses sparse CCA to find correlated sub-dimensions in the two data sets, and bases the GSEA statistic on the weight vectors from this analysis. We tested our methodology using multiple real and simulated data sets and compared it with standard approaches in the literature based on pairwise correlation analysis or the intersection of gene lists from differentially expressed up and down-regulated mRNAs / miRNAs.

In the simulation study for integrated GSEA of miRNA and mRNA expression data, the PWC method has larger power than SCCA when the number of targeting miRNAs is small. But the SCCA method outperforms the PWC approach as the number of targeting miRNAs increases. This separation is greatest when the sample size is small and the standard deviation is large, with the power of the methods converging to each other as the sample size increases. As we expected, the whole simulation results support that the power of two methods increase as the sample size of the simulated data and the number of correlated miRNAs and mRNAs increase. Inversely, the power decreases as the error rate of the simulated data increases. In general we found that the SCCA method had better performance (higher power) than PWC.

In the real data analysis of miRNA / mRNA expression, the SCCA-GSEA method may give a more reasonable number of pathways compared to DAVID analysis (gene-set analysis) using all putative targets of differentially expressed (DE) miRNAs and the intersection of this list with DE mRNAs. Since the number of putative target genes of DE miRNAs is usually quite large (several thousands), the number of significantly enriched pathways based on this list is correspondingly large as well. And intersecting this list with DE mRNAs based on a hard p-value threshold may result in the opposite problem of too few genes inthe list. Hence, integrated analysis using our SCCA- GSEA approach may result in a nice compromise of obtaining a focused list of germane pathways and biological gene sets.

In our research, we have introduced sparse canonical correlation analysis as a method for doing integrated miRNA / mRNA analysis. There are also other methods for this purpose, including MMIA(Nam, et al., 2009), mirAct (Liang, et al., 2011), and MAGIA (Sales, et al., 2010). A drawback of the above approaches is that evaluating all potential miRNA / mRNA interactions using pairwise correlations can lead to significant reduction in power due to the number of comparisons involved. And methods that focus on pairing significantly up-regulated mRNAs with down-regulated miRNA counterparts (and vice-versa) (e.g., MMIA) potentially lose information by using a hard threshold for determining the differentially expressed (DE) list of mRNAs and miRNAs. As an alternative, sparse CCA is a data reduction technique that has been effective in the high-throughput setting for integrating gene expression and other types of 'omics' data.

There are quite a few other methods for doing integrated GSEA based on multiple data sets. Poisson et al. (Poisson, et al., 2011) introduce two methods of integrated GSEA using both gene expression and metabolite information. The first is logistic regression analysis with 2-df Wald test, a multivariate extension of the competitive logistic regression test. In this method, they first separately modeled genes and metabolites with absolute per-element t-statistic. The null hypothesis of the joint test is that both regression coefficients are zero. Under the null hypothesis, the test statistic follows chi-square distribution with two degrees of freedom. The second is sum of squared statistics with a 2-dimensional permutation test, a multivariate extension of the self-contained sum of squared statistics. They create observed pair and null pair enrichment test statistics for genes and metabolites, then calculate the Mahalanobis distance from observed and null statistic pairs to the centroid of the sets of null pairs. Then, they calculate the joint permutation p-value as the proportion that the Mahalanobis distances corresponding to observed statistics are larger than or equal to the null statistic Mahalanobis distance.

One obvious issue for both methods is that they are joint assessment approaches which connect per-gene and per-metabolite test statistics as a single vector. In most cases, gene expression and metabolites have a different sample size, so two-sample t statistics are not directly comparable (due to differing degrees of freedom). Use of p-values solves the comparability problem, but in this case the empirical p-value will lack precision and lose directionality. Our method avoids this issue by creating test statistics separately from both data sets and mapping them to the gene level. For miRNA / mRNA integrated analysis this is done by

incorporating the targeting matrix of miRNAs. For mRNA / methylation analysis, we map MRIs back to correlated genes and use averaged values of methylation probes to obtain two lists with the same size. Although there are many differences between our method and that of Poisson et al., the overall procedure is similar. First, create per-element test statistics. Then, define gene sets and create a GSEA score. Lastly, calculate a permutation based p-value corresponding to each gene set.

Jiang and Gentleman (Jiang and Gentleman, 2007) start from the original GSEA which is described in (Subramanian, et al., 2005) and (Tian, et al., 2005). Then, they extend the method of obtaining the test statistic with linear modeling and posterior probabilities. They further extend the gene set aggregation function by using the median and sign-test rather than the mean gene set score. In the paper, they also apply the method on acute lymphoblastic leukemia (ALL) data to produce an incidence matrix for showing the association between genes and phenotypes (pre-defined gene sets). However, they do not integrate any other kind of data with mRNA expression data. In our research, we create putative targeting gene matrix of miRNA, in this case, we generate the per-gene statistic with the information both miRNA and mRNA data sets. In their paper they create gene statistic with three method, two-sample t-statistic, linear modeling and posterior probability as gene statistic. All of these methods depend on hard threshold, but our SCCA method using a soft-threshold to detecting the weight vectors for both data sets. In the Jiang's paper, they create gene set statistics by three different summaries of the evidence for each gene set, mean, median and sign-test. Following, they produce a competitive permutation test with 5000 permutations for obtaining P-values for each pathway. In our research, we use square root mean as the evidence of gene set, the reason we choose this has been stated in previous section. For real data analysis in our research, we produce a self-contained test for GESA method.

Lai et al. (Lai, et al., 2014) introduce a concordant integrative gene set enrichment analysis method. This research focuses on low-sample expression data, and they apply the method on two microarray gene expression data sets. They choose traditional two-sample student's t-test to screen genes. Then, they propose a mixture model with three normal distributions which represent three conditions (genes not differentially expressed, up-regulated and down-regulated) for each individual gene expression data set. The model is estimated by the E-M algorithm. After this, they derive the probability for a gene to be in a pre- defined gene set, then calculate the concordant gene set enrichment score by a partial concordance/discordance (PCD)

model. For computational convenience, a Monte Carlo approximation is produced. Lastly, a p-value is obtained based on the likelihood ratio test. This method is quite different from ours, in that they create a mixture modeling statistical method for concordant integrated gene set enrichment analysis. So, before integrated analysis they can statistically test for genome-wide concordance. Second, it is convenient for calculating the FDR due to using a probabilistic framework for integrated analysis of gene sets. But on the other hand, there are several disadvantages. The mixture model is simple due to being restricted to the two-sample situation. Second, they assume that genes are independent. In our research, we do not need the restrictive assumption of independent among genes, and we use SCCA to investigate the correlation trend throughout the entire data set.

There are several advantages for the sparse canonical correlation method. First, it does not consider individual pairwise correlations but rather the correlation pattern on a global (genome-wide) scale. Sparse CCA can provide the main characteristics of the data by condensing the variables into a smaller dimension. In our case, for instance, genes contained in the same pathway may have a similar effect from the variations in multiple regulatory elements (e.g., miRNA expression and methylation patterns). Second, the GSEA score contains contributions from both mRNA expression and miRNA expression / gene methylation. For integrating miRNA and mRNA expression into a GSEA score we create putative target matrix of the miRNAs and then combine this with the measurements from the miRNA expression using SCCA. Hence this procedure explicitly takes miRNA expression into account in the GSEA score. In previous methods (e.g. MMIA(Nam, et al., 2009), mirAct (Liang, et al., 2011), and MAGIA (Sales, et al., 2010)), they generally simply take list of up regulated differentially expressed genes intersected with targeting genes of down regulated differentially expressed miRNAs.

Some limitations of our approach include the computational burden associated with calculating the optimal penalty for SCCA (Witten and Tibshirani, 2009). In the GSEA test for this research, we first calculated the permuted p-value of each gene set using the competitive test. This approach permuted the statistic associated with each gene and re-calculated the gene-set scores for each permutation, and had the advantage of being computationally fast. However, the competitive test evaluates the null hypothesis that the composite score for a given gene set is *different* from the other gene sets (hence the term competitive). In contrast, the self-

contained test (which permutes the samples for calculating the null distribution) addresses the null hypothesis of more direct biological interest, that the gene set does not contain genes whose expression levels are associated with the phenotype of interest. Further, the re-sampling strategy of the competitive test is gene-based (as opposed to sample-based), which is not in-line with the experimental design and has the underlying assumption that the statistics associated with all the genes are independent. Hence in our real data analysis we opted for using the self-contained test to evaluate the alternative that the gene set contains genes that are differentially expressed (associated with the phenotype) and also regulated by miRNAs. But since there is an additional permutation procedure to obtain the optimal shrinkage penalty in SCCA (e.g., using the *CCA.permute* function), the self-contained test is quite computationally burdensome (i.e. one permutation for finding multiple canonical vectors takes 10-20 minutes). So, despite the limitations we used the competitive test method for the simulation study. And though we used the self-contained test for real data analysis, we used the same shrinkage penalty from the original data set for each permutation.

In our simulation study involving multiple gene sets, we found that when the sample size increased the permuted optimal parameter λ for SCCA might be very restrictive. This means that very few miRNAs and mRNAs will have non-zero loadings on the canonical correlation weight vectors. In this case, the power of SCCA will start to decrease for larger sample sizes. However, this situation is addressed by using the one standard deviation rule to select a larger λ and hence involve more miRNAs and mRNAs.

Another possible extension to our SCCA-GSEA approach is to directly incorporate covariates. In our current research, we first detected down regulated differentially expressed miRNAs and then applied SCCA on these DE miRNAs and whole mRNA data sets to calculate canonical weight vectors. But in CCA we can directly incorporate a phenotype to identify features that are correlated across the whole miRNA and mRNA data sets and also correlated with the phenotype (Witten and Tibshirani, 2009). Then, we could use SCCA on the complete miRNA and mRNA data sets coupled with the covariates. The alternative approach allows extensions to survival data, multiple class data or quantitative data.

SCCA often overlooks the structural or group effect within genomic data, which can be important (e.g., methylation probes in a MRI interact and work together as a group). In this case, group sparse canonical correlation analysis (Lin, et al., 2013) is introduced to analyze the relationship between two different types

of genomic data (i.e., methylated probes and gene expression in our research). We did do a preliminary application of the group SCCA method (Lin, et al., 2013) on our mRNA expression / gene methylation data. We start by obtaining initial canonical weight vectors by applying SCCA on the two data sets. Then, we treat each MRI as a group, and use the iterative group sparse CCA algorithm to obtain updated canonical weight vectors starting from the initial estimates from SCCA. Next, we create a GSEA score with the new canonical vectors based on the group SCCA approach by the same method we described in Chapter V. In future studies, we will conduct a simulation study for the group SCCA method and compared it with SCCA for detecting enriched gene sets.

In conclusion, in this research we applied sparse canonical correlation analysis for both integrated analysis of mRNA expression / miRNA expression and mRNA expression / gene methylation. We then developed two novel gene set enrichment analysis statistics based on these integrated analysis using SCCA, and evaluated the performance on both real and simulated data sets. The performance of the proposed statistics shows promise for identifying biological pathways enriched for genes regulated by miRNA expression or gene methylation. Potential future extensions include using more sophisticated penalty functions and incorporating phenotypes directing into the GSEA statistic based on SCCA.

REFERENCES

Abatangelo, L., *et al.* Comparative study of gene set enrichment methods. *BMC bioinformatics* 2009;10:275-275.

Baylin, S.B. DNA methylation and gene silencing in cancer. *Nature clinical practice. Oncology* 2005;2 Suppl 1:S4-11.

Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995;57(1):289-300.

Bird, A. DNA methylation patterns and epigenetic memory. 2002(0890-9369 (Print)).

Corney, D.C., *et al.* MicroRNA-34b and MicroRNA-34c are targets of p53 and cooperate in control of cell proliferation and adhesion-independent growth. *Cancer research* 2007;67(18):8433-8438.

Efron, B. and Tibshirani, R. On testing the significance of sets of genes. 2007:107-129.

Ferretti, E., *et al.* MicroRNA profiling in human medulloblastoma. *International journal of cancer. Journal international du cancer* 2009;124(3):568-577.

Flynt, A.S. and Lai, E.C. Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nature reviews. Genetics* 2008;9(11):831-842.

Gifi, A. Nonlinear Multivariate Analysis. Wiley; 1990.

Gonzalo, S. Epigenetic alterations in aging. *Journal of applied physiology (Bethesda, Md. : 1985)* 2010;109(2):586-597.

Gottardo, F., *et al.* Micro-RNA profiling in kidney and bladder cancers. *Urologic Oncology: Seminars and Original Investigations* 2007;25(5):387-392.

Haines, T.R., Rodenhiser, D.I. and Ainsworth, P.J. Allele-specific non-CpG methylation of the Nf1 gene during early mouse development. *Developmental biology* 2001;240(2):585-598.

Herman, J.G. and Baylin, S.B. Gene silencing in cancer in association with promoter hypermethylation. *The New England journal of medicine* 2003;349(21):2042-2054.

Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* 1936;28:321.

Jiang, Z. and Gentleman, R. Extensions to gene set enrichment. *Bioinformatics (Oxford, England)* 2007;23(3):306-313.

John, B., *et al.* Human MicroRNA Targets. *PLoS Biol* 2004;2(11):e363.

Jones, P.A. and Takai, D. The role of DNA methylation in mammalian epigenetics. *Science (New York, N.Y.)* 2001;293(5532):1068-1070.

Kanehisa, M., *et al.* The KEGG resource for deciphering the genome. *Nucleic Acids Research* 2004;32(Database issue):D277-D280.

Kozomara, A. and Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research* 2014;42(Database issue):D68-D73.

Kren, B.T., *et al.* microRNAs identified in highly purified liver-derived mitochondria may play a role in apoptosis. *RNA biology* 2009;6(1):65-72.

Lai, Y., *et al.* Concordant integrative gene set enrichment analysis of multiple large-scale two-sample expression data sets. *BMC Genomics* 2014;15(Suppl 1):S6.

Lakshmipathy, U., *et al.* Micro RNA expression pattern of undifferentiated and differentiated human embryonic stem cells. *Stem cells and development* 2007;16(6):1003-1016.

Lewis, B.P., *et al.* Prediction of Mammalian MicroRNA Targets. *Cell* 2003;115(7):787-798.

Liang, Z., *et al.* mirAct: a web tool for evaluating microRNA activity based on gene expression data. *Nucleic Acids Research* 2011;39(Web Server issue):W139-W144.

Lin, D., *et al.* Group sparse canonical correlation analysis for genomic data integration. *BMC bioinformatics* 2013;14(1):1-16.

Lister, R., *et al.* Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science (New York, N.Y.)* 2013;341(6146).

Lister, R., *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;462(7271):315-322.

Lu, J., *et al.* MicroRNA expression profiles classify human cancers. *Nature* 2005;435(7043):834-838.

Nam, S., *et al.* MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Research* 2009;37(Web Server issue):W356-W362.

Nelson, K.M. and Weiss, G.J. MicroRNAs and cancer: past, present, and potential future. *Molecular cancer therapeutics* 2008;7(12):3655-3660.

Nilsen, T.W. Mechanisms of microRNA-mediated gene regulation in animal cells. *Trends in genetics : TIG* 2007;23(5):243-249.

Parkhomenko, E., Tritchler, D. and Beyene, J. Sparse Canonical Correlation Analysis with Application to Genomic Data Integration. In, *Statistical Applications in Genetics and Molecular Biology*. 2009. p. 1.

Poisson, L.M., Taylor, J.M. and Ghosh, D. Integrative set enrichment testing for multiple omics platforms. *BMC bioinformatics* 2011;12(1):1-11.

Ritchie, M.E., *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015.

Sales, G., *et al.* MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. *Nucleic Acids Research* 2010;38(Web Server issue):W352-W359.

Schaefer, C.F., *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Research* 2009;37(Database issue):D674-D679.

Seelan, R.S., *et al.* Developmental profiles of the murine palatal methylome. *Birth defects research. Part A, Clinical and molecular teratology* 2013;97(4):171-186.

Smyth, G.K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. In, *Statistical Applications in Genetics and Molecular Biology*. 2004. p. 1.

Subramanian, A., *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 2005;102(43):15545-15550.

Tatsuguchi, M., *et al.* Expression of microRNAs is dynamically regulated during cardiomyocyte hypertrophy. *Journal of molecular and cellular cardiology* 2007;42(6):1137-1141.

Tian, L., *et al.* Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America* 2005;102(38):13544-13549.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 1996:267-288.

Tibshirani, R., *et al.* Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005;67(1):91-108.

Ting, A.H., *et al.* Short double-stranded RNA induces transcriptional gene silencing in human cancer cells in the absence of DNA methylation. *Nature genetics* 2005;37(8):906-910.

Toedling, J., *et al.* Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC bioinformatics* 2007;8:221.

Tucker, K.L. Methylated cytosine and the brain: a new base for neuroscience. *Neuron* 2001;30(3):649-652.

Witten, D.M., Tibshirani, R. and Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 2009;10(3):515-534.

Witten, D.M. and Tibshirani, R.J. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Statistical Applications in Genetics and Molecular Biology* 2009;8(1):Article 28.

Wossidlo, M., *et al.* 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature communications* 2011;2:241.

Zhan, M., *et al.* MicroRNA expression dynamics during murine and human erythroid differentiation. *Experimental hematology* 2007;35(7):1015-1025.

CURRICULUM VITA

Dake Yang

Department of Biostatistics & Bioinformatics Phone: (502) 432-2129

University of Louisville Email:d0yang03@Louisville.edu

School of Public Health and Information Sciences

485 E. Gray St, Louisville, KY 40202

EDUCATION

- Ph.D. in Biostatistics Sep. 2011 - present

  Department of Biostatistics & Bioinformatics, University of Louisville

  Advisor: Guy Brock, Ph.D

- M.S. in Biostatistics Sep. 2009 - May 2011

  Department of Biostatistics & Bioinformatics, University of Louisville

  Advisor: Guy Brock, Ph.D

- B.S. in Statistics Sep. 2004 - May 2008

Department of Mathematics, Beijing Institute of Technology, P. R. China.

RESEARCH INTERESTS

- Integrated analysis of miRNA-mRNA data and mRNA-methylation data

- Empirical evaluation of methods to detect differentially expressed genes
- Statistical bioinformatics, genomics, and high dimensional data

PUBLICATIONS

Published

1. Yang, D., Parrish, R. S., & Brock, G. N. (2014). Empirical Evaluation of Consistency and Accuracy of Methods to Detect Differentially Expressed Genes Based on Microarray Data. *Computers in Biology and Medicine*, 46, 1–10. (Honorable Mention Paper (Top 10%), Computers in Biology and Medicine, 2014)
2. Li, M., Yang, D., Brock, G. N., Knipp, R. J., Bousamra, M., Nantz, M.H., and Fu, X. A. (2015). Breath Carbonyl Compounds as Biomarkers of Lung Cancer. Accepted to *Lung Cancer*.
3. Alton, T., Brock, G. N., Yang, D., Wilkings, D. A., Hertweck, S. P., Loveless, M. B. (2012). Retrospective Review of Intrauterine Device in Adolescent and Young Women. *Journal of Pediatric and Adolescent Gynecology*, 25(3), 195-200.

In progress

- Yang, D., Mukhopadhyay, P., Greene, R. M., Pisano, M. M., and Brock, G.N. (2015+).

Integrated Analysis of miRNA-mRNA Expression profiles using Sparse Canonical Correlation Analysis. (planned submission to *BMC Bioinformatics*).

- Mukhopadhyay, P., Brock, G., Yang, D., Greene, Robert. M., and Pisano, M. M. (2015+). Developmental gene expression profiling of mammalian, embryonic neural tube. (planned submission to *Birth Defects Research, part A*).

PRESENTATIONS

- Poster presentations

  — "Integrated analysis of miRNA-mRNA expression profiles," at *Joint Statistical Meetings*, Seattle, Washington. (August 2015, scheduled)

  — "Integrated analysis of miRNA-mRNA expression profiles," at *UT-KBRIN Bioinformatics summit*, Buchanan, Tennessee. (March 2015)

  — "Consistency of Differentially Expressed Gene Rankings Based on Subsets of Microarray Data," at *UT-ORNL-KBRIN Bioinformatics Summit*, Memphis, Tennessee. (March 2011)

RESEARCH EXPERIENCE

- Graduate Research Assistant Sep 2011 - present

Department of Biostatistics & Bioinformatics, University of Louisville

Methodological Research

— Develop methods for integrated analysis of miRNA- mRNA expression profiles. MicroRNAs (miRNAs) are a large number of small endogenous non-coding RNA molecules (18-25 nucleotides in length) which regulate expression of genes post-transcriptionally. While a variety of algorithms exist for determining the targets of miRNAs, they are generally based on sequence information and frequently produce lists consisting of thousands of genes. Canonical correlation analysis (CCA) is a multivariate statistical method that can be used to find linear relationships between two data sets, and here we apply CCA to find the linear combination of differentially expressed miRNAs and their corresponding target genes having maximal negative correlation. Due to the high dimensionality, sparse CCA is used to constrain the problem and obtain a solution. A novel gene set enrichment analysis statistic is proposed based on the sparse

CCA results for estimating the significance of predefined gene sets. The methods are illustrated with both a simulation study and real miRNA-mRNA expression data concerning the murine embryonic developing neural tube, and also other cancer data sets.

—

Collaborative Research

— Statistical support for the project “Confirmation of a VOC profile characteristic of lung cancer from exhaled human breath using chemoselective silicon microreactors,” (PI: Dr. Xiaoan Fu, J.B., Speed School of Engineering, University of Louisville), including developing classification models to assess the relationship between VOC markers and lung cancer cases / controls, and making ROC plots.

— Statistical support for Dr. Emma M. Sterrett (Family Therapy Program Kent School of Social Work, University of Louisville) to assess the relationship between various types of adolescent support and delinquency / drinking age.

— Statistical support for the project “Effects of Educational Intervention on Long-Term Outcomes of Hospitalized Children with Asthma,” (PI: Dr. Tania Condurache, Department of Pediatrics, University of Louisville School of Medicine), including testing for differences between the control and intervention groups on frequency of hospital visits, number of missed school days, etc.

— Statistical support for the project “Predicting Percent Weight Loss in BARIA Surgical Patients Using EGG 7 Weeks Following Surgery” (PI: Dr. Thomas L. Abell., Department of Medicine, Gastroenterology, Hepatology and Nutrition, University of Louisville) including analyzing the BARIA study data to develop a multivariable model to predict the percent weight loss for various study groups.

—

COMPUTATIONAL SKILLS

- Statistical software: proficient in R, SAS.
- Applications: proficient in MS Office and LaTeX.

HONORS AND AWARDS

- University Fellowship at University of Louisville. 2011 - 2013
- Outstanding Student Award, Beijing Institute of Technology, P. R. China. 2005 - 2008

REFERENCES

- Guy Brock, Ph.D.

Associate Professor, Department of Bioinformatics and Biostatistics, University of Louisville.

guy.brock@louisville.edu, (502) 852-3444.

- KB Kulasekera, Ph.D.

Chair and Professor, Department of Bioinformatics and Biostatistics, University of Louisville.

kb.kulasekera@louisville.edu, (502) 852-6422.

- Maiying Kong, Ph.D.

Associate Professor, Department of Bioinformatics and Biostatistics, University of Louisville.

maiying.kong@louisville.edu, (502) 852-3988.

- Dongfeng Wu, Ph.D.

Associate Professor, Department of Bioinformatics and Biostatistics, University of Louisville.

dongfeng.wu@louisville.edu, (502) 852-1888