A novel data mining method to identify differentially expressed gene signatures

by

Ai-Ling Teh

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE

Major: Statistics

Program of Study Committee:
Derrick K. Rollins, Major Professor
Dan Nettleton
Laura Jarboe

Iowa State University
Ames, Iowa
2010

Table of Contents

CHAPTER 1: INTRODUCTION

I. MICROARRAY

Living organisms are composed of complex gene structures that are hard to assign biological functions to. Even more challenging is attempting to understand their fundamental networking inside living cells. This raises the million dollar question. Scientists are interested in finding the various pathways and directions needed to solve the problems that are revealed. A molecular biology field called functional genomics (FG) was developed by biologists to study the behavior of complicated gene structures. FG represents genome analysis and more specifically a study of gene function through different high technology experiments [1]. FG involves the study of at least three areas: transcriptomics, proteomics, and metabolomics [2]. One of the popular experimental techniques used to study transcriptional genetic responses in this field is the microarray experiment technique [3-4]. The microarray experiment technique was first developed as a two-channel technology [3, 5] and is now widely used to explore and study the genome.

The Deoxyribonucleic acid (DNA) microarray typically involves hybridization of two messenger ribonucleic acid (mRNA) samples, in which they are converted into a cDNA sample. It is used primarily to study the difference in expression levels between two mRNA samples. Microarray data contains a wealth of information about thousands of genes that are hidden behind high noise levels and low signal levels. Thus, the main goal in this area is to explore and understand the biological behaviors and functions of these genes that play an important role in creating living organisms.

A good statistical method is always preferred to efficiently explore and extract the information hidden behind high noise levels and low signal levels. There are various

statistical methods proposed to analyze microarray data, yet scientists and statisticians are still collaborating to develop a more powerful method that is able to effectively and efficiently analyze high noise level microarray data. Several popular statistical methods that are widely used to analyze high dimensionality microarray data in the field today are principal component analysis (PCA) [2, 6-7], the simple linear model or mixed-effects linear model [4, 8], Student's *t*-test controlling for false discovery rate (FDR) based on p value [9-12], network component analysis (NCA) [13- 15], and the Wilcoxon Mann-Whitney [13, 16] and Bayesian method [17-19].

II. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a powerful statistical method used to analyze multivariate and large dimensionality data in a reduced dimensionality. PCA uses mathematical formulas and procedures to orthogonally transform the original data to a new coordinate data system that is represented by principal components (PCs) [20]. PCs are the linear combinations of eigenvectors and the variables. The PCA approach uses the covariance matrix (scaled sum of squares and cross products) or correlation matrix (sums of squares and cross products from standardized data) to generate the PCs. The PCA approach will rank the largest eigenvalue (i.e. maximum variance) as the first principal component (PC 1) and the smallest eigenvalue as the last PC. These PCs represent the contribution of each gene to the data and thus they are used in data analysis as weight factors. The main concern of PCA is explaining the variance-covariance structure of a set of variables through linear combinations of the variables [20].

PCA is a non-parametric analysis and the answers are not based on any hypothesis or probability distribution. The only assumption that PCA has is that all the

data are a linear combination of certain basis vectors [21]. Most statistical methods tend to eliminate partial important information and thus convey a misleading message. PCA, however, does not possess this problem because it is able to maximize the variability of data while minimizing the dimensionality of the data set.

III. MOTIVATION AND THESIS ORGANIZATION

Rollins et. al. (2006) proposed a novel PCA-based method to analyze microarray data and identify assay specific gene signatures in functional genomic studies. The proposed PCA-based method used PCA in two ways to explore the data: Eigengene (EG) where the genes are treated as variables and Eigenassay (EA) where the assays are treated as variables. This method is unique because it is the only method that uses gene contribution (product of loading and gene expression level) to create rank ordered assay specific gene signatures.

Overall, this method performed very well in analyzing microarray data and in identifying gene(s) that are important to a particular physiological condition of interest. But one limitation of this method is its ability to produce a good differentially expressed gene signature that expressed the differences between two assay groups. We are looking to improve the assay-specific gene signature produced from this method so that we can obtain a signature that looks like an “L-shape” to indicate the top genes that are really showing the differences while the rest of the genes show zero differences.

Hence, with these ideas in mind, we have extended the work of Rollins et al. (2006) and developed a method that can improve the performance of the PCA-based method in identifying assay-specific and differentially expressed gene signatures in

functional genomic studies. This research not only improved the PCA-based method proposed by Rollins et. al.(2006) but also developed two new test statistics that are able to identify differentially expressed gene signatures. More specifically, this work proposed two new PCA-based test statistics that have high statistical power and low false discovery rate.

To this end, this thesis is organized as follows: Chapter 2 gives brief reviews on various statistical methods that are adopted for mining of microarray data. Chapter 3 provides the ideas and details on the two new test statistics developed. Chapter 4 gives an illustration of applying the proposed method to a real data set followed by results and discussion. Last but not least, Chapter 5 gives general conclusions about the work of this research and also suggests a few ideas for future research.

IV. REFERENCES

1. Hieter P, Boguski M: **Functional Genomics: It's All How You Read It**, *Functional Genomics* **1997, 278:** 601-602.

2. Rollins D K, Zhai D, Joe AL, Guidarelli JW, Murarka A, Gonzalez R: **A novel data mining method to identify assay-specific signatures in functional genomic studies**, *BMC Bioinformatics* **2006, 7**:377.

3. Cordero F, Botta M, Calogero RA: **Microarray Data Analysis and Mining Approaches**, *Briefings in Functional Genomics and Proteomics* **2007, 6**: 265-281.

4. Kerr MK, Churchill GA: **Statistical Design and the Analysis of Gene Expression Microarray Data**, *Genetic Research* **2001, 77:** 123-128.

5. Shalon D, Smith SJ, Brown PO: **A DNA Microarray System for Analyzing Complex DNA Samples Using Two-Color Fluorescent Probe Hybridization**, *Genome Res* **1996, 6**: 639-45.

6. Raychaudhuri S, Stuart JM, Altman RB: **Principal Component Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series**, *Pac Symp Biocomput* **2009, 2000:** 455-466.

7. Liu A, Zhang Y, Gehan E, and Clarke R: **Block Principal Component Analysis with Application to Gene Microarray Data Classification**, *Statist. Med* **2002, 21:** 3465-3474.

8. Nettleton D: **A Discussion of Statistical Methods for Design and Analysis of Microarray Experiments for Plant Scientists**, *The Plant Cell* **2006, 18:** 2112-2121.

9. Steelman CA, Recknor JC, Nettleton D, and Reecy JM: **Transcriptional profiling of myostatin-knockout mice implicates Wnt signaling in postnatal skeletal muscle growth and hypertrophy**, *The FASEB Journal* **2006, 20**: 580-582.

10. Storey JD, and Tibshirani R: **Statistical significance for genomewide studies**, *Proc. Natl. Acad. Sci. USA* **2003, 100**: 9440-9445.

11. Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments**, *Stat. Sinica* **2002, 12**: 111–139.

12. Ge Y, Sealfon SC, Speed TP: **Multiple Testing and Its Applications to Microarray,** *Statistical Methods in Medical Research* **2009, 18:** 543.

13. Hyduke DR, Jarboe LR, Tran LM, Chou KJY, and Liao JC: **Integrated network analysis identifies nitric oxide response networks and dihydroxyacid dehydratase as a crucial target in *Escherichia coli***, *Proceedings of the National Academy of Sciences of the United States of America* **2007**, **104**: 8488-8489.

14. Liao JC, Boscolo R, Yang YL, Tran LY, SAbatti C, Roychowdhury VP: **Network component analysis: Reconstruction of regulatory signals in biological systems**, *Proceedings of National Academy of Sciences* **2003, 100:** 15522-15527.

15. Tran LM, Brynildsen MP, Kao KC, Suen JK, Liao JC: **gNCA: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation,** *Metabolic Engineering* **2005, 7:** 128-141.

16. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniem OP, Wilfond B, Borg A, Trent J: **Gene-expression profiles in hereditary breast cancer,** *The New England Journal of Medicine* **2001, 344:** 539-548.

17. Sartor MA, Tomlinson CR, Wesselkamper SC, Sivaganesan S, Leikauf GD, Medvedovic M: **Intensity-based Hierarchical Bayes Method Improves Testing for Differentially Expressed Genes in Microarray Experiments**, *BMC Bioinformatics* **2006, 7:** 538.

18. Kendziorski CM, Newton MA, Lan H, Gould MN: **On Parametric Empirical Bayes Methods for Comparing Multiple Groups Using Replicated Gene Expression**

Profiles, *Statistics in Medicine* **2003, 22**: 3899-3914.

19. Efron B, Tibshirani R, Storey JD, Tusher V: **Empirical Bayes Analysis of a Microarray Experiment**, *Journal of American Statistical Association* **2001, 96:** 456.

20. Johnson RA and Wichern DW: **Applied Multivariate Statistical Analysis**. 6th edition. Pearson Prentice Hall; 2008: 430 – 470.

21. Shlens J: **A Tutorial on Principal Component Analysis** [http://www.snl.salk.edu/~shlens/pca.pdf]

CHAPTER 2: LITERATURES REVIEW

This chapter will discuss a few common statistical methods applied by the scientists to perform microarray data analysis and interpretation. This chapter is broken down into several sections to briefly describe and discuss some general statistical methods such as PCA, the simple linear model or mixed effect model, the Student's *t*-test controlling for FDR, the Bayesian method, and network component analysis (NCA). There are a lot more statistical methods packages that are available for microarray data analysis but for the purpose of this thesis, I will only discuss a few methods that I overviewed.

I. PRINCIPAL COMPONENT ANALYSIS

Rollins et. al(2006) used the PCA-based method to analyze large dimensionality microarray data to identify assay-specific gene signatures in functional genomics studies. This method creatively used gene contribution, which is a product of the linear combination of the loading and gene expression level to obtain the desired signatures. Their method also employed the PCA approach in two different methods. The first method treated genes as variables and was called the Eigengene (EG) method, while the second method treated assays as the variables and was called the Eigenassay (EA) method. This method was proven to successfully identify assay-specific gene signatures for functional genomics studies.

Liu et. al. (2002) proposed block PCA to analyze microarray data. The uniqueness of this method is that PCA was applied to the principal components instead of the original variables [2]. The variables or genes in the original data were first grouped into

several blocks based on their similarity in behavior and functionality. The PCA was then performed on each block to obtain a small number of variance-dominating PCs. These PCs were then combined together to form a data set and PCA was performed on this "data" again. They showed that block PCA was an effective method to identify genes that showed insight into cancer phenotypes in their studies.

A PCA based approach is widely applied because this method does not assume a distribution for the data and the only assumption it has is that all the data can be represented as a linear combination of certain basis vectors [3]. The PCA approach also provides researchers with the percentage of variation accounted for by the data by choosing a particular PC [4].

II. LINEAR MODEL, MIXED EFFECT MODEL, STUDENT'S *T*- TEST

The second method described here is the simple linear model and the mixed effect model incorporating the Student's *t* test when a study is looking to identify genes that expressed differentially from other genes across two or more different conditions. A *t*-statistic is used by some researchers to test for significant differences expressed by a gene across two different conditions. The main idea behind this method is to obtain a p-value for each of the genes involved and then make a decision with the help of test statistics and the p-value. Ge et. al. (2009) used Welch *T*-statistics, which are an adaptation from the Student's *t*-test, to perform the data analysis and control the false discovery rate (FDR).

In this particular study, the p-value typically corresponds to a null hypothesis stating that the genes are not differentially expressed across two or more assay groups.

Since the p-value is often misunderstood to be the probability that the null hypothesis is correct, most researchers believe that a large p-value corresponds to a gene that is not differentially expressed from the other across two conditions. Even though this allows for straight forward decision making, the interpretation of the p-value is not correct because p-value is computed under the assumption that the null hypothesis is true.

Storey and Tibshirani (2003) have converted the p-value to the q-value to control for false discovery rate (FDR) [6]. Let me describe a simple example that will help to illustrate the meaning of the FDR. For example, let 200 genes, identified to be differentially expressed compared to all other genes, have a 5% FDR. Then there is a chance that 10 out of the 200 genes are not really differentially expressed as claimed. The q-value on the other hand can control the FDR easily because the q-value is inversely related to the p-value. Thus, if a researcher wishes to have a 5% FDR, he or she can declare all genes with a q-value less than 0.05 to be differentially expressed. Hence, controlling the gene signature using the FDR is a very popular approach since it can help to minimize the false positive rate.

III. NETWORK COMPONENT ANALYSIS (NCA)

The next method that I would like to describe here is the network component analysis (NCA) method. This method was first proposed by Liao et. al. (2003) after validating this method experimentally through spectra absorbance of network components for various hemoglobin species [7]. This method has the ability to uncover the hidden regulatory signals from the output of network systems [7]. This method decomposes the data set into two matrices, where the first matrix consists of regulatory

signals and the second matrix encodes the connectivity strengths between the regulatory layer and the output signals. The advantage of this method is its ability to properly handle the prior knowledge of a structure characterizing a given system without making any assumptions on the statistical properties.

Liao et. al. (2003) applied this method to microarray data generated from the yeast *Saccharamyces cerevisiae* and the activity of various transcription factors during the cell cycle. Using the NCA method, they successfully generated a network consistent representation of the regulatory signals [7]. Tran et. al. (2005) have further extended the NCA method to the generalized NCA (gNCA) method. gNCA incorporates the regulatory signal constraints arising from genetic knockouts [8]. The gNCA method was applied on a microarray data set to study the effect of transcription factor activities for an *E. coli* regulator (*arcA*) deletion mutant during the glucose acetate transition [8]. Incorporating both network structure constraints and signal constraints imposed by genetic knockouts, gNCA was able to identify transcription factors that expressed differently between wild type strain and mutant strain.

IV. OTHER METHODS

Besides the three statistical methods described above, there are many more statistical methods that are being used to analyze microarray data with high dimensionality. Some researchers have selected the Bayesian approach for data analysis because this method is able to effectively analyze the data while reducing the dimensionality of the original data set [9-10]. Efron et. al. (2001) and Kendsiorski et. al. (2003) used the empirical Bayesian method to analyze their microarray data. The

Empirical Bayesian method is similar to the regular Bayesian method. The only difference is that the prior distribution is not independent of the data as in the typical Bayesian method. For the empirical Bayesian method, the prior distribution is estimated from the data itself.

Non-parametric models such as the Wilcoxon Mann-Whitney [12] and Wilcoxon Signed Rank [13] tests are favored when the data are not normally distributed or cannot be assumed to have a normal distribution. The Wilcoxon Mann-Whitney test, sometimes called the Wilcoxon rank-sum test, is used when two samples are independent of each other, while the Wilcoxon signed rank test is applied when two samples are related or a repeated sample is taken on the same sample.

V. REFERENCES

1. Rollins D K, Zhai D, Joe AL, Guidarelli JW, Murarka A, Gonzalez R: **A novel data mining method to identify assay-specific signatures in functional genomic studies**, *BMC Bioinformatics* **2006, 7**:377.

2. Liu A, Zhang Y, Gehan E, and Clarke R: **Block Principal Component Analysis with Application to Gene Microarray Data Classification**, *Statist. Med* **2002, 21:** 3465-3474.

3. Shlens J: **A Tutorial on Principal Component Analysis** [http://www.snl.salk.edu/~shlens/pca.pdf]

4. Raychaudhuri S, Stuart JM, Altman RB: **Principal Component Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series**, *Pac Symp Biocomput* **2009, 2000:** 455-466.

5. Ge Y, Sealfon SC, Speed TP: **Multiple Testing and Its Applications to Microarray,** *Statistical Methods in Medical Research* **2009, 18:** 543.

6. Storey JD, Tibshirani R: **Statistical Significance for Genomewide Studies**, *Proceedings of the National Academy of Sciences* **2003, 100:** 9440-9445.

7. Liao JC, Boscolo R, Yang YL, Tran LY, Sabatti C, Roychowdhury VP: **Network component analysis: Reconstruction of regulatory signals in biological systems**,

Proceedings of National Academy of Sciences **2003, 100:** 15522-15527.

8. Tran LM, Brynildsen MP, Kao KC, Suen JK, Liao JC: **gNCA: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation,** *Metabolic Engineering* **2005, 7:** 128-141.

9. Efron B, Tibshirani R, Storey JD, Tusher V: **Empirical Bayes Analysis of a Microarray Experiment**, *Journal of American Statistical Association* **2001, 96:** 456.

10. Kendziorski CM, Newton MA, Lan H, Gould MN: **On Parametric Empirical Bayes Methods for Comparing Multiple Groups Using Replicated Gene Expression Profiles**, *Statistics in Medicine* **2003, 22**: 3899-3914.

11. Sartor MA, Tomlinson CR, Wesselkamper SC, Sivaganesan S, Leikauf GD, Medvedovic M: **Intensity-based Hierarchical Bayes Method Improves Testing for Differentially Expressed Genes in Microarray Experiments**, *BMC Bioinformatics* **2006, 7:** 538.

12. Hyduke DR, Jarboe LR, Tran LM, Chou KJY, and Liao JC: **Integrated network analysis identifies nitric oxide response networks and dihydroxyacid dehydratase as a crucial target in *Escherichia coli***, *Proceedings of the National Academy of Sciences of the United States of America* **2007**, **104**: 8488-8489.

13. Liu W, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho M, Baid J, Smeekens SP: **Analysis of High Density Expression Microarrays With Signed-Rank Call Algorithms**, *Bioinformatics* **2002, 18**: 1593-1599.

CHAPTER 3: AN EXTENDED DATA MINING METHOD FOR IDENTIFYING DIFFERENTIALLY EXPRESSED ASSAY-SPECIFIC SIGNATURES IN FUNCTIONAL GENOMIC STUDIES

A paper to be submitted to *BMC Bioinformatics*

Derrick K Rollins, AiLing Teh, and Dan Nettleton

ABSTARCT

Background: Microarray data sets provide relative expression levels for thousands of genes for a small number, in comparison, of different experimental conditions called *assays*. Data mining techniques are used to extract specific information of genes as they relate to the assays. The multivariate statistical technique of principal component analysis (PCA) has proven useful in providing effective data mining methods. This article extends the PCA approach of Rollins et al. (2006) to the development of ranking genes of microarray data sets that ***express most differently*** between two biologically different grouping of assays. This method is evaluated on real and simulated data and compared to a current approach on the basis of false discovery rate (FDR) and statistical power (SP) which is the ability to correctly identify important genes.

Results: This work developed and evaluated two new test statistics based on PCA and compared them to a popular method that is not PCA based. Both test statistics were found to be effective as evaluated in three case studies: (i) exposing *E.* coli cells to two different ethanol levels; (ii) application of myostatin to two groups of mice; and (iii) a simulated data study derived from the properties of (ii). The proposed method (PM) effectively identified critical genes in these studies based on comparison with the current method (CM). The simulation study supports higher identification accuracy for PM over CM for

both proposed test statistics when the gene variance is constant and for one of the test statistics when the gene variance is non-constant.

Conclusions: PM compares quite favorably to CM in terms of lower FDR and much higher SP. Thus, PM can be quite effective in producing accurate signatures from large microarray data sets for differential expression between assays groups identified in a preliminary step of the PCA procedure and is, therefore, recommended for use in these applications.

I. INTRODUCTION

It is well known that living organisms have complicated gene structures. However, while major advancements have been made in recent years, understanding of the biological functions of each individual gene is still quite limited. Active research is strongly focused on understanding the behavior of genes and as well as the highly complex metabolism and regulatory network inside living cells [1]. This effort falls under a molecular biological field called functional genomics (FG). There are at least three areas in which experimental techniques are widely applied in FG: transcriptomics, proteomics, and metabolomics [2]. A combination of leading scientific techniques as well as powerful mathematical and statistical tools for data analysis makes the task of identifying important transcriptome, proteome, and metabolome corresponding to a biological effect promising. Typical studies in these areas involve the identification of possible behavior and responses of species under various genetic backgrounds as well as environmental factors (i.e. assay).

There are different high technology techniques applied in FG field to advance understanding of the transcriptional genetic response of many organisms in various environmental perturbations [1]. One of the techniques that have been adopted in this field is a multiplex technology called DNA microarray [3]. A new technique that is becoming popular and will probably displace array-based measurement in FG is next-generation sequencing (RNAseq) [4-5]. These techniques have the ability to generate data sets that consist of expression levels of thousands of genes, providing a wealth of information that is hidden by high noise levels, low signal levels, and a relatively small number of experimental units to the number of genes studied. More specifically, since the data set containing the gene expression measurements consists of a lot more genes than assays, analytical techniques are needed to provide accurate gene identification under a large number of gene candidates that is much greater than the number of experimental runs.

To achieve this objective, traditional statistical methods, such as principal component analysis (PCA) [2, 3-8], the focus of this article, are being retrofitted to provide effective statistical inference in this challenging context of microarray data analysis. Other methods included linear model analysis [9-14], Bayesian method [17 - 19] and neural network analysis (NCA) [18-20]. Thus, statistics is playing a critical role through the development of methodologies that give high statistical power (SP) (i.e., accurate identification), and low false discovery rate (FDR) [21] (i.e. low misidentification). To this end, this article introduces two new PCA based statistics for determining gene rank for *differential expression* between two PCA identified assay groups. This work extends the technique introduced by Rollins et al. (2006) that

determines gene rank for a *single* PCA identified assay group. Thus, the proposed method (PM) in this work is aimed at finding the genes with high expression levels in one group and low expression levels in the other group.

The PM method uses PCA to first establish the existence of the assay groupings of interest. Then using the results that established the grouping, the differential contribution for each gene is determined using a statistic based on eigenvalues. This article proposes and evaluates two statistics. The first one is the group averaged difference of eigenvalue linear combinations that we call T_{diff}. The second one divides T_{diff} by its estimated pooled standard deviation that we call T_{scaled}. The genes are ranked based on the largest absolute value of these statistics. The PM is evaluated against the ranking determined by the well known Student's *t*-statistic [14] that we call T_{pooled} in this work. We will refer to T_{pooled} as the current method (CM) which is actually a subclass of the PM that weighs each assay equally in each group. Note that for the CM the assay members in each group is not established based on the data but by á priori considerations. In contrast, for the PM the data drives the assay weight as well as group assignment of the assays.

The CM and PM are applied in the following three case studies to compare their effectiveness (i.e., power) in identifying assay-specific signature: (i) exposure of *E. coli* cells to two different levels of ethanol concentration [2]; (ii) the use of myostatin as inhibitor of skeletal muscle growth for five 5-weeks-old myostatin and non-treated mice [9]; and (iii) a simulation study based on statistical properties of the second case study.

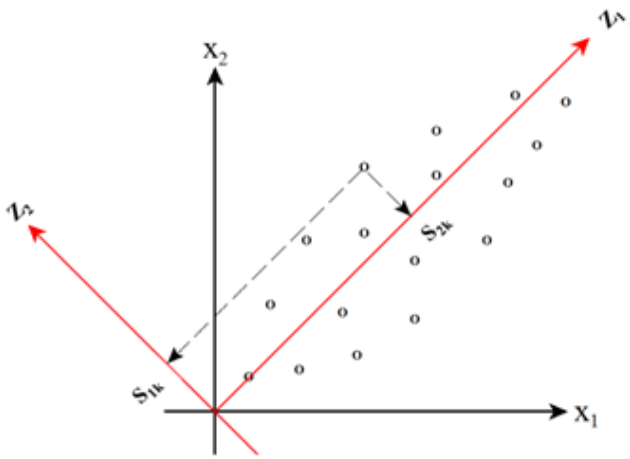
This work is organized into the following section. The Background Section gives a brief review of PCA and connects it to our application in FG's data analysis. This section is followed by the Methods Section that derives and presents the test statistics of the CM and PM. These test statistics are evaluated and compared in three studies in the Results and Discussion Section. The final section summarizes the results and gives concluding remarks on the contribution of this work.

II. BACKGROUND

The microarray data set is given as an m by n matrix $\mathbf{X}$ where n is the number of assays expressed along columns (i.e. variables) and m represents the number of genes expressed along rows. The cells in this matrix are given as x_{ij} which is the expression level of the i^{th} gene for the jth assay (i.e. condition). Principal component analysis (PCA) is a multivariate technique that mathematically transforms (rotates) the original coordinate system to a new orthogonal coordinate system based on correlations among the variables [20]. The principal components (PCs) are eigenvectors generated from either the covariance matrix (scaled sum of squares and cross products) or the correlation matrix (sums of squares and cross products from standardized data) of the variables involved. They are used to construct n linear combinations of the n variables that can be thought of as n pseudo variables [20]. A PC is rank ordered by the amount of variation in the original data set that it captures.

An illustration is given in Fig. 1 that shows a visual representation of a two-dimensional data system (x) and a rotated data system (z). As shown, the new coordinated system points z_1 in the direction with the greatest spread in the data. The other variable, z_2,

points in a direction that is orthogonal to z_1, but also seeks to maximize spread in this direction. The first PC determines z_1 and the second PC determines z_2. A data matrix of rank *n* will give *n* PCs that are linear combinations of the variables in the original data matrix that can be described as *n pseudo* variables. The goal in this application of PCA is to obtain at least one pseudo variables that represent the biological behavior of interest. This can be a PC that represents a small portion of the total variation making it a potentially very powerful data mining approach.



$$Cov(X_1, X_2) = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

$$Cov(Z_1, Z_2) = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

Figure 1. Visual representation of original data system and the rotating data system.

The figure represented the original data system on the horizontal and vertical axis while the new rotating data system is represented as z_1 and z_2. Variance-covariance matrices for original data system and rotating data system are showed on the right of the figure.

The top of Fig. 2 shows the relationship between the original data matrix, **X**, the *n* by *n* PC loading matrix, **L**, and the *m* by *n* pseudo data matrix, called the scores matrix, **S**. The PCs derived from **X** are called eigengenes (EG) because the elements of **S** represent pseudo values for gene expression. In Fig. 2 the bottom set of matrices are derived from the transpose of **X** which is an *n* by *m* matrix. In this case the loading matrix is *m* by *n* in

dimension and the scores matrix is *n* by *n* in dimension. The PCs derived from the transpose of **X** are called eigenassays (EA) because the elements of the scores matrix represent pseudo assays. The proposed method (PM), following Rollins et al. (2006), uses both EG and EA approaches to develop signatures sets of ranked genes. In the next section we derive the EG and EA statistics for determining gene contribution for the PM.

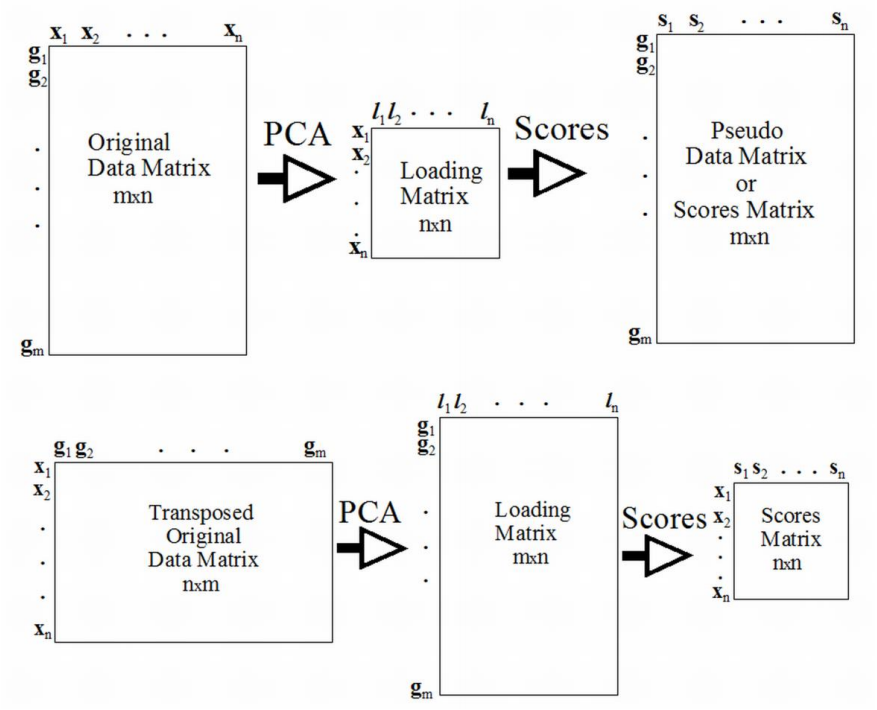


Figure 2. Visual representation of data, loading, and score matrices for X and X^T

The top matrices are for **X** and determining EGs and the bottom ones for $\mathbf{X}^{\mathbf{T}}$ and for determining EAs.

III. METHODS

(a) Eigengene Contribution Approach

The first step in the eigengene (EG) approach of the PM is to standardize the elements of **X** to give the standardized matrix **Z** with each element equal to

$z_{ij} = \left(x_{ij} - \bar{x}_j\right) / s_j$, where $\bar{x}_j$ and s_j are the sample mean and sample standard deviation of the data in column *j*, respectively. The following distributional assumptions apply: $x_{ij} \overset{indep}{\sim} N\left(\mu_{x_j}, \sigma^2_{x_i}\right)$, $z_{ij} \overset{indep}{\sim} N\left(\mu_{z_j}, \sigma^2_{z_i}\right)$, $\bar{x}_j \overset{indep}{\sim} N\left(\mu_{x_j}, \sigma^2_j / m\right)$, and $\mathrm{E}[\, s_j^2\,] = \sigma_j^2 \;\forall\; j$. These assumptions indicate that each assay can have it own mean expression level, $\mu_{x_j}\ (j = 1, \ldots, n)$, and that the variance of each gene is constant across assays but can be different for different genes. Also, $\mu_{z_j} = 0 \;\forall i, j$ since $\mathrm{E}[x_{ij}] = \mathrm{E}[\,\bar{x}_j\,] = \mu_{x_j} = \forall\, i, j$. These assumptions will be utilized later after proposing the test statistics.

The elements of the EG scores matrix, $\mathbf{S}^{\mathrm{EG}}$, are determined by

$$\begin{aligned} s_{ij}^{EG} &= \ell_{1j}^{EG} z_{i1} + \ell_{2j}^{EG} z_{i2} + \ldots + \ell_{nj}^{EG} z_{in} = \sum_{k=1}^{n} \ell_{kj}^{EG} z_{ik} \\ &= g_{ij1}^{EG} + g_{ij2}^{EG} + \ldots + g_{ijn}^{EG} = \sum_{k=1}^{n} g_{ijk}^{EG}; \qquad i = 1,\ldots,m;\ j = 1,\ldots,n;\ k = 1,\ldots,n \end{aligned} \tag{1}$$

where s_{ij}^{EG} is the score for the i^{th} gene using the j^{th} vector of EG loadings, ℓ_{ij}^{EG} is the i^{th} loading for the j^{th} EG vector, and g_{ijk}^{EG} is the contribution for the i^{th} gene, on the k^{th} assay from the j^{th} EG loading vector. Let A = Group A with n_A assay members and B = Group B with n_B assay members with no members in common with Group A such that

$$2 \le n_A + n_B \le n \tag{2}$$

The mean contribution for i^{th} gene from the j^{th} EG loading vector for Groups A and B, respectively are

$$\bar{g}_{ij}^{EG_A} = \frac{1}{n_A} \sum_{over\ k'} \ell_{k'j}^{EG} z_{ik'} = \frac{1}{n_A} \sum_{over\ k'} g_{ijk'}^{EG} \tag{3}$$

$$\bar{g}_{ij}^{EG_B} = \frac{1}{n_B} \sum_{over\ k''} \ell_{k'' j}^{EG} z_{ik''} = \frac{1}{n_B} \sum_{over\ k''} g_{ijk''}^{EG} \tag{4}$$

where k' and k'' are the assay members in Groups A and B, respectively. Finally, the EG differential gene contribution between Groups A and B for the i^{th} gene from the j^{th} EG loading vector is given as

$$d\bar{g}_{ij}^{EG} = \bar{g}_{ij}^{EG_A} - \bar{g}_{ij}^{EG_B} \tag{5}$$

The basic difference between the method in Rollins et al. (2006) and this extension is that work developed gene signatures for individual group using equations of the form given by (3) and (4) and this work uses equation of the form given by Eq. 5.

(b) Eigenassay Contribution Approach

As stated above, the EA approach uses the transpose of **X** as the data matrix treating the genes as the variables. Following Rollins et al. (2006), $\mathbf{X}^T$ is not standardized in the EA approach as in the EG approach. The elements of scores matrix, $\mathbf{S}^{EA}$, are determined from Eq. 6 as follows:

$$\begin{aligned} s_{ij}^{EA} &= \ell_{1j}^{EA} x_{1i} + \ell_{2j}^{EA} x_{2i} + \ldots + \ell_{mj}^{EA} x_{mi} \\ &= \sum_{p=1}^{m} \ell_{pj}^{EA} x_{pi} = \sum_{p=1}^{m} g_{ijp}^{EA}; \qquad i = 1,\ldots,n;\ j = 1,\ldots,n;\ p = 1,\ldots,m \end{aligned} \tag{6}$$

where s_{ij}^{EA} is the score for the i^{th} assay using the j^{th} vector of EA loadings, ℓ_{ij}^{EA} is the i^{th} loading for the j^{th} EA vector, and g_{ijp}^{EA} is the contribution for the p^{th} gene, on the i^{th} assay from the j^{th} EA loading vector. As above, for A = Group A with n_A assay members and B

= Group B with n_B assay members with no members in common with Group A, we obtain the contribution expressions as follows. The mean contribution for p^{th} gene from the j^{th} EA loading vector for Groups A and B are

$$\bar{g}_{jp}^{EA_A} = \frac{\ell_{pj}^{EA}}{n_A} \sum_{over\, i'} x_{pi'} = \frac{1}{n_A} \sum_{over\, i'} g_{i'jp}^{EA} \tag{7}$$

$$\bar{g}_{jp}^{EA_B} = \frac{\ell_{pj}^{EA}}{n_B} \sum_{over\, i''} x_{pi''} = \frac{1}{n_B} \sum_{over\, i''} g_{i''jp}^{EA} \tag{8}$$

respectively, where i' and i'' represent the assay members in Groups A and B, respectively. Finally, the EA differential gene contribution between Groups A and B for the p^{th} gene from the j^{th} EG loading vector is given as

$$d\bar{g}_{jp}^{EA} = \bar{g}_{jp}^{EA_A} - \bar{g}_{jp}^{EA_B} \tag{9}$$

(c) Test Statistics

The next step after deriving the gene contribution equations is to define the decision or test statistics based on these derivations. T_{diff} for EG and EA are equivalent to Eqs. 5 and 9, respectively. More specifically,

$$T_{diff_{ij}}^{EG} = d\bar{g}_{ij}^{EG} = \bar{g}_{ij}^{EG_A} - \bar{g}_{ij}^{EG_B} = \frac{1}{n_A} \sum_{over\, k'} \ell_{k'j}^{EG} z_{ik'} - \frac{1}{n_B} \sum_{over\, k''} \ell_{k''j}^{EG} z_{ik'} \tag{10}$$

$$T_{diff_{jp}}^{EA} = d\bar{g}_{jp}^{EA} = \bar{g}_{jp}^{EA_A} - \bar{g}_{jp}^{EA_B} = \frac{\ell_{pj}^{EA}}{n_A} \sum_{over\, i'} x_{pi'} - \frac{\ell_{pj}^{EA}}{n_B} \sum_{over\, i''} x_{pi''} = \ell_{pj}^{EA} \left(\bar{x}_{Ap} - \bar{x}_{Bp} \right) \tag{11}$$

The variances for the components of these equations are given below by treating the loadings as fixed variables:

$$V\left(\bar{g}_{ij}^{EG_A}\right)=V\left(\frac{1}{n_A}\sum_{over\ k'}\ell_{k'j}^{EG}z_{ik'}\right)=\frac{1}{n_A^2}\sum_{over\ k'}\left(\ell_{k'j}^{EG}\right)^2\sigma_{z_i}^2=\frac{\sigma_{z_i}^2}{n_A^2}\sum_{over\ k'}\left(\ell_{k'j}^{EG}\right)^2 \quad (12)$$

$$V\left(\bar{g}_{ij}^{EG_B}\right)=V\left(\frac{1}{n_B}\sum_{over\ k''}\ell_{k''j}^{EG}z_{ik''}\right)=\frac{1}{n_B^2}\sum_{over\ k''}\left(\ell_{k''j}^{EG}\right)^2\sigma_{z_i}^2=\frac{\sigma_{z_i}^2}{n_B^2}\sum_{over\ k'}\left(\ell_{k''j}^{EG}\right)^2 \quad (13)$$

$$V\left(\bar{g}_{jp}^{EA_A}\right)=V\left(\frac{\ell_{pj}^{EA}}{n_A}\sum_{over\ i'}x_{pi'}\right)=\frac{\left(\ell_{pj}^{EA}\right)^2}{n_A^2}n_A\sigma_{xp}^2=\left(\ell_{pj}^{EA}\right)^2\frac{\sigma_{xp}^2}{n_A} \quad (14)$$

$$V\left(\bar{g}_{jp}^{EA_B}\right)=V\left(\frac{\ell_{pj}^{EA}}{n_B}\sum_{over\ i''}x_{pi''}\right)=\frac{\left(\ell_{pj}^{EA}\right)^2}{n_B^2}n_B\sigma_{xp}^2=\left(\ell_{pj}^{EA}\right)^2\frac{\sigma_{xp}^2}{n_B} \quad (15)$$

Thus, combining Eqs. 10-11, the variances for $T_{diff_{ij}}^{EG}$ and $T_{diff_{jp}}^{EA}$ respectively are:

$$V\left(T_{diff_{ij}}^{EG}\right)=\frac{\sigma_{z_i}^2}{n_A^2}\sum_{over\ k'}\left(\ell_{k'j}^{EG}\right)^2+\frac{\sigma_{z_i}^2}{n_B^2}\sum_{over\ k''}\left(\ell_{k''j}^{EG}\right)^2=\sigma_{z_i}^2\left[\frac{1}{n_A^2}\sum_{over\ k'}\left(\ell_{k'j}^{EG}\right)^2+\frac{1}{n_B^2}\sum_{over\ k''}\left(\ell_{k''j}^{EG}\right)^2\right] \quad (16)$$

$$V\left(T_{diff_{pj}}^{EA}\right)=\left(\ell_{pj}^{EA}\right)^2\frac{\sigma_{xp}^2}{n_A}+\left(\ell_{pj}^{EA}\right)^2\frac{\sigma_{xp}^2}{n_B}=\left(\ell_{pj}^{EA}\right)^2\sigma_{xp}^2\left[\frac{1}{n_A}+\frac{1}{n_B}\right] \quad (17)$$

The scale test statistic in the EG case can now be given by dividing Eq. 10 by the estimated standard deviation using Eq. 16:

$$T_{scale_{ij}}^{EG}=\frac{T_{diff_{ij}}^{EG}}{\left[\hat{V}\left(T_{diff_{ij}}^{EG}\right)\right]^{1/2}}=\frac{\frac{1}{n_A}\sum_{over\ k'}\ell_{k'j}^{EG}z_{ik'}-\frac{1}{n_B}\sum_{over\ k''}\ell_{k''j}^{EG}z_{ik'}}{s_{pooled\ z_i}\sqrt{\left[\frac{1}{n_A^2}\sum_{over\ k'}\left(\ell_{k'j}^{EG}\right)^2+\frac{1}{n_B^2}\sum_{over\ k''}\left(\ell_{k''j}^{EG}\right)^2\right]}} \quad (18)$$

where

$$s_{pooled\ z_i}^2=\frac{n_A-1}{n_A+n_B-2}s_{Az_i}^2+\frac{n_B-1}{n_A+n_B-2}s_{Bz_i}^2 \quad (19)$$

$s_{Az_i}^2$ and $s_{Bz_i}^2$ are the sample variances for the standardized expression levels for Groups A and B, respectively, corresponding to the i^{th} gene. Note that when

$x_{ij} \overset{indep}{\sim} N\left(\mu_{x_j}, \sigma^2\right), \forall\, i, j,$ then $\bar{x}_j \overset{indep}{\sim} N\left(\mu_{x_j}, \sigma^2 / m\right) \forall\, j.$ Therefore,

$z_{ij} = \left(x_{ij} - \bar{x}_j\right) / s_j \overset{indep}{\sim} N(0,1)$, since $\bar{x}_j \approx \mu_{x_j}$ and $s_j^2 \approx \sigma^2\ \forall\ j$ because *m* is very large. In this case where the variation of the assays are all similar, V(z_{ij}) is taken to equal 1 and

$$T_{scale_{ij}}^{EG} = \frac{\frac{1}{n_A}\sum_{over\,k'} \ell_{k'j}^{EG} z_{ik'} - \frac{1}{n_B}\sum_{over\,k''} \ell_{k''j}^{EG} z_{ik'}}{\sqrt{\left[\frac{1}{n_A^2}\sum_{over\,k'}\left(\ell_{k'j}^{EG}\right)^2 + \frac{1}{n_B^2}\sum_{over\,k''}\left(\ell_{k''j}^{EG}\right)^2\right]}} \tag{20}$$

Similarly, the scaled test statistic in the EA case can also be given now by dividing Eq. 11 by the estimated standard deviation using Eq. 17:

$$T_{scale_{jp}}^{EA} = \frac{T_{diff_{jp}}^{EA}}{\left[\hat{V}\left(T_{diff_{jp}}^{EA}\right)\right]^{1/2}} = \frac{\ell_{pj}^{EA}\left[\frac{1}{n_A}\sum_{over\,i'} x_{pi'} - \frac{1}{n_B}\sum_{over\,i''} x_{pi''}\right]}{\ell_{pj}^{EA} s_{pooled\,x_p}\sqrt{\left[\frac{1}{n_A} + \frac{1}{n_B}\right]}} = \frac{\bar{x}_{A_p} - \bar{x}_{B_p}}{s_{pooled\,x_p}\sqrt{\left[\frac{1}{n_A} + \frac{1}{n_B}\right]}} \tag{21}$$

where

$$s_{pooled\,x_p}^2 = \frac{n_A - 1}{n_A + n_B - 2} s_{Ax_p}^2 + \frac{n_B - 1}{n_A + n_B - 2} s_{Bx_p}^2 \tag{22}$$

$s_{Ax_p}^2$ and $s_{Bx_p}^2$ are the sample variances for the un-standardized expression levels for Groups A and B, respectively, corresponding to the p^{th} gene. Note that $T_{scale_{JP}}^{EA}$ is independent PCA loadings and thus, does not benefit from PCA. In actuality, Eq. 21 is the commonly known Student's *t*-statistics [14]; thus,

$$T_{pooled,p} = T_{scale_{JP}}^{EA} \tag{23}$$

From Eq. 23 it is clear that scaling the EA differential contribution is not providing any new technique in PCA and therefore is not a useful result under the PM. Thus, we do not propose scaling for the EA approach.

The steps for applying the PM are as follows:

1. Standardize **X** to obtain **Z**.
2. Obtain the loading and scores matrices for **X** (EG) based on correlation.
3. Obtain the loading and scores matrices for $\mathbf{X}^T$ (EA) based on covariance.
4. For each of the *n* EG loading vectors, plot its loadings against the assay number. Select the plot(s) that separate the assays into desired or interesting groups for further analysis.
5. For each *n* EA score vectors, plot its scores against the assay number. Select the plot(s) that separate the assays into desired or interesting groups for further analysis.
6. For each selected EG loading vector in Step 4, using **Z** and Eq. 5 determine the differential EG contribution for each gene.
7. For each selected EA loading vector in Step 5, using **X** and Eq. 9 determine the differential EA contribution for each gene.
8. For each case in Steps 6 and 7, rank order the differential contribution and then table (with the corresponding gene)and plot these values against the rank. These signature plots can be used to determine where to make cutoffs as described in Rollins et al. (2006).

In the next section we evaluate the proposed test statistics that we have derived in this section against a current method that uses the Student's *t*-test statistic. This work also includes an evaluation to determine when it is better to choose T_{diff} or T_{scale}.

IV. RESULTS AND DISCUSSIONS

The best choice for a test statistics is the one that has the highest statistical power (SP) and the lowest false discovery rate (FDR) [21]. This section presents three case

studies to evaluate the proposed test statistics against one another and against a current method (CM) that uses T_{pooled}. The first study revisits the single group analysis in Rollins et al. (2006) involving exposure of *E. coli* cells to two different levels of ethanol concentration [2]. The second study applies the proposed method (PM) to data from Steelman et al. (2006) [9]. This data set involves the use of myostatin as an inhibitor of skeletal muscle growth for five 5-weeks-old myostatin (called "mutant") and non-treated (called "wild-type") mice in each group. The third study is a mathematically simulated data study using characteristics of the data from Study 2.

Exposure of *E. coli* cells Study

The data set for the first case study contains *E. coli* cells that were exposed to two different ethanol concentrations. In Rollins et al. (2006) ranked signatures were obtained for non-ethanol (i.e., non-treated) (Group A) and ethanol (Group B) separately. Thus, these signatures ranked the genes based on their contribution to the score of their group. However, the goal of this work is to obtain a ranked signature of the genes that is based on the ***difference of gene contribution*** between the two groups. Therefore, under this objective, genes with high contribution in both groups would not be ranked high; whereas, genes with low contribution in one group and high contribution in the other group could be ranked high based on the greatest negative, positive, or absolute difference, depending on the interests of the experimenter. For this study, we ranked the genes based on absolute difference for evaluative purposes.

The results of this study using the PM are given in Table 1 and Fig. 3. These results were obtained from the first PC for an EA analysis using $T_{diff_{i1}}^{EA}$ only to determine

differential gene contribution. This PC was selected, as supported by Fig. 3, because it separated the two groups in the score plot quite well. The plot on the right in Fig. 3 gives the differential contribution calculated from $T_{diff_{i1}}^{EA}$ by rank with the rank decreasing with increasing value on the horizontal axis. As this figure shows, the top genes clearly standout by their distinct separation and how they line up almost vertically along the vertical axis. Table 1 gives the top 20 genes. This list contains some of the top genes in the ethanol and non-ethanol signatures in Rollins et al. (2006) as indicated. In addition, it contains genes that were not ranked very high in either signature. However, note that each gene is at opposite ends of the signatures in Rollins et al. (2006) in support of their differential significance. Thus, the PM has potentially found genes that might express relatively low within assays of similar conditions but quite differently between assays of different conditions. Follow up experiments would be necessary to verify these findings which is beyond the scope of this work.

Skeletal Muscle Growth in Mice Study

The second study is a data set that involved the use of myostatin as inhibitor of skeletal muscle growth for five 5-week-old myostatin (called "mutant") and non-treated (called "wild-type") mice in each group. A powerful method for ranking genes and determining the size of signatures is the Q-method developed by Storey and Tibshirani (2003). The Q-method uses T_{pooled} and a novel method for achieving high SP and low FDR. The Q-method first uses T_{pooled} to obtain p-values then convert to q-values in order to determine where to cut-off signatures based on a maximum q-value. Given that the q-value is related to the p-value, one could also rank genes based on p-values or their T_{pooled} values which are inversely related. Since we are primarily interested in ranking genes in

this work we will compare the techniques based on the abilities of T_{pooled} and the PM to find top ranked genes.

Table 1: Top 20 genes that showed distinct difference between ethanol and non-ethanol along with their ranking

The "*" gives the rank in the ethanol and non-ethanol signature in Rollins et al. (2006).

Rank	Gene Name	EtOH Rank*	Non-EtOH Rank*	Rank	Gene Name	EtOH Rank*	Non-EtOH Rank*
1	*b2387*	729	2001	11	*argT*	925	2330
2	*ybdO*	558	2182	12	*argH*	2	4286
3	*b1455*	959	2120	13	*ycbE*	317	3626
4	*gltD*	2884	151	14	*b0538*	328	2658
5	*appY*	360	2457	15	*citB*	372	2642
6	*caiA*	5	3810	16	*wbbH*	2952	408
7	*b0960*	2664	787	17	*ccmD*	2605	1083
8	*yaiD*	1	4284	18	*agaA*	885	2483
9	*b1815*	3178	43	19	*ymcC*	568	2587
10	*ydaK*	375	2548	20	*abc*	389	2705

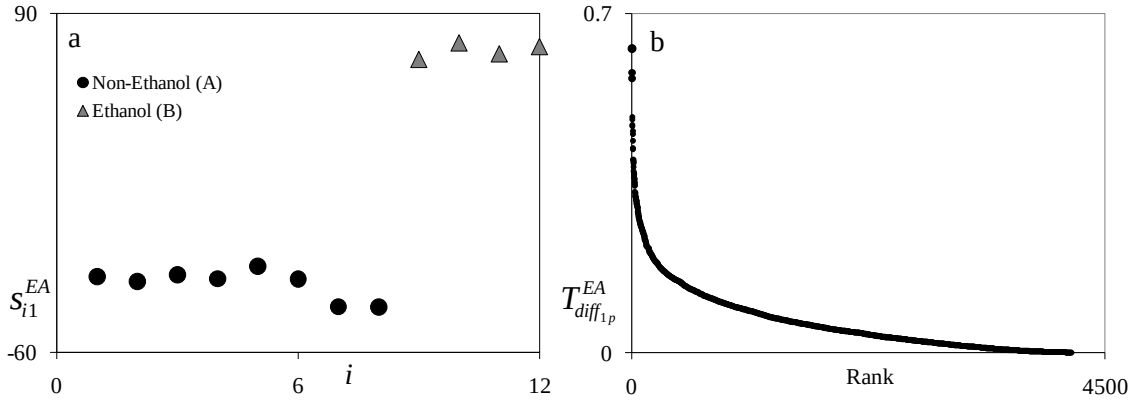


Figure 3. EA1 Score plot (a) and gene signature plot (b) for the *E. coli* in ethanol and non-ethanol study

The score plot shows excellent grouping for the non-ethanol (A) and ethanol (B) assays (B). For the signature plot on the right, the rank decreases as the number increases. The top ranking genes seen on the upper left in plot (b) show excellent ranking as evidenced by the vertical line and clear separation of the most highly ranked genes.

PCA results for PM are given in Fig. 4. These results were obtained from the first PC for an EA analysis using $T_{diff_{i1}}^{EA}$ only to determine differential gene contribution. This PC was selected, as supported by Fig. 4, because it separated the two groups in the score plot quite well. As shown by the $T_{diff_{i1}}^{EA}$ plot on the right, the top genes clearly standout by their distinct separation and by how they line up along the vertical axis. The top genes that the PM identified were genes identified in Steelman et al. (2006). In addition, it also identified genes that were not previously identified in their work.

A comparison of the PM and the CM is given in Table 2. In this table, the top 200 genes of the CM are selected as the base set. The number and percentage of the top 10, 20, . . .,100 genes of the PM in this set are given. This analysis is represented by the first three columns in the table. In addition, this table gives results that switch the roles of the PM and CM. More specifically, the top 200 genes of the PM are selected as the base set and the number and percentage of the top 10, 20, . . ., 100 genes of the CM in this set are determined. This analysis is represented by the last three columns in Table 2. With the CM as the base set, the results range from 70% of the top 10 genes to 22% of the top 100 genes of the PM being in set of the top 200 genes of the CM. Similarly, with the PM as the base set, the results range from 50% of the top 10 genes to 22% of the top 100 genes of the CM being in the set of the top 200 genes of the PM. Thus, while there is agreement between the two approaches, the lack of agreement warrants further investigation on the

best choice of method based on the criteria of highest SP and lowest FDR. Our last study is a Monte Carlo simulation data study to compare these two approaches under these criteria.

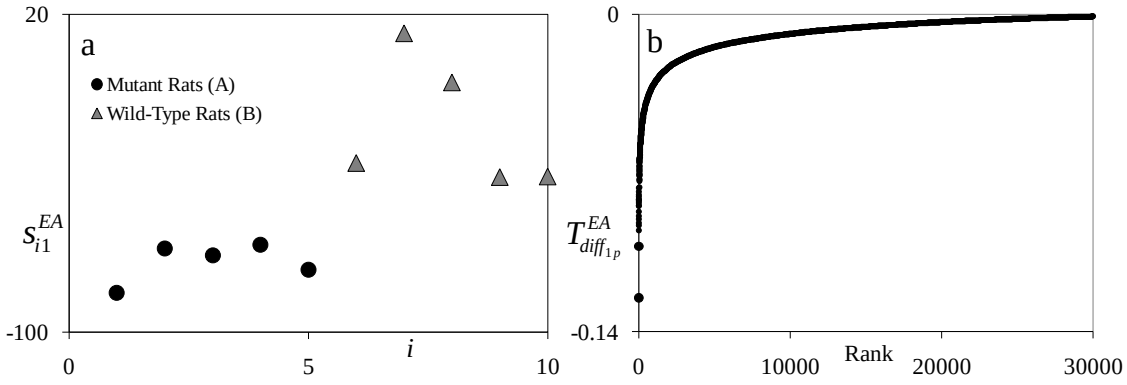


Figure 4. EA1 Score plot (a) and gene signature plot (b) for the skeletal muscle growth in mice study

The score plot shows excellent grouping for the mutant mice (A) and the wild-type mice (B) assays. For the signature plot on the right, the rank decreases as the number increases. The top ranking genes seen on the left in this plot show excellent ranking as evidenced by the vertical line and clear separation of the most highly ranked genes.

Table 2: Top ranked genes of one method in the top 200 genes of the other method in skeletal muscle growth in mice study

This table shows how many of the top genes for one method are in the top 200 genes of the other method with x = # of top PM genes and y = # of top CM genes. For example, the result for x = 30 means that 10 (33%) of the top 30 genes of the PM where in the top 200 genes of the CM.

x	x in top 200 CM genes	%x in Top 200 CM genes	Y	y in top 200 PM genes	%y in Top 200 PM genes
10	7	70	10	5	50
20	9	45	20	7	35
30	10	33	30	7	23

40	10	25	40	11	28
50	11	22	50	15	30
60	17	28	60	17	28
70	19	27	70	17	24
80	20	25	80	19	24
90	20	22	90	20	22
100	21	21	100	21	21

Simulation Study

As stated above, the purpose of the simulated data study is to evaluate and compare the PM and CM to identify genes with significant differential effects. We simulated several data sets based on the statistical properties of the data matrix from the second study. Each data matrix contained 40,000 genes with 10 assays of five samples in each group. The distribution for the simulated data can be described as follows:

$$x_{ij} \overset{indep}{\sim} N\left(\mu_{x_j}, \sigma^2_{x_i}\right), \quad \forall i, j \tag{24}$$

such that

$$\mu_{x_j} = \begin{cases} 5.3 + \delta, & \delta > 0; i = 1, ..., 200; \quad j = 1, ..., 5 \\ 5.3, & otherwise \end{cases} \tag{25}$$

Thus, 200 of the genes for each of the assays in Group A had the largest mean and were significantly different than all the other genes that had a mean of 5.3. The study will evaluate the ability of the CM and PM to identify these 200 genes when the variance for all the data in the data matrix is the same (Part 1) and when the variance differ from gene

to gene (Part 2). Each result in the simulation study is an average of five trials. All the results for PM will be based on eigengene (EG) principal components (PCs).

Simulation Study -- Part 1

In the first simulation study we evaluated the techniques under different levels of $\sigma^2_{x_i}$ with $\delta = 1$. There were seven levels of these values that ranged from 0.04 to 1.0. Thus, the range of the ratio of σ_{x_i} to the differential effect (δ) was also 0.2 to 1.0. The PCA results for one trial of the PM at the lowest level of σ_{x_i} are given in Fig. 5. As shown, the loading plot indicates excellent separation of Group A and Group B indicating that PCA was able to pick up a difference of $\delta = 1$ quite well for 200 of the 40,000 genes using the $T^{EG}_{diff_{i1}}$ test statistic. The signature plot reveals a distinct signature for these genes as evidenced by the large gap. For this case the percents of the 200 significantly different genes (SDG) ranked in the top 200 by $T^{EG}_{diff_{i1}}$, $T^{EG}_{scaled_{i1}}$ and T_{pooled_i}, were 100.0%, 99.9% and 90.9%, respectively. These percentages for all the cases for this part of the simulation study for these three test statistics are given in Fig. 6. In addition, this figure gives results for percentages of the SDG in the top 300 and top 400 for these test statistics. As shown, T^{EG}_{diff} has the best performance, followed closely by T^{EG}_{scaled} at the extremes and poorly by CM statistic T_{pooled}. Thus, when the variability of the assays is similar, T^{EG}_{diff} appears to be the best choice for identifying the most significant genes.

Simulation Study -- Part 2

In the second simulation study we evaluated the techniques by varying levels of $\sigma_{x_i}^2$ for each gene and two levels of δ: 1 and 3. More specifically, the distribution for σ_{x_i} was log normal with mean 0.37 and variance 0.37^2. Thus, for each data table a σ_{x_i} was randomly generated for each gene *i*, *i* = 1, ..., *m*, and then ten simulated expression values, one for each assay, were generated according to Eqs. 24 and 25 for the given level of δ.

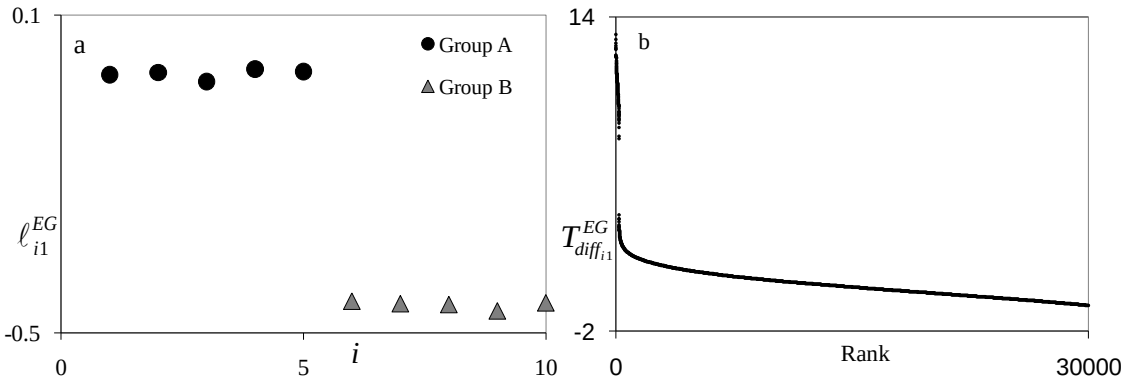


Figure 5: Loading1 plot when δ = 1 and σ = 0.20 (a) and its sharp gene signature (b)

Figure a shows a nice clean separation of Group A from Group B. Figure b shows the nice assay-specific gene signature plotted against their rank.

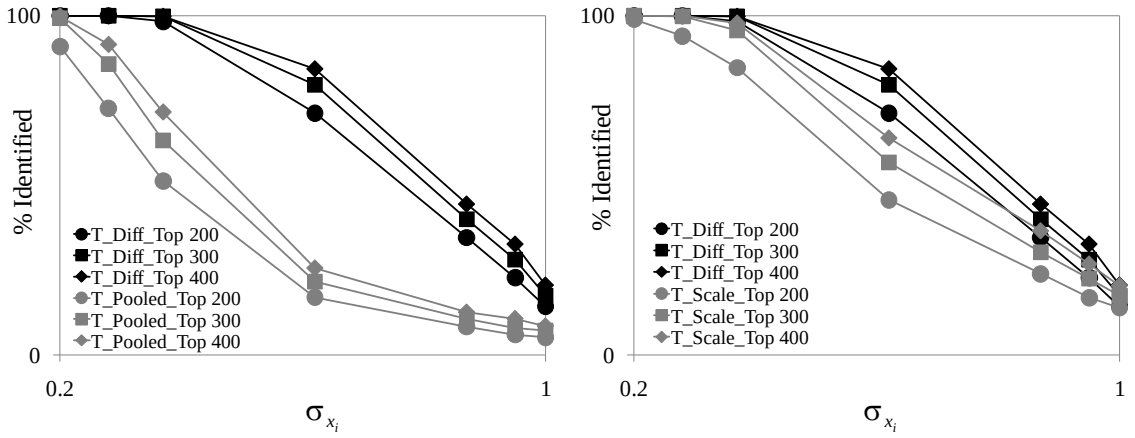


Figure 6. The % of the 200 significantly different genes (SDG) in the top lists for the three statistics in Simulation Study Part 1

These results for T_{diff}^{EG} and T_{pooled} are on the left and the ones for T_{diff}^{EG} and T_{scaled}^{EG} are on the right.

Identification results for this part of the study are given in Fig. 7 as the percent of the SDG that are in the top 200 and top 400 ranks determined by the three test statistics. The best performing method this time is T_{scaled}^{EG}, followed by T_{pooled}, and then by T_{diff}^{EG}. At δ = 3, all three methods are close but spread out at δ = 1. While the spread at δ = 1 for T_{scaled}^{EG} and T_{pooled} is significant, the spread for T_{pooled} and T_{diff}^{EG} is quite large. Thus, T_{diff}^{EG} does not appear to be the best choice when δ is small and there is significant variation between the genes across the assays. Since T_{scaled}^{EG} consistently did the best, when the gene variation is significant across the assays, it is our recommendation.

Our final analysis in this study evaluated performance in signature determination. The CM is the Q-method developed by Storey and Tibshirani (2003) that uses the p-values of the pooled *t*-test (i.e., T_{pooled}) and cuts the list off at a maximum Q-value, commonly 0.05, the value used in this analysis. The PM is the Inflection Method (IM) that is described in Rollins et al. (2006) that cuts the list off at the greatest change in the signature plot of the ranked genes. The results are from Part 1 of the simulation study with a constant σ_{x_i} for all the genes in a data table.

The results of this analysis are given in Fig. 8. The plot gives the signature size (SS) (i.e., the number of genes in the signature) and the SDG against σ_{x_i}. Statistical Power is seen by the height of the SDG curve. As typical, SP, as indicated by this line, decreases as σ_{x_i} increases. Hence, the PM signature performance is seen to be significantly better than the CM in terms of SP. An indication of the false discovery rate (FDR) of the methods can be compared by the separation of their two lines in Fig. 8.

These lines for the PM are very close except at the highest levels of σ_{x_i}. This indicates that the number of insignificant genes in the signatures of the PM is quite small and hence, has a small FDR. The FDR of the CM appears to be much higher for low values σ_{x_i} and the SS drops to zero relatively quite fast so that performance at low σ_{x_i} is not too meaningful since there are very few genes in the signature. Thus, the IM with the test statistics of the PM for determining signature cutoffs appears to have merit as a viable approach.

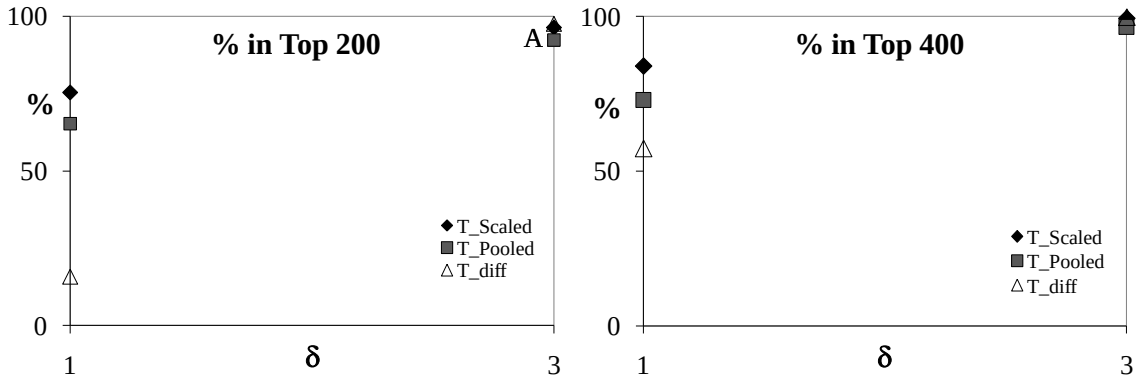


Figure 7. The % of SDG in the top 200 (left) and in the top 400 (right)

The plot shows percent of the 200 significantly different genes (SDG) that are in the top 200 and 400 lists at two different values of δ for the three test statistics in Part 2 of the simulation study.

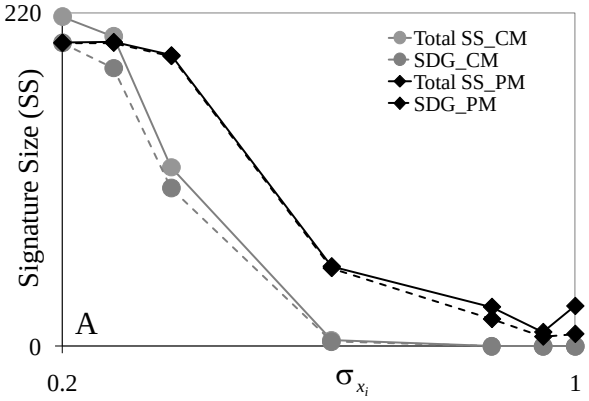


Figure 8. Signature Size (SS) performance for the CM and PM.

The plot gives the signature sizes (i.e., number of genes the method determines as being significant) and the number of the 200 significantly different genes (SDG) in the signatures for the CM and PM. Higher statistical power (SP) is observed by the higher height of the SDG plot and the higher false discovery rate (FDR) is observed by the greater separation of the lines of the same color.

V. CONCLUSION

This work proposed a new principal component analysis (PCA) method for analyzing large dimensional data set such as gene expressions data set. The strength of the proposed method (PM) comes from its data driven nature. PCA is first used to verify the existence of the assay grouping of interests. From the principal components (PCs) that provide this verification, the contribution of each gene providing the greatest differential expressions based on linear combination from the PCs are ranked. Thus, a PM signature is not just a difference of expression levels for genes but differences in a direction verified to have the characteristics of interests. This approach distinguishes PM for methods that do not form groups on the basis of data analysis and develop signatures from the differences between two groups in the original data space.

Following Rollins et al. (2006), the PM develops test statistics treating the assays as variables (eigengene, EG) and the genes as variables (eigenassay, EA). These test statistics are linear combinations of these variables (i.e., pseudo variables) as determined from the elements of the eigenvectors. One test statistic, called T_{diff} is the difference of the average expression levels between two groups of pseudo variables. The other test statistics, called T_{scaled}, is T_{diff} divided by the estimated pooled standard deviation. We compared the performance of these two test statistics with the common and popular

Student's *t*-statistic, T_{pooled} that we called the current method (CM). Two real data studies provided evidence in support of the PM as a viable technique. A simulation study provided the strongest supportive evidence for the use of T_{diff} when the gene variability is fairly uniform throughout a data table and for T_{scaled} when the variability is not fairly uniform. Finally, with the PM test statistics, the inflection method (IM) introduced by Rollins et al. (2006), indicated strong promise in determining signature cutoffs in terms of statistical power (SP) and false discovery rate (FDR) as compared to CM.

We are applying the PM in a variety of applications involving biological as well as physical phenomenon, with promising results. These applications include: 1. Nitric Oxide- and S-nitrosoglutathione- responsive genes in *E-coli*; 2. analysis of DNA microarray data for juvenile small round blue cell tumors; 3. analysis of metabolite data from corn tissues (silk, pollen, coleoptile, and seedlings) for differential expression levels between the wild type and genetic mutations; 4. analysis of spectroscopy data for super alloys; and 5. the enhancement of nondestructive tests for ceramic armor in the resistance of ballistic penetration. Thus, the PM has potential application in a variety of situations where differential analysis is needed on large data tables with a relatively small number of different conditions or assays. It appears to have promise for these applications for high SP and low FDR as compared to other currently available methods.

VI. ABBREVIATION

FG, functional genomics; PCA, Principal Component Analysis; PC, Principal Component; EG, Eigengene; EA, Eigenassay; *l*, loading; *S*, score; PM, proposed method; CM, current method; T_{diff}, difference of the average expression levels between two groups of pseudo

variables; T_{scaled}, scaled statistics by dividing T_{diff} by its estimated pooled standard deviation; T_{pooled}, Student's *t*-statistics; g_i, gene contribution for i^{th} gene; δ, differential effect; SS, signature size; SDG, statistically different genes; SP, statistical power; FDR, false discovery rate; IM, inflection method

VII. REFERENCES

1. Dharmadi Y, & Gonzalez R: **DNA Microarrays: Experimental Issues, Data Analysis, and Application to Bacterial Systems,** *Biotechnol. Progress* 2004, **5**: 1309-1324.

2. Rollins D K, Zhai D, Joe AL, Guidarelli JW, Murarka A, Gonzalez R: **A novel data mining method to identify assay-specific signatures in functional genomic studies**, *BMC Bioinformatics* 2006, **7**:377.

3. Zhang W, Carriquiry A, Nettleton D, Dekkers JCM: **Pooling mRNA in Microarray Experiments and Its Effect on Power,** *Gene Expression* **2007, 23:**1217-1224.

4. Shendure J, Ji H: **Next-generation DNA sequencing**, *Nature Biotechnology* **2008**, 26: 10.

5. Morozova O, Marra MA: **Applications of next-generation sequencing in functional genomics**, *Genomics* **2008, 92:** 255-264.

6. Raychaudhuri S, Stuart JM, Altman RB: **Principal Component Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series**, *Pacific Symposium on Biocomputing* **2000, 5**: 452-463.

7. Misra JW, Hwang D, Hsiao LL, Gullans S, Stephhanopoulos G: **Interactive exploration of microarray gene expression patterns in a reduced dimensional space**, *Genome Res* **2002, 12**: 1112-1120.

8. Sharov AA, Dudekula DB, and Ko MSH: **A web-based tool for principal component and significance analysis of microarray data,** *Bioinformatics* **2005, 21**: 2548-2549.

9. Nettleton D: **A discussion of Statistical Methods for Design and Analysis of Microarray Experiments for Plant Scientists,** *The Plant Cell* **2006, 18:** 2112-2121.

10. Steelman CA, Recknor JC, Nettleton D, and Reecy JM: **Transcriptional profiling of myostatin-knockout mice implicates Wnt signaling in postnatal skeletal muscle growth and hypertrophy**, *The FASEB Journal* **2006, 20**: 580-582.

11. Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments**, *Stat. Sinica* **2002, 12**: 111–139

12. Storey JD, and Tibshirani R: **Statistical significance for genomewide studies**, *Proc. Natl. Acad. Sci. USA* **2003, 100**: 9440-9445.

13. Ge Y, Sealfon SC, and Speed TP: **Statistical Methods in Medical Research**, *Statistical Method in MedicalResearch* **2009, 18:** 543.

14. Devore JL.: **Probability and Statistics for Engineering and the Sciences**. 7th edition. Duxbury Press; 2007.

15. Sartor MA, Tomlinson CR, Wesselkamper SC, Sivaganesan S, Leikauf GD, Medvedovic M: **Intensity-based Hierarchical Bayes Method Improves Testing for Differentially Expressed Genes in Microarray Experiments**, *BMC Bioinformatics* **2006, 7:** 538.

16. Kendziorski CM, Newton MA, Lan H, Gould MN: **On Parametric Empirical Bayes Methods for Comparing Multiple Groups Using Replicated Gene Expression Profiles**, *Statistics in Medicine* **2003, 22**: 3899-3914.

17. Efron B, Tibshirani R, Storey JD, Tusher V: **Empirical Bayes Analysis of a Microarray Experiment**, *Journal of American Statistical Association* **2001, 96:** 456.

18. Liao JC, Boscolo R, Yang YL, Tran, LM, Sabatti C, Rpoychowdhury VP: **Network component analysis: Reconstruction of regulatory signals in biological systems**, *The National Academy of Sciences* **2003, 100:** 15522-15527.

19. Tran LM, Brynildsen MP, Kao KC, Suen JK, Liao JC: **gNCA: A framework for determining transcription factor activity baed on transcriptome: identifiability and numerical implementation**, *Metabolic Engineering* **2005, 7:** 128-141.

20. Hyduke DR, Jarboe LR, Tran LM, Chou KJY, and Liao JC: **Integrated network analysis identifies nitric oxide response networks and dihydroxyacid dehydratase as a crucial target in *Escherichia coli***, *Proceedings of the National Academy of Sciences of the United States of America* **2007**, **104**: 8488-8489.

21. Benjamini Y, and Hochberg Y: **Controlling the false discovery rate: apractical and powerful approach to multiple testing**, *Journal of Royal Statistical Society Series B* **1995, 57:** 289-300.

22. Gonzalez R, Tao H, Purvis JE, Shanmugam KT, York SW, and Ingram LO: **Gene Array-Based Identification of Changes That Contribute to Ethanol Tolerance in Ethanologenic *Escherichia coli*: Comparison of KO11 (Parent) to LY01**

(Resistant Mutant), *Biotechnol. Prog* **2003, 19**: 612- 623.

23. Johnson RA and Wichern DW: **Applied Multivariate Statistical Analysis**. 6th edition. Pearson Prentice Hall; 2008: 430 – 470.

CHAPTER 4: DETERMINATION OF NITRIC OXIDE- AND S-NITROSOGLUTHATHIONE- RESPONSIVE GENES IN *ESCHERICHIA COLI* USING DIFFERENTIAL PRINCIPAL COMPONENT ANALYSIS

A paper to be submitted to *BMC Bioinformatics*

AiLing Teh, Laura R. Jarboe, and Derrick K. Rollins, Sr.

ABSTRACT

Background: *Escherichia coli*(*E. coli*) resides inside the human body and during the course of infection is exposed to reactive nitrogen species (RNS) such as nitric oxide (NO) and S-nitrosoglutathione (GSNO). NO plays an important role in the mammalian immune system and regulation of blood flow. GSNO sometimes serves as a donor of NO, but its thiol chemistry also has effects distinct from NO. This work applies the recently developed Principal Component Analysis (PCA) technique of Rollins et al. (in review) in determining NO and GSNO response genes in an *E. coli* network. PCA is a powerful data mining method that is able to analyze and extract critical information from large microarray data sets even though the number of genes is much greater than the number of experimental runs.

Result: The application of Rollins et al. technique identified a signature of NO-specific and GSNO- specific genes, containing those that were previously identified by Hyduke et. al. (2007) and Jarboe et. al. (2008) as well as new genes which may affect NO and GSNO activity. Three genes not previously identified as affecting NO and GSNO response were selected for experimental analysis to evaluate the identification ability of

the method. With control groups showing no significance, these genes were found to be significant at the 0.002 test level.

Conclusions: Although the three chosen genes investigated in this work were not ranked very high in the list identified from the PCA method, the results support the ability of the method to successfully identified genes that affect NO and GSNO activity. Thus, the method appears to be robust and we recommend its use in new gene discovery applications using gene expression data.

I. INTRODUCTION

E. coli is one of the bacterial species that resides in the human digestive tract, but it is also the causative agent of diseases such as cystitis and pyelonephritis. Studies showed that during the course of infection, *E. coli* was affected by reactive nitrogen species (RNS) such as NO and GSNO [1- 3]. Moreover, NO is found to play an important role of fighting the infection inside the human body. Network Component Analysis (NCA), a method based on known regulatory networks, was previously used to identify the molecular targets and response network to NO and GSNO [1-2]. The purpose of this work is to evaluate the ability of a new Principal Component Analysis (PCA) technique developed by Rollins et al (2010), to identify new genes that show effect on a chemical activity. More specifically, the identification ability of this method was evaluated in this work using NO and GSNO response networks.

PCA has been widely used by practitioners to analyze microarray data set [5-8]. Two common strengths are its ability to reduce dimensionality and to create data driven methods by exploiting its ability to find natural relationships among correlated variables. The latter strength is the one we exploit in this work. Rollins et al. (2006) developed a

technique that uses ***gene contribution*** from biologically related groups identified by PCA to obtain ranked gene lists. Rollins et al. (2010) extended this technique to obtain ranked gene lists that use ***differential*** gene contribution between different biologically related groups identified by PCA. Rollins et al. (2010) evaluated this technique by comparing it to a previously identified list in a real data study and a known list in a simulated data study. In contrast, the work in this article evaluates the ability of this technique to identify new genes that affect a specific type of biological response and then to evaluate the accuracy of the identification results by experimentation. This study uses the responses to NO- and GSNO in *E. coli* as the biological system. It applies the PCA technique to identify ranked order list of genes and then select three genes from the lists for experimental validation of their responsive ability.

II. METHODS

The data matrix for this work has $m = 4166$ rows representing 4166 genes and $n = 4$ columns or assays (i.e., experimental conditions). The cells in this matrix are given as x_{ij} which is the expression level of the ith gene for the jth assay. This microarray data matrix is from [1-2] and it gives the 5-minute response of *E. coli* strain BW25113 to four chemical treatments: the RNOS nitric oxide (NO, $j = 1$), and S-nitrosoglutathione (GSNO, $j = 2$), potassium cyanide (KCN, $j = 3$) and serine hydroxamate (SeOH, $j = 4$). Potassium cyanide is known to cause respiration inhibition, one of the known effects of NO [4]. Serine hydroxamate induces amino acid starvation [10], which was concluded to be an effect of both NO and GSNO treatment (Hyduke, Jarboe). KCN and SeOH are included to eliminate some of the indirect effects from NO and GSNO.

The PCA technique used in this work can be understood and appreciated by considering the following case. Suppose one is interested in finding a ranked order list of genes that express most differently for NO response versus the other three responses. One way of expressing this difference mathematically is:

$$g_i = x_{i1} - \left(\frac{x_{i2} + x_{i3} + x_{i4}}{3} \right) \qquad i = 1, \ldots, m \tag{1}$$

where g_i gives the differential expression of gene *i* according to Eq. 1. However, one problem with Eq. 1 is that it was written without evidence of the relationship it expresses. Thus, without evidence one cannot truly have confidence in its use to rank genes. Another problem is that it was not developed based on supporting empirical or much theoretical support and is therefore, likely far from optimal for ranking the genes.

The PCA method of Rollins et al. (2010) does not suffer from these limitations. It treats either the assays or genes as variables and obtains Eigenvalues (called "*Eigengenes*," EG treating assay as variables or "*Eigenassays*," EA treating genes as the variables) that are used to validate the existence of the biological relationships of interests. It then uses this information to establish the mathematical expression to calculate the contribution of each gene to this relationship. More specifically, for this case, as described in Rollins et al. (2010), an EG analysis would plot the four (4) loadings for each principal component against the assay numbers. Similarly, an EA analysis would plot the product of the vector of the loadings for each principal component (PC) by each of the four genes vectors (producing four *scores*) against the assay numbers. The plots that are acceptable for this case would separate the loading for Assay 1 in the EG analysis from the loading for the other assays or the score for Assay 1 in the EA from the scores for the other assays. The EG or EA corresponding to plots that give the most distinct

separation would be used to determine the corresponding gene contribution equations as follow for EG and EA, respectively:

$$g_{EG_i} = \ell_1 z_{i1} - \left(\frac{\ell_2 z_{i2} + \ell_3 z_{i3} + \ell_4 z_{i4}}{3} \right) \qquad i = 1, \ldots, m \tag{2}$$

$$g_{EAi} = \ell_i \left(x_{i1} - \left(\frac{x_{i2} + x_{i3} + x_{i4}}{3} \right) \right) = \ell_i g_i \qquad i = 1, \ldots, m \tag{3}$$

where in Eq. 2, ℓ_j is the *j*th loading or weight for this EG vector, $j = 1, \ldots, 4$; and

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad i = 1, \ldots, m; \quad j = 1, \ldots, 4 \tag{4}$$

While, ℓ_i in Eq. 3 is the *i*th loading for this EA vector. Note that in Eq. 2, each z_{ij} for gene *i* has a different loading that does not change from gene to gene and that in Eq. 3, each x_{ij} for gene *i* has the same loading that changes from gene to gene. Thus, by using loadings associated with PCs that distinguishes NO response from the other three, Eqs. 2 and 3 will be more affective in ranking genes for NO responsiveness. This same process can be applied for other responses of interests as long as the PCs can be found that reflect the behavior of interest. Next the procedures are presented to obtain the ordered lists from application of Eqs. 2 and 3.

Eigengene Procedure

As stated above, the PCA technique has two ways of creating ranked order list of genes. In the context of the application of this work, we first describe the Eigengene (EG) procedure and then the Eigenassay(EA) procedure. For the EG procedure, the assays are

treated as the variables. The first step is to standardize the data matrix as described in Rollins et al. (2006). Next, using an appropriate software package the four loading and four score vectors are obtained from the standardized data matrix. The four elements of each loading vector are plotted against the four assay number creating four loading scatter plots. The loading plot showing sufficient and the greatest separation of Assay 1 from the other assays is selected for the NO response. Similarly, the plot showing sufficient and the greatest separation of Assay 2 from the other assays is selected for the GSNO response. Using the selected loading vectors, the contributions for each assay group are obtained using equations similar to Eq. 2 with the appropriate modifications based on the scattered plot for both NO and GSNO. Note that, as described in Rollins et al. (2006) the standardized values, defined as z_{ij}, are used in the contribution equations. The values of g_{EGi} are then used to rank their associated genes. A line plot (called a *signature plot* in Rollins et al., 2006) of g_{EGi} provides a visual representation of the sharpness of the ranking.

Eigenassay Procedure

For the EA procedure the genes are treated as the variables. This is done by taking the transpose of data matrix which is an n x m matrix. Since the genes are spread along the columns it might not be possible to display this matrix in a spread sheet. In addition, because of the large number of variables, m, not all software packages will be able to do PCA on this matrix. We have found *R* [11-12], a free open statistical software package, to be quite effective for obtaining the PCs for the EA procedure. This software has several different ready packages that are written by professional which will help to obtain the loading and score vectors once the data is read into the workspace. The four elements of

each score vector are plotted against the four assay number creating four loading scatter plots. The plot showing sufficient and the greatest separation of Assay 1 from the other assays is selected for the NO response. Similarly, the plot showing sufficient and the greatest separation of Assay 2 from the other assays is selected for the GSNO response. Using the selected loading vectors, the contributions are obtained using equations similar to Eq. 3 with the appropriate modifications based on the scattered plot for both NO and GSNO. The values of g_{EAi} are then used to rank their associated genes. A line (signature) plot of g_{EAi} provides a visual representation of the sharpness of the ranking.

III. RESULTS AND DISCUSSION

As discussed on the previous section, we generated four loading plots using the EG procedure and four score plots using the EA procedure. The loading plots for Loading Vectors 2 (L2) and 4 (L4) are shown in Fig. 1. As shown, the plot for L2 indicates separation of GSNO and the one for L4 indicates separation for NO. Thus, L2 and L4 were selected to create the ranked list of genes for GSNO and NO responses, respectively. The EG contribution equations for GSNO and NO are given below by Eqs. 5 and 6, respectively.

$$g_{EG-GSNOi} = \ell_{22} z_{i2} - \left(\frac{\ell_{12} z_{i1} + \ell_{32} z_{i3} + \ell_{42} z_{i4}}{3} \right) \qquad i = 1, \ldots, m \tag{5}$$

$$g_{EG-NOi} = \ell_{14} z_{i1} - \left(\frac{\ell_{24} z_{i2} + \ell_{34} z_{i3} + \ell_{44} z_{i4}}{3} \right) \qquad i = 1, \ldots, m \tag{6}$$

where ℓ_{ij} is the *i*th loading for the *j*th EG vector. Similarly, under the EA procedure, based on the score plots S1 and S3 in Fig. 2, we selected EA1 and EA3, to create the ranked list of genes for GSNO and NO, respectively where s_{ij} is the *i*th score for the *j*th

EA score vector. The EA contribution equations for GSNO and NO are given below by Eqs. 7 and 8, respectively.

$$g_{EA-GSNO_i} = \ell_{i1}\left(x_{i2} - \left(\frac{x_{i1} + x_{i3} + x_{i4}}{3} \right) \right) \qquad i = 1, \ldots, m \tag{7}$$

$$g_{EA-NO_i} = \ell_{i1}\left(x_{i1} - \left(\frac{x_{i2} + x_{i3} + x_{i4}}{3} \right) \right) \qquad i = 1, \ldots, m \tag{8}$$

where ℓ_{ij} is the *i*th loading for the *j*th EA vector.

Figure 3 provides the EG signature plots for Eqs. 5 and 6. Similarly, Fig. 4 provides EA signature plots for Eqs. 7 and 8. From the rank ordered differential gene contributions shown in the signature plots, differences plots below, the distinct gaps for the top genes are clearly revealed.

The top 25 genes that we identified to be important to direct NO target using EG and EA are showed in Tables 1 and 2 respectively. On the other hand, Tables 3 and 4 show genes that identified to be important to direct GSNO target using EG method and EA method respectively. From these tables, the agreement is seen to be quite as most of the top genes identified from EG method matched with those that identified from EA method. Some of the top genes that we have identified have also been identified in other RNOS studies, as reviewed by (Jarboe et. al. NO: Biology and Pathobiology). In addition, our list includes a number of genes that, to our knowledge, have not been previously identified as important to the NO or GSNO response. Therefore, we selected three of these genes and investigate their ability to affect NO tolerance.

The eigengene and eigenassay methods both identified known NO-responsive genes *hmp*, *norV*, *norW* and *ytfE* as affecting NO response. *Hmp* and *NorV* both have

demonstrated NO-consuming activity and each of these four genes play an important role in NO defense [9]. Additionally, the NO-important genes include Fe-S cluster repair genes *iscRSUA* and many amino acid biosynthesis genes. This is consistent with the previous report that NO damages the Fe-S center of many proteins, including branched-chain amino acid biosynthesis protein IlvD, causing amino acid starvation (Hyduke et. al.). For the GSNO response, the eigengene and eigenassay methods both identified many genes related to methionine and cysteine biosynthesis. This is consistent with reports that GSNO depletes the levels of cysteine and the methionine precursor homocysteine (Jarboe et. al.). Together, these findings validate the methodology presented here. However, this method also identified many genes that are outside the previously-reported NO and GSNO response models. This includes several genes with minimal available functional or annotation data, such as *yahN, ycbR, yciE, yeaR*.

We have selected three genes that were not previously reported for NO responsive or GSNO responsive and worthwhile to experiment to verify the results. These three genes are *ycbR, yfhA, and yahN*. Note that, as seen from Tables 1 through 4, these genes are not ranked very high. The highest rank is for *yahN* -- 16 in Table 2 for NO. The rank for *yfhA* is 28 and is only ranked for the NO response in Table 2. While *ycbR* rank appears in three of the four lists, these ranks are 20, 34 and 45. Thus, these low ranks will strengthen the confidence in the method if these genes are found to have an effect.

It was proposed that genes that are important to NO defense might have a positive effect on NO tolerance. Wild-type *E. coli* shows a growth lag in response to treatment with 10 μM DEA/NO and this sensitivity is pronounced in the absence of the *hmp* gene (Hyduke et. al.). Therefore, we increased the expression of the candidate genes, *ycbR*,

yfhA and *yahN* in BW25113 Δ*hmp* and tested the growth response of the resulting strain to NO challenge. Figure 5 shows the growth response for the control (BW25113 Δ*hmp* + blank TOPO vector) and the strains overexpressing *yahN*. The data shows a decrease in cell density relative to the control strain when *yahN* are overexpressed relative to the control vector. Note that some other genes were also tested, but no significant difference was seen relative to the control.

To evaluate the significance of the linear change over time, we formally tested hypotheses for no slope versus the existence of a slope for both control and test cases. Note this is a simple t-test in classical simple linear regression. The results of analysis are given in Table 5 which indicates three runs for the *ycbR* and *yfhA* cases and one run for the yahN case. Three runs were performed for the *yahN* case but two are not reported due the experimental problems that caused these results to be inconclusive. Focusing on the average results, in agreement with PCA results, all three genes indicate significance effect as the slopes are indicated to be significant at the 0.002 test level with P-values of 0.00122, 0.00119 and 0.00000196 for *ycbR, yfhA* and *yahN*, respectively. In contrast, none of the control cases have significant slopes for the averaged results with P-values of 0.281, 0.372, and 1.00. The negative slopes suggest that these genes may play an antagonistic role in NO and RNOS defense. However, further research will be needed to understand their affects on NO and RNOS defense which is beyond the scope of this work. Figure 5 is a plot of *yahN* case in Table 5 with fitted trend lines. As shown, the test run indicates a strong negative slope while the control run is quite flat.

IV. CONCLUSION

This work evaluated the ability of the PCA technique of Rollins et al. (2010) to find genes that are differentially expressed *between* assay groups. This technique is an extension of the Rollins et al. (2006) technique that focused strictly on ranking genes for a subset of assays with a common biological interpretation. This work provided the first test of this PCA technique in validation through experimentation on genes not known to have the identified effect. None of the selected genes had very high rank. However, since all three genes strongly indicated significance either on NO or GSNO response, the technique appears to have a degree of sensitivity which translates into power in statistical language. Power is the ability to find genes that are statistically significant. As power increases, the ability to find genes with weaker effect increases. In Rollins et al. (2010) the power of the technique was evaluated in a simulation study and the results compared well with a popular method. While much more experimental work is required to establish the strength of the technique evaluated in this work, these results are quite promising, and we, therefore, encourage its use.

V. ABBREVIATIONS

NO, nitric oxide; GSNO, S-nitrosoglutathione; KCN, potassium cyanide; SeOH, Serine hydroxamate; PCA, Principal Component Analysis; PM, proposed method; NCA, network component analysis; T_{diff}, absolute different in gene contribution of two assay groups; *E.* coli, Escherichia *coli*; EG, Eigengene; EA, Eigenassay; principal components, PCs; RNS, reactive nitrogen species; DeaNO, diethylamine nitric oxide; IPTG, Isopropyl β-D-1-thiogalactopyranoside

VI. REFERENCES

1. Hyduke DR, Jarboe LR, Tran LM, Chou KJY, and Liao JC: **Integrated network analysis identifies nitric oxide response networks and dihydroxyacid**

dehydratase as a crucial target in *Escherichia coli*, *Proceedings of the National Academy of Sciences of the United States of America* **2007**, **104**: 8488-8489.

2. Jarboe LR, Hyduke DR, Tran LM, Chou KJY, and Liao JC: **Determination of S-nitrosoglutathione targets and response networks in *Escherichia coli* using integrated biochemical and systems analysis**, *The Journal of Biological Chemistry* **2008, 283:** 5148-5157.

3. Jarboe LR, Hyduke DR, Liao JC: Nitric Oxide: Systems approaches to unraveling nitric oxide response networks in prokaryotes. In *Biology and pathobiology.* 2nd edition. Edited by Ignarro LJ. California: Academic Press; 2010: 103-136.

4. Voskuil MI, Schnappinger D, Visconti KC, Harrell MI, Dolganov GM, Sherman DR, Schoolnik GK: **Inhibition of respiration by nitric oxide indeuces a *mycobacterium tuberculosis* dormamcy program**, *The journal of experimental medicine* **2003, 198:** 705-713.

5. Raychaudhuri S, Stuart JM, Altman RB: **Principal Component Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series**, *Pacific Symposium on Biocomputing* **2000, 5**: 452-463.

6. Misra JW, Hwang D, Hsiao LL, Gullans S, Stephhanopoulos G: **Interactive exploration of microarray gene expression patterns in a reduced dimensional space**, *Genome Res* **2002, 12**: 1112-1120.

7. Sharov AA, Dudekula DB, and Ko MSH: **A web-based tool for principal component and significance analysis of microarray data,** *Bioinformatics* **2005, 21**: 2548-2549.

8. Derrick K. Rollins, Sr., Dongmei Zhai, Alrica L Joe, Jack W. Guidarelli, Abhishek Murarka, Ramon Gonzalez: **A novel data mining method to identify assay-specific signatures in functional genomic studies**, *BMC Bioinformatics* 2006, **7**:377.

9. Jennifer A. Synder, Brian J. Haugen, Eric L. Buckles, C. Virginia Lockatell, David E. Johnson, Michael S. Donnenberg, Rodney A. Welch, Harry L. T. Mobley: **Transcriptome of Uropathogenic *Escherichia coli* during Urinary Tract Infection**, *Infection and Immunity* 2004, 72:6373.

10. Dmitrii G Rodionov, Edward E Ishiguro: **Ampicillin-induced bacteriolysis of *Escherichia coli* is not affected by reduction in levels of anionic phospholipids**, *FEMF Microbiology Journal* 1997, **156:** 85-89.

11. Using R for a Principal Component Analysis [http://fs6.depauw.edu:50080/~harvey/Chem%20351/PDF%20Files/Handouts/RDocs/Using%20R%20for%20a%20Principle%20Component%20Analysis.pdf].

12. Janet Flatley, Jason Barrett, Steven T. Pullan, Martin N. Hughes, Jeffrey Green, Robert K. Poole: **Transcriptional Responses of *Escherichia coli* to *S*-Nitrosoglutathione under Defined Chemostat Conditions Reveal Major Changes in Methionine Biosynthesis**, *The Journal of Biological Chemistry* 2005, **280**: 10065–10072.

13. **Atlas of Genetics and Cytogenetics in Oncology and Haematology** [http://atlasgeneticsoncology.org]

14. **EcoliWiki** [http://ecoliwiki.net/colipedia/index.php/Welcome_to_EcoliWiki]

15. **EcoCyc** [http://ecocyc.org/]

VII. FIGURE LEGENDS

Figure 1 - Loading plots generated for EG analysis

Loading 2 is used to identify GSNO-response genes and Loading 4 is used to identify NO-response genes. Blue square circle represents one group and the black circles represent the other group.

Figure 2 - Score plots generated for EA analysis

S1 is used to identify GSNO-response gene and S3 is used to identify NO-response genes. Blue square circle represents one group and the black circles represent the other group.

Figure 3 - Signature plots showing gene contribution ranking for NO (left) and GSNO (right) from EG method

Signature plot for NO response is generated using Loading 4 while GSNO response is generated using Loading 2 respectively from EG method.

Figure 4 - Signature plots showing gene contribution ranking for NO(left) and GSNO(right) from EA method

Signature plot for NO response is generated using Score 3 while GSNO response is generated using Score 1 respectively from EA method.

Figure 5 – NO resistance for the *yahN* case in Table for the Test and Control run with10 μm DeaNO pumped and 0.01mm Isopropyl IPTG.

Plot on the left show the response growth versus time for averages of control and *ycbR* while plot of the right corresponds to the averages of *yfhA* and control. From these plots, there are apparent changes in response growth as time increases.

VIII. TABLES

Table 1 – Top 25 genes for NO direct target identified from Eigengene method.

Ranking	Gene	Function
1	*hmp*	Flavohemoglobin; nitric oxide dioxygenase; dihydropteridine reductase; GSNO and nitrite reductase;
2	*norW*	NADH:flavoruberedoxin reductase
3	*norV*	Anaerobic nitric oxide reductase flavorubredoxin
4	*ytfE*	Iron-sulfur cluster repair protein; confers resistance to nitric oxide and hydrogen peroxide; di-iron center
5	*sdhC*	Oxidation of succinate, carries electrons from FADH to CoQ
6	*cydA*	Contains the heme b558 component of cytochrome bd-I
7	*carA*	Component of carbamoyl phosphate synthetase
8	*sdhA*	Catalytic subunits in the four subunit enzyme; contains the FAD cofactor
9	*yeaR*	Conserved protein yeaR
10	*cydB*	Bind the heme b595 component and iron-chlorin component of cytochrome bd-I
11	*sdhB*	Transfer redox centers, delivery electrons from the FAD cofactor in SdhA to the ubiquinone
12	*sdhD*	sdhABCD operon is negatively regulated by ryhB RNA as part of indirect positive regulation by Fur.
13	*iscR*	Transcriptional repressor for isc operon; contains Fe-S cluster; binds RNA in vitro
14	*b0725*	Phantom gene, meaning that at a previous time it was thought to be a gene, but more recent analyses indicate it is not a gene
15	*ndk*	Catalyzes the reaction in which the terminal phosphate of a nucleoside-triphosphate is transferred to a nucleoside-diphosphate
16	*oppA*	Component of oligopeptide ABC transporter; Oligopeptide transport, periplasmic oligopeptide binding protein
17	*cvpA*	Required for wild-type production of colicin V from an episomal gene
18	*yaiB*	Anti-adaptor protein required for stabilization of the alternative sigma factor
19	*sucA*	E1(0) component of the oxoglutarate dehydrogenase complex.
20	*pyrD*	Component of dihydroorotate oxidase
21	*yeeF*	Putative amino acid transporter of the APC superfamily
22	*rnpA*	Ribonucleaus P protein component
23	*glnL*	Sensor-histidine kinase; transmembrane protein composed of three domains
24	*rplM*	50S ribosomal subunit protein L13
25	*fliR*	Integral membrane components of the flagellar export apparatus

Table 2 – Top 25 genes for NO direct target identified from Eigenassay method.

Genes bolded are genes selected to study in the experiment.

Ranking	Gene	Function
1	*norW*	NADH:flavoruberedoxin reductase
2	*norV*	Anaerobic nitric oxide reductase flavorubredoxin
3	*hmp*	Flavohemoglobin; nitric oxide dioxygenase; dihydropteridine reductase; GSNO and nitrite reductase;
4	*ytfE*	Iron-sulfur cluster repair protein; confers resistance to nitric oxide and hydrogen peroxide; di-iron center
5	*yeaT*	Regulate the expression of *yeaU*, which encodes a D-malate dehydrogenase essential for growth on D-malate
6	*hscB*	Involved in iron-sulfur cluster assembly. HscB physically interacts with HscA and with IscU
7	*yfjL*	CP4-57 prophage; predicted protein; function unknown
8	*yoaG*	*yoaG* is thought to form an operon together with *yeaR*. Transcription of *yoaG* is induced in response to nitrate
9	*yeaR*	Conserved protein
10	*argC*	ArgC catalyzes the NADPH-dependent reduction of *N*-acetylglutamyl-phosphate to yield *N*-acetyl-L-glutamate 5-semialdehyde
11	*hcp*	The hybrid cluster protein (HCP) exhibits hydroxylamine reductase activity, possibly functioning as a scavenger of toxic by-products of nitrogen metabolism
12	*iscU*	IscU is a scaffold protein for assembly and transfer of iron-sulfur clusters.
13	*nuoC*	NADH:ubiquinone oxidoreductase subunit C, complex I; NADH dehydrogenase I
14	*ilvC*	Ketol-acid reductoisomerase
15	*iscS*	Component of cysteine desulfurase
16	***yahN***	Predicted neutral amino acid efflux system
17	*fdx*	Ferredoxin, an iron-sulfur protein; involved in assembly of other Fe-S clusters
18	*hypF*	Hydrogenase maturation protein
19	*ilvB*	Component of acetohydroxybutanoate synthase I
20	***ycbR***	Putative periplasmic pilus chaperone, induced by AI-2 pheromone, function unknown
21	*iscX*	Protein with possible role in iron-sulfur cluster biogenesis
22	*puuD*	Component of y-glutamyl-y-aminobutyrate hydrolase
23	*argG*	Component of argininosuccinate synthase
24	*yciE*	Conserved protein
25	*ilvN*	Component of acetohydroxybutanoate synthase I

Table 3 - Top 25 genes for GSNO direct target identified from Eigengene method.

Ranking	Gene	Function
1	*yedN_1*	Hypothetical protein
2	*metR*	Positive regulatory gene for metE and metH; autogenous regulation
3	*metA*	Homoserine O-transsuccinylase
4	*yigM*	Predicted inner membrane protein
5	*carA*	Component of carbamoyl phosphate synthetase
6	*cspB*	Qin prophage; cold shock protein
7	*uspB*	Predicted universal stress (ethanol tolerance) protein B
8	*hmp*	Flavohemoglobin; nitric oxide dioxygenase; dihydropteridine reductase; GSNO and nitrite reductase;
9	*fliR*	Integral membrane components of the flagellar export apparatus
10	*ybdL*	Methionine aminotransferase, PLP-dependent
11	*metN*	L,D-methionine transporter, ATP-binding protein; methionine sulfoximine sensitivity
12	*yaiB*	Anti-adaptor protein that is required for stabilization of the alternative sigma factor
13	*rnpA*	Ribonucleaus P protein component
14	*dnaK*	Chaperone Hsp70; DNA biosynthesis; autoregulated heat shock proteins
15	*rpsA*	30S ribosomal subunit protein S1
16	*mmuP*	S-methylmethionine permease, CP4-6 putative prophage remnan
17	*metF*	Component of 5,10-methylenetetrahydrofolate reductase
18	*glnA*	Adenylyl-[glutamine synthetase], GlnA
19	*mmuM*	S-methylmethionine:(seleno)homocysteine methyltransferase; CP4-6 putative prophage remnant
20	*metB*	Cystathionine gamma-synthase; homotetrameric
21	*cysJ*	Sulfite reductase [NADPH] flavoprotein alpha-component; binds FMN and FAD
22	*rpsU*	30S ribosomal subunit protein S21
23	*glnL*	Sensor-histidine kinase; transmembrane protein composed of three domains
24	*ytfE*	Iron-sulfur cluster repair protein; confers resistance to nitric oxide and hydrogen peroxide; di-iron center
25	*ybdH*	Predicted oxidoreductase

Table 4 - Top 25 genes for GSNO direct target identified from Eigenassay method.

Ranking	Gene	Function
1	*yedN_1*	Hypothetical protein
2	*metR*	Positive regulatory gene for metE and metH; autogenous regulation
3	*metF*	Component of 5,10-methylenetetrahydrofolate reductase
4	*yigM*	Predicted inner membrane protein
5	*metA*	Homoserine O-transsuccinylase
6	*ygcN*	Predicted oxidoreductase with FAD/NAD(P)-binding domain
7	*cysJ*	Sulfite reductase [NADPH] flavoprotein alpha-component; binds FMN and FAD
8	*cysH*	Component of 3'-phospho-adenylylsulfate reductase
9	*cysN*	Component of SULFATE-ADENYLYLTRANS-CPLX; sulfate adenylyltransferase
10	*mmuP*	S-methylmethionine permease, CP4-6 putative prophage remnan
11	*ybdL*	Methionine aminotransferase, PLP-dependent
12	*mmuM*	S-methylmethionine:(seleno)homocysteine methyltransferase; CP4-6 putative prophage remnant
13	*cysD*	Component of Sulfate-Adenylyltrans-CPLX; sulfate adenylyltransferase
14	*sbp*	Component of sulfate ABC transporter
15	*yeeE*	Putative transport system permease protein
16	*metN*	L,D-methionine transporter, ATP-binding protein; methionine sulfoximine sensitivity
17	*metB*	Cystathionine gamma-synthase; homotetrameric
18	*uspB*	Predicted universal stress (ethanol tolerance) protein B
19	*yjdK*	Hypothetical protein; predicted protein
20	*cspB*	Qin prophage; cold shock protein
21	*cysI*	Sulfite reductase [NADPH] hemoprotein beta-component; has 4Fe-4S iron-sulfur center
22	*ybdH*	Predicted oxidoreductase
23	*ydcD*	Hypothetical protein
24	*thrA*	Component of aspartate kinase I
25	*dnaK*	Chaperone Hsp70; DNA biosynthesis; auto-regulated heat shock proteins

Table 5 – P-values for Test versus Control for evaluating the significant effect for three identified genes in Tables 1 and 2

	ycbR		*yfhA*		*yahN*	
Run	Test	Control	Test	Control	Test	Control
1	$3.48(10^{-3})$	$6.73(10^{-1})$	$8.66(10^{-6})$	$1.21(10^{-1})$	$1.96(10^{-6})$	$1.00(10^{-1})$
2	$6.68(10^{-6})$	$7.87(10^{-3})$	$3.40(10^{-4})$	$9.18(10^{-1})$		
3	$1.60(10^{-4})$	$1.65(10^{-1})$	$3.21(10^{-3})$	$7.70(10^{-2})$		
Mean	$1.22(10^{-3})$*	$2.81(10^{-1})$	$1.19(10^{-3})$*	$3.72(10^{-1})$	$1.96(10^{-6})$*	$1.00(10^{-1})$

*****significant at the 0.002 level

Figure 1

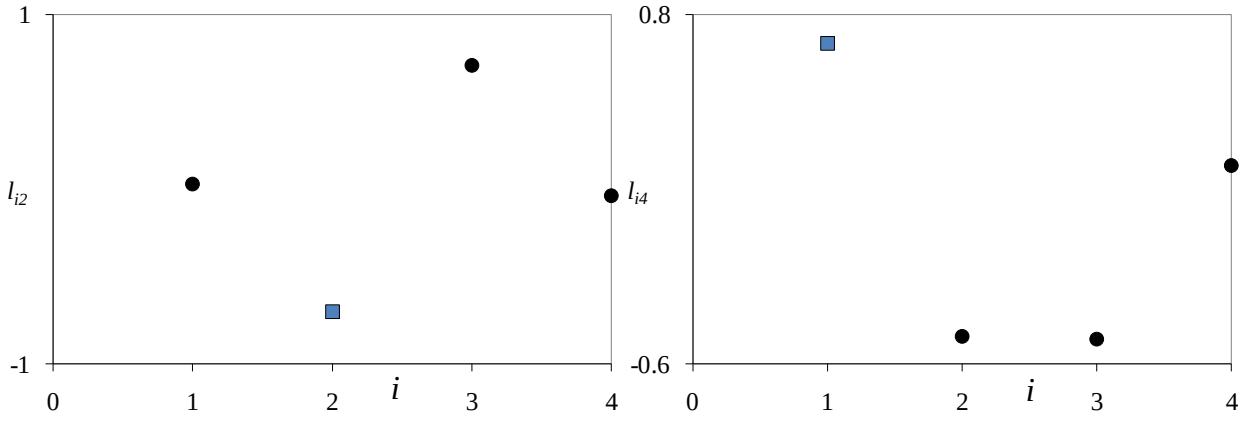


Figure 2

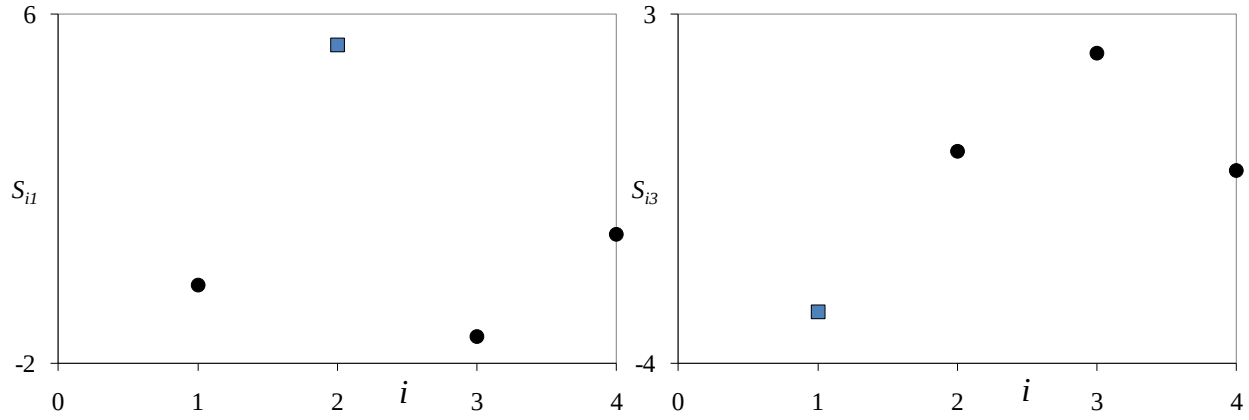


Figure 3

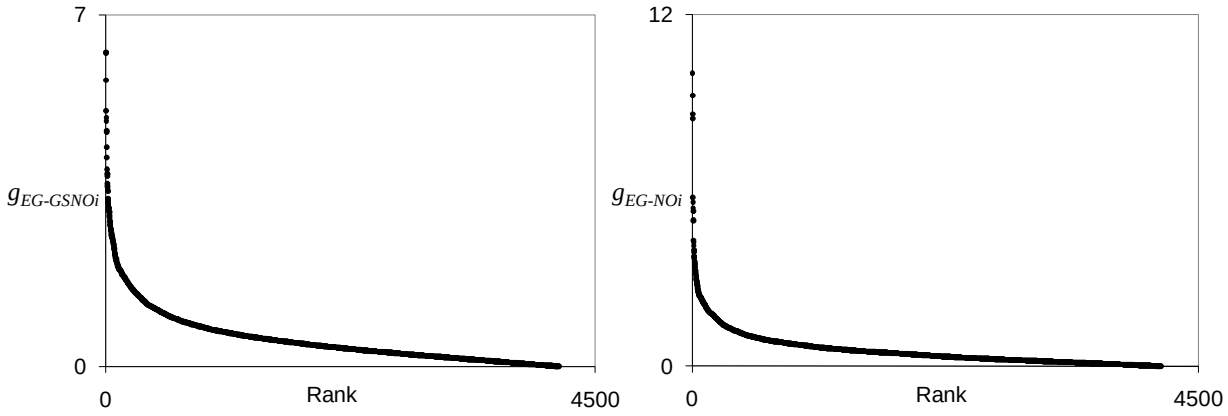


Figure 4

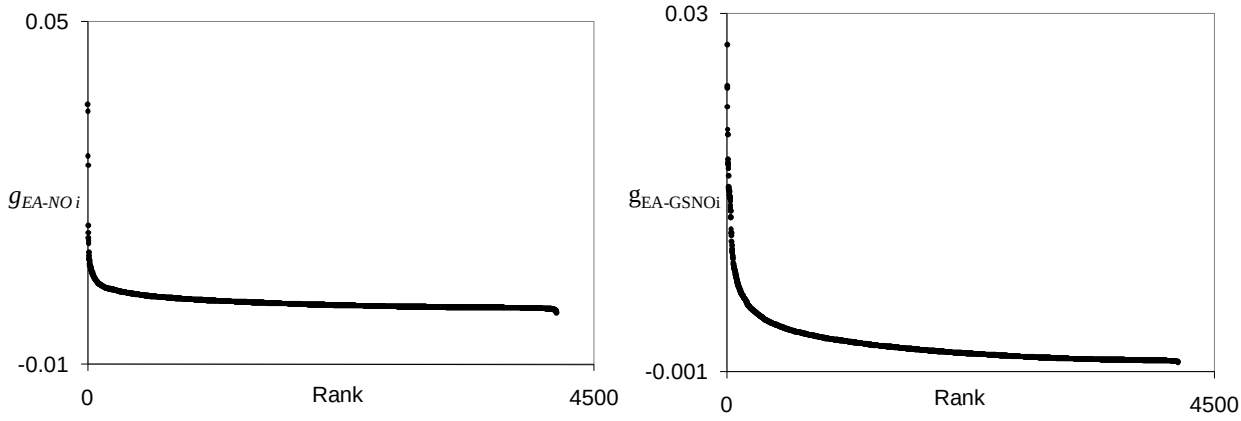
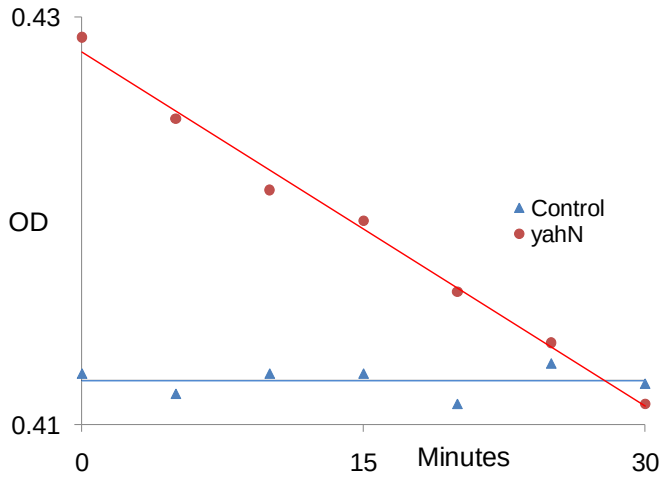


Figure 5



CHAPTER 5. SUMMARY AND FUTURE WORK

I. SUMMARY

The differential PCA method described is a powerful method for analyzing large dimensionality data, t such as microarray data. This method is unique and effective for several reasons. First, it is able to effectively generate an assay specific gene signature that shows a distinct difference for assay groups. Second, it is able to control the FDR and also achieve high power at the same time when performing data analysis. This differential PCA method, which is also an extension of work from Rollins et. al.(2006), is proven to be highly effective through all the real data case studies as well as a simulation study.

We have successfully applied this method to two *E. coli* data sets where one of them deals with the study of two different ethanol levels to the genes [2] and the second study identifies Nitric Oxide (NO) specific and S-nitrosoglutathione (GSNO) specific genes in *E. coli* [3 - 4]. This method also performed very well in a data set that involved the study of the effect of myostatin on mice [5].

This approach is currently being tested on a large dimensionality data set that is not microarray experiment data. At the moment, it is showing great success in the trial runs completed thus far. This indicates that the differential PCA method is not only applicable to microarray experiment data set but also to all other large dimensionality data sets that contain a wealth of information where differential analysis is needed.

II. FUTURE WORK

Principal component analysis is proven to be a powerful statistical tool to analyze microarray data. The current approach uses different statistical software as well as mathematical procedures to perform the analysis. Since this method is performed using different statistical software and procedures, it would be valuable to combine all the individual steps into a single package. This methodology can be further extended by developing a user friendly and easy to operate software package incorporating all methods and steps taken to perform the data analysis.

As verified, this is a powerful statistical method for analyzing large dimensionity microarray data. It may also be worthwhile to try applying this method to other large dimensionality data sets other than microarray data sets when the objective of the research is to identify subjects that express differentially between two or more different conditions. I believe this statistical approach is able to effectively extract important information as well as make a sound interpretation of the data.

III. REFERENCES

1. Rollins D K, Zhai D, Joe AL, Guidarelli JW, Murarka A, Gonzalez R: **A novel data mining method to identify assay-specific signatures in functional genomic studies**, *BMC Bioinformatics* **2006, 7**:377.

2. Gonzalez R, Tao H, Purvis JE, Shanmugam KT, York SW, and Ingram LO: **Gene Array-Based Identification of Changes That Contribute to Ethanol Tolerance in Ethanologenic *Escherichia coli*: Comparison of KO11 (Parent) to LY01 (Resistant Mutant),** *Biotechnol. Prog* **2003, 19**: 612- 623.

3. Hyduke DR, Jarboe LR, Tran LM, Chou KJY, and Liao JC: **Integrated network analysis identifies nitric oxide response networks and dihydroxyacid dehydratase as a crucial target in *Escherichia coli***, *Proceedings of the National Academy of Sciences of the United States of America* **2007**, **104**: 8488-8489.

4. Jarboe LR, Hyduke DR, Tran LM, Chou KJY, and Liao JC: **Determination of S-nitrosoglutathione targets and response networks in *Escherichia coli* using**

integrated biochemical and systems analysis, *The Journal of Biological Chemistry* **2008, 283:** 5148-5157.

5. Steelman CA, Recknor JC, Nettleton D, and Reecy JM: **Transcriptional profiling of myostatin-knockout mice implicates Wnt signaling in postnatal skeletal muscle growth and hypertrophy**, *The FASEB Journal* **2006, 20**: 580-582.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my greatest thanks to my dearest major professor, Dr. Derrick Rollins who is always so patient and gentle in guiding and leading me through my course of master study. His indomitable spirit has inspired and motivated me immensely during my masters program as well as my daily life.

Secondly, I would like to say thank you to my parents and my family members who love and care for me so much. They are always there to support and encourage me. Wei-Chyin, to me you are like my family, and I am grateful that you always stand by me to give me your full support and care whenever I need it.

Finally, to my academic committee members, I really appreciate your time and patience in guiding me through my research. To all my group-mates and friends who are always there to help me and support me. Thank you!