Spring 4-5-2013

Identification of Differential Gene Pathways with Sparse Principal Component Analysis

Yichao Yin

IDENTIFICATION OF DIFFERENTIAL GENE PATHWAYS WITH SPARSE PRINCIPAL COMPONENT ANALYSIS

by

YICHAO YIN

Under the Direction of Ruiyan Luo

ABSTRACT

The development of the technology makes it possible to measure large amount of genes expressions simultaneously. Since biological functions are mostly coordinated by multiple genes, called “gene pathway”, it is interesting to identify differential gene pathways which are associated with clinical phenotype. Principal component analysis has been proposed to identify differential gene pathways in several literatures, while sparse principal component analysis (SPCA) has not drawn any attention. We proposed to use SPCA to identify differential gene pathways. The results show that, comparing to PCA, SPCA could identify more differential expressed gene pathways, especially when the higher-order interactions among genes are considered.

INDEX WORDS: Gene pathways, Sparse principal component analysis

IDENTIFICATION OF DIFFERENTIAL GENE PATHWAYS

WITH SPARSE PRINCIPAL COMPONENT ANALYSIS

by

YICHAO YIN

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2013

IDENTIFICATION OF DIFFERENTIAL GENE PATHWAYS

WITH SPARSE PRINCIPAL COMPONENT ANALYSIS

by

YICHAO YIN

Committee Chair: Dr. Ruiyan Luo

Committee: Dr. Ruiyan Luo

Dr. Gengsheng Qin

Dr. Yichuan Zhao

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2013

ACKNOWLEDGEMENTS

I am heartily thankful to my respected supervisor, Dr. Ruiyan Luo, who offers her guidance and support to me with great patience.

Many thanks to Dr. Gengsheng Qin, Dr. Yichuan Zhao, Dr. Xu Zhang, Dr. Jiawei Liu, Dr. Jun Han and Dr. Xin Qi for their generous help and insights they shared with me.

I would also like to thank all my classmates and colleagues in the department of math and statistics. Together we make this journey fun and rewarding. I wish them all the best in their future.

TABLE OF CONTENTS

LIST OF TALBES

LIST OF FIGURES

1. INTRODUCTION

The field of bioinformatics has developed rapidly during the past decade. One of its most important problems is to identify particular genomic characters which are related to different phenotypes. With the development of microarray technology, it is possible to measure expressions of thousands of genes and study their relationship with clinical outcomes. If particular genes or gene sets are associated with clinical outcomes significantly in a regression model, the genes or gene sets could be treated as "differentially expressed" (Ma & Kosorok, 2009).

The methodologies to identify differential genes generally involve two steps. First, a significant statistic should be calculated based on the gene expression data for each gene. Because of randomness, there might be some extreme measurements. To remove those outliers, several methods have been proposed, such as shrinkage, penalization and thresholding (Benjamini & Yekutieli, 2001). Second, multiple testing will be generated to adjust and determine which genes should be treated as differential ones. Specifically, the false discovery rate (FDR) controlling procedure is a good alternative providing satisfying results for such problems (Allison, Cui, Page, & Sabripour, 2006).

In fact, based on recent studies, variations of certain phenotypes are not only related to just one differentially expressed gene but multiple ones, especially for those clinical outcomes from complex diseases, like cancer. So far, various approaches have been proposed to investigate the association between multiple genes and clinical outcomes. Generally speaking, there are two main analysis directions.

One direction is, taking genes as sampling units and testing whether the gene set of interest is overrepresented in the list of genes. Ackermann and Strimmer (2009) summarized a common modular structure for most published methods in this direction. The modular structure involves five steps: calculating a statistic for each gene, making necessary transformation, choosing a null hypothesis,

calculating a statistic for gene set, and assessing the significance. One of the most well-known approaches with this structure is the gene set enrichment analysis (GSEA) (Subramanian, et al., 2005). It uses the enrichment score which is related to a weighted Kolmogorov-Smirnov-like statistic to measure the significance of specific gene set in different phenotypes. After two years, a more powerful approach was suggested by altering the enrichment score into maxmean statistic (Efron & Tibshirani, 2007). It is superior to other methods especially for gene set analysis when the experiment has few replicates, such as cases for prokaryotes (Tintle, et al., 2008).

Another direction in investigating the association between gene sets and phenotypes uses clinical subjects as samples. This direction is more realistic on performing the biological experiment, since a biological replication always takes a new sample of subjects, not a new sample of genes (Goeman & Buhlmann, Analyzing gene expression data in terms of gene sets: methodological issues, 2007). The global test (Goeman, Van de Geer, De Kort, & Van Houwelingen, 2004) is a well-known example in this direction. It is based on an empirical Bayesian generalized linear model proposed to test whether the global expression pattern of a group of genes is strongly related to some clinical outcome of interest. Nettleton, Recknor and Reecy (2008) proposed a nonparametric multivariate method by considering the joint expression distribution of gene sets across two or more conditions. Ma and Kosorok (2009) proposed using principal component analysis (PCA) to detect the differentially expressed gene pathways. They pointed out that when using clinical outcomes and gene expressions from pathways to fit the regression model, the number of genes in a pathway may be larger than the sample size. In that case, the models would be saturated and the significance based on the regression model between clinical outcomes and gene expression from pathways might be improper. To deal with this problem, they use the linear combinations of gene expressions (which are the principal components of the gene expression data) as “representative features”, and use the clinical outcomes and the representative features to fit regression models. Based on the property of principal components, those representative features would

reduce the dimension of gene expression data and still keep most of the gene information. In addition, by reducing the dimension of gene expression data, it is possible to include the higher-order interactions between gene expressions into the regression models. Ma & Kosorok (2009) identified new differential gene pathways by including second-order terms of gene expressions.

However, classical PCA has some main drawbacks. One of them is that the principal components are linear combination of all original variables and most of the coefficients are non-zero. This often makes it hard to interpret the principal components, especially when the original dimension is high. To address this, several sparse principal component analysis methods have been proposed (Qi et al., Zou et al., Shen and Huang, Witten et al., Johnstone and Lu). We propose to use Qi et al.'s sparse principal component analysis (SPCA) to identify differentially expressed gene pathways, which have not obtained any attraction yet.

2. METHOD

2.1 Sparse Principal Component Analysis (SPCA)

Principal component analysis is a popular method to reduce data dimension and interpret the data. Theoretically, getting the first principal component could be identified as follows: We have a $n \times p$ data matrix $\mathbf{X}$ where n and p are the number of the observations and the number of the variables, respectively. Without loss of generality, we assume that the column means (sample means) are all zeros. Let $\mathbf{\Sigma} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}$ be the $p \times p$ sample covariance matrix, where $\mathbf{X}^{\mathrm{T}}$ denotes the transpose of the data matrix $\mathbf{X}$. For any vector $\mathbf{v} = (v_1, \cdots, v_p)^T \in \mathbb{R}^p$, let $\|\mathbf{v}\|_2 = \left(\sum_{i=1}^{p} {v_i}^2\right)^{\frac{1}{2}}$ be the l_2-norm of $\mathbf{v}$. The coefficient vector $\mathbf{v} \in \mathbb{R}^p$ of the first principal component is the solution to the following optimization problem,

$$\max_{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|_2=1} \frac{\mathbf{v}^T\mathbf{\Sigma}\mathbf{v}}{{\|\mathbf{v}\|_2}^2} = \max_{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|_2=1} \mathbf{v}^T\mathbf{\Sigma}\mathbf{v}$$

Figure 1 illustrates the solution to the above problem in a two dimensional case, while the blue circle indicates the restriction $\|\mathbf{v}\|_2 = 1$ and the series of ellipses represent $\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} = \mathrm{R}^2, \mathrm{R} \in \mathbb{R}$. The red ellipse and the blue circle are tangent, which gives the solution at the point of tangency. Usually, the solution point would not lie on any axis, which means the coefficient vector $\mathbf{v}$ would have most of the components to be non-zero.

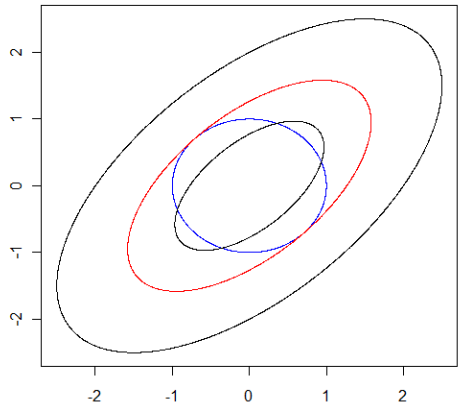

Figure 1: Case for PCA

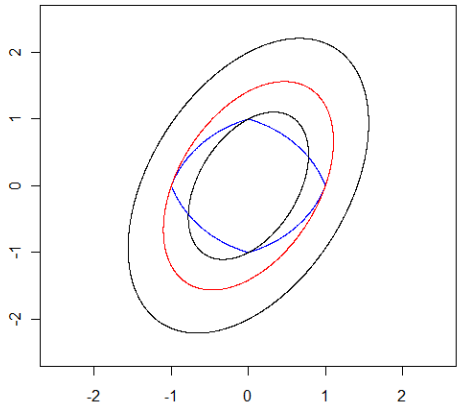

Figure 2: Case for SPCA

However, when it comes to the gene expression data, the high dimension non-zero coefficient vector $\mathbf{v}$ will make it difficult to interpret the contribution from genes to the principal component. So SPCA is proposed to revise PCA method, aiming at finding the linear combination with only a few non-zero coefficients and explain most of the variance.

In this thesis, we suggest to use the SPCA proposed by Qi, Luo & Zhao (2013), which can efficiently obtain uncorrelated principal components. They replaced the l_2-norm of $\mathbf{v}$ in the optimization problem with a "mixed norm". Let $\|\mathbf{v}\|_1 = \sum_{i=1}^{p}|v_i|$ be the l_1-norm of $\mathbf{v}$. For any $\lambda \in [0,1]$, they defined the "mixed norm" $\|\cdot\|_\lambda$ in $\mathbb{R}^p$ space. For any $\mathbf{v} \in \mathbb{R}^p$,

$$\|\mathbf{v}\|_\lambda = \left[(1-\lambda)\|\mathbf{v}\|_2{}^2 + \lambda\|\mathbf{v}\|_1{}^2\right]^{\frac{1}{2}} \qquad (*)$$

It is obvious that if $\lambda = 0$, this norm is just the l_2-norm of $\mathbf{v}$, and if $\lambda = 1$, it is the l_1-norm. Using this norm, the optimization problem solving the coefficient vector $\mathbf{v} \in \mathbb{R}^p$ of the first principal component becomes as follows,

$$\max_{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|_2=1} \frac{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}{(1-\lambda_1)\|\mathbf{v}\|_2{}^2 + \lambda_1\|\mathbf{v}\|_1{}^2} = \max_{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|_2=1} \frac{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}{\|\mathbf{v}\|_{\lambda_1}{}^2}$$

where $0 \le \lambda_1 \le 1$ is the tuning parameter for the first principal component. Similarly as Figure 1, Figure 2 illustrates the solution to the above problem in a two dimensional case. It is clear to see that the points of tangency might just be the vertexes of the blue curve lying on axises. The tuning parameter controls the sparseness of the solution. With a larger λ_1, the points of tangency are more likely to be the vertexes, leading to sparser coefficients but less proportion of variations explained by the first principal component.

For higher-order principal components, since the SPCA could only satisfy at most one of the following two properties from PCA: orthogonal coefficient vectors of different principal components or uncorrelated principal components, there are two alternative definitions for higher-order principal components for SPCA. We choose the definition with uncorrelated principal components for our SPCA. Suppose that we already have coefficient vectors $\mathbf{v}_j$, $1 \le \mathrm{j} \le \mathrm{k}-1$, then the optimization problem to solve $\mathbf{v}_k$ is as follows,

$$\max_{\substack{\|\mathbf{v}\|_2=1, \mathbf{v}\perp\boldsymbol{\Sigma}\mathbf{v}_j, \\ \mathrm{j}=1,\ldots,\mathrm{k}-1}} \frac{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}{\|\mathbf{v}\|_{\lambda_k}{}^2}$$

We refer to Qi, Luo and Zhao (2013) for detailed algorithms for solving the above optimization problems. We set the tuning parameter λ based on p, the dimension of sample covariance matrix $\boldsymbol{\Sigma}$. Our goal is to get the sparse PCs with most of the coefficients being zero while they still explain most of the variations comparing with classical PCs. Due to the computational cost, we use the same tuning parameter λ within calculating PCs for the same $\boldsymbol{\Sigma}$. In the numerical study, we select λ as follows:

Table 1: Selection of Tuning Parameter

Dimension of Sample Covariance Matrix **Σ**	[0,20]	(20,100]	(100,500]	(500,1000]	>1000
Tuning Parameter λ	0	0.02	0.01	0.005	0.001

2.2 Construction of representative features

Our goal is to identify differentially expressed pathways using a small number of representative features from the gene expression data for each pathway. Considering the PCA with gene expression data and the non-linear effects, Ma and Kosorok (2009) suggested the following three ways to construct the representative feature for gene pathways. For a pathway with m genes, denote $X_1, \dots, X_m$ as the gene expressions.

(R1) They took the first c PCs of these m gene expression measurements as the representative features. They suggested to consider c^* different sets ($\mathrm{c} = 1, \dots, c^*$) of representative features, since we don't know how many PCs should be chosen. In their numerical study, they chose $c^* = 5$, and so would we. Comparing different sets of PCs could partly tell us how many PCs to choose. Since they would compare those pathways with at least c^* genes and smaller pathways (less than c^* genes) could be investigated by individual gene, they excluded those smaller pathways in their study.

(R2) They considered the second-order expanded gene set for each pathway, like $\{X_1, \dots, X_m, X_1X_1, \dots, X_1X_m, X_2X_2, \dots, X_2X_m, \dots, X_mX_m\}$, then consider the first c PCs, while $\mathrm{c} = 1, \dots, c^*$. In this case, the PCs are the linear combination of the gene expression and their second-order terms.

(R3) They also considered another way to compose the second-order terms. They first chose d PCs, then constructed the representative features with these PCs and their second-order terms. To compare the results from different number of PCs, they considered d^* different sets ($\mathrm{d} = 1, \dots, d^*$) of representative

features. In their numerical study, they chose $d^* = 3$. The representative feature set would look like $\{PC_1, \dots, PC_d, PC_1PC_1, \dots, PC_1PC_d, PC_2PC_2, \dots, PC_2PC_d, \dots, PC_dPC_d\}$.

These three different ways provide totally $c^* + c^* + d^*$ different sets of representative features for each pathway, while those from (R1) only consider the linear effects of genes and those from (R2) and (R3) consider the contributions from higher order terms in two different ways.

We would use SPCA to find the sparse PCs and then use the above method to construct representative features.

2.3 Identification of differential pathways

We collected gene expression data along with clinical outcome Y to identify differential pathways as follows.

(1) Construct gene pathways using information from public gene databases. We use the KEGG (Kyoto Encyclopedia of Genes and Genomes) (http://www.genome.ad.jp/kegg/) to get all the pathway information, and construct gene pathways by matching genes in all the KEGG pathways and genes from the gene expression data. As a result, all the pathways in our study will only contain genes with gene expression data.

(2) For each pathway, based on section 2.2, we can construct $c^* + c^* + d^*$ different sets of representative features. For each set of representative features, we fit a regression model while Y is the response and the representative features are covariates. We compute a summary statistic T to reflect the relationship between Y and representative features. Then we do a permutation study by randomly permuting Y for B times, and then fitting the same regression model and computing the summary statistic for each time. Finally, we compute a permutation P-value for T with all the B summary statistics. In our study, we set $\mathrm{B} = 50000$.

(3) For each set of representative features, we have permutation P-values for all the pathways. Using the

approach suggested by Ma and Kosorok (2009) to control the false discovery rate (FDR), we set the expected FDR to $\mathrm{q} = 0.1$, order the P-values of N pathways as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$, and let r be the largest i such that $p_{(i)} \leq \mathrm{i/N} \times \mathrm{q}/\left(\sum_{i=1}^{N} 1/i\right)$. Then, those pathways corresponding to $p_{(1)} \ldots p_{(r)}$ will be defined as differentially expressed.

3. DATA ANALYSIS

3.1 Data

DLBCL (diffuse large B-cell lymphoma) is a cancer of B-cell and is the most common type of non-Hodgkin lymphoma. We studied the clinical data for 240 patients with untreated DLBCL from http://llmpp.nih.gov/DLBCL/. The median of followup years is 2.8, and 138 observations died. A total of 7399 genes expressions were measured for all the observations. We get pathway information from KEGG. A total of 1614 genes belong to 178 KEGG pathways, with size ranging from 1 to 214 and median size 24.5. Among the 178 pathways, 142 have sizes equal to or larger than 5 and will be studied in our analysis.

The gene expression data have about 10.4% missing value. We input the mean expression value of the same gene from other samples, so that these genes or samples will get a small variance and less weight in the analysis (Goeman, Van de Geer, De Kort, & Van Houwelingen, 2004).

3.2 Process

For each pathway, there is a corresponding gene expression data matrix along with the clinical outcome, which is survival data in DLBCL's case. We conduct the following analysis.

(1) For each pathway, use SPCA to construct $c^* + c^* + d^*$ different sets of representative features with the gene expression data matrix.

(2) For each set of representative features, fit the regression model with clinical outcomes as the response variable and the representative features as the covariates. Since in this case, the clinical outcomes are survival data, the Cox proportional hazards model will be used and the statistic of the score test is chosen as the summary statistic. Then compute the permutation P-value for each set of representative features.

(3) For each set of representative features, combine and analyze the permutation P-values from all the pathways using the FDR approach. Summarize the identified differentially expressed pathways.

(4) Compare all the results for different sets of representative features.

(5) Construct representative features with classical PCA (simply set the tuning parameter as zero in Qi et al.'s SPCA (2013)), repeat step (2)-(4). Compare the results with SPCA.

3.3 Results

Ma and Kosorok (2009) conducted PCA on this data set. With only linear effects, they suggested that using the first three principal components would identify the most differentially expressed pathways (29) while keeping the number of representative features small. We set $\lambda = 0$ in (*) and identify differentially expressed pathways with the first 1, 2… and 5 principal components on the linear effects, respectively. Note that when $\lambda = 0$, the SPCA proposed by Qi et al. (2013) reduces to the classical PCA. We compare our list of identified pathways with those reported in Ma and Kosorok (2009) and list the result in Table 2. When we use the first three principal components obtained under scheme (R1), we also identified the most differentially expressed pathways (39). Here we set $\mathrm{q} = 0.1$ in FDR, where $\mathrm{q} = 0.2$ in Ma and Kosorok (2009). We would get even more pathways if $\mathrm{q} = 0.2$. Among the 39 pathways identified with the first three principal components of the linear effects, 13 pathways match Ma and Kosorok (2009)'s, and 8 more are confirmed to be associated with DLBCL in literature. The differences between our results and Ma and Kosorok (2009)'s should be due to updated pathway information and/or

probably different pre-processing on data. Since the pathway information and procedure of pre-processing the data in Ma and Kosorok (2009) is unavailable, in the following, we compare our results from SPCA and the reduced SPCA with $\lambda = 0$.

Table 2: Comparison of Identified Pathways Results

Method	Ma & Kosorok's PCA (R1)	Our PCA (R1)				
Number of PCs	3	1	2	3	4	5
Number of Differential Expressed Pathways	29(29)*	20(9)	26(15)	39(21)	25(14)	23(12)

*: number in parentheses is the number of pathways which have been confirmed the association with DLBCL in literature or match Ma & Kosorok's result.

Our PCA approach does identified some new pathways. For example, we identified the Antigen processing and presentation pathway and the Systemic lupus erythematosus pathway as differential, while these two pathways are not in Ma and Kosorok's results. Recent researches show that, the Antigen processing and presentation pathway is associated with differential methylated genes between two different type of DLBCLs (Shaknovich, et al., 2010) and patients with systemic lupus erythematosus have an increased risk of DLBCL (Löfström, Backlin, Pettersson, Lundberg, & Baecklund, 2011). Also, we identified the Cell-cycle pathway as differential with both linear and non-linear effects with PCA approach, while it is only identified in Ma & Kosorok's paper with non-linear effects. Ma & Kosorok mentioned that the Cell-cycle pathway contains some important genes which are associated with lymphomas. In addition, cell cycle deregulation found in DLBCLs (Monti, et al., 2012) confirms the relationship between them.

Considering the difference in the pathway information and procedure of pre-processing the data, we compare our results from SPCA and the reduced SPCA with $\lambda = 0$, while the latter is actually PCA. The number of pathways identified in each case is listed in Table 3.

Table 3: Comparison of Identified Pathways between PCA and SPCA

Construction of Representative Feature	R1						R2						R3		
Number of PCs	1	2	3	4	5		1	2	3	4	5		1	2	3
PCA	20	26	39	25	23		3	0	0	0	0		14	20	11
SPCA	27	31	34	27	29		10	0	0	0	0		20	25	18

When constructing the PCs with only linear effects (R1), both SPCA and PCA identified most differentially expressed pathways (34 and 39, respectively) with the first three PCs. Using the first one, two, four, or five PCs as representative features, SPCA identified more differential pathways. When including gene interactions in constructing the PCs (R2 and R3), SPCA found more differentially expressed pathways than PCA for each of the number of PCs included in the study. Generally speaking, SPCA identified more differential pathways than PCA. We further studied the pathways that are only found by SPCA, and searched literatures supporting their association with DLBCL. The following three pathways are found to be associated with DLBCL: the Hedgehog (HH) signaling pathway, the Axon guidance pathway and the Toll like receptor signaling pathway.

Singh, et al. (2010) found that dysregulation of HH pathway is involved in the biology of DLBCL. They found that the HH signaling inhibition induces predominantly cell-cycle arrest and apoptosis in DLBCL cells of germinal center (GC) B-cell type and activated B-cell (ABC) type, respectively. After HH signaling inhibition in DLBCL cells, apoptosis of ABC type was associated with the downregulation of BCL2. Functional inhibition of BCL2 significantly increased apoptosis induced by HH inhibition in DLBCL cells. It is found that DLBCL cells synthesize, secrete and respond to endogenous HH ligands, supporting for the existence of an autocrine HH signaling loop. Their findings imply HH signaling as a potential therapeutic target in DLBCL, in particular for those lymphomas expressing the HH receptor smoothened.

The Axon guidance pathway is differential because several genes involved in axon guidance are directly regulated by ZBTB7A, which specifically represses the transcription of a major tumor suppressor gene

(p14ARF), and was expected to play a significant role in carcinogenesis. In transgenic mice, the oncogenic nature of ZBTB7A was examined in vivo and ZBTB7A was overexpressed in T and B lymphoid lineage cells. These transgenic mice developed thymic lymphomas. Other studies also suggest that ZBTB7A is overexpressed in DLBCL, follicular lymphomas, anaplastic large cell lymphoma and angioimmunoblastic lymphoma (Apostolopoulou, Pateras, Kotsinas, & Gorgoulis, 2011).

These two pathways are identified as differential with linear effects only in SPCA. In addition, the Toll like receptor (TLR) signaling pathway is identified with non-linear effects in SPCA while study on the TLR9 implies potential relationship between the Toll like receptor signaling pathway and the DLBCL (Huang, Weng, Huang, Lin, Tsai, & Chuang, 2012).

Additionally, table 3 shows that more pathways are identified with (R3) than with (R2). Recall that both (R2) and (R3) involve gene interactions in PCs. But in (R2), PCs are linear combinations of both linear effects and gene-gene interactions. In (R3), PCs only consist of linear effects and products of PCs are regarded as interactions between genes. When the number of genes in a pathway is big, scheme (R2) could easily lead to higher dimension than sample size, and the large amounts of second order terms would bring in quite a lot of noise information, which distort the information contained in PCs. However, (R3) constructs the higher-order terms with PCs which already extract most information from the gene expression data, so those representative features in (R3) would contain more representative information from pathways than those in (R2). This can also explain why the SPCA finds more pathways than PCA in (R2) and (R3). Restricting some coefficients be zero, PCs obtained from SPCA are linear combinations of genes (R3) and/or gene-gene interactions (R2) that contribute most of the variations in the expression data by excluding the “noise” signals. The representative features in (R3) only include the “important” genes and their interactions. These denoised features better capture the true variations in the original dataset and hence serve better as representative features.

At last, the results show that SPCA with (R3) would identify most differential pathways among all the

methods considering non-linear effects. The probable reason is that sparse-PCs are linear combination of a small part of genes and the interaction of Sparse-PCs will only contain the information from those genes and their interactions, so the representative features exclude most of the noise while keeping acceptable variations.

4. DISCUSSION

4.1 Interpretation of differential pathways

In this study, we focus on using PCA and SPCA to construct "representative features" of pathways and then based on the associations of these features with phenotypes to identify differentially expressed pathways. PCA has a drawback since PCs are linear combination of all genes, and it is difficult to interpret the "representative feature" for each pathway (Ma & Kosorok, 2009). Sparse-PCA constructs the PCs with linear combination of only small part of all genes and makes it possible to interpret the representative features and further study the mechanism of the identified differentially expressed pathways.

4.2 Non-linear effects in pathway analysis

In this study, we follow the idea from Ma & Kosorok (2009) and use SPCA to construct "representative features" to investigate the non-linear effects in our analysis. Two mechanisms are applied to incorporate the non-linear effects. In (R2), we apply SPCA on the covariance matrix of gene expression levels and products of gene expression levels, with the latter being non-linear effects. In (R3), we first apply SPCA on the covariance matrix of gene expression levels and then get the products of PCs, with the latter capturing non-linear effects. Analysis on DLBCL suggests that, incorporating the non-linear effects does identify some new pathways which are missed by using linear effects only. Based on our analysis,

(R3) is a better way to construct the representative feature with non-linear effects than (R2), implying that products of PC scores better capture the important information in gene-gene interactions, while information in PCs obtained from PCA and/or SPCA on the covariance matrix of gene expression levels and products of gene expression levels is weakened by the noises introduced by the huge amount of product terms. In our analysis, we only considered the second order non-linear effects. In fact, with the advantage of SPCA shown in our case, it is feasible to consider even higher order non-linear effects. Using fewer genes to construct the representative feature by SPCA will also solve the lack of interpretability problem (Ma & Kosorok, 2009).

5. CONCLUSION

In this study, we propose to identify differential gene pathways using SPCA instead of classical PCA. Our case studies suggest that (i) SPCA could more effectively capture the pathway information, assess the association between pathways and clinical outcomes, and then identify more pathways, compared to classical PCA; (ii) when non-linear gene effects are considered, constructing PCs first draws better results than constructing transformations of gene expressions first; (iii) when non-linear gene effects are considered, SPCA shows great advantages as it effectively extracts gene information and makes it possible to interpret the results biologically.

We only include second order non-linear effects (first order gene-gene interaction in model). With the ability of SPCA in dealing with high dimensional data, it is possible to include even higher order interactions to obtain representative features of pathway, which will be our future study.

BIBLIOGRAPHY

Ackermann, M., & Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics, 10*, 47.

Allison, D. B., Cui, X., Page, G. P., & Sabripour, M. (2006, January). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics, 7*, 55-65.

Apostolopoulou, K., Pateras, I. S., Kotsinas, A., & Gorgoulis, V. G. (2011, December). ZBTB7A (zinc finger and BTB domain containing 7A). *Atlas of Genetics and Cytogenetics in Oncology and Haematology, 15*, 1066-1074.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*, 1165-1188.

Cillessen, S. A., Hess, C. J., Hooijberg, E., Castricum, K. C., Kortman, P., Denkers, F., et al. (2007, December 1). Inhibition of the Intrinsic Apoptosis Pathway downstream of Caspase-9 Activation Causes Chemotherapy Resistance in Diffuse Large B-Cell Lymphoma. *Clinical Cancer Research, 13*(23), 7012-21.

Efron, B., & Tibshirani, R. (2007, June). On Testing the Significance of Sets of Genes. *The Annals of Appiled Statistics, 1*, 107-129.

Goeman, J. J., & Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics, 23*, 980-987.

Goeman, J. J., Van de Geer, S. A., De Kort, F., & Van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioimformatics, 20*, 93-99.

Gupta, M., Han, J., Stenson, M., Maurer, M., Wellik, L., Hu, G., et al. (2012, March 22). Elevated serum IL-10 levels in diffuse large B-cell lymphoma: a mechanism of aberrant Janus kinase 2 activation. *Blood, 119*(12), 2844-53.

Huang, W., Weng, S., Huang, C., Lin, H.-C., Tsai, P.-C., & Chuang, J.-H. (2012, November). Expression of Toll-like

receptor9 in diffuse large B-cell lymphoma: further exploring CpG oligodeoxynucleotide in NFkB pathway. *APMIS, 120*(11), 872-81.

Löfström, B., Backlin, C., Pettersson, T., Lundberg, I., & Baecklund, E. (2011, Sepember). Expression of APRIL in diffuse large B cell lymphomas from patients with systemic lupus erythematosus and rheumatoid arthritis. *The Journal of rheumatology, 38*(9), 1891-7.

Ma, S., & Kosorok, M. R. (2009). Identification of differential gene pathways with principal component analysis. *Bioinformatics, 25*, 882-889.

Monti, S., Chapuy, B., Takeyama, K., Rodig, S. J., Hao, Y., Yeda, K. T., et al. (2012, September 11). Integrative Analysis Reveals an Outcome-Associated and Targetable Pattern of p53 and Cell Cycle Deregulation in Diffuse Large B Cell Lymphoma. *Cancer Cell, 22*(3), 359-372.

Nettleton, D., Recknor, J., & Reecy, J. M. (2008). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics, 24*, 192-201.

Qi, X., Luo, R., & Zhao, H. (2013). Sparse principal component analysis by choice of norm. *Journal of Multivariate Analysis, 114*, 127-160.

Rai, D., Kim, S.-W., McKeller, M. R., Dahia, P. L., & Aguiar, R. C. (2010, February 16). Targeting of SMAD5 links microRNA-155 to the TGF-β pathway and lymphomagenesis. *Proceedings of the National Academy of Sciences of the United States of America, 107*(7), 3111-6.

Shaknovich, R., Geng, H., Johnson, N. A., Tsikitas, L., Cerchietti, L., Greally, J. M., et al. (2010, November 18). DNA methylation signatures define molecular subtypes of diffuse large B-cell lymphoma. *Blood, 116*, e81-e89.

Singh, R. R., Kim, J., Davuluri, Y., Drakos, E., Cho-Vega, J., Amin, H. M., et al. (2010). Hedgehog signaling pathway is activated in diffuse large B-cell lymphoma and contributes to tumor cell survival and proliferation. *Leukemia, 24*, 1025-1036.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.

Proceedings of the National Academy of Sciences, 102, 15545-15550.

Tintle, N. L., Best, A. A., DeJongh, M., Van Bruggen, D., Heffron, F., Porwollik, S., et al. (2008). Gene set analyses for interpreting microarray experiments on prokaryotic organisms. *BMC Bioinformatics, 9*, 469.