Sparse Principal Component Analysis (SPCA) of Wheat Microarray Data Identifies Co-Expressed Genes Differentially Regulated by Cold Acclimation

Amrit Karki[1,3], Xijin Ge[1], Din Chen[2], and Fedora Sutton[3]

[1]Department of Mathematics and Statistics, South Dakota State University, Harding hall 211, Brookings, SD 57007
[2]School of Nursing, University of Rochester Medical Center, 601 Elmwood Avenue, NY 14642
[3]Department of Plant Science, South Dakota State University, 1205 Jackrabbit Dr., Box 2108, Brookings, SD 57007

Abstract

DNA microarray technology is a powerful tool for high-throughput analysis that has been used for the purpose of monitoring expression levels of thousands of genes simultaneously and identifying those genes that are differentially expressed. The high dimensionality of microarray data, the expression of thousands of genes in a much smaller number of samples, presents challenges that affect the validity of analytical results. A main issue in microarray transcription profiling is data mining and analysis. Statistical methods are vital for these scientific endeavors. We utilized data obtained from a transcriptome analysis of cold acclimation effects on two winter wheat mutant lines varying in freeze survival. The line with 75 % survival was designated freeze resistant (FR) and the line with 30 % survival was designated freeze susceptive (FS). After pre-processing of the microarray data, we compared the results obtained with and without Sparse Principal Component Analysis (SPCA) on the annotated gene sets. From a starting dataset of 61,115 genes, the significantly differentially expressed genes (DEGs) obtained without using SPCA were 15 in FR, 246 in both (FR and FS), and 36 in FS. However, the significantly DEGs identified with SPCA were 14 in FR, 226 in both (FR and FS), and 36 in FS. Given the small dataset, SPCA was still able to reduce the starting annotated dataset from 1321 to only 1211.

Key Words: Microarray, transcriptome, sparse principal component, differentially expressed, p-value

1. Introduction

The high dimensionality of microarray data, the expression of thousands of genes in a much smaller number of samples, presents challenges that affect the validity of analytical results (Nikulin et al., 2009). Principal component analysis (PCA) is a classical tool widely used in data processing and dimensionality reduction. Basically, PCA consists of finding a few orthogonal directions in the data space, which preserve the most information in the data. This is accomplished by finding directions that would maximize the variance of the projections of the data points along these directions (He et al., 2011). PCA generally produces mostly non-zero entries and each principal component is a linear combination of all the original variables (Zou et al., 2004). The main motivation for SPCA is that the largest PCA component is difficult to interpret as usually all components are nonzero.

This article used the SPCA method for dimension reduction of wheat microarray transcriptome data. The main objective of the SPCA in wheat microarray data analysis is to approximate the properties of regular PCA keeping the number of non-zero loadings small (Sjostrand et al., 2006). The SPCA imposes extra constraints or penalty terms to the standard PCA to achieve sparsity and non-zero values reflect the high variance of the standard methods. We propose that SPCA explains a large part of the total variance of the gene expression levels of the wheat microarray data and lead to identification of the subset of genes representing the principal component which are considered important.

2. Method

2.1 Datasets: Biological samples

To generate the microarray data, RNA was isolated from two Winoka winter wheat mutant lines differing in freeze survival (Sutton et al. 2009). These lines designated FR and FS displayed freeze survival of 75% and 30% respectively. Treatments consisted of six replicates of four plants/pot for each line grown in the green house at 22-28 o C with 14 hours photoperiod. At the fourth leaf stage, three replicates of each line representing non-treated control plants (12 plants/line) were removed from the green house, the crown tissue, the most freeze resistant part of the plant, was excised and frozen in liquid N2 for later RNA isolation. The remaining six replicates of 12 plants/line were transferred to a 4 o C cold room for 4 wk to achieve cold acclimation (CA) as described (Kenefick et al., 2002). At the end of cold acclimation the crown tissue was harvested at 4 o C and frozen in liquid N2 for later RNA isolation. The classic CsCl gradient RNA isolation procedure of Chirgwin et al., (1979) was used to prepare total RNA from the frozen crown tissue. Resulting RNA was treated with DNase to remove any possible contaminating genomic DNA.

2.2 Microarrays

Eight Affymetrix wheat microarrays (CEL files) data were used in the analyses. Each array is composed of 61,127 (Additional file 1) probe sets representing 55,052 transcripts for all 42 chromosomes in the wheat genome (affymetrix.com). cDNA synthesis and hybridizations with the microarrays were performed at UC Riverside genomics facility (http://genomics.ucr.edu/facility/genomics/instruments/affymetrix.html).

2.3 Data analysis

The overall approach for the microarray data analyses is presented in Figure 1. All of the analysis was performed based on the Affymetrix GenChip Manual (Affymetrix Inc) using the statistical program R 2.12.0 (http://www.R-project.org) with affy, RMA Bioconductor packages (Irizarry et al., 2003) for the pre-processing. We used the model developed by the Bolstad et al., 2004 to measure the expression which is formulated as;

Expression values = $y_{ij} = log_2(S(N\left(B\left(PM_{ij}\right)\right))$ (1)

Background subtraction (B), Normalization (N) to facilitate between-array comparison

Summarization of 11-20 probe pair (PM/MM) intensities to one probe set value (S)

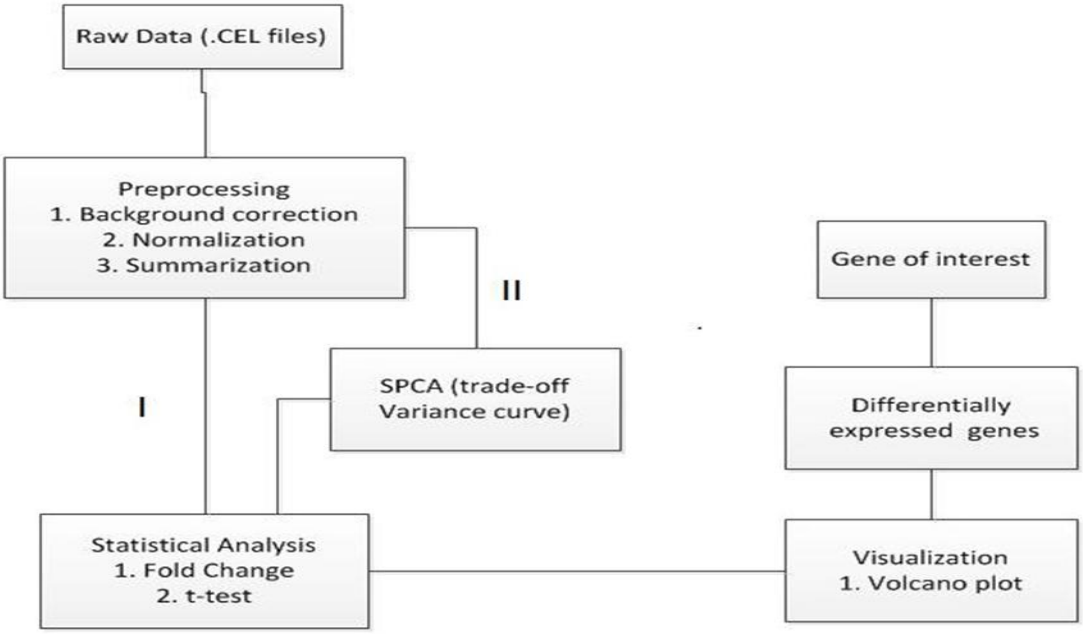

Figure 1: Work flow for a microarray data analysis

Statistical methods for identifying differentially expressed genes

After normalization of the microarray data, we compared the results (I) Without SPCA; only fold change then test statistics were applied (II) With SPCA; SPCA technique for dimensionality reduction was applied followed by fold change and test statistics.

2.3.1 Fold Change

One of the simplest methods for identifying differentially expressed gene is to evaluate the log ratio between two conditions (Churchill et al., 2003). In our dataset log transformation was performed in the data pre-processing then two replications were averaged and calculated difference between CA and Untreated (UT) in FR and similarly in FS, and results are taken to be differentially expressed if the expression under one condition is one-fold greater or less than that under the other condition.

$$FC = \log\left(\frac{\bar{X}}{\bar{Y}}\right) \qquad (2)$$

Where,

FC = fold change

$\bar{x}_i = i^{th}$ genes of CA treatment
$\bar{y}_i = i^{th}$ genes of UT

i= 1, 2,…, n

If CA = UT, effects of fold change is 0, If CA = 2UT then effect of fold change is 1.
2-fold up-regulated = log ratio of +1, 2-fold down-regulated genes = log ratio of -1
Non-differentially expressed genes = log ratio of 0.

2.3.2 Student t-test

Gene expression data are usually given in terms of the base-2 logarithm of the expression ratio, defined as the expression level of a gene relative to its level in some control condition (deHoon et al., 2002). We used two samples t-test with unequal variances to detect differentially expressed genes:

$$t = \frac{\overline{x}_A - \overline{x}_B}{s} \tag{3}$$

$$\overline{x}_A = \frac{\sum_{k=1}^{n_A} x_k}{n_A} \tag{4}$$

$$\overline{x}_B = \frac{\sum_{k=1}^{n_B} x_k}{n_B} \tag{5}$$

$$S = \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}} \tag{6}$$

$$S_A = \sqrt{\frac{\sum_{k=1}^{n_A} (x_k - \overline{x}_A)^2}{n_A - 1}} \tag{7}$$

The data with very small variance due to its low expression level contributes to large absolute t-values regardless of the mean difference between two conditions, and thus these genes can be selected as differentially expressed genes although they are not truly differentially expressed. To overcome this problem of the traditional t-test, we applied false discovery rate (FDR) proposed by Benjamini and Hochberg, 1995 to estimate significance levels as well as to control family- wise Type I error rates. FDR = FP / (FP + TP). FP = false positive, TP = true positive.

2.3.3 Sparse Principal Component Analysis

SPCA can be described as an extension of PCA, where a constraint of the number of nonzero loadings is added. SPCA method is used to reduce the number of nonzero coefficients from high-dimension data in this research. Since interpretation depends on comparing the relative sizes of the loading vectors, the sparse loadings in SPCA are much easier to interpret than PCA described by Zou et al., 2004.

The regression methods used in the calculation of SPCA all originate from ordinary least squares (OLS) approximations. The independent variable Y is approximated by a linear combination of the dependent variables in X (Sjostrand et al., 2007). The coefficients for each variable (column) of X contained in β

$$\beta_{OLS} = \arg\min_{\beta} ||Y - X\beta||^2 \qquad (8)$$

Where, ||.|| represents the l_2 norm. This is the best linear unbiased estimator given a number of assumptions, such as independent and identically distributed residuals. However, if some bias is allowed, estimators can be found with lower mean square error than OLS when tested on an unseen set of observations (Sjostrand et al., 2007). A common way of implementing this is by introducing some constraint on the coefficients in β. The method described here use constraints on either the l_1 norm or the l_2 norm of β, or both, adding the l_2 constraint give

$$\beta_{ridge} = \arg\min_{\beta} ||Y - X\beta||^2 + \gamma||\beta||^2 \qquad (9)$$

This is known as ridge regression (Hoerl et al. 1970). The theme of this theorem is to show that the connection between PCA and a regression method is possible. More about ridge regression was discussed in Hoerl et al., 1970, Zou et al., 2004, and Sijostrand et al., 2007 where sufficiently large values of γ will shrink the coefficients of β. The shrinkage introduces bias but lowers the variance of the estimates. After normalization, the coefficients are independent of the parameter γ, so, ridge penalty does not penalize the regression coefficients but ensure the reconstruction of PCs. Lasso penalty is added to the problem to penalize for the absolute values of coefficients to make the coefficients vector sparse. Replacing the l_2 norm in the constraint with the l_1 norm gives

$$\beta_{LASSO} = \arg\min_{\beta} ||Y - X\beta||^2 + \gamma_1||\beta||_1 \qquad (10)$$

Where, $||\beta||_{1=} \sum_{i=1}^{p} |\beta|$. This is the LASSO method by Tibshirani (1996). Using the l_1 norm not only shrinks the coefficients, but also drives the one by one to exactly zero as γ increases. This implements a form of variable selection, as minor coefficients will be set to zero in a controllable fashion, while the remaining coefficients will be used to minimize the size of the regression residuals.

Another possibility is to use a combination of the constraints from ridge regression and the LASSO. This approach is known as the elastic net (Zou et al., 2004) and has the form

$$\beta_{EN} = \arg\min_{\beta} ||Y - X\beta||^2 + \gamma||\beta||^2 + \gamma_1||\beta||_1 \qquad (11)$$

The main benefit of the elastic net is that it better handles cases where p > n (Sjostrand et al., 2006).

To the remaining data sets, SPCA was applied to reduce the number of nonzero coefficients. The equation by Zou et al., (2004) was applied to obtain sparse loading; Suppose we are considering the first k-principal components. Let α and βbe p x k matrices. Let x_i be the i^{th} row of the data matrix x. For any γ > 0, let

$$\left(\hat{\alpha}, \hat{\beta}\right) = \arg\min_{\beta} \sum_{i=1}^{n} |X_i - \alpha\beta^T X_i|^2 + \gamma \sum_{j=1}^{k} |\beta_j|^2 + \sum_{j=1}^{k} \gamma_1 |\beta_j|_1 \qquad (12)$$

$\sum_{j=1}^{k} \gamma_1 |\beta_j|_1$ is a lasso penalty.

Subject to $\alpha^T \alpha = I_k$

Where, γ is used for all k components, different γ_{1j} are allowed for penalizing the loadings of different principal components. If p > n, a positive is required in order to get exact PCA when the sparsity constraint (the lasso penalty) ($\gamma_{1j} = 0$) vanishes (Zou et al., 2004).

3. Results

3.1 Pre-processing and presence calls selection

The shape of the box plot for all the arrays looked similar, thus we concluded that there were less systematic biases in the data. Log transformation was used to remove the biases that were present.

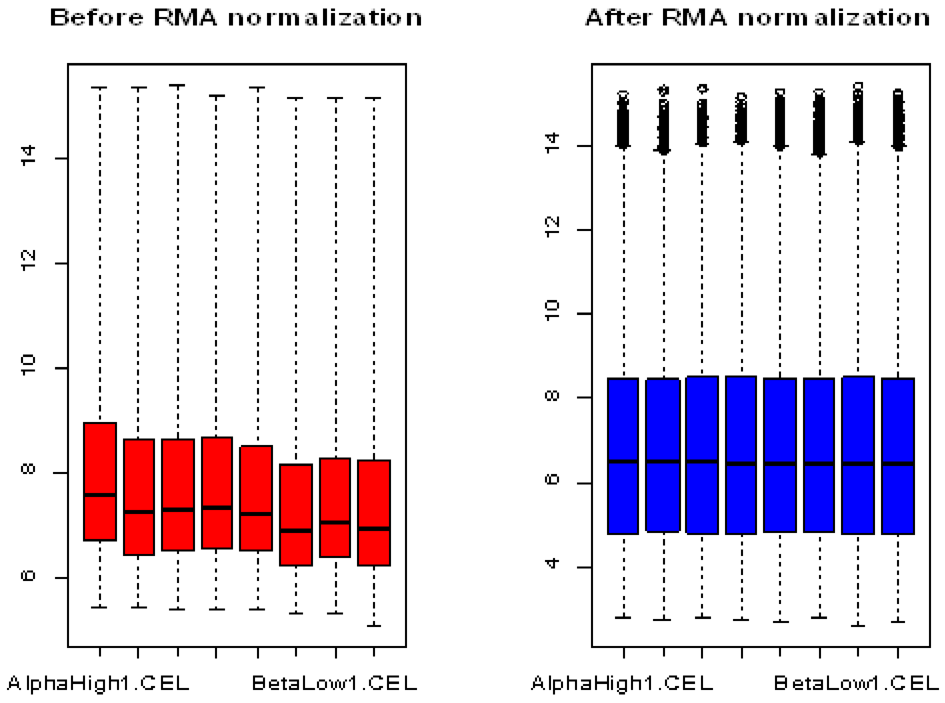

Figure 2: Box plot representation of chip-wise PM log intensity distributions. (a) Raw data before normalization. Chips 1 and 2, Chips 5 and 6 deviate strongly. (b) After RMA normalization, all eight intensity distributions appear similar. Because of the inconsistent distributions before normalization, it is recommended to carefully investigate the impact of the deviate chips on further analysis steps.

The box plot of raw intensities of the data across the eight chips is depicted in Figure 2A. The raw intensities differed between Chip 1 and Chip 2, and Chip 5 and Chip 6. The results of using RMA normalization as described in methods is depicted in Figure 2B. All intensities after transformation were distributed similarly.

The overall result of the microarray data is reflected in Figure 3 which is explained in detail in the following topics:

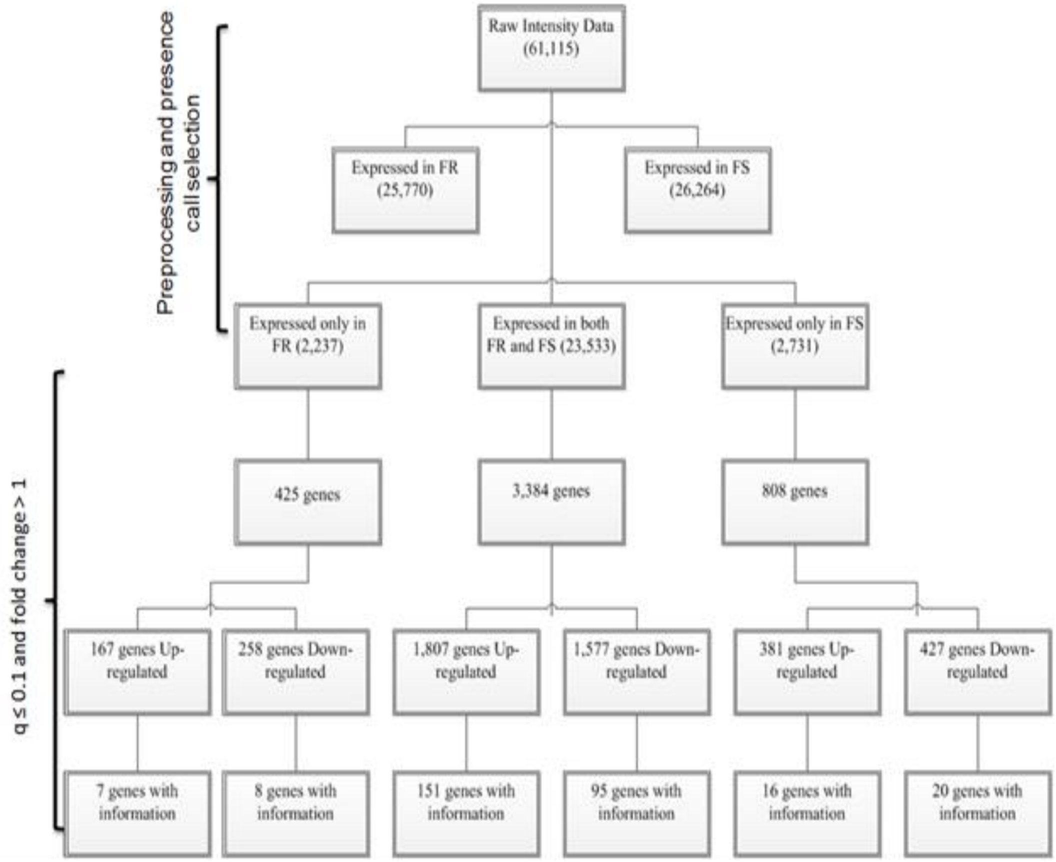
Preprocessing and presence call selection
q ≤ 0.1 and fold change > 1
Raw Intensity Data (61,115)
Expressed in FR (25,770)
Expressed in FS (26,264)
Expressed only in FR (2,237)
Expressed in both FR and FS (23,533)
Expressed only in FS (2,731)
425 genes
3,384 genes
808 genes
167 genes Up-regulated
258 genes Down-regulated
1,807 genes Up-regulated
1,577 genes Down-regulated
381 genes Up-regulated
427 genes Down-regulated
7 genes with information
8 genes with information
151 genes with information
95 genes with information
16 genes with information
20 genes with information

Figure 3: Overall results of microarray data analysis before SPCA

3.2 Fold change determination

Genes for which the fold change in mean expression levels of each group by the symbol (+ and -) were reported as up-regulated or down-regulated (see eq.2). After accomplishing the background correction, quantile normalization, and median polish summarization, 25,770 and 26,264 expressed genes remained for FR and FS lines respectively. Of these, 2,237 responded to cold acclimation only in FR, 23,533 in both FR and FS, and 2,731 in only FS. In FR, there were 1021 up-regulated and 1216 down-regulated genes. Of the genes expressed in both FR and FS, 11,241 were up-regulated, 12,292 were down-regulated. In FS, there were 1317 up-regulated and 1414 down-regulated genes.

3.3 Differentially expressed genes (DEGs)

Using Test- statistics such as The Benjamin and Hochberg False Discovery Rate (FDR) method for multiple testing corrections with a FDR of significance level at α = 0.1, significantly DEGs were characterized as q ⩽ 0.1 and fold change level greater than one. The 1021 up - and 1216 down-regulated genes expressed only in FR were reduced to 167 and 258. Similarly, the 1317 up- and 1414 down-regulated genes expressed only in FS were reduced to 381 and 427. FDR and fold change reduced the 11,241 up - and 12,292 down - regulated genes in both FR and FS to 1,807 and 1, 577 respectively.

Of the DEGs reported in FR, FS, and both FR and FS, only 7 up-and 8 down-regulated genes with information (gene symbol, gene title, and go biological function) were identified in FR, 16 up - and 20 down-regulated genes were identified in FS. Similarly, 151 up - and 95 down-regulated genes in both FR and FS were identified.

3.4 SPCA for dimension reduction

Our analysis of principal components (PC) indicated that the first PC accounts for ~90% of the information present in the entire data set. Trade off curves between explained variances and cardinality in both (FR and FS), FR, and FS is explained in Figure 4 and Table 1. It was observed that γ gradually changes the SPCA algorithm in FR and FS. The variance of the SPC increased with the number of non-zero loadings, but after γ = 0.05, the growth flattened markedly in common (expressed in both FR and FS) set. Therefore, we chose γ = 0.05for genes expressed in both FR and FS and γ = 0.07 for genes expressed only in FR and only in FS as the minimum number of non-zero loadings for which adding more variables does not give a significant contribution.

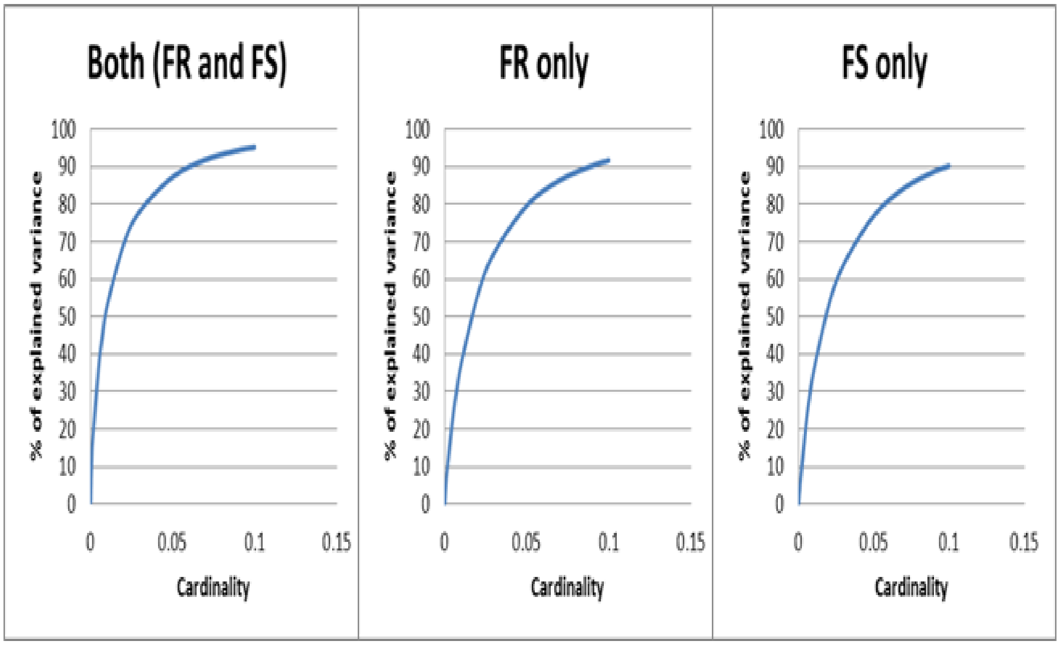

Figure 4: Trade-off curves between explained variance and cardinality. The vertical axis is the Percentage explained variance and horizontal axis is the sparsity. A) In both (FR and FS), B) in FR only, and C) in FS only.

Table 1: Explained variance as a function of lambda both in FR and FS, FR only, and FS only.

γ	Both FR and FS var. (%)	No. of non - zero loadings	FR var. (%)	No. of non - zero loadings	FS var. (%)	No. of non - zero loadings
0	0.6	8	0.3	4	0.3	4
0.00001	1	14	0.3	6	0.3	4
0.00005	2.7	30	0.7	13	0.5	10
0.0001	3.8	45	1.2	25	0.9	17
0.0003	7.6	95	2.9	53	2.2	44
0.0005	10.3	123	4.5	79	3.2	66
0.0007	12.7	158	5.7	100	4.2	107
0.001	15.9	199	7.4	135	5.9	150
0.005	37.5	498	23.8	520	22.1	623
0.007	44.6	605	29.8	682	28.2	798
0.01	52.9	712	37.6	894	35.7	1030
0.02	69.3	964	55.8	1337	53.2	1552
0.03	78.2	1088	67	1620	64.1	1861
0.05	87.3	1211	79.7	1908	77	2267
0.07	91.9	1278	86.4	2057	84.3	2481
0.09	94.4	1298	90.3	2154	88.7	2594
0.1	95.2	1305	91.7	2175	90.2	2635

The overall result using SPCA is depicted in Fig. 5. This treatment resulted in a reduction of total genes from 2,237 to 2,057 in FR, 1321 to 1,211 in both (FR and FS), and 2,731to 2,481 in FS. We further analyzed the resulting genes and reported 415(163 up-and 252 down-regulated), 226 (142 and 84 genes as up - regulated and down - regulated), and 776 (372 up-and 404 down-regulated) significantly DEGs in FR, both, and FS respectively.

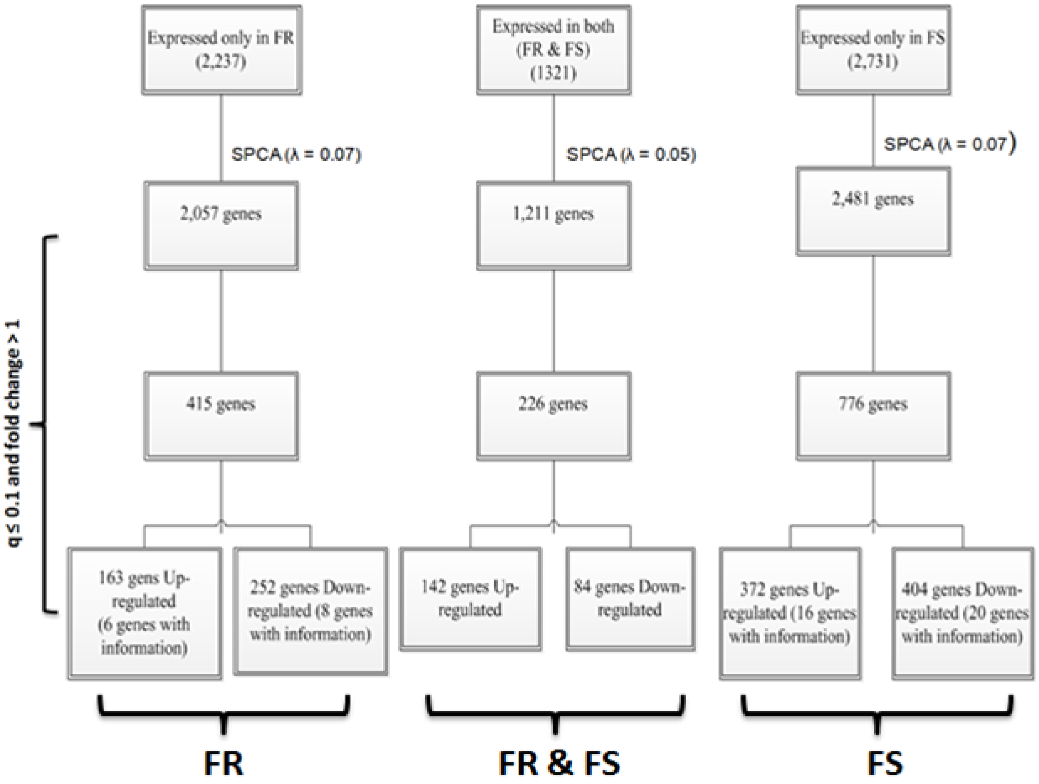
Expressed only in FR (2,237)
Expressed in both (FR & FS) (1321)
Expressed only in FS (2,731)
SPCA (λ = 0.07)
SPCA (λ = 0.05)
SPCA (λ = 0.07)
2,057 genes
1,211 genes
2,481 genes
q ≤ 0.1 and fold change > 1
415 genes
226 genes
776 genes
163 gens Up-regulated (6 genes with information)
252 genes Down-regulated (8 genes with information)
142 genes Up-regulated
84 genes Down-regulated
372 genes Up-regulated (16 genes with information)
404 genes Down-regulated (20 genes with information)
FR
FR & FS
FS

Figure 5: Overall results of microarray data analysis after SPCA

3.5 Visualization of gene expression changes

Volcano plot

The 'volcano plot' is a scatter plot of the negative log_{10} -transformed p-values from the gene specific t-test against the log_2 fold change (see Figure 6). Genes with statistically significant differential expression according to the gene-specific t-test will lie above a horizontal threshold line and genes with large fold-change values will lie outside a pair of vertical threshold lines (Churchill et al., 2003). Of the 1321 genes expressed in both FR and FS, the volcano plot algorithms with cut-off at $p \leq 0.01$ and fold change greater than 2, identified 187 and 217 genes as significantly regulated by cold acclimation in FR and FS respectively. Of the 1211 expressed genes obtained after SPCA treatment, 176 and 207 genes were identified as significantly expressed in FR and FS.

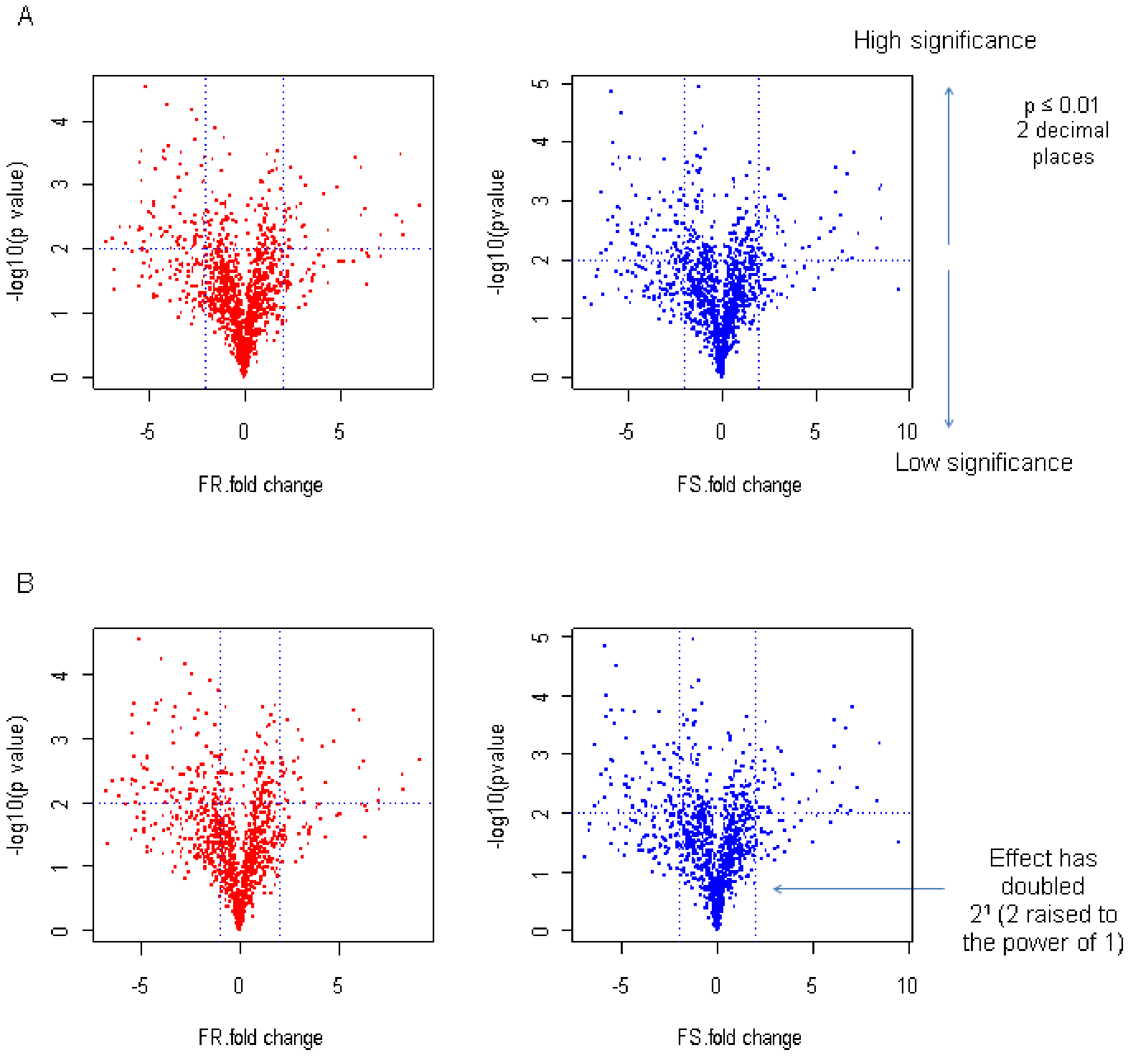

Figure 6: Volcano Plot. The negative log10-transformed p-values are plotted against (A) the log ratios (log2 fold change) in FR and FS with response to cold before SPCA and (B) after SPCA

4. Conclusions

This paper focuses on the implementation of the sparse principal component analysis in microarray data analysis. We applied SPCA algorithm developed by Zou et al., 2004 that efficiently deals with p >> n data. However, this research represents the result from the wheat microarray data analysis we compared SPCA only a set of data that have information (GO terms, gene titles, and gene symbols). The dimension reduction from 1321 to 1211 using SPCA is significant in the small data set that resulted to 20 significantly DEGs differences between with SPCA and without SPCA. In this work we aimed to filter correlated genes to identify significantly DEGs in response to cold. In future studies we will verify significantly DEGs using RT-PCR and cluster them. We will also attempt to analyze sequences and find regulatory elements of selected genes.

Acknowledgements

The South Dakota State University Agriculture Experimental Station provided the funds used to purchase the Affymetrix wheat microarrays and cover the contract to the University of CA Riverside Genome center to perform the hybridizations and scanning of the microarrays and Sub-Award to F. Sutton from UMN NSF 05-603 to J. Cohen

We also acknowledge Dr. Jun Ye, Mathematics and Statistics Department, SDSU for introducing SPCA techniques in the class and encouragement to apply in this research.

References

Nikulin, V. and G. J. McLachlan (2009). "Penalized Principal Component Analysis of Microarray Data." proceedings of the 6th international conference on Computational intelligence methods for bioinformatics and biostatistics.

He, Y., M. R.D.C., et al. (2011). "An algorithm for sparse PCA based on a new sparsity control criterion." Proceedings of the SIAM International Conference on Data Mining, Mesa, AZ, April 28-30, 2011.

Zou, H., T. Hastie, et al. (2004). "Sparse Principal Components Analysis." technical report, Department of Statistics at Stanford University.

Sjostrand, K., M. B. Stegmann, et al. (2006). "Sparse Principal Component Analysis in Medical Shape Modeling." International Symposium on Medical Imaging 2006, San Diego, CA, USA 6144.

Sutton, F., D. G. Chen, et al. (2009). "Cbf genes of the Fr-A2 allele are differentially regulated between long-term cold acclimated crown tissue of freeze-resistant and – susceptible, winter wheat mutant lines." BMC Plant Biology 9:34.

Kenefick, D. G., F. Sutton, et al. (2002). "Plant water uptake by hard red winter wheat (Triticum aestivum L.) genotypes at 2_C and low light intensity." BMC Plant Biology 2:8.

Chirgwin, J., A. Przybyla, et al. (1979). "Isolation of Biologically Active Ribonucleic Acid from Sources Enriched in Ribonuclease." American Chemical Society 18(24): 5294-5299.

Irizarry, R., B. Bolstad, et al. (2003). "Summaries of Affymetrix GeneChip probe level data." Nucleic Acids Res 31(4).

Bolstad, B. M., R. A. Irizarry, et al. (2004). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics 19(2).

Churchill, G. and X. Cui (2003). "Statistical tests for differential expression in cDNA microarray experiments." Genome Biology 4:210.

deHoon, H. J. L., S.Imoto, et al. (2002). "Statistical analysis of a small set of time-ordered gene expression data using linear splines." Bioinformatics 18(11): 1477-1485.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." JRStatSocB 57(1): 289-300.

Sjostrand, K., E. Rosttrup, et al. (2007). "Sparse Decomposition and Modeling of Anatomical Shape Variation." IEEE TRANSACTIONS ON MEDICAL IMAGING 26, No.12.

Hoerl, E. and R. W. Kennard (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." Technometrics 12, No. 1: 55-67.

Tibshirani, R. (1996). "Regression Shrinkage and Selcection via the Lasso." Journal of the Royal Statistical Society. Series B (Methodological) 58(1): 267-288.