18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Validating a gene expression signature of invasive ductal carcinoma of the breast and detecting key genes using neural networks

Samarasinghe, S., D. Kulasiri

Centre for Advanced Computational Solutions, Lincoln University, New Zealand
Email: sandhya.samarasinghe@lincoln.ac.nz

Abstract: Breast cancer is one of the leading causes of death in women in the world. It is a complex disease with challenges to accurate diagnosis due to cancer subtypes that are difficult to distinguish. The most common subtype is Invasive Ductal Carcinoma (IDC), a cancer in ductal cells that line the milk ducts in the breast. In depth understanding of the genetic basis of IDC can help treat it more effectively.

Microarray based gene expression analysis is making new grounds in accurate diagnosis of diseases including cancer. Microarray experiments are designed to measure the expression levels of thousands of genes in various cells/tissues of interest and they are analysed to decipher a small set of genes that constitutes the gene signature of a particular disease. The few studies on breast cancer gene expression compare cancer subtypes but very few have compared gene expression between matched cancer and healthy tissues in the breast (Turashvili et al., 2007). The few studies that have compared different subtypes have little agreement on the gene signatures (Turashvili, 2007; Zhao et al., 2004, Sorlie, et al., 2001). Therefore, it is highly beneficial to further assess the validity of genes identified as differentially expressed, in order to boost confidence in the usefulness of the genes in various medical applications including diagnosis, prognosis and drug development. In this study, the validity of differentially expressed genes pertaining to a carefully conducted experiment on breast tissues affected by Invasive Ductal Carcinoma (IDL) and matched healthy tissues is conducted using neural networks and statistical methods. The data was obtained from NCBI database and deposited by Turashvili et al (2007) from their experiments on breast cancer. The original authors extracted a 326 gene signature for IDC using statistical methods. In our study, the ability of this gene set to discriminate the disease state from healthy state is investigated and validated using two sets of independent datasets.

Our visual and qualitative exploration using Self organizing maps (SOM) followed by statistical tests indicated that the validation data supported 80% of the original gene signature. Another SOM results declared that the original gene set is able to classify patients as being healthy or having IDC. Original gene set was optimally clustered into two classes based on correlation of expression patterns of genes by SOM /Ward clustering. The two classes and genes in them were supported by 60% of the validation data. As an alternative, PCA was used to determine genes with correlated expressions in the original gene signature and 4 PCs accounted for 86% of the variation in the data with the first 2 PCs accounting for around 70%. Top most important 100 genes in PC1 and PC2 provided 52% support for the two SOM classes with PC1 dominating class 1 and PC2, class 2. Genes that were validated by independent data in the two SOM classes were used in conjunction with PC1 and PC2 to extract highly influential genes from the top 6%, 18% and 57% of the original genes represented by PC1 and PC2. These key genes may prove to be the most crucial in identifying ductal tumor from healthy tissues. Four new genes were among key genes that may shed more light onto the disease mechanism. The key genes as well as overall set of validated genes may provide further support to understand or refine genetic networks that these genes are part of in the next stage of our study.

Keywords: *Breast cancer, Invasive Ductal Carcinoma (IDC), gene expression profiles, neural networks, Self Organising Maps (SOM), Principal Component Analysis (PCA)*

1. INTRODUCTION

Breast cancer is one of the leading causes of death in women in the world. It is a complex disease with challenges to accurate diagnosis due to cancer subtypes that are difficult to distinguish. The most common subtype is Invasive Ductal Carcinoma (IDC), a cancer in ductal cells that line the milk ducts in the breast. In depth understanding of the genetic basis of IDC can help treat it more effectively.

Microarray based gene expression analysis is making new grounds in accurate diagnosis of diseases including cancer. In this study, an assessment of the efficacy of differentially expressed genes pertaining to a carefully conducted experiment on breast tissues affected by Invasive Ductal Carcinoma (IDL) is conducted using neural networks and statistical methods. Microarray experiments are designed to measure the expression levels of thousands of genes in various cells/tissues of interest and they are analysed to decipher a small set of genes that makes the gene signature of a particular disease. In the case of cancer, RNA is extracted from normal (benign) and cancer tissues and expression levels are measured separately in order to identify the genes with significant changes in expression between normal and cancer tissues. Few studies have already used microarray data to extract highly expressed genes in breast cancer subtypes (Turashvili et al., 2007; Zhao et al., 2004, Sorlie et al. 2001). However, agreement between these studies in terms of the genes identified is minimum. There are several reasons for this: the major one is the small sample size and the large number of genes, and another is the non-standardized analysis of gene expression. A recent survey of 7 widely used analysis methods highlights that results vary between the methods (Jiang et al., 2008). Most common methods include statistical tests for means that is greatly affected by the sample size. Therefore, it is highly beneficial to assess further the genes identified as differentially expressed, in order to boost confidence in the usefulness of the genes in various medical applications such as diagnosis and prognosis. This study involves an extensive further investigation of the differentially expressed genes in IDC obtained from statistical analysis of microarray data by Turashvili et al. (2007).

2. BREAST CANCER MICROARRAY DATA

The data for the study was obtained from NCBI database and deposited by Turashvili et al (2007) from their experiments on 5 IDC and 5 ILC patients (referred to as 'original data' in our paper). ILC refers to invasive lobular carcinoma, a cancer in the lobules where milk is secreted in the breast. ILC is the second most common subtype after IDC. From each patient, normal ductal and normal lobular cells were also extracted in addition to their own particular cancer (IDC or ILC) cells. Thus there were 3 microarrays per patient and altogether 30 microarrays. Expression levels were measured on HGU133Plus 2.0 Affymetrix arrays consisting of 54,675 genes on the chip. The original authors presented results from a comparison between healthy and tumor ductal tissues using GCOS (Affymetrix) pairwise comparison and rank product (RP). The 326 genes that are common to the two methods have been declared as the gene signature distinguishing between healthy and ductal tumor tissues. In our study, the ability of this gene set to discriminate the disease state from healthy state is investigated and validated using two sets of independent data: one set is a totally independent dataset from NCBI database and contain gene expression levels for 7 healthy ductal and 7 matched ductal tumor tissues measured on HGU133Plus 2.0 Affymetrix arrays (NCBI, 2008). The other validation set came from the healthy ductal data from the 5 ILC patients in the original study.

3. OBJECTIVES

The goal of this study was to further scrutinise the discriminatory power of a set of differentially expressed genes to distinguish between healthy and ductal cancer (IDC) tissues. The task is subdivided into: (i) study the expression pattern of the selected gene signature and verify with independent data; (ii) use self-organising maps (SOM) to cluster patients (tumor ductal and normal ductal) based on the expression of the gene set and validate using independent data; (iii) cluster correlated genes using self organising maps (SOM) yielding possible sets of genes with similar function and/or belonging to same signaling pathway; and (iv) identify key genes that carry significant information using Principal Component Analysis (PCA) and SOM.

4. GENE EXPRESSION ANALYSIS AND MODEL RESULTS

4.1 Exploration and validation of differentially expressed genes between healthy and ductal tumor

The original authors, Turashvili et al. (2007), reported 326 genes differentially expressed between tumor ductal and normal ductal cells. This gene set was obtained by comparing gene expression in tumor and

and normal ductal cells obtained from the 5 ductal tumor patients. Gene expression is a measure of the concentration of mRNA produced by a gene in a tissue. Log ratio of tumor to healthy gene expression (i.e., logarithm of the ratio of expression of a gene in a tumor tissue to the expression of the same gene in the corresponding healthy tissue is used in selecting differentially expressed genes. The above 326 genes displayed more than two-fold up or down regulation (i.e. log ratio greater than 1 or less than 0.707). The log ratio of 326 genes for the 5 patients is shown in Fig. 1(a) and the mean log ratio and 95% confidence interval are presented in Fig. 1(b) where the 326 genes are ordered from highly down-regulated to highly up-regulated along the horizontal axis (from left to right) in Fig. 1. Thus the genes above the horizontal axis are up-regulated and those below it are down-regulated. One colour line represents expression of the 326 genes in one patient. The left figure shows reasonable variation among patients; yet they agree broadly on the up-regulated and down-regulated genes. The right figure shows that for some genes, confidence interval contains zero and they may not be differentially expressed.

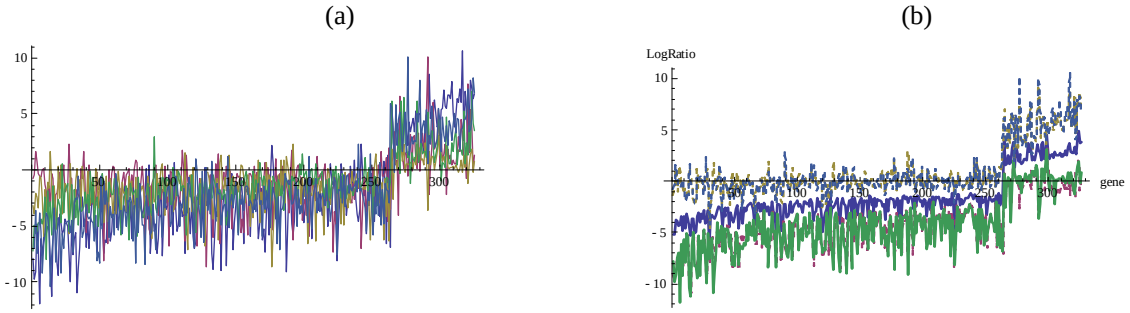


Figure 1. Log ratio of tumor ductal to healthy ductal gene expression for the 326 differentially expressed genes: (a) expression patterns for the 5 ductal patients, and (b) mean log ratio and 95% Confidence Interval for the data in Fig. 1(a).

In order to assess the validity of patterns in Figure 1, the expression levels for the same 326 gene set in the independent microarray data obtained for the 7 IDC and corresponding healthy ductal tissues were used. The log ratio for the 7 patients is shown in Figure 2(a) which indicates better agreement with the highly upregulated genes in Fig.1(a) than with the down regulated genes. Fig. 2(b) presents the mean log ratio from the two datasets: blue representing the original data and pink, the validation data. Again, there is closer agreement between most of the upregulated genes and discrepancy between some of the down regulated genes.

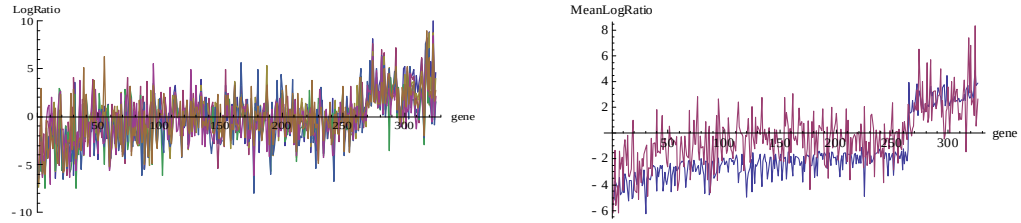


Figure 2. Log ratio for 7 independent IDC patients: (a) individual patient data; (b) mean log ratio for the 7 independent IDC patients (pink colour line) and mean log ratio for the original 5 IDC patients (blue line).

A hypothesis that the means are different for the original and validation data shown in Fig 2(b) was tested using two sample t-tests (two-tailed) and 61 (19%) genes failed the hypothesis test at 0.05 significance level as shown in Fig.3 where solid line indicates the critical p value (Mathematica, 2008). Correlation between original and validation mean log ratio was 0.63 indicating a reasonably strong agreement. After removing the genes that failed the hypothesis, correlation increased to 0.75.

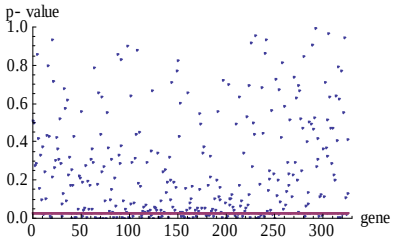


Figure 3. p-values from hypothesis testing for equality of means between original and validation data for each gene. Solid line indicates the cut-off point of 0.025 for a two-sided test with significance level of 0.05.

The 326 gene set was further analysed using SOM to assess in a more meaningful way how they feature in individual patients. SOM, a powerful nonlinear clustering method, projects multidimensional data onto a

two-dimensional grid of neurons in such a way that input vectors that are in close proximity in the data space are projected onto neurons that are in close proximity on the grid. This allows a realistic view of data and their cluster characteristics and relations. This is achieved by incremental adjustment of neuron weight vectors each representing a cluster of input vectors using a distance metric such as Euclidean,

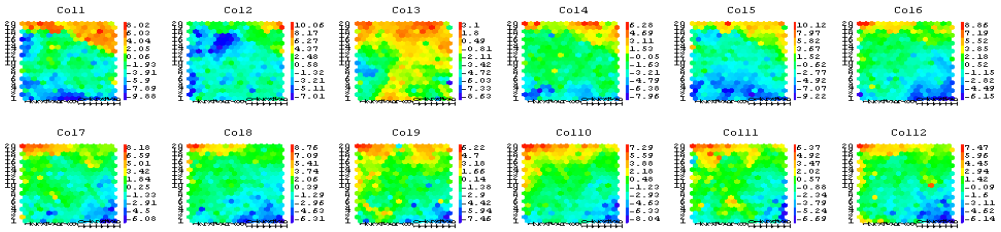
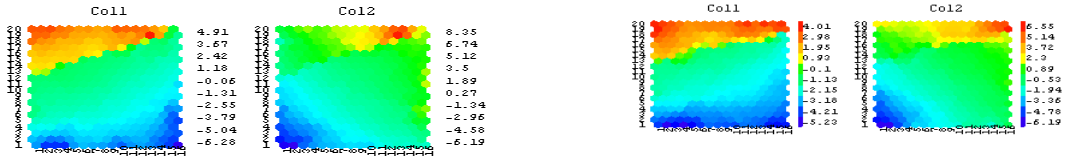


Figure 4. Log ratio for original and validation data visualized on an SOM: Col 1 to Col 5 refer to gene expression in original patients and Col 6 to Col 12 represent the expression in patients in validation data.

correlation distance etc. For example, Euclidian distance brings closer genes with similar expression levels and correlation distance brings closer genes with similar trends (rate) of gene expression. Starting with initial random neuron weight vectors, closest matching neuron to each input vector is found and its weight vector is brought closer to that input vector. Weights of its neighbour neurons are also brought a little closer to the same input according to a neighbourhood function (Gaussian, exponential etc.) to preserve proximity of original data on the map. Training continues until a stopping criterion is met. In this study, the genes were clustered on a 20x16 neuron map using a Gaussian neighbourhood function and Euclidian distance to bring closer genes with similar expression levels (Machine learning Framework, 2007). Here, log ratios from the original 5 patients and 7 patients in the independent validation set were combined so that gene expression between the two data sets can be compared directly. There were 12 log ratios in each input vector with the first 5 corresponding to original 5 patients and the remaining 7 corresponding to the 7 patients in the validation set. The expression patterns of the 326 genes for these 12 patients are shown in the maplets of the trained SOM in Figure 4. Here, red colour indicates high up regulation and blue indicates high down regulation. Each maplet shows distribution of gene expression levels in one patient and corresponding points in each maplet refer to the same gene or group of genes so by following the maplets across, similarity or otherwise of gene expression between the patients can be visually ascertained. The figure shows that patients 1, 4 and 5 (maplets indicated by Col,1,4,5) in the original data have very similar expression patterns. Patient 3 (Col.3) is an anomaly with much lower log ratios for the highly upregulated genes. Original patient 2 (Col 2) has fewer up regulated genes and the original authors also found from hierarchical clustering that patient 2 did not cluster together with the other 4 patients but was grouped with healthy data. This is because it is more similar to healthy tissues than to tumor tissues but whether it is a possible outcome for a tumor tissue or an experimental error is not clear. Overall, there is agreement between the two datasets; however, the independent set (Col 6 to Col 12) shows more consistent patterns across the 7 patients.

Another view of the data was taken by using mean log ratio distribution before (Figure 5(a)) and after removing the 61 genes that did not pass the t-test (Figure 5(b)). In both Figures 5(a) and 5(b), left maplet is the mean log ratio for the 5 patients in original data and the right maplet is that for the 7 patients in the



(a) (b)

Figure 5. Mean log ratio for original and validation data: (a) all 326 genes and (b) after removing incompatible genes

validation data. In Figure 5(a) all 326 genes are used whereas in Figure 5(b), 265 genes are used after removing the 61 genes. They show that there is better agreement between up regulated genes (red colour) than the down regulated genes (blue colour) in both cases but this is further improved after removing the 65 suspicious genes.

4.2 SOM clustering of healthy and tumor patients using the gene expression signature

In the next stage of analysis, another SOM (4x4) was trained to cluster original patients based on their expression pattern of the 326 genes. The results are shown in Figure 6(a) where labels indicate the type of tissue (nd indicate normal (healthy) ductal, td indicate tumor ductal, and numbers 1 to 5 refer to the patient number). Each neuron in the trained map represents the center of gravity of the gene expression vectors that fall in its vicinity. These neurons were further clustered using Ward clustering (Mlf, 2007) that displayed two optimum clusters indicated by red (healthy) and blue (tumor) on the map. In Figure 6(a), healthy and tumor tissues form two clearly separate clusters. Only tumor patient 2 (td2) has been wrongly clustered with normal patient nd3 and this was also reported by the original authors from hierarchical clustering. All the other normal samples have been clustered in close proximity. Tumor patients 1, 4 and 5 are in close proximity and tumor patient 3 (td3) has been placed far from them. These results are consistent with the patterns shown in Figure 4 (Col 1 to 5).

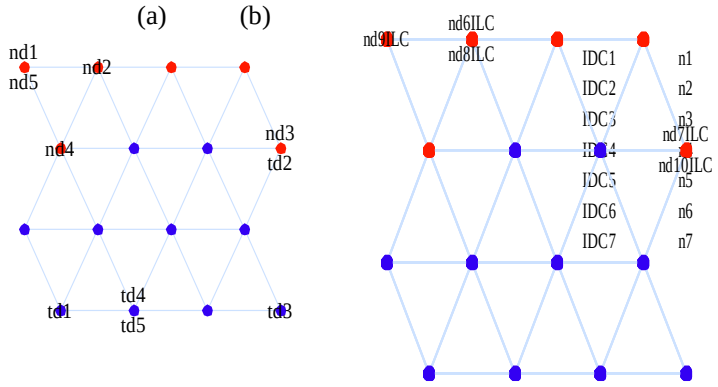


Figure 6. Clustering of healthy and cancer patients based on gene expression profile of 326 genes: (a) SOM map generated from the original 5 normal ductal (labeled nd1 to nd5) and 5 tumor ductal patients (labeled td1 to td5), (b) validation of map developed in (a) by projecting 3 sets of data: 5 normal ductal samples (nd6ILC to nd10ILC) extracted from the 5 Invasive lobular carcinoma (ILC) patients in the original data; normal ductal (n1 to n7) and ductal cancer (IDC1 to IDC7) samples in the independent validation data.

To verify the expression patterns (patients) clustered by the map, three independent data sets were used. The first is the healthy ductal data obtained from the 5 lobular cancer patients. (Note that normal ductal cells were extracted from all 5 invasive lobular carcinoma (ILC) patients). These samples are labeled nd6ILC to nd10ILC. The gene expression patterns of the 326 genes were extracted from the data for these 5 healthy ductal tissues and projected onto the trained SOM as shown in Figure 6(b). There is 100% accuracy of prediction for these patients. The second verification came from the independent validation dataset for the 7 ductal and matched healthy samples already described previously. Healthy samples are labeled n1 to n7 and tumor samples are labeled IDC1 to IDC7. The expression levels for the 326 genes were extracted from this data and were projected onto the map giving the results shown in Figure 6(b). It shows 100% classification accuracy for both normal and tumor samples. It also confirms the consistency of normal and tumor data in this dataset. As the map has classified all the new patterns accurately, the 326 gene signature identified by the authors appears to be robust.

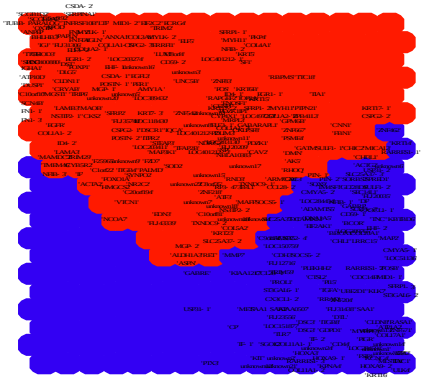
A new SOM was trained with the reduced set of 265 genes using the original 5 IDC and corresponding normal healthy ductal data. Again, td2 was clustered with healthy data. Validation with healthy ductal data from the 5 ILC patients produced perfect results as were the healthy ductal data from the independent ductal dataset. However, 4 out the 7 ductal tumor samples (IDC1 to IDC3 and IDC6) were wrongly classified as healthy; but they were all projected to a neuron closest to the 'tumor representing neurons' indicating that this set may prove to be useful in future as more patient data become available.

4.3 Assessing correlated gene clusters using SOM

Gene expression normally involves co-regulation of related genes working together in genetic regulatory networks. Furthermore, there may be several networks acting independently or in concert with other networks to produce the expression levels measured in the experiments. Therefore, assessing the relationships between genes can provide vital clues to the mechanism of operation of gene regulatory networks. As microarrays provide a snapshot of action of a possibly large number of networks in an instance of time, correlated gene expression may indicate possible co-regulated gene expression. In this section, SOM is used for clustering genes based on their expression across the patients. As it is expected to validate the results using the independent dataset, normalized mean expression for the 5 healthy and 5 tumor tissues is used and correlation distance metric is used that brings closer genes that have similar rates (positive or negative) of change of expression. Neighbourhood function was Gaussian. Figure 7 (a) shows the results where the names of the genes are projected on the map. In order to find the cluster structure, Ward clustering was applied to the trained map neurons and the results in Figure 7(b) indicate strong support for the 2

optimum clusters followed by 8 and 11 clusters. In this study, we present the results for the 2 clusters depicted in red and blue in Figure 7(a). In this figure, genes that are closer have similar rates of change of expression levels. Class 1 (blue) had 123 genes and Class 2 (red) had 203 genes (Table 1).

The map was validated by projecting independent data for the 7 patients. 102 genes fell in Class 1 and 224 in Class 2. However, the number of genes common to validation and original data in the 2 classes were 49 (40%) and 146 (72%), respectively, based on the number of genes from the original dataset falling in the two classes (Table 1). Overall, there is support for 195 (60%) out of the original 326 genes in the 2 classes. We will refer to the rest of the columns in Table 1 in the next section.



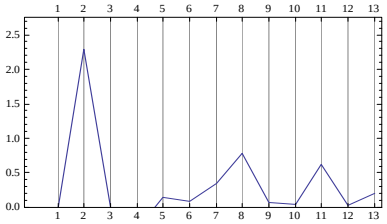


Figure 7: (a) (left) trained map with genes projected onto it; (b) (right) Results from Ward clustering indicating the number of clusters (x axis) and Ward index (y-axis).(2 clusters are shown to be the optimum)

Table 1. Clustering and validation results for the SOM

Class	No of genes in the original dataset	No of genes in the validation dataset	Number (and %) common genes between original and validation data	No.of Top 100 PC1 genes matching original data (and valid. data)	No.of Top 100 PC2 genes matching orig data (and valid. data)	Total unique PC1 and PC2 genes matching orig. data (and %)
1	123	102	49 (40%)	52 (54)	45 (30)	80 (65%)
2	203	224	146 (72%)	46 (44)	55 (70)	89 (44%)

4.4 Analysing correlated genes and highly influential genes using PCA

For diagnosis of tumor, it is highly useful if a small number of key genes that help detect cancer from healthy tissues can be found. As stated earlier, knowledge of corregulated genes linked to the disease are also of importance for understanding gene networks. To look into these 2 aspects from a different perspective, principal component analysis (PCA) was conducted. Here, expression data from tumor ductal and normal ductal tissues of the original 5 IDC patients were combined with the normal ductal expression in original 5 ILC patients to enlarge the database for PCA. (There was a slight improvement in the results from the combined data compared to those based on the 5 IDC patients alone). PCA is a linear method where correlated inputs are combined into independent

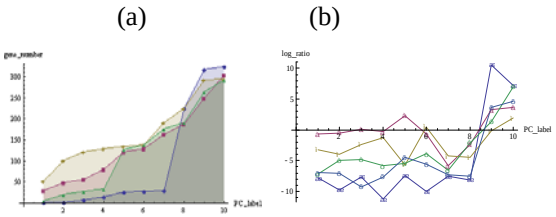


Figure 8. (a) The top ten genes represented in the first 4 PCs (from bottom to top lines: PC1, PC4, PC2, PC3); (b) expression patterns of the top 10 genes in PC1 (These 10 genes are : KRT14, KRT17, KRT5, LOC118430, SCGB1D2, TFAP2B, TFPI2, KRT16, TOP2A, and COL11A1);

principal components. For this data, the top 4 principal components (PCs) captured 86% of variance in the data (PC1: 59%, PC2: 11%, PC3: 10%, PC4: 6%) with PC1 roughly contributing 60% and PC1 and PC2 together contributing 70% to the coverage. Figure 8 shows the labels of the top 10 most influential genes covered by the first 4 PCs. It indicates that PC1 (bottom blue line) has incorporated mainly the highly upregulated and highly down regulated genes whereas the other three PCs cover genes in the whole range. The plots for top (largest) 30 and top 100 PC coefficients, showed trends consistent with those in Figure 8(a) (eg. PC1 capturing the genes with high fold ratio etc.). The log ratios of the top 10 (most influential) genes covered by PC1 are shown in Figure 8(b) for the 5 IDC patients. It shows that 3 out of 5 patients demonstrate highly up or down regulation, one patient moderate up and down regulation and one with

relatively high up regulation. Moreover, the most influential 30 and 100 genes in PC1 were also strongly supported by 4 out of 5 patients.

Next, PCA results were viewed in conjunction with the SOM results in Figure 7. First, the two methods were compared by assessing the number of the most influential top 100 genes in PC1 and PC2 falling in the 2 SOM classes (Table 1 (columns 5,6,7)). 52 PC1 and 45 PC2 genes were in class 1 and 46 PC1 and 55 PC2 genes in class 2. In each class, there were overlapping genes between PC1 and PC2 so, altogether there were 80 unique genes in class1 and 89 unique genes in class 2. These represent 65% and 44% of the original genes in class1 and class 2, respectively. Agreement of the top 100 PC1 and PC2 genes with the validation genes falling into class 1 and 2 were also evaluated. Results shown in Table 1 (within brackets in columns 5 and 6) indicate that 54 PC1 and 30 PC2 genes agree with class 1 validation genes and 44 PC1 and 70 PC2 genes agree with class 2 genes. The above discussion highlights that SOM class 1 is dominated by PC1 genes and class 2 by PC2 genes. The final step of the comparison between SOM and PCA was conducted to pick key genes in breast cancer. This was done by assessing the number of top 10, 30 and 100 PC1 and PC2 genes in the SOM class 1 and class 2 genes validated by the independent data and the results for the top 10 and 30 PC1 and PC2 genes are shown in Table 2. It shows that one each of top 10 PC1 (KRT14) and PC2 (RARRES1) genes are in class 1 (total 2) and 1 PC1 and 3 PC2 genes are in class 2 (total 4). Unknown 9 indicate a gene on the chip presently without a name. These 6 genes are the most highly influential genes within the top 6% of the key genes. As for top 30 PC1 and PC2 genes, there were 7 in class 1 and 17 in class 2. These 24 comprise the most influential genes in the top 18% of key genes. Interestingly, there are no overlapping PC1 and PC2 genes in either top 10 or top 30 category.

Table 2. Most influential validated genes in the Top 10 and 30 list of PC1 and PC2 factor loadings

PC1 and PC2 genes	Class 1	Class 2
Top 10 genes in PC1&PC2 (6 genes)	2-genes: KRT14, RARRES1	4-genes: SCGB1D2, CAV2, unknown9, SPIN3
Top 30 genes in PC1&PC2 (24 genes)	7-genes:KRT14, SFRP1, BBOX1, DTL, GABRP, RARRES1, GDPD1	17-genes: SCGB1D2, SCGB2A2, VTCN1, MAOB, CAV2, unknown9, SPIN3, MRPL3, HMGCS2, SYNPO2, C8orf4, FZD7, LAMA3, LOC203411, NS3TP2, unknown3, ELF3

Note: black represents genes represented by PC1 and red indicates those represented by PC2.

5.0 CONCLUSIONS

We validated a set of 326 genes differentially expressed in ductal carcinoma as reported by Turashvili et al.(2007) using an independent dataset. Validation data supported 80% of the genes. The gene set was optimally clustered into two classes by SOM /Ward clustering classifying healthy and cancer tissues. The two classes and genes in them were supported by 60% of the validation data. Four PCs accounted for 86% of the variation in the data with the first 2 PCs accounting for 70%. Top 100 genes in PC1 and PC2 provided 52% support for the two classes and PC1 dominated class 1 and PC2, class 2. SOM and PCA results were used to extract highly influential genes from the top 6%, 18% and 57% of the original genes. These key genes may prove to be crucial in identifying ductal tumor in a diagnostic setting. Four new genes were among key genes that may shed more light onto the disease mechanism.

REFERENCES

Jiang, N., L.J. Leach, X. Hu, E. Potokina, T. Jia, A. Druka, R. Waugh, M.J. Kearsey, Z. Luo, (2008). Methods for evaluating gene expression from Affymetrix datasets. *BMC Bioinformatics*, www.biomedcentral.com/1471-210519/282

Machine Learning Framework for Mathematica, 2007, uni software plus, www.unisoftwareplus.com.

Mathematica, Wolfram Research, Inc. Champaign, IL, 2008.

NCBI gene expression Omnibus. www.ncbi.nlm.nih.gov/geo (orig.data: GSE 7564.Valid. data: GDS 2046)

Samarasinghe, S. (2007), Neural networks for applied sciences and engineering- From fundamentals to complex pattern recognition, Taylor and Francis Group, Florida, USA

Truahvili, G., J. Bouchal, K. Baumforth, W. Wei, M. Dziechciarkova, J. Ehrmann, J. Klein, E. Fridman, J. Skarda, J. Srovnal, M. Hajduch, P. Murray and Z. Kolar, (2007), Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer*, 7: 55 doi:10.1186/1471-2407-7-55.

Zhao, H., A. Langerod, Y. Ji, K.W. Nowels, J.M. Nesland, R. Tibshirani, I.K. Bukholm, R. Karesen, D. Botstein, A. Borresen-Dale, S.S. Jeffrey, (2004), Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Molecular Biology of the Cell*, 15, 2523-2536.

Sorlie, T., C.M.Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, M.van de Rijn, and S.S. Jeffrey. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implication. Proc. Natl. Acad. Sci. USA, 98: 10869-10874.