Integrated Analysis of Gene Expression and Copy Number Data on Gene Shaving Using Independent Component Analysis

Jinhua Sheng, Hong-Wen Deng, Vince D. Calhoun, and Yu-Ping Wang

Abstract—**DNA** microarray gene expression and microarray-based comparative genomic hybridization (a**CGH**) have been widely used for biomedical discovery. Because of the large number of genes and the complex nature of biological networks, various analysis methods have been proposed. One such method is "gene shaving," a procedure which identifies subsets of the genes with coherent expression patterns and large variation across samples. Since combining genomic information from multiple sources can improve classification and prediction of diseases, in this paper we proposed a new method, "**ICA gene shaving**" (**ICA**, independent component analysis), for jointly analyzing gene expression and copy number data. First we used **ICA** to analyze joint measurements, gene expression and copy number, of a biological system and project the data onto statistically independent biological processes. Next, we used these results to identify patterns of variation in the data and then applied an iterative shaving method. We investigated the properties of our proposed method by analyzing both simulated and real data. We demonstrated that the robustness of our method to noise using simulated data. Using breast cancer data, we showed that our method is superior to the Generalized Singular Value Decomposition (**GSVD**) gene shaving method for identifying genes associated with breast cancer.

Index Terms—Clustering technique, comparative genomic hybridization (**CGH**), copy number variation (**CNV**), generalized singular value decomposition (**GSVD**), gene expression, gene shaving, independent component analysis (**ICA**).

1 Introduction

The human genome is estimated to have about 20,000 to 25,000 protein-coding genes [1]. A variety of techniques for the analysis of gene expression data have evolved to exploit the huge amount of information obtained with oligonucleotide arrays [2] and complementary deoxyribonucleic acid (c**DNA**) microarrays [3], [4]. **DNA** microarray technology has been proven to be an effective approach for identifying genes which are potential therapeutic molecular targets [5]. This technique lacks the power for detecting regional variations of the genome. On the other hand, array comparative genomic hybridization (a**CGH)** allows assessment of changes in chromosomal **DNA** sequence copy numbers across the genome and provides valuable information regarding genetic alternations in diseases such as cancers [6], [7]. The **aCGH** technology is an invaluable tool in oncology, which uses microarrays to perform high resolution and genome-wide screening of **DNA** copy number changes. Several important applications of a**CGH** have been reported in cancer research [8], and clinical genetics [9].

With the vast increase in biological information, the problem of integrating different types of genomic measurements has become a great challenge. The integration of chromosomal copy number variation (**CNV**) with gene expression will probably identify new therapeutic targets that could not be identified by analysis of independent platforms alone [10]. Recent investigations [11], [12], [13], [14] have shown the promise of integrated analysis of **CNV** and gene expression. Most studies demonstrate that copy number variation affects the expression levels of those genes contained within that **CNV**. Copy number variations are both directly and indirectly correlated with changes in expression and it is beneficial to examine the indirect effects of **CNV**s [11]. Optimal power to find such associations can only be achieved by analyzing copy number and gene expression jointly [12]. By combining genomic data from different sources, it is possible to obtain an integrated genome-wide view of gene aberration and their effects on gene expression [13], [14]. Gene over or underexpressions usually correspond to increased or decreased copy numbers, respectively (e.g., see Fig. 1). An integrated analysis of gene expression data with copy number data can reveal their intrinsic connections.

Combined analysis of copy number and gene expression microarrays of the same or similar tumor samples has revealed a major and direct effect of allelic imbalance on gene expression in a variety of cancer types, including breast [15], [16], pancreatic [17], colorectal [18], prostate [19], and lung [20] cancer. On a global level, 40-60 percent of the genes at higher level of amplification showed elevated

- *J. Sheng is with the Department of Radiology and Imaging Sciences, Indiana University School of Medicine, 950 West Walnut Street, Indianapolis, IN, 46202. E-mail: jinhua_sheng@yahoo.com.*
- *H.-W. Deng is with the Department of Biostatistics and Bioinformatics, Tulane University School of Public Health and Tropical Medicine, 1440 Canal Street, Suite 2011, New Orleans, LA, 70112. E-mail: hdeng2@tulane.edu.*
- *V.D. Calhoun is with the The Mind Research Network, 1101 Yale Boulevard, Albuquerque, NM 87131. E-mail: vcalhoun@unm.edu.*
- *Y.-P. Wang is with the Department of Biomedical Engineering and Department of Biostatistics and Bioinformatics, Tulane University, 533 Lindy Boggs Bldg., New Orleans, LA 70118, and is also affiliated with Shanghai University for Science and Technology. E-mail: wyp@tulane.edu.*

Manuscript received 23 Mar. 2010; revised 21 July 2010; accepted 21 Nov. 2010; published online 14 Apr. 2011.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2010-03-0078.
Digital Object Identifier no. 10.1109/TCBB.2011.71.

1545-5963/11/$26.00 © 2011 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

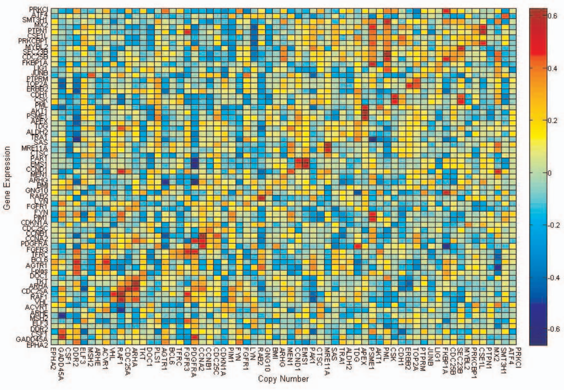

Fig. 1. Display of the Pearson's correlation analysis between copy number and gene expression level across the NCI-60 cell lines. This indicates correlations existed along the diagonal line where the copy number variations cause the corresponding gene expression changes.

expression, while 10 percent of highly over-expressed genes were amplified. In low-level copy number aberrations, only about 10 percent of the genes have been reported to show coherent changes in gene expression [21]. Fig. 1 displays the Pearson correlation coefficients for all possible combinations of gene expression and copy number changes from the NCI-60 cell lines [22], indicating that a correlation exists between the expression levels of genes and copy number changes around the same locations of the genome (along the diagonal line). Variations in gene expression and gene copy number are strongly linked to diseases such as breast cancer and have a bit positive over negative correlations [23]. Genes in tumorigenesis show associations between copy numbers and expression levels. Some copy number changes extend over larger chromosomal regions.

Integrating data from different sources such as gene expression and copy number can increase the reliability of the analysis results and the prediction of prognosis. Association between copy number changes and gene expression levels have been studied in [16], [21], [22], and ~12 percent of gene expression variation can be explained by differences in copy numbers [19]. Integration of **DNA** copy number alterations and gene expression profiling may also result in improved classification and prognosis in breast cancer. For example, Chin et al. [24] found that the accuracy of risk stratification according to the outcome of breast cancer disease can be improved by joint analyses of gene expression and **DNA** copy number. Several approaches have been described to identify a subset of genes, whose expression levels are most significantly associated with copy number changes in the corresponding genomic region [25]. The singular value decomposition (**SVD**) or the principal component analysis (**PCA**) has been a popular method for analyzing and reducing the dimension of gene data [26], [27]. The **SVD** model describes the overall observed genome-scale molecular biological data as the outcome of a simple linear network. However, the gene expression and copy number data are separately analyzed using the **SVD** method. The generalized singular value decomposition (**GSVD**) model describes the two genome-scale molecular biological data sets as the outcome of a simple linear comparative network, where a few independent sources, some common to both data sets whereas some are exclusive to one data set or the other, affect all the genes in both data sets. In 2006, Berger et al. [28] applied an iterative shaving method based on the **GSVD** of their joint data sets to identify subsets of genes with similar gene expression or copy number patterns. The **SVD** and **GSVD** models are usually used to model **DNA** microarray data. The **GSVD** is already a trusted method for analyzing and reducing the dimension of gene data in two breast cancer cell line and tumor data sets for the identification of gene subsets that are biologically validated. The independent component analysis (**ICA)** and **PCA** are very similar in some respects; however, the goals of the two methods are different. The **ICA** finds the statistically independent components and is more suitable for separating mixed signals and uncovering hidden biological processes from the observed measurements.

The **GSVD**-based approach assumes that gene expression or gene copy number data are generated by the linear combination of a set of biological processes. However, this assumption might not be realistic. The **ICA** uses a more general statistical assumption (as described in Section 2.2), which is more appropriate for modeling and analysis of genomic data. **ICA** has been recently successfully used for the joint analysis of f**MRI**, EEG, and genomic imaging data [29], [30]. Motivated by these facts, we used the **ICA** technique to jointly analyze gene expression and copy number data and the preliminary results were encouraging [31]. In this paper, we present our recent results on the development of an **ICA**-based iterative dimension reduction method and apply it to analyze both gene expression and copy number data in order to identify subsets of genes with coherent expression patterns and large variation across subjects. We examine the robustness of the method to noise and its convergence properties using simulated data. We apply the method to breast cancer cell line and breast cancer tumor studies and demonstrate the effectiveness of the method. With our proposed algorithm, we can identify a list of variant genes and select genes that correspond to functionally related groups. When compared with the **GSVD**-based method, improved performance is obtained in identifying genes that are known to contribute to the progression of breast cancers.

2 Method

We introduce our **ICA**-based method for the integrated analysis of gene expression and copy number change data and then apply it to the identification of gene subsets in the breast cancer cell and breast tumor data in combination with a gene shaving method.

2.1 Gene Shaving

Large scale gene expression studies, such as those conducted using c**DNA** arrays, often provide millions of data points. A **PCA**-based statistical method called "gene shaving" was introduced in [27] to identify groups of genes that have coherent patterns of expression with large variance across samples, or groups of genes that optimally separate the sample into predefined classes. Gene shaving differs from hierarchical clustering and other widely used methods for analyzing gene expression studies in that genes may belong to more than one clusters, and the clustering may be supervised by an outcome measure. Fig. 2 shows a schematic procedure of the gene shaving process based on the **PCA**. The goal of gene shaving is to extract coherent and typically small clusters of genes that vary as much as possible across the samples. The first principal component of the current cluster of genes is computed. This eigen-gene is the linear combination of genes with largest the variance across samples. We

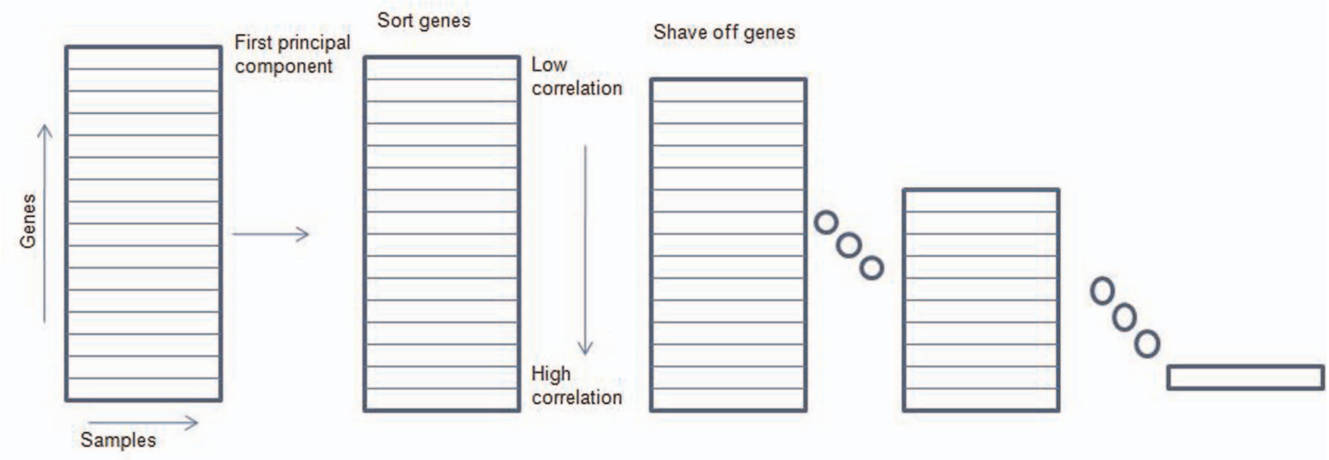

Fig. 2. The procedure of the “gene shaving” method for isolating interesting genes from a set of **DNA** microarray experiments as used in [27].

compute the correlation of each gene with the eigen-gene, and shave off the genes having lowest correlation. The process is then repeated on the reduced cluster of genes.

The shaving process shown here requires repeated computation of the largest component of a large set of variables and retains the typically 90-95 percent of genes with the highest variance at each iteration until all clusters (such as the top 5-10 percent highest variant genes) are found. The gene shaving method is a potentially useful tool for the exploration of gene expression data and for identification of interesting clusters of genes whose expressions are highly predictive of certain cancers and patient survival.

2.2 ICA Approach

The ICA is a recently developed method in which the goal is to find a linear representation of unknown non-Gaussian data so that the components are statistically independent, or as independent as possible. Such a representation seems to capture the essential structure of the data in many applications, including feature extraction and signal separation. The **ICA** is becoming an increasingly popular tool for analyzing biomedical data. Liebermeister [32] proposed using the linear **ICA** for microarray analysis to extract expression modes, where each mode represents the linear influence of a hidden cellular variable. However, to our knowledge, no results have been reported to use **ICA** for the combined analysis of gene expression and copy number data sets.

Consider an observed m-dimensional random vector denoted by $X = (x_1, \ldots, x_m)^T$, which is generated by the source signals **S** with an unknown process [33]

$$X = A{\cdot}S + N_t, \tag{1}$$

where $S = (s_1, \ldots, s_n)^T$ is an n-dimensional vector, and is not observable; $A_{m\times n}$ is an unknown mixing matrix; and N_t is Gaussian noise. Typically $m >= n$, so A is usually of full rank. A typical **ICA** model assumes that the elements in the source signal S are statistically independent, and are mostly non-Gaussian, with an unknown but linear mixing process.

The goal of **ICA** model is to estimate a separation matrix $W_{n\times m}$ such that Y is a good approximation to the true sources S

$$Y = W{\cdot}X. \tag{2}$$

The separation matrix W is the approximate inverse of the mixing matrix A and can be estimated from the observed data to ensure independent coefficients S, with non-Gaussian distributions. Therefore, the **ICA** is an approach for solving the blind source separation (**BSS**) problem. This approach has been used to solve the cocktail party problem, where several people are speaking simultaneously in the same room. The problem is to separate the voices of different speakers from their mixed voices recorded by a few microphones in the room. The **ICA** model for blind source separation is shown in Fig. 3.

Some classical approaches to solving **BSS** problem include the maximization of information transformation, maximization of non-Gaussianity, mutual information minimization, and tensorial methods. Some of the most commonly used **ICA** algorithms are the **FastICA** [34], Infomax [35] and joint approximate diagonalization of eigen-matrices (**JADE**) [36]. In this paper, the **FastICA** algorithm was utilized, which has been proven to be effective for our data. It performs centering and whitening as a preprocessing step.

We now apply the **ICA** model to our gene expression or gene copy number change data and (1) can be generalized as:

$$R = A{\cdot}U + N_t, \tag{3}$$

where the input matrix $R_{m\times p}$ contains gene expression or gene copy number data; $U_{n\times p}$ is an $n \times p$ matrix containing all unknown source signals; p is the number of genes and m is the number of experiments.

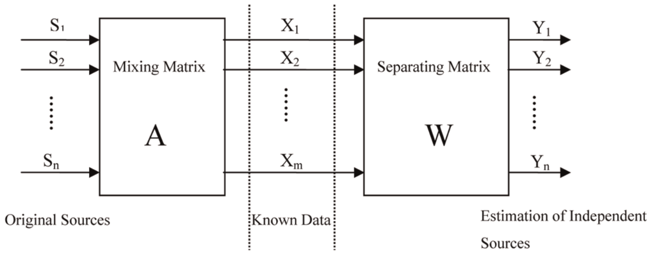

Fig. 3. A basic **ICA** model for blind source separation.

We project each input set onto the kth column of A corresponding to the direction of the highest variance to find the highest parallel contribution from data R

$$R^T \cdot \boldsymbol{a_k} = (\boldsymbol{a_k}^T \cdot A \cdot U)^T, \tag{4}$$

where $\boldsymbol{a_k}$ is a $m \times 1$ vector, i.e., the kth column of A and T denotes matrix transposition.

The projection direction, the kth column of A can be sought, corresponding to the maximum value of the sum of the kth row of matrix $A^{T \cdot} A$.

2.3 Joint ICA

The common technique used to analyze the input data is to project the original data on a lower-dimensional subspace expanded by orthogonal components of the decomposition and find clusters that are tight and far away from other clusters. Instead of the orthogonal ones, here we get a subspace spanned by statistically independent components based on the **ICA**. We apply the **ICA** model to uncover the complex biological process that lead to two different measurements, e.g., gene expression and gene copy number variations. Based on the **ICA** analysis of these two joint data sets, we accomplish the goal of "gene shaving." An iterative dimension reduction method based on **ICA** is proposed to analyze both gene expression and copy number data in order to locate functionally related gene subsets.

Joint **ICA** [29], [30] is an approach that enables us to jointly analyze data from multiple modalities collected in the same set of subjects. The gene expression and copy number data can be better analyzed in a unified framework in which the two sets of data are fused. We assume the independence of gene expression and copy number data, respectively, using the following generative models for the data:

$$\begin{cases} \mathrm{R_A} = \mathrm{A}_A \cdot \mathrm{U}_A, \\ \mathrm{R_B} = \mathrm{A}_B \cdot \mathrm{U}_B, \end{cases} \tag{5}$$

where $\boldsymbol{R_A}$ and $\boldsymbol{R_B}$ represent the matrix of gene expression and copy number changes, respectively; $\boldsymbol{U_A}$ and $\boldsymbol{U_B}$ represent their source signals, and $\boldsymbol{A_A}$ and $\boldsymbol{A_B}$ are their mixing matrices. Our idea is motivated by the algorithm for fusion of f**MRI** and **ERP** data proposed by Calhoun et al. [29], [30] but applied to gene expression and copy number separately. When the **ICA** is applied to the union of gene expression and copy number, it is similar to the algorithm by Calhoun et al. [30].

Because aberrations in gene expression and gene copy number are correlated, the elements of the mixing matrices should be correlated. The idea of creating snapshots of the **ERP** and f**MRI** data can be translated into fusing the mixing matrices of gene expression and copy number in our case. Both mixing matrixes can be interacted to find the direction of the highest variance on both data sets. The joint contribution from $\boldsymbol{R_A}$ and $\boldsymbol{R_B}$ can be computed as

$$\begin{cases} M_A = |A_B| \cdot A_A^T, \\ M_B = A_B \cdot |A_A|^T. \end{cases} \tag{6}$$

We compute the top 5 percent percent of genes with the highest parallel contribution from $\boldsymbol{R_A}$ and $\boldsymbol{R_B}$ corresponding to the highest variances. We project the original data in the kth direction as

$$\begin{cases} P_A = R_A^T \cdot m_{A_k}, \\ P_B = R_B^T \cdot m_{B_k}, \end{cases} \tag{7}$$

where $m_{A_k}$and $m_{B_k}$are the kth column of $\boldsymbol{M_A}$ and $\boldsymbol{M_B}$, corresponding to the direction of the largest variance from the matrix pair R_A and R_B, respectively.

2.4 Joint ICA-Based Gene Shaving Algorithm

The genes are iteratively projected onto the vector corresponding to the independent component with highest variance. The projection corresponds to the direction of highest variation in the original data. The joint **ICA** method can be extended to accomplish the goal of "shaving" based on the chosen direction. We proposed the following algorithm and its two variants for clustering genes where the genes may be of different significance in both data sets. Ninty to ninty-five percent of the genes are retained from data sets with joint **ICA** in the direction of the highest variance, from which the corresponding genes that contribute to cancer progression are identified.

Algorithm 1. Gene shaving is based on the selection of genes from the a**CGH** data. The schematic procedure of this algorithm is shown in Fig. 4, where **e**ach individual procedure is connected with solid lines.

Given the matrix $\mathbf{R_A}$ of a**CGH** and the matrix $\mathbf{R_B}$ of gene expression for the same organisms or the same clones of the same samples, we perform the following steps:

1. Preprocess microarray data, quality filtering, normalization, and data transformation.
2. Form the matrix $R = [^{\mathrm{R_A}}_{\mathrm{R_B}}]$.
3. Compute the mixing matrix M_A using the **FastICA** algorithm, analyze and select the direction of projection.
4. Project **R** onto the independent component according to the chosen direction, which corresponds to largest variance.
5. Retain the top $\eta = 95\%$ of genes with the highest contribution from R_A and select the related genes from R_B corresponding to retained a**CGH** data.
6. Reform the matrix R after shaving.
7. Repeat Steps 3-6 if the number of genes is greater than or equal to the set number of samples.
8. Analyze the clusters with the top 5 percent highest variant genes through visualization and functional assessment.

There are two variants of Algorithm 1, depending on the selection of genes in terms of a**CGH** and/or c**DNA** data.

Algorithm 2. Joint **ICA** gene shaving is based on the selection of genes from c**DNA** data. Algorithm 2 is similar to algorithm 1, but genes are selected in terms of **cDNA** data. The schematic procedure of this algorithm is shown in Fig. 4, in which each individual procedure is connected through solid and dotted lines.

Algorithm 3. Joint **ICA** gene shaving is based on the selection of genes from both the matrix $\mathbf{R_A}$ of a**CGH** and the matrix $\mathbf{R_B}$ of **cDNA**. The genes with the lowest correlation from $\mathbf{R_A}$ or $\mathbf{R_B}$ are all shaved off. The schematic procedure of this algorithm is shown in Fig. 4, in which **e**ach individual procedure is connected through solid and dashed lines.

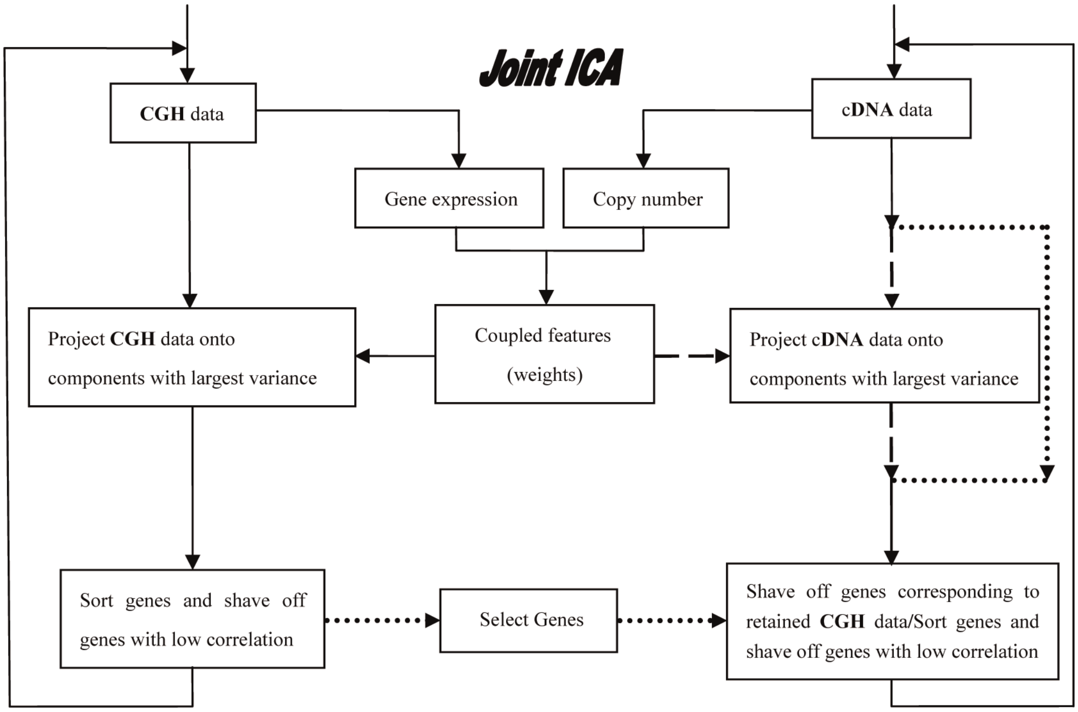
Joint ICA
CGH data
cDNA data
Gene expression
Copy number
Coupled features (weights)
Project CGH data onto components with largest variance
Project cDNA data onto components with largest variance
Sort genes and shave off genes with low correlation
Select Genes
Shave off genes corresponding to retained CGH data/Sort genes and shave off genes with low correlation

Fig. 4. The schematic procedure of joint **ICA** gene shaving to identify gene subsets.

These algorithms are appropriate for different data sets, which is similar to the **GSVD** method when using different project angle parameters [28]. Algorithm 1 depends more on copy number data; Algorithm 2 depends more on gene expression; and Algorithm 3 depends on both of them. We apply these iterative procedures in the following section to locate functionally related gene subsets, corresponding to similar and dissimilar patterns of variations in gene expression and/or gene copy number changes.

3 Results and Discussion

We applied the **ICA** gene shaving method for dimension reduction and clustering analysis of combined a**CGH** and **cDNA** expression data. In order to test the robustness of the method to noise, we generated simulation data as described in Berger et al. [28] and compared **ICA** gene shaving and **GSVD** gene shaving when the data contain noise. Our proposed algorithms were applied to demonstrate efficacy to real data from breast cancer cell lines and a breast cancer tumors, which were preprocessed by normalization and log_2-transformation. The algorithms were implemented in Matlab and the codes and data are available for download on the website [37].

3.1 Test on Simulation Data

Copy number data were generated using the model proposed by Wang et al. [38], which defined three states: **amplified** (**a**), **deleted** (**d**), and **normal** (**z**). Gene expression data were generated based on the model of Attoor et al. [39]. Gene expression was defined as: **over** (**o**), **under** (**u**), and **constant** (**c**) expression state. The relation between copy number and gene expression states was modeled using a simple state flow. The connection between the data was modeled by the transition probability matrix [22]

$$P = \begin{bmatrix} P_{du} & P_{dc} & P_{do} \\ P_{zu} & P_{zc} & P_{zo} \\ P_{au} & P_{ac} & P_{ao} \end{bmatrix}. \tag{8}$$

In our simulations, we assumed a strictly correlative model between copy number and gene expression states using the transition probability matrix, $P = I_{3\times3}$.

By increasing the noise variance, different groups of genes were observed after the shaving iterations were completed. In order to evaluate the robustness of the method to noise, the gene list percentage similarity (**PS**) was computed by counting the number of genes obtained from noisy data (**ND**) intersecting with that obtained from the original data (**OD**) [28]

$$PS = \frac{\#ND \cap \#OD}{\#Tot} \times 100\%, \tag{9}$$

where **Tot** is the number of total genes in the list.

We compared our proposed **ICA** gene shaving method with the **GSVD** gene shaving by analyzing of an ensemble of 1,000 expression and copy number data sets in a simulation study. Each set has N = P = 1,500 genes in three samples. We analyzed 75 remaining genes. Additive random noise was generated 1,000 times for each variance level. We compared the two methods based on the percentage similarity index. The results were shown in Fig. 5.

The results in Fig. 5 show that the ranges of **PS** for both gene expression and gene copy numbers decrease with the increase of noise level, regardless of the shaving method used. The **PS** value with **ICA** gene shaving method is always higher than that of **GSVD** gene shaving, which indicates that the **ICA** gene shaving method is more robust to the noise.

3.2 Cell Line Case Study

After the proposed **ICA** gene shaving method has been proven to be effective on simulated data, it was then tested on

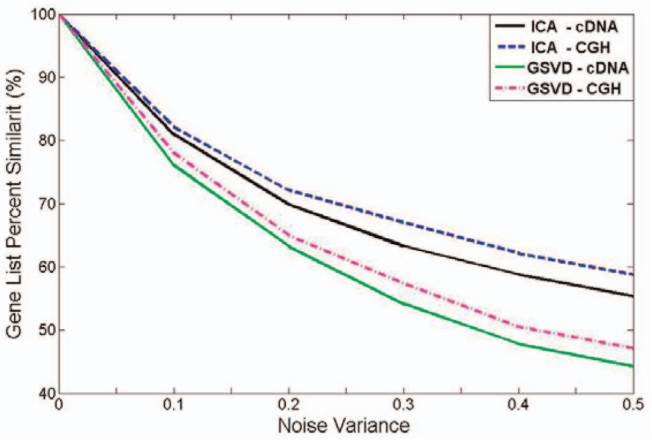

Fig. 5. The effects of additive noise on **PS** value in c**DNA** and a**CGH** data using **GSVD** gene shaving and **ICA** gene shaving algorithm, respectively.

real biological data. Three breast cancer cell lines with similar copy number profiles on chromosome 17 were analyzed [40]. The **SKBR3**, **BT474**, and **UACC812** cell lines all have amplified regions around the **ERBB2** gene, which is known to play roles in the progression of breast cancers [15].

From the original data set from Hyman et al. [15], we parse out genes from chromosome 17. Each set has $N = P = 619$ genes in three samples. We retained the top 5 percent of the most interesting genes in chromosome 17. We detected genes and genomic locations from gene expressions and copy numbers with high variations, as shown in Figs. 6 and 7, respectively. We obtained a list of genes and copy numbers that captured the highest shared variation with our proposed method. Fig. 8 shows the list of gene subsets from the **ICA** and **GSVD** gene shaving, respectively, based on gene expression data, while Fig. 9 displays the list of gene subsets based on gene copy number changes. Fig. 10 displays the top 15 highest variant genes from combined gene expression and copy number changes using the **ICA** and **GSVD** methods, respectively.

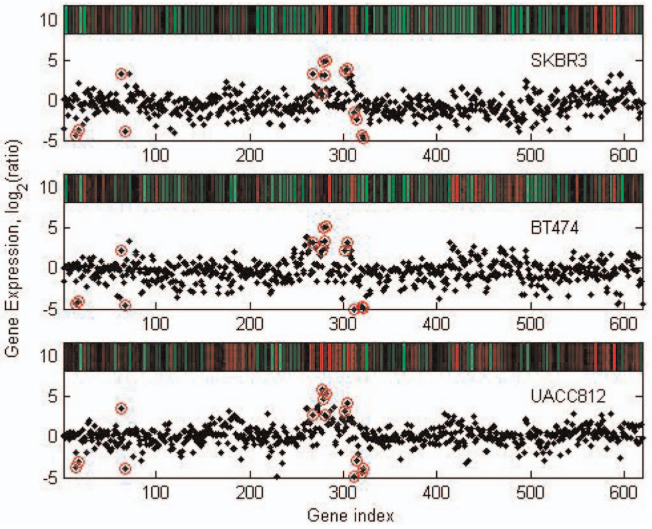

Fig. 6. Plot of selected genes from **cDNA** gene expression data. This plot shows the original cell line expression data for the **SKBR3**, **BT 474**, and **UACC812** cell lines over chromosome 17. The circled genes were selected using our **ICA** gene shaving method.

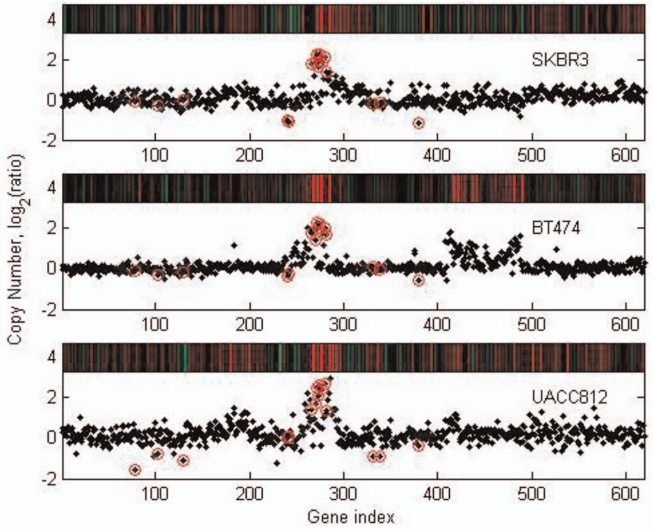

Fig. 7. Plot of selected genes from a**CGH** copy number data. This plot shows the original cell line copy number data for the **SKBR3**, **BT 474**, and **UACC812** cell lines over chromosome 17. The circled genes were selected using our **ICA** gene shaving method.

From the gene list provided, we observe that all **ERBB2** genes were successfully extracted using our **ICA** gene shaving method while one **ERBB2** gene was extracted using the **GSVD** gene shaving method. Our method was also able to uncover several **HOX** family genes (**HOXB3**, **HOXB6**, and **HOXB7**), which have been found to contribute to the progression of several cancer types [41]. Thus, our **ICA** gene shaving method found more genes related to breast cancers than the **GSVD** gene shaving method.

3.3 Analyzing Breast Cancer Cell Lines and Breast Cancer Tumors

We present another case study using the data from breast cancer cell lines [15] and breast tumors [42].

Our **ICA** gene shaving method was applied to the breast cancer cell lines [15] with Algorithms 1-3. We report the top 50 of the highest variant genes corresponding to algorithm 3 in Figs. 11 and 12 in terms of gene expression and copy number ratios, respectively. We can observe the correlation across the samples for over- or underexpressed genes, in addition to amplified or deleted genes. The genes in Fig. 11 capture the highest expression variations, which represent extremely over- and underexpression with similar transcriptional responses. Similarly, the genes in Fig. 12 capture the highest variation in the copy number changes. We can isolate the groups of genes that have similar and dissimilar patterns of gene expression and copy number. The genes with high copy number changes show highly similar expression characteristics. Figs. 11 and 12 demonstrate the ability of our algorithms to locate genes with highest variation and with the strongest correlation across all the samples.

In the study of 37 breast tumors conducted by Pollack et al. [42], it was reported that the copy number changes played a direct role in the transcriptional program of human breast tumors [42]. Based on the analysis of breast tumor data, we show the top 50 highest variant genes using the **ICA** gene shaving (Algorithms 1-3), respectively, on both gene expression and copy number data as shown in Fig. 13. We also compared with the **GSVD** gene shaving method of different relative significance as shown in Fig. 14.

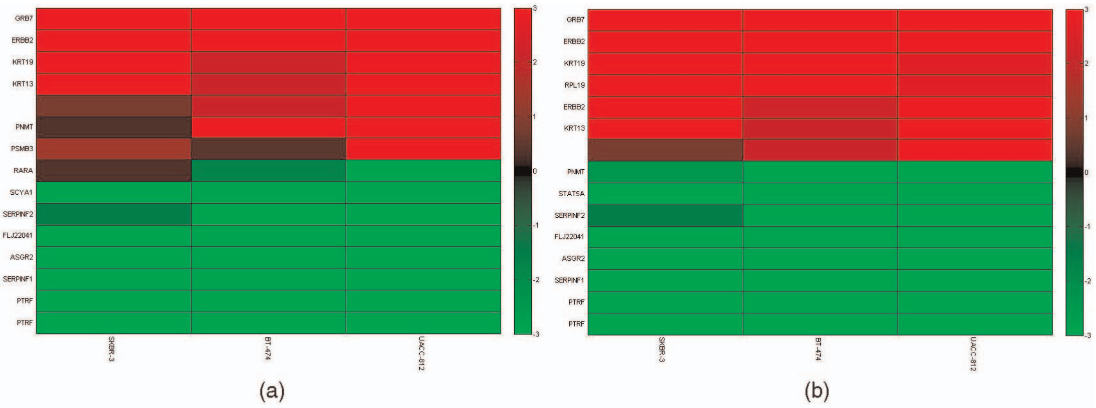

Fig. 8. These plots show the selected genes using (a) The **GSVD** gene shaving method. (b) The **ICA** gene shaving method, respectively, based on c**DNA** gene expression.

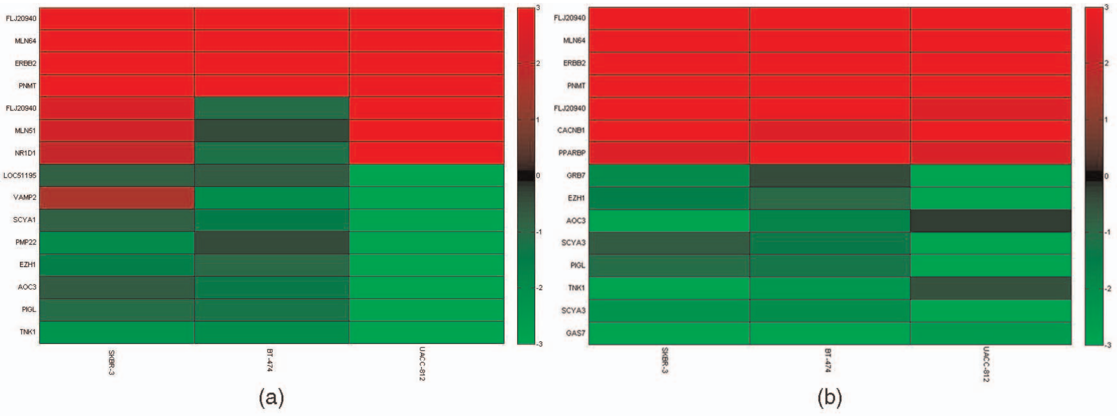

Fig. 9. These plots show the selected genes using (a) The **GSVD** gene shaving method. (b) The **ICA** gene shaving method, respectively, based on a**CGH** copy number data.

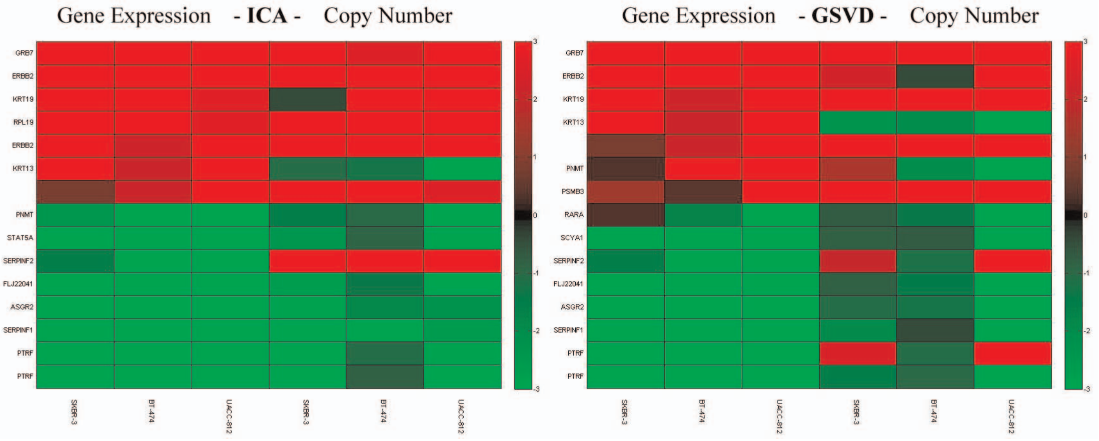

Fig. 10. We retain the gene expression values of the top 15 highest variant genes from combined gene expression and copy number changes using the **ICA** and **GSVD** methods, respectively.

From Figs. 13 and 14, our **ICA** gene shaving method has better ability to locate genes with highest variation in copy numbers than using the **GSVD** gene shaving method. The subsets of genes with similarly higher and lower gene copy number changes can be identified with the **ICA** gene shaving method. No patterns of similar gene expressions were observed in the list of genes with the top 25 highest (positive or negative) variant gene expression using either the **GSVD** gene shaving or the **ICA** gene shaving method.

We summarize parameters such as p-values in selecting genes used in the **ICA** and **GSVD**-based gene shaving methods, as in Tables 1 and 2. They are for analyzing both gene expression and copy number data, and for analyzing breast cancer cell lines and breast cancer tumors, respectively. The lower P-value is, the more statistically significant the detected cluster is. Table 2 and Figs. 13, 14 all show that even though **ICA** gene shaving method has better quality in detecting the clusters than the **GSVD** method, it is not good enough to distinguish clearly the top highest gene expressions for the study of breast cancer tumors [42].

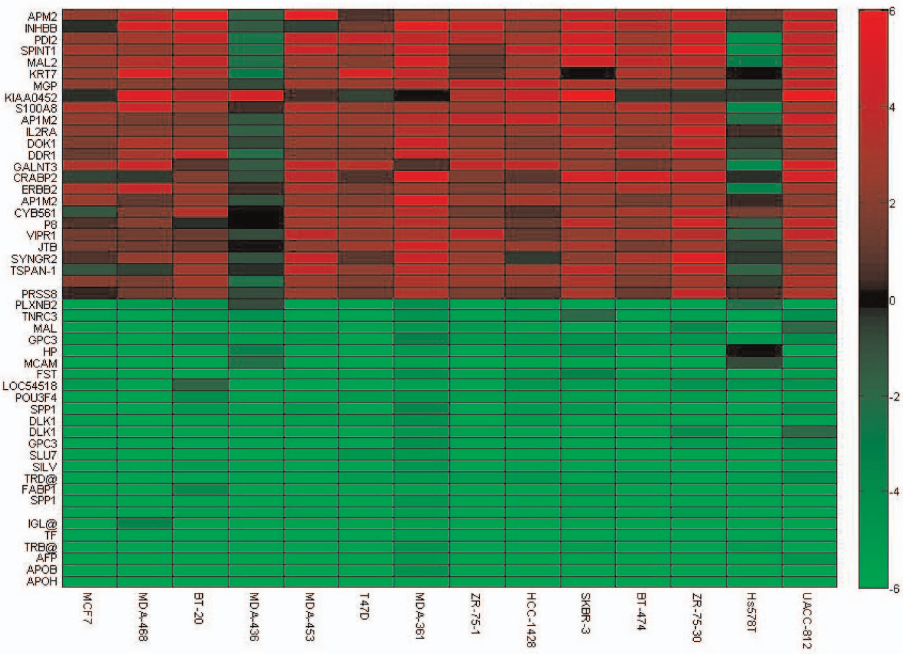

Fig. 11. The top highest variant genes of gene expression in 14 samples are retained using algorithm 3 in the study of breast cancer cell lines [15]. The pattern shows the highest parallel contributions to the iterative projections with gene shaving.

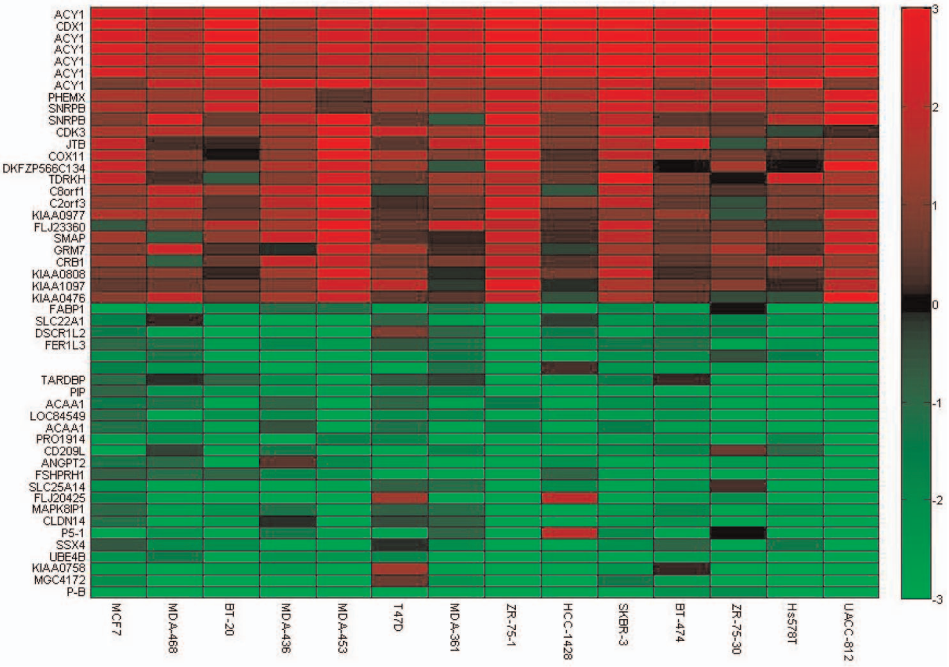

Fig. 12. The top highest variant genes with gene copy number changes in 14 samples are retained using algorithm 3 in the study of breast cancer cell line [15].

We also applied our method to identify gene subsets that contribute to breast cancer tumors. Genes with the highest statistical significance include **ERBB2**, **MUC1**, and **GRB7** with concomitant changes in copy number and expression levels. For the tumor samples, our **ICA** gene shaving method was able to locate known or candidate oncogenes successfully. The **GSVD** gene shaving method obtained all three oncogenes (**ERBB2**, **CCND1**, and **MYC**) and two candidate oncogenes (**GRB2** and **TPD51**) corresponding to projection angle *"max"*; two oncogenes (**ERBB2** and **MYC**) and two candidate oncogenes (**TPD52** and **GRB7**) corresponding to *"min"*; and two oncogenes (**ERBB2** and **MYC**) and a candidate oncogenes (**GRB7**) corresponding to *"zero."* Our **ICA** gene shaving method obtained all three oncogenes (**ERBB2**, **CCND1**, and **MYC**), and three candidate oncogenes (**GRB2, TPD52**, and **GRO1**) corresponding to "Algorithm 1"; two candidate oncogenes (**GRB2** and **GRO1**) corresponding to "Algorithm 2"; and three candidate oncogenes (**GRB2**, **TPD52**, and **GRB7**) corresponding to "Algorithm 3." These genes were known to contribute to the progression of breast cancer tumors but were missed by the **GSVD** gene shaving method.

Our method was successfully used to locate important genes that exhibit patterns of similar and dissimilar

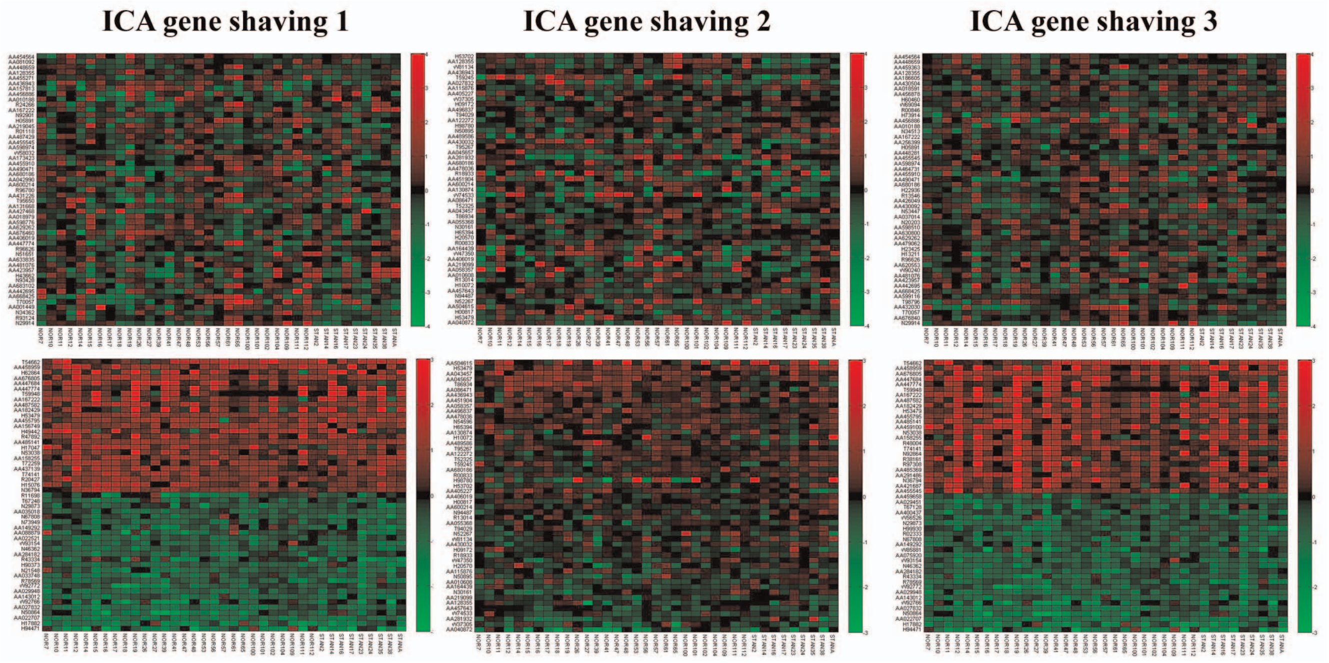

Fig. 13. The top three pictures are the lists of genes with the top 50 highest variant gene expression using three **ICA** gene shaving methods, respectively. The bottom three pictures are the list of genes with the top 50 highest variant copy numbers using three **ICA** gene shaving methods, respectively. The subsets of genes which have similar gene copy number changes can be identified. The data are from the study of breast tumors [42].

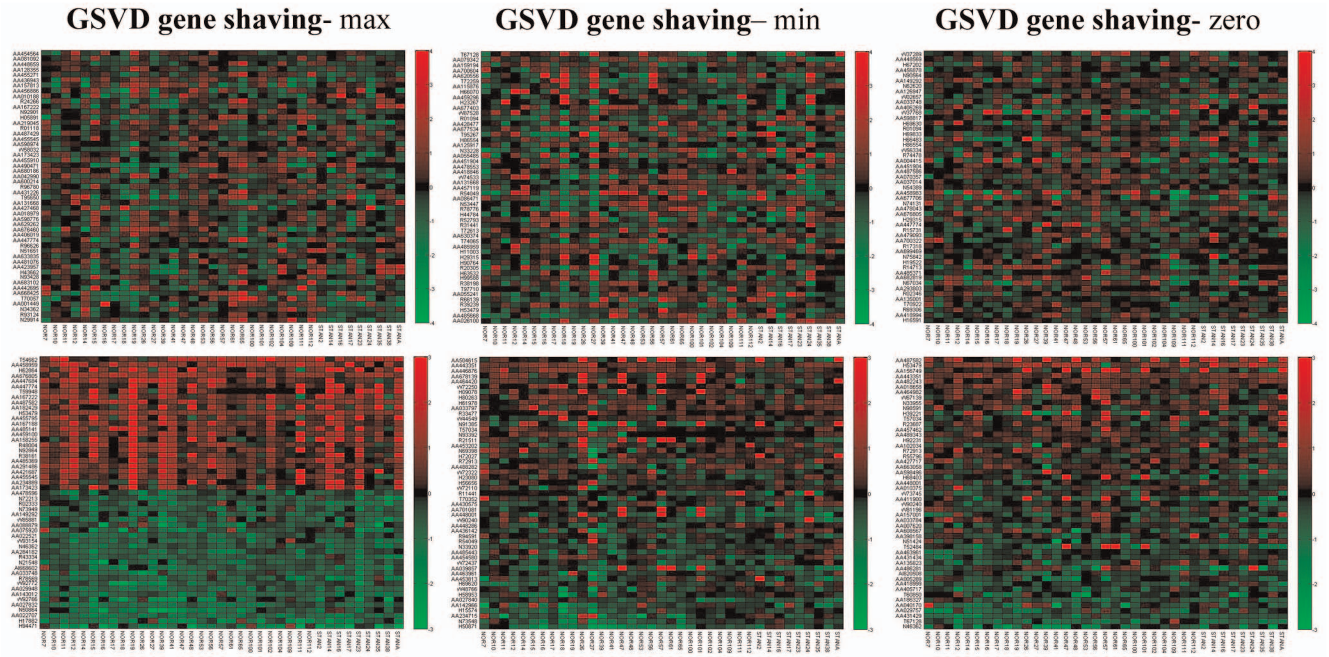

Fig. 14. The top three pictures are the lists of genes with the top 50 highest variant gene expression using three **GSVD** gene shaving methods, respectively. The bottom three pictures are the lists of genes with the top 50 highest variant copy numbers using three **GSVD** gene shaving methods, respectively. “max” indicates no significance in the copy number data set relative to the gene expression data set; “min” indicates no significance in the gene expression data set relative to the gene copy number data set; “zero” indicates that genes may be of equal significance in both data sets. The data are from the study of breast tumors [42].

variations. All three oncogenes and more candidate oncogenes are obtained by the three algorithms of the **ICA** gene shaving method, even if no patterns of similar gene expressions are observed. “Algorithm 1” depends more on the gene copy number data set, and “Algorithm 2” depends more on the gene expression data set. “Algorithm 3” uses both the gene expression and copy number data sets equally. These algorithms are appropriate for different data sets, which is similar to the **GSVD** method when using different projection angles [28].

4 Conclusion

Combining genomic data from different sources promises to be a very robust, reliable, and efficient technique. In this paper, we integrate gene copy number changes with gene

TABLE 1
A Comparison of Parameters Used for the Study of Breast Cancer Cell Lines [15]

Methods	Parameter/Algorithm	P-value(gene expression)	P-value(copy number)
GSVD	$\theta_{\max}$	< 0.001	< 0.001
	$\theta_{\min}$	< 0.001	< 0.001
	θ_0	< 0.001	< 0.001
ICA	algorithm 1	< 0.001	< 0.001
	algorithm 2	< 0.001	< 0.001
	algorithm 3	< 0.001	< 0.001

TABLE 2
A Comparison of Parameters Used for the Study of Breast Cancer Tumors [42]

Methods	Parameter/Algorithm	P-value(gene expression)	P-value(copy number)
GSVD	$\theta_{\max}$	0.7748	<0.001
	$\theta_{\min}$	0.8766	< 0.001
	θ_0	0.8968	< 0.001
ICA	algorithm 1	0.4156	< 0.001
	algorithm 2	0.5321	< 0.001
	algorithm 3	0.3432	< 0.001

expression for locating subsets of genes with similar and dissimilar patterns of variations. The combined data sets result in more accurate identification of gene subsets associated with cancers and diseases. We compared the **ICA**-based gene shaving method with the **GSVD** based one. When tested on simulated data, the **ICA** gene shaving method increased performance by about 10 percent over that of the **GSVD** gene shaving in terms of the gene list percentage similarity value, which indicates the improved robustness of the method to noise. Statistical analysis was performed using both copy number and expression data to identify genes, showing differential expressions associated with copy number alterations.

The **SVD** method has been used for the analysis of gene expression and copy number data [26], which are, however, not analyzed in an integrated manner. The **GSVD**-based gene shaving method was proposed in [28] to integrate the two data sets. It has been used to identify gene subsets in breast cancer cell lines and breast cancer tumors, but also has limitations. Our proposed **ICA** gene shaving method improves this method by using a more realistic model, as demonstrated in our simulation study. Furthermore, testing on real breast cancer cell and breast tumor data shows that the **ICA** gene shaving method can identify genes that were missed by the **GSVD** gene shaving method, which are known to contribute to the progression of breast cancers. All three oncogenes and more candidate oncogenes can be obtained with our **ICA** gene shaving method. This method will contribute to better medical diagnosis and prognosis with improved identification of gene subsets associated with diseases and cancers.

The **ICA** method appears to be useful for gene data analysis, but it also has some inherent limitations. If gene component processes exhibit saturation or other nonlinear properties, it may not be appropriate for analysis using a wholly linear model. The **ICA** algorithm assumes that the distribution for each signal component is statistically independent. This criterion provides an essentially unique decomposition of the data, but it may not necessarily be the desired representation for all purposes. There are new developments or other variants of **ICA** methods such as the group **ICA** [29] and we are currently exploring their use in integrated genomic data analysis.

ACKNOWLEDGMENTS

This work has been supported by the US National Science Foundation (NSF) Advanced Biological Informatics Program, and the National Institutes of Health (NIH) under the grant R21 LM010042.

REFERENCES

[1] Int'l Human Genome Sequencing Consortium, "Finishing the Euchromatic Sequence of the Human Genome," *Nature,* vol. 431, pp. 931-945, Oct. 2004.

[2] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown, "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays," *Nature Biotechnology,* vol. 14, pp. 1675-1680, Dec. 1996.

[3] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," *Science,* vol. 270, pp. 467-470, Oct. 1995.

[4] G.B. Bezerra, G.M.A. Cançado, M. Menossi, L.N. de Castro, and F.J. Von Zuben, "Recent Advances in Gene Expression Data Clustering: A Case Study with Comparative Results," *Genetics and Molecular Research,* vol. 4, pp. 514-524, 2005.

[5] J. Chen and Y.-P. Wang, "A Statistical Model-Based Approach for the Identification of DNA Copy Number Changes in Array CGH Data Sets," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 6, no. 4, pp. 529-541, Oct.-Dec. 2009.

[6] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, and T.R. Golub, "Prediction of Central Nervous System Embryonal Tumor Outcome Based on Gene Expression," *Nature,* vol. 41, pp. 436-442, 2002.

[7] O.P. Kallioniemi, A. Kallioniemi, J. Piper, J. Isola, F.M. Waldman, J.W. Gray, and D. Pinkel, "Optimizing Comparative Genomic Hybridization for Analysis of DNA Sequence Copy Number Changes in Solid Tumors," *Genes Chromosomes Cancer,* vol. 10, pp. 231-243, 1994.

[8] A. Kallioniemi, "CGH Microarrays and Cancer," *Current Opinion in Biotechnology,* vol. 19, pp. 36-40, 2008.

[9] M. Shinawi, S.W. Cheung, "The Array CGH and Its Clinical Applications," *Drug Discovery Today,* vol. 13, pp. 760-770, 2008.

[10] M.R. Speicher and N.P. Carter, "The New Cytogenetics: Blurring the Boundaries with Molecular Biology," *Nature Reviews Genetics,* vol. 6, pp. 782-792, 2005.

[11] H. Lee, S.W. Kong, and P.J. Park, "Integrative Analysis Reveals the Direct and Indirect Interactions between DNA Copy Number Aberrations and Gene Expression Changes," *Bioinformatics,* vol. 24, pp. 889-896, 2008.

[12] R.X. Menezes, M. Boetzer, M. Sieswerda, G.B. Ommen, and J.M. Boer, "Integrated Analysis of DNA Copy Number and Gene Expression Microarray Data Using Gene Sets," *BMC Bioinformatics,* vol. 10, pp. 203-217, 2009.

[13] M. Schäfer, H. Schwender, S. Merk, C. Haferlach, K. Ickstadt, and M. Dugas, "Integrated Analysis of Copy Number Alterations and Gene Expression: A Bivariate Assessment of Equally Directed Abnormalities," *Bioinformatics,* vol. 25, pp. 3228-3235, 2009.

[14] C. Soneson, H. Lilljebjörn, T. Fioretos, and M. Fontes, "Integrative Analysis of Gene Expression and Copy Number Alterations Using Canonical Correlation Analysis," *BMC Bioinformatics,* vol. 11, pp. 191-211, 2010.

[15] E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringneár, G. Sauter, O. Monni, A. Elkahloun, O.P. Kallioniemi, and A. Kallioniemi, "Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer," *Cancer Research,* vol. 62, pp. 6240-6245, 2002.

[16] J.R. Pollack, T. Sørlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.L. Børresen-Dale, and P.O. Brown, "Microarray Analysis Reveals a Major Direct Role of DNA Copy Number Alteration in the Transcriptional Program of Human Breast Tumors," *Proc. Nat'l Academy of Sciences USA,* vol. 99, pp. 12963-12968, 2002.

[17] A.J. Aguirre, C. Brennan, G. Bailey, R. Sinha, B. Feng, C. Leo, Y. Zhang, J. Zhang, J.D. Gans, N. Bardeesy, C. Cauwels, C. Cordon-Cardo, M.S. Redston, R.A. DePinho, and L. Chin, "High-Resolution Characterization of the Pancreatic Adenocarcinoma Genome," *Proc. Nat'l Academy of Sciences USA,* vol. 101, pp. 9067-9072, 2004.

[18] D. Tsafrir, M. Bacolod, Z. Selvanayagam, I. Tsafrir, J. Shia, Z. Zeng, H. Liu, C. Krier, R.F. Stengel, F. Barany, W.L. Gerald, P.B. Paty, E. Domany, and D.A. Notterman, "Relationship of Gene Expression and Chromosomal Abnormalities in Colorectal Cancer," *Cancer Research,* vol. 66, pp. 2129-2137, 2006.

[19] J.L. Phillips, S.W. Hayward, Y. Wang, J. Vasselli, C. Pavlovich, H. Padilla-Nash, J.R. Pezullo, B.M. Ghadimi, G.D. Grossfeld, A. Rivera, W.M. Linehan, G.R. Cunha, and T. Ried, "The Consequences of Chromosomal Aneuploidy on Gene Expression Profiles in a Cell Line Model for Prostate Carcinogenesis," *Cancer Research,* vol. 61, pp. 8143-8149, 2001.

[20] G. Tonon, K.K. Wong, G. Maulik, C. Brennan, B. Feng, Y. Zhang, D.B. Khatry, A. Protopopov, M.J. You, A.J. Aguirre, E.S. Martin, Z. Yang, H. Ji, L. Chin, and R.A. DePinho, "High-Resolution Genomic Profiles of Human Lung Cancer," *Proc. Nat'l Academy of Sciences USA,* vol. 102, pp. 9625-9630, 2005.

[21] R. Mao, X. Wang, E.L. Spitznagel, L.P. Frelin, J.C. Ting, H. Ding, J.W. Kim, I. Ruczinski, T.J. Downey, and J. Pevsner, "Primary and Secondary Transcriptional Effects in the Developing Human Down Syndrome Brain and Heart," *Genome Biology,* vol. 6, pp. R107.1-R107.20, 2005.

[22] K.J. Bussey, K. Chin, S. Lababidi, M. Reimers, W.C. Reinhold, W.L. Kuo, F. Gwadry, Ajay, H. Kouros-Mehr, J. Fridlyand, A. Jain, C. Collins, S. Nishizuka, G. Tonon, A. Roschke, K. Gehlhaus, I. Kirsch, D.A. Scudiero, J.W. Gray, and J.N. Weinstein, "Integration Data on DNA Copy Number with Gene Expression Levels and Drug Sensitivities in the NCI-60 Cell Line Panel," *Molecular Cancer Therapeutics,* vol. 5, pp. 853-867, 2006.

[23] W.N. van Wieringen and M.A. van de Wiel, "Nonparametric Testing for DNA Copy Number Induced Differential mRNA Gene Expression," *Biometrics,* vol. 65, pp. 19-29, 2009.

[24] K. Chin, S.D. Vries, J. Fridlyand, P.T. Spellman, R. Roydasgupta, W.-L. Kuo, A. Lapuk, R.M. Neve, Z. Qian, T. Ryder, F. Chen, H. Feiler, T. Tokuyasu, C. Kingsley, S. Dairkee, Z. Meng, K. Chew, D. Pinkel, A. Jain, B.M. Ljung, L. Esserman, D.G. Albertson, F.M. Waldman, and J.W. Gray, "Genomic and Transcriptional Aberrations Linked to Breast Cancer Pathophysiologies," *Cancer Cell,* vol. 10, pp. 529-41, 2006.

[25] H.M. Horlings, C. Lai, D.S.A. Nuyten, H. Halfwerk, P. Kristel, E. Beers, S.A. Joosse, C. Klijn, P.M. Nederlof, M.J.T. Reinders, L.F.A. Wessels, and M.J. Vijver, "Integration of DNA Copy Number Alterations and Prognostic Gene Expression Signatures in Breast Cancer Patients," *Clinical Cancer Research,* vol. 16, pp. 651-663, 2010.

[26] O. Alter, P.O. Brown, and D. Botstein, "Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling," *Proc. Nat'l Academy of Sciences USA,* vol. 97, pp. 10101-10106, Aug. 2000.

[27] T. Hastie, R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, and P. Brown, "'Gene Shaving' as a Method for Identifying Distinct Sets of Genes with Similar Expression Patterns," *Genome Biology,* vol. 1, no. 3, pp. 1-20, 2000.

[28] J.A. Berger, S. Hautaniemi, S.K. Mitra, and J. Astola, "Jointly Analyzing Genes Expression and Copy Number Data in Breast Cancer Using Data Reduction models," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 3, no. 1, pp. 2-16, Jan.-Mar. 2006.

[29] V. Calhoun, J. Liu, and T. Adali, "A Review of Group ICA for fMRI Data and ICA for Joint Inference of Imaging, Genetic, and ERP Data," *NeuroImage,* vol. 45, pp. S163-S172, 2009.

[30] V.D. Calhoun, T. Adali, G.D. Pearlson, and K.A. Kiehl, "Neuronal Chronometry of Target Detection Fusion of Hemodynamic and Event-Related Potential Data," *NeuroImage,* vol. 30, pp. 544-553, 2006.

[31] Y.-P. Wang, "Integration of Gene Expression and Gene Copy Number Variations with Independent Component Analysis," *Proc. IEEE Int'l Conf. Eng. Medicine and Biology Soc. (EMBS),* pp. 5700-5703, 2008.

[32] W. Liebermeister, "Linear Modes of Gene Expression Determined by Independent Component Analysis," *Bioinformatics,* vol. 18, pp. 51-60, 2002.

[33] A. Hyvärinen, "Independent Component Analysis: Algorithms and Applications," *Neural Networks,* vol. 13, nos. 4/5, pp. 411-430, 2000.

[34] A.J. Bell and T.J. Sejnowski, "An Information Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation,* vol. 7, no. 6, pp. 1129-1159, 1995.

[35] A. Hyvärinen and E. Oja, "A Fast Fixed-Point Algorithm for Independent Component Analysis," *Neural Computation,* vol. 9, no. 7, pp. 1483-1492, 1997.

[36] J.F. Cardoso and A. Souloumiac, "Blind Beamforming for Non-Gaussian Signals," *IEE Proc.-F,* vol. 140, no. 6, pp. 362-370, 1993.

[37] J. Sheng, H.-W. Deng, V. Calhoun, and Y.-P. Wang, "Webpage: Integrated Analysis of Gene Expression and Copy Number Data on Gene Shaving Using Independent Component Analysis," http://sites.google.com/site/geneticimaging/file-cabinet, 2011.

[38] P. Wang, Y. Kim, J. Pollack, B. Narasimhan, and R. Tibshirani, "A Method for Calling Gains and Losses in Array CGH Data," *Biostatistics,* vol. 6, pp. 45-58, Jan. 2005.

[39] S. Attoor, E.R. Dougherty, Y. Chen, M.L. Bittner, and J.M. Trent, "Which is Better for cDNA-Microarray-Based Classification: Ratios or Direct Intensities," *Bioinformatics,* vol. 20, pp. 2513-2520, Nov. 2004.

[40] O. Monni, M. Bärlund, S. Mousses, J. Kononen, G. Sauter, M. Heiskanen, P. Paavola, K. Avela, Y. Chen, M.L. Bittner, and A. Kallioniemi, "Comprehensive Copy Number and Gene Expression Profiling of the 17q23 Amplicon in Human Breast Cancer," *Proc. Nat'l Academy of Sciences USA,* vol. 98, pp. 5711-5716, May. 2001.

[41] P. Kauraniemi, S. Hautaniemi, R. Autio, J. Astola, O. Monni, A. Elkahloun, A. Kallioniemi, "Effects of Herceptin Treatment on Global Gene Expression Patterns in HER2-Amplified and Non-Amplified Breast Cancer Cell Lines," *Oncogene,* vol. 23, pp. 1010-1013, Jan. 2004.

[42] J.R. Pollack, T. Sørlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.L. Børresen-Dale, and P.O. Brown, "Microarray Analysis Reveals a Major Direct Role of DNA Copy Number Alteration in the Transcriptional Program of Human Breast Tumors," *Proc. Nat'l Academy of Sciences USA,* vol. 99, pp. 12963-12968, 2002.

Jinhua Sheng (SM'06) received the bachelor's and master's degree in electronic engineering from Hefei University of Technology, China, and the PhD degree in nuclear electronics from University of Science and Technology of China, respectively. He joined China Academy of Telecommunications Technology as an associate professor, and an associate dean of graduate school in 1997. Since 2001, he has served as a postdoctoral fellow, a research associate, and a research scientist at University of Illinois, Rush University, University of Wisconsin, University of Missouri and Indiana University, respectively. He has published about forty papers and been granted two US patents. His research works have been reported in some professional journals or media, such as *Science Daily*, *EurekAlert*, and *First Science eBioNews*. He is an active reviewer for many peer-reviewed journals such as *Medical Engineering and Physics*, *Neurocomputing*, *BioMed Central Bioinformatics*, *IEEE Transactions on Systems, Man and Cybernetics*, *IEEE Transactions on Signal Processing*, *EURASIP Journal on Advances in Signal Processing*, etc., and some international conferences. His research interests include image processing, medical imaging, nuclear electronics, bioinformatics, and genomic signal processing. He is a senior member of the IEEE and a member of Senior Member Review Panel Meeting of the IEEE.

Hong-Wen Deng received the bachelor's degree in ecology and environmental biology and the master's degree in ecology and entomology from Peking University. He received the master's degree in mathematical statistics and the PhD degree in quantitative genetics from the University of Oregon. He was a postdoctoral fellow in the Human Genetics Center at the University of Texas in Houston where he conducted postdoctoral research in molecular and statistical population/quantitative genetics. He also served as a Hughes fellow in the Institute of Molecular Biology at the University of Oregon. He previously served as a professor of medicine and biomedical sciences at Creighton University Medical Center, a professor of orthopaedic surgery and basic medical science, and the Franklin D. Dickson/Missouri endowed chair in orthopaedic surgery at the School of Medicine of University of Missouri-Kansas City. He is currently the chair of the Tulane Biostatistics Department and the director of the Center of Bioinformatics and Genomics. He is the holder of multiple NIH RO1 awards and recipients of multiple honors for his research. He has published more than 300 peer-reviewed articles, 10 book chapters, 3 books. His area of interest is in the genetics of osteoporosis and obesity.

Vince D. Calhoun received the bachelor's degree in electrical engineering from the University of Kansas, Lawrence, Kansas, in 1991, master's degree in biomedical engineering and information systems from Johns Hopkins University, Baltimore, in 1993 and 1996, respectively, and the PhD degree in electrical engineering from the University of Maryland Baltimore County, Baltimore, in 2002. He worked as a senior research engineer at the psychiatric neuroimaging laboratory at Johns Hopkins from 1993 until 2002. He then served as the director of medical image analysis at the Olin Neuropsychiatry Research Center and as an associate professor at Yale University. He is currently the chief technology officer and the director of Image Analysis and MR Research at the Mind Research Network and is a professor in the Departments of Electrical and Computer Engineering (primary), Neurosciences, Psychiatry and Computer Science at the University of New Mexico. He is the author of more than 160 full journal articles and more than 300 technical reports, abstracts, and conference proceedings. Much of his career has been spent on the development of data driven approaches for the analysis of brain imaging data. He has won more than $18 million in NSF and NIH grants on the incorporation of prior information into independent component analysis (ICA) for functional magnetic resonance imaging, data fusion of multimodal imaging and genetics data, and the identification of biomarkers for disease. He is a chartered grant reviewer for NIH. He has organized workshops and special sessions at multiple conferences. He is currently serving on the IEEE Machine Learning for Signal Processing (MLSP) technical committee and previously served as the general chair of the 2005 meeting. He is a reviewer for many journals and is on the editorial board of the Human Brain Mapping and Neuroimage journals. He is a senior member of the IEEE, the Organization for Human Brain Mapping, the International Society for Magnetic Resonance in Medicine, and the American College of Neuropsychopharmacology.

Yu-Ping Wang (SM'06) received the BS degree in applied mathematics from Tianjin University, China, in 1990, and the MS degree in computational mathematics and the PhD degree in communications and electronic systems from Xi'an Jioatong University, China, in 1993 and 1996, respectively. After his graduation, he had visiting positions at the Center for Wavelets, Approximation, and Information Processing of the National University of Singapore and Washington University Medical School in St. Louis. From 2000 to 2003, he worked as a senior research engineer at Perceptive Scientific Instruments, Inc., and then Advanced Digital Imaging Research, LLC, Houston, Texas. In the Fall of 2003, he returned to academia as an assistant professor of computer science and electrical engineering at the University of Missouri-Kansas City. He is currently an associate professor of Biomedical Engineering and Biostatistics at Tulane University and a member of Tulane Center of Bioinformatics and Genomics and Tulane Cancer Center. His research interests lie in the interdisciplinary biomedical imaging and bioinformatics areas, where he has about 100 publications. He has served on numerous program committees and NSF/NIH review panels. He was a guest editor for the *Journal of VLSI Signal Processing Systems* on a special issue on genomic signal processing. He is a member of Machine Learning for Signal Processing technical committee of the IEEE Signal Processing Society. He is a senior member of the IEEE.