Biomarker Insights

METHODOLOGY

LSOSS: Detection of Cancer Outlier Differential Gene Expression

Yupeng Wang[1,2] and Romdhane Rekaya[1,2,3]

[1]Department of Animal and Diary Science, University of Georgia, Athens, GA 30602, USA. [2]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA. [3]Department of Statistics, University of Georgia, Athens, GA 30602, USA. Corresponding author email: wyp1125@uga.edu

Abstract: Detection of differential gene expression using microarray technology has received considerable interest in cancer research studies. Recently, many researchers discovered that oncogenes may be activated in some but not all samples in a given disease group. The existing statistical tools for detecting differentially expressed genes in a subset of the disease group mainly include cancer outlier profile analysis (COPA), outlier sum (OS), outlier robust *t*-statistic (ORT) and maximum ordered subset *t*-statistics (MOST). In this study, another approach named Least Sum of Ordered Subset Square *t*-statistic (LSOSS) is proposed. The results of our simulation studies indicated that LSOSS often has more power than previous statistical methods. When applied to real human breast and prostate cancer data sets, LSOSS was competitive in terms of the biological relevance of top ranked genes. Furthermore, a modified hierarchical clustering method was developed to classify the heterogeneous gene activation patterns of human breast cancer samples based on the significant genes detected by LSOSS. Three classes of gene activation patterns, which correspond to estrogen receptor (ER)+, ER− and a mixture of ER+ and ER−, were detected and each class was assigned a different gene signature.

Keywords: differential gene expression, cancer, outlier

Biomarker Insights 2010:5 69–78

This article is available from http://www.la-press.com.

Introduction

The most widely used method for detecting differential gene expression in comparative microarray studies is the two-sample *t*-statistic. A gene is determined to be significant if the absolute *t*-value exceeds a certain threshold *c*, which is usually determined by its corresponding *P*-value or false discovery rate. Recently, Tomlins et al[1] introduced the cancer outlier profile analysis (COPA) method for detecting cancer genes which are differentially expressed in a subset of disease samples. Heterogeneous patterns of oncogene activation were observed in the majority of cancer types considered in their studies. Thereafter, several further studies in this direction have been proposed. Tibshirani and Hastie[2] introduced the outlier sums (OS) method, Wu[3] proposed the outlier robust *t*-statistic (ORT), and Lian[4] introduced the maximum ordered subset *t*-statistics (MOST) procedure.

In this study, a simple statistical test named Least Sum of Ordered Subset Square *t*-statistic (LSOSS) is proposed for detecting cancer outlier differential gene expression. The performance of LSOSS was compared to existing procedures using both simulated and real data sets. Furthermore, we extended previous studies by classifying heterogeneous gene activation patterns of human breast cancer.

Existing statistical methods

Assuming case-control microarray data were generated for detecting differentially expressed genes consisting of *n* samples from a normal group and *m* samples from a cancer group. Let x_{ij} be the expression value for gene $i = (1, 2, ..., p)$ and sample $j = (1, 2, ..., n)$ in the normal group and y_{ij} be the expression value for gene $i = (1, 2, ..., p)$ and sample $j = (1, 2, ..., m)$ in the cancer group. In this study, and without loss of generality, we are only interested in 1-sided tests where the activated genes from cancer samples are over-expressed or up-regulated.

The two-condition *t*-statistic for gene *i* is defined by:

$$t_i = \frac{\bar{y}_i - \bar{x}_i}{s_i}$$

where $\bar{y}_i$ is the mean expression value in cancer samples, $\bar{x}_i$ is the mean expression value in normal samples for gene i and s_i is the pooled standard error estimate given by:

$$s_i^2 = \frac{\sum_{1\le j\le n}(x_{ij}-\bar{x}_i)^2 + \sum_{1\le j\le m}(y_{ij}-\bar{y}_i)^2}{n+m-2}.$$

The *t*-statistic is powerful when most cancer samples are activated.

Tomlins et al[1] defines the COPA statistic as

$$copa_i = \frac{q_r(\{y_{ij} : 1\le j\le m\}) - med_i}{mad_i}$$

Where $q_r(.)$ is the *rth* percentile of the expression data, and med_i is the median expression value for all samples

$$med_i = median(\{x_{ij} : 1 \le j \le n\}, \{y_{ij} : 1 \le j \le m\}),$$

and mad_i is the median absolute deviation of expression values in all samples and is given by:

$$mad_i = 1.4826 \times median(\{(x_{ij} - med_i) : 1 \le j \le n\}, \{(y_{ij} - med_i) : 1 \le j \le m\}).$$

The COPA statistic uses a fixed *rth* sample percentile, which is determined by users. This limitation was overcome by the OS statistic[2] defined by:

$$OS_i = \frac{\sum_{y_{ij}\in R_i}(y_{ij} - med_i)}{mad_i}$$

where $R_i = \{y_{ij} : y_{ij} > q_{75}(\{x_{ij} : 1 \le j \le n\}, \{y_{ij} : 1 \le j \le m\}) + IQR(\{x_{ij} : 1 \le j \le n\}, \{y_{ij} : 1 \le j \le m\})\}$ and IQR(•) is the inter-quantile range of the expression data

$$IQR(\{x_{ij} : 1 \le j \le n\}, \{y_{ij} : 1 \le j \le m\})\} = q_{75}(\{x_{ij} : 1 \le j \le n\}, \{y_{ij} : 1 \le j \le m\}) - q_{25}(\{x_{ij} : 1 \le j \le n\}, \{y_{ij} : 1 \le j \le m\}).$$

Wu[3] modified the OS statistic by proposing the ORT statistic which consists mainly in changing the definition of R_i as:

$$R_i = \{y_{ij} : y_{ij} > q_{75}(\{x_{ij} : 1 \le j \le n\}) + IQR(\{x_{ij} : 1 \le j \le n\})\}.$$

and replacing med_i in OS by med_{ix}, which is the median expression value in normal samples. Further, mad_i was replaced by

$$mad_i' = 1.4826 \times median(\{(x_{ij} - med_{ix:}) : 1 \le j \le n\}, \{(y_{ij} - med_{iy}) : 1 \le j \le m\}),$$

where med_{iy} is the median expression value in cancer samples. Lian[4] argued that OS and ORT statistics used arbitrary outliers and proposed the MOST statistic which consider all possible values for outlier thresholds. The MOST procedure requires cancer sample expression data be sorted in descending order and the following statistic calculated:

$$MOST_i = \max_{1 \le k \le m} \left(\frac{\sum_{1 \le j \le k} (y_{ij} - med_{ix})}{mad_i'} - \mu_k \right) / \delta_k .$$

μ_k and δ_k are obtained from the order statistics of m samples generated from a standard normal distribution and are used to make different values of the statistic comparable for different values of k.

Methods

The least sum of ordered subset variance *t*-statistic

In our proposed method, least sum of ordered subset square *t*-statistic (LSOSS), mean expression values in normal and cancer samples were considered instead of median expression values. Our hypothesis was that if outliers are present among cancer samples, the distribution of gene expression values in cancer samples will have two peaks. The higher peak corresponds to activated samples while the lower peak indicates inactivated samples. Consequently, this outlier issue can be addressed through the idea of detecting a "change point" or "break point" in the ordered gene expression values of the cancer group. A model related to fitting least squares should be effective for this goal. For each gene, an optimal change point in its expression can be detected and could be used to investigate potential outliers in cancer samples. To this end, we propose the Least Sum of Ordered Subset Square *t*-statistic (LSOSS). The general idea of LSOSS is to use the sum of squares of two ordered subsets of cancer samples to estimate the square sum of the *t*-statistic and to use the mean value of the appealing subset of cancer samples to estimate the mean value of cancer samples of the *t*-statistic.

The proposed LSOSS method involves the following steps:

a) For each gene i, the expression levels in cancer samples are sorted in descending order and then divided into two subsets:

$$S_{ik1} = \{y_{ij} : 1 \le j \le k\},$$
$$S_{ik2} = \{y_{ij} : k+1 \le j \le m\}.$$

b) For the two subsets, the mean and sum of squares for each gene i are calculated:

$$\bar{y}_{S_{ik1}} = mean(\{y_{ij} : 1 \le j \le k\}),$$
$$\bar{y}_{S_{ik2}} = mean(\{y_{ij} : k+1 \le j \le m\}),$$
$$SS_{S_{ik1}} = \sum_{1 \le j \le k} (y_{ij} - \bar{y}_{S_{ik1}})^2,$$
$$SS_{S_{ik2}} = \sum_{k+1 \le j \le k} (y_{ij} - \bar{y}_{S_{ik2}})^2.$$

The only issue left to be solved is the value k that divided the two subsets. For that purpose an exhaustive search was implemented for all possible values ranging from 1 to m-1. The optimum value of k is obtained by minimizing the pooled sum of squares for cancer samples given by:

$$\arg \min_{1 \le k \le m-1} (SS_{S_{ik1}} + SS_{S_{ik2}}).$$

Let s_{ix}^2 be the sum of squares for normal samples given by:

$$s_{ix}^2 = \sum_{1 \le j \le n} (x_{ij} - \bar{x}_i)^2.$$

The pooled standard error estimated for gene i is defined by

$$s_i^2 = \frac{s_{ix}^2 + SS_{S_{ik1}} + SS_{S_{ik2}}}{n+m-2}.$$

c) The LSOSS statistic for declaring a gene i with outlier differential expression in case samples is computed as:

$$LSSV_i = k\frac{\bar{y}_{S_{ik1}} - \bar{x}_i}{s_i}$$

($LSSV_i = (m-k)(\bar{y}_{S_{ik2}} - \bar{x}_i)/s_i$, if repressed gene expression is of interest), where k could be interpreted as the number of outlier samples for gene i.

A modified hierarchical clustering method for classification of heterogeneous gene activation patterns of human breast cancer

We developed a modified hierarchical clustering method for classification of heterogeneous gene activation patterns of human breast cancer samples. 100 permutations were conducted in order to assign a P-value for each gene. The top d genes detected by LSOSS, at the P-value <0.05, were selected for further analysis. For each gene i, the cancer samples that were selected as outliers were marked by 1 and the rest were marked by 0:

$$y_{iw}' = \begin{cases} 1, & \text{if gene } i \text{ has an outlier in sample } w \\ 0, & \text{otherwise,} \end{cases}$$
$$1 \le w \le m.$$

Thus, each cancer sample w can be represented by a vector with a rank d consisting of 0 or 1:

$$z_w = (y_{iw}', 1 \le i \le d).$$

For each cancer sample, the number of 1's indicates the number of genes with outlier expression in that sample compared to other case samples. The similarity between any two cancer samples w and v was denoted by the number of common outlier expression, which can be obtained by counting the number of 1's computed by $\mathbf{z}_w \cdot \mathbf{z}_v^T$. Then, a hierarchical clustering method was adopted to cluster cancer samples. A bootstrap re-sampling method with 5000 replicates was used to assign a P-value to each sub tree of the clustering. The common outliers in a sub-tree with a P-value < 0.05 were highlighted. Then cancer samples were re-ordered according to the proposed clustering method. These vectors of re-ordered samples formed a $d \times m$ two-dimension array. We used a color image to display this array.

Results

Simulation studies

Simulation studies were conducted to compare the performance of LSOSS with those of MOST, ORT, OS, COPA and the t-statistic. To this end, the R source code from Lian[4] was used. The simulation was conducted assuming equal number of normal and cancer samples (n = m = 20) and the expression data was generated from a standard normal distribution. Expression for 2000 genes were simulated, of which 1000 genes were assumed to be differentially expressed and their data was generated by adding a constant, u, to their expression in the first k cancer samples.

The receiver operating characteristic (ROC) curve was used for evaluating the performance of the different statistical methods. Figure 1 shows the ROC curves for different combinations of k and u. When k = 10 and u = 2, LSOSS clearly outperforms others methods and was second best when k = 5 or 15 and u = 2. When k = 20 and u = 2, LSOSS was comparable to ORT and better than OS and COPA. When u is decreased to 1 with k = 10, LSOSS is the only method comparable to the t-statistic. LSOSS shows a low sensitivity when k = 2. However, the case where only one or two samples are activated within a large number of cancer samples may be less realistic. Overall, the performance of LSOSS is appealing in terms of detection power.

Application to human breast cancer data

The breast cancer microarray data from West et al[5] is available at http://data.cgt.duke.edu/west.php. The data were normalized by the quantile method[6] and the log transformation of the expression values were used for the following analysis. There are in all 7129 genes

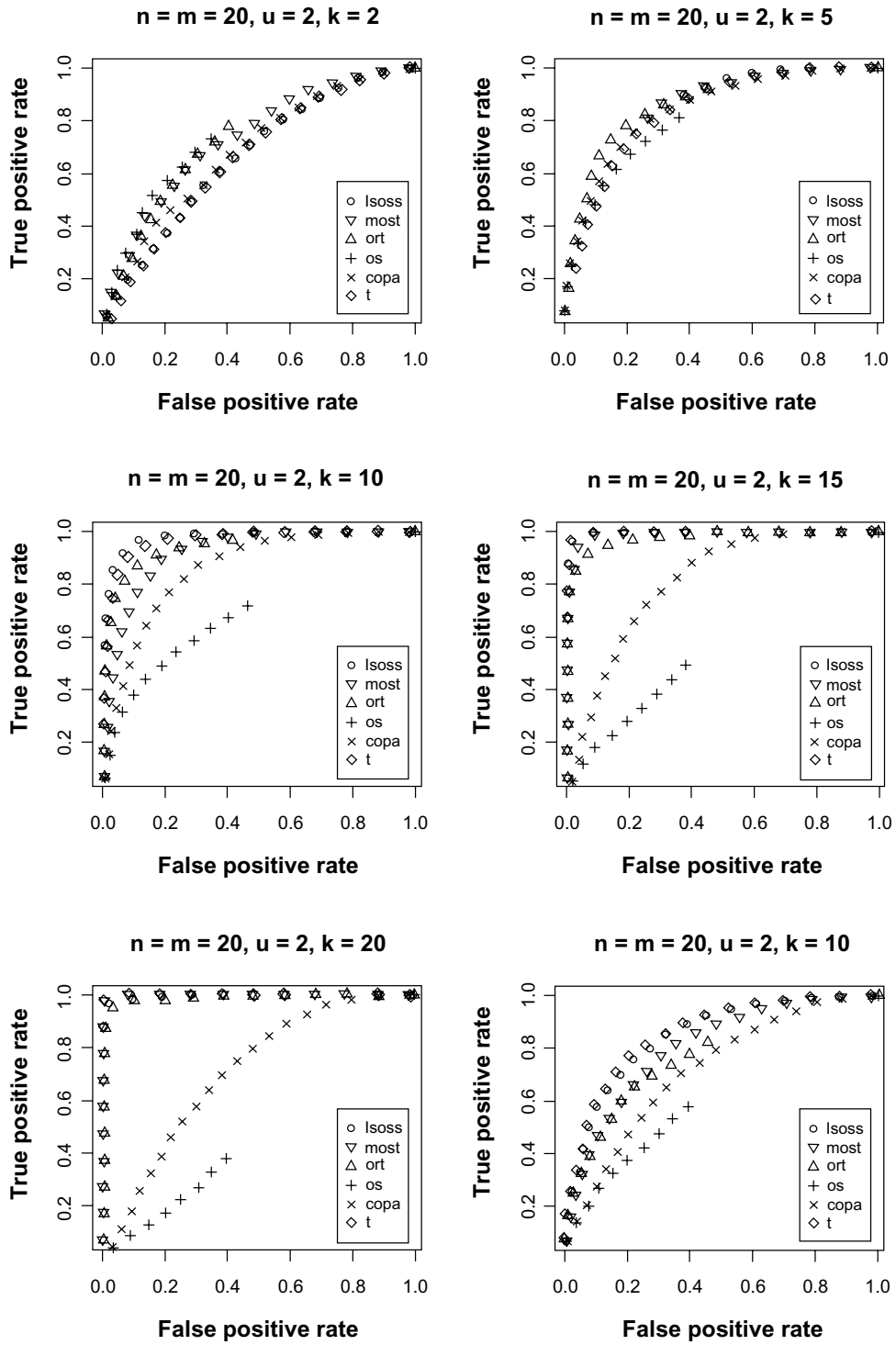

Figure 1. ROC curves comparing different statistical methods.

and 49 tumor samples in this dataset. Among them, 25 tumor samples have negative lymph nodes (LN–) and 24 tumor samples have positive lymph nodes (LN+). We treated the negative LN samples as the control group and the positive LN samples as the cancer group. Genes with expression below a certain threshold (log(10)) in at least 20 samples were removed from the analysis. When evaluating LSOSS based on human breast cancer data, we studied how many genes among the top 25 genes selected separately by different statistical

Table 1. Genes confirmed to be associated with breast cancer that are ranked on the top 25 identified using different cancer outlier detection approaches.

t	COPA	OS	ORT	MOST	LSOSS
ATM	IL6	IL6	ATM	SLC3 A2	KCNH2
FRAP1	LCN2	AGTR1	ERBB4	CGA	NEO1
SOD2		PAK1	THRA	MUC5B	MAGEA3
		CASC3	SMARCA4	CENPB	ENG
			TRADD	HDC	GABRG2
			CTAG1B	IGFBP5	ATM
			AGTR1	FOLR1	NUP88
			CASC3	CKB	CYP3 A7
					PMP22

approaches showed biological relevance in the literature. The numbers of breast cancer related genes identified by existing methods (Table 1) were 8, 8, 4, 3, and 2 for MOST, ORT, OS, the *t*-statistics, and COPA, respectively. However, our proposed method (LSOSS) has identified 9 breast cancer related genes: KCNH2,[7] NEO1,[8] MAGEA3,[9] ENG,[10] GABRG2,[11] ATM,[12] NUP88,[13] CYP3A7[14] and PMP22.[15] Although it should not be treated as a golden standard method for evaluating different statistical tools, this type of analysis generally validates the statistical results and highlights their biological relevance.

Table 2. Genes confirmed to be associated with prostate cancer that are ranked on the top 25 identified using different cancer outlier detection approaches.

t	COPA	OS	ORT	MOST	LSOSS
UBE2E3	ELF1	ELF1	ELF1	ELF1	RB1
BRCA2	CTCF	CAV2	RB1	PAK2	UBE2E3
		CFTR			BMI1
		CTCF			BTG2
					ELF1

Application to human prostate cancer data

To further assess the performance of LSOSS on real data, we downloaded a human prostate cancer dataset.[16] This dataset, generated by the Affymetrix HG-U95av2 chip, consists of 52 prostate tumor samples and 50 normal adjacent samples. The raw data were converted to expression values using a robust multi-array average (RMA) approach.[17] Different statistical methods were run on this dataset and their performances was evaluated by the number of genes among the top 25 genes selected by each approach known to have biological relevance according to the National Cancer Institute Cancer Gene Index, available at https://cabig.nci.nih.gov/inventory/data-resources/cancer-gene-index/. The

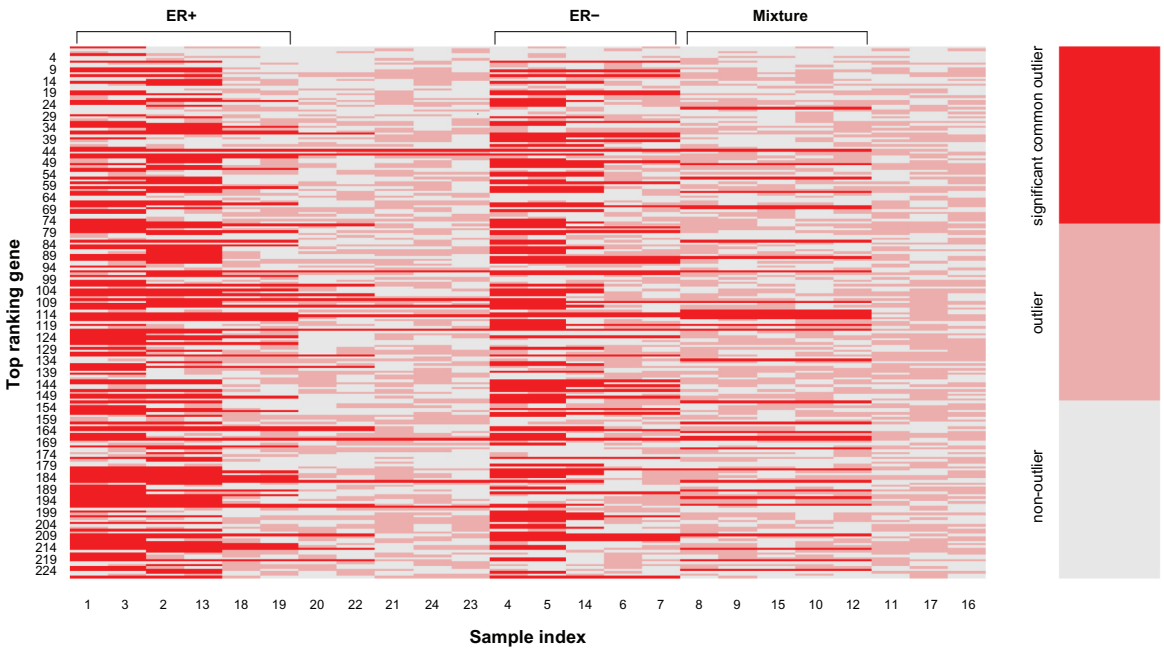

Figure 2. Color image for classification of heterogeneous gene activation patterns of human breast cancer.

Table 3. Classes and biomarkers of heterogeneous gene activation patterns of human breast cancer.

	Classes of heterogeneous activation patterns		
	ER+	**ER–**	**Mixture**
Involved samples[a]	1 (ER+/LN+/Nevins4), 2 (ER+/LN+/Nevins5), 3 (ER+/LN+/Nevins6), 13 (ER+/LN+/Nevins46), 18 (ER+/LN+/Marks206), 19 (ER+/LN+/Marks207)	4 (ER–/LN+/Nevins7), 5 (ER–/LN+/Nevins8), 6 (ER–/LN+/Nevins9), 7 (ER–/LN+/Nevins11), 14 (ER–/LN+/Nevins47)	8 (ER+/LN+/Nevins13), 9 (ER+/LN+/Nevins19), 10 (ER+/LN+/Nevins20), 12 (ER+/LN+/Nevins41), 15 (ER–/LN+/Nevins98)
Gene signatures[b]	24 (CYP3A7) 35 (P2RX4) 37 (DHFR) 38 (UBB) 45 (CTBP1) 47 (RAB35) 48 (RAC1) 53 (SERPINB6) 61 (ROS1) 68 (LRRC14) 77 (SLC35D1) 80 (HOXB8) 84 (STAT5B) 86 (NGF) 97 (MAPK14) 99 (MNAT1) 103 (CYP2D7P1) 105 (MSMB) 107 (ACOT2) 109 (ERBB3) 112 (CASP8) 115 (NPY1R) 116 (GPR68) 117 (FBP1) 118 (THBS4) 122 (BRD2) 125 (KRR1) 128 (SLC39A6) 133 (PKLR) 138 (C11orf58) 151 (MDS1) 157 (PSMC5) 164 (RPL26) 165 (RPL34) 169 (CLPS) 172 (TCEAL1) 183 (GYPE) 185 (SEMA3F) 186 (CYFIP2) 187 (NDST1) 191 (ESR1) 197 (ADH6) 198 (BRD2) 210 (ICAM3) 214 (COX6C) 215 (APBB2) 216 (IRF7) 221 (NA)	7 (TALDO1) 11 (NEO1) 13 (RDBP) 20 (ATM) 21 (CLEC10A) 33 (SRM) 38 (UBB) 39 (APBA2) 41 (SOX3) 45 (CTBP1) 50 (GRK5) 59 (HRK) 69 (DLG3) 78 (TAX1BP1) 80 (HOXB8) 91 (PTPN1) 92 (RPL24) 93 (F2RL1) 97 (MAPK14) 98 (KRTAP5–9) 110 (CALM2) 115 (NPY1R) 116 (GPR68) 120 (ZNF138) 122 (BRD2) 125 (KRR1) 133 (PKLR) 144 (ADAM3B) 146 (ERG) 148 (MYOD1) 151 (MDS1) 156 (SMPD1) 158 (SFTPD) 165 (RPL34) 169 (CLPS) 177 (PPA2) 182 (CTRL) 192 (BCL2L1) 202 (GNB2L1) 210 (ICAM3) 211 (FGFR2) 212 (IL8RB) 228 (KRT4)	27 (GYPA) 45 (CTBP1) 51 (WNT5 A) 57 (PTHLH) 63 (COPS6) 69 (DLG3) 70 (FZD2) 84 (STAT5B) 91 (PTPN1) 97 (MAPK14) 110 (CALM2) 114 (LPO) 115 (NPY1R) 116 (GPR68) 117 (FBP1) 120 (ZNF138) 122 (BRD2) 135 (TCL6) 153 (SLC6A11) 162 (SMG1) 166 (POU2F2) 168 (UBE2H) 169 (CLPS) 173 (MMP11) 182 (CTRL) 187 (NDST1) 191 (ESR1) 194 (FMO1) 197 (ADH6) 210 (ICAM3) 216 (IRF7) 221 (NA) 225 (ASGR2)

Notes: [a]Data are shown in the format of "sample index (sample name)"; [b]Data are shown in the format of "gene ranking (gene symbol)".

Table 4. Classification of the cancer samples lacking significant common outliers.

Samples	Probability			Predicted class
	ER+	ER–	Mixture	
11 (ER+/LN+/Nevins40)	0.328	0.326	0.346	Mixture
16 (ER–/LN+/Nevins99)	0.329	0.332	0.339	Mixture
17 (ER+/LN+/Marks205)	0.348	0.328	0.324	ER+
20 (ER+/LN+/Marks208)	0.333	0.304	0.363	Mixture
21 (ER-/LN+/Marks214)	0.307	0.362	0.331	ER–
22 (ER-/LN+/Marks215)	0.337	0.288	0.375	Mixture
23 (ER-/LN+/Marks216)	0.357	0.261	0.382	Mixture
24 (ER-/LN+/Marks217)	0.300	0.304	0.396	Mixture

comparison of these different statistical approaches is summarized in Table 2. LSOSS, which identifies 5 prostate cancer related genes RB1,[18] UBE2E3,[19] BMI1,[20] BTG2[21] and ELF1,[22] was the best approach with this dataset.

Classification of heterogeneous gene activation patterns of human breast cancer

Breast cancer is a heterogeneous disease.[23,24] Although a number of candidate cancer outliers were identified by existing tools, the heterogeneous gene activation patterns of cancer samples were not addressed after the usage of such methods. LSOSS was applied to the human breast cancer data set from West et al.[5] At a *P*-value cutoff of 0.05, 228 genes were selected for further analysis. The hierarchical clustering method described in the Methods section was then implemented. Three main classes of heterogeneous activation patterns of human breast cancer were observed (Fig. 2). The samples and common outliers in each class are shown in Table 3. Interestingly, we found that the first class consists of 6 ER+ samples, the second class consists of 5 ER– samples, and the third class is a mixture of 4 ER+ and 1 ER– samples. The common outlier genes in each class are regarded as its genetic signature. It is worth noting that although some genes may be part of the genetic signature of different classes of cancer samples, each class has a unique gene signature. For the remaining 8 cancer samples without significant common outliers, their classes were assigned according to their coverage of the gene signatures for different classes (Table 4). Among them, 6 were classified into the mixture group and two others were classified into ER+ and ER– groups.

Discussion and Conclusions

Unraveling the heterogeneous patterns of cancer samples is an important goal in medical research, especially for clinical diagnosis and the molecular understanding of cancer mechanisms. The heterogeneous patterns of oncogene activation have been well studied and several useful statistical tools have been proposed. LSOSS is a reasonable model to detect cancer outlier differential gene expression. For each gene, LSOSS tries to find an optimal "change point" in the ordered expression values of cancer samples. If one gene is expressed heterogeneously in cancer samples, the variance of gene expression values in cancer samples is overestimated by the *t*-statistic while LSOSS gives an appropriate estimate. Furthermore, LSOSS uses the mean value of the appealing subset instead of the overall mean value of the cancer samples. Thus, LSOSS detects cancer outliers more easily. If one gene is expressed homogeneously in cancer samples, LSOSS still works well because it behaves similarly to the *t*-statistic because the mean values of two subsets are expected to be very close in this case.

However, a single oncogene with heterogeneous expression cannot fully account for the heterogeneous gene activation patterns of cancer samples as the synergic and epistatic effects among multiple oncegenes should not be neglected. Thus, it is necessary to classify cancer samples and assign each class a specific gene signature. This goal, if achieved, will definitely facilitate the understanding of different

underlying pathologies and genetics for heterogeneous cancers. Our proposed scheme could be a useful tool toward this goal. Three classes of heterogeneous gene activation patterns of human breast cancer were detected with specific gene signatures. In addition, these heterogeneous gene activation patterns may be regarded as the signatures for subtypes of human breast cancer. Thus, the procedure presented could also be useful in detecting and classifying breast cancer subtypes. The classification of breast cancer subtypes has been well discussed.[25–28] Our approach, however, differed from previous studies mainly in that the classification is based on different combinational activation patterns of candidate genes instead of clustering their expression values. The detection of specific gene interactions accounting for heterogeneous gene activation patterns of cancers is our next goal in this direction.

Acknowledgements

We thank Jamie Williams for critical reading of the manuscript. This study was supported in part by resources and technical expertise from the University of Georgia Research Computing Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer.

Disclosure

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Tomlins SA, Rhodes DR, Perner S, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*. 2005; 310:644–8.
2. Tibshirani R, Hastie T. Outlier sums for differential gene expression analysis. *Biostatistics*. 2007;8:2–8.
3. Wu B. Cancer outlier differential gene expression detection. *Biostatistics*. 2007;8:566–75.
4. Lian H. MOST: detecting cancer differential gene expression. *Biostatistics*. 2008;9:411–8.
5. West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A*. 2001;98:11462–7.
6. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics*. 2003;19: 185–93.
7. Kuznetsova EB, Kekeeva TV, Larin SS, et al. Novel methylation and expression markers associated with breast cancer. *Mol Biol (Mosk)*. 2007;41:624–33.
8. Lee JE, Kim HJ, Bae JY, et al. Neogenin expression may be inversely correlated to the tumorigenicity of human breast cancer. *BMC Cancer*. 2005; 5:154.
9. Gaugler B, van den Eynde B, et al. Human gene MAGE-3 codes for an antigen recognized on a melanoma by autologous cytolytic T lymphocytes. *J Exp Med*. 1994;179:921–30.
10. Gómez-Esquer F, Agudo D, Martínez-Arribas F, Nuñez-Villar MJ, Schneider J. mRNA expression of the angiogenesis markers VEGF and CD105 (endoglin) in human breast cancer. *Anticancer Res*. 2004;24: 1581–5.
11. Garib V, Lang K, Niggemann B, Zänker KS, Brandt L, Dittmar T. Propofol-induced calcium signalling and actin reorganization within breast carcinoma cells. *Eur J Anaesthesiol*. 2005;22:609–15.
12. Ye C, Cai Q, Dai Q, et al. Expression patterns of the ATM gene in mammary tissues and their associations with breast cancer survival. *Cancer*. 2007; 109:1729–35.
13. Schneider J, Linares R, Martínez-Arribas F, et al. Developing chick embryos express a protein which shares homology with the nuclear pore complex protein Nup88 present in human tumors. *Int J Dev Biol*. 2004;48: 339–42.
14. Calaf GM, Roy D. Human drug metabolism genes in parathion-and estrogen-treated breast cells. *Int J Mol Med*. 2007;20:875–81.
15. Kunz-Schughart LA, Heyder P, Schroeder J, Knuechel R. A heterologous 3-D coculture model of breast tumor cells and fibroblasts to study tumor-associated fibroblast differentiation. *Exp Cell Res*. 2001; 266:74–86.
16. Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*. 2002;1:203–9.
17. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4:249–64.
18. Cooney KA, Wetzel JC, Merajver SD, Macoska JA, Singleton TP, Wojno KJ. Distinct regions of allelic loss on 13q in prostate cancer. *Cancer Res*. 1996;56:1142–5.
19. Bull JH, Ellison G, Patel A, et al. Identification of potential diagnostic markers of prostate cancer and prostatic intraepithelial neoplasia using cDNA microarray. *Br J Cancer*. 2001;84:1512–9.
20. Berezovska OP, Glinskii AB, Yang Z, Li XM, Hoffman RM, Glinsky GV. Essential role for activation of the Polycomb group (PcG) protein chromatin silencing pathway in metastatic prostate cancer. *Cell Cycle*. 2006;5:1886–901.
21. Ficazzola MA, Fraiman M, Gitlin J, et al. Antiproliferative B cell translocation gene 2 protein is down-regulated post-transcriptionally as an early event in prostate carcinogenesis. *Carcinogenesis*. 2001;22:1271–9.
22. Takai N, Miyazaki T, Nishida M, Nasu K, Miyakawa I. The significance of Elf-1 expression in epithelial ovarian carcinoma. *Int J Mol Med*. 2003;12: 349–54.
23. Bertucci F, Birnbaum D. Reasons for breast cancer heterogeneity. *J Biol*. 2008;7:6.
24. Anderson WF, Matsuno R. Breast cancer heterogeneity: a mixture of at least two main types? *J Natl Cancer Ins*. 2006;98:948–51.
25. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406:747–52.
26. Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001;98:10869–74.

27. Sørlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*. 2003;100:8418–23.

28. Kapp AV, Jeffrey SS, Langerød A, et al. Discovery and validation of breast cancer subtypes. *BMC Genomics*. 2006;7:231.