Biostatistics (2007), **8**, 3, *pp.* 566–575
doi:10.1093/biostatistics/kxl029
Advance Access publication on October 4, 2006

Cancer outlier differential gene expression detection

BAOLIN WU

Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, Minneapolis, MN 55455, USA
baolin@biostat.umn.edu

SUMMARY

We study statistical methods to detect cancer genes that are over- or down-expressed in some but not all samples in a disease group. This has proven useful in cancer studies where oncogenes are activated only in a small subset of samples. We propose the outlier robust *t*-statistic (ORT), which is intuitively motivated from the *t*-statistic, the most commonly used differential gene expression detection method. Using real and simulation studies, we compare the ORT to the recently proposed cancer outlier profile analysis (Tomlins *and others*, 2005) and the outlier sum statistic of Tibshirani and Hastie (2006). The proposed method often has more detection power and smaller false discovery rates. Supplementary information can be found at http://www.biostat.umn.edu/∼baolin/research/ort.html.

Keywords: Cancer outlier profile analysis; Differential gene expression detection; Microarray; Robust; T-statistic.

1. INTRODUCTION

Recently, Tomlins *and others* (2005) have proposed the "cancer outlier profile analysis" (COPA) method for detecting cancer genes which show increased expressions in a subset of disease samples. They argue that in the majority of cancer types, oncogene has heterogeneous activation patterns; traditional analytical methods, for example, *t*-statistic, which search for common activation of genes across a class of cancer samples, will fail to find such oncogene expression profiles. Instead, we should search for overexpression only in a subset of cases. Through applications to public cancer microarray data sets, they have shown that the proposed COPA can perform better than the commonly used *t*-statistic.

More recently, Tibshirani and Hastie (2006) proposed the outlier sum (OS) statistic to detect cancer gene outlier expressions. The OS and COPA are similarly defined using robust location and scale estimates of the gene expression values (more details in Section 2). Through simulation studies and applications, they have shown that the OS can perform better than the COPA, for example, having smaller false discovery rates (Benjamini and Hochberg, 1995).

In this paper, we consider the statistical methods to detect cancer genes with a subset of over- or down-expressed outlier disease samples. Many methods have been proposed to detect differentially expressed genes (see, e.g. Dudoit *and others*, 2002; Troyanskaya *and others*, 2002). Among them, the *t*-statistic is the most commonly used method. We will discuss several problems associated with the *t*-statistic for cancer gene outlier expression detection, which will motivate the development of the outlier robust *t*-statistic (ORT). We will further establish the connection of the OS, COPA, and ORT statistics to the *t*-statistic from a robustness consideration. Through simulation studies and applications to a

public breast cancer microarray data, we empirically evaluate and compare the different outlier detection statistics.

2. Statistical methods

Consider a 2-class, for example, cancer/normal tissues, microarray data. Let x_{ij} be the observed expression values for samples $i = 1, \ldots, n$ and genes $j = 1, \ldots, p$. Without loss of generality, assume that the first n_1 samples are from the normal group and the last n_2 samples are from the cancer group, where $n = n_1 + n_2$. In the following discussion, we assume that the outlier disease samples are overexpressed. Similar arguments will carry through to detect genes with down-expressed outlier disease samples.

The 2-sample t-test statistic for gene j is defined as

$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{s_j}\sqrt{\frac{n_1 n_2}{n}}, \quad \text{where } \bar{x}_{1j} = \frac{\sum_{i \leqslant n_1} x_{ij}}{n_1}, \ \ \bar{x}_{2j} = \frac{\sum_{i > n_1} x_{ij}}{n_2}. \tag{2.1}$$

Here s_j is the pooled standard error estimate for gene j

$$s_j^2 = \frac{\sum_{i \leqslant n_1}(x_{ij} - \bar{x}_{1j})^2 + \sum_{i > n_1}(x_{ij} - \bar{x}_{2j})^2}{n - 2}.$$

The t-statistic is based on the assumption that all disease samples are overexpressed. While in cancer gene outlier analysis, only a subset of the disease samples are assumed to be overexpressed. Intuitively, we want to make inference only using those overexpressed samples (outliers).

In the following, we first study the recently proposed COPA method (Tomlins *and others*, 2005) and the OS statistic (Tibshirani and Hastie, 2006) for detecting cancer gene outliers. We will make some intuitive connections between these 2 outlier detection statistics and the t-statistic. The t-statistic (2.1) will be studied from a robustness (against outlier) perspective, which shows its dependence on all disease samples and the inappropriate variance estimate. We then propose an ORT to remove the "all disease samples" dependence and appropriately reduce the outlier effects on the variance.

2.1 *t-statistic, COPA and OS*

Note that we can equivalently write the t-statistic (2.1) as

$$t_j = \frac{\sqrt{n_1 n_2 (n-2)}}{n} \frac{\text{avg}_{i > n_1}(x_{ij} - \text{avg}_{1j})}{\sqrt{\text{avg}\{(x_{ij} - \text{avg}_{1j})^2_{i \leqslant n_1}, (x_{ij} - \text{avg}_{2j})^2_{i > n_1}\}}}, \tag{2.2}$$

where avg$(\cdot)$ means the sample average; avg_{1j} and avg_{2j} are the normal and disease group sample means. According to our assumption, only a subset of those disease samples ($i > n_1$) is overexpressed. So the $\text{avg}_{i > n_1}(\cdot)$ in the numerator, which sums over all disease samples, will introduce some extra noise. Another problem is the variance estimate, which might overestimate the true value since we already know that there is a subset of outlier disease samples. The COPA and OS statistics address these 2 problems with their different approaches. They are defined as follows:

First, (robustly) standardize the data

$$\tilde{x}_{ij} = \frac{x_{ij} - \text{med}_j}{\text{mad}_j}, \quad i = 1, \ldots, n, \ \ j = 1, \ldots, p, \tag{2.3}$$

where med_j is the median and mad_j is the median absolute deviation of gene j's expression values

$$\text{med}_j = \text{median}_{i=1,\ldots,n}(x_{ij}), \quad \text{mad}_j = 1.4826 \times \text{median}_{i=1,\ldots,n}(|x_{ij} - \text{med}_j|),$$

where the constant 1.4826 makes mad_j approximately equal to the standard error for normally distributed random variables. Here the medians are used due to the robustness consideration.

Let $q_r(\cdot)$ be the rth percentile of the data. The COPA statistic (Tomlins *and others*, 2005) is defined as the rth percentile of the disease samples' standardized expression values $q_r(\tilde{x}_{ij}: i > n_1)$, where the authors have used $r = 75, 90$, or 95. Note that the subtraction and scaling would not change the order of the observed values. So it is easily checked that the COPA statistic is equivalent to

$$q_r(\tilde{x}_{ij}: i > n_1) = \frac{q_r(x_{ij}: i > n_1) - \text{med}_j}{\text{mad}_j}. \tag{2.4}$$

Compared to the t-statistic, intuitively the COPA replaces the normal sample mean $\bar{x}_{1j}$ by the all-sample median med_j, the sample standard error s_j by the median absolute deviation mad_j, and the disease sample mean $\bar{x}_{2j}$ by the rth percentile $q_r(x_{ij}: i > n_1)$. Here, mad_j can be viewed as a scaling factor to make the COPA statistics comparable across different genes.

Immediately, we can see that the COPA statistic might not be efficient, since a fixed rth sample percentile is approximately equivalent to using the information from only one sample. We expect to see improved power if instead we sum over, ideally, all outlier disease samples. The OS statistic (Tibshirani and Hastie, 2006) proposed to replace the rth percentile with a sum over the outlier disease samples identified with some heuristic criterion. The OS statistic is defined as

$$W_j = \sum_{i>n_1} \tilde{x}_{ij} \times I\{\tilde{x}_{ij} > q_{75}(\tilde{x}_{kj}: k = 1, \ldots, n) + \text{IQR}(\tilde{x}_{kj}: k = 1, \ldots, n)\},$$

where $I(\cdot)$ is the indicator function and $\text{IQR}(\cdot)$ calculates the interquartile range

$$\text{IQR}(\tilde{x}_{kj}: k = 1, \ldots, n) = q_{75}(\tilde{x}_{kj}: k = 1, \ldots, n) - q_{25}(\tilde{x}_{kj}: k = 1, \ldots, n).$$

It is commented that values greater than the limit $q_{75} + \text{IQR}$ are defined to be outliers in the usual statistical sense.

Similarly, since the subtraction and scaling would not change the order of the observed values, it is easily checked that the OS statistic is equivalent to

$$W_j = \frac{\sum_{i \in R}(x_{ij} - \text{med}_j)}{\text{mad}_j}, \tag{2.5}$$

where R is the set of "outlier disease samples" defined by the following heuristic criterion:

$$R = \{i > n_1: x_{ij} > q_{75}(x_{kj}: k = 1, \ldots, n) + \text{IQR}(x_{kj}: k = 1, \ldots, n)\}. \tag{2.6}$$

2.2 *Outlier robust t-statistic*

Besides the inefficiency of the COPA statistic owing to its use of a fixed rth sample percentile, a second problem is that the median over all samples, med_j, is not quite the right statistic to replace the normal sample mean, avg_{1j}. It might overestimate the normal group mean owing to the contamination by disease samples if a majority of them have outlier expressions. A more intuitive and appropriate quantity might be, for example, the normal sample median.

Another problem is the median absolute deviation estimation. Since we already know that the disease and normal samples are different, it might not be the best approach to use the overall median as a common estimate for the 2 group medians. Intuitively, it might help to base our estimate on, for example, the group median–centered expression values

$$|x_{ij} - \text{med}_{1j}|, \quad i = 1, \ldots, n_1; \quad |x_{ij} - \text{med}_{2j}|, \quad i = n_1 + 1, \ldots, n,$$

where med_{1j} and med_{2j} are the sample medians for the normal and disease groups

$$\text{med}_{1j} = \text{median}_{i \leqslant n_1}(x_{ij}), \quad \text{med}_{2j} = \text{median}_{i > n_1}(x_{ij}).$$

An intuitive and reasonable estimate for the median absolute deviation might then be, for example,

$$1.4286 \times \text{median}\{|x_{ij} - \text{med}_{1j}|_{i \leqslant n_1}, |x_{ij} - \text{med}_{2j}|_{i > n_1}\} \tag{2.7}$$

which is in spirit very similar to the pooled sample variance estimate

$$\frac{n}{n-2} \times \text{avg}\{(x_{ij} - \text{avg}_{1j})^2_{i \leqslant n_1}, (x_{ij} - \text{avg}_{2j})^2_{i > n_1}\}.$$

In essence, we use the sample median to replace average, and the absolute difference to replace squared difference in order to obtain a more robust variance estimate.

Summarizing previous discussions, we propose the following ORT to detect cancer genes with over-expressed outlier disease samples

$$t^*_j = \frac{\sum_{i \in U_j}(x_{ij} - \text{med}_{1j})}{\text{median}\{|x_{ij} - \text{med}_{1j}|_{i \leqslant n_1}, |x_{ij} - \text{med}_{2j}|_{i > n_1}\}}, \quad j = 1, \ldots, p, \tag{2.8}$$

where U_j is the set of "outlier disease samples" for gene j defined by

$$U_j = \{i > n_1 : x_{ij} > q_{75}(x_{kj} : k = 1, \ldots, n_1) + \text{IQR}(x_{kj} : k = 1, \ldots, n_1)\}. \tag{2.9}$$

Note that here we explicitly calculate the outlying measures using only the normal group samples. We use permutations to estimate the ORT's null distribution and calculate the P-values. For simplicity, we omit those constants in the statistic definition, since they would not affect the significance testing based on the permutations.

In the following, we use simulation studies and applications to a public breast cancer microarray data to empirically evaluate and compare the detection power of previously discussed 4 methods: the t-statistic, COPA, OS, and the proposed ORT.

3. Simulation studies

Simulation studies are conducted to evaluate the power of various outlier detection statistics. We also compare their false discovery rates (Benjamini and Hochberg, 1995).

Suppose we have $n = 25$ normal and disease samples. There are in total $p = 1000$ genes with their expression values simulated from the standard normal distribution. The first gene contains $k = 1, 5, 10, 15, 20, 25$ outlier disease samples with their expression values being added constant $\mu = 2$. For each simulated data, we can calculate the P-value for the first gene, which is the proportion of the other (null) genes with the absolute test statistics bigger than the first gene. The P-values from the simulations can be used to estimate the true/false-positive rates, that is, the sensitivity and 1 − specificity, which are then used to construct the receiver operating characteristic curve for power comparison.

Figure 1 shows the estimated true/false-positive rates based on 1000 simulations. In the extreme situation with only one outlier disease sample ($k = 1$), the OS statistic performs the best, the ORT has comparable performance as the OS, and the t-statistic and COPA have almost no detection power. When increasing to $k = 5$ outlier disease samples, the ORT, OS, and COPA have similar power, all better than the t-statistic. For $k = 10$ outlier disease samples, the ORT performs the best. The detection power of both the ORT and t-statistic increases with more outlier disease samples. While the performance of the

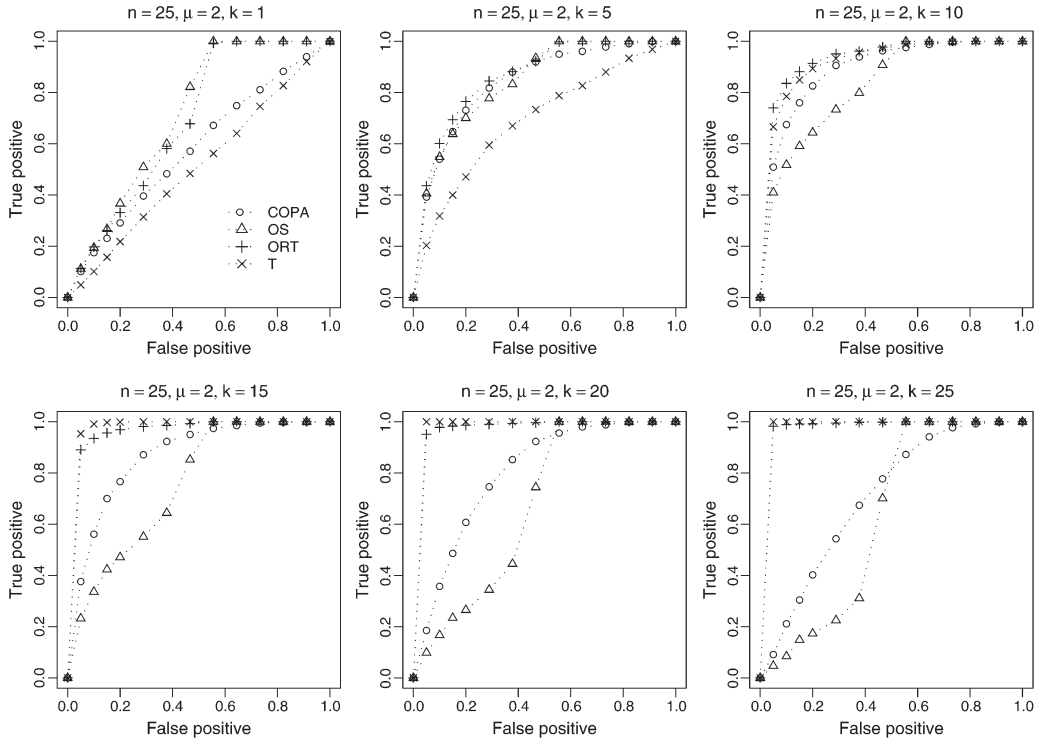

Fig. 1. Detection power estimation based on 1000 simulations. There are $n = 25$ disease and normal samples, and 999 null genes with their expression values simulated from standard normal distribution. The first gene contains a subset of k outlier disease samples with their expression values added constant $\mu = 2$.

COPA and OS decreases a little bit when the outlier disease samples approach the full set ($k = 20, 25$). Overall, the ORT performs the best. It seems to be able to automatically adapt to the unknown number of outlier samples, and combine the strength of both the OS and t-statistic.

Next we evaluate and compare the false discovery rates of the 4 methods based on the simulation. We set $m = 100, 200, 300$ of the $p = 1000$ genes as differentially expressed with $k = 1, 5, 15, 20, 25$ outlier disease samples with their expression values being added constant $\mu = 2$. Figure 2 shows the estimated false discovery rates based on 1000 simulations for $m = 200$ differentially expressed genes. Similar patterns as the true/false-positive rates estimation (see Figure 1) are observed. The ORT has the overall best performance with the smallest false discovery rates.

Very similar patterns have been observed for $m = 100, 300$. We also did the simulation studies for $n = 15, 25$; $k = 1, 3, 6, 9, 12, 15$ or $k = 1, 5, 10, 15, 20, 25$; and $\mu = 1, 2$. We consistently observe that the ORT has the overall best performance. Complete simulation results are available at the supplementary web site (http://www.biostat.umn.edu/~baolin/research/ort.html).

In Section 4, we apply the 4 cancer gene outlier detection statistics to a public breast cancer microarray data and empirically compare their performance.

4. APPLICATION TO THE BREAST CANCER MICROARRAY DATA

The breast cancer microarray data reported by West *and others* (2001) contained the expression levels of 7129 genes from 49 breast tumor samples. Each sample had a binary outcome describing the status of

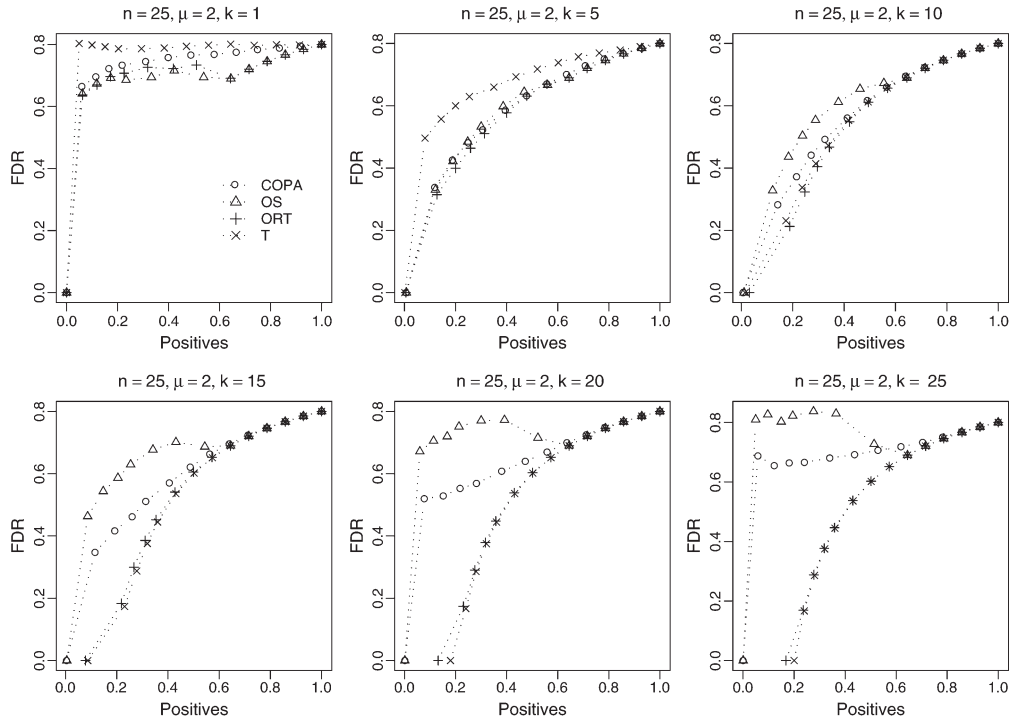

Fig. 2. False discovery rate estimation based on 1000 simulations. There are $n = 25$ disease and normal samples, and $p = 1000$ genes with their expression values simulated from standard normal distribution. The first $m = 200$ genes contain a subset of k outlier disease samples with their expression values being added constant $\mu = 2$. The x-axis is the positive rates: the proportion of genes called significant.

lymph node involvement in breast cancer. Among them, 25 tumor samples had no positive lymph nodes discovered and 24 tumor samples had identifiably positive nodes. The gene expressions, obtained from the Affymetrix human HuGeneFL GeneChip, can be downloaded from http://data.cgt.duke.edu/west.php. We normalize the data using quantile normalization (Bolstad *and others*, 2003), and then log transform the intensities for follow-up statistical analysis. In the cancer gene outlier detection, we treat the negative group as the normal class. We applied the t-statistic, COPA, OS, and the proposed ORT to detect genes with overexpressed disease samples. We rank the genes based on each test statistic. For those top 25 genes identified by each method, we mapped their Affymetrix identifiers to the UniGene cluster identifiers using the Bioconductor (Gentleman *and others*, 2004) annotation package hu6800, which were then used to search for relevant literature in the PubMed. There are in total 13 genes identified that have been studied previously and shown related to breast cancer.

Table 1 lists the confirmed breast cancer–related genes ranked in top 25 for each outlier detection statistic. ORT identified 8 genes, 5 of them were not selected by other statistics. There were 5 genes that were missed by the ORT but identified by the others. Also listed in the table is the ranking of each gene by the 4 test statistics. The genes identified by the OS were ranked generally high by the ORT. Among those genes identified by the ORT, some were ranked low by the OS but relatively higher by the t-statistic, for example, ATM and ERBB4; while several others were ranked low by the t-statistic but relatively higher by the OS, for example, AGTR1 and CASC3. It seems likely that the proposed ORT could combine the

Table 1. *Genes ranked in top* 25 *by the outlier detection statistics and confirmed to be associated with breast cancer in previous studies. The last* 4 *columns also list the ranking of each gene by the* 4 *methods*

Methods	Rank	UniGene ID	Gene name	*t*	COPA	OS	ORT
t	18	Hs.435561	ATM		819	4296	7
	23	Hs.338207	FRAP1		4507	4296	4376
	24	Hs.487046	SOD2		3670	4296	401
COPA	17	Hs.512234	IL6	3447		5	126
	21	Hs.204238	LCN2	4744		4296	4375
OS	5	Hs.512234	IL6	3447	17		126
	14	Hs.477887	AGTR1	2191	98		21
	15	Hs.435714	PAK1	4744	125		32
	16	Hs.350229	CASC3	731	105		22
ORT	7	Hs.435561	ATM	18	819	4296	
	9	Hs.390729	ERBB4	82	1842	1203	
	17	Hs.724	THRA	817	121	69	
	18	Hs.327527	SMARCA4	84	196	55	
	19	Hs.460996	TRADD	380	483	415	
	20	Hs.534310	CTAG1B	1883	292	176	
	21	Hs.477887	AGTR1	3291	98	14	
	22	Hs.350229	CASC3	731	105	16	

strength of both the OS and *t*-statistic (see also Figures 1 and 2 in Section 3). Overall, the ORT had the best detection power.

Figure 3 shows the expression profiles of the 8 genes that were identified by the ORT and confirmed associated with the breast cancer in previous studies. Figure 4 shows the expression profiles of the other 5 confirmed breast cancer–related genes that were missed by the ORT but identified by the other 3 methods. We have added some jittering to the horizontal positions to distinguish among close points. The title lists the gene names. Within the parentheses are those outlier statistics that have ranked the gene in top 25.

5. DISCUSSION

Previous discussions have focused on detecting genes with overexpressed outlier disease samples. The proposed ORT can be adapted to detect cancer genes with down-expressed outlier disease samples as follows:

$$t_j^* = \frac{\sum_{i\in D_j}(x_{ij} - \text{med}_{1j})}{\text{median}\{|x_{ij} - \text{med}_{1j}|_{i\leqslant n_1}, |x_{ij} - \text{med}_{2j}|_{i>n_1}\}}, \quad j = 1, \ldots, p,$$

where D_j is the set of down-expressed "outlier disease samples" for gene j defined by

$$D_j = \{i > n_1 : x_{ij} < q_{25}(x_{kj} : k = 1, \ldots, n_1) - \text{IQR}(x_{kj} : k = 1, \ldots, n_1)\}.$$

Similarly, here we have used the intuition that values less than the limit $q_{25} - \text{IQR}$ are defined to be outliers in the usual statistical sense. When applied to the breast cancer microarray data to detect cancer genes

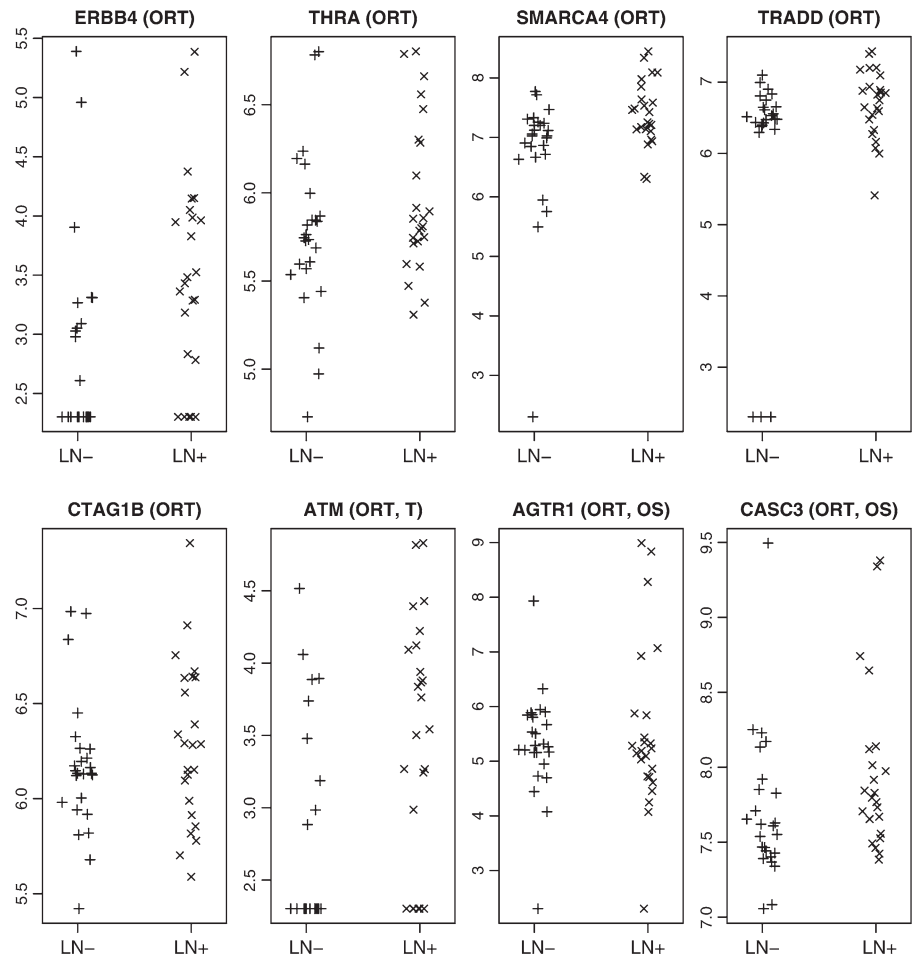

Fig. 3. Cancer gene outlier detection for breast cancer microarray data: plotted are 8 top-ranking genes that were identified by the ORT and confirmed associated with the breast cancer in the literature. The lymph node–negative samples (LN−) serve as the normal group, and we look for outlier samples in the lymph node–positive (LN+) group. We have added some jittering to the horizontal positions to distinguish among close points. The title lists the gene names. Within the parentheses are those outlier statistics that have ranked the gene in top 25.

with down-expressed outlier disease samples, the OS, COPA, and ORT have very similar performance. Overall, ORT has the best detection power. Complete lists of all the identified genes for different methods are available at the supplementary web site (http://www.biostat.umn.edu/∼baolin/research/ort.html).

The proposed ORT is intuitively motivated from the widely used t-statistic with the robustness consideration. Compared to the COPA and OS, ORT more appropriately takes into account the difference between the normal and disease groups, for example, the proper estimation of median absolute deviation (2.7) and the use of normal group median instead of the overall median (2.8). Through simulation studies and application to public cancer microarray data, we have illustrated the competitive performance of the proposed ORT. In this paper, we have focused on comparing 2 groups. The study of multigroup comparisons will be reported in the future.

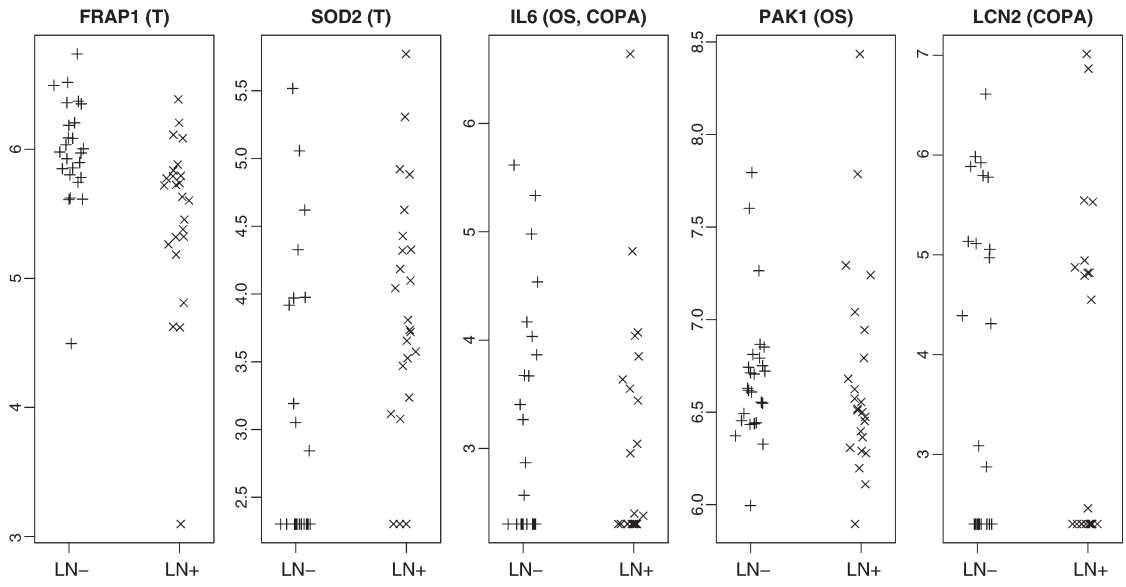

Fig. 4. Cancer gene outlier detection for breast cancer microarray data: plotted are 5 top-ranking genes missed by the ORT but identified by the other 3 methods that were confirmed related to the breast cancer in the literature. The lymph node–negative samples (LN−) serve as the normal group, and we look for outlier samples in the lymph node–positive (LN+) group. We have added some jittering to the horizontal positions to distinguish among close points. The title lists the gene names. Within the parentheses are those outlier statistics that have ranked the gene in top 25.

ACKNOWLEDGMENTS

This research was partially supported by a University of Minnesota artistry and research grant and a research grant from the Minnesota Medical Foundation. *Conflict of Interest:* None declared.

REFERENCES

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* **57**, 289–300.

BOLSTAD, B., IRIZARRY, R., ASTRAND, M. AND SPEED, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193.

DUDOIT, S., YANG, Y. H., CALLOW, M. J. AND SPEED, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.

GENTLEMAN, R., CAREY, V., BATES, D., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J. *and others* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80.

TIBSHIRANI, R. AND HASTIE, T. (2006). Outlier sums for differential gene expression analysis. *Biostatistics* **8**, 2–8.

TOMLINS, S. A., RHODES, D. R., PERNER, S., DHANASEKARAN, S. M., MEHRA, R., SUN, X. W., VARAMBALLY, S., CAO, X., TCHINDA, J., KUEFER, R. *and others* (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648.

TROYANSKAYA, O. G., GARBER, M. E., BROWN, P. O., BOTSTEIN, D. AND ALTMAN, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* **18**, 1454–1461.

WEST, M., BLANCHETTE, C., DRESSMAN, H., HUANG, E., ISHIDA, S., SPANG, R., ZUZAN, H., OLSON, J. A. J., MARKS, J. R. AND NEVINS, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 11462–11467.

[*Received June 15, 2006; revised September 11, 2006; accepted for publication September 29, 2006*]