Multi-class cancer outlier differential gene expression detection
Supplementary Information

Fang Liu, Baolin Wu*

1 Breast cancer microarray data analysis

The breast cancer microarray data (West *et al.*, 2001), obtained from the Affymetrix human HuGeneFL GeneChip, contained the expression levels of 7129 genes from 49 breast tumor samples. The raw data can be downloaded from http://data.cgt.duke.edu/west.php. Each sample had two binary outcomes: the status of lymph node involvement in breast cancer (negative and positive, denoted as LN-/LN+) and the estrogen receptor status (negative and positive, denoted as ER-/ER+). Among them, there are 13 ER+/LN+ tumors, 11 ER-/LN+ tumors, 12 ER+/LN- tumors, and 13 ER-/LN- tumors. We process the data using the quantile normalization (Bolstad *et al.*, 2003) and log transformation for followup statistical analysis. In the cancer gene outlier detection, we treat the ER-/LN- group as the normal class, and apply the F-statistic, the proposed ORF and OF to detect genes with over-expressed disease samples. The genes were ranked based on each test statistic. For the top 25 genes identified by each method, we use the Bioconductor (Gentleman *et al.*, 2004) annotation package **hu6800** to search for related literature in the PubMed .

1.1 Cancer genes with over-expressed outlier disease samples

Table 1 to 3 list the top 25 genes identified by each outlier detection statistic.

1.2 Cancer genes with down-expressed outlier disease samples

Table 4 shows the expression profiles of the identified genes with down-expressed disease samples that have been confirmed associated with the breast cancer in previous studies. Figure 1 shows the expression profiles of these genes.

Table 5 to 6 list the top 25 genes identified by each outlier detection statistic.

*Email: baolin@biostat.umn.edu, phone: (612)624-0647, fax: (612)626-0660

Figure 1: Oncogene outlier detection for breast cancer microarray data: 9 additional top ranking genes that are identified by ORF/OF and shown related to breast cancer in the literature are plotted. The ER-/LN- samples serve as the normal 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 ranked the gene in top 25.

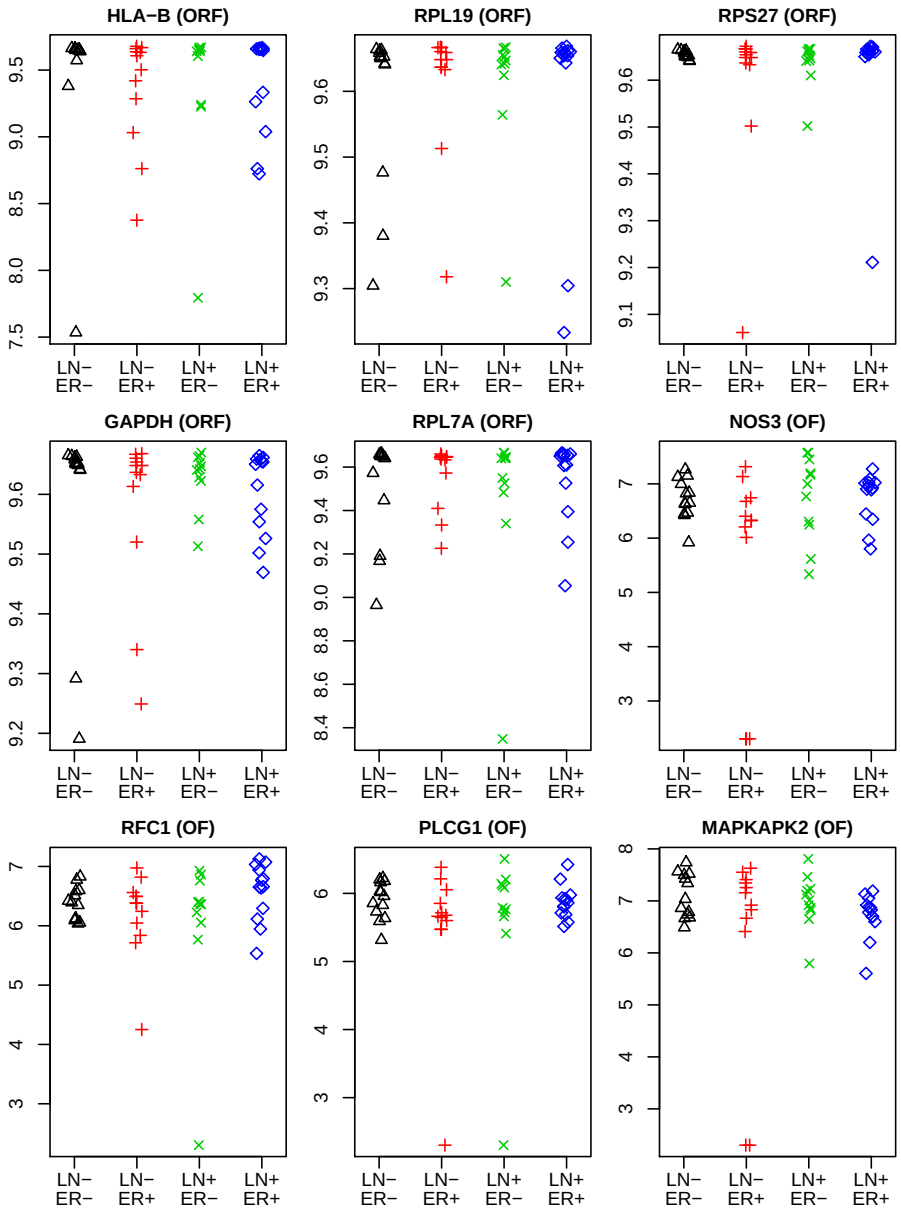

Table 1: Top 25 (over-expressed) genes identified by the outlier robust F-statistic (ORF). Those in bold face font have been studied previously and confirmed associated with breast cancer in the literature.

Ranking	Affymetrix ID	Gene Abbrevations
1	Y08976_at	FEV
2	HG3319-HT3496_s_at	HIRA
3	S75313_at	ATXN3
4	U48936_at	SCNN1G
5	K02882_cds1_s_at	IGHD
6	X96969_at	SLC14A2
7	M87770_at	FGFR2
8	U17579_rna1_at	GHRHR
9	M93119_at	INSM1
10	J03242_s_at	IGF2
11	U07664_at	HLXB9
12	X78817_at	ARHGAP4
13	D13897_rna2_at	PYY
14	M31525_at	HLA-DOA
15	M96739_at	NHLH1
16	U00968_at	SREBF1
17	X71428_at	FUS
18	M24248_at	MYL3
19	D50663_at	DYNLT1
20	X66417_at	CSN3
21	X92518_s_at	HMGA2
22	HG3513-HT3707_at	MYH8
23	M58285_at	NCKAP1L
24	S82198_at	CTRC
25	HG2566-HT4867_at	MAPT

References

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** (2), 185–193.

Gentleman,R., Carey,V., Bates,D., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J., Hornik,K., Hothorn,T., Huber,W., Iacus,S., Irizarry,R., Leisch,F., Li,C., Maechler,M., Rossini,A., Sawitzki,G., Smith,C., Smyth,G., Tierney,L., Yang,J. and Zhang,J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology,* **5** (10), R80.

Table 2: Top 25 genes identified by the outlier F-statistic. Those in bold face font have been studied previously and confirmed associated with breast cancer in the literature.

Ranking	Affymetrix ID	Gene Abbrevations
1	Y08976_at	FEV
2	HG3319-HT3496_s_at	HIRA
3	S75313_at	ATXN3
4	U48936_at	SCNN1G
5	K02882_cds1_s_at	IGHD
6	X96969_at	SLC14A2
7	M87770_at	FGFR2
8	U17579_rna1_at	GHRHR
9	M93119_at	INSM1
10	J03242_s_at	IGF2
11	U07664_at	HLXB9
12	X78817_at	ARHGAP4
13	D13897_rna2_at	PYY
14	M31525_at	HLA-DOA
15	M96739_at	NHLH1
16	U00968_at	SREBF1
17	X71428_at	FUS
18	M24248_at	MYL3
19	D50663_at	DYNLT1
20	X66417_at	CSN3
21	X92518_s_at	HMGA2
22	HG3513-HT3707_at	MYH8
23	M58285_at	NCKAP1L
24	S82198_at	CTRC
25	HG2566-HT4867_at	MAPT

West,M., Blanchette,C., Dressman,H., Huang,E., Ishida,S., Spang,R., Zuzan,H., Olson,John A.,J., Marks,J.R. and Nevins,J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS,* **98** (20), 11462–11467.

Table 3: Top 25 genes identified by the F-statistic. Those in bold face font have been studied previously and confirmed associated with breast cancer in the literature.

Ranking	Affymetrix ID	Gene Abbrevations
1	U41060_at	SLC39A6
2	U27185_at	RARRES1
3	X17059_s_at	NAT1
4	L08044_s_at	TFF3
5	U39840_at	FOXA1
6	U22376_cds2_s_at	MYB
7	M99701_at	TCEAL1
8	X83425_at	BCAM
9	X55037_s_at	GATA3
10	X58072_at	GATA3
11	X13238_at	COX6C
12	X87212_at	CTSC
13	L40401_at	ACOT2
14	X59834_at	GLUL
15	U09770_at	CRIP1
16	U62325_at	APBB2
17	U68385_at	MEIS3P1
18	Z29083_at	TPBG
19	U96113_at	WWP1
20	M31627_at	XBP1
21	U42408_at	LAD1
22	X75861_at	TEGT
23	U03886_at	PNPLA4
24	U11791_at	CCNH
25	M16038_at	LYN

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

Methods	Rank	Affymetrix ID	Gene Abbreviation	F	OF	ORF
F	1	U41060_at	SLC39A6		2272	3345
	9	X55037_s_at	GATA3		6717	6717
	10	X58072_at	GATA3		3034	2132
	20	M31627_at	XBP1		6200	6200
	24	U11791_at	CCNH		6380	6380
OF	5	M93718_at	NOS3	320		465
	8	L24783_at	RFC1	2402		1781
	12	M34667_at	PLCG1	3143		1002
	14	U12779_at	MAPKAPK2	2012		894
ORF	1	D49824_s_at	HLA-B	4683	3645	
	18	X63527_at	RPL19	4707	2013	
	20	HG3214-HT3391_at	RPS27	3361	1890	
	24	X01677_f_at	GAPDH	4085	3508	
	25	M36072_at	RPL7A	4287	580	

Table 5: Top 25 (down-expressed) genes identified by the outlier robust F-statistic (ORF). Those in bold face font have been shown to be related to breast cancer in the literature.

Ranking	UniGene ID	Gene Name
1	D49824_s_at	HLA-B
2	AFFX-HSAC07/X00351_M_at	ACTB
3	HG3364-HT3541_at	RPL37
4	X98482_r_at	ACTB
5	X56932_at	RPL13A
6	Z28407_at	RPL8
7	M17733_at	TMSB4X
8	L04483_s_at	RPS21
9	D79205_at	RPL39
10	U27185_at	RARRES1
11	U06155_s_at	RPL23AP7
12	M17886_at	RPLP1
13	L76191_at	IRAK1
14	X15940_at	RPL31
15	L06505_at	RPL12
16	HG1800-HT1823_at	RPS20
17	HG3549-HT3751_at	RPL10
18	X63527_at	RPL19
19	M11147_at	FTL
20	HG3214-HT3391_at	RPS27
21	U05340_at	CDC20
22	D86974_at	LOC440345
23	HG2873-HT3017_at	RPL30
24	X01677_f_at	GAPDH
25	M36072_at	RPL7A

Table 6: Top 25 (down-expressed) genes identified by the outlier F-statistic (OF). Those in bold face font have been shown to be related to breast cancer in the literature.

Ranking	UniGene ID	Gene Name
1	X57351_s_at	IFITM2
2	Z70759_at	WNT2B
3	X70944_s_at	SFPQ
4	D87442_at	NCSTN
5	M93718_at	NOS3
6	X73113_at	MYBPC2
7	HG3039-HT3200_at	ARL2
8	L24783_at	RFC1
9	U78678_at	TXN2
10	HG110-HT110_s_at	HNRPAB
11	M73547_at	REEP5
12	M34667_at	PLCG1
13	X96484_at	DGCR6
14	U12779_at	MAPKAPK2
15	Y11681_at	MRPS12
16	M63483_at	MATR3
17	X89985_at	BCL7B
18	D87435_at	GBF1
19	Z50781_at	TSC22D3
20	D83778_at	KIAA0194
21	U22055_at	SND1
22	U31176_at	GFER
23	M55531_at	SLC2A5
24	U93237_rna2_at	MEN1
25	D26069_at	CENTB2