

Short Communication

Identifying pre-post chemotherapy differences in gene expression in breast tumours: a statistical method appropriate for this aim

EL Korn^{*,1}, LM McShane¹, JF Troendle², A Rosenwald³ and R Simon¹

¹Biometric Research Branch, EPN-8128, National Cancer Institute, Bethesda MD 20892, USA; ²Biometry and Mathematical Statistics Branch, National Institute of Child Health and Human Development, Bethesda MD 20892, USA; ³Metabolism Branch, Bldg. 10/4N114, National Cancer Institute, Bethesda MD 20892, USA

Although widely used for the analysis of gene expression microarray data, cluster analysis may not be the most appropriate statistical technique for some study aims. We demonstrate this by considering a previous analysis of microarray data obtained on breast tumour specimens, many of which were paired specimens from the same patient before and after chemotherapy. Reanalysing the data using statistical methods that appropriately utilise the paired differences for identification of differentially expressed genes, we find 17 genes that we can confidently identify as more expressed after chemotherapy than before. These findings were not reported by the original investigators who analysed the data using cluster analysis techniques. *British Journal of Cancer* (2002) **86**, 1093–1096. DOI: 10.1038/sj/bjc/6600216 www.bjcancer.com
© 2002 Cancer Research UK

Keywords: cluster analysis; doxorubicin; gene expression; microarray; multiple comparisons; statistical methods

Gene expression profiles of tumour specimens, such as obtained by cDNA microarray experiments, can be studied to address a variety of different scientific aims. One aim is to classify the specimens into newly-formed groups so that the gene expression profile is similar within groups and different between groups; statistical cluster analysis techniques have been used for this type of aim (Eisen *et al*, 1998; Tamayo *et al*, 1999). When the specimens come from pre-specified groups, then there are other possible aims. The aim on which we are focusing in this brief communication is the identification of genes that are expressed differentially between pre-specified groups. Identifying such genes can lead to an understanding of how the groups are different at the cellular-functional level (when the identified genes have known function) and can also lead to clues about the function of identified genes with unknown function. Other possible aims with pre-specified groups include demonstrating global differences between groups using multivariate analysis (without identifying individual genes that are differentially expressed), and developing predictors of group membership (Golub *et al*, 1999; Hedenfalk *et al*, 2001). As we will demonstrate here, it is important to use the appropriate statistical methods to address the particular aim under consideration.

Perou *et al* (2000) studied gene expression profiles measured by cDNA microarrays using specimens from 65 breast tumours from 42 individuals. Among the profiles, data from 20 individuals with specimens taken before and after a 16-week course of doxorubicin chemotherapy were included. Based on a cluster analysis, Perou *et al* (2000) note 'Gene expression patterns in two tumour samples from the same individual were almost always more similar to each other than either was to any other sample.' This similarity does not eliminate the equally interesting possibility of finding large and statistically significant differences

in pre vs post chemotherapy gene expression in the 20 paired specimens. A cluster analysis is not the appropriate statistical analysis for examining this possibility. To find genes that are differentially expressed, we will perform an analysis that is appropriate for this aim. We end with a discussion of the biological significance of the identified genes.

Table 1 Genes showing statistically significant pre-post chemotherapy differences in expression for breast cancer patients (original cDNA microarray data from Perou *et al*. (2000))

Accession no. ^a	No. patients	Gene expression (geometric mean)			Adjusted P value ^b
		Pre-chemo	Post-chemo	Ratio of Post/Pre	
AA478553	19	0.74	2.21	2.98	0.0006
N23941	20	1.27	2.18	1.72	0.0014
W96134	20	1.66	3.08	1.85	0.0018
N95402	20	1.19	2.07	1.74	0.0032
AA040944	20	0.42	1.79	4.21	0.0033
AA442853	20	1.53	2.61	1.71	0.0035
AA134757	20	2.32	4.69	2.02	0.0067
AA418077	20	1.75	3.38	1.94	0.0084
R12840	20	0.50	1.77	3.55	0.0166
A1831083	20	1.24	2.48	2.00	0.0178
AA044993	20	0.65	1.37	2.09	0.0180
AA0318596	20	1.19	2.06	1.73	0.0217
AA454868	18	2.95	5.05	1.71	0.0218
AA598794	20	0.72	1.52	2.10	0.0254
T74141	19	8.28	16.44	1.98	0.0326
H210741	20	0.95	1.83	1.93	0.0374
AA133129	20	0.68	1.37	2.01	0.0409

*Correspondence: EL Korn; E-mail: korne@ctep.nci.nih.gov
Received 8 August 2001; revised 10 January 2002; accepted 24 January 2002

^aThe names of the genes are given in Table 2. ^bTwo-sided adjusted P-value from step-down permutation paired t-test, based on 100 000 randomly chosen permutations taken from the set of all possible permutations at each step.

MATERIALS AND METHODS

The primary data were obtained at <http://genome-www.stanford.edu/molecularportraits/> and were pre-processed in a standard manner: Data from spots flagged by the original investigators as not useable or which were labelled 'EMPTY' were omitted here. In each channel, signal for a spot was calculated as foreground intensity minus background. Spots for which signal was less than 100 in both channels were not used. If the signal

was less than 100 in only one channel, the spot was used with the signal set in that channel to 100. The expression ratio was formed as channel 2 divided by channel 1 signal. Ratios were median normalised within each array by dividing the ratios by the median of the ratios for that array. All analyses were performed on log transformed median-normalised expression ratios. Genes for which data were missing from more than half of the 20 paired tumour specimens were eliminated from consideration. This left 8029 genes for analysis.

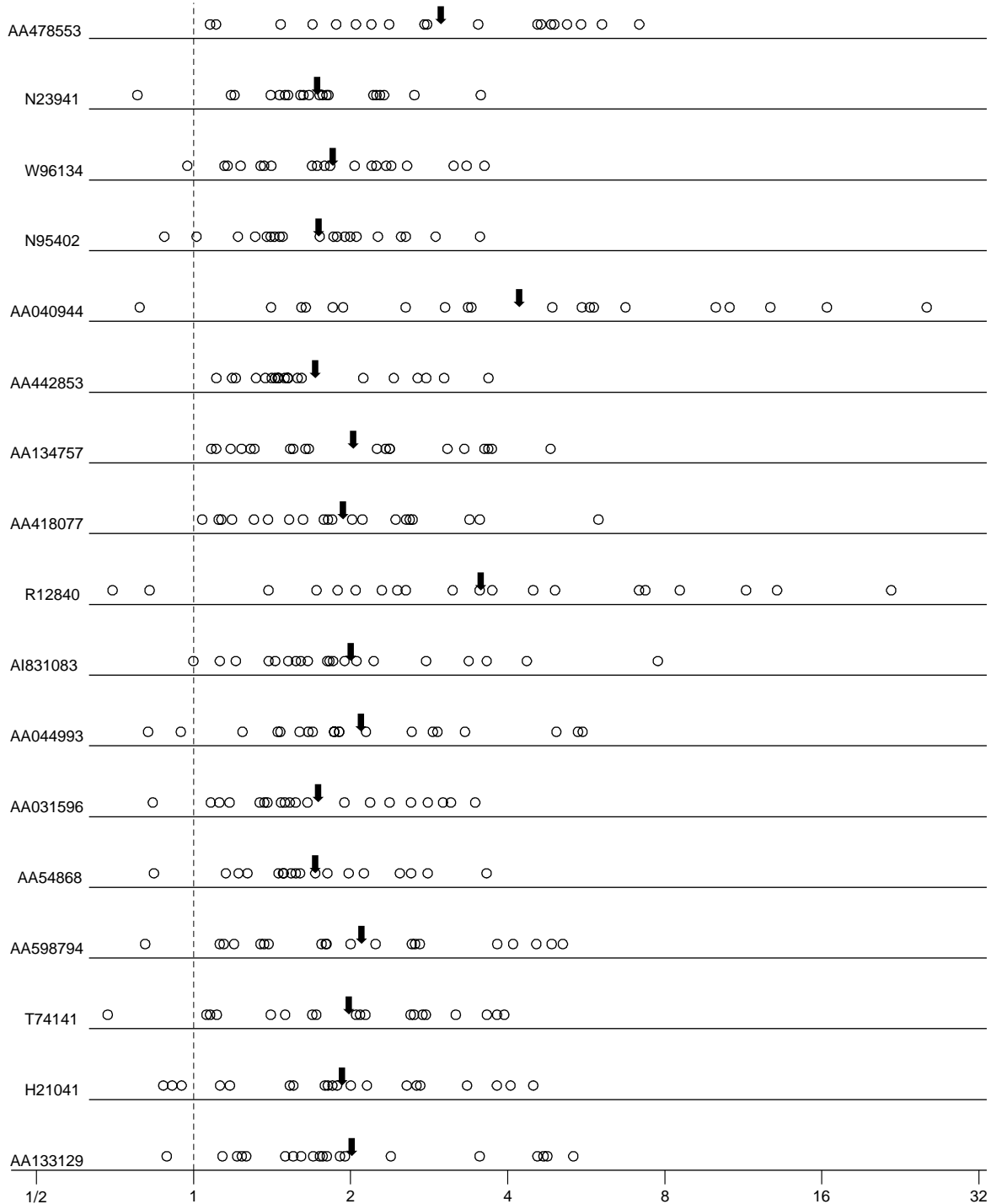


Figure 1 For each gene given in Table 1, plotted points are the ratios of the post-chemotherapy to pre-chemotherapy gene expression ratios for each of 18–20 patients, and arrows are the geometric means of the post/pre ratios.

One statistical method for identifying genes that are differentially expressed is to perform many univariate analyses, testing genes one at a time for differential expression between groups, and then identify the genes which show the most statistically significant differences. In the present application, since there are two groups with paired specimens, an appropriate univariate analysis is a paired *t*-test. One could naively perform 8029 paired *t*-tests, and then identify the genes whose *P*-values were <0.05 . There are two problems with this approach. The first is that one would expect $401=0.05 \times 8029$ genes to show statistically significant ($P < 0.05$) mean group differences even if the expression data were random numbers. The phenomenon of increasing numbers of 'false positives' with increasing numbers of hypothesis tests is known as a 'multiple comparisons problem.' The second problem with this approach is that standard parametric *t*-tests assume that data are normally distributed. This is not usually a problem in applications with large numbers of samples, but in the present application with small numbers of specimens and where interest is in very small (unadjusted) *P*-values, the normality assumption can be important (Ringland, 1983). We deal with both problems simultaneously by using a step-down permutation approach (Westfall and Young, 1993), an approach that has been used previously to identify differentially expressed genes (Callow *et al*, 2000). This approach does not require normal distributions and controls for the multiple comparisons. In fact, it is less conservative than the frequently used Bonferroni adjustment (Miller, 1981) for multiple comparisons.

Note that the proposed statistical analysis involving 20 pairs of data points for each gene automatically accounts for any noise in the data (e.g., due to mRNA extraction, labelling, hybridisation, and spot-to-spot variations within a microarray). Therefore, it is not necessary to perform replicate microarrays on specimens, replicate clones on each microarray, or provide data on intra- and inter-assay variability measurements; all the data required are the 20 pairs of data points for each gene available on the website. However, the extent to which the sources of random variation are controlled or minimised will affect the power to detect true differential effects. Thus, the quality of the data on the website will affect our ability to detect interesting findings, but the reported statistical significance of findings are accurate regardless of this quality. (This is in contrast to an experiment involving two cell lines, in which one would need data on replicate assays or inter-assay variability to be able to conduct statistical inference.) However, if there are replicate clones or genes spotted on the

microarrays, it is of obvious interest to see if they yield similar results. We address this question by examining the differential expression of any genes which have the same name as genes found to have differential expression that is statistically significant.

RESULTS

Table 1 and Figure 1 show the genes identified as being differentially expressed at an adjusted significance level of $P < 0.05$ by the step-down permutation paired *t*-test. The *P*-values are adjusted for the multiple comparisons, so that by chance we would expect

Table 3 Gene expression of genes with names matching those given in Table 2 (original cDNA microarray data from Perou *et al* (2000))

Accession no. ^a	No. patients	Gene expression (geometric mean)			Unadjusted <i>P</i> value ^b
		Pre-chemo	Post-chemo	Ratio of Post/Pre	
N23941 ^c	20	1.27	2.18	1.72	<0.001
N95402 ^c	20	1.19	2.07	1.74	<0.001
W96134 ^c	20	1.66	3.08	1.85	<0.001
AA293362	20	1.71	2.94	1.72	<0.001
AA040944 ^c	20	0.42	1.79	4.21	<0.001
R12840 ^c	20	0.50	1.77	3.55	<0.001
N36944	20	0.81	1.58	1.95	<0.001
AA485377	20	0.52	1.33	2.55	<0.001
AA134757 ^c	20	2.32	4.69	2.02	<0.001
AA614680	20	11.22	16.58	1.48	0.006
AA035156	18	2.04	3.52	1.72	<0.001
AA044993 ^c	20	0.65	1.37	2.09	<0.001
AA598794 ^c	20	0.72	1.52	2.10	<0.001
AA031596 ^c	20	1.19	2.06	1.73	<0.001
H95959	20	2.08	3.10	1.49	<0.001
AA045463	20	1.78	2.64	1.48	0.002
N66035	20	1.58	2.22	1.41	<0.001
AA454868 ^c	18	2.95	5.05	1.71	<0.001
AA461197	20	1.62	2.54	1.56	0.002

^aThe names of the genes are given in Table 2. ^bTwo-sided paired *t*-test, unadjusted for multiple comparisons. ^cAppears in Table 2.

Table 2 Names of genes given in Table 1

Accession no. ^a	Gene name ^b
AA478553	dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2)
N23941	cyclin-dependant kinase inhibitor 1A (p21, Cip1)
W96134	<i>v-jun</i> avian sarcoma virus 17 oncogene homologue
N95402	cyclin-dependant kinase inhibitor 1A (p21, Cip1) ^c
AA040944	<i>v-fos</i> FBJ murine osteosarcoma viral oncogene homologue
AA442853	cyclin-dependant kinase 5, regulatory subunit 1 (p35)
AA134757	fibulin 1
AA418077	GTP-binding protein overexpressed in skeletal muscle
R12840	<i>v-fos</i> FBJ murine osteosarcoma viral oncogene homologue
A1831083	dihydropyrimidinase-like 3
AA044993	connective tissue growth factor
AA031596	secreted protein, acidic, cysteine-rich (osteonectin)
AA454868	platelet-derived growth factor receptor-like
AA598794	connective tissue growth factor
T74141	Duffy blood group
H21041	activating transcription factor 3
AA133129	transcription elongation factor B (5III), polypeptide 3 (110 kD, elongin A)

^aAccession number as given in Perou *et al* (2000) database. ^bGene names associated with accession numbers as given by <http://www.ncbi.nlm.nih.gov/UniGene/>. ^cThe gene name is given as 'copine V' in the Perou *et al* (2000) database.

to identify incorrectly any of the genes as differentially expressed (at adjusted $P < 0.05$) less than once in 20. The average expression for each of the 17 identified genes increased after chemotherapy, and all of the specimens for four of the genes showed more expression after the chemotherapy. The names of the genes are given in Table 2; more information can be found by searching on the accession numbers at <http://www.ncbi.nlm.nih.gov/UniGene/>.

There were no spots in the data base with the same accession number as any of the 17 genes identified in Table 2; we would have expected very close agreement on all gene expression values for such spots. There were nine spots with replicate gene names besides the three pairs of replicate gene names in the identified 17 genes. Gene expressions for all the replicates are displayed in Table 3, along with univariate P -values representing the strength of evidence that the clone is differentially expressed (unadjusted for multiple comparisons). The results suggest that the ratio of post/pre gene expression are roughly similar for replicate gene names, and definitely in the same direction (i.e., greater than 1.0). Interestingly, the individual pre and post gene expressions do not always agree well, e.g., AA614680 is expressed at 4–5 times higher levels as compared to the reference sample than the other two clones associated with fibulin 1.

DISCUSSION

An appropriate statistical analysis has identified genes differentially expressed between pre and post chemotherapy specimens, a task

REFERENCES

- Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res* **10**: 2022–2029
- Daschner PJ, Ciolino HP, Plouzek CA, Yeh GC (1999) Increased AP-1 activity in drug resistant human breast cancer MCF-7 cells. *Breast Cancer Res Tr* **53**: 229–240
- Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95**: 14863–14868
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**: 531–537
- Hedenfalk I, Duggan D, Chen YD, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J (2001) Gene-expression profiles in hereditary breast cancer. *N Engl J Med* **344**: 539–548
- Miller Jr RG (1981) *Simultaneous Statistical Inference* 2nd edn, pp 67–70 New York: Springer-Verlag
- Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Aksien LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D (2000) Molecular Portraits of human breast tumours. *Nature* **406**: 747–752
- Ringland JT (1983) Robust multiple comparisons. *J Am Statist Assoc* **78**: 145–151
- Sethi T, Rintoul RC, Moore SM, MacKinnon AC, Salter D, Choo C, Chilvers ER, Dransfield I, Donnelly SC, Strieter R, Haslett C (1999) Extracellular matrix proteins protect small cell lung cancer cells against apoptosis: A mechanism for small cell lung cancer growth and drug resistance in vivo. *Nat Med* **5**: 662–668
- Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* **96**: 2907–2912
- Vousden KH (2000) p53: Death star. *Cell* **103**: 691–694
- Westfall PH, Young SS (1993) *Resampling-Based Multiple Testing*, pp 72–74 New York: Wiley

for which cluster analysis is not well suited. The genes identified by this approach reveal important biological insights into the response of breast cancer tumours to doxorubicin treatment. The transcriptional up-regulation of the cyclin-dependent kinase inhibitor p21 reflects the p53-dependent response to doxorubicin induced DNA damage and leads to cell cycle arrest (Vousden, 2000). The up-regulation of *c-fos* and *c-jun* as well as higher expression of genes involved in the stromal reaction and extracellular matrix composition (fibulin 1, connective tissue growth factor, osteonectin) might explain, at least in part, the incomplete response to cytotoxic chemotherapy in some of the tumour cells. In particular, elevated mRNA levels of *c-jun* and *c-fos* have been observed in MCF-7 human breast cancer cells with resistance to doxorubicin as compared to drug-sensitive MCF-7 wild type cells (Daschner *et al*, 1999). Moreover, the adhesion of tumour cells to extracellular matrix proteins may provide a survival signal and confer resistance to chemotherapy-induced apoptosis. In small cell lung cancer, this effect was recently shown to be mediated by the integrin family of receptors (Sethi *et al*, 1999). Therefore, it appears plausible that similar mechanisms might exist in breast cancer.

This communication demonstrates the benefits of providing published microarray data on a website for possible reanalysis by other investigators using different methods or seeking to address different questions.