BIOMARKERS, VOL. 9, NO. 6 (NOVEMBER–DECEMBER 2004), 447–460

Taylor & Francis
healthsciences

Identification of transcriptional biomarkers induced by SERMS in human endometrial cells using multivariate analysis of DNA microarrays

JESSICA C. POLE[1], PAUL CARMICHAEL[2] and JULIAN L. GRIFFIN*

Molecular Toxicology, Biomedical Sciences, Faculty of Medicine, Imperial College London, UK

Received 14 June 2004, revised form accepted 5 November 2004

Functional genomic tools such as DNA microarrays are having a major impact on the drug-discovery process. Potentially, compounds with toxic responses can be identified and removed at a relatively early stage in the drug-development process before tests on animals or humans solely on the gene expression profiles produced in cell culture. The study examined the expression profiles of primary cultured human endometrial cells treated with 17β-oestradiol and the SERMs (selective oestrogen receptor modulators) tamoxifen and raloxifene. Primary cultures from three individuals were split into glandular epithelial cells and stromal cell types. Principal components and partial least-squares-discriminate analyses were employed to examine the transcript profile as a whole, identifying genes responsible for patient separation and those that might have an important role in tamoxifen-associated carcinogenesis. Using multivariate data analysis, transcriptional responses were identified in epithelial cells but not in stromal cells for the three SERMs examined. However, it was demonstrated that a major problem associated with using primary cultures derived from human patients is the large transcriptional differences that might exist between the different cultures.

Keywords: 17β-oestradiol, tamoxifen, raloxifene, multivariate data analysis, glandular epithelial cells, stromal cell.

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

Introduction

The potential impact of functional genomic techniques, such as DNA microarrays, in the drug-discovery process are significant, and there has been widespread interest in the use of so-called 'omic technologies' as early screening tools. For example, compounds with toxic responses could potentially be removed at a relatively early stage in the drug-development process, before tests on animals or humans, solely on the gene expression profiles produced in cell culture. The feasibility and validity of the use of microarrays to cluster similar compounds based on their gene expression profiles has been shown in a number of studies (Burczynski *et al*. 2000, Cunningham *et al*. 2000, Scherf *et al*. 2000, Gerhold *et al*. 2001, Waring *et al*. 2001). These studies show a strong correlation between mechanism of action, histopathology, clinical chemistry and gene-expression

[1] Present address: Department of Pathology, Hutchison, MRC Research Centre, University of Cambridge, Cambridge CB2 2XZ, UK.
[2] Present address: Safety & Environmental Assurance Centre, Unilever, Sharnbrook, Bedfordshire MK44 1LQ, UK.

* Corresponding author: Julian L. Griffin, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, UK. Tel: 44 (0)1223 333657; Fax: 44 (0)1223 760002; e-mail: jlg40@mole.bio.cam.ac.uk

Biomarkers ISSN 1354-750X print/ISSN 1366-5804 online © 2004 Taylor & Francis Group Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/13547500400022192

RIGHTSLINK

profiles induced by the agents, suggesting that microarray assays might prove a highly sensitive approach for safety screening of drug candidates. Furthermore, the technology may be particularly appropriate for screening carcinogens, as arrays have been used to identify genes whose expression levels were altered significantly in many different cancers, including prostate (Luo *et al*. 2002), hepatocellular (Goldenberg *et al*. 2002), pancreatic (Iacobuzio-Donahue *et al*. 2002), ovarian (Haviv and Campbell 2002), endometrial (Mutter *et al*. 2001) and breast tumours (Perou *et al*. 2000). Identification of key genes implicated in tumourigenesis also allows the identification of molecular fingerprints to facilitate the classification of different tumour types (Alizadeh *et al*. 2000, Bertucci *et al*. 2000).

However, for the technology to be viable in the drug safety assessment process, both data acquisition and processing must be relatively quick with a high throughput. Furthermore, the data must be processed in such a manner that key changes can still be identified amongst the myriad of transcripts measured using a DNA microarray. Multivariate analysis of these large datasets, such as by principal components analysis (PCA), hierarchical clustering analysis (HCA) and linear discriminate analysis (LDA), is one solution to the data-processing issue, as these approaches analyse the global transcriptional profile simultaneously. PCA also has the benefit that the technique copes with 'long and lean' data sets where the number of experiments (observations) is very much smaller than the number of transcripts (variables). The approach is also robust to random variation and experimental error (noise) in the system, multicollinearity (where variables may appear to be collinear as a result of the size of the matrix) and a degree of missing data (Wold *et al*. 1984).

The present paper examines the expression profiles of primary cultured human endometrial cells treated with 17β-oestradiol and the SERMs tamoxifen and raloxifene. Primary cultures from three individuals were split into glandular epithelial and stromal cell types. Multivariate analyses were employed to examine the transcript profile as a whole, identifying genes responsible for patient separation and those that might have an important role in tamoxifen-associated carcinogenesis.

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

Materials and methods

Patient details

Endometrial samples were obtained from three premenopausal patients undergoing laporoscopic sterilization. Patients had endometria that had not been influenced by hormonal drug intervention. Samples were taken in the proliferative stage of the menstrual cycle. The three patient ages were 45 (LG15), 38 (LG20) and 42 (LG21) years, where the codes in parentheses signify the different patients. Tissue samples were placed in sterile DMEM/F12 media (Invitrogen, Paisley, UK), supplemented with 100 μg ml^{-1} kanamycin sulphate (Sigma-Aldrich, Poole, UK) and 2 mM glutamine (Invitrogen). The samples were then transferred on ice to the laboratory.

Ethical approval for the acquisition of endometrial samples from individuals undergoing gynaecological surgery was obtained from the appropriate hospital authorities, as was patient consent.

Microarrays

Human Affymetrix GeneChips® U95Av2 were used. They contain about 12 000 previously characterized sequences representing the known full-length gene complement of the human genome.

RIGHTSLINK

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

Cell culture

Separate cultures of stromal and epithelial cells were prepared from the tissue samples. For this, the method described by Satyaswaroop *et al*. (1979) was used. In brief, samples were dissected into 2 mm^3 pieces and incubated with a digestion solution of sterile filtered (0.22 μm Gelman Acrodisc filter; Fisher Scientific, Loughborough, UK) 15 ml Hanks buffered saline solution, or HBSS; Invitrogen) containing 20 mg collagenase type II and 15 mg DNase I (both Sigma-Aldrich) for 30–60 min at 37°C, with gentle agitation every 10 min. The incubation generated a single-cell suspension of stromal cells whilst maintaining the epithelial cells as whole glands. Dissociation was aided by pipetting the solution through a 1 ml pipette.

To remove undigested fragments, the suspension was filtered through sterile 350 μm steel mesh (Cadisch Precision Meshes Ltd., London, UK), and washed with 50 ml HBSS to dilute the digestion enzymes and prevent cell lysis. The suspension was centrifuged (100*g*, 10 min, 4°C) and washed once more with 20 ml HBSS. The mixed cell suspension was passed through sterile 35 μm steel mesh (Cadisch Precision Meshes Ltd.) to separate the stromal cells from the gland fragments, which remain on the filter. Glands were harvested by inverting the mesh into a Petri dish and washing back through with complete media containing 5% foetal bovine serum (FBS; Invitrogen); complete media consisted of DMEM/F-12 media (Invitrogen), supplemented with 100 μg ml^{-1} kanamycin sulphate (Sigma-Aldrich) and 2 mM glutamine (Invitrogen).

The separated cell types were centrifuged (100*g*, 10 min, 4°C), resuspended in 5 ml complete media and the stromal cells counted using a haemocytometer. Trypan blue (Sigma-Aldrich) was used to assess cell viability. The stromal cells were seeded at 0.7×10^6 cells per 25 cm^2 culture flask (Nunc, Fisher Scientific, Loughborough, UK) in 4 ml complete media. Glands were plated in half the number of 25 cm^2 culture flasks used for the stromal cells in 4 ml complete media.

Drug treatment

Cell cultures were exposed to a final concentration of 0.1 μM 17β-oestradiol, 0.1 μM tamoxifen or 0.1 μM raloxifene by adding a stock solution dissolved in ethanol to the cell culture media or ethanol alone. This dose is non-toxic but influences cell homeostasis in previous experiments conducted by the authors. A time point of 24 h was used to eliminate the potentially confusing contribution of non-specific immediate early stress responses of cells exposed to stimuli.

RNA extraction

Total RNA was isolated directly from the plated cell cultures using RNAStat 60 reagent (Tel-test, Inc., Texas, USA). RNA samples were resuspended in 100 μl nuclease-free water. Qiagen RNeasy spin columns were used to clean up RNA, and RNA was extracted using the RNeasy Mini Kit (Qiagen, Crawley, UK). RNA concentration and quality were assessed by a ultraviolet light spectrophotometer. All RNA samples had an absorbance ratio at 260/280 nm of 1.9–2.1, indicating pure RNA. Samples were also run on a 1% agarose gel to check total RNA integrity.

Gene chip hybridization

First- and second-strand cDNA synthesis and *in vitro* transcription reactions were performed according to Affymetrix recommendations. First-strand cDNA synthesis was primed with a T7(dT)24 primer (100 pmol μl^{-1}; MWG Biotech, Milton Keynes, UK) using SuperScriptII reverse transcriptase (Invitrogen). Phenol:chloroform:isoamyl alcohol (25:24:1) (Sigma-Aldrich) purification of cDNA was performed using Phase Lock™ gel tubes (Eppendorf, Cambridge, UK). From 2 μg cDNA, cRNA was synthesized and biotin-labelled nucleotides were incorporated using Enzo BioArray High-Yield RNA Transcript Labelling kit (Enzo Life Sciences, New York, USA). After a 37°C incubation step for 6 h, the labelled cRNA was cleaned up using an RNeasy Mini kit (Qiagen). cRNA was fragmented in buffer (40 mM Tris-acetate, pH 8.1, 100 mM potassium acetate, 30 mM magnesium acetate; Sigma-Aldrich) for 35 min at 94°C. Fragmented cRNA (55 μg) was hybridized on the Human Genome U95Av2 chip (Affymetrix, California, USA) for 18 h at 60 rpm rotation in a 45°C Genetic® hybridization oven (Model 640; Affymetrix). Chips were washed and stained with streptavidin phycoerythrin (SAPE; Invitrogen) in Affymetrix fluidics stations. To amplify staining, SAPE solution was added twice with an anti-streptavidin biotinylated antibody (Vector Laboratories, Peterborough, UK) staining step in between.

Hybridization of the probe arrays was detected using a Hewlett-Packard GeneArray® fluorometric scanner (Hewlett Packard Corp., California, USA) at a wavelength of 488 nm. The microarray images were analysed for quality control.

RIGHTSLINK

Affymetrix analysis of the data

The data were initially analysed using Affymetrix GeneChip software. Three chips were hybridized for each treatment condition (vehicle control, 17β-oestradiol, tamoxifen and raloxifene) and each cell type, resulting in a total number of 24 chips. The chips contain 16–20 oligonucleotide probe pairs per gene or cDNA clone, and use a match/mismatch protocol for determining expression values. This takes into consideration variability in hybridization among probe pairs and other hybridization artefacts that could affect fluorescence intensities.

The data set was exported into Microsoft Excel® where minimal and negative expression values were set to 20 (floored) in order to eliminate genes with negative expression levels and to reject uninterpretable information (so-called flooring of the data). All data were $\log_2$ transformed before pattern recognition analysis.

Univariate data analysis

The data sets for the complete transcriptional changes for stromal and epithelial cells across all four treatments and three patients were imported into Matlab (www.mathworks.com). An ANOVA test was performed using the Statistical toolbox to identify transcripts that were altered by a given treatment in either stromal or epithelial cells for all three patients ($p < 0.05$).

Multivariate data analysis

PCA, as part of the software package SIMCA (Umetrics, Umea, Sweden), was used as a multivariate pattern recognition tool. It is a multivariate analysis method that highlights inherent clustering behaviour for samples based on the similarity of the gene expression profiles. The supervised multivariate analysis approach of partial least-squares analysis (PLS) was also used to identify significant changes correlating to treatment exposure. PLS is a regression extension of PCA (Eriksson *et al*. 1999) and is useful when analysing noisy data with a large number of variables. A PLS model is formed by expressing the variables measured as a set of X-score and X-weight vectors, and in turn also expressing the responses in terms of a Y-score and Y-weight vectors. Each dimension in the subsequent PLS model expresses a linear relation between X- and Y-score vectors. In this way, the data are modelled as a set of 'factors' in X and Y, simplifying the relationship between the two data matrices.

PLS models were built from the data set by representing the gene changes as X variables and the treatments as a Y variable, correlating gene changes to treatment. The validity of the model was determined by sequentially removing one sample in every seven and predicting its class membership using the PLS model. A model was deemed robust when the quality of fit (Q^2) exceeded that by chance (0.097 for these models), indicating that the minimum number of components describes the most information. Loading plots, which show the inherent clustering of the variables (gene transcripts) following pattern recognition, were also generated for PCA and PLS models. Contribution plots could then be interrogated to highlight coefficients; in this case, the individual genes that contributed to the clustering. Genes that had contributed greatly to the clustering were then selected and identified.

PLS was also applied in the form of the spectral filter orthogonal signal correction. This filter removes variation from the X matrix, which is unrelated (orthogonal) to Y. Such approaches have previously been applied to a number of biological problems including metabonomics and signal processing of infrared spectroscopy data (Beckwith-Hall *et al*. 2002, Gavaghan *et al*. 2002, Luypaert *et al*. 2002).

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

Results

Overview of data

Data sets from each treated array were plotted against the respective control to indicate the spread of gene expression changes. Control and treated groups for each patient and means across the patients were compared (figure 1A). The plots indicate a fairly tight spread of gene changes, with few outliers. This basic analysis was repeated for cells from different patients exposed to the same drug (figure 1B) and stromal cells against epithelial cells (figure 1C). A large spread of gene changes was evident in both sets of comparison, indicating a large variation between cells from different patients as well as expected differences between cell types.

RIGHTSLINK

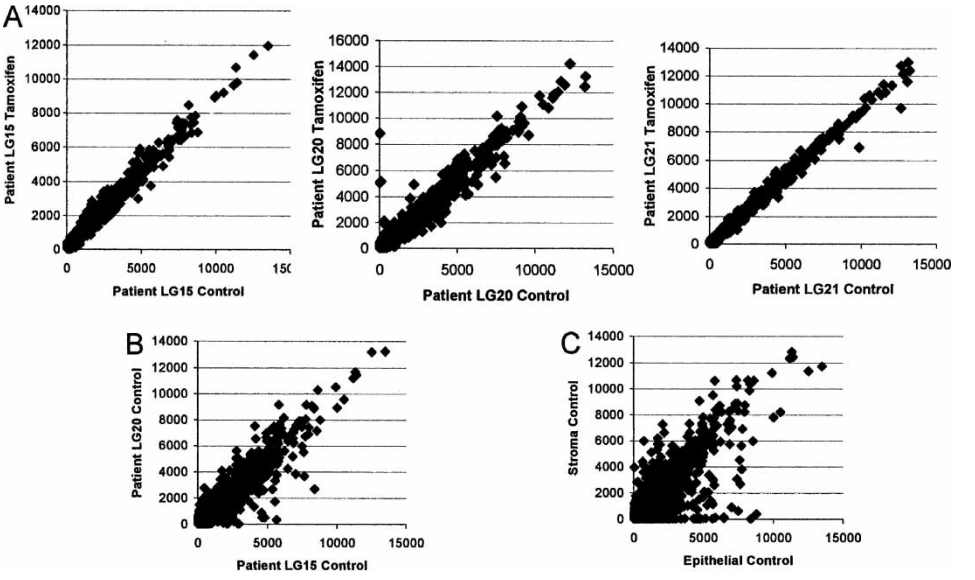

Figure 1. Scatter plots for gene transcript expression across various two-way comparisons. Each plot represents the expression of a transcript in two different samples, with deviation from the diagonal indicating differential expression of that transcript. (A) Three two-way comparisons of the effect of tamoxifen on glandular epithelial cells compared with the control experiment (dosing vehicle alone) for the three patients. (B) Comparison of glandular epithelial cells derived from two patients (both cell cultures under control conditions). (C) Comparison of expression in glandular epithelial and stromal cells.

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

ANOVA test of variance

Using an ANOVA filter to identify transcripts that were altered in all three patients in epithelial cells, 194 transcripts were identified as being significantly altered with <0.05. The ten transcripts with the highest significance, given with both Affymetrix ID (ID) and LocusLink ID (LL), were pericentrin 2 (kendrin) (ID: 32097_at; LL: 5116), Rho guanine nucleotide exchange factor (GEF) 12 (ID: 35772_at; LL: 23365), ubiquitin-specific protease 15 (ID: 34295_at; LL: 9958), KIAA0007 (ID: 38250_at; LL: 23160), cAMP-responsive element binding protein-like 2 (ID: 39438_at; LL: 1389), α-2-HS-glycoprotein (ID: 36621_at; LL: 197), glutamine phosphoribosyl pyrophosphate amidotransferase (ID: 34341_at; LL: 5471), vesicle amine transport protein 1 homologue (*T. californica*) (ID: 40147_at; LL: 10493), DNA sequence from Fosmid 27C3 on chromosome 22q11.2-qter (ID: 39374_at) and an unknown gene. However, applying a similar analysis to the stromal cell cultures, only 77 transcripts were significantly different according to an ANOVA test. The ten transcripts with the highest significance were transcripts from membrane-associated tyrosine- and threonine-specific cdc2-inhibitory kinase (ID: 480_at; LL: 9088), zinc finger protein 154 (pHZ-92) (ID: 31390_at; LL: 7710), tumour protein D52-like 2 (ID: 40076_at; LL: 7165), P97 antigen (melanoma specific; ID: 939_at), killer cell immunoglobulin-like receptor, 4 (ID: 38231_f_at; LL: 3805), sulfite oxidase (ID: 32122_at; LL: 6821), protein kinase, lysine deficient 1 (ID: 39313_at; LL: 65125), and hypothetical protein LOC65243 (ID: 35557_at; LL: 65243).

RIGHTSLINK

Furthermore, only two transcripts were altered in both cell lines: mRNA for diacylglycerol kinase (ID: 39044_s_at; LL: 8527) and tumour protein D52-like 2 (ID: 40076_at). However, this univariate approach is prone to false-positives, and a simple power analysis suggested that significant numbers of the identified transcripts were altered in transcriptional profile purely by chance. Applying a Bonferroni correction to this analysis, only 21 and eight transcripts were significantly altered in the epithelial and stroma data sets, respectively. Thus, the data were further examined using multivariate analysis.

PCA of the complete data set

PCA was initially used to verify that data obtained from the Affymetrix analyses was not influenced by external factors such as buffer solutions, the batches in which the chips were made or the day the analysis was performed. PCA of the total about 12 000 transcripts classified gene arrays according to the cell type and patient from which they were selected. This also indicated that the separately cultured cells were relatively pure, demonstrating the expected difference in transcriptional profiles between the two cell types. In subsequent analyses, the cell cultures were analysed separately.

The PCA of the epithelial cell lines demonstrated clear separation according to the patients from which the cells were derived, regardless of treatment (figure 2A). Examining the loadings plots to identify which genes were responsible for this clustering, a number of transcripts were identified that were up- or down-regulated in cell cultures from one of the patients, but largely did not influence the other sets

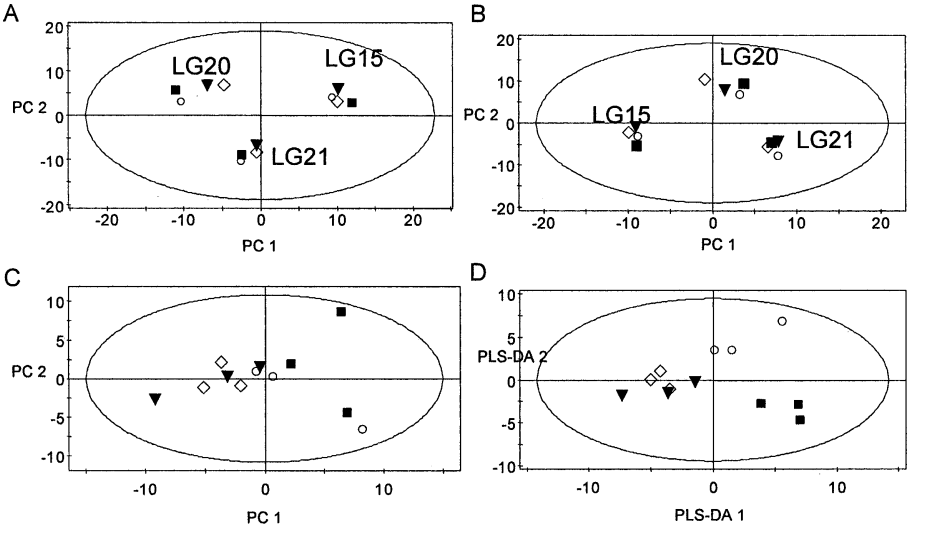

Figure 2. Multivariate analysis of the complete transcriptional changes recorded in glandular epithelial and stromal cells. (A) PCA of the epithelial cell lines, demonstrating clear separation according to patient. (B) PCA of the stromal cell data set similarly showing clustering according to patient. To identify changes associated with treatment, orthogonal signal correction was applied to the epithelial data set. (C) PCA of the post-OSC data set from epithelial cells. (D) PLS-DA analysis of the post-OSC data set from epithelial cells. ■, Control (vehicle only); ○, 0.1 μM 17β-oestradiol; ▼, 0.1 μM tamoxifen; ◇, 0.1 μM raloxifene.

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

RIGHTSLINK

of cultured cells. In patient LG20, the major increased transcripts compared with the other two patients were mRNA for L-3-phosphoserine-phosphatase homologue (twice; gene ID: 37208_at, 37209_at; LL: 5723), cathepsin G (ID: 679_at; LL: 1511), mevalonate (diphospho) decarboxylase (ID: 34410_at; LL: 4597), two unknown genes, glutathione *S*-transferase M1 (ID: 39054_at; LL: 2944), DEAD (Asp-Glu-Ala-Asp) box polypeptide 17 (ID: 41260_at; LL: 10521), and mRNA for HL23 ribosomal protein homologue ribosomal protein L23 (twice; ID: 32394_s_at; 32395_s_at; LL: 9349). Similarly, transcripts could be identified for the other cell cultures that distinguished cell cultures from patient LG15 (human granulocyte-macrophage colony-stimulating factor (CSF1) gene (twice; ID: 1400_at; 1401_at; LL: 1437), keratin 6 isoform (ID: 39015_f_at; LL: 3853), human complement factor B (ID: 35822_at; LL: 629), cathepsin E (CTSE) gene (twice; ID: 271_s_at; 206_at; LL: 1510), RAD mRNA (ID: 1776_at; LL: 6236), matrix metalloproteinase 14 (membrane-inserted) (ID: 34747_at; LL: 4323), and pregnancy specific β-1-glycoprotein 9 (ID: 38180_f_at; LL: 5678) all increased in expression) and for patient LG21 (apolipoprotein E (ID: 608_at; LL: 348), three unknown genes, a putative lymphocyte G0/G1 switch gene (ID: 38326_at; LL: 50486), lymphocyte-specific protein 1 (LSP1; ID: 36493_at; LL: 4046), pregnancy specific β-1-glycoprotein 7 (ID: 35956_s_at; LL: 5676), transporter protein family 38, member 3 (g17; ID: 40389_at; LL: 10991), β-tubulin folding cofactor D (ID: 39399_at; LL: 6904), all increased in expression). Furthermore, PCA of the stromal cell data set similarly demonstrated that the variation associated from which patient the cells were derived was the dominant effect across the data set (figure 2B).

To investigate the transcriptional changes associated with drug treatment, a discriminate form of pattern recognition is required to exclude variations in the data that is associated with which patient the cell cultured were derived. For such multivariate supervised approaches, the complete data set of about 12 000 transcripts is too large to allow any robust cross-validation. However, the data-filtering approach OSC could be used. This filter removes variation orthogonal to the axes that describe the drug treatments. Following OSC, both PCA and PLS-DA separated the three treatment groups from each other and the control group (figure 2C, D). Using the PLS-DA model, the variables most responsible for the separation were identified using the variable importance in the projection (VIP). These VIP scores of the variables reflect the importance of a transcript to the overall multivariate separation produced by the model, and were used to identify the most significant transcriptional changes for classifying the different microarrays according to treatment (table 1). When applying this analysis to the stromal cell data set, even with the OSC filter, no separation could be produced using PCA or PLS-DA.

The OSC filter produces a new data set compared with the original input and, thus, subsequent analysis using PCA and PLS-DA is not cross-validated with respect to the original data. To avoid the OSC filter and to examine whether there were common responses to a drug treatment, the stromal and epithelial data were filtered in one of three ways to reduce the number of variables included in any subsequent analysis. The resultant reduced data sets were examined by

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

RIGHTSLINK

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

Table 1. Major transcriptional changes responsible for the separation of epithelial cultures according to drug treatment. The 30 most significant transcripts identified by the PLS-DA model using the variable indicator for the projection model (VIP) are shown across all drug treatments. Drug treatment identifies which drug caused the maximum change in transcription for that gene. Fold changes are calculated from the $\log_2$ of the ratio of the mean transcripts in the treatment group compared with that from the control group.

Affymetrix probeset ID	LocusLink ID	Gene description	VIP score	Drug treatment	Fold changes
2018_at	2697	gap junction protein, α-1, 43 kDa (connexin 43)	5.84	EST	2.0±1.4
1486_at	5439	polymerase (RNA) II (DNA directed) polypeptide J, 13.3 kDa	5.41	TAM	3.9±3.5
38964_r_at	7454	Wiskott–Aldrich syndrome (eczema-thrombocytopenia)	5.33	RAL	−3.36±5.6
1069_at	5743	prostaglandin-endoperoxide synthase 2 and cyclooxygenase	5.09	EST	2.2±1.8
37649_at	3145	hydroxymethylbilane synthase	4.97	TAM	−1.3±1.1
32550_r_at	1050	CCAAT/enhancer binding protein (C/EBP), α	4.87	RAL	−1.6±0.9
32115_r_at	135	adenosine A2a receptor	4.58	RAL	1.3±0.8
38875_r_at	9687	GREB1 protein	4.54	EST	2.0±1.2
37013_at	1636	angiotensin I-converting enzyme (peptidyl-dipeptidase A) 1	4.5	RAL	−2.8±1.2
35090_g_at	9542	neuregulin 2	4.46	RAL	2.1±1.7
38860_at	5143	phosphodiesterase 4C, cAMP-specific	4.42	EST	−2.1±1.3
1535_at	1112	checkpoint suppressor 1	4.38	TAM	1.4±1.2
41537_r_at	4784	nuclear factor I/X (CCAAT-binding transcription factor)	4.15	TAM	−1.2±2.1
33503_at	23746	aryl hydrocarbon receptor interacting protein-like 1	4.12	TAM	−1.1±1.7
1577_at	367	androgen receptor	4.08	EST	1.9±0.5
32897_at	4524	5,10-methylenetetrahydrofolate reductase (NADPH)	4.05	TAM &	−1.7±0.8
				RAL	−1.3±0.7
31410_at	23495	tumour necrosis factor receptor superfamily, member 13B	4.04	EST	2.3±1.7
37940_at	–	cluster including AA806768: ob91d06.s1	4.02	RAL	2.4±1.3
33648_at	157697	hypothetical protein LOC157697	4.01	EST	−2.0±1.1
37184_at	6804	syntaxin 1A (brain)	3.97	RAL	0.3±0.3
38487_at	23166	stabilin 1	3.97	TAM	1.03±0.8
33277_at	8898	myotubularin-related protein 2	3.92	TAM	0.63±0.9
40186_at	1852	dual-specificity phosphatase 9	3.91	EST	−1.8±2.4
36567_at	57030	solute carrier family 17 (Na-dependent inorganic phosphate co-transporter)	3.88	RAL	−1.9±2.5
40044_at	8178	elongation factor RNA polymerase II	3.85	RAL	1.2±0.8
38298_at	3779	potassium large conductance calcium-activated channel	3.82	EST	1.7±1.3
31396_at	–	cluster including AB012851: Musashi	3.8	RAL	2.7±1.6
2030_at	7481	wingless-type MMTV integration site family, member 11	3.78	RAL	−1.2±1.5
31667_r_at	–	cluster including W27698: 36f8	3.78	EST	−2.5±1.7
36728_at	146	adrenergic, α-1D-, receptor	3.78	EST	−1.9±2.7

RIGHTSLINK

PCA and PLS-DA to examine whether the three drugs produced distinct transcriptional patterns and also whether there were common transcriptional responses for 17β-oestradiol and the two SERMs. Using these smaller data sets, the models produced could be cross-validated using the goodness of fit algorithm (Q^2) applied to determine how many components a PCA or PLS model uses, using a prediction approach to assess these models.

PLS-DA modelling of the most expressed transcripts

A data set was derived containing the 25 most expressed transcripts for each drug treatment as well as the control cell population. These transcripts represent those that are less affected by the inherent noise in the system associated with detecting hybridization. This produced a data set containing only 29 transcripts for the epithelial cell cultures, demonstrating that drug treatment did not significantly or greatly change the expression of this set of transcripts. Applying PCA to this data set again identified clustering associated with the patient from which the cells were derived. Most of the variation in this data set (94%) could be explained by two PCs that were influenced by the patient from which the cells were derived. PLS-DA failed to identify a trend associated with drug treatment. In a similar manner, the most expressed transcripts from the cultured stromal cells did not separate the arrays according to treatment for the three patients.

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

PLS-DA modelling of the biggest fold changes in the data set

To examine the transcripts most influenced by drug treatment, a data set was produced for transcripts where the $\log_2$-fold change was greater than 1.5 compared with the control expression for a given patient. This produced a data set containing 92 different transcripts. For epithelial cells, PCA of the data revealed the clustering of the control cell culture population away from the three drug treatments, largely along PC1 (figure 3A). The most prominent gene changes were increases in transcripts from the genes fascin homologue 2, actin-bundling protein (ID: 34490_f_at; LL: 25794), elaC homologue 2 (*E. coli*) (ID: 39869_at; LL: 60528), microtubule-associated protein 1A (ID: 35917_at; LL: 4130), immunoglobin light chain (ID: 32361_s_at; LL: 8940), tumour necrosis factor (ligand) superfamily, member 4 (ID: 32319_at; LL: 7292), katanin p80 subunit B 1 (ID: 40976_at; LL: 10300), erythroid K-Cl co-transporter splicing isoform 2 (ID: 38624_at; LL: 6560), Musashi homologue 1 (ID: 31395_i_at; LL: 4440), RNA polymerase II polypeptide J, 13.3 kDa (ID: 1486_at; LL: 5439) and glycine C-acetyltransferase (2-amino-3-ketobutyrate coenzyme A ligase) (ID: 33579_i_at; LL: 23464), and decreases in Wiskott–Aldrich syndrome protein (WASP; ID: 38963_i_at; LL: 7454), CCAAT/enhancer binding protein α (ID: 32550_r_at; LL: 1050), Hep27 protein (ID: 35004_at; LL: 10202), 15-lipoxygenase (ID: 34636_at; LL: 246), α-A1-adrenergic receptor (ID: 36729_g_at; LL: 146), oestrogen receptor mRNA, alternatively spliced transcript E (ID: 33670_at), ankyrin 1 (variant 2.1; ID: 32152_at; LL: 286), A2 mRNA (ID: 39803_at), DNA topoisomerase III mRNA (ID: 41004_at; LL: 7156) and an unknown transcript.

Applying PLS-DA to the data set, the four treatment groups could be separated across four PLS-DA components (figure 3B, C). Furthermore, a highly predictive

RIGHTSLINK

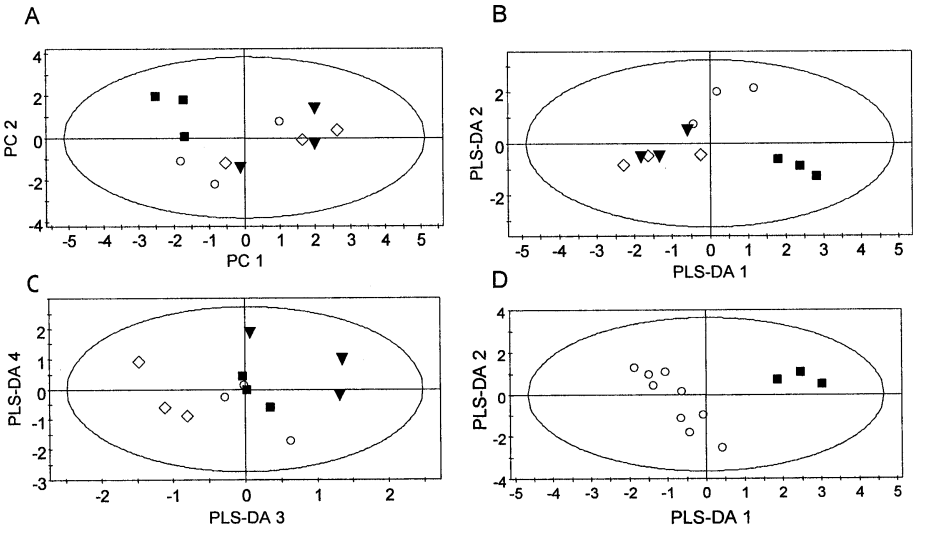

Figure 3. Transcripts (92 different genes in total) with a $\log_2$-fold change greater than 1.5 compared with the control expression were compared using PCA and PLS-DA. (A) PCA of the data revealed the clustering of the control cell culture population away from the three drug treatments. PLS-DA analysis of this data set clustered the four treatment groups across four components. (B) PLS-DA components 1 and 2. (C) PLS-DA components 3 and 4. ■, Control (vehicle only); ○, 0.1 μM 17β-oestradiol; ▼, 0.1 μM tamoxifen; ◇, 0.1 μM raloxifene. (D) A PLS-DA model of the data set discriminating the control data set from the three drug treatments considered as a single group. ■, Control (vehicle only); ○, drug treatment.

PLS-DA model could be built to cluster the combined three treatment groups away from the control group (figure 3D). Table 1 shows the transcripts responsible for this classification. However, a similar analysis of stromal cells failed to build a predictive model, suggesting the three drugs did not perturb the transcriptional profiles of cells in a statistically significant manner.

Discussion

With the advent of global transcriptional analysis using DNA microarrays, the capability of identifying transcriptional profiles associated with the adverse effects of drugs will potentially revolutionize drug safety assessment in the pharmaceutical industry. The major problem with applying univariate statistical analysis to microarray data is the large number of variables, resulting in a number of variables being identified as significantly altered in expression according to univariate statistical tests, such as the Student's *t*-test, purely by chance. This is exacerbated in situations where sample numbers are limited, such as in the present study where primary human cell cultures were used. As part of this study, an ANOVA test was performed to identify transcripts that were significantly altered following exposure of the cultured cells to one of three different SERMs. While a number of genes were identified as being altered in cultures of both stromal and epithelial cells, an attempt to correct for false-positives in the data set suggested that the majority of transcripts identified could have resulted purely by chance.

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

RIGHTSLINK

To avoid the problems associated with univariate analysis of these large data sets, a series of multivariate pattern recognition tools based on PCA was employed. Applying PCA to either the complete transcriptional data set from stromal or epithelial cells, the most apparent trend was that associated with the patient. Thus, to detect any effect associated with the drugs, the differences associated with the cultured cells must first be subtracted from the data set. Across five PCs, representing about 94% of the variation, no trend was identified in epithelial cells associated with the four drug treatment groups. Using the variable loadings of these PCs, the transcripts most responsible for this separation were identified, and an important aspect of using cultured human cells may be initially screening these transcripts for predisposed sensitivity or unresponsiveness to a given class of compounds. Thus, this innate variation in the data sets associated with the different patients is potentially a serious problem in data sets derived from human tissue. A further source of variation will also be associated with the steps required to collect and culture the cells used in this study. Significant alterations in gene expression will result in normal variation in these processes. This may also exacerbate the differences between stromal and epithelial cells.

In the various supervised pattern recognition approaches applied to the data to identify transcriptional changes associated with a drug response, more robust models were built for epithelial compared with stromal cells, indicating the three SERMs had a more profound effect on the former. In fact, no model could be produced that separated the stromal cells according to treatment group without prior selection of transcripts according to fold changes associated with a given drug. This was also supported by the ANOVA test, where over twice as many transcripts were significantly altered by one or more treatments in epithelial compared with the stroma cells.

Despite the successful identification of transcriptional profiles associated with the cell types, the patients from which the cells were derived and the treatment group (epithelial cells only), the major transcriptional changes identified for each classification were associated with a wide range of pathways as well as with a number of transcripts with putative or unknown function. Thus, rather than identifying a particular pathology or response associated with a drug, the multivariate techniques identified 'barcodes' of transcripts associated with a given separation. As PCA and PLS-DA identify changes with respect to all the other changes in the data sets, this approach may be useful for identifying transcriptional changes that are only significant in terms of identifying a response when they occur in conjunction with another or as part of a series of changes. Furthermore, the approach might be good at identifying which transcripts need to be re-examined by other analyses, such as RT-PCR, to confirm a change that might not be significant in terms of a Student's *t*- or ANOVA test when considering a given transcript in isolation.

A number of transcripts previously implicated in endometrial cancer were identified as being significantly altered in expression in this study. Increased expression of cyclooxygenase-2 (COX-2) has been detected in a number of cancers including prostate, colorectal, uterine adenocarcinoma and endometrial (Tong *et al*. 2000, Ferrandina *et al*. 2002, Landen *et al*. 2003). The enzyme is rate limiting

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

RIGHTSLINK

in prostaglandin biosynthesis, with COX-2-derived prostacyclin participating in blastocyst implantation through activation of peroxisome proliferator-activated receptor delta. As well as increased COX-2 expression, a number of transcripts associated with hormonal responses were detected as a result of the oestrogen-like properties of the drugs examined. These transcripts included those of the α-A1-adrenergic receptor, the oestrogen receptor gene and the androgen receptor. Oestrogen receptor expression is high in normal cyclical endometria, hyperplasias and carcinomas (Bozdogan *et al*. 2002), and in keeping with previous reports, the three SERMS increased the expression of the oestrogen receptor compared with normal endometrial cells (Schutze *et al*. 1992). However, expression of the androgen receptor is much lower in both normal and cancerous tissue (Brys *et al*. 2002). A number of other transcripts have roles in proliferation or control of the cell cycle, including polymerase II and Checkpoint suppressor 1, and of course have been implicated in cancerous cell growth. Furthermore, Wiskott–Aldrich syndrome increases the likelihood of a number of cancers as a result of compromised immunity (Model 1977, Cotelingam *et al*. 1985, Shcherbina *et al*. 2003), but why SERM action induces the protein product of Wiskott–Aldrich syndrome gene is unknown.

Examining the table of transcripts found to be altered by drug treatment of epithelial cells as identified by multivariate analysis, many of the fold changes were comparable in size with the standard deviations calculated, indicating the transcripts were on the borderline in terms of significance. However, the standard deviation represents a univariate analysis of the data. Thus, many of the transcripts identified might not be significantly altered when considered alone, but their expression may be highly characteristic of the drug treatment when considered in conjunction with the global transcriptional changes. This highlights a further advantage of using multivariate analysis of data as well as classical univariate approaches such as Student's *t*-test and one-way ANOVA.

In conclusion, the use of multivariate data analysis transcriptional responses was identified in epithelial cells but not stromal cells for the three SERMs examined. However, a major problem associated with using primary cultures derived from human patients is the large transcriptional differences that might exist between the different cultures.

Supplemental data: for the DNA microarray data, see http://www.bio.cam.ac.uk/aboutjp339/

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

Acknowledgements

J. L. G. is a Royal Society University Research Fellow.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. Jr, Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. 2000, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

RIGHTSLINK

BECKWITH-HALL, B. M., BRINDLE, J. T., BARTON, R. H., COEN, M., HOLMES, E., NICHOLSON, J. K. and ANTTI, H. 2002, Application of orthogonal signal correction to minimise the effects of physical and biological variation in high resolution 1H NMR spectra of biofluids. *Analyst*, **127**, 1283–1288.

BERTUCCI, F., HOULGATTE, R., BENZIANE, A., GRANJEAUD, S., ADELAIDE, J., TAGETT, R., LORIOD, B., JACQUEMIER, J., VIENS, P., JORDAN, B., BIRNBAUM, D. and NGUYEN, C. 2000, Gene expression profiling of primary breast carcinomas using arrays of candidate genes. *Human Molecular Genetics*, **9**, 2981–2991.

BOZDOGAN, O., ATASOY, P., EREKUL, S., BOZDOGAN, N. and BAYRAM, M. 2002, Apoptosis-related proteins and steroid hormone receptors in normal, hyperplastic, and neoplastic endometrium. *International Journal of Gynecology and Pathology*, **21**, 375–382.

BRYS, M., SEMCZUK, A., BARANOWSKI, W., JAKOWICKI, J. and KRAJEWSKA, W. M. 2002, Androgen receptor (AR) expression in normal and cancerous human endometrial tissues detected by RT-PCR and immunohistochemistry. *Anticancer Research*, **22**, 1025–1031.

BURCZYNSKI, M. E., MCMILLIAN, M., CIERVO, J., LI, L., PARKER, J. B., DUNN, R. T. II, HICKEN, S., FARR, S. and JOHNSON, M. D. 2000, Toxicogenomics-based discrimination of toxic mechanism in HepG2 human hepatoma cells. *Toxicology Science*, **58**, 399–415.

COTELINGAM, J. D., WITEBSKY, F. G., HSU, S. M., BLAESE, R. M. and JAFFE, E. S. 1985, Malignant lymphoma in patients with the Wiskott–Aldrich syndrome. *Cancer Investigation*, **13**, 515–522.

CUNNINGHAM, M. J., LIANG, S., FUHRMAN, S., SEILHAMER, J. J. and SOMOGYI, R. 2000, Gene expression microarray data analysis for toxicology profiling. *Annals of the New York Academy of Sciences*, **919**, 52–67.

ERIKSSON, L., JOHANSSON, E., KETTANEH-WOLD, N. and WOLD, S. 1999, *Introduction to Multi- and Megavariate Data Analysis Using Projection Methods (PCA & PLS)* (Umea: Umetrics).

FERRANDINA, G., LEGGE, F., RANELLETTI, F. O., ZANNONI, G. F., MAGGIANO, N., EVANGELISTI, A., MANCUSO, S., SCAMBIA, G. and LAURIOLA, L. 2002, Cyclooxygenase-2 expression in endometrial carcinoma: correlation with clinicopathologic parameters and clinical outcome. *Cancer*, **95**, 801–807.

GAVAGHAN, C. L., WILSON, I. D. and NICHOLSON, J. K. 2002, Physiological variation in metabolic phenotyping and functional genomic studies: use of orthogonal signal correction and PLS-DA. *FEBS Letters*, **530**, 191–196.

GERHOLD, D., LU, M., XU, J., AUSTIN, C., CASKEY, C. T. and RUSHMORE, T. 2001, Monitoring expression of genes involved in drug metabolism and toxicology using DNA microarrays. *Physiology and Genomics*, **5**, 161–170.

GOLDENBERG, D., AYESH, S., SCHNEIDER, T., PAPPO, O., JURIM, O., EID, A., FELLIG, Y., DADON, T., ARIEL, I., DE GROOT, N., HOCHBERG, A. and GALUN, E. 2002, Analysis of differentially expressed genes in hepatocellular carcinoma using cDNA arrays. *Molecular Carcinogenesis*, **33**, 113–124.

HAVIV, I. and CAMPBELL, I. G. 2002, DNA microarrays for assessing ovarian cancer gene expression. *Molecular and Cellular Endocrinology*, **191**, 121–126.

IACOBUZIO-DONAHUE, C. A., MAITRA, A., SHEN-ONG, G. L., VAN HEEK, T., ASHFAQ, R., MEYER, R., WALTER, K., BERG, K., HOLLINGSWORTH, M. A., CAMERON, J. L., YEO, C. J., KERN, S. E., GOGGINS, M. and HRUBAN, R. H. 2002, Discovery of novel tumor markers of pancreatic cancer using global gene expression technology. *American Journal of Pathology*, **160**, 1239–1249.

LANDEN, C. N. JR, MATHUR, S. P., RICHARDSON, M. S. and CREASMAN, W. T. 2003, Expression of cyclooxygenase-2 in cervical, endometrial, and ovarian malignancies. *American Journal of Obstetrics and Gynecology*, **188**, 1174–1176.

LUO, J. H., YU, Y. P., CIEPLY, K., LIN, F., DEFLAVIA, P., DHIR, R., FINKELSTEIN, S., MICHALOPOULOS, G. and BECICH, M. 2002, Gene expression analysis of prostate cancers. *Molecular Carcinogenesis*, **33**, 25–35.

LUYPAERT, J., HEUERDING, S., DE JONG, S. and MASSART, D. L. 2002, An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream. *Journal of Pharmacological and Biomedical Analysis*, **30**, 453–466.

MODEL, L. M. 1977, Primary reticulum cell sarcoma of the brain in Wiskott–Aldrich syndrome. Report of a case. *Archives in Neurology*, **34**, 633–635.

MUTTER, G. L., BAAK, J. P., FITZGERALD, J. T., GRAY, R., NEUBERG, D., KUST, G. A., GENTLEMAN, R., GULLANS, S. R., WEI, L. J. and WILCOX, M. 2001, Global expression changes of constitutive and hormonally regulated genes during endometrial neoplastic transformation. *Gynecology and Oncology*, **83**, 177–185.

PEROU, C. M., SORLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H., AKSLEN, L. A., FLUGE, O., PERGAMENSCHIKOV, A., WILLIAMS, C., ZHU, S. X., LONNING, P. E., BORRESEN-DALE, A. L., BROWN, P. O. and BOTSTEIN, D. 2000, Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

RIGHTSLINK

Satyaswaroop, P. G., Bressler, R. S., De la Pena, M. M. and Gurpide, E. 1979, Isolation and culture of human endometrial glands. *Journal of Clinical Endocrinology and Metabolism*, **48**, 639–641.

Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O. and Weinstein, J. N. 2000, A gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, **24**, 236–244.

Schutze, N., Kraft, V., Deerberg, F., Winking, H., Meitinger, D., Ebert, K., Knuppen, R. and Vollmer, G. 1992, Functions of estrogens and anti-estrogens in the rat endometrial adenocarcinoma cell lines. *International Journal of Cancer*, **52**, 941–949.

Shcherbina, A., Candotti, F., Rosen, F. S. and Remold-O'Donnell, E. 2003, High incidence of lymphomas in a subgroup of Wiskott–Aldrich syndrome patients. *British Journal of Haematology*, **121**, 529–530.

Tong, B. J., Tan, J., Tajeda, L., Das, S. K., Chapman, J. A., DuBois, R. N. and Dey, S. K. 2000, Heightened expression of cyclooxygenase-2 and peroxisome proliferator-activated receptor-delta in human endometrial adenocarcinoma. *Neoplasia*, **2**, 483–490.

Waring, J. F., Ciurlionis, R., Jolly, R. A., Heindel, M. and Ulrich, R. G. 2001, Microarray analysis of hepatotoxins *in vitro* reveals a correlation between gene expression profiles and mechanisms of toxicity. *Toxicology Letters*, **120**, 359–368.

Wold, S., Dunn, Albano C., Edlund, W. J., Esbensen, U., Geladi, K., Hellberg, P., Johansson, S., Lindberg, E. and Sjostrom., W. 1984, Multivariate data analysis in chemistry. In *Chemometrics: Mathematics and Statistics in Chemistry*, edited by B. R. Kowalski (Dordrecht: D. Reidel).

Biomarkers Downloaded from informahealthcare.com by Hacettepe Univ. on 11/18/12
For personal use only.

RIGHTSLINK