COMPUTATIONAL ANALYSIS OF MICROARRAY DATA

KOSTAS LYKOSTRATIS, DEPARTMENT OF BIOLOGY, UNIVERSITY OF YORK

INTRODUCTION

Urothelium is a specialised transitional epithelium, which lines the urinary tract and gives rise to the most common form of bladder cancer, transitional cell carcinoma (TCC). In the body, the urothelium is in constant contact with the urine. It consists of three layers (figure 1) and forms a permeability barrier, protecting underlying cells from urine, which potentially contains many genotoxins and non-genotoxins. Urothelium also has the ability to accommodate changes in bladder volume therefore relieving pressure from other organs such as kidneys.

The working hypothesis is that in order to maintain the urinary barrier and protect underlying cells, the urothelium, will respond to genotoxic damage through DNA repair and cell survival, rather than cell death by apoptosis and this will be reflected in characteristic changes in gene expression. The work taking place in the Jack Birch unit aims to use an in vitro model of human urothelial cells to characterise the response to a range of genotoxic agents in terms of both cell behaviour and gene expression changes which may be used as markers of genotoxic exposure.

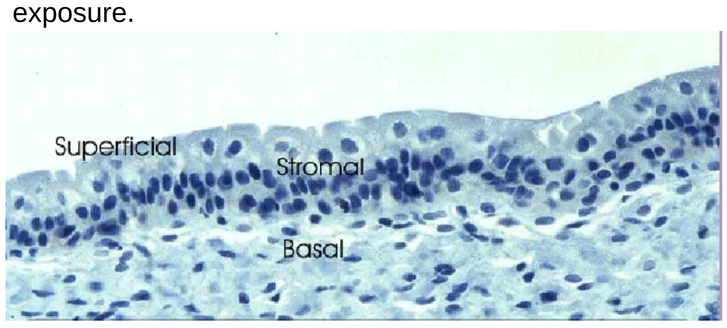


Figure 1. The urothelium

AIMS OF THIS PROJECT

- Process the data obtained by microarray experiment in an appropriate manner.
- Understand the responses to different damaging agents by recognising gene expression levels and profiles
- Identify a selected number of genes that can be used as markers of damage

METHODS

RNA was extracted from NHU (normal human urothelial) monocultures (figure 2) that were grown in serum free medium with a high proliferation rate.

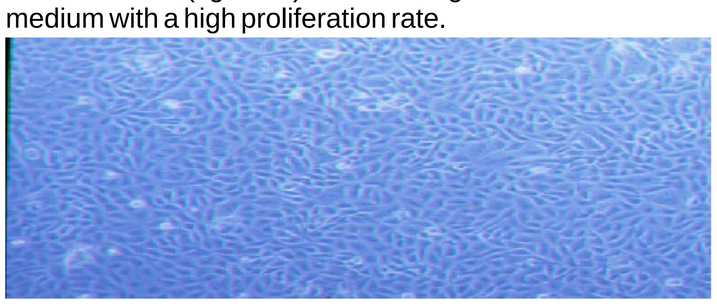


Figure 2. NHU monocultures

Cells were treated with 3 DNA damaging agents:
1) Hydrogen peroxide (h2o2)
2) Mitomycin C (mmc), used to treat superficial bladder cancer
3) N-methyl-N-nitrosourea (mnu), mutagenic, used in Chemotherapy

DNA microarrays: GSK Tox 2.2 model was used (GlaxoSmithkline).Grids contained 3000 DNA sequences. RNA from treated NHU cells were labelled with Green fluorescence. Affymetrix scanner was used to scan the grids and the median spot intensity was measured using the ROBIOVISION program (total intensity normalisation was used to normalise the intensities-get rid of the noise). Correlation of spot positions with sequences was done by a GlaxoSmithkline inhouse program.
From the original set of 3000 sequences used, a large amount was removed due to experimental errors (empty signals, etc). The data set used for the Microarray data analysis consisted of 1611 genes.
For each test sample, the intensity obtained was divided by that of the control. This produced an average ratio of gene expression which was then converted to logarithmic ratios (base 2) to avoid limiting values between 1 and 0 (Microsoft Excel was used for this function). The log (ratios) varied from 1.5 to +~3.
The dataset was modified appropriately to be compatible with formats used by the Microarray programs (tab delimited-txt files: incorporated GWEIGHT, EWEIGHT and GORDER).
The dataset was examined using hierarchical clustering, k-means clustering and PCA analysis.
Four software tools were used overall including Cluster-Treeview (EisenLab), Jexpress, AMANDA and TIGRMultipleExperimentViewer. The last of these was the one used primarily for visualisation of results.

REFERENCES

- Eisen M.B, Spellman P.T, Brown P.O, Botstein D, (1998). Cluster analysis and diplay og genome wide expression patterns. *Proceedings of Natural Academy of Science USA*, **95**:14863-68

- Quackenbush J, (2001). Computational analysis of microarray data. *Nature Genetics Reviews,* **2**(6):418-427

- Thyjaer T, Workman C, Kruhoffer M, Demtroder K, Wolf H, Andersen L.D, Frederiksen C.M, Knudsen S, Orntoft T.F, (2001). Identification of Gene Expression Patterns in Superficial and Invasive Human Bladder Cancer. *Cancer Research*, **61**:2492-299

RESULTS

Hierachical clustering was performed using the Euclidean distance by both TIGR-MultipleExperimentViewer and Jexpress. Due to the nature of the expression dataset used (limited variation of fold differences) the clusters were undefined in a multiple chain manner without revealing any distinct pattern clamps. PCA analysis was performed to reveal any present classification patterns but the unimodal distribution of expression in the most informative two dimensions suggested that the genes do not fall into well-defined clusters (figure 3).

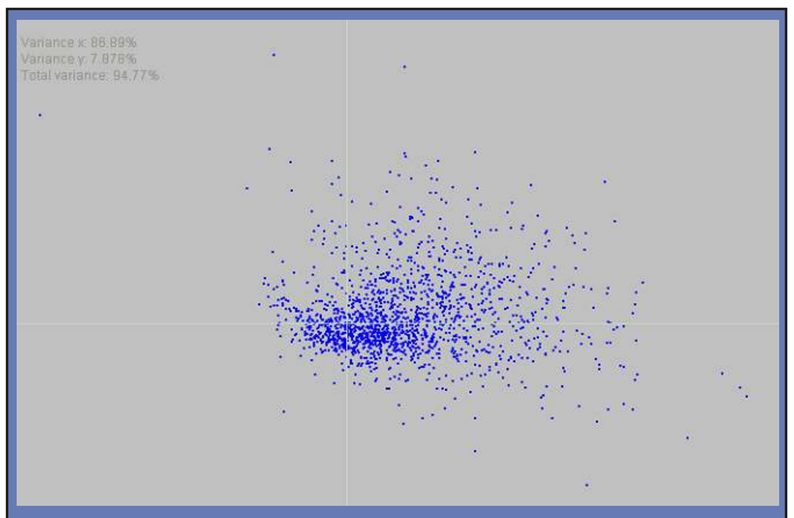


Figure 3. PCA analysis. Unimodal distribution of expression

An SQL query (via Microsoft Access) was used to isolate only the most upregulated (more then 2-fold increase) and downregulated (more that 1-fold decrease) genes for the analysis.
h2o2>=1 OR mmc>=1 OR mnu>=1 AND h2o2<= -0.5 OR mmc<= -0.5 OR mnu<= -0.5;
The new set of genes was used again for hierarchical clustering analysis (average-linkage clustering was used since the complete and single linkage methods are inappropriate in the absence of distinct clamps of clusters). This analysis performed well and 3 major clusters were identified (figure 4). PCA analysis also suggested the same results (data not shown).

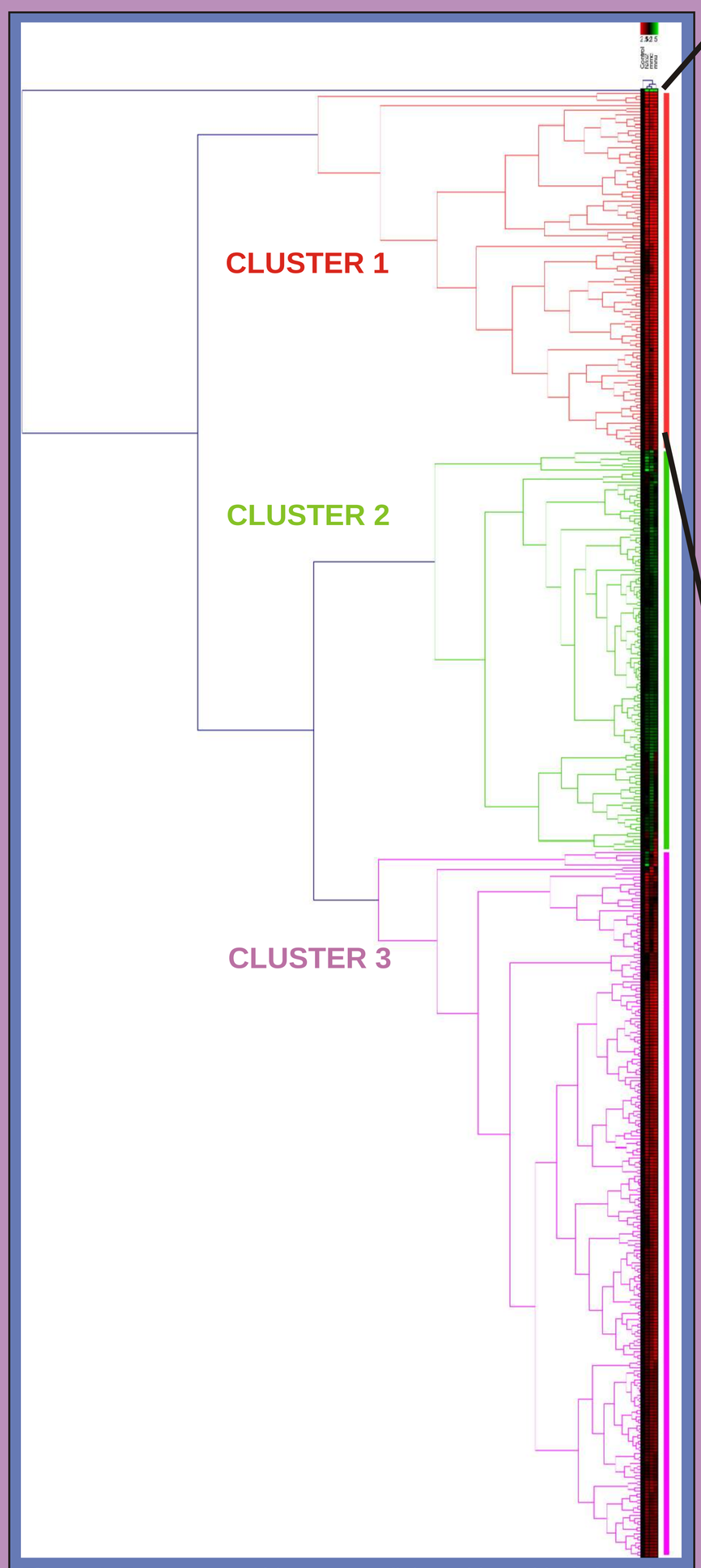


Figure 4. Tree generated by average-linkage hierarchical clustering analysis (Euclidean distance). The three clusters are shown in three different colors: red, green and purple. These indicate high upregulation, low downregulation and low upregulation respectively. The tree was constructed using the TIGRMultipleExperimentViewer.

Since 3 obvious clusters were generated by hierarchical clustering and PCA analysis, the same dataset was used for k-means clustering with k=3. The clustering results showed distinct patterns of gene expression (figure 5).

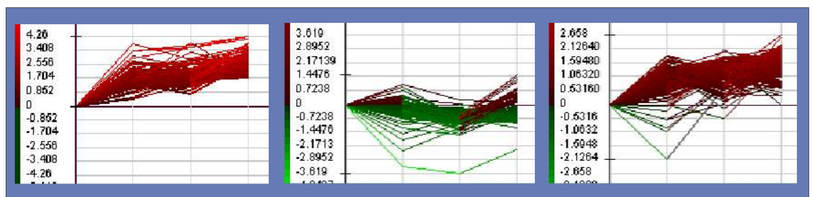


Figure 5. K-means using JEXPRESS: The expression profiles of the three clusters containing 125, 232 and 119 genes respectively.

For a more accurate analysis of gene expression, cluster 1 was selected for further examination since was the only one containing significant fold differences (high upregulation).

The expression profile of cluster 1 (figure 7) in respect to the different treatments is shown in figure 6. Mnu results in the greatest increase of gene expression levels overall compared to the levels generated from the two other treatments.

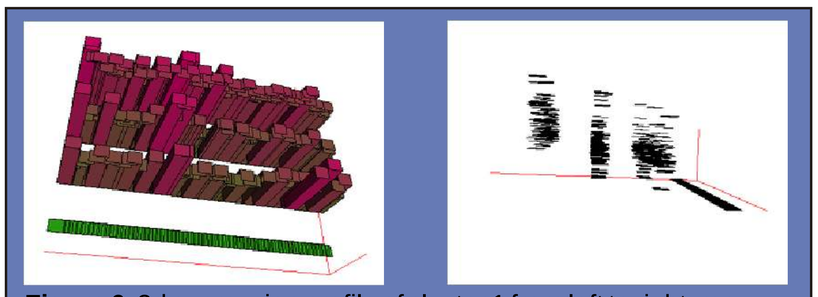
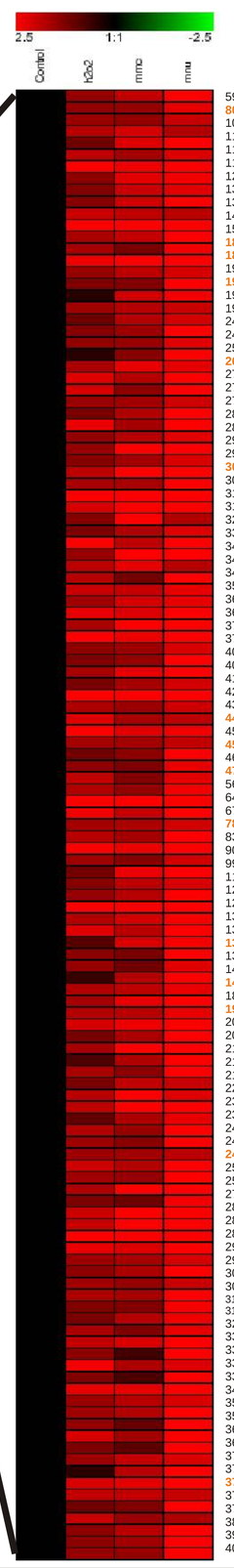


Figure 6. 3d expression profile of cluster 1 from left to right: mnu mmc and h2o2.

Figure 7. Cluster 1 shown in detail



DISCUSSION

Cluster 1 shows upregulated groups of genes that are involved in different cell processes. Presence of DNA repair genes such as MSH6, PMS2 and PCNA, supports the working hypothesis for urothelium repair following DNA damage. However, genes such as Caspase-2, caspase-6 and ICAD-s indicate that apoptosis might be also occurring (in cells beyond repair).
Genes involved in cell cycle and signalling (e.g. CyclinD2, cdc-2, PI3Ks, B-raf) also show upregulation, possibly indicating cell recovery. Finally, genes involved in the immune response (TNF alpha, interferon inducible factor 10) are also upregulated as expected after DNA damage.
The selected set of genes can be further tested (Self Organising Maps) or coupled experimentally with Northern blotting or real time PCR to isolate a small number of genes that can be used as markers of DNA damage.
To date, there is no standard method for microarray data analysis and the currently available tools need to be optimised for more efficient data interpretation.

Many thanks to Prof J.Southgate and R.Grallan for their advise, discussions, and providing the data and tools needed for the analysis.