

ArrayShine: An Excel Program for Transforming Gene Expression Data into Color-Coded Molecular Signatures or Fingerprints

H Khan

Citation

H Khan. *ArrayShine: An Excel Program for Transforming Gene Expression Data into Color-Coded Molecular Signatures or Fingerprints*. The Internet Journal of Oncology. 2003 Volume 2 Number 1.

Abstract

The advent of microarray has revolutionized the pace of research in understanding the complex pathophysiology of cancer and developing novel diagnostic and prognostic markers. Microarray gene expression profiling can be used to define molecular 'signatures' or 'fingerprints', which are supposed to be the most powerful tools for cancer management in the near future. The analysis and interpretation of microarray data are tedious, time-consuming and error-prone tasks. Moreover, a continuous upsurge in the gene expression data has hampered the application of microarrays in routine clinical practice. Graphics is a powerful tool to simplify data presentation, and could also be helpful in reducing the prevailing complexities of tabular expression data. The aim of this study was to develop a procedure for transforming numeric expression data into color-coded graphical output for better visualization and ease in comparison. The software, ArrayShine, developed in Microsoft Excel has been validated for converting numeric expression data into color-coded arrays, sorting genes in ascending order of expression, filtering differentially expressed genes and creating gene signatures for visual comparison. The versatility of software in dealing with a single gene to several thousand genes furthers its utilization for expression data from various sources including high-density microarrays, custom macroarrays and reverse-transcription polymerase chain reaction (RT-PCR). The software will be particularly useful for small laboratories that cannot afford the expense of skilled computer professional and of professionally developed software tools.

INTRODUCTION

The development of cDNA microarray technology for rapid expression profiling of thousands of genes in a single hybridization step has tempted the researchers to utilize this technique for unfolding the mysteries of genetic diseases. One of the major areas in which microarrays have been extensively utilized is cancer (1). The pattern of expressed genes on a microarray demonstrates a typical profile in relation to cancer type or disease severity. These unique sets of genes defining specific pathophysiology are regarded as 'molecular signatures' or 'fingerprints'. Tumors with closely related genetic lesions will have similar signatures and also will be expected to have similar clinical behaviors (2). The information encoded in gene signatures can provide valuable insights in cancer diagnosis and prognosis (3,4,5,6,7).

In the recent years, numerous software tools have been developed to reduce the complexities of microarray data and to extract meaningful interpretation of results (8,9,10,11,12,13). However, only fewer attempts have been made towards

exploring the potentials of graphical presentation of expression data (14,15). The commercial software for the same purpose are costly and often beyond the reach of small laboratories of developing countries. Whereas, simple and inexpensive software could help to bring the benefits of microarray-based clinical research to many less-developed centers. By realizing the fact that usage of tabular expression data for routine clinical practice might be more complex and tedious, a computer program, ArrayShine, has been developed for transformation of numeric gene expression data into color-coded graphical output. This multi-purpose software performs following tasks: (i) to convert numeric expression data into color-coded arrays, (ii) to sort genes in ascending order of expression, (iii) to filter differentially expressed genes and (iv) to create array signatures or fingerprints for ease in visual comparison.

METHODS

SOFTWARE DESIGN

The ArrayShine program has been developed in Microsoft

Excel (Version 2000). The program is composed of two worksheets, one for data entry and the other for report display.

DATA ENTRY WINDOW

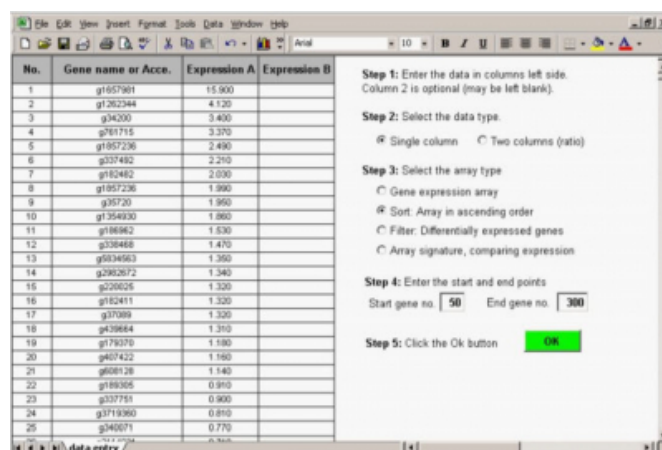
The data entry window (Fig. 1) contains columns for data entry, two option buttons for choosing appropriate data type, four option buttons for selecting output type, two input boxes to specify the target range (start and end points) within the expression data, and a command button to which a macro has been assigned to execute the program. There are four columns (Excel worksheet columns) for data entry. Column 1 is for serial number of gene and column 2 for gene name (or accession number). Columns 3 and 4 are designated for gene expression data. Based on the option chosen, the program either uses the data in column 3 alone or computes the ratios of values entered in column 3 (always numerator) and column 4, in order to generate the array sets. Usually single column option has to be used for normalized data, however, if ratios of various gene probes to housekeeping gene (standard) or ratios of sample to control are intended, two-column option should be selected.

REPORT WINDOW

The color-coded gene expression profiles are displayed on a new Excel worksheet. There are four types of array outputs including (i) expression of all the genes selected, in the same order (ii) sorting genes in ascending order of expression, (iii) filtering differentially expressed genes; only filtered genes are displayed in the array and (iv) visual comparison of gene signatures, multiple clusters are displayed for a comparative view. The graphic output of gene expression data is a collection of color-coded squares, spanning horizontally left to right (10 squares in each row) and expanding vertically downwards. The gene expression ratio has been classified into seven categories (different color codes), three for down-regulation (light to dark blue), three for up-regulation (light to dark red) and one for norm-regulation (gray); yellow color is used to identify missing data. The report also shows a table of genes in the array, number of genes in various categories of expression, and the scale of color-coding (Fig. 2). There are two buttons (Next and Reset) on the report window, the former is used to display data entry window and the later for clearing the report window (array signature mode).

Figure 1

Figure 1: Data entry window of the ArrayGraphic software.



PROCEDURE FOR CREATING VISUAL ARRAYS

The steps involved in creating a color-coded array of gene expression data are also shown in Fig. 1. Briefly, enter the data in the respective columns of data entry sheet. Large expression data can be conveniently copy-pasted from another source file. Then choose the array type by selecting one of the option buttons, specify the data range by inputting start and end points, and run the program by clicking the 'OK' button. For each execution, the expression profiling of one sample is processed and the procedure has to be repeated for multiple analysis. It is convenient to keep the original data file open for transferring (copy-paste of entire column is most suitable) data to ArrayShine datasheet.

SOFTWARE VALIDATION

Three different types of data sets were selected from the published studies (_{16,17,18}) to validate the applications of ArrayShine software. To ensure the functionality of software with large data, the compressed Excel file containing the normalized expression data for 13638 genes (₁₆) was downloaded from the website. The data in column 1 (Spot No.), column 2 (Genbank Accession No.) and column 4 (Day 0 vs 2) of the data file were copied and pasted to ArrayShine worksheet. The program was executed and successfully validated for all the 4 array options. A representative output of option 3 (filtered genes) is shown in Fig. 2. The data reported by Bull et al (₁₇) were used to validate the software application for small data sets as the case of RT-PCR. The information about serial number, gene identification number and expression level of 28 genes specific to prostate cancer were entered in 'data entry' window to create the respective graphical output (Fig. 3).

Figure 2

Figure 2: Output window: showing color-coded array of filtered genes, summary of genes in various categories of expression, color-scale, and tabular data.

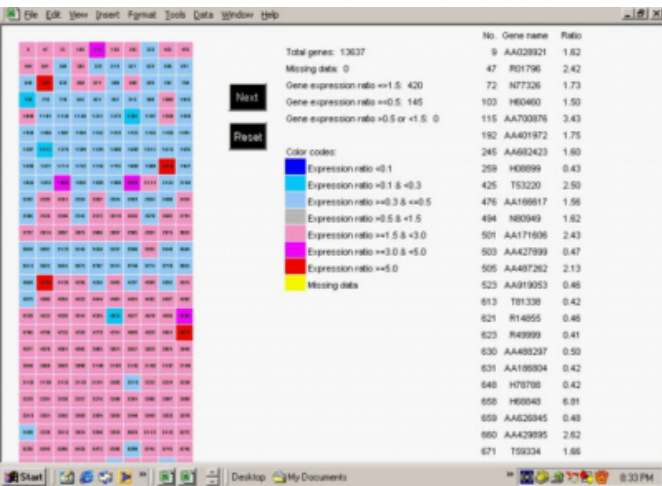


Figure 3

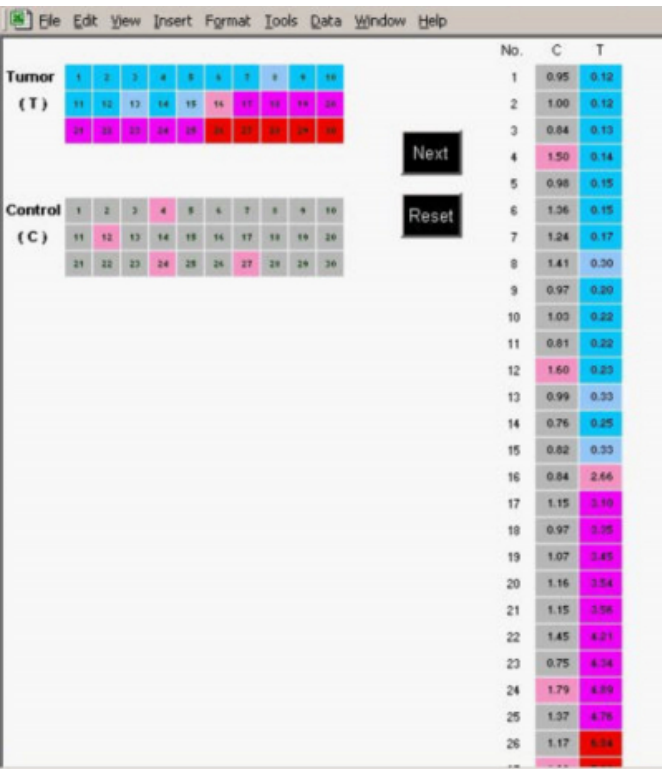
Figure 3: Output window: showing the color-coded array of all 28 genes entered in the program; partial view of other information is also seen.



In order to validate the creation of multiple array signatures (option 4, for visual comparison), expression data of 30 differentially expressed genes in ovarian cancer samples were used (18). The average signal ratios with control probes were entered and the program was run to create array signature followed by clicking the 'next' button and running the program again after inputting the signal ratios with tumor probes. The difference in the two sets can be clearly observed from the resulting output (Fig. 4).

Figure 4

Figure 4: Output window: showing a comparative view of gene expression in two groups (control and tumor specimens). Clicking the 'Next' button prompts for the entry of next group; 'Reset' button is used to clear the screen.



DISCUSSION

High-density microarrays possess unmatched supremacy for preliminary screening of differentially expressed genes, which can be used to design a simpler and less expensive diagnostic chip for rapid molecular characterization of cancers (2). Since only selected genes constitute molecular signatures, they can also be analyzed by conventional RT-PCR, without using the sophistication of microarray technology, which is still beyond the reach of many third world laboratories. Bull et al (17) have suggested that a smaller number of potential markers could be assessed more conveniently in biopsy samples using RT-PCR. Recently published data (17,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37) on cancer genetics clearly indicate that only a small fraction of total genes on a microarray show differential expression (Table 1). Based on these findings, the number of genes with differential expression ranged between 3 and 176 (mean, 59 13.19) as compared to total genes on microarrays (range, 425-25000; mean, 4924 1367.33). However, microarray expression profiling could be efficiently utilized to shortlist genes that could be helpful in molecular diagnosis of cancers

(17). Thus, designing of macro-arrays or RT-PCR methods might be more realistic approaches in generating useful data for accurately classifying the tumor type and improving therapeutic decisions (4,17,38).

Figure 5

Table 1: Comparative view of differentially expressed genes in microarray studies (during 2000-2002) on various cancer types.

Cancer type	Total number of genes in microarray	Differentially expressed genes	
		Number	%
Breast cancer (19)	5361	176	3.28
Colorectal cancer (20)	2280	8	0.35
Colorectal cancer (21)	4608	59	1.28
Colorectal cancer (22)	1200	14	1.16
Drug resistant cancer cells (23)	1176	28	2.38
Esophageal cancer (24)	9216	52	0.56
Gastric cancer (25)	6800	162	2.38
Gastric cancer cells (26)	6800	32	0.47
Hepatocellular carcinoma (27)	597	3	0.50
Hepatocellular carcinoma (28)	14000	156	1.11
Lung cancer cell lines (29)	588	163	27.72
Lung cancer cell lines (30)	600	17	2.83
Lung cancer cell lines (31)	425	3	0.70
Malignant mesothelioma cells (32)	588	39	6.63
Ovarian cancer (33)	25000	56	0.22
Ovarian cancer (34)	9121	103	1.13
Prostate cancer (17)	1877	17	0.90
Prostate cancer (35)	588	19	3.23
Pulmonary metastasis (36)	7070	15	0.21
Thyroid cancer (37)	588	58	9.86

The major steps involved in configuring molecular signatures include the extraction of useful information from microarrays and its subsequent transformation into a format suitable for routine application. The results of our software validation clearly show that ArrayShine program can be efficiently used to convert numeric gene expression data into color-coded arrays and to filter useful information from large microarray data (Fig. 2). The program automatically generates array signatures using the filtered information of differentially expressed genes. Alternatively, small data sets, or pre-filtered data can be directly fed in the program for construction (Fig. 3) and comparison of array signatures (Fig. 4). The visual fingerprints defining a particular cancer type, sub-class or severity stage, can be stored in computer drives or made available on a hard copy for future reference.

The selection of Microsoft Excel spreadsheet for the development of ArrayShine was based on the fact that Excel provides sufficient computational and visualization power for robust analysis of microarray data (15,39). Moreover, the availability of microarray data in Excel format and the familiarity of clinicians and scientists with Excel were also the important considerations while choosing the Excel platform for developing this program. Recently, Schageman et al (15) have also used Excel to develop a tool for the

analysis and visualization microarray data. However, the graphical output with the use of ArrayShine is a typical prototype of microarray signals, in contrast to color-coded scatter plots reported earlier (15). Since the ArrayShine program is capable of dealing with a minimum of one gene to several thousand genes it finds wider application for dealing with expression data from high/low density microarrays, custom manual arrays, and conventional RT-PCR experiments. ArrayShine software may also be helpful in developing a virtual library of cancer prognostic or diagnostic markers in the form of color-coded collection of specific probes. Further studies are in progress to upgrade this tool with interactive selection and statistical evaluation of microarray data.

AVAILABILITY OF SOFTWARE

The ArrayShine software (1.6 MB) can be obtained from the Author by sending a letter of request.

ACKNOWLEDGMENTS

The author expresses sincere thanks to Dr. John M. Mariadason, Albert Einstein Cancer Center, New York (USA), Dr. J.H. Bull, CMC International, Cheshire (UK), and Dr. Kai Wang, Chiroscience R&D, Inc., Bothell (USA), whose valuable data have been used to validate the functionality of ArrayShine software.

CORRESPONDENCE TO

Haseeb Ahmad Khan PhD, MRSC (UK) Research Center, Armed Forces Hospital T-835, P.O. Box 7897, Riyadh 11159 Kingdom of Saudi Arabia. E-mail: khan_haseeb@yahoo.com

References

1. Rew DA. DNA microarray technology in cancer research. *Eur J Surg Oncol* 2001;27: 504-8.
2. Ladanyi M, Chan WC, Triche TJ, Gerald WL. Expression profiling of human tumors: the end of surgical pathology. *J Mol Diagnos* 2001;3:92-7.
3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286:531-7.
4. Martin KJ, Kritzman BM, Price LM, Koh B, Kwan CP, Zhang X, Mackay A, O'Hare MJ, Kaelin CM, Mutter GL, Pardee AB, Sager R. Linking gene expression patterns to therapeutic groups in breast cancer. *Cancer Res* 2000;60:2232-8.
5. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JJ, Yang L, Marti GE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;403:503-11.

6. Chuma M, Sakamoto M, Yamazaki K, Ohta T, Ohki M, Asaka M, Hirohashi S. Expression profiling in multistage hepatocarcinogenesis: identification of HSP70 as a molecular marker of early hepatocellular carcinoma. *Hepatology* 2003;37:198-207.
7. Tan ZJ, Hu XG, Cao GS, Tang Y. Analysis of gene expression profile of pancreatic carcinoma using cDNA microarray. *World J Gastroenterol* 2003;9:818-23.
8. Covell DG, Wallqvist A, Rabow AA, Thanki N. Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data. *Mol Cancer Ther* 2003;2:317-32.
9. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 2003;4:R28.
10. Martoglio AM, Miskin JW, Smith SK, MacKay DJ. A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics* 2002;18:1617-24.
11. Dudoit S, Gentleman RC, Quackenbush J. Open source software for the analysis of microarray data. *Biotechniques* 2003;Suppl:45-51.
12. Breitkreutz BJ, Jorgensen P, Breitkreutz A, Tyres M. AFM 4.0: a toolbox for DNA microarray analysis. *Genome Biol* 2001;2: software 1.1-1.3.
13. Peterson LE. CLUSFAVOR 5.0: hierarchical cluster and principal-component analysis of microarray-based transcriptional profiles. *Genome Biol* 2002;24:SOFTWARE0002.
14. Pavlidis P, Noble WS. Matrix2png: a utility for visualizing matrix data. *Bioinformatics* 2003;19:295-6.
15. Schageman JJ, Basit M, Gallardo TD, Garner HR, Shohet RV. MarC-V: a spreadsheet-based tool for analysis, normalization, and visualization of single cDNA microarray experiments. *Biotechniques* 2002;32:338-44.
16. Mariadason JM, Arango D, Corner GA, Aranes MJ, Hotchkiss KA, Yang W, Augenlicht LH. A gene expression profile that defines colon cell maturation in vitro. *Cancer Res* 2002;62:4791-804.
17. Bull JH, Ellison G, Patel A, Muir G, Walker M, Underwood M, Khan F, Paskins L. Identification of potential diagnostic markers of prostate cancer and prostatic intraepithelial neoplasia using cDNA microarray. *Br J Cancer* 2001;84:1512-9.
18. Wang K, Gan L, Jeffery E, Gayle M, Gown AM, Skelly M, Nelson PS, Ng WV, Schummer M, Hood L, Mulligan J. Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* 1999;229:101-8.
19. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J. Gene-expression profiles in hereditary breast cancer. *N Eng J Med* 2001;344:539-48.
20. Otsuka M, Kato M, Yoshikawa T, Chen H, Brown EJ, Masuho Y, Omato M, Seki N. Differential expression of the L-plastin gene in human colorectal cancer progression and metastasis. *Biochem Biophys Res Commun* 2001;289:876-81.
21. Takemasa I, Higuchi H, Yamamoto H, Sekimoto M, Tomita N, Nakamori S, Matoba R, Monden M, Matsubara K. Construction of preferential cDNA microarray specialized for human colorectal carcinoma: molecular sketch of colorectal cancer. *Biochem Biophys Res Commun* 2001;285:1244-9.
22. Hernandez A, Smith F, Wang Q, Wang X, Evers BM. Assessment of differential gene expression patterns in human colon cancers. *Ann Surg* 2000;232:576-85.
23. Wang W, Marsh S, Cassidy J, McLeod HL. Pharmacogenomic dissection of resistance to thymidylate synthase inhibitors. *Cancer Res* 2001;61:5505-10.
24. Kihara C, Tsunoda T, Tanaka T, Yamana H, Furukawa Y, Ono K, Kitahara O, Zembutsu H, Yanagawa R, Hirata K, Takagi T, Nakamura Y. Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene-expression profiles. *Cancer Res* 2001;61:6474-9.
25. Yoshitaka H, Hirokazu T, Shuichi T, Naoko M, Ja-Mun C, Masashi F, Tatsuhiko K, Hiroyuki A. Global gene expression analysis of gastric cancer by oligonucleotide microarrays. *Cancer Res* 2002;62:233-40.
26. Hippo Y, Yashiro M, Ishii M, Taniguchi H, Tsutsumi S, Hirakawa K, Kodama T, Aburatani H. Differential gene expression profiles of scirrhous gastric cancer cells with high metastatic potential to peritoneum or lymph nodes. *Cancer Res* 2001;61: 889-95.
27. Daniel G, Suhail A, Tamar S, Orit P, Oded J, Ahmed E, Yakov F, Tikva D, Ilana A, Nathan G, Abraham H, Eithan G. Analysis of differentially expressed genes in hepatocellular carcinoma using cDNA arrays. *Mol Carcinogen* 2002;33:113-24.
28. Xu L, Hui L, Wang S, Gong J, Jin Y, Wang Y, Ji Y, Wu X, Han Z, Hu G. Expression profiling suggested a regulatory role of liver-enriched transcription factors in human hepatocellular carcinoma. *Cancer Res* 2001;61:3176-81.
29. Okabe S, Fujimoto N, Sueoka N, Suganuma M, Fujiki H. Modulation of gene expression by (-)-epigallocatechin gallate in PC-9 cells using a cDNA expression array. *Biol Phram Bull* 2001;24:883-6.
30. Hellmann GM, Fields WR, Doolittle DJ. Gene expression profiling of cultured human bronchial epithelial and lung carcinoma cells. *Toxicol Sci* 2001;61:154-63.
31. Kiguchi T, Niiya K, Shibakura M, Miyazono T, Shinagawa K, Ishimaru F, Kiura K, Ikeda K, Nakata Y, Harada M. Induction of urokinase-type plasminogen activator by the anthracycline antibiotic in human RC-K8 lymphoma and H69 lung-carcinoma cells. *Int. J. Cancer* 2001;93:792-7.
32. Kettunen E, Nissen AM, Ollikainen T, Taavitsainen M, Tapper J, Mattson K, Linnainmaa K, Knuutila S, El-Rifai W. Gene expression profiling of malignant mesothelioma cell lines: cDNA array study. *Int J Cancer* 2001;91:492-6.
33. Sridhar V, Lee J, Pandita A, Iturria S, Avula R, Staub J, Morrissey M, Calhoun E, Sen A, Kalli K, Keeney G, Roche P et al. Genetic analysis of early- versus late-stage ovarian tumors. *Cancer Res* 2001;61:5895-904.
34. Ono K, Tanaka T, Tsunoda T, Kitahara O, Kihara C, Okamoto A, Ochiai K, Takagi T, Nakamura Y. Identification by cDNA microarray of genes involved in ovarian carcinogenesis. *Cancer Res* 2000;60:5007-11.
35. Elek J, Park KH, Narayanan R. Microarray-based expression profiling in prostate tumors. *In Vivo* 2000;14:173-82.
36. Clark EA, Golub TR, Lander ES, Hynes RO. Genomic analysis of metastasis reveals an essential role for RhoC. *Nature* 2000;3:532-5.
37. Chen KT, Lin JD, Chao TC, Hsueh C, Chang CA, Weng HF, Chan EC. Identifying differentially expressed genes associated with metastasis of follicular thyroid cancer by cDNA expression array. *Thyroid* 2001;11:41-6.
38. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF, Hampton GM. Molecular classification of human

carcinomas by use of gene expression signatures. Cancer Res 2001;61:7388-93.

39. Conway T, Kraus B, Tucker DL, Smalley DJ, Dorman AF, McKibben L. DNA array analysis in a Microsoft windows environment. Biotechniques 2002;110:112-9.

Author Information

Haseeb Ahmad Khan, PhD, MRSC (UK)

Research Center, Armed Forces Hospital