Spring 2004

Non-parametric algorithms for evaluating gene expression in cancer using DNA microarray technology

Virginie Aris
New Jersey Institute of Technology

ABSTRACT

NON-PARAMETRIC ALGORITHMS FOR EVALUATING GENE EXPRESSION IN CANCER USING DNA MICROARRY TECHNOLOGY

by

Virginie Aris

Microarray technology has transformed the field of cancer biology by enabling the simultaneous evaluation of tens of thousands mRNA expression levels in a single experiment. This technology has been applied to medical science in order to find gene expression markers that cluster diseased and normal tissues, genes affected by treatments, and gene network interactions. All methods of microarray data analysis can be summarized as a study of differential gene expression. This study addresses three questions, 1) the roles of selectively expressed genes for the classification of cancer, 2) issues of accounting for both experimental and biological noise, and 3) issues of comparing data derived from different research groups using the Affymetrix GeneChip™ platform. A key finding of this study is that selectively expressed genes are very powerful when used for disease classification. A model was designed to reduce noise and eliminate false positives from true results. With this approach, data from different research groups can be integrated to increase information and enable a better understanding of cancer.

NON-PARAMETRIC ALGORITHMS FOR EVALUATING GENE EXPRESSION IN CANCER USING DNA MICROARRY TECHNOLOGY

by
Virginie Aris

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey - Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Biology

Federated Biological Sciences Department

May 2004

APPROVAL PAGE

NON-PARAMETRIC ALGORITHMS FOR EVALUATING GENE EXPRESSION IN CANCER USING DNA MICROARRY TECHNOLOGY

Virginie Aris

Dr. Michael Recce, Dissertation Advisor — Date
Associate Professor, Information Systems, NJIT

Dr. Peter P. Tolias, Dissertation Co-Advisor — Date
Vice President, Advanced Research and Technology Assessment Worldwide,
Ortho-Clinical Diagnostics,
Johnson & Johnson Company

Dr. Marvin Schwalb, Committee Member — Date
Professor of Microbiology and Molecular Genetics,
UMDNJ- New Jersey Medical School

Dr. Farzan Nadim, Committee Member — Date
Associate Professor, Dept. Mathematics, NJIT
Dept. Biology, Rutgers University

Dr. Ronald P. Hart, Committee Member — Date
Professor, Cell Biology & Neuroscience, Rutgers University

BIOGRAPHICAL SKETCH

Author: Virginie Aris

Degree: Doctor of Philosophy

Date: May 2004

Undergraduate and Graduate Education

- Doctor of Philosophy in Computational Biology
  New Jersey Institute of Technology, Newark, NJ. 2004

- Masters of Sciences in Plant Pathology
  North Carolina State University, Raleigh, NC, 1999

- Agronomic Engineering Degree
  Ecole Nationale Superieure d'Agronomie et des Industries Alimentaires, Nancy, France, 1997

- DEUG in Biology and Life Sciences
  Universite Montpellier II, Montpellier France, 1994

Major: Computational Biology

Presentations and Publications:

Aris, V. and Recce, M. A Method to Improve Detection of Disease Using Selectively Expressed Genes in Microarray Data.. in Methods of Microarray Data Analysis, Papers from CAMDA '00. (eds. Lin, S. M. and Johnson, K. F.) Kluwer Academic Publishers, Boston, USA., 2002.

Ramanathan Y., Haibo Zhang, Virginie Aris, Patricia Soteropoulos, Stuart A. Aaronson, and Peter P. Tolias, 2002. Functional Cloning, Sorting and Expression Profiling of Nucleic-Acid Binding Proteins. *Genome Research.*12, 1175-1189.

Aris, V.M., J. Hu, P. Soteropoulos, M. Recce, and P.P. Tolias, 2002. Application of the Directional change Assessment Algorithm on microarray data to underscore markers that distinguish ovarian, breast and prostate cancer. Oncogenomics May1-5, 2002, Dublin, Ireland.

Aris, V. and Recce, M., 2000. A Method to Improve Detection of Disease Using Selectively Expressed Genes in Microarray Data. in Methods of Microarray Data Analysis, CAMDA '00. Duke University, NC.

Aris, V. M., S. Leath, J. E. Bailey, and C. L. Campbell. 1999. Modeling the vertical Spread of *Stagonospora nodorum* Epidemics on Winter Wheat. Phytopathology, Vol. 86 No. 6 (Supplement) p. S3.

Aris, V. M., and J. E., Bailey, 1998. Adapting a Weather-Based Leafspot Advisory on Peanut to Partially resistant Genotypes. Presentation at the APRES meeting Norfolk Virginia. APRES Proceedings Vol. 30 p. 22.

I would like to dedicate this dissertation to my family, cancer patients and survivors.

Your strength and persistence inspired me.

Thank you.

"Try with all your might -- work very, very hard -- to make the world a better place.

But if all your efforts are to no avail -- no hard feelings!"

Dalai Lama

ACKNOWLEDGMENT

I would like to thank Dr. Michael Recce, who not only steered me into the field of Computational Biology or more specifically into the analysis of microarrays, but also was a great source of inspiration and ideas. I would like to express my deepest appreciation to Dr. Peter Tolias who funded this project with the NIH grant CA83213 from the National Cancer Institute and let me work in his lab with an amazing staff and cutting edge technology. Many thanks to Dr. James Dermody and Dr. Marvin Schwalb for financially supporting my last year of graduate school. Special thanks to Dr. Ron Hart, Dr. Farzan Nadim, and Dr. Marvin Schwalb who actively participated in my committee.

I am very grateful to have been able to work at the Center for Applied Genomics. I would like to thank the Director Dr. Patricia Soteropoulos, the staff members (past and present) Saleena Ghanny, Anthony Galante, Michael Cody, Donna Wilson, Tongsheng Wang, Jun Hu, and Anbing Shi, the postdocs Dr. Ramanathan and Dr. Papasotiropoulos, and my fellow students Jeff Cheng, Haibo Zhang and Filipo Posta.

I would like to thank Karen Gansner, Amy Trimarco and Clarisa González-Lenahan for helping me with the "administrative" side of this doctorate.

I also want to thank Jason Lambert for all his help and moral support this past year.

TABLE OF CONTENT

TABLE OF CONTENT
(Continued)

Chapter **Page**

TABLE OF CONTENT
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Cancer Background

The term cancer encompasses many conditions characterized by uncontrolled proliferation of cells. Almost all organs and cell types can undergo oncogenic transformations, with an array of mechanisms and outcomes. Cancer arises through a succession of events making it more preeminent in the aging population. The causes are many and varied including: genetic predisposition, environmental influences, diet, cigarette smoking, infectious agents and aging. The complexity and diversity of the regulatory and downstream effector pathways affected by cancer has hindered the development of effective and specific therapies.

Cancer accounts for 23% of the deaths in the United States annually. It is the second leading cause of death behind heart disease (Jemal et al., 2003; Simmonds, 2003). Approximately 1,334,100 new cases of invasive cancer were diagnosed in 2003, more than one million cases of squamous cell skin cancer, 55,700 cases of breast carcinoma, and 37,700 cases of melanoma. An estimated 556,500 people died from cancer in 2003. The three most commonly diagnosed cancers for men are: prostate (33%), lung (14%), and colon (11%). For women, the three most commonly diagnosed cancers are: breast (33%), lung (12%) and colon (11%). The leading causes of cancer deaths are the same three as above for each gender. More specifically, lung cancer surpasses all other cancers in terms of fatalities being responsible for 31% of the death by cancer for men, and 25% for women. Prostate cancer accounts for 10% of the cancer death toll for men, and breast

cancer accounts for 15% of the death by cancer in women. Colon cancer accounts for 10% of the deaths for men and 11% for women. Cancer is the primary cause of death among women aged from 40 to 79 and men aged from 60 to 79.

Due to early detection through increased screening, the survival rate has increased and the death toll decreased slightly in the last few years (Black and Welch, 1993). Eighty-five percent of the prostate cancers are diagnosed at the local or regional stage with a five-year survival rate of 100%. Unfortunately, most cancer therapies have a limited efficacy when the disease is treated in its later stages.

To improve survival, the key points are: 1) early diagnosis 2) more effective drugs to treat later stages 3) a better prediction of the treatment response and 4) chemo-prevention of tumorgenesis (Ochs and Godwin, 2003). Biomarkers can help with all those key points. Gene expression markers, by definition, are genes that are consistently up-regulated or down-regulated in cancer samples compared to normal samples. The hypothesis is that the consistency is the sign that the genes in question are part or downstream of regulatory pathways necessary for tumorgenesis. Gene markers are obviously important for early diagnosis and can help define the therapeutic course of action i.e. treatment for tumor with hormone receptors in breast cancer (Shenkier et al., 2004). Biomarkers can become targets for drug development and can help predict treatment response. A detailed study of their function might explain the events leading to their de-regulation and help to design chemo-prevention therapies. Two factors are important in identifying biomarkers: the ability to define the cancer group and subgroups and the use of high-through-put methods such as microarray technology to discover genes

that are consistently up-regulated or down regulated. The second issue is discussed in this dissertation.

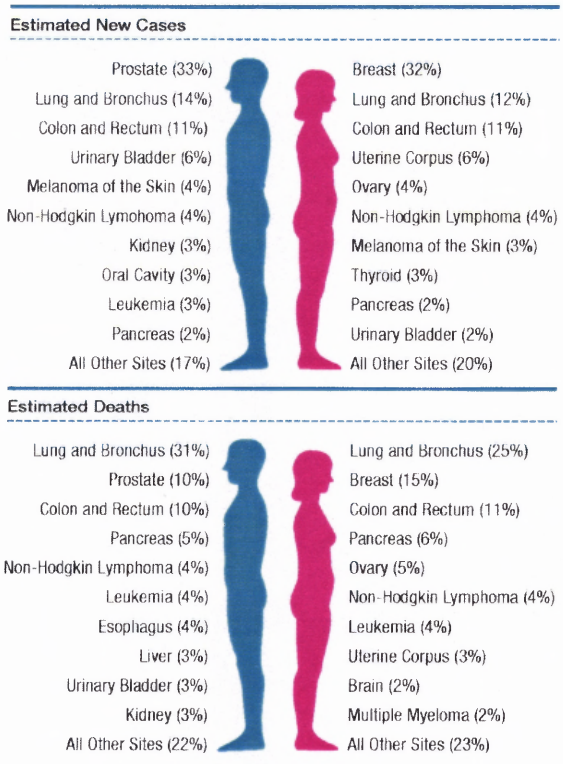


Figure 1.1 Ten leading cancer types for the estimated new cancer cases and deaths, by sex, US, 2003* excluding skin cancers and in situ carcinomas except urinary bladder (Jemal et al., 2003).

The normal cell can be viewed as a system in which functions are tightly regulated. There are even fail-safe and redundant checks balances to ensure that the cell functions the way it should. In contrast cancer cells present an accumulation of replication disorders arising through gradual accumulation of genetic changes. Typical cancer cells contain combinations of genetic changes that alter gene expression causing the cell to escape the checks and controls that prevent proliferation and metastasis. A single mutation event is generally not enough to circumvent all the different safeguards a

complex organism has built in. Instead, two or more sequential events are needed to initiate the transformation of a normal to a malignant cell (Vogelstein and Kinzler, 1993).

Cancer is thought to arise from the clonal proliferation of a single cell, requiring the activation of the cell cycle and the overriding of some cell functions, in particular the cell death program (apoptosis). During proliferation, mutations, deletions and amplifications of chromosomes occur due to genomic instabilities. Specific cellular malfunctions associated with cancer are the loss of the tumor suppressor p53 gene (Sherr, 2004) thereby preventing apoptosis, the gain of function or amplification of the cell growth regulators such as c-KIT (Muller-Tidow et al., 2004), and the activation of proto-oncogenes such as c-MYC (Nilsson and Cleveland, 2003). During these changes, the RNA levels of other genes are also affected, thereby disturbing the normal cellular equilibrium within the background of the other process and stochastic events. The cell as a result obtains another state of equilibrium in which it becomes "immortal". Specifically, it becomes independent of its environment for survival and proliferation. However, this is not enough for the development of a solid tumor. Many people carry *in situ* tumors, that are very small and do not develop into disease (Folkman and Kalluri, 2004). In order to proliferate, tumors need to recruit their own blood supply through angiogenesis (Hanahan and Folkman, 1996).

A major problem when studying a few genes in a pathway is that these genes are also influenced by others omitted in the study. This results in sometimes inconsistent findings where the same perturbation can lead to two different outcomes in different cells: i.e. over-expression of Myc can lead to cell proliferation or cell death (Nilsson and Cleveland, 2003). The knowledge of pathways is too incomplete to omit the other genes.

To be able to comprehend cancer, a global approach at the genome expression level must be taken. Microarray technology is particularly well suited for this holistic approach as it analyzes the expression of thousands of genes at the same time.

Genetic changes associated with cancer can involve chromosomal deletion, amplification, mutation, or deregulation of expression. Cancers can rise from a multitude of cellular events, and have very heterogeneous genetic changes. Cancer cells that seem histologically identical may respond differently to therapy, and may evolve very differently from indolent tumors to invasive metastatic tumors. Many of these changes can be observed at the mRNA level. Current cancer classification techniques and treatment decisions rely on subjective judgments of tumor histology by pathologists. Multiple DNA microarray studies have been proven useful for classification (Golub et al., 1999) and prognosis (Shipp et al., 2002; van 't Veer et al., 2003; van 't Veer et al., 2002). Global analysis of gene expression will allow classification of morphologically similar human cancer and will help tailor treatment maximizing the therapeutic effect and minimizing the toxicity (van 't Veer and De Jong, 2002). Another way microarray data can help treatment of cancer is through the screening of new drugs. It should be possible to observe the effect of a drug on gene expression profiles and create a model to predict the expected therapeutic response (Hughes et al., 2000).

However, the holistic approach of using microarray technology has brought up some issues. In addition to the significant cost of performing microarray experiments, there is a signal to noise issue as most genes are not involved in the process/condition studied. Also problematic, on the hardware/software level, are the data storage and the handling of large datasets. In order to be able to save the information in a format that can

be mined, databases and standard annotations were created. Also data analysis packages and methods were developed to handle the amount of data collected. Another limitation of microarray technology is the purity of the sample. Tumors are a mixture of heterogeneous cell types with malignant cells at different stages of differentiation, normal epithelial cells, blood vessels and cells involved in the inflammatory process. This mix can mask the differences from one cancer to another. Laser-capture-microdissection resolves this problem but creates the need to amplify the RNA sample to be able to detect the different mRNAs, with the risk of introducing bias in the measurements.

Once produced in a cell, RNA is stabilized, spliced, polyadenylated, exported towards the cytoplasm, translated into proteins and ultimately degraded (Lodish et al., 1995a). The regulation of mRNA decay rate is important for determining transcript abundance. The decay rate of individual mRNAs varies greatly form a short half life of a few minutes to half lives spanning several life cycles (Herrick et al., 1990; Wilusz et al., 2001). This difference may be important if samples are not processed in a consistent manner. Another unresolved issue in microarray data results is the effect of cross hybridization. This phenomenon occurs when there is non specific binding or binding of related sequence to the wrong microarray probes. This affects greatly the estimated mRNA levels.

1.2 Tools Available for Cancer Study at the DNA/RNA Level

Many tools were available to the study of cancer before the advent of microarrays. Most of our knowledge about cancer today still stems from experiments done with traditional techniques. The most utilized techniques for discovering mutations at the genome level are Fluorescence In Situ Hybridization (FISH) and Comparative Genomic Hybridization (CGH). Fluorescence In Situ Hybridization analysis uses either whole chromosome probes or specific probes to label chromosomes and see if there is any obvious deletion or amplification of the genetic material. Using this method, gain or loss of chromosomes can be shown (Bernell et al., 1998). Comparative Genomic Hybridization also helps detect regions of gain or loss of DNA at the genomic scale. The DNA of a cancer cell and a normal cell in metaphase, are labeled with a different fluorochrome and hybridized to each other. Differences in the ratio of intensity of the two fluorochromes along the chromosomes, helps detect the regions where amplification or deletion occurred (Kallioniemi et al., 1992).

Another commonly used technique for looking at DNA mutations is Restriction Fragment Length Polymorphism (RFLP). Initially DNA is digested into fragments using a cocktail of restriction enzymes and then run on a gel. By studying multiple samples on a gel, one can find particular fragments (bands) that are specific to either normal or cancer samples (Dracopoli et al., 1985). Once such markers are found, the gel "bands" can be cut out and sequenced to find out the identity of the genes involved. A derivative from this technique the amplified restriction fragment polymorphism (AFLP) has an extra step of amplification after the restriction of the DNA, helping with the issue of small sample amounts. A variant, the cDNA-AFLP first transcribes the mRNA into cDNA and

then the standard AFLP protocol is applied (Bachem et al., 1996). This last technique allows the identification of multiple differentially regulated transcripts at a time.

Another method designed to locate a particular sequence of DNA within a complex mixture is the Southern blot, named after Edward M. Southern. The method is as follows: first DNA fragments are separated by electrophoresis on an agarose gel, then the fragments are transferred (blotted) onto a membrane, the membrane is then soaked in a solution with a labeled probe of sequence complementary to DNA that needs to be located, and then after washing and imaging one can see if the particular sequence sought was in the sample (Southern, 1975).

The major techniques to study mRNA expression levels are: Northern blotting, *in situ* hybridization and quantitative PCR. The Northern blot uses the same basic concept as Southern blotting on RNA. This method locates specific sequences that the experimenter has a probe for. Even though one can look at a few samples at a time, only two genes are usually probed: the gene sequence of interest and one for a control sequence (Actin, GAPDH). *In situ* hybridization helps locate the intracellular localization of a specific mRMA or protein (Korabiowska et al., 2004). For this, a radioactively labeled probe is hybridized to a fixed cell, and the result is developed on x-ray sensitive film. Quantitative polymerase chain reaction (PCR), after reverse transcription of mRNA into cDNA, can help find if a particular mRNA is present in the sample (Kondo et al., 1992; Myers and Gelfand, 1991). An improvement of this technique: the quantitative real-time PCR, is the most precise technique so far for the quantization of mRNA (Bernard and Wittwer, 2002; Mocellin et al., 2003).

With the sequencing of the human genome, more genes and predicted gene sequences are now available. This changes the approach of molecular biology from the study of a single gene to the study of a system of thousands of genes. The Serial Analysis of Gene Expression (SAGE) amplify tag sequences from known and unknown genes, and sequences them (Velculescu et al., 1995). The advantages of this technique are many, the instrumentation cost is small if the facility already has a sequencer and it isolates and quantitates known and unknown transcripts. The disadvantages are also numerous as some of the tags are not specific to one gene and might differ in length posing problems when analyzing the sequenced concatenated tags.

Another very powerful method to select differentially regulated mRNA and isolate rare mRNA species is the subtractive hybridization technique (Lee et al., 1991). The mRNA from the investigated cells is reversed transcribed into cDNA. Common mRNA sequences between the studied cells and the chosen standard are subtracted. The standard cells have their mRNA reverse transcribed and biotinylated. cDNA that hybridize to the mRNA of the studied cells and are separated from the rest of the mRNAs by binding to streptavidin. The resulting subset of subtracted mRNA is then cloned into a library for further screens and investigation. This powerful method is very lengthy; two to three months of work is required in order to obtain a subtracted library.

DNA microarrays are rapidly becoming a fundamental tool in genomic research. Publications based on microarray findings increase each year (Ochs and Godwin, 2003). This technique is fast; less than a week is needed from isolation of the RNA to the acquisition of the results, and multiple samples can be processed at the same time. The description of the technology is going to be explained in detail in the next section.

1.3 Microarrays: The Fundamentals

Deoxyribonucleic acid (DNA) is the cellular molecule that carries the information required to build a cell or an organism. This information is transcribed into ribonucleic acid (RNA), which carries the instructions from the DNA for the correct amino acid sequence to synthesize proteins, the main effectors of cellular functions. Proteins are involved in cell processes from DNA replication to the regulation of cell structure and function. Proteins can be regulated by a tightly controlled production, activation/inactivation and degradation in the cell. Ubiquitinization and degradation of mRNAs can also lead to a regulation of the amount of proteins in a cell. The cell has a tight quantitative and qualitative control on transcription of genes (DNA → mRNA). These messenger RNA levels in a cell can give an indication of which proteins in the cells are being produced at the time of sampling and can also reflect changes in the genetic code due to disease (Lodish et al., 1995b). Cancer cells can have deletion or amplification of DNA resulting in change of gene expression. Gene expression can also be modulated through methylation or mutation of the promoter sequence, deregulation of transcription factors (Jones, 1996; Lodish et al., 1995b).

The DNA microarrays technology estimates the amount of mRNA for thousands of genes at the same time in contrast to the "one gene in one experiment" approach. An array consists of an orderly arrangement of gene specific poly-nucleotides strand that match known and unknown nucleotide strands of DNA (i.e. human, rat...). There are design considerations in choosing the oligonucleotides for hybridization on microarrays. First, the oligonucleotides must have a common thermodynamic profile of melting temperature, in order to properly hybridize or bind to the target while reducing non-specific binding.

As the melting temperature increases with the length of the oligonucleotides and GC content, the hybridization temperature will have to be adjusted depending on these factors. A second consideration is to avoid probe that might form secondary structures preventing the hybridization of the target. And the last consideration is to find probe that are not homologues to other sequences. As the probe length shortens, there is an increase in the probability to find an homology to another sequence. A 75-80% homology or more than 15 continuous complementary bases in non targeted sequence will produce a false signal as it hybridizes partially to the probe. One way to remedy this is to use BLAST[1] software to select probes in parts of the genes that do not have homology to other genes. These considerations are the same as the ones for designing PCR primers. Temperature and salt concentration can be optimized for a given set of primers for PCR, but with microarrays these conditions need to be the same and close to optimal for thousands of probes at the same time.

Two microarrays technologies are dominating the field: the Affymetrix GeneChip (Lockhart et al., 1995) and the in-house printed arrays (Schena et al., 1995).

1.3.1 Affymetrix GeneChip Technology

Affymetrix GeneChips are manufactured by synthesizing oligonucleotides of defined sequences on a glass substrate chip. The synthesis technology is a light directed solid phase DNA synthesis. This technique allows the synthesis of tens of thousands of unique oligonucleotides per square centimeter on a glass surface (Fodor et al., 1991). A probe cell is a specific area on the chip containing millions of copies of a specific 25-nucleotide

[1] http://www.ncbi.nlm.nih.gov/BLAST/ March 2004

long oligonucleotide. Arrays are manufactured in a series of cycles represented in Figure 1.2. The glass substrate is first coated with linkers containing photolabile protective groups. Then, a mask is applied exposing selected portions of the probe array to ultraviolet light which removes the photolabile protecting groups. This enables selective nucleoside phosphoramidite addition only at the exposed sites. This is repeated to allow specific sets of oligonucleotide probes to be synthesized for each probe cell.

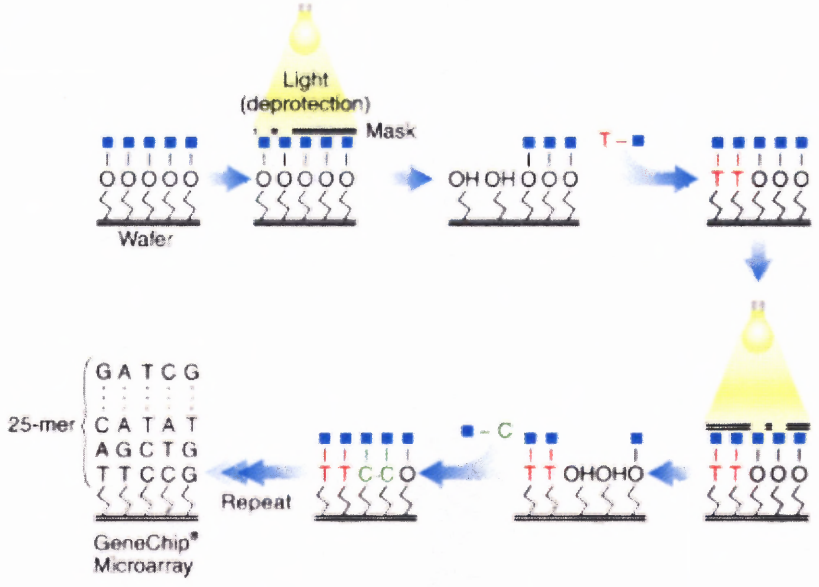


Figure 1.2 Construction of the Affymetrix GeneChips using photolithography technique. Five-inch square quartz wafers are coated with silane and linker molecules to provide a uniform density and the ability to be activated by light. A mask is applied on the wafer and the UV light activates the surface uncovered. The added nucleotide (A, T, G, C) will couple with the activated linkers, and the wafer is washed to remove uncoupled nucleotide. The process is repeated until the probes reach the length of 25 nucleotides. (Figure courtesy of Affymetrix, Inc.)

Labeling of the RNA sample is achieved by first reverse transcription of the mRNA into cDNA and then into cRNA through in-vitro transcription with biotin labeled nucleotides. The cRNA is then fragmented and hybridized overnight onto the probe array. The hybridized probe array is then stained with streptavidin phycoerythrin

conjugate and scanned with a laser emitting a wavelength of 488 nm and the amount of light emitted at 570 nm is recorded as proportional to the bound target (Figure 1.3). The signal can be amplified with a second round of staining with goat anti-streptavidin antibody and biotinylated Goat IgG.

These arrays are able to measure the absolute expression of genes in cells or tissues with the precision of one transcript in 1,000,000 (Lipshutz et al., 1999). The disadvantages of the Affymetrix GeneChip technology are a high cost per array and the inability to compare two samples on the same chip rending comparison susceptible to normalization processes.

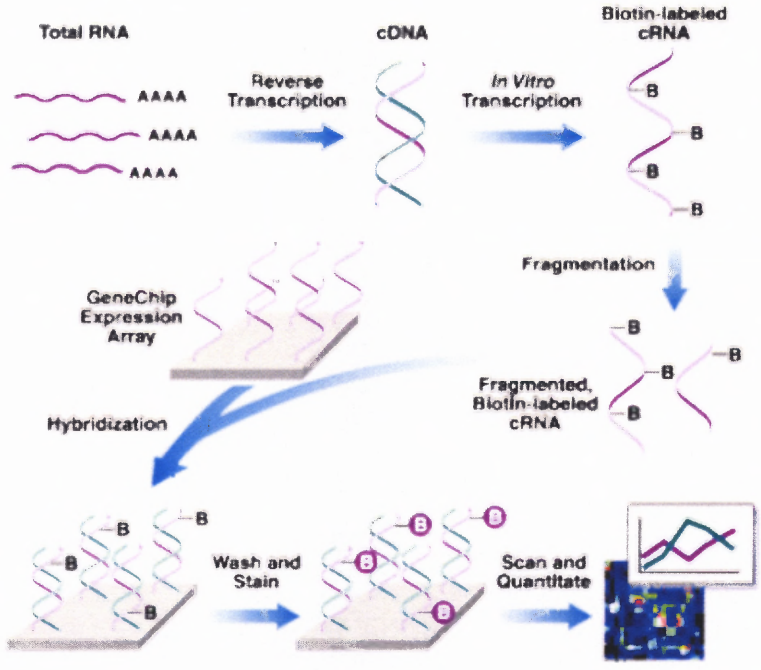


Figure 1.3 Labeling of the RNA sample and hybridization of the Affymetrix GeneChips. The RNA sample is transcribed into double stranded cDNA using a T7-oligo dT primer. After clean up and *in vitro* transcription is performed using a T7 RNA polymerase and biotin labeled nucleotides, the cRNA is then fragmented and hybridized overnight onto the probe array. The hybridized probe array is then stained with streptavidin phycoerythrin conjugate. A second round of staining with goat anti-streptavidin antibody and biotinylated Goat IgG is used to amplify the signal received when scanning.
(Figure courtesy of Affymetrix, Inc.)

1.3.2 Spotted Microarrays

The other widely used technology is the spotted microarray. Synthetic oligonucleotides or PCR products derived from cDNA are spotted onto a microscopic glass slide coated with poly-L-lysine, Amino Silane, Super Amine, or Super Aldehyde. Those arrays are usually printed using a high density arrayer (Cheung et al., 1999). Spatial resolution of the arrayer determines the density of the array. Most instruments use pins or needles to transfer/spot oligonucleotides or cDNA probes from 96 or 384 microtiter plates. The pin shape, diameter and the spotting solution determine the size of the spot on the array. The final array is no bigger than 3 square inch and the spot diameter is of the order of 0.1mm to 70 microns. The labeling of total mRNA is typically done by synthesizing single stranded DNA with a reverse transcriptase incorporating nucleotides with a fluorescent molecule attached to them. Cyanine 3 and Cyanine 5 dUTP are the most commonly incorporated fluorescent molecules. Two samples, each labeled with a different dye are hybridized to the array (usually overnight). Labeled gene products bind to their complementary sequences in the spots. The array is then washed and the dyes enable the amount of sample bound to a spot to be measured by the level of fluorescence emitted when excited by laser. From the fluorescence intensities in the channels scanned for each spot, the relative expression levels of the genes in both samples can be estimated. Spotted microarrays are usually analyzed with the ratio of relative expression between the samples hybridized on the slide.

1.4 Summary

Microarray technology is a powerful tool to analyze the changes at the mRNA levels in cancer. This technology allows the analysis of thousands of genes in a single experiment in less than a week. It is thus considered to be the best technology to find gene expression markers in cancer. However the development of microarrays has preceded statistical methods for analysis of the results. Major problems due to signal estimation, normalization, and multi-testing are still the topic of many publications each year.

CHAPTER 2

CURRENT STATUS OF MICROARRAY DATA ANALYSIS

The use of microarrays has become common in biological sciences. The query "DNA microarray and cancer" yielded 226 papers in the literature search engine Pubmed[2]. In this Chapter, the different steps and methods in the analysis of microarray data are going to be presented. First, a signal intensity has to be extracted from the image for each gene/feature and the transcript expression level has to be estimated. Then, the results from each array have to be normalized in order to be able to compare the results from one array to another. Finally, analysis methods can be applied to find the genes that are differentially regulated between conditions.

2.1 Image Analysis and Signal Extraction

For both types of microarray, a specialized scanner is used to asses the hybridization signal. Analysis of gene expression data is usually done in a two-step process: image processing and data analysis. There has been a plethora of techniques and software developed for detecting and delineating the target spots on the array, eliminating background from the intensity, correcting for bias due to different dye affinities, and scaling or normalization in order to compare multiple arrays (Bolstad et al., 2003; Li and Wong, 2001a; Quackenbush, 2002; Yang et al., 2001). These techniques are usually specific to the array platform although some correspondence can be drawn between them.

[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi March 23rd 2004

For spotted arrays, cDNA or oligonucleotides corresponding to the complementary sequence of the cDNA obtained by *in vitro* reverse transcription are deposited on slides as spots. A lot of emphasis has been placed on the feature/spot recognition with different ways to identify the arrayed spots on the chip and measure the background. In most cases the spots are circled and the pixel intensities for each wavelength within the circle are averaged, or a median is taken, and intensities outside the circle are also averaged for local background estimation i.e. GenePix® Pro 4.0 from Axon Instruments Inc. A newer version of this software allows the delineation of spots that are not shaped into circles (GenePix® Pro 5.0). Another software package, ImaGene® from BioDiscovery Inc. allows the detection of pixel artifacts inside the spots and removes them from the intensity estimation of the spot. There are multiple software packages and methods for extracting the signal intensity. Future software must place more emphasis on the quality assessment of spots and arrays. In the image processing step, the quality of the printing, the alignment of the grid, the annotation, and scanner calibration are essential for the reproducibility of the experiment (Johnson and Lin, 2003).

Affymetrix GeneChips are hybridized with only one sample per chip. The dynamic intensity of the probes range from 0 to 65,000. As the oligonucleotide sequence is short: 25 nucleotides, a gene is represented by a probe-set composed of a series of preferably non-overlapping nucleotides. A probe set, usually contains between 16 to 20 probe pairs. A probe pair is composed of 2 probe cells. One of the probe-cells contains DNA sequences complementary to specific mRNA sequences and is called the perfect match probe or PM probe. The other probe cell is composed of an oligomer identical to

the PM probe except for a single base difference in the central position (mismatch or MM probe). The mismatch probe is used to estimate the background hybridization. Many different software packages have been developed for the estimation of gene intensity. The Microarray Analysis Suite 4.0 (MAS 4), produced by Affymetrix Inc, is empirical in nature (Affymetrix, 2000). Subsequently a model-based estimate for the expression levels of genes (MBEI) was designed by Li and Wong, in 2001, as an alternative to MAS 4. In 2002 Affymetrix Inc. released a new version of its Microarray Analysis Suite, MAS 5, based on statistical algorithms (Affymetrix, 2001; Affymetrix, 2002a; Hubbell et al., 2002). Another method called robust multi-array average (RMA) was developed in Dr. Speed's lab (Irizarry et al., 2003), and is basically a PM only model with a local background subtraction. These commonly used techniques are reviewed in the following sections.

2.1.1 MAS 4

The MAS4 software first extracts the intensity for all probe cells by aligning a grid to match the delimitation of the features and then averages the pixel intensity within a delineated probe cell (Affymetrix, 2000). With MAS4 the Average difference (Avg Diff) is calculated for each probe set as the average of the differences between every PM probe cell and its control MM probe cell. This is considered to be directly related to the level of expression of the transcript. However, this method can give negative values for a probe set if there is more hybridization signal in the mismatched sequences than in the perfectly matched ones. The software also estimates the presence or absence of each transcript by giving an absolute call using a decision matrix. For example, in order for a gene to be

called present, the positive fraction, number of time PM>MM in the probe set divided by the number of probe pairs used, has to be above 0.43, the ratio of positive to negative probe pairs has to be above 4.0 and the Log Avg Ratio of the PM to MM intensities for each probe pair has to be above 1.3. All these default threshold values have been established through empirical testing (Affymetrix, 2000).

2.1.2 MAS 5

The updated version of MAS 4.0 released in 2002 and designated Microarray Analysis Suite 5.0 (MAS 5) uses a statistical inference of the signal of a transcript and its presence or absence instead of an empirical one (Hubbell et al., 2002). Their detection call is based on the Wilcoxon signed rank test to determine whether the discriminant score (R=(PM-MM)/(PM+MM)) of the probe set is greater than 0.015. For the expression value, if few MM probe intensities exceed the PM probe intensity in the probe set, the algorithm creates an adjusted MM value based on the average difference intensity between $\log_2$PM and $\log_2$MM. When most of the MM probe intensities exceed the PM value, the adjusted MM value is set to some fraction of PM and the transcript is considered absent. This eliminates the negative signal. Then MAS 5.0 uses the One-Step Tukey's Biweight Estimate as a quantitative measure of the mean mRNA abundance for each gene. This weighted estimate is less sensitive to outliers. The log of the background-adjusted PM-MM difference for each probe pair is weighted by its distance from the median value for the entire probe set. The weighted log(PM-MM) values are then used to calculate the overall mean of the probe set. The signal output in MAS 5 is this adjusted mean converted back into linear scale.

2.1.3 MBEI

Li and Wong designed, in 2001, a model-based estimate for the expression levels of genes as an alternative to MAS 4. They observed using an Analysis of Variance analysis (ANOVA) that the residual mean squares at the probe level is greater than the ones between replicate arrays. Their model accounts for this probe variability. For I arrays hybridized with different samples, each gene has J probe pairs (10-16 depending on the arrays) associated with it, giving Ix(2xJ) probe intensities (PM + MM) intensity values. They assumed that the intensity for a particular probe increases linearly with the concentration of the transcript in the sample, that this rate of increase can be different from one probe to another, and that the rate of increase is higher in the PM probes instead of the MM probes (Li and Wong, 2001a). With θi representing the concentration of a transcript in the ith sample they formulated the following model for the PM and MM intensities:

$$MM_{ij} = b_j + \theta_i \alpha_j + \varepsilon$$

$$PM_{ij} = \mathrm{b}_j + \theta_i \alpha_j + \theta_i \varphi_j + \varepsilon$$

b_j represents the background intensity for jth probe, α_j the rate of increase of the jth MM probe due to non specific hybridization, φ_i is the specific rate of increase for the jth PM probe due to the target concentration in the sample and ε represents random error following a normal distribution $N(0,\sigma^2)$. By subtracting the MM intensities from the PM we obtain:

$$y_{ij} = PM_{ij} - MM_{ij} = \theta_i \varphi_j + \varepsilon_{ij}$$

The average difference is a linear function of the target concentration plus or minus a random error. Given replicate data, we can estimate the mean and standard

errors of the θ and φ of each probe set. Given the standard error, we can identify outlier arrays and probes and exclude them from the analysis for the estimation of the target concentration. They also noticed that cross-hydridization is more likely to occur on the MM probes than with the PM probes. The target not only hybridizes to the PM probe but also cross-hybridizes with the MM probe. They added an option to the dChip 1.1 program for a PM-only model (Li and Wong, 2001a), which is more robust to cross-hybridization than the PM-MM difference model as it does not subtract the MM probe to the PM signal intensities.

$$y_{ij} = PM_{ij} = \theta_i \varphi_j + \varepsilon_{ij}$$

The PM-only model estimates the expression value θ and the probe sensitivity index (PSI) φ using an iterative algorithm, and eliminates the outliers using the same method as the PM-MM model. The Li and Wong model was found to give a more accurate estimation of the signal intensity than the Average Differences given by the Affymetrix MAS4 software (Lemon et al., 2002). However, Rajagoplan demonstrated that the dChip algorithm gave inferior results than MAS5 for estimating true changes and presented a higher percentage of false positive on the Latin square dataset (Hubbell et al., 2002; Rajagopalan, 2003). He also found the dChip algorithm PM only to be very weak at estimating lower target concentrations because it does not subtract the background hybridization.

2.1.4 RMA

The Robust Multi-array Average (RMA), developed in Dr. Speed's lab (Irizarry et al., 2003), corrects the arrays by first using a signal dependant background estimate, then normalizes the arrays using quantile normalization and then fits each normalized, $\log_2$ based background corrected probe set to a linear model:

$$Y_{ijn}=\mu_{in}+\alpha_{jn}+\varepsilon_{ijn}$$

Where α_{jn} is the probe affinity effect for the probe set n, μ_i represents the log scale expression level for array i, and ε_{ij} is a random error (with a mean of 0). They use a median polish method to estimate the parameters in a robust manner. This method is available in the Bioconductor package, an open source and open development software project for the analysis and comprehension of genomic data (Ihaka and Gentleman, 1996). The authors showed that RMA reduced the variance and bias compared to MAS5 and Li and Wong PM-MM model (Irizarry et al., 2003). It also showed greater sensitivity in detection of differential expression.

2.2 Normalization Strategies

After image processing, it is necessary to normalize the data obtained, in order to compare the different sample intensities. In the GeneChip system, samples are hybridized on different chips and normalization is applied to enable comparison of one chip to another. For spotted arrays, two samples are labeled with different dyes resulting in a need to do a within-chip normalization to eliminate bias due to uneven starting RNA level, difference in labeling efficiency due to a dye bias or RNA quality, and systematic bias when measuring the expression levels.

GeneChips are frequently normalized using a global scaling procedure. Every gene-expression value is multiplied by a constant so that the median intensity on the array equals a set threshold. Affymetrix MAS 4 trims the top and bottom 2% of intensities in order to take a more robust mean measure. MAS 5 uses a global normalization to a set baseline intensity and perturbs it by multiplying or dividing by 1.1. This helps overcome the fact that a normalization factor does not necessarily fit all the probe sets. A related approach in spotted arrays is to fit the two color intensities on a line with a slope of 1 and an intercept of 0. These approaches rely on the fact that a cell will express about the same amount of RNA at any given time or conditions, and that most of the genes expressed are not changing. This assumption might not be true in a lot of cases, for example, if you compare different cell types that expresses less RNA than the other or if the treatment has an effect on the production or degradation of RNA. Another problem with the global scaling is that it does not account for the non-linearity of the data. The highly expressed RNA levels might need a different scaling factor than the lower expressed ones. This can be seen when plotting the intensities of the two samples (two Genechips, or one two color spotted) against each other and the result has a non linear trend, and in the case of most spotted array a bias towards the Cy3 in the low expression range (green tail). In order to better see the bias, one can use an M-A plot (Dudoit et al., 2000) where M represents the log ratio of the two dyes; M=log (Cy5/Cy3), and A is the average of the logarithm of the intensities; A=[log (Cy5)+log (Cy3)]/2. This plot is a rescaled 45-degree rotation compared to Cy3 plotted against Cy5. In the case of no bias, one should see an even scatter of the spots around zero on the x-axis. If the bias is

intensity dependant, the cloud of points will not be evenly distributed and will have a curved shape.

To correct for the non linearity of global scaling approach, non linear scaling methods (Li and Hung Wong, 2001; Li and Wong, 2001a); (Stuart et al., 2001) can be performed with the assumption that an "invariant set" of genes exists, and that their rank in expression within a chip does not change significantly. They use this invariant set for the non-linear regression at either the feature (Li and Hung Wong, 2001; Li and Wong, 2001a) or the probe set level (Stuart et al., 2001). This method corresponds roughly to the Lowess normalization used on spotted microarrays; the genes found in the invariant set follow the Lowess normalization curve (Yang et al., 2002b). The global normalization Lowess (Dudoit et al., 2000; Quackenbush, 2002; Yang et al., 2002a; Yang and Speed, 2002) performs a locally weighted linear regression as a function of the log(Cy3* Cy5) and correct the log ratios to an average of zero in that local region. This method performs a linear approximation of a non-linear regression. The user can define the fraction of data used for smoothing each point, usually 40%, rendering the linear approximation robust to a small percentage of differentially expressed genes. Another variant of this method uses a rank invariant method to select the genes that are not differentially expressed and applied the Lowess normalization to the chip using this set of genes (Tseng et al., 2001). This algorithm can also be applied locally to each sub arrays produced by individual pins, correcting for spatial variation such as gradient and differences between the pins (Yang et al., 2002b).

Another approach is to scale to a set of controls that can be either genes whose expression is not affected by the treatment or spike-in controls. The problem with these

methods is that: 1) there is no gene whose RNA levels never fluctuate, 2) spiking RNA or DNA is subject to pipeting errors and the end result may not correlate with the quality of the RNA that influenced the hybridization.

The normalization procedure has a greater effect on the subsequent detection of differentially expressed genes than was anticipated in the past (Hoffmann et al., 2002). In their study Hoffman et al., 2002, found that the effect of the normalization was greater than the effect of the analysis algorithm for finding differentially expressed genes.

2.3 Analysis of Affymetrix GeneChips

Analysis methods for microarray data can be classified in two categories: supervised and unsupervised learning. Unsupervised methods can be applied without the prior knowledge of the sample states: i.e. cancer or normal. They identify patterns in the gene expression profile and group similar samples together. For example, hierarchical clustering (Eisen and Brown, 1999) is an agglomerative process in which expression profiles are joined in function of their distance into a tree. The distances between members of a tree, clusters or single expression profiles, can be computed in many ways (Quackenbush, 2001). Because of its simplicity, and graphical display, this method has become one of the most widely used analysis of gene-expression data. Other preeminent unsupervised classification techniques are k-means clustering (Tavazoie et al., 1999), self-organizing maps (SOM) (Kohonen, 1990) and the principle component analysis (PCA) (Raychaudhuri et al., 2000). Principle Component Analysis is a mathematical technique to reduce the dimensionality of the data without a significant loss of information. It reduces the redundant information. This method is sometimes used

before the k-means clustering and self-organizing maps as it provides an estimate of the number of clusters. K-means clustering and self-organizing maps use iterative approaches to find clusters, increasing the distance between clusters while decreasing the distances of the gene expression profiles within the cluster. With unsupervised clustering, sometimes the classification of the samples does not correspond to expected classification. This can happen, for example, when cancer and normal samples are taken from the same patient. Although there is a difference in the state of the samples, cancer samples can cluster with their respective patient normal sample, because most of the background genes stay unaffected by the change of state (normal to tumor).

The supervised methods for analysis present a powerful alternative when sample state/treatment is known. Those methods are very useful for finding gene markers. Each gene is evaluated for its ability to separate the sample states. Genes that are consistently differentially regulated between the conditions are very good candidate markers. Methods to find differentially regulated genes are presented below.

2.3.1 Standard Supervised Analysis Methods

The most straightforward approach is to define a fold change threshold or cut-off that needs to be exceeded in order to consider a gene differentially expressed. The problem with this method is that it is arbitrary and does not account for noise in the data. This method, therefore, generally yields a greater number of false positives.

Affymetrix MAS 5 has its own algorithm for finding differentially regulated genes. MAS5 performs three global normalizations: to a set baseline intensity and the same baseline perturbed by multiplying or dividing by an error factor. Then a Wilcoxon

signed rank tests is performed on the probe level data, on for the three different normalizations (Liu et al., 2002). Finally, a difference call (increase or decrease) is issued only if the three p-values from the Wilcoxon signed rank tests fall under a user defined threshold.

Another method frequently used (Roberts et al., 2000) is the Student's t-test for comparison between two experimental conditions. If the samples come from two populations that are normal and with equal variance, the t value for testing the difference between to two population means is computed as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_2}}$$

With $\overline{X}_1 - \overline{X}_2$ being the differences of the means and $S_{\overline{X}_1 - \overline{X}_2}$ is the standard error (Zar, 1999). The hypothesis of no differences in the means is rejected when |t| is above the thresholds of the t distribution for a specific Type I error rate α (false positive rate) and degree of freedom. This test assumes that the samples came at random from a normal population with equal variance. This assumption is often wrong, it has been observed that some genes behave normally while others do not (Hoffmann et al., 2002). The test is however quite robust to the normality assumption especially when the number of samples of the conditions are equal and numerous. Due to the expensive nature of the technology, many experiments contain few samples and an unequal number of them per condition. Adapting a parametric test to analysis of microarray data violates the theory on which the test was based.

Non-parametric tests do not require the estimation of the population variance or mean and do not state any hypothesis about the nature of their distribution. The most

commonly used two sample rank test is Mann-Whitney test. The actual measurements are not used but their rank is. The data is ranked from the highest value to the lowest and average ranks R_1 and R_2 are taken for each population of n_1 and n_2 samples. The Mann-Whitney statistic is calculated as follow:

$$U = n_2 n_1 + \frac{n_1(n_1+1)}{2} - R_1$$

U is then compared to the table of the critical values of the Mann-Whitney U distribution at the Type I error rate α and number of samples in each population. Non-parametric statistics can be used for the analysis of microarray data but require a greater number of replicates to obtain the same probability of Type II error (false negative) than a parametric test for the same significant p-values (Zar, 1999). These tests used alone produce a high rate of false positives due to multiple comparisons performed on the data. With a type I error α equal to 0.01, one hundred genes out of 10,000 can identified as significant just by chance. One hundred false positive genes in the result set are too many and require tedious follow up and confirmatory experiments to be able to find the true positives.

2.3.2 Correction for Multiple Testing

Multiple comparison procedures are designed to control the familywise error rate, i.e. the combination of the type I errors from the multiple tests. The most commonly used technique, the Bonferroni correction consists in dividing α by the number of comparisons and use this value as the cut-off for the significant p-values (Bonferroni, 1936). However, this simple correction method has attracted many criticisms (Perneger, 1998). The Bonferroni correction is too conservative, each comparison is held to an

unreasonable standard. In the case of a microarray with 12,000 genes, the p-value cut off would be 0.001 or 0.05 divided by 12,000, $8*10^{-8}$ and $4*10^{-6}$ respectively. This might be an unreasonable high standard and it definitely increases the Type II error were legitimate results are failed to be detected. Other methods have tried to control the false discovery rate in a little less stringent manner. The Simes and the Hochberg procedures have been shown to increase the power of the false discovery rate control compared to the Bonferroni correction (Benjamini and Hochberg, 1995). For both procedures the p-values are ordered in ascending order and the rejection of the null hypothesis depends in part on the rank of the p-value. For the Simes procedure a gene is considered significantly up or down regulated if its p-value is below or equal to its p-value rank divided by the total number of comparisons and multiplied by the type I error α.

A gene ranked i is significant if:

$$p_{(i)} \leq \frac{i}{m}\alpha$$

With $p_{(i)}$ being the p-value of the ordered i^{th} gene, m the total number of tests performed on the data and α the type I error rate. Similarly for the Hochberg procedure a gene ranked i is significant if:

$$p_{(i)} \leq \frac{i}{m+1-i}\alpha$$

Those methods might still be too stringent when the common number of tests on micoarray data is between 12,000 and 20,000.

The significance analysis of microarrays method (SAM) (Tusher et al., 2001) has been described for finding differentially expressed genes. In this method the means and

standard deviations are used to compute the relative difference in gene expression. This relative difference d(i) (for a gene i) is very similar to the t value of a student t-test.

$$d(i) = \frac{m_A(i) - m_B(i)}{s(i) + s_0}$$

Where $m_A(i)$ and $m_B(i)$ are the mean expressions of the gene i for the state or treatment A and B respectively, s(i) is the standard deviation and s_o is a small constant. The small constant s_o helps ensure that the distribution of d(i) is independent of expression, preventing genes with a very small standard deviation and expression from being considered differentially expressed. The standard deviation is computed as follow:

$$s(i) = \sqrt{c\{\sum_a [x_a(i) - m_A(i)]^2 + \sum_b [x_b(i) - m_B(i)]^2\}}$$

Where Σ_a and Σ_b are the sum over the samples in A and B, c=(1/n+1/m)/(n+m-2) with n and m being the number of samples in A and B. The novelty of this method is that they use balanced permutations of the dataset to estimate the relative difference of those genes in a random case. Genes are ranked in descending order of their d(i) value (i.e.: d(1) is the largest value in the dataset) and the same is done for the permutation $d_p(i)$ and averaged. The average of the $d_p(i)$ is the expected relative difference. By plotting the relative difference (from the treatment effect) against the expected relative difference we can observe that for most genes those values are equal. When the difference between those two measures exceeds a certain threshold Δ, the genes are considered differentially expressed. One can substitute one of the balanced comparisons to the treatment comparison, compare it to the expected relative difference and count how many genes cross the threshold. By repeating this procedure for all balanced comparisons, one can obtain an average number of false positives. This method provides the flexibility of an

asymmetric cut-off for significant genes due to the comparison with random effects. This method can also be extended to the analysis of 3 or more states/treatments using a Fisher's linear discriminant to obtain the parameter d(i).

2.4 Analysis of Spotted Arrays (two or more samples per arrays)

When analyzing two or more samples per array, one is really looking at the hybridization of those samples for the different spots (i.e: for example normal *vs.* malignant). The amount of oligonucleotides spotted on the chips is in excess of the samples to be probed. Ratios are inherently independent from the quantities of cDNA or oligo spotted which can be quite variable from the first slide printed to the last one, or from one slide lot to another. The question, regardless of the experiment, is still finding genes that are differentially expressed. Reducing the dataset to the genes that are differentially expressed helps eliminate noise that could perturb further analysis such as clustering. Fixed fold change cut-offs of 2 or more can be used but as was observed many times, the variation in fold-change increases as the signal decreases. A more powerful way to identify the genes that are differentially expressed within a slide is to calculate a z score with the mean and standard deviation of the log(ratio). This method defines a global standard for the chip for fold change difference and confidence. It has the same inconvenience as the fold change method for it does not take into account the greater variability of the measurements at lower expression. An alternative approach is to calculate the z-score in a local manner. Using a sliding window, the mean and standard deviation of the genes surrounding each data point T_i can be calculated, and the local z-score can be computed:

$$Z_i^{local} = \frac{\log_2(T_i)}{\sigma_{\log_2(T_i)}^{local}}$$

Genes with an absolute local z-score above 1.96 are considered significantly differentially expressed (95% confidence level) (Quackenbush, 2002).

2.5 Synopsis on Current Status of Microarray Data Analysis

There are many combinations of the methods described above for the analysis of microarray data from signal estimation, normalization to analysis of significant genes. The major problem with analysis is that standard statistical tests are ill adapted to the data. Standard analysis produces too many false positives and when correction for multiple testing is applied, due to the high number of tests, the corrections tend to be too conservative. A different approach to analysis is presented in Chapters 3, 4 and 5. In Chapter 3, a nonparametric method is used to separate cancer samples according to their selectively expressed genes. This method disregards the normalization techniques that influence the results by changing/correcting signal intensities. In Chapter 4, a noise boundary model is described, to eliminate spurious fold change, reducing the number of false positives in further analysis. The issue of using data from different sources to identify differentially specific cancer biomarkers will be examined in Chapter 5.

CHAPTER 3

ANALYSIS OF THE ROLE OF SELECTIVE EXPRESSION IN CLASSIFICATION

3.1 Introduction

Application of microarray technology depends on accurate comparison of the level of gene expression across a set of microarrays. In general, this process requires normalization or scaling of measurements made on each microarray, so that differential changes in gene expression can be observed. In the most frequently used normalization method, the data are scaled by the ratio of the mean of the frequency distribution of gene expression levels on one microarray to that on a control microarray. This method works well as long as the gene expression distributions are linear and have a similar form on all microarrays. Furthermore, the dependence on the accuracy of this normalization process is exacerbated by the magnitude of the corrections, which often increase or decrease measured expression level values by a multiplication factor larger than 6.

This dissertation presents non-parametric methods for microarray data analysis. This Chapter describes the role of selectively expressed genes for classification as an alternative method for comparison of gene expression across a set of microarrays that is not dependent on the standard normalization process. Instead of using differential gene expression, the usefulness of selective expression was evaluated. Selectively expressed genes are generally present in microarrays from one group, and are less likely to be present in the group that is being distinguished from. This method was tested on data from the Golub et al (1999) study, whose aim was to classify and predict classes of leukemia by determining which genes have expression levels that are most correlated

with different disease states (Golub et al., 1999). Their analysis method was based on the differential expression levels of all the genes on the microarrays. The expression data was derived from bone marrow and peripheral blood samples from patients suffering with acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). These childhood acute leukemia are blood related cancers arising in the bone marrow. These cancers are progressing rapidly resulting in the accumulation of immature cells in the bone marrow and blood. The bone marrow is saturated and can no longer produce enough normal red blood cells, white blood cells and platelets, leading to anemia, the lack of immune defense and easy bruising and bleeding. The acute lymphoblastic leukemia are comprised of two subtypes: T-cell and B-cell leukemia. The original dataset was split into two sets, a training set and an independent evaluation set. The training set was used to develop a statistical separation of the two groups, and was composed only of bone marrow samples. The independent set was then used to evaluate the separation, but some of the independent set samples were derived from peripheral blood and therefore added heterogeneity to the analysis. All of the samples were hybridized on GeneChips, produced by Affymetrix Inc.

The Affymetrix software MAS 4 employs a decision matrix based on metrics comparing the intensity of the PM to the MM to determine if a transcript is present (P), marginal (M) or absent (A). The Golub et al. (1999) study did not take into consideration the present and absent calls. Using a neighborhood analysis, they were able to classify 36 of the 38 samples in the training set and 29 of the 34 independent samples. They also used self-organizing maps (Kohonen, 1990) for automatic discovery of the classes. In

contrast, in this study, present and absent calls were used as indicator of selective expression, and evaluated for separation of the two types of leukemia.

3.2 Methodology Development

In the initial analysis, the level of gene expression is not used. The samples in the training set are placed into two groups (27 ALL, 11 AML) and the absent and present calls of each of the genes in each group are converted into binary numbers, with one corresponding to present and zero corresponding to absent. Marginal genes and the present genes with an average difference below zero were also considered to be absent. The selectivity of each of the genes is computed as the absolute value of the difference between the real-valued averages of the (expressed/not expressed) binary values for each of the two disease groups. With this metric, a gene that is present in all of the samples in one group and absent in all of the samples in the other group is maximally selective, with a selectivity of one. In contrast a gene that is absent in both groups, present in both groups, or present in the same percentage in each of the two groups is not selective at all, and has a selectivity of zero. A gene would be considered significant if it is twice as likely to be expressed in one group as in the other. In other words, a gene is considered significant if it has a selectivity, or absolute difference, that is larger than 0.5. The group average represents the ideal behavior of a sample from this group and is called an exemplar. The 7129 genes are sorted into a ranked list from the most selective to the least selective, and the most selective genes are used to construct an exemplar vector for each of the two groups. The dimensionality of the exemplar vector is set by the number of genes that have been included in its definition.

The likelihood of selective genes occurring by chance was evaluated by randomly shuffling the data. The samples were randomly assigned to the two groups, which had the same size as the true AML/ALL groups and the same process was used to construct a ranked list of selective genes. The average selectivity for each rank on this list was computed for 45 shuffles of the data, and this is compared with the selectivity of the genes in the true AML/ALL groups.

The usefulness of selective genes as a method for classifying samples was evaluated by computing the Euclidian distance from each microarray to each of the two exemplar vectors. These distances were computed for all of the microarrays in both the training and independent sets. Members of a group should be closer (smaller distance) to the exemplar of that group than to the exemplar of the other group. The exemplars computed from the training set are used to classify the data from the independent set.

A simple form of normalization can be applied to selectively expressed genes. This normalization seeks to make corrections to errors in the assignment of absent and present calls, which may have occurred due to variation in the processing of the samples. In replicate experiments using the same sample on microarrays, some genes with low expression levels were found that have absent calls on GeneChips with low hybridization or high background (data not shown). Samples that are more successfully processed might be more likely to have an increased number of present calls. This hypothesis is supported by the correlation shown in Figure 3.1. This Figure shows the relationship between the number of genes with present calls on a microarray and the scaling factor computed by the Affymetrix software. There is a linear relationship between the inverse of the scaling factor and the number of genes expressed in the chips (R^2= 0.51).

Therefore chips with fewer genes expressed require a larger scaling factor, and chips with more genes expressed have a smaller scaling factor. In general, higher scaling factors might be needed if the processing of the microarray was less successful. This implies that microarrays with lower average expression levels might have genes that should be present, but their expression level is not discernable from the background. This is a detection threshold problem; the sensitivity of the microarray changes depending on the background noise, quality of the RNA and labeling reaction. Refinements of this technique might lead to greater detection threshold/intensity cut-off at the frontier between an absent and a present gene. For normalization, it is impossible to scale up an absent call into a present call. Instead, microarrays with more genes expressed can be scaled down by setting the genes with lowest expression level to absent. Provided that the genes with the lowest expression are the genes masked in the background noise, this solves the detection threshold problem. This method bases its trust more in the comparisons of expression levels within an array than across arrays.

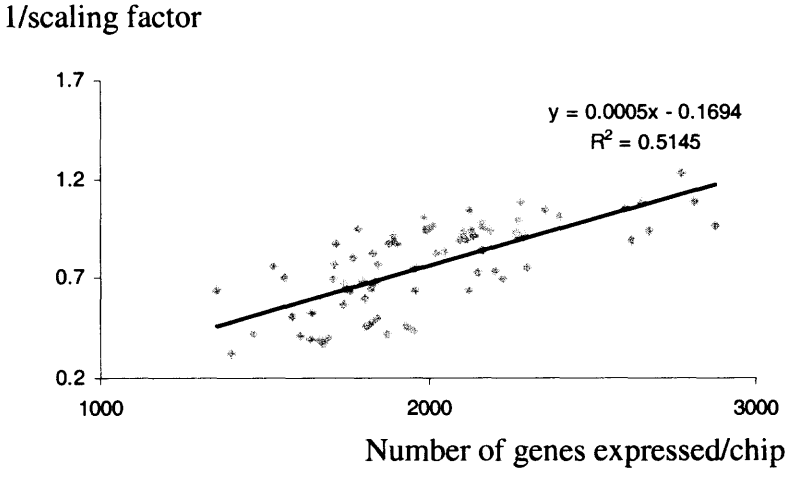


Figure 3.1 Linear regression of the number of expressed genes (on the X-axis) to the inverse of the scaling factor (on the Y-axis).

More precisely, the expression levels of all of the genes in each microarray are ranked from the highest expression level to the lowest expression level. The gene with the lowest ranked expression level on the microarray with the largest number of genes is set to absent and the selectivity of the genes is computed. This process can be continued until the selectivity starts to decrease or until all of the chips have at least as many genes present as the average number of genes present at the start of the process. The selectivity of a gene increases if it is usually absent in one group, present in the other group, and one or more of the present calls in the usually absent group are switched to absent because of low expression values on microarrays with a larger than average number of genes expressed. For example, a gene that is absent in all of the ALL microarrays except for one, and is present in the all of the AML microarrays, has a selectivity that is less than one. If the one ALL present call has a low expression level on a microarray with more than the average number of present genes, then this normalization will set the ALL present call to absent, and the selectivity will increase to be exactly one. This method also reduces the number of genes poorly detected by the technique from the set of genes used to classify the samples. Some genes might be better detected as present in some samples and might gain a spurious selectivity index by chance. By changing the call of those low expressed genes to absent on arrays with a high number of genes found present, the selectivity of those low expressed genes will be decreased compared to the selectivity of genes with higher expression values.

In the initial analysis of the training data, the number of present calls ranges from 1352 to 2877. The normalization is applied in the first instance keeping the highest ranked 2500 expression levels from each of the microarrays. Any of the microarrays that

has more than 2500 genes present has the genes with the lowest expression levels now forced to be absent. The genes with selectivity greater than 0.5 prior to this process are called the selective genes. The difference between the number of selective genes that increase their selectivity and the number of selective genes that decrease their selectivity is computed. The same process is repeated in which the highest ranked 2250, 2000, 1750, 1500, and 1352 genes are kept as present, and find the number of genes that optimizes the selectivity of the selective genes.

3.3 Results

It can be observed in Figure 3.2 that a substantial fraction of the genes are absent and this suggests that selective expression might usefully distinguish between groups of samples. If the majority of the genes are always present then the data must be normalized to look for differential expression. Also, there is a large variation in the number of genes present from one slide to another, from 1352 genes found present in sample 27 to 2890 in sample 5, with an average of approximately 2000 out of over 7000 genes.

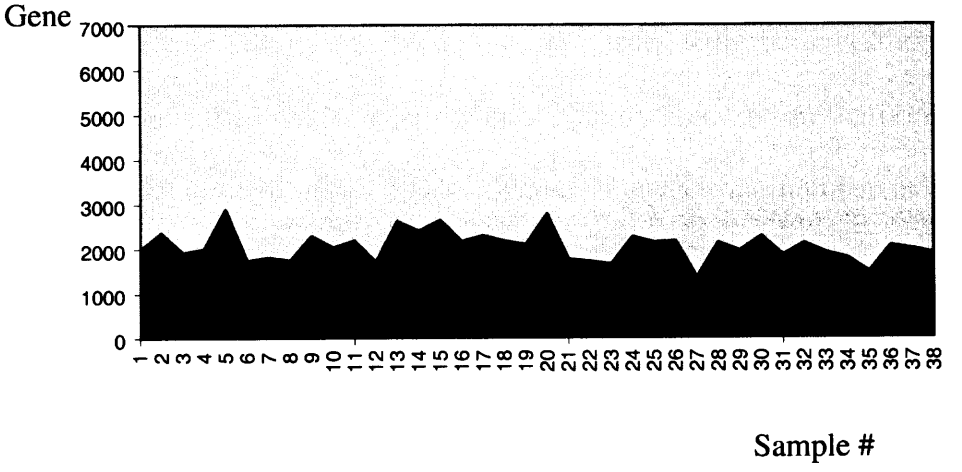


Figure 3.2 Number of genes Present ■ or Absent □ across the 38 training samples (on the X-axis).

When the selectivity metric described above, is applied to the two leukemia groups, 121 genes are found to be selective. This included several genes that were also found to be important in the distinction between the two types of leukemia by Golub et al. (1999); MB-1 required for the B-cell antigen receptor (Payelle-Brogard et al., 1999), zyxin, HOXA9 involved in myeloid differentiation (Casas et al., 2003; Zeisig et al., 2004), and cyclin D3 involved in cell cycle and proliferation (Table 3.1). In the list of the 24 most selective genes we see that some genes, such as the myosin light chain, are more expressed in the ALL leukemia. Others, such as zyxin, are expressed more often in the AML group. As described in the methods section, this selectivity was compared to 45 random shuffling of the samples into the two groups. Those selective genes had significantly higher difference in expression than the randomly shuffled sets (last column on the table). Figure 3.3 shows the selectivity of the genes, after normalization. The x-axis is the position of the gene in this ranked list, and the y-axis is the selectivity.

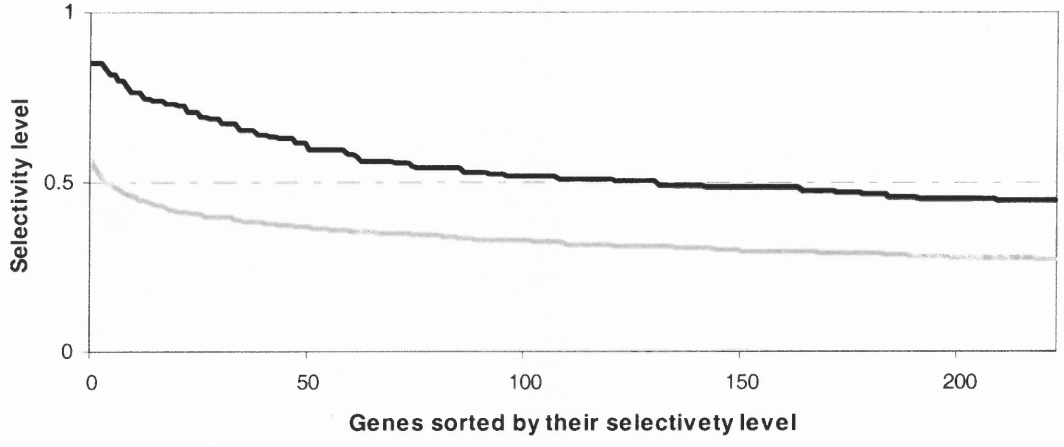


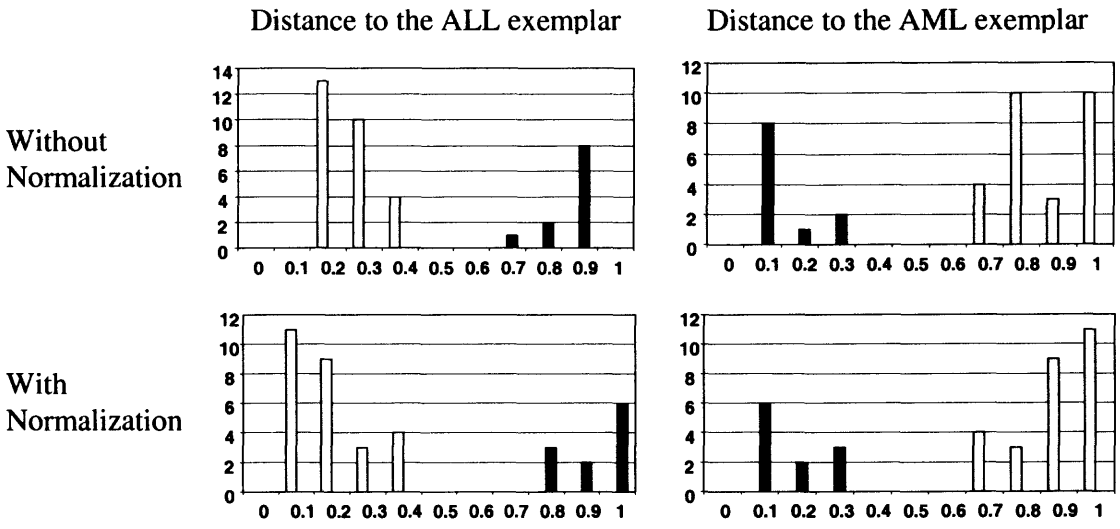
Figure 3.3 The selectivity level is the absolute difference between the ALL exemplar and the AML exemplar ■ . The selectivity level between the shuffled samples □ is represented as a comparison.

Table 3.1 Most Selectively Genes Expressed in the AML/ALL Training Set
The average expression for the ALL/AML groups and the absolute difference, are displayed as well as the absolute difference of the shuffled set with the standard deviation and the number of standard deviation away from the AML-ALL real grouping.

		ALL-AML grouped selectively			Shuffle comparison		
Rank	Gene Description	Average ALL	Average AML	Absolute Difference	Mean Absolute Difference	Standard Deviation (SD)	No. SD
1	DAGK1 Diacylglycerol kinase	0.85	0	0.85	0.60	0.07	3.51
2	CYSTATIN A	0.15	1	0.85	0.56	0.05	5.36
3	KIAA0035 gene	0.85	0	0.85	0.54	0.05	6.25
4	MYL1 Myosin light chain	0.93	0.09	0.84	0.52	0.05	6.61
5	LEPR Leptin receptor	0.19	1	0.81	0.51	0.04	6.98
6	NUCLEOLYSIN TIA-1	0.81	0	0.81	0.50	0.04	7.83
7	Terminal transferase mRNA	0.89	0.09	0.8	0.49	0.04	7.59
8	Inducible protein mRNA	0.89	0.09	0.8	0.48	0.04	8.42
9	CHRNA7 Cholinergic receptor	0.04	0.82	0.78	0.48	0.04	8.33
10	PPBP	0.15	0.91	0.76	0.47	0.04	8.04
11	CST3 Cystatin C	0.15	0.91	0.76	0.47	0.04	8.01
12	RB1	0.85	0.09	0.76	0.46	0.03	8.51
13	Zyxin	0.07	0.82	0.74	0.46	0.03	8.19
14	BTF5	0.93	0.18	0.74	0.46	0.03	8.33
15	GLUTATHIONE S-TRANSFERASE	0.26	1	0.74	0.45	0.03	8.58
16	MB-1	0.74	0	0.74	0.45	0.03	8.77
17	KIAA0230	0.74	0	0.74	0.44	0.03	9.06
18	adipsin	0	0.73	0.73	0.44	0.03	8.81
19	PTX3	0	0.73	0.73	0.44	0.03	8.97
20	IGIF	0	0.73	0.73	0.43	0.03	9.26
21	RABAPTIN-5	0.81	0.09	0.72	0.43	0.03	9.25
22	UBIQUITIN-LIKE PROTEIN GDX	0.81	0.09	0.72	0.43	0.03	9.15
23	HOXA9	0.11	0.82	0.71	0.43	0.03	8.61
24	Cyclin D3	0.89	0.18	0.71	0.42	0.03	8.58

In order to classify the samples in one group or the other, the Euclidian distance of their binary call (1 for present and 0 for absent) to the exemplar was computed. This distance among the 10, 20, 30 most selective genes, was then averaged and the closest exemplar was determined. With the ten most selective genes, the 2 groups in the training set were completely separated (Figure 3.4).

After applying the normalization process described above, the selectivity was maximized by retaining, as present, the 2000 genes on each microarray that have the highest expression level on that microarray. The lower two panels of Figure 3.4 show that this normalization improves the separation of the two groups. Using the ten most selective genes, the AML samples are closer to the AML exemplar and that the ALL samples are further from the AML exemplar.



X-axis: Distance to the exemplar, ALL on the left and AML on the right.
Y-axis: Frequency distribution

Figure 3.4 Frequency distribution of the distance of the independent samples to the AML and ALL exemplar before and after normalization using the 10 most selective genes on the training set. The histogram in each panel has one entry for each sample, and a different color for the two types of leukemia ALL samples □ and AML samples ■ .

Figure 3.5 shows the effect of changing the dimensionality of the exemplar vector on cluster separation. The prior analysis was based on the ten most selective genes, or equivalently on ten dimensional exemplar vectors. The x-axis of this Figure is the dimensionality of the exemplar vector. Again the included genes have been ranked, and the genes are added to the exemplar in rank order from the most selective to the least selective. The y-axis is the distance of each of the samples from the ALL exemplar. Each sample is plotted as a distinct line. The top panel shows the distances prior to applying the normalization process and the lower panel shows the distances after applying the normalization process. Over most of the range shown in the Figure the normalization process has increased the separation of the two groups. As more genes are included in the exemplar vector the graphs gradually move closer together.

The same analysis was applied to the independent set, but the distances were computed to the exemplar that was computed from the training data. As shown in Figure 3.6, this technique was successful at separating the two groups. Only one sample, number 66, was incorrectly classified. The top panel shows the distances of the two groups from the ALL exemplar prior to normalization, and the lower panel shows the distances after the same normalization process had been applied to the independent evaluation set. In this normalization no additional optimization was performed, and as before the 2000 genes with highest expression level from each microarray were kept as present. Note that again the normalization increases the separation of the two groups.

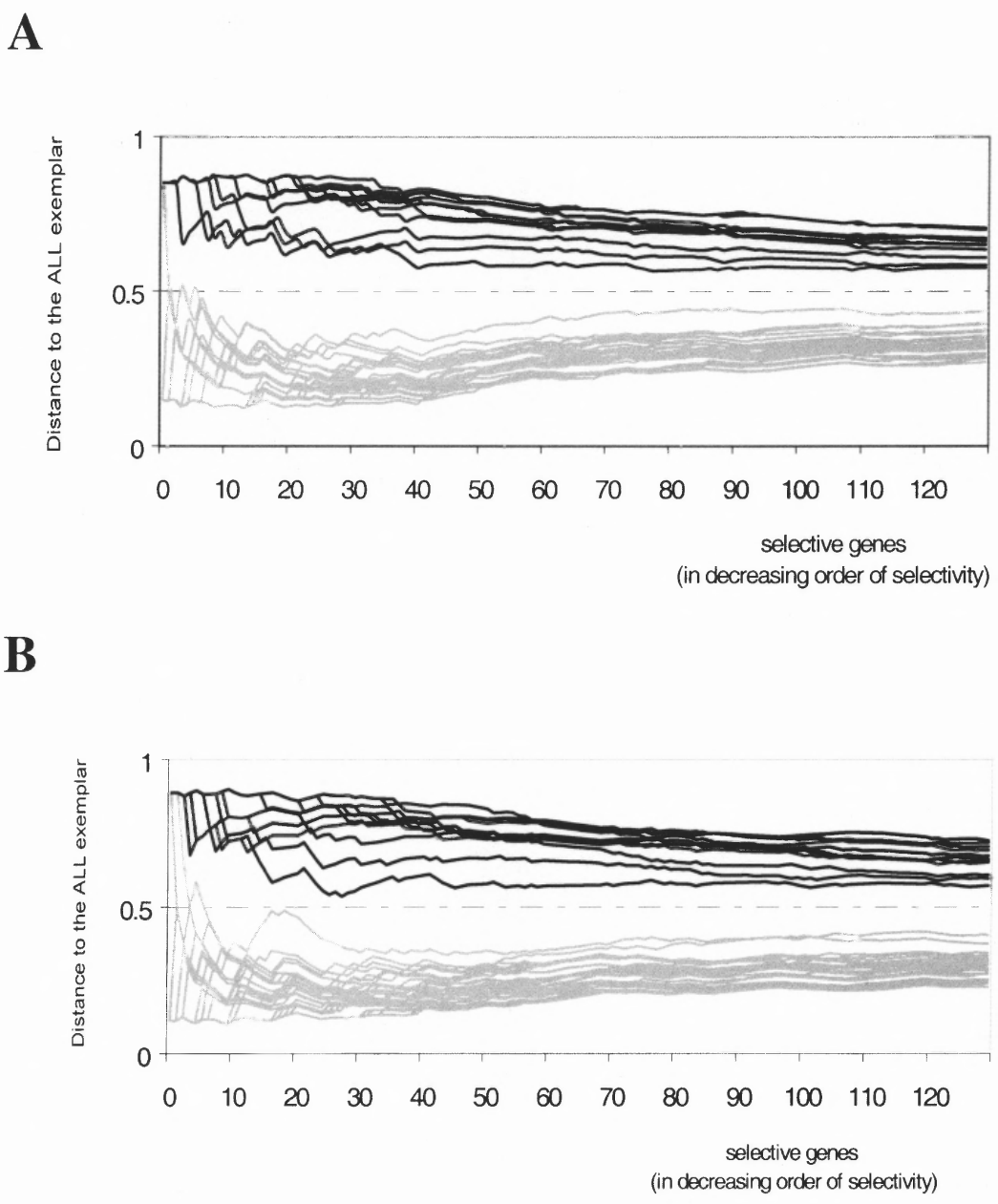


Figure 3.5 ALL samples ▢ and AML samples ■ , (A) Distance of the training set samples to the ALL exemplar before normalization. (B) Distance of the training set samples to the ALL exemplar after normalization.

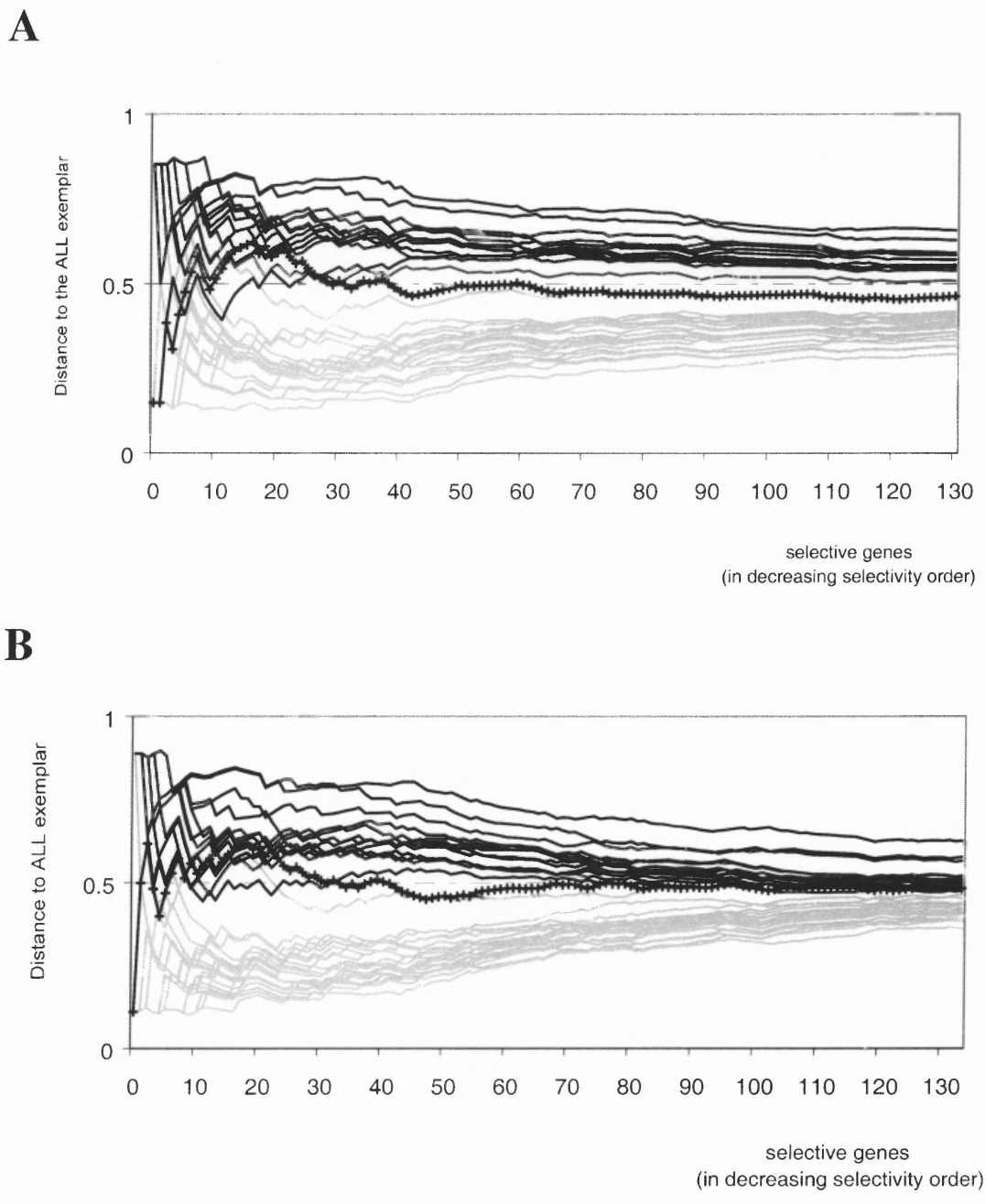


Figure 3.6 ALL samples ☐ and AML samples ■, +++ misclassified sample number 66, (A) Distance of the independent set samples to the ALL exemplar before normalization. (B) Distance of the independent set samples to the ALL exemplar after normalization.

The same method was used to examine other possible classifications of the data. There was information in the data about the sub grouping of the ALL samples into T-cell and B-cell type leukemia. Just as was found in the ALL-AML classification, genes were identified in the T-cell to B-cell groups that are more selective than expected by chance, as evaluated by shuffling the samples. The resulting exemplar vectors were able to completely separate the T-cell and B-cell groups in the training set without error. In the independent set, one out of the 20 ALL independent samples, number 67, which was a peripheral blood sample, was misclassified. The method was also applied to the classification of success and failure of treatment and to female and male subjects. In both of these two cases there was no significant difference between the selectivity of genes found in the real groups from that found in the randomized groups. The absence of a correlation in the treatment response case may be due in part to the small number of samples (11 in the training set).

3.4 Conclusion

Selective gene expression is sufficient for separating the AML vs. ALL samples, and T-cell vs. B-cell samples. Selective expression does not have the same dependency as differential expression on normalization of microarray data, and for this reason it may be more robust regardless of variations in processing. Some of the most selective genes were differentially expressed in the Golub et al. (1999) study (e.g. MB-1, zyxin, HOXA9, cyclin D3). Other genes like Catalase (X04085) which were differentially expressed in both ALL and AML, but not selectively expressed were not selected by this method.

However, TCL1 was not found to be differentially expressed in the prior study, and was found to be selectively expressed with this method.

The shuffling method provided a useful verification of the use of selective genes to classify the data into groups. The AML and ALL groups, and the T-cell the B-cell groups were far from the randomly constructed datasets, while the success–failure classification and female–male classification was close to chance.

This approach is orthogonal to prior studies and complementary, since it focuses on genes that are selectively expressed rather than on differential expression levels. This simple approach is very robust, powerful and easy to use in diagnostic microarray development as it contains strongly expressed genes. A synthesis of the qualitative and quantitative methods should improve the classification performance, and add understanding to the mechanism of action of the genes involved in some pathology.

This work demonstrated the possibility of circumventing the normalization problems by using only selectively expressed genes. However, it then raises the issue of definition of the expression versus no expression. There is no real cut-off and finding expression depends on the background noise and sensitivity of the technique. There is also quantitative information overlooked with this method. In the next Chapter a noise boundary model is presented. This model is designed to enable the analyses of differential gene expression.

CHAPTER 4

ASSESSING THE NOISE LEVEL AND TRUST THRESHOLD FOR DIFFERENTIAL EXPRESSION ON AFFYMETRIX GENECHIPS.

4.1 Introduction

In this Chapter the characteristics of the noise of the GeneChip microarray data is evaluated. When noise is consistent and reproducible it can be filtered from the data, and some false positives can be eliminated. There are two major sources of noise in microarray data. The first source of noise is biological which comprises variations between different patients, tumor location, variation of cellular composition among tumors, heterogeneity of the genetic material within tumor due to genomic instability, and differences in sample preparation. Biological noise cannot be corrected but it can be accounted for with statistics using replicates of the treatment conditions. The second source of noise is the microarray technical noise that comprises nonspecific cross hybridization, efficiency of the labeling reaction and differences between microarrays. The noise derived from experimental technique is reproducible and the boundary of the noise can be modeled. It has been observed, that in differential comparisons of any given gene, there is a greater variance in the fold-change calculation at lower signal intensities (Mutch et al., 2002; Tu et al., 2002). When comparing replicate samples, lower expression values tend to have the greatest variance in signal intensity. This suggests that larger errors can occur when lower signals are used to compute fold-changes in differential comparisons. Fold change, computed in this way, can lead to extraneous inclusions in lists of significantly up-regulated or down regulated genes. For example, a fold change of two calculated from intensities of 25 and 50 may not be as trustworthy as

a fold-change of two determined between intensity values of 2,000 and 4,000. Thus, the purpose of error boundary modeling is to reduce the influence of less trustworthy fold-change calculations in the differential analysis of microarray data. The technique of coupling a noise boundary model to an analysis method has previously been shown to be useful for two color cDNA arrays (Baggerly et al., 2001; Hughes et al., 2000; Ideker et al., 2002; Yang et al., 2002a).

4.2 Development of a Noise Boundary Model

It would be ideal to construct the noise model to measure the technological noise with replicated data, but this type of data is not always available. One can use replicates for the same condition: i.e. different normal biopsies of tissues from different patients. In this case the noise model not only measures the technique error but also some of the biological variability. This could be useful in the sense that it accounts for tissue variability in the design of the error model. This approach is in a way similar to the work done by Mutch et al., 2002. Their analysis relied on 2 assumptions: 1) the signal is more variable closer to the low intensity detection threshold, 2) empirically, from the literature, regardless of the method of gene selection, only a few percent of genes change due to treatment (less than 5%). Based on these assumptions, they have created a limit fold change model by making a histogram of absolute fold changes as a function of minimum intensity. Absolute fold changes were binned in the size of 200. Only the top 5% probe sets of each bin were selected to be considered differentially regulated. Their data contained no biological replicates, instead they pooled samples, and they had only one technical replicate. Their method of selecting the up-regulated or down-regulated genes

consisted in taking the top 5% of the absolute fold changes in a uniform manner across the intensity range. In contrast, the error boundary model was created to eliminate untrustworthy fold changes from the analysis. This boundary was designed to be used as a preprocessing step before evaluation of the data. In Chapter 5, publicly available data (Bhattacharjee et al., 2001; Su et al., 2001; Toruner et al., 2004; Welsh et al., 2001a; Welsh et al., 2001b) is going to be used to find cancer markers. This Chapter is going to describe how the noise boundary model was designed for those data sets. First, two questions were asked: which probe-set intensity measurement method should be used, and can a noise boundary model be designed for each method? It was noticed that for each method, MAS5, dChip PM-only and RMA, for any combination of normal biopsies, there was an increase of the fold changes for a decrease in the average intensity of the probe-set. Figure 4.1 displays a scatter plot of the fold-change plotted against the average intensity for each probe set from arrays from two normal lung tissues using MAS5 for signal estimation. This plot, commonly named a volcano plot due to its shape, shows a considerable increase in signal variation at smaller signal amplitudes. The initial broadening in fold change starts at a signal intensity of 200 with 61.6% of the genes having an average signal intensity lower than this value. Significant broadening occurs for signal values with less than an average signal amplitude of 100, which corresponds to over 44.8% of the measurements of a chip scaled at 300.

To model the noise, a fold-change threshold boundary was drawn for each comparison between normal biopsies for each cancer studied. This was accomplished by binning the data into fixed width bins. Each bin was set to include 200 expression values. A percentile of the fold-change was calculated for each bin, and considered to be the

error boundary. Figure 4.2 shows the 80th percentile error boundary values for each of the bins of the replicate data displayed in Figure 4.1.

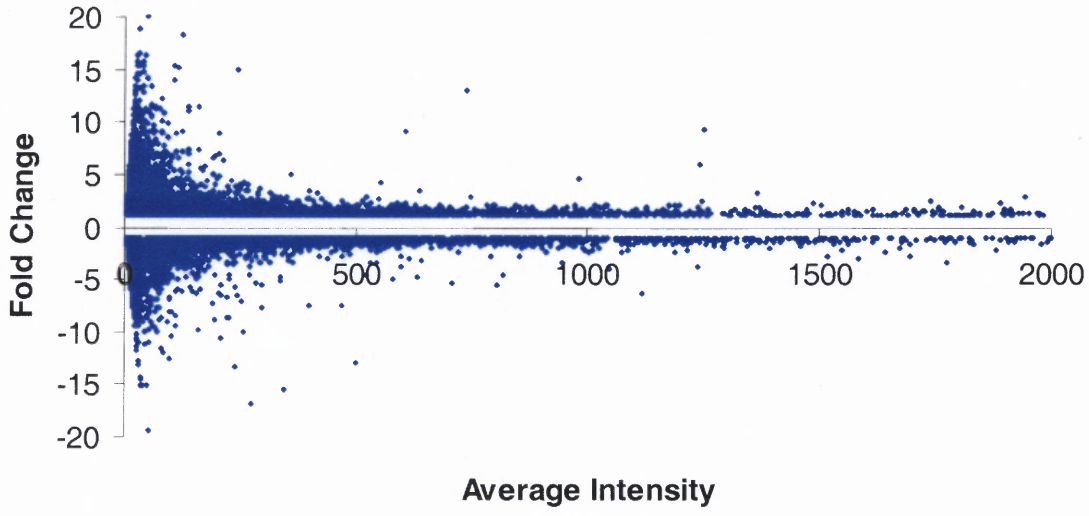


Figure 4.1 Relationship of fold-change, on the y-axis, to the average signal intensities, on the x-axis, for two normal lung samples, (MAS 5 data).

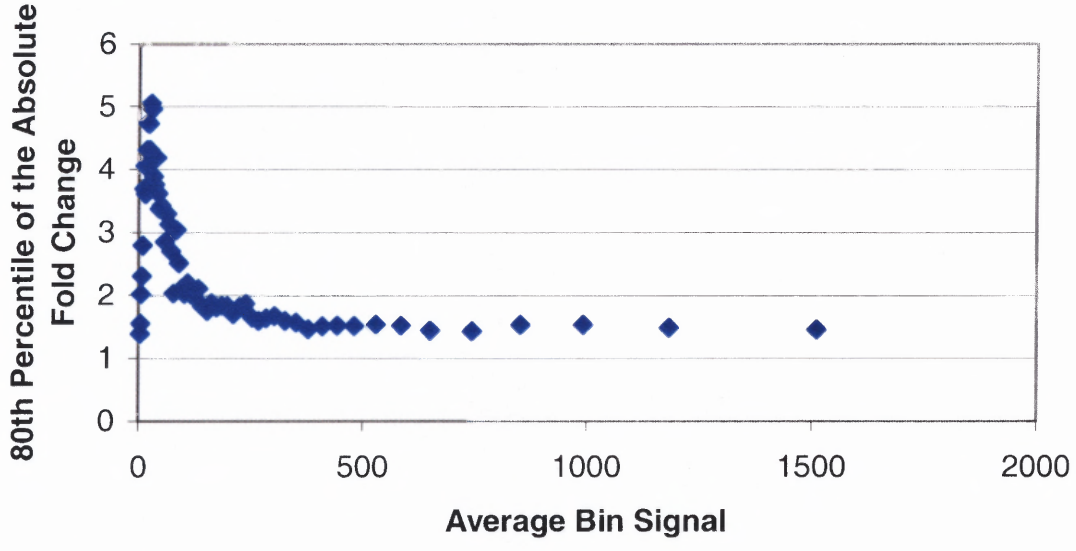


Figure 4.2 The 80th percentile of the absolute fold change (y-axis), from two lung normal samples, is plotted against the average signal (x-axis) for each bin of 200 genes.

For modeling purposes, the percentile was plotted against the inverse of the average bin revealing a linear relationship that can be characterized with a slope and intercept. This linear relationship was seen with all the probe-set signal estimation methods (MAS5, dChip PM-only and RMA). In Figure 4.3, this linear relationship can be seen for the same two lung samples as Figure 4.2. The linearity stops for an average bin intensity of 25 (1/0.04) at which point the gene expression may not be reliable because it is too close to the detection limit of the technology. Because of this, a minimum intensity cutoff is going to be set to preserve the linearity. This minimum intensity cutoff is another parameter, after the percentile, which is going to greatly influence the model. The next section will present a sensitivity analysis on these two parameters; expression cut off and percentiles.

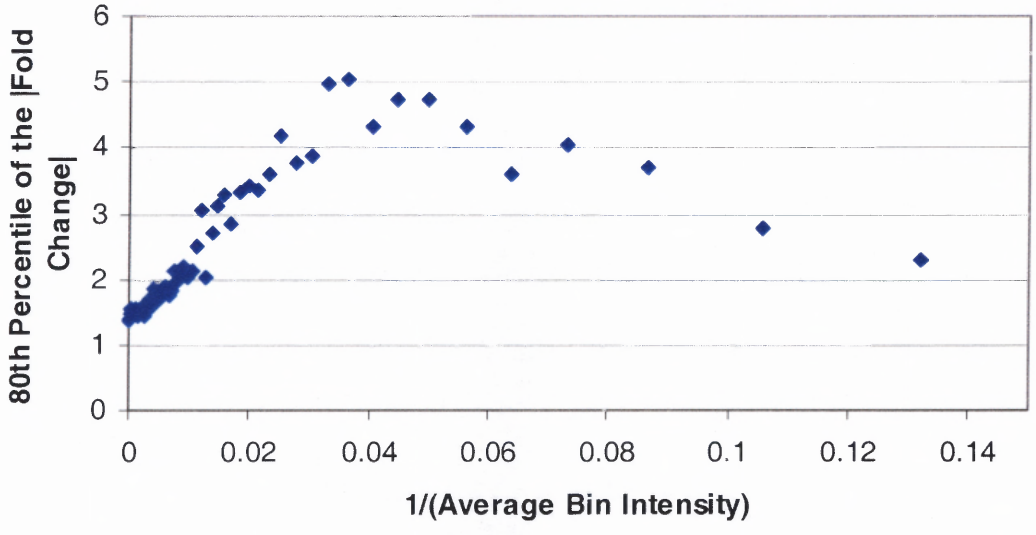


Figure 4.3 The 80th percentile of the absolute fold change (y-axis), from two lung normal samples, is plotted against the inverse of the average bin signal intensity (x-axis).

4.3 Sensitivity Analysis of the Parameters (cut off and percentile) and the Probe Set Intensity Extraction Methods

To be able to perform a sensitivity analysis on the parameters, a standard dataset was chosen. The Latin square replicate data set (Hubbell et al., 2002) was used for this purpose as it contains genes replicated for their expression levels and spiked in genes with different concentrations on different chips. Since the concentration of these genes was the only difference between the arrays, the true and false positive rates could be accurately determined. The 16 samples of the replicate set in the Latin square dataset were considered replicates for modeling the noise as the number of probe-sets spiked in represents only a small portion, less than 0.0012%, of the total number of probe-sets. For this replicate set, a sensitivity analysis was performed around two parameters: percentile taken for each bin for regression and minimum intensity cutoff. This sensitivity analysis was also performed for the different methods of probe set intensity extraction: dChip PM only, MAS 5 and RMA. The sensitivity analysis consists in observing the variation of the slope and intercept obtained from the linear regression on the boundary defined by the percentile and cut-off. The average slopes and intercepts were then graphed as a function of those two parameters.

For dChip PM only and RMA, the parameter that influences most the average slope is the minimum intensity cutoff (Figure 4.4 and 4.5). Only the highest percentiles, above 94, have an effect on the slope and decrease it drastically due to the introduction of noisy data. On the other hand, the slopes increase steadily with the minimum intensity cutoff for RMA and with a delay for dChip PM only. This can be due to non linear increase of the absolute fold change with the decrease in bin intensity. The increase has concave shape just like in Figure 4.2, the more the low end data is cut, the higher the

slope of the regression. Also as an artifact of the transformation of the x-axis, inverse of the average bin intensity, there is also a non-equidistant repartition of the bin percentile fold change. This result in bin with a low average intensity having more weight on the regression than the bins with higher intensities: i.e. more bins are regrouped in the 0 to 0.02 range than 0.02 and 0.04 range. As the average intensity gets higher, the differences of the fold change percentiles decrease and the inverse of the average intensity decreases further. Eliminating the low intensity bins has a greater effect on the slope. Also, slopes with a cut-off higher than 500 are probably not reliable as the regressions were performed on less than 20% of the original data.

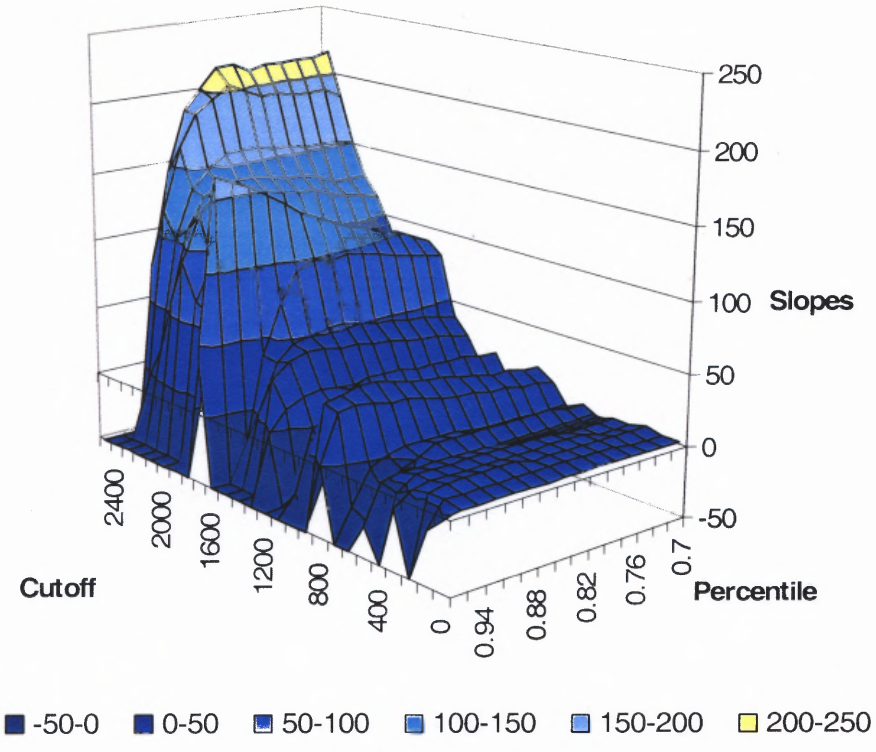


Figure 4.4 Three dimensional graph of the effect of the minimum intensity cutoff and percentile on the slope of the regressed percentile to the average intensity of the bins, with data obtained with the RMA (Ihaka and Gentleman, 1996).

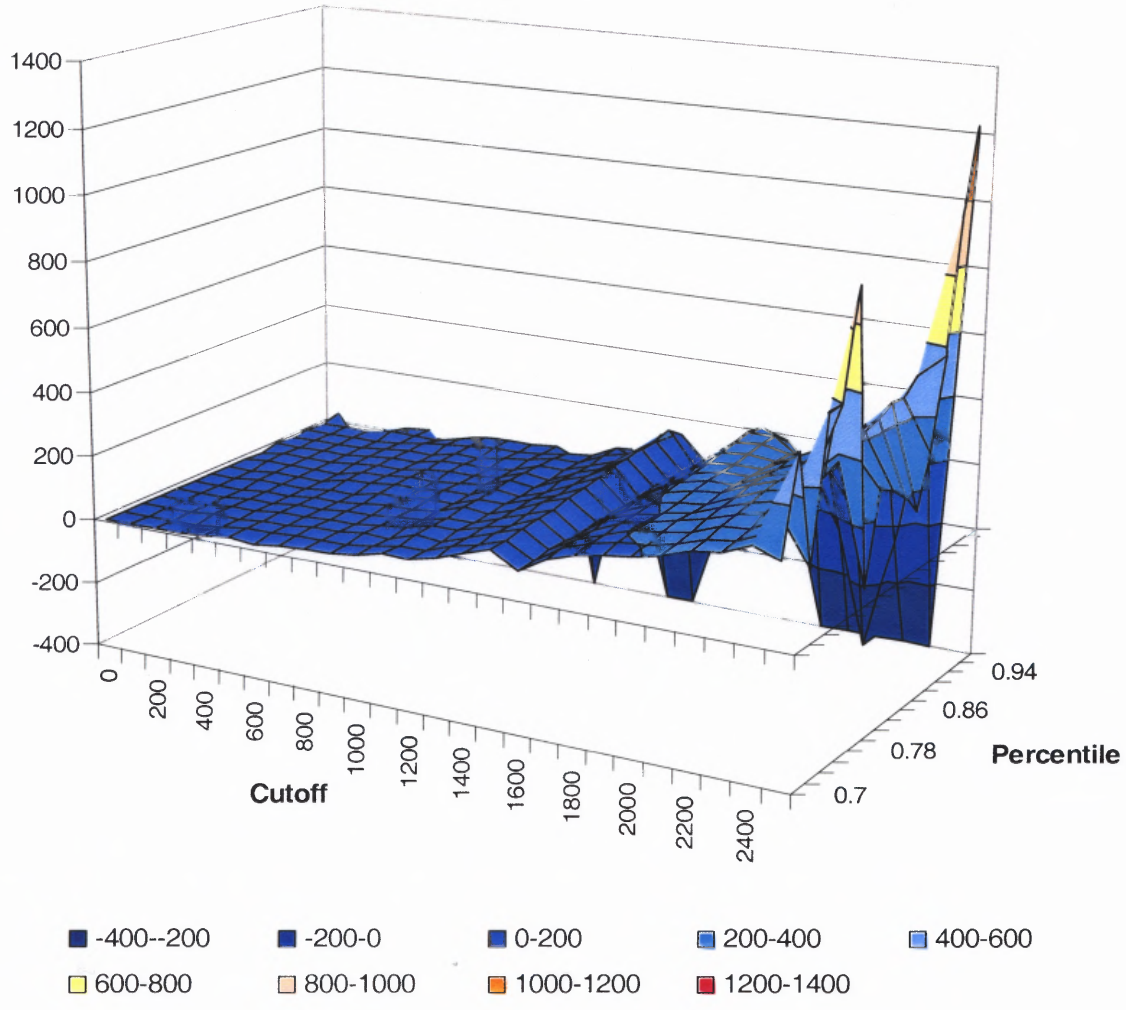


Figure 4.5 Three dimensional graph of the effect of the minimum intensity cutoff and percentile on the slope of the regressed percentile to the average intensity of the bins, with data obtained with the dChip PM only.

In Figure 4.6, MAS 5 average derived slopes are both influenced by the minimum intensity cutoff and the percentile. The slope increases in a step like manner increasing with a larger percentile and a larger minimum intensity cutoff. The magnitude of the slope is also very different between MAS 5, dChip PM only and RMA. For the 80th percentile and a minimum intensity cutoff of 100, the average slope is 157.6 for MAS 5, 7.7 for dChip PM only and 2.15 for RMA. Since the regression is performed on the inverse of the average intensity, the slope gives an indication of the noise in the low

intensity range. Because the Latin square GeneChips can be considered as replicates, it can be inferred that RMA and dChip PM only are better controlling the noise in the low intensity range compared to MAS 5. One explanation could be the fact that dChip PM only and RMA attenuates the signal compared to MAS5, reducing both noise and true fold change (unpublished data, appendix A). In all three cases there are conditions where the slope is stable for minor variations of the cutoff or percentile.

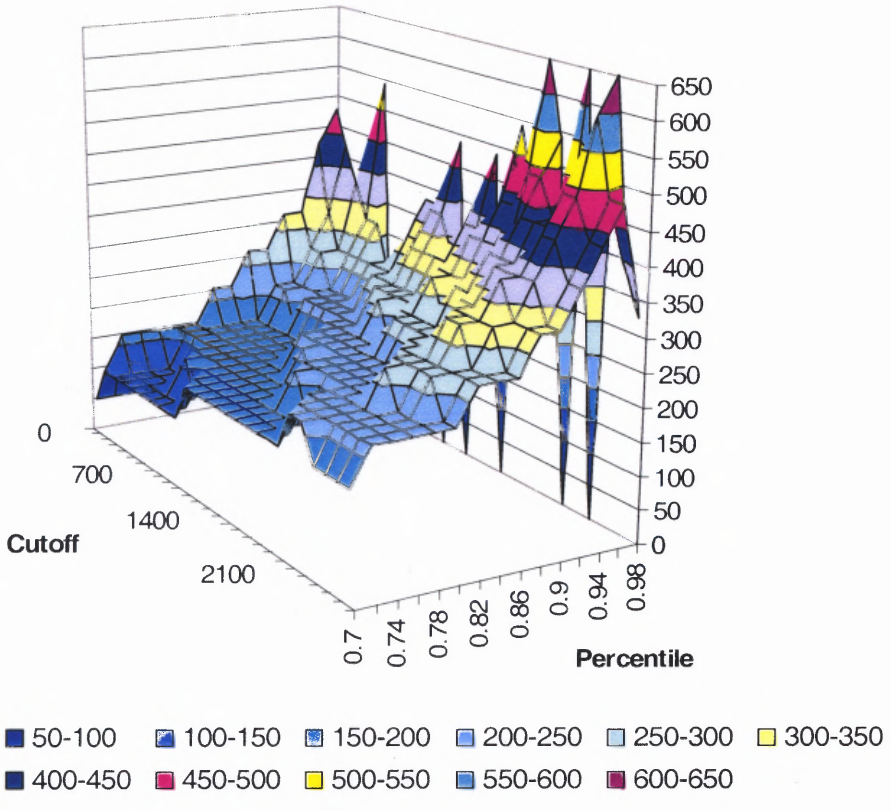


Figure 4.6 Three dimensional graph of the effect of the minimum intensity cutoff and percentile on the slope of the regressed percentile to the average intensity of the bins, with data obtained using MAS5 (Affymetrix).

For all three methods of probe-set intensity extraction, the average intercept is insensitive to the percentile and the minimum value cutoff except for the extreme values of those parameters: i.e. 98 percentile and no minimum value cutoff (see Figures 4.7, 4.8, 4.9). For the 80th percentile and a minimum intensity cutoff of 100, the average intercept is also very similar between the methods: 1.08 for MAS 5, 1.13 for dChip PM only and 1.12 for RMA.

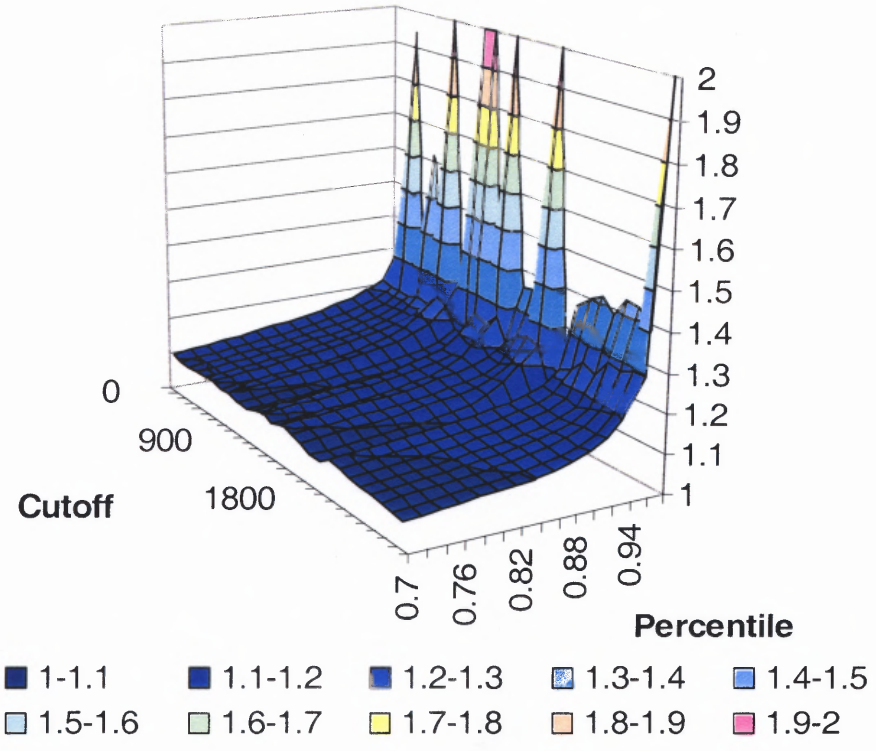


Figure 4.7 Three dimensional graph of the effect of the minimum intensity cutoff and percentile on the intercept of the regressed percentile to the average intensity of the bins, with data obtained using RMA (Ihaka and Gentleman, 1996).

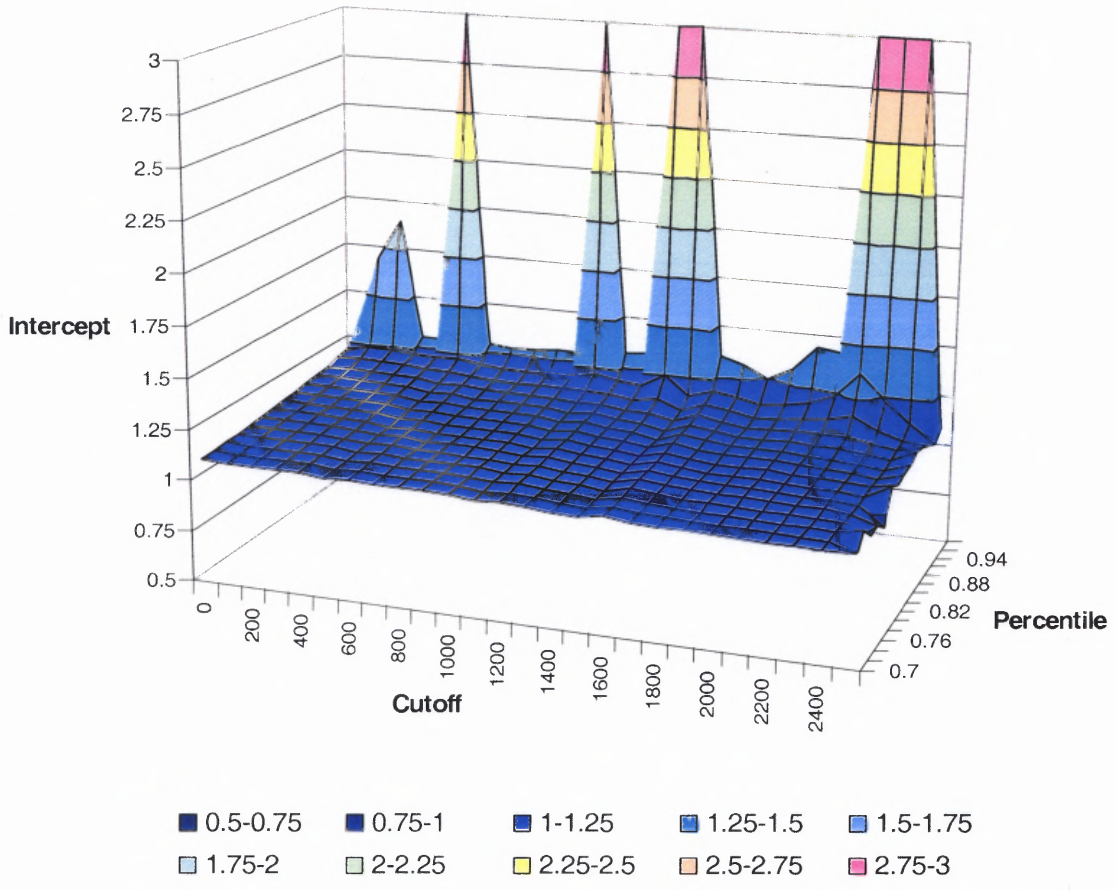


Figure 4.8 Three dimensional graph of the effect of the minimum intensity cutoff and percentile on the intercept of the regressed percentile to the average intensity of the bins, with data obtained using dChip PM only.

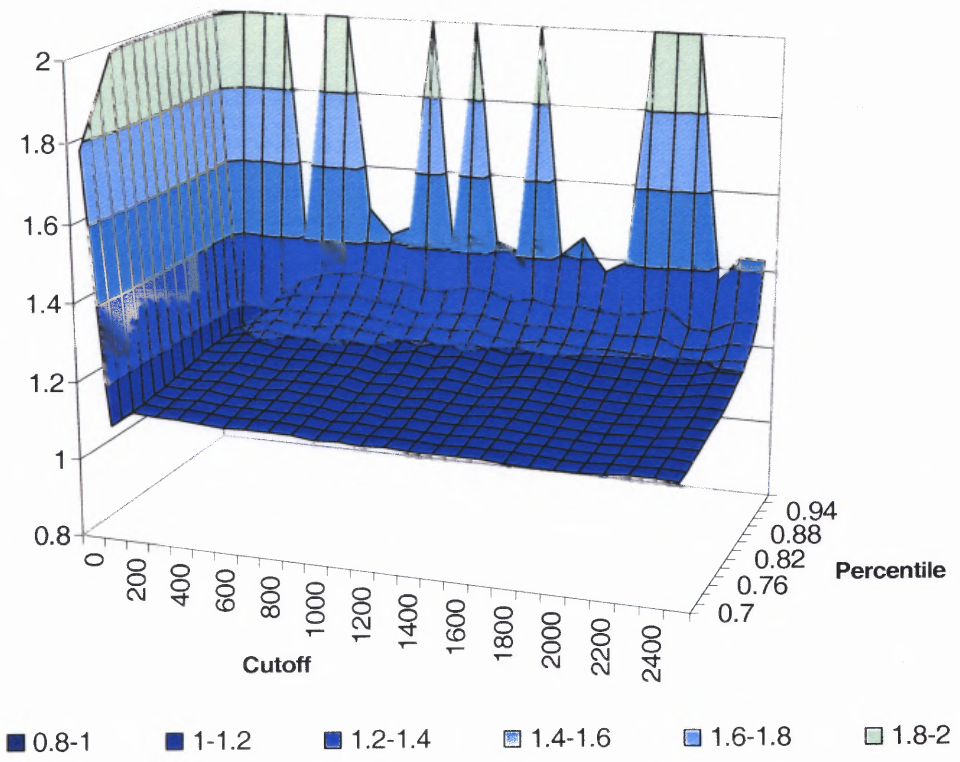


Figure 4.9 Three dimensional graph of the effect of the minimum intensity cutoff and percentile on the intercept of the regressed percentile to the average intensity of the bins, with data obtained using MAS5 (Affymetrix).

To evaluate the performance of the noise boundary model to identify the two fold change spiked-in genes in the results and eliminate false positives from the results, the sum of the rank of 12 out of the 14 spiked probes was evaluated. Two out of the 14 spiked genes were omitted because either their concentration was too low to be detectable (1597_at was spiked at the concentration of 0 to 0.25 pm) or the amount had saturated (1708_at was spiked at 256 and 512 pm). The noise boundary model was applied to any combination of the chips spiked at one concentration to all the other chips spiked at the other concentration. For every gene, up-regulation or down-regulation was then recorded

as the fold change was compared to the noise boundary. The maximum number of fold change directions was then divided by the number of comparisons. This resulted in an index called the Event ratio (Er) which is discussed in more detail in Chapter 5. Those probe-sets were ranked in descending order according to their Er score (Highest Er score is assigned the number 12626: i.e. the number of genes in the chip). The scores obtained were summed for the 12 spiked probe-sets and the results were normalized. The perfect score is 1. The result with the probe signal estimated with MAS 5 is presented in Figure 4.10. A plateau can be observed at 0.99 for most of the range for percentile and cutoff values. However there is a sharp decrease for a low cutoff, i.e. zero. The percentile also had little effect until a percentile higher than 94% was reached. The high percentile and low cutoff introduce more noise in the data setting the noise boundary model too high, therefore reducing the Er score of the spiked-in genes. The graph (Figure 4.11) for the signal estimation with RMA is similar to the MAS 5 in that most of the area covered by the simulation for the percentile and cutoff is a plateau at 0.99 for the sum of rank of the spiked in genes. In the same manner, the sum of ranks decreases for higher percentiles (above 96%). However, in this case, the rank is insensitive to low intensity cutoffs but decreases with higher cutoffs, with a first dip with gene intensities lower than 1600, and a second for intensities of 2100. The higher cutoffs are actually cutting most of the data to construct the model. For perspective, the chips are scaled to an average intensity of 300.

The dChip PM only algorithm did not seem to perform as well as MAS 5 and RMA, as its plateau was smaller and more sensitive to the parameters (Figure 4.12). The plateau average is also 0.99, but it is limited to percentiles lower than 90% and a

minimum intensity cutoff lower than 1100. Overall, MAS 5 seemed to be more robust, as the noise boundary model performance gradually decreased with the percentile.

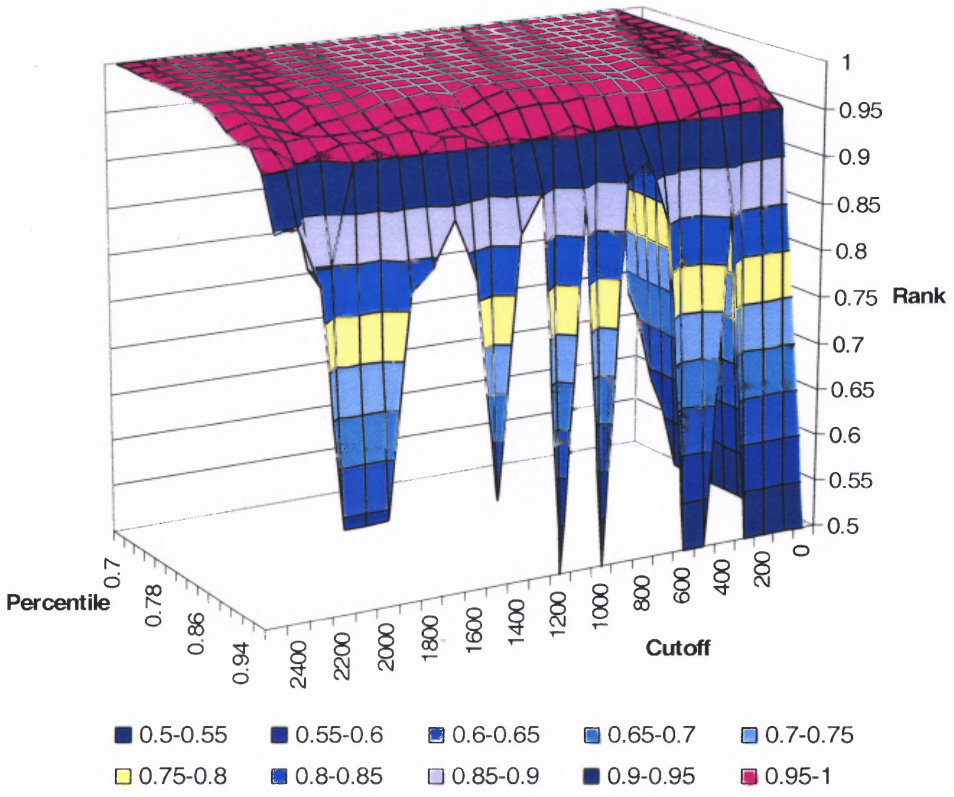


Figure 4.10 Three dimensional graph of the effect of the minimum intensity cutoff and percentile on the score of the spiked in genes of the replicate set of the Latin square dataset, with data obtained using MAS5 (Affymetrix).

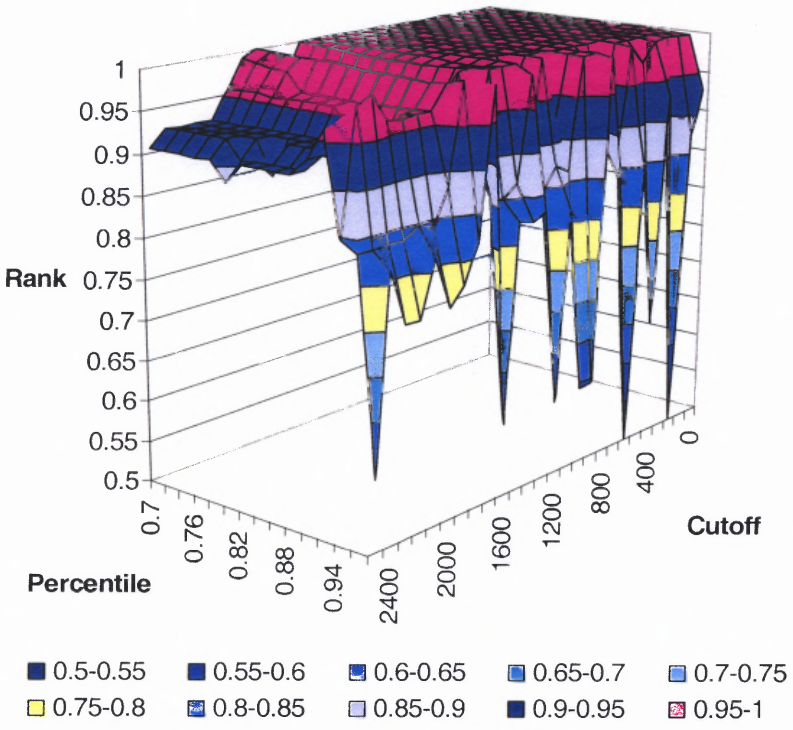


Figure 4.11 Three dimensional graph of the effect of the minimum intensity cutoff and percentile on the score of the spiked in genes of the replicate set of the Latin square dataset, with data obtained using RMA (Ihaka and Gentleman, 1996).

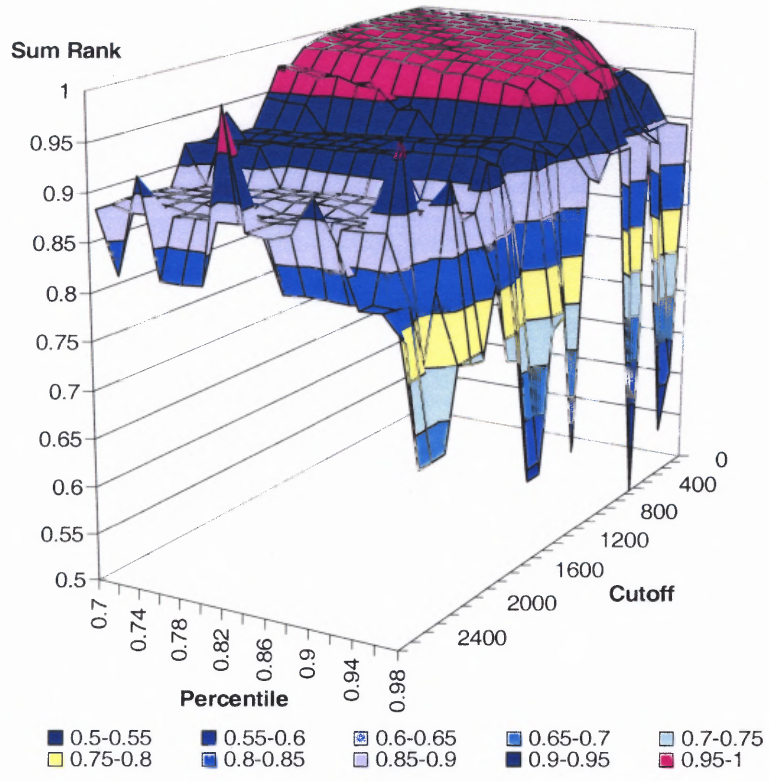


Figure 4.12 Three dimensional graph of the effect of the minimum intensity cutoff and percentile on the score of the spiked in genes of the replicate set of the Latin square dataset, with data obtained using dChip PM only.

A compromise has to be made between setting a noise boundary model to low, finding all the spiked genes with a lot of false positives and setting the noise model too high. At this point the model is so conservative that only a few spiked in genes are found. Figure 4.13 displays the false positive rate when an Er cutoff is set to 0.9, equivalent to a gene being consistently over or under expressed in 90% of the comparisons. The false positive rate is very high for a percentile of 0.98. This artifact is due to the presence of only one gene which is a false positive in the result set. The false positive rate decreases sharply with the percentiles and then increases again as the boundary model becomes less conservative.

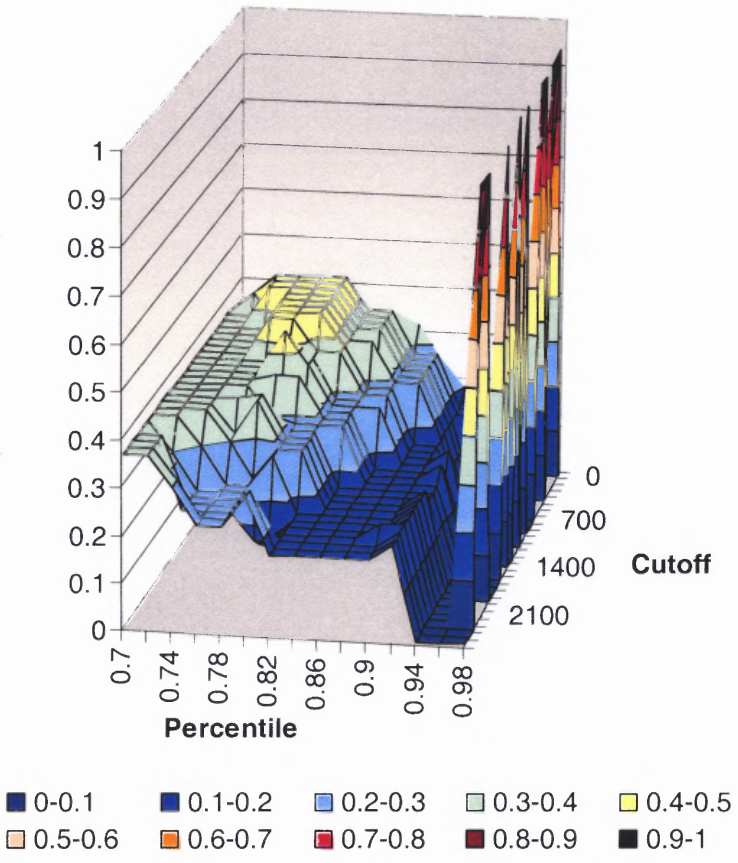


Figure 4.13 Three dimensional graph of the effect of the minimum intensity cutoff and percentile on the false positive rate of the genes with an Er index above 0.9 in the replicate set of the Latin square dataset, data obtained using MAS5 (Affymetrix).

4.4 Sensitivity Analysis Discussion

Noise modeling can be performed for all of the methods. MAS 5 was shown to be noisier for low intensities but the correction using the noise model seemed to perform very well as its performance decreased gracefully for finding spiked genes. MAS 5 will be used in further studies because of its wide use for signal intensity estimation and its robustness for the two parameters tested. A cutoff for the minimum intensity of 100 and the 80th percentile was selected for the model parameters as they are in regions where the slope,

intercept and rank are not very sensitive to change, and where the false positive rate is reasonable.

4.5 Evaluation of the Noise Model on Real Data

Figure 4.14 displays the 80th percentile error boundaries for five different normal tissues as a function of the inverse average bin intensity. Bins with an average intensity lower than 100 (above 0.01 in the Figure) were not displayed. They are below the minimum intensity cutoff and hinder the linearity relation of the percentile to bin intensity. A leveling off of the fold changes at high was also noticed; this leveling is due to saturation on the chip. To decrease the effect of the saturation on the regression, the top 8% of the genes were eliminated i.e. top 5 bins with lowest inverse average intensity. The slope and intercept were then calculated for each cancer dataset as they give an indication of the noise level at low and high expression values respectively. For each comparison of normal samples in a tissue, the slope and intercept were averaged (Table 4.1). There seems to be a negative correlation between the slopes and intercepts. The higher the intercept the lower the slope. If a dataset contains an inherent high background, the signal to noise ratio is decreased. The intercept will increase as the 80th percentile is going to be higher. The slope on the other hand is not going to increase, and might even decrease as the low intensity background noise remains constant. Before using this noise boundary model in Chapter 5 to find cancer markers, the stability of the slope and intercept for the different datasets must be evaluated. One of the differences with the Latin square replicate data set is that these public data consist of biological replicates of normal tissue instead of technical replicates. The cancer biopsies were not used in

designing the noise model as they might be more variable than normal tissue. The first two simulations were then performed for all normal tissue samples to confirm that the minimum intensity cutoff and percentile selected were also in regions where their slope was also insensitive to small changes (Figures 4.15, 4.16, 4.17, 4.18, 4.19). Figures for the simulation of the effect of the cut-off value and percentile on the intercepts are presented in Appendix B. For all the normal tissue, the results from the simulation are very similar.

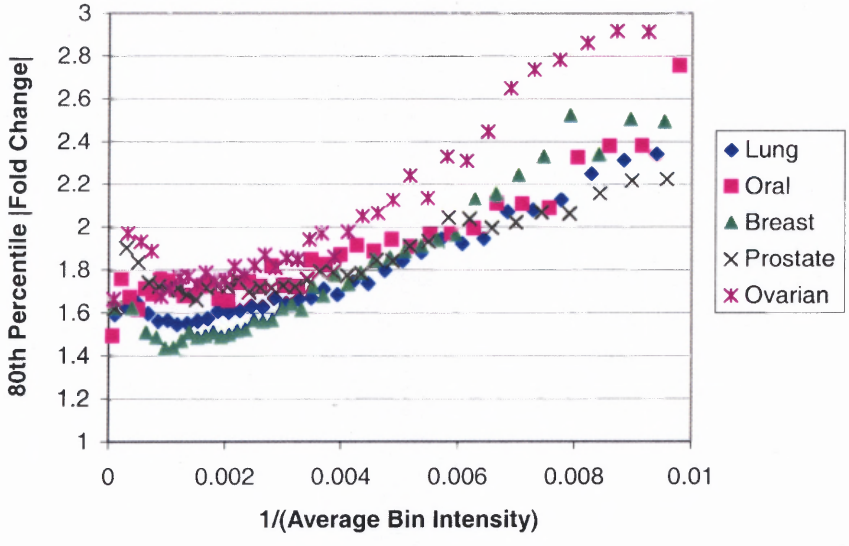


Figure 4.14 This Figure represents the 80th percentile for each of the five tissues plotted against the inverse of the average bin intensity. The different normal tissues are represented in color, ▲ for breast, * for Ovarian, x for Prostate, ■ for Oral and ♦ for lung.

Table 4.1 Average Slopes and Intercepts for the Different Tissue Types
This table displays the average slope and intercept of the regression of the 80th percentile of the bins by the inverse of the average expression per bin. The bin size was 200 and the minimum intensity cutoff was 100.

	Average Slope	Stdev	Average Intercept	Stdev
Lung normal	96	29	1.42	0.15
Breast Normal	139	33	1.24	0.06
Ovarian Normal	154	45	1.48	0.12
Prostate Normal	61	26	1.61	0.26
Oral Normal	89	12	1.55	0.22

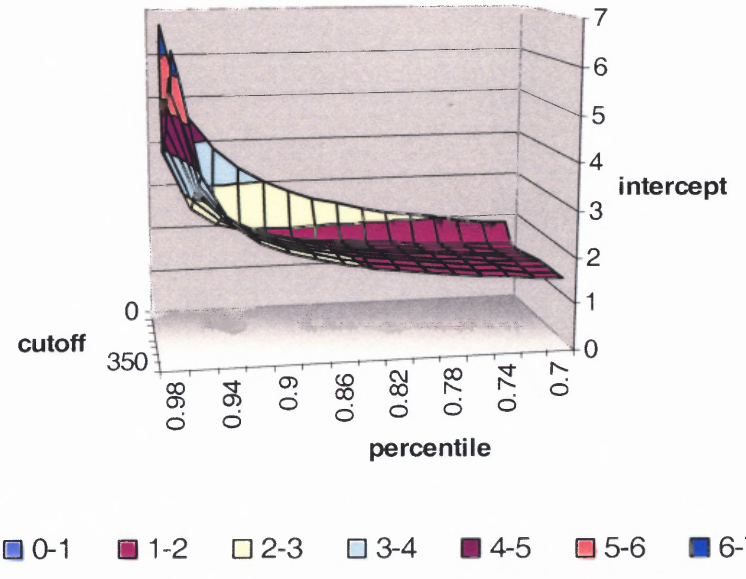


Figure 4.15 Three-dimensional graph of the effect of the minimum intensity cutoff and percentile selected on the slope of the regressed percentile to the average intensity of the bins for the prostate normal biopsies.

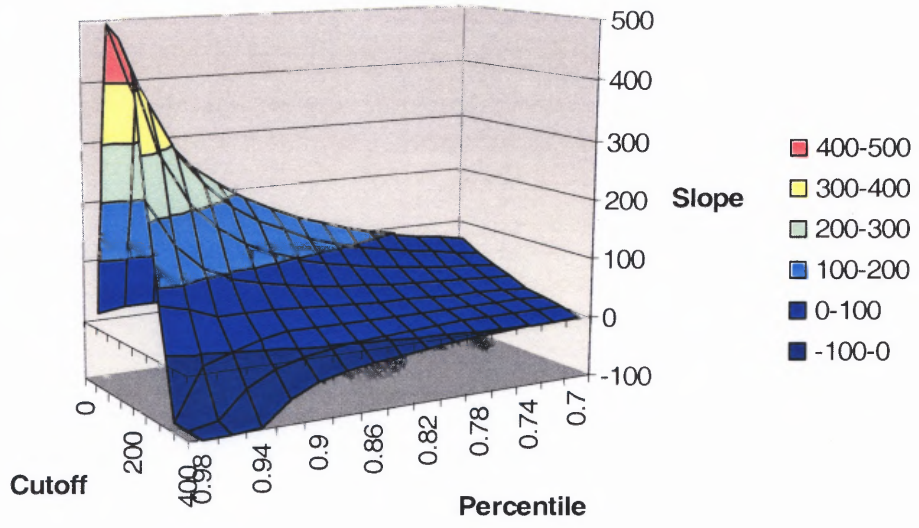


Figure 4.16 Three-dimensional graph of the effect of the minimum intensity cutoff and percentile selected on the slope of the regressed percentile to the average intensity of the bins for the lung normal biopsies.

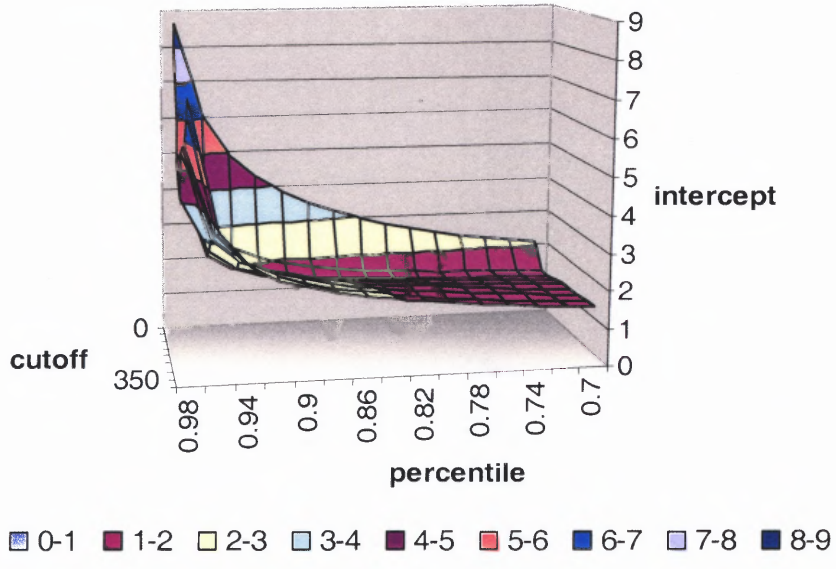


Figure 4.17 Three-dimensional graph of the effect of the minimum intensity cutoff and percentile selected on the slope of the regressed percentile to the average intensity of the bins for the ovarian normal biopsies.

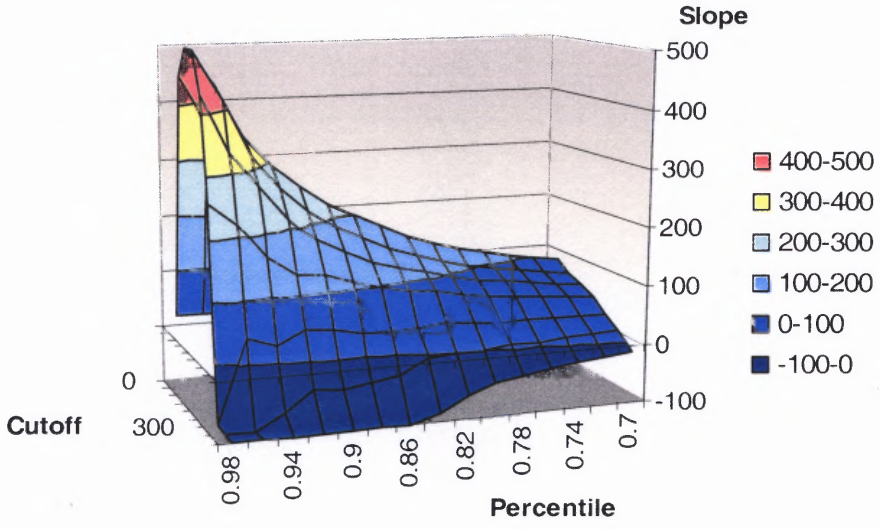


Figure 4.18 Three-dimensional graph of the effect of the minimum intensity cutoff and percentile selected on the slope of the regressed percentile to the average intensity of the bins for the breast normal biopsies.

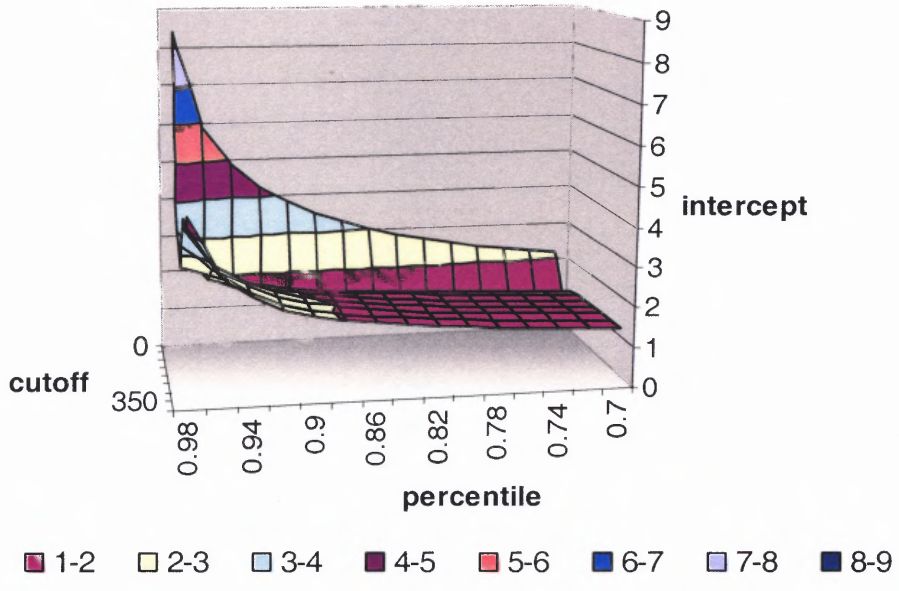


Figure 4.19 Three-dimensional graph of the effect of the minimum intensity cutoff and percentile selected on the slope of the regressed percentile to the average intensity of the bins for the oral normal biopsies.

4.6 Conclusion

There is a characteristic increase in the range of fold change values that occur at lower expression levels on Affymetrix GeneChips replicate arrays. This occurs with all techniques for estimating gene expression levels from GeneChips measurements. This increase in fold change was found to be consistent and could be characterized through regression analysis in the three different, commonly used, probe set intensities extraction methods (MAS5, dChip PM only and RMA). This noise was also shown to be consistent not only on replicate arrays but also on normal tissue replicate data. The boundary of this noise can be modeled for all data extraction methods and is found to fit well with an inverse linear function. From this observation, a noise boundary model was derived. This noise boundary is used in the next Chapter to determine if a fold change is above the noise level and relevant to the analysis, eliminating spurious false positives. It is essential to eliminate the background level fold changes since the multiplicity of the tests introduces hundreds of false positives. This model increases the trust and confidence for identifying differentially expressed gene markers. Another advantage of the noise boundary is that it is tailored to the noise level in the different normal tissues. This Chapter presented an original analysis of the noise. Using this noise boundary as a preprocessing step to an analysis method will help reduce the number of false positives obtained. Chapter 5 will apply this noise boundary model to a non-parametric analysis of microarray data from different research laboratories to delineate cancer markers.

CHAPTER 5

NONPARAMETIC DIRECTIONAL CHANGE ASSESMENT ALGORITHM IDENTIFIES TISSUE SPECIFIC MARKERS FOR DIFFERENT CANCER TYPES

5.1 Hypothesis

To be able to find markers, data from different research groups might be needed as each group has expertise on their cancer studied and access to samples. One particular concern when using microarray data derived from different labs is that chip-to-chip normalization cannot eliminate differences in scanner settings, image processing software, labeling, and hybridization protocols. The laser power on the scanner can be different from one scanner to another causing saturation of some spots. The quality of the RNA isolated can influence the mRNA species present and also the success of the labeling reaction in ways that are not well known, difficult to control and impossible to account for. A few studies (Ramaswamy et al., 2001; Su et al., 2001) successfully classified different cancers by their molecular profile on microarrays using hierarchical clustering and support vector machine (SVM) techniques (Brown et al., 2000). Both studies found that their markers comported a high number of genes that distinguished the normal tissues of origin.

The approach taken in this Chapter is radically different from previous efforts. Here, cancer samples are compared first to corresponding normal tissue, eliminating the tissue effect genes. Then the most discriminating genes for each cancer *vs.* normal cells are compared among cancers. The starting hypothesis is that the most reliable discriminating markers are transcripts that are differentially regulated in a consistent manner between normal and cancer biopsies. Also the normalization problems due to lab specific parameters

(scanner settings, labeling) and even tissue specific artifacts are avoided, as each cancer biopsy is compared to its corresponding normal tissue processed by the same research group, in the same environment. These environmental parameters and artifacts are assumed to be the same for the normal and cancer biopsies and should cancel out when compared. This allows the selection of genes that best separate normal biopsies from tumors. These classifiers were then evaluated to see if they are specific to the different types of cancers. Since gene expression measurements of individual Affymetrix GeneChips probe sets do not always follow a normal distribution, a non-parametric method was used.

The method to find discriminating markers uses an unweighted voting scheme. This non-parametric method for marker selection was chosen to avoid making any assumptions on the shape of the data distribution. However, one drawback is that errors in microarray expression measurements cannot be accounted for as they vary in scale across the dynamic range of the technique's sensitivity. To remedy this problem, the noise boundary model described in Chapter 4 was used. The computed boundary for the noise makes the selection criteria more stringent, eliminating many false positives signals, and highlighting genes that are consistently differentially expressed in comparisons between a cancer and its corresponding normal tissue. This integrative approach can highlight sets of distinct transcripts distinguishing a variety of solid tumors.

5.2 Materials and Methods

All of the microarray data used in this analysis was derived from RNA isolated from biopsies and hybridized on Affymetrix GeneChips HG-U95A, HG-U95Av2 or HG-U133A. All the research groups used the same standard procedure for labeling the cRNA, hybridization and scanning (Wodicka et al., 1997). The datasets were obtained from several different sources: Data from 24 breast cancer biopsies were from Su et al.(Su et al., 2001), and the three corresponding normal breast tissue biopsies were provided by Garret Hampton from the Genomics Institute of the Novartis Research Foundation. For prostate cancer, the dataset was derived from 21 tumors and 8 normal biopsies (Welsh et al., 2001a) whereas the ovarian cancer dataset originated from 14 tumor and four normal biopsies (Welsh et al., 2001b). Finally, the lung cancer dataset consisted of biopsies from 61 samples of lung adenocarcinoma, 20 lung carcinoids, six small cell lung cancer, 21 squamous lung cancers, and 17 normal lung tissues (Bhattacharjee et al., 2001). Out of the 61 adenocarcinoma samples, 19 were replicates and 52 were sub-divided into five categories according to Bhattacharjee et al.(2001) (Bhattacharjee et al., 2001): seven in cluster 1, nine in cluster 2, 15 in cluster 3, 13 in cluster 4, and eight samples of colon metastasis. The oral cancer dataset consisted of 4 normal and 16 oral cancer biopsies (Toruner et al., 2004). The directional change assessment and the noise model algorithms were programmed using Python, and the comparison for markers was performed with Excel. Latest annotations and Gene Ontology classification were downloaded from the NetAffx™ Analysis Center[3].

[3] http://www.affymetrix.com/analysis/index.affx March 2004

5.3 Nonparametric Microarray Data Analysis

Several methods have been described for combining data from microarray experiments where there are replicates for both experimental and control conditions. Often, the numbers of replicates are small and the distribution is not normal. For the same difference in mean, depending on the distribution of the data, the overlap of two distributions can be dramatically different. Our ideal markers would be genes with no overlap in their distribution; the consistency of change is therefore more interesting than the amplitude. There is still a problem due to the low number of replicates, the multiple testing of 12,000 genes and the inherently noisy measurement. For this reason, the noise boundary model was incorporated with the non-parametric data mining. The noise boundary helps eliminate some of the noise that is proportional to the probe intensity measured and helps eliminate false positives due to chance, because we are performing multiple testing on more than 12,000 genes with random noise. The combination of the non-parametric voting scheme to find consistently differentially regulated genes with the noise model will be referred as the directional change assessment algorithm in the rest of this Chapter. For each transcript, the ratio of expression intensities (fold change) of each cancer biopsy to all normals was determined. If the absolute value of the ratio is above the noise boundary, up-regulation (+) or down regulation (-) is recorded. If the ratio is below the value given by the noise boundary for the average of the intensities, then the fold-change is considered insignificant and a no-change (0) direction is assigned. The Event ratio (Er) is described by:

$$Er = \frac{\max\{\#positive - \#negative)}{\#comparisons}$$

The closer the Er is to 1, the more consistent the direction of change is for that gene. Conversely, if the Er score is close to 0.5, then the gene is inconsistent with regard to its

directionality, considered noisy and thus cannot be used as a reliable marker for disease classification. To test the validity of this approach, the samples were shuffled 100 times between the categories (cancer and normal) and Er computation was repeated. Each time the data was shuffled, the probe sets were sorted by descending Er scores and the probe set information was discarded and replaced by its rank. The average and standard deviation of the ranks was then computed and compared to the results obtained for the cancer versus normal biopsies. For all the comparisons performed, higher Er scores were obtained in the case of a cancer versus normal classification than with randomly shuffled sets. These results confirm that the genes found significant with a high Er score are not just random in this case. An illustration of the results obtained with the breast cancer versus normal biopsies can be seen in Figure 5.1, the Figures for the other cancers are in Appendix C.

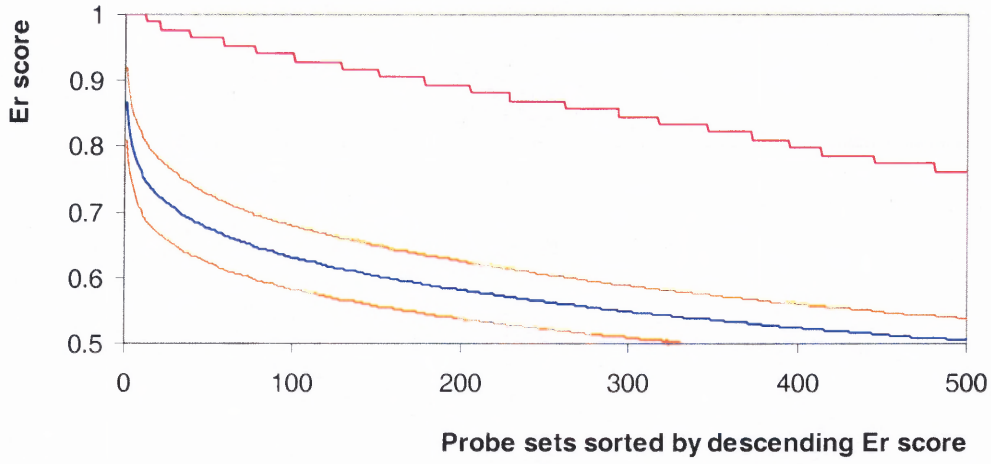


Figure 5.1 Comparison of the Er score of the 500 top ranked probe sets for breast cancer versus normal breast biopsies. Er score for the real breast cancer vs. normal biopsies —, Average Er score of the 500 top ranked probe sets of the 100 shuffling sets —, one standard deviation away form the average shuffled sets —.

5.4 Testing the Er Algorithm

To compare the Er algorithm with the noise model to other commonly used analysis methods, the replicate set from the Latin square dataset was used. In this dataset, fourteen probe sets were spiked in at a two fold level. The T-test performed on this data found 175 probe sets with a p-value below 0.01, including the fourteen spiked genes. Although this method collected all the spiked genes, it also highlighted 161 false positives. The percentage of true to false positives is only 8.6%. There is a multi-testing problem, with 12,000 tests times the type I error rate: 0.01, 120 probe sets are expected to have a p-value below 0.01. To correct for the multiple testing, one common method is the Bonferroni (Benjamini and Hochberg, 1995; Bonferroni, 1936; Perneger, 1998) method which found seven genes to be significant with six true positives. The methods of Hochberg and Simes (Benjamini and Hochberg, 1995), both found 16 significant genes with 11 true positive. The SAM (Tusher et

al., 2001) method also corrects for multiple testing, found 21 genes significant including 12 true positive with a delta of 1.54. Although they were able to find 85% of the spiked genes, their false positive rate was underestimated; they estimated a median false positive rate of 4.58% when it was 42% instead. The Er model described above with a cut-off of 0.9 found 12 genes with 8 of them being true positives. Without the noise model, 93 were found with an Er score above 0.9. The Er model with the noise boundary model is well within the separation levels of the standard techniques for eliminating spurious false positives. It would be interesting to compare these methods on multiple datasets as the results and performance of the different techniques may be dataset dependant. In the Latin square dataset, the spiked genes are independent, but in a real scenario the genes may not be independent and the result of one test may depend on the levels and test results of other genes. It is not known how much this affects the standard multiple test corrections. Also the Latin square data set is a very clean dataset with very little noise and few differences in scaling between chips (Affymetrix, 2002b). Under real world situations, data may be noisier and this may affect the performance of the different methods. It is worth noticing, that the fold change that the probe sets were spiked at was two fold, but the actual read intensity only shows an average of 1.53 fold. These methods decrease the number of false positives compared to the t-test alone but some true positive are occasionally missed. This is partly due to the fact that the spiked genes were added at a range of concentration testing the limit of detection from very low to high signal saturation.

5.5 Differentially Expressed Genes

The Er model was then used to compare each cancer biopsy to its corresponding normal tissue. In the absence of error modeling, the directional change algorithm identified 1,910 probe-sets that had an Er score above 0.9 in ovarian cancer, 1,355 in breast cancer, 1,730 in oral cancer and 322 in prostate cancer. In contrast, incorporation of error modeling dramatically reduced the number of probe-sets with Er scores above 0.9 to 272 for ovarian, 177 for breast, 129 for oral cancer and 2 for prostate cancer. For lung cancer biopsies, the distinct sub-classes were compared against normal tissues and 15 probe-sets with an Er value above 0.9 in all comparisons were uncovered. The following sections will present the results of the algorithm for the different tissues. Some gene markers found are already known markers and or have a biological function that might take part in the oncogenic process. These markers are reviewed below. Other gene markers may sometimes not seem relevant to the condition. These might be genes co-regulated or downstream from a pathway affected in cancer, and they can be very useful for diagnostic and classification. There are many ways to regulate proteins concentration in a cell and it might not always be seen at the RNA level, but its effect can be seen on the RNA levels of the downstream genes expression. Although not reviewed here, these genes are very useful for diagnostic and classification.

5.5.1 Breast Cancer

The breast cancer dataset is composed of 21 infiltrating ductal carcinoma and 2 normal biopsy samples, with 2 technical replicates. One hundred and seventy seven probe sets were found to be differentially regulated between the normal and cancer samples with Er scores above 0.9, most of them being down regulated (151 out of 171). Thirty two of those probe sets encoded for different ribosomal proteins, all down-regulated, averaging 2-3 fold down compared to the normal samples. These genes are located on 19 different chromosomes so a deletion would not be able to explain this down regulation. Methylation of ribosomal DNA has been reported in human breast cancer (Yan et al., 2000) and could be the cause of the decrease in mRNA encoding ribosomal proteins. Other notable genes involved were: thrombospondin 2 (THBS2), a known marker (Clezardin et al., 1999) involved in angiogenesis (de Fraipont et al., 2001) was up regulated, and caveolin 1 (CAV1) was down regulated. Caveolin-1 acts as a tumor suppressor protein. When expressed at high levels, it inhibits cell proliferation (Fiucci et al., 2002).

Table 5.1 Top 30 Most Differentially Expressed Probe-sets in Breast Cancer Compared to Normal Breast Biopsies

Probe Set ID	Description	Symbol	Fold Change	ER Score
40304_at	bullous pemphigoid antigen 1, 230/240kDa	BPAG1	-17.8	1
34203_at	calponin 1, basic, smooth muscle	CNN1	-13.2	1
32666_at	chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)	CXCL12	-8.8	1
38430_at	fatty acid binding protein 4, adipocyte	FABP4	-241.4	1
31719_at	fibronectin 1	FN1	14.1	1
32847_at	myosin, light polypeptide kinase	MYLK	-10.0	1
41178_at	ribosomal protein L11	RPL11	-3.0	1
34643_at	ribosomal protein S4, X-linked	RPS4X	-2.3	1
506_s_at	signal transducer and activator of transcription 5A	STAT5A	-6.9	1
41400_at	thymidine kinase 1, soluble	TK1	6.6	1
36119_at	caveolin 1, caveolae protein, 22kDa	CAV1	-7.2	0.99
37187_at	chemokine (C-X-C motif) ligand 2	CXCL2	-12.3	0.99
41618_at	collagen, type XVII, alpha 1	COL17A1	-6.9	0.99
33862_at	phosphatidic acid phosphatase type 2B	PPAP2B	-4.3	0.99
2016_s_at	ribosomal protein L10	RPL10	-2.6	0.99
31791_at	tumor protein p73-like	TP73L	-4.1	0.99
39864_at	cold inducible RNA binding protein	CIRBP	-3.8	0.98
38700_at	cysteine and glycine-rich protein 1	CSRP1	-3.0	0.98
32648_at	delta-like 1 homolog (Drosophila)	DLK1	-9.8	0.98
40887_g_at	eukaryotic translation elongation factor 1 alpha 1	EEF1A1	-2.3	0.98
38627_at	hepatic leukemia factor	HLF	-7.3	0.98
38737_at	insulin-like growth factor 1 (somatomedin C)	IGF1	-11.8	0.98
33674_at	ribosomal protein L29	RPL29	-2.6	0.98
31722_at	ribosomal protein L3	RPL3	-3.3	0.98
33657_at	ribosomal protein L34	RPL34	-3.4	0.98
33656_at	ribosomal protein L37	RPL37	-2.6	0.98
32466_at	ribosomal protein L41	RPL41	-2.7	0.98
36061_at	semaphorin 5A	SEMA5A	-3.7	0.98
31508_at	thioredoxin interacting protein	TXNIP	-4.7	0.98
1814_at	transforming growth factor, beta receptor II	TGFBR2	-7.6	0.98

5.5.2 Ovarian Cancer

The ovarian dataset is derived from four normal and 14 tumor biopsies (serous papillary ovarian carcinoma) (Welsh et al., 2001b). Two hundred and seventy two probe-sets were identified in the ovarian dataset with an Er score above 0.9, of which the top 30 are presented in table 5-2. Notably, genes are involved in cell-to-cell adhesion and are epithelial cells markers, such as the secreted phosphoprotein 1 (SPP1), claudin 3, 4 and 7 (CLDN 3, 4, 7), keratin 7, 8 and 18 (KRT 7,8,18), agrin, and cahedrin 1 and 6 (CDH 1, 6), were up regulated by at least 3-fold in ovarian cancer samples. As was previously suggested (Adib et al., 2004) this might suggest an epithelial origin of the tumors. Eighty to ninety percent of the ovarian cancers arise from the epithelia (Auersperg et al., 1999). Genes involved in the regulation of cell growth; v-fos (FOS), quiescin Q6 (QSCN6), protease serine 11 (PRSS11), the mortality factor like 2 (MORF4L2) as well as the cyclin dependant kinase inhibitor 1A and 1C (CDKN1A, CDKN1C) were down regulated, whereas cyclin B1 (CCNB1) was up regulated by at least 30 fold over the normal sample levels. The tumor-associated calcium signal transducer 2 (TACSTD2) and the CD24 antigen associated with metastasis and angiogenesis were also up-regulated. These results are in agreement with the results from the original and a similar study (Welsh et al., 2001b), (Adib et al., 2004).

Table 5.2 Top 30 Most Differentially Expressed Probe-sets in Ovarian Cancer Compared to Normal Ovarian Biopsies

Probe Set ID	Description	Symbol	Fold Change	ER Score
34343_at	steroidogenic acute regulatory protein	STAR	-129.1	1
266_s_at	CD24 antigen (small cell lung carcinoma cluster 4 antigen)	CD24	118.5	1
39701_at	paternally expressed 3	PEG3	-97.7	1
291_s_at	tumor-associated calcium signal transducer 2	TACSTD2	91.9	1
32275_at	secretory leukocyte protease inhibitor (antileukoproteinase)	SLPI	73.7	1
33904_at	claudin 3	CLDN3	66.4	1
41294_at	keratin 7	KRT7	58.4	1
35766_at	keratin 18	KRT18	53.4	1
38324_at	liver-specific bHLH-Zip transcription factor	LISCH7	53.1	1
2094_s_at	v-fos FBJ murine osteosarcoma viral oncogene homolog	FOS	-51.1	1
34213_at	KIBRA protein	KIBRA	48.8	1
37394_at	complement component 7	C7	-48.7	1
36869_at	paired box gene 8	PAX8	47.2	1
35717_at	ATP-binding cassette, sub-family A (ABC1), member 8	ABCA8	-46.5	1
41412_at	ortholog of rat pippin	PIPPIN	-46.1	1
35276_at	claudin 4	CLDN4	45.8	1
37247_at	transcription factor 21	TCF21	-37.5	1
32154_at	transcription factor AP-2 alpha	TFAP2A	30.7	1
1507_s_at	endothelin receptor type A	EDNRA	-28.8	1
38482_at	claudin 7	CLDN7	28.6	1
36133_at	desmoplakin	DSP	27.0	1
35168_f_at	collagen, type XVI, alpha 1	COL16A1	-26.3	1
35389_s_at	ATP-binding cassette, sub-family A (ABC1), member 6	ABCA6	-24.6	1
38291_at	proenkephalin	PENK	-24.2	1
35390_at	ATP-binding cassette, sub-family A (ABC1), member 6	ABCA6	-20.4	1
41812_s_at	nucleoporin 210	NUP210	19.8	1
36917_at	laminin, alpha 2 (merosin, congenital muscular dystrophy)	LAMA2	-19.6	1
37628_at	monoamine oxidase B	MAOB	-18.8	1
37985_at	lamin B1	LMNB1	18.7	1
1897_at	transforming growth factor, beta receptor III	TGFBR3	-8.6	1

5.5.3 Oral Cancer

The oral cancer dataset was composed of four normal biopsies and 16 cancerous biopsies. Unlike the other sets, the RNA was obtained from laser capture and was hybridized on Affymetrix Hu133A GeneChips. The major gene classes affected are related to cytoskeletton organization, development and differentiation. Keratin 4 and 13 (KRT4 and 13), sciellin (SEL), and the small proline-rich protein 1A and 1B (SPPR1A and 1B) involved in epidermal differentiation, are all down regulated. This result is consistent with the results from the original study (Toruner et al., 2004) and a study on esophageal squamous cell carcinoma (Luo et al., 2004), a similar tissue of origin. The down regulation of these genes indicates a de-differentiation of the cancer cells compared to the normal epithelial cells. The next family of genes involved included genes having an effect on cell to cell adhesion: desmoglein 1, 2 and 3 (DSG1, 2 and 3) and claudin 17 (CLDN17). Associated with the up regulated matrix metalloproteinase 1 and 13 (MMP1 and 13) that encode collagenases, destroying the interstitial tissue, these changes may allow cells to infiltrate the connective tissue. Two of the selected up regulated genes MMP1, LGALS1, and the down regulated gene KRT4 were confirmed by RT-PCR by the original study (Toruner et al., 2004).

Table 5.3 Top 30 Most Differentially Expressed Probe-sets in Oral Cancer Compared to Normal Biopsies

Probe Set	Description	Symbol	Fold Change	ER Score
206605_at	26 serine protease	P11	-139.9	1
219684_at	28kD interferon responsive protein	IFRG28	17.2	1
206513_at	absent in melanoma 2	AIM2	5.6	1
201753_s_at	adducin 3 (gamma)	ADD3	5.9	1
218876_at	brain specific protein	CGI-38	-28.3	1
220090_at	chromosome 1 open reading frame 10	C1orf10	-293.2	1
204439_at	chromosome 1 open reading frame 29	C1orf29	112.2	1
208747_s_at	complement component 1, s subcomponent	C1S	9.0	1
208126_s_at	cytochrome P450, family 2, subfamily C, polypeptide 18	CYP2C18	-22.0	1
217901_at	desmoglein 2	DSG2	6.0	1
219597_s_at	dual oxidase 1	DUOX1	-8.6	1
218396_at	hypothetical protein FLJ10381	FLJ10381	4.6	1
201163_s_at	insulin-like growth factor binding protein 7	IGFBP7	7.3	1
214599_at	involucrin	IVL	-31.0	1
220782_x_at	kallikrein 12	KLK12	-29.0	1
213050_at	KIAA0633 protein	COBL	-40.3	1
212314_at	KIAA0746 protein	KIAA0746	12.5	1
204777_s_at	mal, T-cell differentiation protein	MAL	-87.6	1
219554_at	Rhesus blood group, C glycoprotein	RHCG	-210.5	1
206008_at	transglutaminase 1	TGM1	-69.3	1
206004_at	transglutaminase 3	TGM3	-55.6	1
210986_s_at	tropomyosin 1 (alpha)	TPM1	13.7	1
201325_s_at	epithelial membrane protein 1	EMP1	-7.0	0.99
211597_s_at	homeodomain-only protein	HOP	-32.0	0.99
213294_at	hypothetical protein FLJ38348	FLJ38348	4.0	0.99
202983_at	SWI/SNF related, matrix associated member 3	SMARCA3	4.4	0.99
221328_at	claudin 17	CLDN17	-8.2	0.97
204750_s_at	desmocollin 2	DSC2	-15.7	0.97
213187_x_at	ferritin, light polypeptide	FTL	4.4	0.97
214091_s_at	glutathione peroxidase 3 (plasma)	GPX3	-11.3	0.97

5.5.4 Prostate Cancer

The prostate dataset was composed of eight normal samples and 17 cancer samples including 3 paired normal/tumor samples from the same patients. Only two genes have an Er above 0.9. This limited number could be due to the fact that this dataset has many more samples, especially normal samples than the breast, ovarian and oral cancers, but the number of samples does not explain everything. By redoing the analysis with only 3 normal samples, only 16-20 probe sets were obtained with an Er ratio above 0.9. This cancer seems to be more specific, with only a few key genes constantly differentially regulated. Top markers for prostate cancer are hepsin (HPN), single-minded homolog 2 (SIM2), LIM protein (LIM) and alpha-methylacyl-CoA racemase (AMACR). Hepsin and alpha-methylacyl-CoA racemase are known markers of prostate cancer (Brooks, 2002; Ernst et al., 2002; Rhodes et al., 2002). PSA (prostate-specific antigen), a protein marker commonly used in the prostate cancer test, was not identified as its probe set was saturated in chips from both prostate normal and cancer biopsies.

Table 5.4 Top 30 Most Differentially Expressed Probe-sets in Prostate Cancer Compared to Normal Prostate Biopsies

Probe Set ID	Description	Symbol	Fold Change	ER Score
37639_at	hepsin (transmembrane protease, serine 1)	HPN	4.9	0.92
39608_at	single-minded homolog 2 (Drosophila)	SIM2	23.3	0.90
37366_at	LIM protein (similar to rat protein kinase C-binding enigma)	LIM	4.4	0.88
41706_at	alpha-methylacyl-CoA racemase	AMACR	15.6	0.88
35330_at	filamin C, gamma (actin binding protein 280)	FLNC	-8.5	0.88
34203_at	calponin 1, basic, smooth muscle	CNN1	-38.9	0.85
40060_r_at	LIM protein (similar to rat protein kinase C-binding enigma)	LIM	4.6	0.85
40776_at	desmin	DES	-11.6	0.85
661_at	growth arrest-specific 1	GAS1	-4.6	0.84
31831_at	smoothelin	SMTN	-4.2	0.83
36780_at	clusterin (testosterone-repressed prostate message 2)	CLU	-3.5	0.83
40674_s_at	homeo box C6	HOXC6	7.2	0.83
32243_g_at	crystallin, alpha B	CRYAB	-4.5	0.83
34320_at	polymerase I and transcript release factor	PTRF	-3.3	0.83
36497_at	chromosome 14 open reading frame 78	C14orf78	-11.5	0.83
36491_at	thymosin, beta, identified in neuroblastoma cells	TMSNB	7.7	0.82
773_at	myosin, heavy polypeptide 11, smooth muscle	MYH11	-29.7	0.82
774_g_at	myosin, heavy polypeptide 11, smooth muscle	MYH11	-17.9	0.82
36432_at	methylcrotonoyl-Coenzyme A carboxylase 2 (beta)	MCCC2	7.5	0.81
38661_at	RNA-binding region (RNP1, RRM) containing 1	RNPC1	-4.4	0.81
36149_at	dihydropyrimidinase-like 3	DPYSL3	-10.6	0.81
38700_at	cysteine and glycine-rich protein 1	CSRP1	-4.1	0.81
1276_g_at	RNA binding protein with multiple splicing	RBPMS	-3.8	0.80
32313_at	tropomyosin 2 (beta)	TPM2	-11.5	0.80
34377_at	ATPase, Na+/K+ transporting, alpha 2 (+) polypeptide	ATP1A2	-4.7	0.80
36119_at	caveolin 1, caveolae protein, 22kDa	CAV1	-3.1	0.80
37765_at	leiomodin 1 (smooth muscle)	LMOD1	-6.4	0.80
39145_at	myosin, light polypeptide 9, regulatory	MYL9	-8.2	0.80
39544_at	desmuslin	DMN	-7.7	0.80
39790_at	ATPase, Ca++ transporting, cardiac muscle, slow twitch 2	ATP2A2	-2.6	0.80

5.5.5 Lung Cancer

The lung cancer dataset consisted of biopsies from 61 samples of lung adenocarcinoma, 20 lung carcinoids, six small cell lung cancer, 21 squamous lung cancers, and 17 normal lung tissues (Bhattacharjee et al., 2001). Out of the 61 adenocarcinoma samples, 19 were replicates and 52 samples could be classified into five distinct categories of adenocarcinomas (Bhattacharjee et al., 2001): seven in cluster 1, nine in cluster 2, 15 in cluster 3, 13 in cluster 4, and eight samples of colon metastasis. The goal of this study was to try to find markers specific to cancer located in the lung. To be able to achieve that without running into problems with under-representating some of the different cancers, every lung cancer type was compared to the normal lung samples. The individual Er scores were then multiplied and the genes with the highest score were selected (Table 5.5). All the genes selected were down-regulated. Nineteen of the 23 genes that had a Gene Ontology[4] annotation have their respective protein located inside the plasma membrane or in the extra-cellular space. By comparing many different types of cancer (i.e. adenocarcinoma, small cells, colon metastasis) to the normal samples, common genes related to the interaction with the lung milieu were isolated. The down regulation of a macrophage receptor with collagenous structure (MARCO) reflects the presence of less macrophages in the tumor tissue compared to the normal lung epithelium. Other mRNAs encoding for proteins involved in cell to cell signaling (TEK, GPRK5), and heterophilic cell adhesion (FCN3, TNA, MFAP4) are down-regulated and may be an evasion sign of the tumor cells from the immune system.

---

[4] http://www.geneontology.org/ March 2004

Table 5.5 Top 30 Most Differentially Expressed Probe-sets in Lung Cancer
The ER score was multiplied for the different lung cancer subcategories to obtain the ER* indicative of the genes the most differentially expressed in all the lung cancers.

Probe Set ID	Description	Symbol	Av. Fold Ch.	ER*	Min ER
36569_at	tetranectin (plasminogen binding protein)	TNA	-9.3	0.99	1.00
34708_at	ficolin (collagen/fibrinogen domain containing) 3	FCN3	-14.8	0.97	0.99
38430_at	fatty acid binding protein 4, adipocyte	FABP4	-70.5	0.95	0.98
35868_at	advanced glycosylation end product-specific receptor	AGER	-24.8	0.95	0.96
1596_g_at	TEK tyrosine kinase	TEK	-14.4	0.93	0.96
32542_at	four and a half LIM domains 1	FHL1	-15.1	0.92	0.96
37398_at	platelet/endothelial cell adhesion molecule (CD31 antigen)	PECAM1	-7.2	0.91	0.92
37247_at	transcription factor 21	TCF21	-14.2	0.87	0.89
39066_at	microfibrillar-associated protein 4	MFAP4	-30.5	0.87	0.92
39631_at	epithelial membrane protein 2	EMP2	-10.1	0.86	0.94
36119_at	caveolin 1, caveolae protein, 22kDa	CAV1	-15.6	0.85	0.89
40331_at	macrophage receptor with collagenous structure	MARCO	-26.1	0.85	0.93
40282_s_at	D component of complement (adipsin)	DF	-6.4	0.85	0.92
36156_at	aquaporin 1	AQP1	-9.8	0.84	0.86
41096_at	S100 calcium binding protein A8 (calgranulin)	S100A8	-12.0	0.83	0.89
1814_at	transforming growth factor, beta receptor II	TGFBR2	-11.9	0.82	0.92
34210_at	CDW52 antigen (CAMPATH-1 antigen)	CDW52	-24.1	0.80	0.90
38177_at	receptor (calcitonin) activity modifying protein 2	RAMP2	-8.1	0.80	0.86
38995_at	claudin 5	CLDN5	-16.2	0.79	0.92
35730_at	alcohol dehydrogenase IB (class I), beta polypeptide	ADH1B	-29.6	0.79	0.88
37967_at	leukocyte specific transcript 1	LST1	-14.1	0.78	0.87
40560_at	T-box 2	TBX2	-11.3	0.77	0.90
607_s_at	von Willebrand factor	VWF	-6.5	0.77	0.88
37168_at	lysosomal-associated membrane protein 3	LAMP3	-17.5	0.73	0.80
38026_at	fibulin 1	FBLN1	-16.4	0.71	0.86
40994_at	G protein-coupled receptor kinase 5	GPRK5	-4.8	0.70	0.86
38239_at	claudin 18	CLDN18	-39.8	0.68	0.81
32527_at	adipose specific 2	APM2	-31.6	0.67	0.87
39350_at	glypican 3	GPC3	-10.7	0.66	0.83
37027_at	hypothetical protein MGC5395	MGC5395	-4.5	0.65	0.87

These results represent a filtered molecular portrait of the reliable transcripts that are differentially regulated in a tumor compared to its corresponding normal tissue. A major strength of our mining strategy is that ranking differentially expressed transcripts by decreasing Er scores instead of fold-changes enables the filtering of false positives (attributed to noisy genes or poorly designed probe sets on the GeneChip) that are missed by clustering algorithms.

5.6 Cancer-Specific Biomarkers

A major advantage of providing Er scores for differentially expressed cancer transcripts is that it provides an easy statistical metric that can be used to underscore markers that are unique to a particular cancer. In this case, although the Er is not a statistical test and the same Er score can vary in its significance depending on the number of samples studied, we accomplished the marker selection by simply sorting genes with a high Er index in one cancer type (Er>=0.9) and low in the others (Er<0.6). As Affymetrix HG-U95A and Hu133A contain different probe-set numbers for the same gene, the SOURCE software[5] from Stanford University was used to match the probe set to their cluster ID using the UniGene Built 167. Cluster ID were then matched between chip types using Microsoft Access. No universal marker encompassing all the cancer vs. their normal tissue was found. This result is compliant with the result from Ramaswamy et al., 2001, using 14 common tumor types including breast, prostate, ovarian and lung cancer. Nonetheless, caveolin-1 (CAV1) was found down regulated at least in 90% of the breast, ovarian, and lung cancer, and in at least 80% of the prostate cancers. Other findings found this gene also down-regulated in large

[5] http://genome-www5.stanford.edu/cgi-bin/source/sourceSearch March 2004

diffuse B-cell lymphoma (Nishiu et al., 2002). CAV-1 is associated with a region of the chromosome 7 q31 frequently deleted in tumors (Fra et al., 1999), and has been shown in many studies to have a tumor suppressing activity when restored (Fiucci et al., 2002; Wiechen et al., 2001).

The number of genes found to be markers varied greatly between cancer types (Table 5.6). Prostate and lung cancer had the smallest number of markers and were the 2 datasets with the most samples. The only prostate marker, SIM2, is a transcription factor located in the nucleus, involved in the following functions according to the Gene Ontology provided by NetAffx™ Analysis Center: regulation of transcription, neurogenesis, embryogenesis and morphogenesis, development, and signal transduction. SIM2 is located on chromosome 21 and may have an influence in Down 's syndrome. This gene has also been found differentially expressed in colon and pancreatic cancer (DeYoung et al., 2003), and antisense inhibition of *SIM2-s* expression in a colon cancer cell line caused inhibition of gene expression, growth inhibition, and apoptosis.

Two genes were found specifically down-regulated in all the lung cancer types compared to other cancer types: AGER and MARCO. The advanced glycosylation end product-specific receptor (AGER or RAGE) has been previously reported to be down-regulated in non-small cell lung cancer (Schenk et al., 2001). AGER is a receptor for amphoterin, highly expressed in the lung, and mediates cell differentiation (Schraml et al., 1997). Down-regulation of AGER may be considered a critical step in lung tumor formation initiating a de-differentiation of the lung cells as it is down regulated in all the different subtypes of lung cancer studied here. On the other hand AGER seems to be up-regulated in pancreatic cancers and its level correlates to the metastatic potential of the cancer cell line

(Takada et al., 2001). The second gene specific to lung cancer is MARCO which is expressed by alveolar macrophages in the lung. Macrophages are involved in inflammation and pathogen clearance in the lung (Bin et al., 2003) (Kraal et al., 2000). The decrease of MARCO RNA in the sample could be due to a decrease of the number of macrophages inside the tumor compared to the normal tissue.

In Ovarian cancer, 39 probe sets were found to have an Er score above 0.9 that were lower than 0.6 in the other cancers. Notably two genes involved in the TGF β signaling pathway, in charge of blocking cell growth, were down-regulated: Janus kinase 1 (JAK1) and a zinc finger homeobox (ZFHX1B). Also found in another study (Schaner et al., 2003), PAX8 involved in development was up regulated. Three other genes involved in the cell growth or maintenance were down-regulated: MLLT2, PRSS11, FOXO3A.

Breast cancer has the most puzzling marker profile, with 16 RNA for ribosomal proteins down-regulated. L34 has been involved in translational control (Moorthamer and Chaudhuri, 1999), S27 in signal transduction, and RPS4X in development and cell cycle control. The down regulation of all these ribosomal proteins could be due, as stated earlier, to methylation at the DNA level. All of the markers for breast cancer are down regulated except for inosine monophosphate dehydrogenase 1 (IMPDH1), up regulated by two fold, which is involved in the biosynthesis of purine nucleotide. Breast cancer has very distinct sub-groups with some cancer being hormone dependant for growth, others being very aggressive with an Her-2 amplification. The cancer samples are probably a mix of these cancer subtypes. This might explain why the well known markers for a particular sub group does not appear in these results. Unfortunately the particular sub-classification of the 16 breast cancer samples is not known (Su et al., 2001).

In oral cancer (Table 5.5), most of the same genes involved in the differentiation of the cell into epithelial cells are found to be markers for this cancer. Keratin 4 and 13 (KRT4 and 13), and the small proline-rich protein 1B (SPPR1B) involved in epidermal differentiation, are all down regulated. Desmoglein 1 and 3 (DSG1 and 3) involved in cell to cell adhesion are also down regulated in a specific manner for oral cancer compared to the other cancer type studied. The matrix metalloproteinase 13 (MMP13) encoding for collagenases, destroying the interstitial tissue was also specifically up regulated in this cancer.

Table 5.6 Gene Markers for Prostate, Lung and Ovarian Cancer that Distinguish between Prostate, Breast, Ovarian, Oral and Lung Cancer
Distinguishing markers are those that are consistently expressed in a given cancer compared to its normal tissue (Er>0.9) but not in any of the other four cancers (Er<0.6). Fold Change (FC)

UniGene ID	Affy probe-set	Gene Name	Description	FC	Er
Prostate					
Hs.27311	39608_at	SIM2	single-minded homolog 2 (Drosophila)	23	0.90
Lung					
Hs.184	35868_at	AGER	advanced glycosylation end product-specific receptor	-25	>0.9
Hs.67726	40331_at	MARCO	macrophage receptor with collagenous structure	-26	>0.9
Ovarian					
Hs.381282	1063_s_at	TYRO3	TYRO3 protein tyrosine kinase	-7	0.91
Hs.308061	121_at	PAX8	paired box gene 8	4	0.95
Hs.496511	1603_g_at	PRKCI	protein kinase C, iota	8	0.95
Hs.32963	1620_at	CDH6	cadherin 6, type 2, K-cadherin (fetal kidney)	21	0.98
Hs.86859	1680_at	GRB7	growth factor receptor-bound protein 7	9	0.95
Hs.153678	31880_at	D8S2298E	reproduction 8	-9	0.91
Hs.288720	32057_at	LRRC17	leucine rich repeat containing 17	-19	0.95
Hs.31653	32215_i_at	RHOBTB3	Rho-related BTB domain containing 3	-12	1
Hs.149900	32779_s_at	ITPR1	inositol 1,4,5-triphosphate receptor, type 1	-6	0.91
Hs.224262	33340_at	PJA2	praja 2, RING-H2 motif containing	-3	0.91
Hs.323079	33910_at	PTPRD	protein tyrosine phosphatase, receptor type, D	-9	0.95
Hs.243987	34241_at	GATA4	GATA binding protein 4	-6	1
Hs.440760	34343_at	STAR	steroidogenic acute regulatory protein	-129	1
Hs.512555	34388_at	COL14A1	collagen, type XIV, alpha 1 (undulin)	-8	1
Hs.173802	34672_at	TBC1D4	TBC1 domain family, member 4	-7	0.96
Hs.14845	34740_at	FOXO3A	forkhead box O3A	-3	0.96

Table 5.6 Cont. Gene Markers for Ovarian Cancer that Distinguish between Prostate, Breast, Oral and Lung Cancer Distinguishing markers are those that are consistently expressed in a given cancer compared to its normal tissue (Er>0.9) but not in any of the other four cancers (Er<0.6). Fold Change (FC)

UniGene ID	Affy probe-set	Gene Name	Description	FC	Er
Ovarian cont.					
Hs.14845	34740_at	FOXO3A	forkhead box O3A	-3	0.96
Hs.48998	34853_at	FLRT2	fibronectin leucine rich transmembrane protein 2	-6	1
Hs.272499	35004_at	DHRS2	dehydrogenase/reductase (SDR family) member 2	-11	0.93
Hs.15780	35390_at	ABCA6	ATP-binding cassette, sub-family A (ABC1), member 6	-20	1
Hs.34871	35681_r_at	ZFHX1B	zinc finger homeobox 1b	-6	0.95
Hs.6454	35756_at	RGS19IP1	regulator of G-protein signalling 19 interacting protein 1	3	0.96
Hs.129673	36234_at	EIF4A1	eukaryotic translation initiation factor 4A, isoform 1	-13	0.93
Hs.75335	36596_r_at	GATM	glycine amidinotransferase	-9	0.96
Hs.308061	36869_at	PAX8	paired box gene 8	47	1
Hs.76798	37205_at	FBXL7	F-box and leucine-rich repeat protein 7	-3	0.93
Hs.78068	37248_at	CPZ	carboxypeptidase Z	-5	1
Hs.458291	38120_at	PKD2	polycystic kidney disease 2 (autosomal dominant)	-4	1
Hs.95243	38317_at	TCEAL1	transcription elongation factor A (SII)-like 1	-8	1
Hs.99824	38364_at	---	Homo sapiens BCE-1	-7	0.91
Hs.372651	38749_at	MGC29643	hypothetical protein MGC29643	24	0.93
Hs.438037	38875_r_at	GREB1	GREB1 protein	-16	1
Hs.114765	39037_at	MLLT2	myeloid/lymphoid translocated to 2	-3	0.91
Hs.172772	39184_at	TCEB2	transcription elongation factor B (SIII)	3	1
Hs.347991	39397_at	NR2F2	nuclear receptor subfamily 2, group F, member 2	-7	1
Hs.438702	39400_at	KIAA1055	KIAA1055 protein	-3	0.91
Hs.117060	39674_r_at	ECM2	extracellular matrix protein 2	-10	1
Hs.40968	41556_s_at	HS3ST1	heparan sulfate (glucosamine) 3-O-sulfotransferase 1	-8	1
Hs.436004	41594_at	JAK1	Janus kinase 1 (a protein tyrosine kinase)	-4	0.93
Hs.75111	718_at	PRSS11	protease, serine, 11 (IGF binding)	-5	1

Table 5.6 Cont. Gene Markers for Breast Cancer that Distinguish between Prostate, Ovarian, Oral and Lung Cancer Distinguishing markers are those that are consistently expressed in a given cancer compared to its normal tissue (Er>0.9) but not in any of the other four cancers (Er<0.6). Fold Change (FC)

UniGene ID	Affy probe-set	Gene Name	Description	FC	Er
Hs.6241	1269_at	PIK3R1	phosphoinositide-3-kinase, regulatory subunit, polypeptide 1	-4	0.95
Hs.410817	31509_at	RPL13	ribosomal protein L13	-3	0.90
Hs.446522	31907_at	RPL14	ribosomal protein L14	-2	0.93
Hs.416566	31952_at	RPL6	ribosomal protein L6	-2	0.94
Hs.356502	31956_f_at	RPLP1	ribosomal protein, large, P1	-2	0.96
Hs.433701	31962_at	RPL37A	ribosomal protein L37a	-3	0.96
Hs.356794	32315_at	RPS24	ribosomal protein S24	-2	0.90
Hs.8102	32438_at	RPS20	ribosomal protein S20	-2	0.93
Hs.381172	32466_at	RPL41	ribosomal protein L41	-3	0.98
Hs.337307	32748_at	RPS27	ribosomal protein S27 (metallopanstimulin 1)	-3	0.94
Hs.437444	33626_at	CACNA1E	calcium channel, voltage-dependent, alpha 1E subunit	-19	0.90
Hs.250895	33657_at	RPL34	ribosomal protein L34	-3	0.98
Hs.469653	33660_at	RPL5	ribosomal protein L5	-3	0.96
Hs.433427	34592_at	RPS17	ribosomal protein S17	-3	0.92
Hs.433427	34593_g_at	RPS17	ribosomal protein S17	-2	0.96
Hs.5662	34608_at	GNB2L1	guanine nucleotide binding protein (G protein)	-3	0.96
Hs.5662	34609_g_at	GNB2L1	guanine nucleotide binding protein (G protein)	-4	0.95
Hs.446628	34643_at	RPS4X	ribosomal protein S4, X-linked	-2	1
Hs.386384	347_s_at	RPS23	ribosomal protein S23	-3	0.94
Hs.288467	34778_at	LRRC15	leucine rich repeat containing 15	5	0.94
Hs.449070	35119_at	RPL13A	ribosomal protein L13a	-3	0.93
Hs.6241	35373_at	PIK3R1	phosphoinositide-3-kinase, regulatory subunit, polypeptide 1	-5	0.95
Hs.90858	35638_at	CBFA2T1	core-binding factor, runt domain; cyclin D-related	-3	0.94
Hs.438	36010_at	MEOX1	mesenchyme homeo box 1	-21	0.94
Hs.439109	38280_s_at	---	cDNA clone EUROIMAGE 1630957	-26	0.92
Hs.308053	38737_at	IGF1	insulin-like growth factor 1 (somatomedin C)	-12	0.98
Hs.22500	39222_at	PIK3C2G	phosphoinositide-3-kinase, class 2, gamma polypeptide	-10	0.92
Hs.32916	39739_at	NACA	nascent-polypeptide-associated complex alpha polypeptide	-2	0.93
Hs.32916	39740_g_at	NACA	nascent-polypeptide-associated complex alpha polypeptide	-2	0.98
Hs.306382	40239_g_at	MGC35048	hypothetical protein MGC35048	-3	0.93
Hs.443518	40304_at	BPAG1	bullous pemphigoid antigen 1, 230/240kDa	-18	1
Hs.317095	40695_at	IMPDH1	IMP (inosine monophosphate) dehydrogenase 1	2	0.90
Hs.23719	41124_r_at	ENPP2	ectonucleotide pyrophosphatase/phosphodiesterase 2	-21	0.96

Table 5.6 Cont. Gene Markers for Oral Cancer that Distinguish between Prostate, Ovarian, Breast and Lung Cancer Distinguishing markers are those that are consistently expressed in a given cancer compared to its normal tissue (Er>0.9) but not in any of the other four cancers (Er<0.6). Fold Change (FC)

UniGene ID	Affy probe-set	Gene Name	Description	FC	Er
Hs.324470	201753_s_at	ADD3	adducin 3 (gamma)	6	1
Hs.512628	202313_at	PPP2R2A	protein phosphatase 2, regulatory subunit B (PR 52)	-3	0.91
Hs.3068	202983_at	SMARCA3	SWI/SNF related, actin dependent regulator of chromatin	4	0.99
Hs.287721	204415_at	G1P3	interferon, alpha-inducible protein	15	0.96
Hs.75716	204614_at	SERPINB2	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin)	-34	0.96
Hs.1690	205014_at	HBP17	heparin-binding growth factor binding protein	-10	0.93
Hs.1076	205064_at	SPRR1B	small proline-rich protein 1B (cornifin)	-16	0.94
Hs.1925	205595_at	DSG3	desmoglein 3 (pemphigus vulgaris antigen)	-4	0.93
Hs.75219	205694_at	TYRP1	tyrosinase-related protein 1	-7	0.97
Hs.115263	205767_at	EREG	epiregulin	-51	0.94
Hs.2936	205959_at	MMP13	matrix metalloproteinase 13 (collagenase 3)	122	0.93
Hs.2022	206004_at	TGM3	transglutaminase 3	-56	1
Hs.105115	206513_at	AIM2	absent in melanoma 2	6	1
Hs.997	206605_at	P11	26 serine protease	-140	1
Hs.2633	206642_at	DSG1	desmoglein 1	-59	0.94
Hs.1200	207206_s_at	ALOX12	arachidonate 12-lipoxygenase	-17	0.94
Hs.185726	207332_s_at	TFRC	transferrin receptor (p90, CD71)	3	0.91
Hs.433871	207935_s_at	KRT13	keratin 13	-65	0.94
Hs.511872	208126_s_at	CYP2C18	cytochrome P450	-22	1
Hs.436986	209682_at	CBLB	Cas-Br-M	5	0.93
Hs.49500	212314_at	KIAA0746	KIAA0746 protein	12	1
Hs.420584	212717_at	KIAA0356	KIAA0356 gene product	-3	0.91
Hs.115176	213135_at	TIAM1	T-cell lymphoma invasion and metastasis 1	-6	0.97
Hs.371139	213240_s_at	KRT4	keratin 4	-559	0.91
Hs.511963	213294_at	FLJ38348	hypothetical protein FLJ38348	4	0.99
Hs.371139	214399_s_at	KRT4	keratin 4	-18	0.91
Hs.511872	215103_at	CYP2C18	cytochrome P450, family 2, subfamily C, polypeptide 18	-32	0.97
Hs.154103	216804_s_at	LIM	LIM protein	-4	0.96
Hs.384944	216841_s_at	SOD2	superoxide dismutase 2, mitochondrial	5	0.91
Hs.237028	219789_at	NPR3	natriuretic peptide receptor C/guanylate cyclase C	-8	0.94
Hs.418127	220017_x_at	CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9	-6	0.94

5.7 Discussion

The method described here extends previous non-parametric approaches to microarray data analysis. After applying the noise boundary model described in Chapter 4, markers were selected according to their consistency for up-regulation or down-regulation using a voting scheme when comparing normal versus cancer biopsies. Genes, differentially expressed in a tissue-specific manner, were eliminated by comparing the cancer samples to the normal biopsies from the same tissue. The genes which were most differentially regulated between cancer and normal biopsies were then compared among different cancer types. Using this method, genes that were tissue specific were eliminated unlike previous studies comparing different cancers (Ramaswamy et al., 2001; Su et al., 2001). Markers with consistent differential expression in ovarian, breast, prostate and lung cancer were found. Among those markers, a high number of them were related to the de-differentiation of the tissue, and were highly specific to their tissue of origin. Cancer arising from cells with the same embryogenic origin, i.e. differentiated at the same time by the same processes, tends to have the same genes involved in de-differentiation needed for cancer. This reflects an oncodevelopmental connection described before (Taipale and Beachy, 2001). With the increasing growth in microarray data publications, this method mines the already performed experiments and finds new information and helps make relevant observations. Most of the genes found as markers were confirmed in the literature. An interesting finding and not widely reported previously was the down regulation of RNA ribosomal proteins in breast cancer, the significance of which is worth further investigation.

CHAPTER 6

DISCUSSION AND CONCLUDING REMARKS

Cancer, the second leading cause of death behind heart disease, accounts for 23% of the deaths in the United States annually (Jemal et al., 2003; Simmonds, 2003). Cancer is characterized by an uncontrolled proliferation of cells. Almost all organs and cell types can undergo an oncogenic transformation, with an array of mechanisms and outcomes. Cancer arises through a succession of events including: chromosomal deletion, promoter methylation, gene fusion, amplification, mutation, alternative splicing or deregulation of expression. The complexity and diversity of the regulatory and downstream effector pathways affected by cancer has hindered the development of effective and specific therapies. During oncogenesis, RNA expression levels of numerous genes are affected and the normal cellular equilibrium is disturbed within the surrounding background processes and stochastic events. The cell, as a result, obtains another state of equilibrium in which it becomes "immortal", becoming partly independent of its environment for survival and proliferation. A major problem in cancer research has been the ability to study only a few genes at a time in a pathway. These genes were also influence by others omitted in the study and could lead to different treatments outcomes in different cells. Thus, the knowledge of pathways and cross-talk between pathways has been too incomplete to omit the other genes from the analysis. To be able to comprehend cancer, a global approach at the genome expression level must be taken. Microarray technology is particularly well suited for this holistic approach as it analyzes the expression of thousands of genes at the same time.

Microarray-based gene expression profiling provides a robust signature of the molecular phenotype of the cell in a specific state. This technology is a powerful tool for determining the changes at the mRNA levels in cancer, allowing the analysis of thousands of genes in a single experiment. Although changes associated with cancer can involve chromosomal deletion, amplification, mutation, alternative splicing or deregulation of expression, gene expression profiles can give an indication of the state of the cell. These profiles are very valuable in finding the common traits pertaining to different cancers arising from a multitude of cellular events with heterogeneous genetic changes. This high through-put technology estimates the expression level of thousands of genes in a single experiment in less than a week. It is thus, so far, considered to be the best technology to find gene expression markers in cancer.

Microarray technology was first developed to analyze gene expression of a complex population of RNA (DeRisi et al., 1997; Lashkari et al., 1997; Lipshutz et al., 1999; Schena et al., 1995). A refinement of this method allows the analysis of copy number imbalances, gene amplification or deletion at the DNA level (Pollack et al., 1999) and deletions or small insertion in tumor suppressor genes (Frolov et al., 2002). This technology has been applied in a similar way for a systematic analysis of protein levels in the cells (Haab, 2001). Proteins perform most of the functions in the cell and make up for the majority of the cellular structures. Protein microarrays have a great potential as they can be used for protein profiling and high-throughput function determination. Profiling determinates the abundance, modification, localization, activity, and interaction of proteins in a given cell or tissue. Function determination teases apart the possible interactions and binding of one protein with other proteins to construct a network of

interactions. Their application is very powerful for protein function studies, screening the production of antibodies and recombinant proteins, discovery of proteins implicated in disease, potential drug targets and rapid detection or diagnosis of disease. One of the major drawbacks in the production of such arrays is the need for sets of cloned genes that can be used for high-throughput expression and purification of recombinant proteins, or access to large sets of purified proteins. The human genome is estimated to contain 30,000 to 40,000 genes, and the proteome is estimated to be at least three times larger (Harrison et al., 2002). The protein interactions are also more complex, happening in a three dimensional space whereas DNA-DNA hybridization is a two dimensional process. Also binding the protein to a support (i.e. a slide) might change its properties. Another problem is the detection range: proteins concentration varies greatly from the huge amount of cell scaffolding proteins to the small regulatory proteins that are present in very low concentration. For all these reasons, the development of protein microarrays is more complex and they are not yet widely available.

Current cancer classification techniques and treatment decisions rely on subjective judgments of tumor histology by pathologists. Cancer cells that seem identical from a histological point of view can respond differently to therapy and may evolve very differently from indolent tumors to invasive metastatic tumors. Tumors with very similar histology can be differentiated with their expression profile. Application of a variety of data analysis techniques (hierarchical clustering, neural networks, self organizing maps) on analysis of global gene expression allowed the classification of tumors that are difficult to separate by conventional histopathological microscopic examination (Golub et al., 1999; Khan et al., 2001; van 't Veer et al., 2002). Microarray technology has proven

useful in the classification of tumors. Also new tumor sub-types can be discovered, ie lung adenocarcinoma was found to consist of four distinct sub classes (Bhattacharjee et al., 2001). Classification of morphologically similar human cancer can help tailor treatments, maximizing the therapeutic effect and minimizing the toxicity (O'Neill et al., 2003; van 't Veer and De Jong, 2002). The patient treatment decisions can be tailored to the properties of the expression profile of their tumor instead of the morphology. One example is adjusting chemotherapy regimens based on tumor profiles. Patients presenting aggressive breast tumors with a high risk of metastasis and poor prognosis would benefit from an aggressive chemotherapy regimen whereas most patients 70-80% would survive without it (van 't Veer et al., 2002). A database repository could be created to store patients tumor gene expression profiles information anonymously before, during and after chemotherapy. A newly diagnosed patient's tumor profiles could then be matched to those in the database and treatment options could be selected depending on the success and failure of treated patients with the same profile. The side effects of treatments will be decreased and their efficacy will increase, as the right treatment would be administered sooner. With the help of the Cancer Biomedical Informatics Grid, caBIG[6], from the US National Cancer Institute, this scenario could quickly materialize. This integrative biomedical informatics infrastructure would be an extensible informatics platform that integrates diverse data types and supports interoperable analytic tools. The goal is to allow the collection of data from different centers in a unifying architecture to support the desired interoperability. This repository includes the gene expression profiles of normal, precancerous, and cancer cells from the Cancer Genome Anatomy Project

[6] http://cabig.nci.nih.gov/ March 2004

(CGAP). This will allow researchers to mine a collection of data and further the understanding and treatment of cancer.

Microarray technology is, without a doubt, revolutionizing the study of cancer. Molecular profiles of tumors help elucidate cancer development and the pathways involved in detail. With gene silencing by RNA interference, one can selectively knock down genes and profile the results on microarrays. Pathways and gene function can then be inferred from the genes expression profiles (Berns et al., 2004; Paddison et al., 2004). Another way microarray data can help treatment of cancer is through the screening of new drugs. It should be possible to observe the effect of a drug on gene expression profiles and create a model to predict the expected therapeutic response (Hughes et al., 2000). Finding a drug that acts on a certain gene may be as simple as finding a drug with the same gene expression profiles as the gene expression profiles obtained with the knocked out gene. Many of the current drugs have a broad spectrum of action disturbing many mechanisms in the cell. This technology will guide drug development by helping to define the targets through a better understanding of the targeted biological processes, thereby speeding up the development and screening of the drugs. Drugs side effects can be estimated even before clinical trial as the effects of the drug is analyzed on the whole cell transcriptome, not just on the target. This will make the drugs developed in this manner safer.

The study of cancer has been for the most part removed from the cell environment, ignoring the interactions of cancer cells with their surroundings. The local environment and cell to cell interactions are key factors in tumor genesis. Some tumor cells have an increased number of receptors for cell growth signaling factors. Blocking

these receptors, as in the hormonal responsive breast cancer (Shenkier et al., 2004), or modifying the surrounding environment of the cell can help stop the proliferation of the tumor. In the same manner, used for drug discovery, a network of gene interaction is needed to be able to find which steps in the pathway are the best targets for disruption and induction of cell death. This task will require a large amount of data, from different cell types and conditions, and some hurdles still have to be overcome before this happens.

The use of DNA microarrays has become common in biological research, however, there is still room for improvement of the technology. Current microarray techniques can identify expressed genes at five or higher copies of mRNA per cell. This can become a problem as studies in yeast showed that the dynamic range of mRNA production could be over six orders of magnitude and that most of the genes were expressed (75%) on average at less than one copy per cell (Gygi et al., 1999; Holland, 2002). Transcription factors, critical for regulation of some genes, may be present at only one to two copies per cell under induced conditions. The genes expressed at higher levels are usually associated with the homeostasis of the cell, i.e. metabolism, protein synthesis, cytoskeleton. The most decisive changes in mRNA expression for cancer might happen at low expression levels below the resolution of the current technology. Thus existing microarrays technology may only see the downstream effects of these low expressed transcription factors.

Cancer cells present an accumulation of replication disorders arising through gradual accumulation of genetic changes. Typical cancer cells contain combinations of genetic changes that alter gene expression causing the cell to escape the checks and controls that prevent proliferation and metastasis. A single mutation event is generally

not enough to circumvent all the different safeguards a complex organism has built in. Instead, two or more sequential events are needed to initiate the transformation of a normal to a malignant cell (Vogelstein and Kinzler, 1993). During proliferation, mutations, deletions and amplifications of chromosomes occur due to genomic instabilities. Though gene expression profile correlate to some degree to the changes occurring at the chromosome level (Ulger et al., 2003), other factors such as promoter methylation influence this correlation. The tumor tissue is a mixture of heterogeneous cell types from malignant cells at different stages of differentiation, normal epithelial cells, blood vessels and cells involved in the inflammatory process. There is a dilution of the gene expressions changes seen when the tumor biopsy comport too many non cancerous cells. Another difficulty is studying the development of cancer, i.e. finding the very early stages of cancer. Many people carry *in situ* tumors, that are very small and do not develop into disease (Folkman and Kalluri, 2004). In order to proliferate, tumors need to recruit their own blood supply through angiogenesis (Hanahan and Folkman, 1996). Diagnostics are usually made when the tumors have reached a certain size and genomic instability is often already taking place. It is important to try to find which early mutation produces the switch from indolent to invasive tumor in order to prevent cancer occurrence and produce early detection methods. For this, laser capture micro-dissection is a very useful instrument. It helps isolate the cancer cells from the surrounding cells, provided that the operator is able to distinguish between them. However, this technique yields very little mRNA from a few thousand cells. Amplification techniques of the resulting mRNA often present a bias due to incomplete synthesis of the mRNA and non linear increase of the lowest expressed mRNA. The sensitivity of microarray detection

needs to be improved to be able to use less RNA, possibly only a few selected cells obtained from laser capture.

Another major concern is the lack of correlation in gene expression levels and significant changes between the different microarrays techniques and the protein levels changes. Results from the Affymetrix GeneChips and the diverse spotted arrays techniques have been reported to have very low correlation (Barczak et al., 2003; Tan et al., 2003). It not known which method is right or wrong, or if there is a right and wrong. The differences might be due, in part, to differences in the cDNA regions probed on the array, their differential affinity to splice variants, and cross hybridization that might differ depending on the length of the oligonucleotides used. Validating the results from microarrays is troublesome. It is hardly feasible to validate 20,000 gene expression levels using real time PCR, each validation requiring design of probes and optimization of the reaction. The other issue is that the level of correlation between mRNA levels and protein levels is variable depending on the studies (Griffin et al., 2002; Gygi et al., 1999; Ideker et al., 2001). Large changes in mRNA levels are often paired with much smaller changes at the protein level (5 to 40 fold less), and small changes in mRNA levels sometimes do not correlate at all with changes in protein amount. Given this, no direct inference can be made for the gene expression level to protein levels in the cells. However, the fact that the microarray data can separate cancer samples in relevant sub-classes is enough to prove their usefulness.

The value of expression profiling is still underestimated due to the infancy of the computational tools necessary to analyze large datasets, and the inexperience of modeling systems with such large amount of biological data. Unsupervised classification is

difficult with results depending on the distance metric used and with different metrics working better for different datasets. Successful studies require complex analysis as well as a substantive knowledge. Statistical methods do not provide all the answers and an expert must still analyze the results for biological significance. With the creation of database repositories[7] and annotation standards[8] for sharing and publishing microarray results, there is a great deal of information that is yet unexplored. The most significant improvement provided by databases, beyond just a storage function, is to link different type of data: gene expression profiling with microarrays and Serial Analysis of Gene Expression (SAGE) analysis, results from analysis (i.e. cluster assignment for a gene), functional annotation, metabolic pathway function, chromosomal location, presence of known promoter elements and samples origin and description. A study of this information is certainly going to help produce significant advances in our understanding of biology in the near future. An analysis of this kind will require a major computational and modeling effort. There is still a long way to go before this is accomplished and the research presented here is one step in the process of building the foundation for this "New Biology" at the genomic scale.

The principal aim of this dissertation was to find cancer biomarkers. Gene expression markers, by definition, are genes that are consistently up-regulated or down-regulated in cancer samples compared to normal samples. The hypothesis is that consistency is the sign that the genes in question are part of downstream regulatory pathways necessary for tumorgenesis. Gene markers are obviously important for early diagnosis and can help define the therapeutic course of action. For example, Gleevec™

---

[7] http://cabig.nci.nih.gov/ March 2004
[8] http://www.mged.org/ March 2004

is very effective against the chronic myeloid leukemia with the "Philadelphia" chromosome translocation but ineffective against the other types of leukemia (Kaelin, 2004; van der Kuip et al., 2004). The probability of relapse can be predicted and a more aggressive treatment can be administered (van 't Veer and De Jong, 2002). Biomarkers can become targets for drug development and can help predict treatment response. A detailed study of their function might explain the events leading to their de-regulation and help to design chemo-prevention therapies.

In order to be able to find cancer biomarkers, it is necessary to comprehend the qualitative and quantitative analysis of microarray data. In this dissertation GeneChip array data was used and common methods of analysis were reviewed first. There are many pre-processing steps with different methods to be performed before analysis of sets of chips. The signal intensity has to be estimated first for every probe-set on the chip, and this step alone can be performed using 4 different methods: MAS 4 and MAS 5 (Affymetrix, 2000; Affymetrix, 2002a; Affymetrix, 2002b), the Model Based Expression Indexes (MBEI) with the dChip software (Li and Wong, 2001a; Li and Wong, 2001b), and the Robust Multi-chip Analysis (RMA) (Irizarry et al., 2003). Then, chips intensities need to be normalized to allow comparison between them. The most common methods for normalizing microarrays are global scaling of the average intensity of a chip to a set target intensity, normalizing using a set of house keeping genes whose intensity does not change, normalizing to a set of invariant genes, lowess normalization (Quackenbush, 2002) and quantile normalization (Irizarry et al., 2003). The effects of these two first pre-processing steps have a major effect on the results. The genes selected as significant are different depending on the methods used for signal extraction and normalization (Tan et

al., 2003). This poses a serious problem for selecting the right combination of methods used. Also, while comparing data from the literature, one has to be very careful in interpreting the results where the pre-processing steps were not the same.

The major problem with microarray data analysis is that standard statistical tests are ill adapted to the data. Standard analysis, such as the widely used t-test like statistics, produces too many false positives. Due to the very high number of tests performed, applying correction for multiple testing renders the result too conservative. Also, the tests and corrections for multiple testing assume that the tests are independent. Gene expression profiles usually present clusters of highly correlated genes. Different approaches for analysis of microarray data were developed and presented in Chapters 3, 4 and 5.

In Chapter 3, a nonparametric method is used to separate cancer samples according to their selectively expressed genes. This method disregards the normalization techniques that influence the results by changing/correcting signal intensities. Selectively expressed genes were correlated to different types of acute leukemia and their cell of origin. This was the first study of this kind, which showed the possibility of classify human diseases using selectively expressed gene data from microarrays. The classification using this method was better than in the original study (Golub et al., 1999) where a t-test like statistic was used. Selective genes can serve as very useful markers when the variation in expression levels is not known in the population, and setting expression thresholds is difficult. They might also correlate better with a change in protein levels in the cell. Selectively expressed genes make for a more accurate diagnosis giving less ambiguous results. They can also be used as a simplifying assumption for the

discovery of gene regulatory networks using Boolean probabilities (Shmulevich et al., 2002). This work was performed using the MAS4 call algorithm. Since then, MAS5 and dChip were released with new algorithms to make decisions on the presence or absence of a gene. Future research directions would be to study the influence of the call algorithm that decide which gene is considered present or absent and their performance in separating cancer subtypes.

In Chapter 4, the development of a noise boundary model is described that eliminates spurious fold-changes and reduces the number of false positives in further analysis. This section of the dissertation analyzed the noise of quantitative data by examining technical and biological replicates. An inverse correlation was found between the noise and the expression levels for all the algorithms considered (MAS4, MAS5, dChip, RMA) of probe set intensity extraction. Low levels of estimated transcript expression are not as reliable as high expression levels in inferring fold changes. This noise was consistent from one comparison of biological replicates to another and therefore could be modeled. Two parameters were the most influential in the modeling: the percentile of the fold-changes chosen for the noise boundary and the low intensity cutoff. This noise model was tested on a standard dataset, the Affymetrix Latin square replicate data set (Hubbell et al., 2002), to see how well it eliminated noise from the data. Percentile and the low intensity cutoff parameters were set with a compromise between eliminating a large amount of the false positives (type I error) and finding most the true positives (type II error). The noise was present for the different tissue types; it was modeled for each with the same method as the Latin square replicate data set, but was found to be tissue and/or lab specific. This noise boundary model was designed to set a

threshold for fold change direction trust, and filter out most fold changes that would occur randomly as a result of background noise. Plans are to study the noise in the spotted array microarray platform in a similar manner and develop a noise boundary model.

The last Chapter of this dissertation introduced a new algorithm (Er algorithm) to select consistently up-regulated or down-regulated genes and set up a methodology to compare data issued from different research groups.

The directional change assessment algorithm (Er algorithm) uses an unweighted voting scheme to select transcripts exhibiting consistent fold changes between samples. This algorithm was tested using the Latin square replicate data set provided by Affymetrix (Hubbell et al., 2002), and the results were similar to other techniques known to reduce false positives. No technique was perfect; eliminating false positives from the results always reduced the number of true positives found. The Bonferroni correction for multiple testing (Bonferroni, 1936) for example was found to be too conservative eliminating all but one false positive but only finding 6 out of 14 true positives. Most of the genes were spiked at a two fold change difference from one set chips to the other. This averaged after hybridization to an estimated fold change of 1.5 with most of the probe set intensity estimation techniques. This low fold change is often within the background noise. The other criticism with using this data set for determining the best method is that the Latin square replicate data set has very little noise, the chips can be considered technical replicates as only 14 genes out of 12,000 are spiked. The method described in this chapter was specifically designed to reduce the influence of the noise in

the result. It is not possible so far to draw a conclusion as to which method is better on standard data with noise.

Cancer bio-markers were selected for prostate, breast, ovarian, lung and oral cancer. Genes expressed in a tissue-specific manner (i.e. expressed in breast but not in prostate) were eliminated by comparing the cancer samples to normal biopsies from the same tissue unlike previous studies comparing different cancers (Ramaswamy et al., 2001; Su et al., 2001). The genes which were most differentially regulated between cancer and normal biopsies were then compared among different cancer types. Markers with consistent differential expression in ovarian, breast, prostate and lung cancer were found. Among those markers, a high number of them were related to the differentiation of the tissue, and were highly specific to their tissue of origin. Cancer arising from cells with the same embryogenic origin, i.e. differentiated at the same time by the same processes, tends to have the same genes involved as cancer cells are usually incompletely differentiated. This reflects an oncodevelopmental connection described before (Taipale and Beachy, 2001). Some gene markers reviewed were known markers and/or have a biological function that might take part in the oncogenic process. Other gene markers found had no obvious or known connection to the oncogenic process. These genes could be co-regulated or downstream from a pathway affected in cancer, and they can be very useful for diagnostic and classification. There are many ways to regulate proteins concentration in a cell: transcription, RNA degradation, alternative splicing, translation of the RNA into proteins, and degradation of proteins. The regulation of a protein might not always be seen at the RNA level, but its effect can be seen on the RNA levels of the co-regulated or downstream genes. With the increasing growth in papers presenting

microarray data and databases to exchange data, this analysis provides a methodology for mining the data in experiments already performed, finds new information and helps make relevant observations.

In human cancer, the use of the microarray technology is starting to provide key insights in tumorgenesis, cancer progression and response to therapies. With the availability of the sequence of the human genome, one can now look at the complete transcriptome of normal and cancer cells. This technology is going to change the way cancer will be detected and treated in the future. The novel methods described in this dissertation are ground-breaking in their approach to data analysis. Not only are they robust and parsimonious, but they improved discrimination between cancer subtypes with similar histology and they selected robust cancer biomarkers. These tools for diagnosis can in turn be translated into better patient treatments with therapies tailored to their specific cancer profiles.

APPENDIX A

FOLD CHANGE ESTIMATION WITH MAS5, DCHIP AND RMA FOR SPIKED GENES IN THE LATIN SQUARE DATA SET

Using the Latin Square data set, the fold change of the spiked genes was estimated with the 3 most common methods (MAS5, MBEI/dChip PM only and RMA).

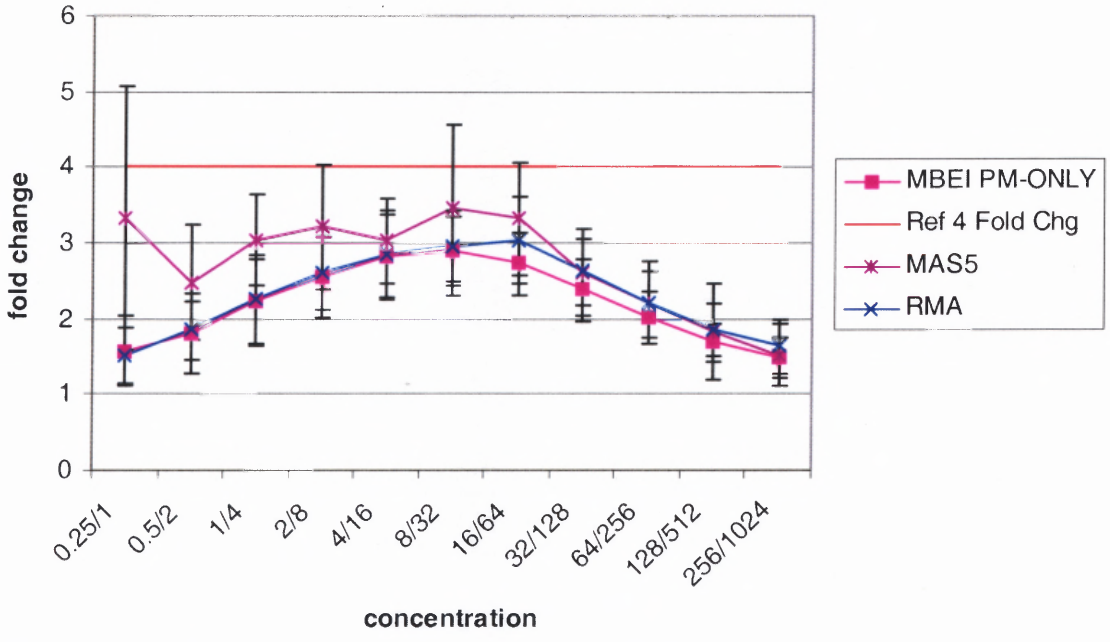


Figure A.1 Average fold change for all the probe-sets in function of their spiked in concentration. This data was obtained from the Latin square data set. The genes were spiked in at a four-fold concentration difference. Overall, all the methods underestimate the fold change, especially MBEI Pm only and RMA. MAS5 has higher standard deviation in its evaluation of the fold change than MBEI Pm only and RMA.
(Figure courtesy of Jeff Cheng, 2003, unpublished results).

APPENDIX B

EFFECT OF THE CUTOFF VALUE AND PERCENTILE ON THE INTERCEPTS

This appendix presents the effect of the minimum intensity cutoff and the percentile chosen on the intercepts of the modeled noise boundary for the five different tissue types.

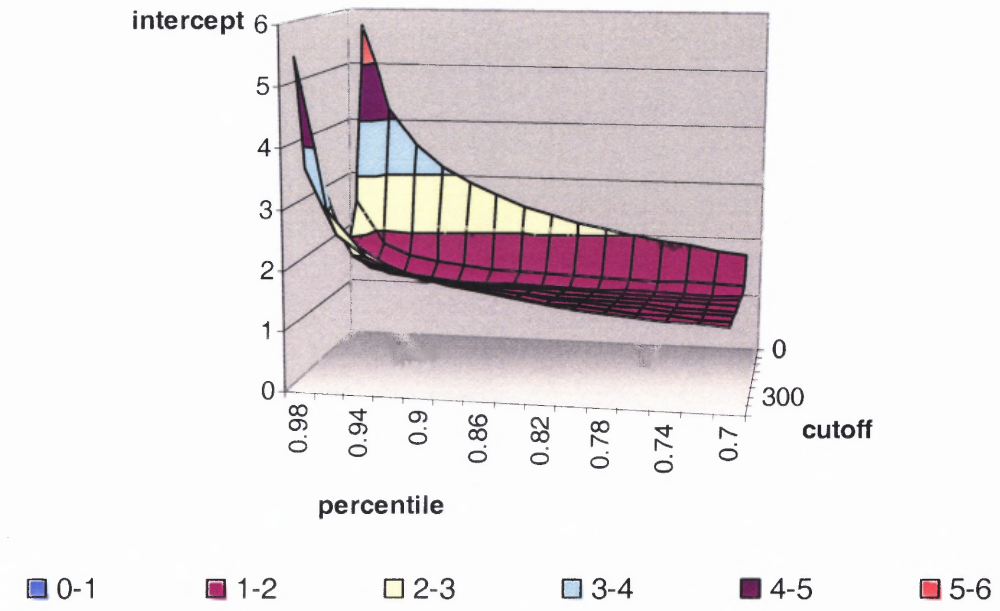


Figure B.1 Three dimensional graph on the effect of the minimum intensity cutoff and percentile on the intercept of the regressed percentile to the average intensity of the bins, for the normal breast tissue data obtained using MAS5 (Affymetrix).

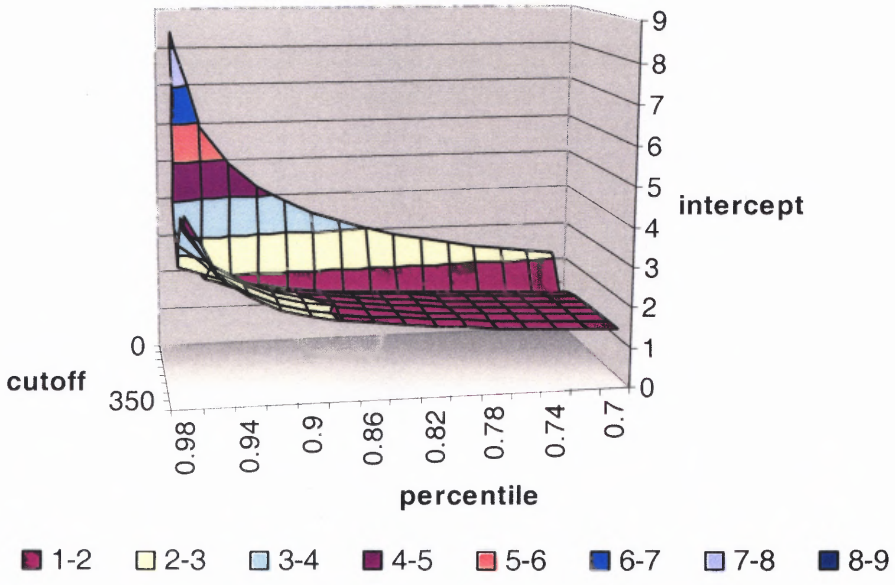


Figure B.2 Three dimensional graph on the effect of the minimum intensity cutoff and percentile on the intercept of the regressed percentile to the average intensity of the bins, for the normal oral epithelium data obtained using MAS5 (Affymetrix).

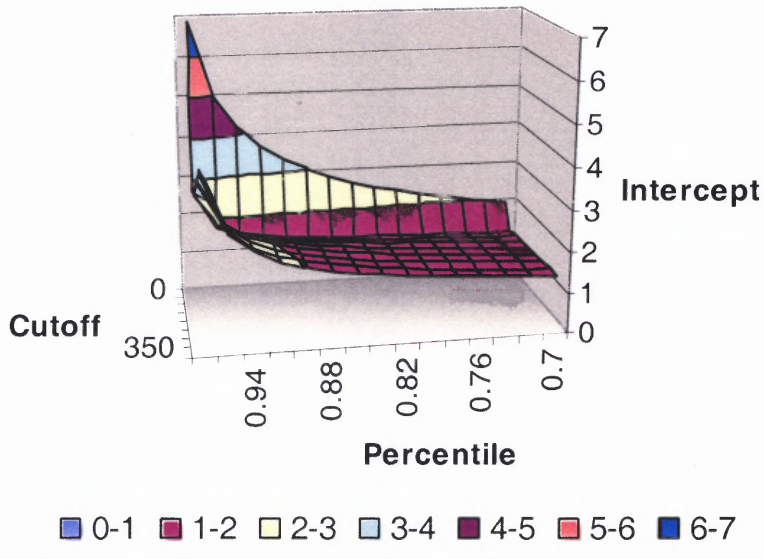


Figure B.3 Three dimensional graph on the effect of the minimum intensity cutoff and percentile on the intercept of the regressed percentile to the average intensity of the bins, for the normal lung tissue data obtained using MAS5 (Affymetrix).

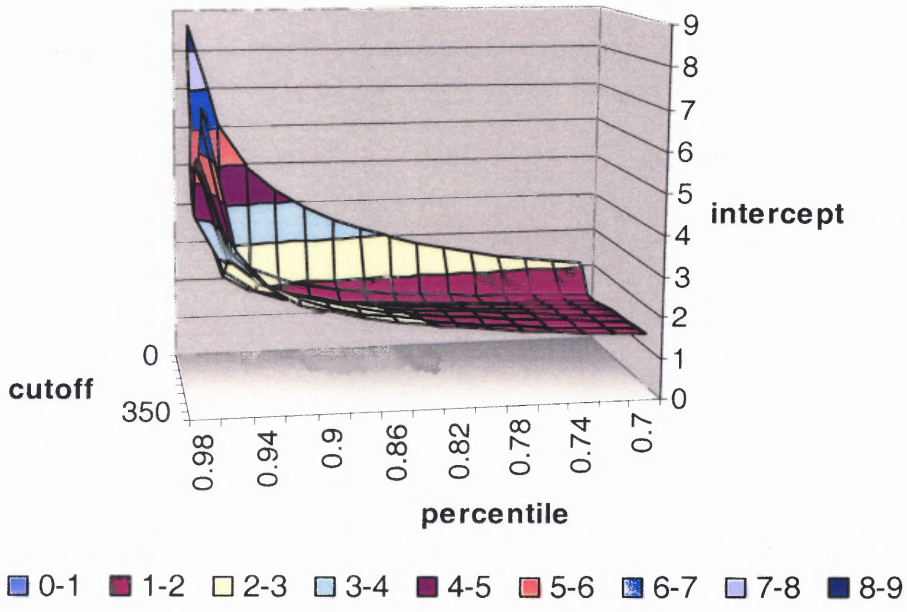


Figure B.4 Three dimensional graph on the effect of the minimum intensity cutoff and percentile on the intercept of the regressed percentile to the average intensity of the bins, for the normal ovarian tissue data obtained using MAS5 (Affymetrix).

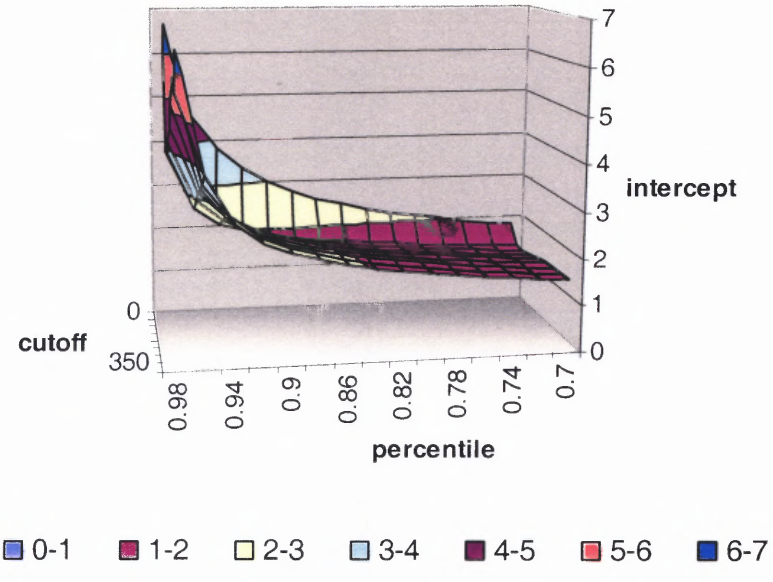


Figure B.5 Three dimensional graph on the effect of the minimum intensity cutoff and percentile on the intercept of the regressed percentile to the average intensity of the bins, for the normal prostate tissue data obtained using MAS5 (Affymetrix).

APPENDIX C
SHUFFLED RESULTS FOR THE ER ALGORITHM

This appendix presents the comparison of the top 500 Er scores for ovarian, lung, prostate and oral cancer compared to Er scores obtained with shuffled cancer and normal samples.

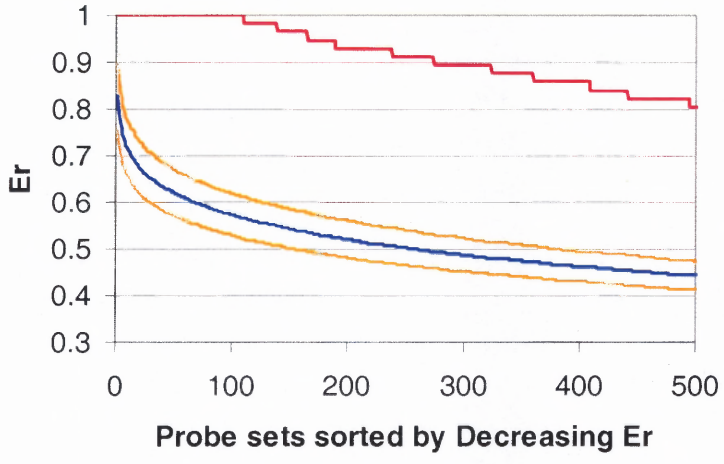


Figure C.1 Comparison of the Er score of the 500 top ranked probe sets for ovarian cancer versus normal biopsies. Er score for the real Ovarian cancer vs. normal biopsies, average Er score of the 500 top ranked probe sets of the 100 shuffling sets , one standard deviation away form the average shuffled sets .

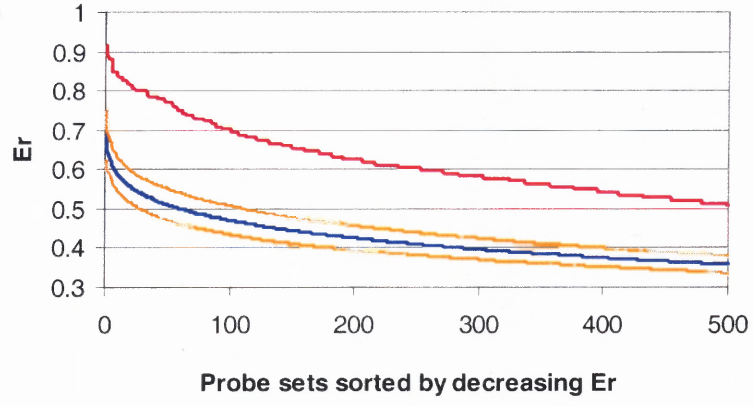


Figure C.2 Comparison of the Er score of the 500 top ranked probe sets for prostate cancer versus normal biopsies. Er score for the real prostate cancer vs. normal biopsies comparisons, average Er score of the 500 top ranked probe sets of the 100 shuffling sets one standard deviation away form the average shuffled sets .

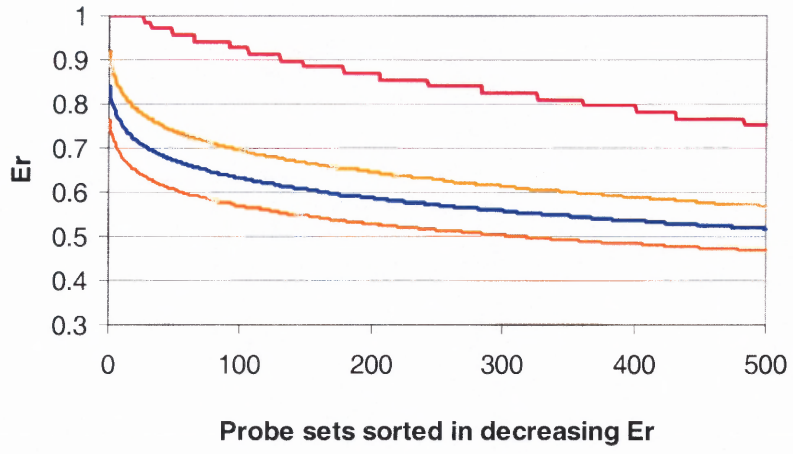


Figure C.3 Comparison of the Er score of the 500 top ranked probe sets for oral cancer versus normal biopsies. Er score for the real oral cancer vs. normal biopsies comparisons, average Er score of the 500 top ranked probe sets of the 100 shuffling sets and one standard deviation away form the average shuffled sets .

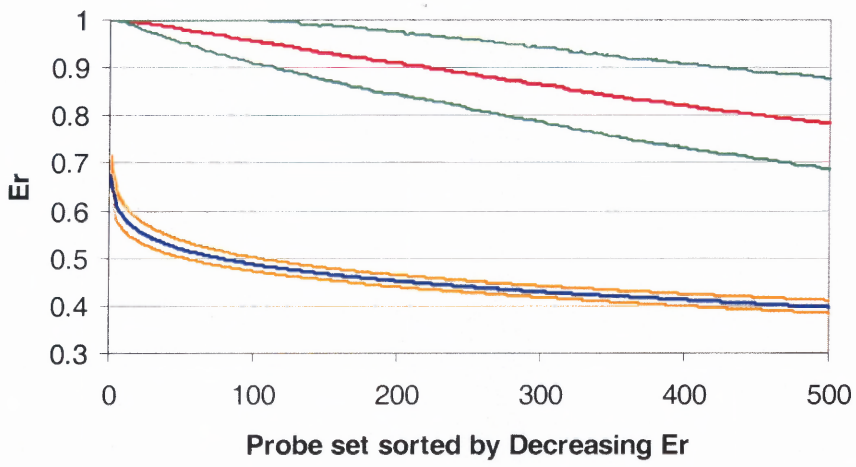


Figure C.4 Comparison of the Er score of the Averaged 500 Top Ranked Probe Sets for Lung Cancers versus Normal Biopsies. Er score for the averaged real lung cancer vs. normal biopsies, standard deviation of this average , average Er score of the 500 top ranked probe sets of the 100 shuffling sets and standard deviation form the average shuffled sets .

REFERENCES

Adib, T. R., Henderson, S., Perrett, C., Hewitt, D., Bourmpoulia, D., Lederman, J., and Boshoff, C. (2004). Predicting biomarkers for ovarian cancer using gene-expression microarrays. Br J Cancer *90*, 686-692.

Affymetrix (2000). Affymetrix Expression Analysis Algorithm Tutorial. Affymetrix Microarrray Suite 40 User Guide Affymetrix, Santa Clara, 295-316.

Affymetrix (2001). Affymetrix Microarray Suite 5.0 User's Guide. Affymetrix, Santa Clara.

Affymetrix (2002a). New statistical algorithms for monitoring gene expression on GeneChip probe arrays. In Affymetrix Technical Note Affymetrix, Santa Clara.

Affymetrix (2002b). Statistical Algorithms Description Document. Affymetrix, Santa Clara *http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf*, March 2004.

Auersperg, N., Pan, J., Grove, B. D., Peterson, T., Fisher, J., Maines-Bandiera, S., Somasiri, A., and Roskelley, C. D. (1999). E-cadherin induces mesenchymal-to-epithelial transition in human ovarian surface epithelium. PNAS *96*, 6249-6254.

Bachem, C. W., van der Hoeven, R. S., de Bruijn, S. M., Vreugdenhil, D., Zabeau, M., and Visser, R. G. (1996). Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. Plant J *9*, 745-753.

Baggerly, K. A., Coombes, K. R., Hess, K. R., Stivers, D. N., Abruzzo, L. V., and Zhang, W. (2001). Identifying differentially expressed genes in cDNA microarray experiments. J Comput Biol *8*, 639-659.

Barczak, A., Rodriguez, M. W., Hanspers, K., Koth, L. L., Tai, Y. C., Bolstad, B. M., Speed, T. P., and Erle, D. J. (2003). Spotted Long Oligonucleotide Arrays for Human Gene Expression Analysis. Genome Res *13*, 1775-1785.

Benjamini, X., and Hochberg, X. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B *57*, 289 - 300.

Bernard, P. S., and Wittwer, C. T. (2002). Real-Time PCR Technology for Cancer Diagnostics. Clin Chem *48*, 1178-1185.

Bernell, Jacobsson, Liliemark, Hjalmar, Arvidsson, and Hast (1998). Gain of chromosome 7 marks the progression from indolent to aggressive follicle centre lymphoma and is a common finding in patients with diffuse large B-cell lymphoma: a study by FISH. Br J Haematol *101*, 487-491.

Berns, K., Hijmans, E. M., Mullenders, J., Brummelkamp, T. R., Velds, A., Heimerikx, M., Kerkhoven, R. M., Madiredjo, M., Nijkamp, W., Weigelt, B., *et al.* (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature *428*, 431-437.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., *et al.* (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A *98*, 13790-13795.

Bin, L.-H., Nielson, L. D., Liu, X., Mason, R. J., and Shu, H.-B. (2003). Identification of Uteroglobin-Related Protein 1 and Macrophage Scavenger Receptor with Collagenous Structure as a Lung-Specific Ligand-Receptor Pair. J Immunol *171*, 924-930.

Black, W. C., and Welch, H. G. (1993). Advances in Diagnostic Imaging and Overestimations of Disease Prevalence and the Benefits of Therapy. N Engl J Med *328*, 1237-1243.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185-193.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze *8*, 3-62.

Brooks, J. D. (2002). Microarray analysis in prostate cancer research. Curr Opin Urol *12*, 395-399.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Jr., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A *97*, 262-267.

Casas, S., Nagy, B., Elonen, E., Aventin, A., Larramendy, M. L., Sierra, J., Ruutu, T., and Knuutila, S. (2003). Aberrant expression of HOXA9, DEK, CBL and CSF1R in acute myeloid leukemia. Leuk Lymphoma *44*, 1935-1941.

Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R., and Childs, G. (1999). Making and reading microarrays. Nature Genetics supplement *21*, 15-19.

Clezardin, P., Bruno-Bossio, G., Fontana, A., Serre, C. M., Magnetto, S., and Frappart, L. (1999). Thrombospondins, tumor angiogenesis and breast cancer. Pathol Biol *47*, 368-374.

de Fraipont, F., Nicholson, A. C., Feige, J. J., and Van Meir, E. G. (2001). Thrombospondins and tumor angiogenesis. Trends in Molecular Medicine *7*, 401-407.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science *278*, 680-686.

DeYoung, M. P., Tress, M., and Narayanan, R. (2003). Identification of Down's syndrome critical locus gene SIM2-s as a drug therapy target for solid tumors. Proc Natl Acad Sci U S A *100*, 4760-4765.

Dracopoli, N., Houghton, A., and Old, L. (1985). Loss of polymorphic restriction fragments in malignant melanoma: implications for tumor heterogeneity. Proc Natl Acad Sci U S A *82*, 1470-1474.

Dudoit, Y., Yang, Y. H., Callow, M. J., and Speed, T. P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.

Eisen, M. B., and Brown, P. O. (1999). DNA arrays for analysis of gene expression. Methods Enzymol *303*, 179-205.

Ernst, T., Hergenhahn, M., Kenzelmann, M., Cohen, C. D., Bonrouhi, M., Weninger, A., Klaren, R., Grone, E. F., Wiesel, M., Gudemann, C., *et al.* (2002). Decrease and gain of gene expression are equally discriminatory markers for prostate carcinoma: a gene expression analysis on total and microdissected prostate tissue. Am J Pathol *160*, 2169-2180.

Fiucci, G., Ravid, D., Reich, R., and Liscovitch, M. (2002). Caveolin-1 inhibits anchorage-independent growth, anoikis and invasiveness in MCF-7 human breast cancer cells. Oncogene *21*, 2365-2375.

Fodor, S. P. A., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991). Light-directed, spacially addressable parallel chemical synthesis. Science *251*, 767-773.

Folkman, J., and Kalluri, R. (2004). Cancer without disease. Nature *427*, 787.

Fra, A. M., Mastroianni, N., Mancini, M., Pasqualetto, E., and Sitia, R. (1999). Human caveolin-1 and caveolin-2 are closely linked genes colocalized with WI-5336 in a region of 7q31 frequently deleted in tumors. Genomics *56*, 355-356.

Frolov, A., Prowse, A. H., Vanderveer, L., Bove, B., Wu, H., and Godwin, A. K. (2002). DNA array-based method for detection of large rearrangements in the BRCA1 gene. Genes Chromosomes Cancer *35*, 232-241.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531-537.

Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L., and Aebersold, R. (2002). Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. Mol Cell Proteomics *1*, 323-333.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol Cell Biol *19*, 1720-1730.

Haab, B. B. (2001). Advances in protein microarray technology for protein expression and interaction profiling. Curr Opin Drug Discov Devel *4*, 116-123.

Hanahan, D., and Folkman, J. (1996). Patterns and Emerging Mechanisms of the Angiogenic Switch during Tumorigenesis. Cell *86*, 353-364.

Harrison, P. M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. (2002). A question of size: the eukaryotic proteome and the problems in defining it. Nucl Acids Res *30*, 1083-1090.

Herrick, D., Parker, R., and Jacobson, A. (1990). Identification and comparison of stable and unstable mRNAs in Saccharomyces cerevisiae. Mol Cell Biol *10*, 2269-2284.

Hoffmann, R., Seidl, T., and Dugas, M. (2002). Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biology *3*, research0033.0031 - 0033.0011.

Holland, M. J. (2002). Transcript abundance in yeast varies over six orders of magnitude. J Biol Chem *277*, 14363-14366.

Hubbell, E., Liu, W. M., and Mei, R. (2002). Robust estimators for expression analysis. Bioinformatics *18*, 1585-1592.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., *et al.* (2000). Functional discovery via a compendium of expression profiles. Cell *102*, 109-126.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics *18 Suppl 1*, S233-240.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science *292*, 929-934.

Ihaka, R., and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics *5*, 299-314.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249-264.

Jemal, A., Murray, T., Samuels, A., Ghafoor, A., Ward, E., and Thun, M. J. (2003). Cancer Statistics, 2003. CA Cancer J Clin *53*, 5-26.

Johnson, K., and Lin, S. (2003). QA/QC as a pressing need for microarray analysis: meeting report from CAMDA'02. Biotechniques *Suppl*, 62-63.

Jones, P. A. (1996). DNA methylation errors and cancer. Cancer Res *56*, 2463-2467.

Kaelin, W. G., Jr. (2004). Gleevec: prototype or outlier? Sci STKE *2004*, pe12.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science *258*, 818-821.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med *7*, 673-679.

Kohonen, T. (1990). The Self-Organizing Map. Proceedings of the IEEE *78*, 1464-1477.

Kondo, K., Umemoto, A., Akimoto, S., Uyama, T., Hayashi, K., Ohnishi, Y., and Y, M. (1992). Mutations in the P53 tumour suppressor gene in primary lung cancer in Japan. Biochem Biophys Res Commun *183*, 1139-1146.

Korabiowska, M., Bauer, H., Quentin, T., Stachura, J., Cordon-Cardo, C., and Brinck, U. (2004). Application of new in situ hybridization probes for Ku70 and Ku80 in tissue microarrays of paraffin-embedded malignant melanomas: correlation with immunohistochemical analysis. Human Pathology *35*, 210-216.

Kraal, G., van der Laan, L. J. W., Elomaa, O., and Tryggvason, K. (2000). The macrophage receptor MARCO. Microbes and Infection *2*, 313-316.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci U S A *94*, 13057-13062.

Lee, S. W., Tomasetto, C., and Sager, R. (1991). Positive selection of candidate tumor-suppressor genes by subtractive hybridization. Proc Natl Acad Sci U S A *88*, 2825-2829.

Lemon, W. J., Palatini, J. J. T., Krahe, R., and Wright, F. A. (2002). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. Bioinformatics *18*, 1470-1476.

Li, C., and Hung Wong, W. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol *2*, research0032.0031 - 0032.0011.

Li, C., and Wong, W. H. (2001a). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA *98*, 31 - 36.

Li, C., and Wong, W. H. (2001b). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol *2*, research0032.0031 - 0032.0011.

Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. Nat Genet *21*, 20-24.

Liu, W.-m., Mei, R., Di, X., Ryder, T. B., Hubbell, E., Dee, S., Webster, T. A., Harrington, C. A., Ho, M.-h., Baid, J., and Smeekens, S. P. (2002). Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics *18*, 1593-1599.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1995). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology *14*, 1675-1680.

Lodish, H., Baltimore, D., Berk, A., Zipursky, S. L., Masudaira, P., and Darnell, J. (1995a). 12. Transcription Termination, RNA Processing, and Posttranscriptional Control. Molecular Cell Biology Third Edition, 485-539.

Lodish, H., Baltimore, D., Berk, A., Zipursky, S. L., Masudaira, P., and Darnell, J. (1995b). Eukaryotic Gene Control: Purpose and General Principles. Molecular Cell Biology Third Edition, 426-430.

Luo, A., Kong, J., Hu, G., Liew, C., Xiong, M., Wang, X., Ji, J., Wang, T., Zhi, H., Wu, M., and Liu, Z. (2004). Discovery of Ca2+-relevant and differentiation-associated genes downregulated in esophageal squamous cell carcinoma using cDNA microarray. Oncogene *23*, 1291-1299.

Mocellin, S., Rossi, C. R., Pilati, P., Nitti, D., and Marincola, F. M. (2003). Quantitative real-time PCR: a powerful ally in cancer research. Trends Mol Med *9*, 189-195.

Moorthamer, M., and Chaudhuri, B. (1999). Identification of ribosomal protein L34 as a novel Cdk5 inhibitor. Biochem Biophys Res Commun *255*, 631-638.

Muller-Tidow, C., Schwable, J., Steffen, B., Tidow, N., Brandt, B., Becker, K., Schulze-Bahr, E., Halfter, H., Vogt, U., Metzger, R., *et al.* (2004). High-Throughput Analysis of Genome-Wide Receptor Tyrosine Kinase Expression in Human Cancers Identifies Potential Novel Drug Targets. Clin Cancer Res *10*, 1241-1249.

Mutch, D. M., Berger, A., Mansourian, R., Rytz, A., and Roberts, M. A. (2002). The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data. BMC Bioinformatics *3*, 17.

Myers, T. W., and Gelfand, D. H. (1991). Reverse transcription and DNA amplification by a Thermus thermophilus DNA polymerase. Biochemistry *30*, 7661-7666.

Nilsson, J. A., and Cleveland, J. L. (2003). Myc pathways provoking cell suicide and cancer. Oncogene *22*, 9007-9021.

Nishiu, M., Yanagawa, R., Nakatsuka, S., Yao, M., Tsunoda, T., Nakamura, Y., and Aozasa, K. (2002). Microarray analysis of gene-expression profiles in diffuse large B-cell lymphoma: identification of genes related to disease progression. Jpn J Cancer Res *93*, 894-901.

Ochs, M. F., and Godwin, A. K. (2003). Microarrays in cancer: research and applications. Biotechniques *Suppl*, 4-15.

O'Neill, G. M., Catchpoole, D. R., and Golemis, E. A. (2003). From correlation to causality: microarrays, cancer, and cancer treatment. Biotechniques *Suppl*, 64-71.

Paddison, P. J., Silva, J. M., Conklin, D. S., Schlabach, M., Li, M., Aruleba, S., Balija, V., O'Shaughnessy, A., Gnoj, L., Scobie, K., *et al.* (2004). A resource for large-scale RNA-interference-based screens in mammals. Nature *428*, 427-431.

Payelle-Brogard, B., Magnac, C., Mauro, F. R., Mandelli, F., and Dighiero, G. (1999). Analysis of the B-Cell Receptor B29 (CD79b) Gene in Familial Chronic Lymphocytic Leukemia. Blood *94*, 3516-3522.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. BMJ *316*, 1236-1238.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet *23*, 41-46.

Quackenbush, J. (2001). Computational analysis of microarray data. Nat Rev Genet *2*, 418-427.

Quackenbush, J. (2002). Microarray data normalization and transformation. Nat Genet *32 Suppl*, 496-501.

Rajagopalan, D. (2003). A comparison of statistical methods for analysis of high density oligonucleotide array data. Bioinformatics *19*, 1469-1476.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., *et al.* (2001). Multiclass cancer diagnosis using tumor gene expression signatures. PNAS *98*, 15149-15154.

Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput, 455-466.

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., and Chinnaiyan, A. M. (2002). Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer. Cancer Res *62*, 4427-4433.

Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., *et al.* (2000). Signaling and Circuitry of Multiple MAPK Pathways Revealed by a Matrix of Global Gene Expression Profiles. Science *287*, 873-880.

Schaner, M. E., Ross, D. T., Ciaravino, G., Sorlie, T., Troyanskaya, O., Diehn, M., Wang, Y. C., Duran, G. E., Sikic, T. L., Caldeira, S., *et al.* (2003). Gene Expression Patterns in Ovarian Carcinomas. Mol Biol Cell *14*, 4376-4386.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary cDNA microarray. Science *270*, 467-470.

Schenk, S., Schraml, P., Bendik, I., and Ludwig, C. U. (2001). A novel polymorphism in the promoter of the RAGE gene is associated with non-small cell lung cancer. Lung Cancer *32*, 7-12.

Schraml, P., Bendik, I., and Ludwig, C. U. (1997). Differential messenger RNA and protein expression of the receptor for advanced glycosylated end products in normal lung and non-small cell lung carcinoma. Cancer Res *57*, 3669-3671.

Shenkier, T., Weir, L., Levine, M., Olivotto, I., Whelan, T., and Reyno, L. (2004). Clinical practice guidelines for the care and treatment of breast cancer: 15. Treatment for women with stage III or locally advanced breast cancer. CMAJ *170*, 983-994.

Sherr, C. J. (2004). Principles of Tumor Suppression. Cell *116*, 235-246.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., *et al.* (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med *8*, 68-74.

Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002). Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics *18*, 261-274.

Simmonds, M. A. (2003). Cancer Statistics, 2003: Further Decrease in Mortality Rate, Increase in Persons Living with Cancer. CA Cancer J Clin *53*, 4.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol *98*, 503-517.

Stuart, R. O., Bush, K. T., and Nigam, S. K. (2001). Changes in global gene expression patterns during development and maturation of the rat kidney. Proc Natl Acad Sci U S A *98*, 5649-5654.

Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., Lapp, H., Schultz, P. G., Powell, S. M., Moskaluk, C. A., Frierson, H. F., Jr., and Hampton, G. M. (2001). Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res *61*, 7388-7393.

Taipale, J., and Beachy, P. A. (2001). The Hedgehog and Wnt signalling pathways in cancer. Nature *411*, 349-354.

Takada, M., Koizumi, T., Toyama, H., Suzuki, Y., and Kuroda, Y. (2001). Differential expression of RAGE in human pancreatic carcinoma cells. Hepatogastroenterology *48*, 1577-1578.

Tan, P. K., Downey, T. J., Spitznagel, E. L., Jr, Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., and Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. Nucl Acids Res *31*, 5676-5684.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. Nat Genet *22*, 281-285.

Toruner, G. A., Celal Ulger, C., Alkan, M., Galante, A., Rinaggio, J., Wilk, R., Tian, B., Soteropoulos, P., Hameed, M. R., Schwalb, M. N., and Dermody, J. J. (2004). Association between gene expression profile and tumor invasion in oral squamous cell carcinoma. Cancer Genetics and Cytogenetics *In press*.

Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., and Wong, W. H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res *29*, 2549-2557.

Tu, Y., Stolovitzky, G., and Klein, U. (2002). Quantitative noise analysis for gene expression microarray experiments. Proc Natl Acad Sci U S A *99*, 14031-14036.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. PNAS *98*, 5116-5121.

Ulger, C., Toruner, G. A., Alkan, M., Mohammed, M., Damani, S., Kang, J., Galante, A., Aviv, H., Soteropoulos, P., and Tolias, P. P. (2003). Comprehensive genome-wide comparison of DNA and RNA level scan using microarray technology for identification of candidate cancer-related genes in the HL-60 cell line. Cancer Genetics and Cytogenetics *147*, 28-35.

van der Kuip, H., Moehring, A., Wohlbold, L., Miething, C., Duyster, J., and Aulitzky, W. E. (2004). Imatinib mesylate (STI571) prevents the mutator phenotype of Bcr-Abl in hematopoietic cell lines. Leukemia Research *28*, 405-408.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Bernards, R., and Friend, S. H. (2003). Expression profiling predicts outcome in breast cancer. Breast Cancer Res *5*, 57-58.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530-536.

van 't Veer, L. J., and De Jong, D. (2002). The microarray way to tailored cancer treatment. Nat Med *8*, 13-14.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. Science *270*, 484-487.

Vogelstein, B., and Kinzler, K. W. (1993). The multistep nature of cancer. Trends Genet *9*, 138-141.

Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F., Jr., and Hampton, G. M. (2001a). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res *61*, 5974-5978.

Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., Lockhart, D. J., Burger, R. A., and Hampton, G. M. (2001b). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc Natl Acad Sci U S A *98*, 1176-1181.

Wiechen, K., Diatchenko, L., Agoulnik, A., Scharff, K. M., Schober, H., Arlt, K., Zhumabayeva, B., Siebert, P. D., Dietel, M., Schafer, R., and Sers, C. (2001). Caveolin-1 is down-regulated in human ovarian carcinoma and acts as a candidate tumor suppressor gene. Am J Pathol *159*, 1635-1643.

Wilusz, C. J., Wormington, M., and Peltz, S. W. (2001). The cap-to-tail guide to mRNA turnover. Nat Rev Mol Cell Biol *2*, 237-246.

Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol *15*, 1359-1367.

Yan, P. S., Rodriguez, F. J., Laux, D. E., Perry, M. R., Standiford, S. B., and Huang, T. H. (2000). Hypermethylation of ribosomal DNA in human breast carcinoma. Br J Cancer *82*, 514-517.

Yang, I. V., Chen, E., Hasseman, J. P., Liang, W., Frank, B. C., Wang, S., Sharov, V., Saeed, A. I., White, J., Li, J., *et al.* (2002a). Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol *3*, research0062.

Yang, Y. H., Buckley, M. J., and Speed, T. P. (2001). Analysis of cDNA microarray images. Brief Bioinform *2*, 341-349.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002b). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res *30*, e15.

Yang, Y. H., and Speed, T. (2002). Design issues for cDNA microarray experiments. Nat Rev Genet *3*, 579-588.

Zar, J. H. (1999). Biostatistical Analysis 4th edn., Upper Saddle River, NJ: Prentice Hall).

Zeisig, B. B., Milne, T., Garcia-Cuellar, M.-P., Schreiner, S., Martin, M.-E., Fuchs, U., Borkhardt, A., Chanda, S. K., Walker, J., Soden, R., *et al.* (2004). Hoxa9 and Meis1 Are Key Targets for MLL-ENL-Mediated Cellular Immortalization. Mol Cell Biol *24*, 617-628.