QUALITY FILTERING AND NORMALIZATION FOR MICROARRAY-BASED CGH DATA

by

Mehrnoush Khojasteh Lakelayeh
B.Sc., SHARIF University of Technology, 2001

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
Master of applied science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

December 2004

ABSTRACT

Altered copy numbers of DNA sequences are a characteristic of solid tumors. Microarray-based Comparative Genomic Hybridization (CGH) has emerged as a promising technology that has the potential to identify minute genomic changes, in the order of single DNA copy number changes, at the gene level.

The data to be extracted from the two microarray images of a 2-color microarray experiment, in the image analysis step, are the ratios of the fluorescent intensities of each spot of the microarray in one image and that of the corresponding spot in the other image. Without identifying the sources of experimental error, and correcting for these errors or removing the data corrupted by significant errors, microarray results can lead to incorrect experimental conclusions.

This research focuses on improving the "image analysis" step of array-CGH experiments. The aim is to reduce the variability and increase the validity of the experimental results. Two issues are addressed in this work: 1) identifying spots likely to be of poor quality, and 2) normalization of the data to remove systematic errors.

In this work, we present a novel approach to quality filtering of microarray spots. We use a variety of shape and image texture measures and design a binary decision tree to discriminate between the spots likely to produce meaningful data and the ones with unreliable measurement data. The proposed procedure is shown to reduce the variability of the data resulting from the low quality spots.

In addressing the second issue, possible sources of systematic variations are examined and accordingly a three-step normalization scheme is used to remove these systematic variations.

The normalization scheme we used consists of the following steps. The spatial bias of the ratio of each spot is estimated using a sliding window centered on each spot and the median of the ratios of the spots inside the window is calculated. The spatial bias is then removed from the data. In the next step, microplate effects are removed from the data. In the final step, the intensity dependent bias is estimated by fitting a LOESS curve to the

logarithm of ratios of spots as a function of the intensities of spots. This bias is then subtracted from the log ratios.

This normalization scheme was shown to increase the accuracy and precision of microarray data.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Calum MacAulay. It was a great honor for me to work under his supervision. His wide knowledge, constant guidance, patience and encouragements have been of great value to me.

I am also very grateful to my other supervisor, Dr. Rabab Ward, for giving me the opportunity to do this thesis and for her support throughout this work.

I also wish to thank my colleagues Spencer Watson, Ronal J. de Leeuw, and Bradley Coe for their useful discussions throughout this work.

Finally, many thanks to my dearest, Ali, for all the precious moments that he shared with me, on my good and bad days.

This thesis is dedicated to my beloved parents, Derakhshande and Ebrahim, whose ever lasting love and support made my greatest dreams come true. Just to say how much I love them and how much I value the sacrifices that they made for me.

This work was supported by the research grant provided by National Cancer Institute of Canada (NCIC) and Genome Canada.

CHAPTER 1 INTRODUCTION

Normal human cells contain two copies of each of the 22 non-sex chromosomes and depending on the sex of the individual either two X chromosomes (female) or an X and a Y chromosome (male). In tumor cells parts of chromosomes (i.e. DNA sequences) may be deleted or amplified. Considering only non-sex chromosomes, DNA sequence copy numbers are defined to be 2 copies for normal cells, 1 copy for a single deletion, 0 for a double deletion, 3 copies for a single copy gain, and greater than 3 copies for higher level amplifications.

In the past, the analysis of the genomes of tumors has been accomplished through the process of Comparative Genomic Hybridization (CGH) of fluorescently labeled DNA to metaphase chromosomes. This technique can only detect very large, in extent, coy number alterations. Recent improvements in the resolution and sensitivity of CGH have been possible through implementation of microarray-based CGH (array-CGH). Microarray-based comparative genomic hybridization (also known as array-CGH) provides a means to quantitatively measure DNA copy-number aberrations at a very high base-pair resolution and to map them directly onto genomic sequence. Ideally the purpose of array-CGH technology is to construct a map of the copy number alterations based on DNA clones (small pieces of DNA) as a function of the position of the clone within the genome. Array-CGH technology is potentially the most powerful tool currently available for research and clinical applications in medical genetics and cancer understanding.

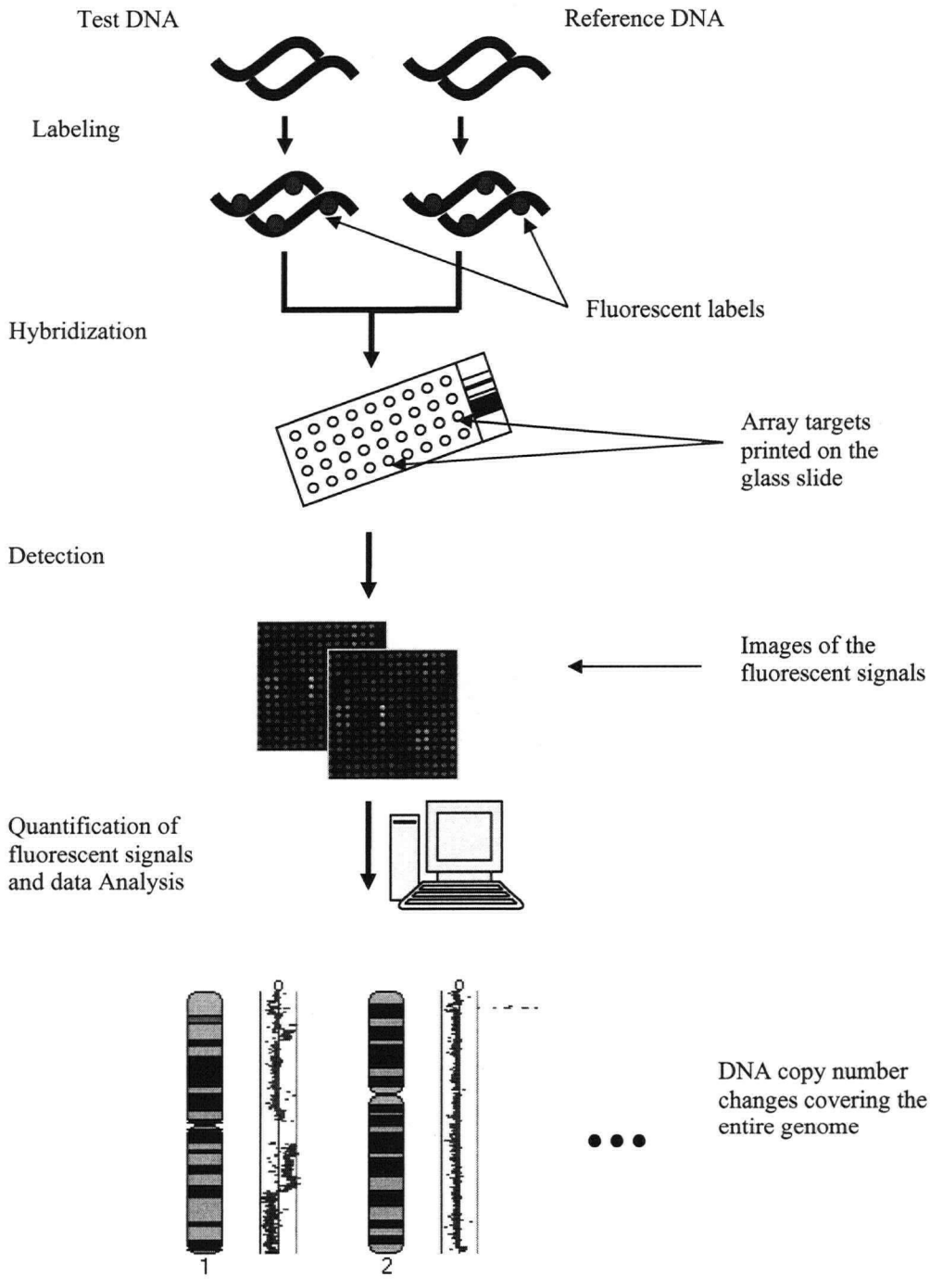

Figure 1-1 Array-CGH experimental steps

In figure 1-1 we summarize the different steps involved in an array-CGH experiment. The DNA clones to be spotted (deposited) on the array are selected based on the experimental design and prepared for robotic printing by an Arrayer. The Arrayer

deposits DNA spots of the same concentration and same volumes at uniform distances on a microscope slide using an array of pins. To test for copy number changes in tumor DNA for example, the test (tumor) and reference (normal) genomic DNA are differentially fluorescently labeled, mixed, and deposited onto the microaray slide. After hybridization (the process of single strands of DNA finding and hybridizing to their complementary strand), the fluorescence signals are detected. Using image-analysis software, chromosomal regions with an abnormal test to reference ratio which indicate a loss or gain of DNA sequences, can be detected. The copy number changes can then be positioned according to their location in the genome to give a full view of the alterations in the genome (shown in figure 1-1).

1.1 Goals of this research

Analyzing and reporting data generated from DNA microarray experiments is a challenging process. There are many potential sources of errors that must be controlled to obtain valid and reproducible results from these experiments and estimate relative copy number of DNA sequences.

The data to be extracted from the microarray images are the test to reference ratios of fluorescent signal intensities for the DNA spots of the array. Before the high level analysis of the data with the goal of finding regional gains or losses of DNA, some intermediate processing steps are needed to ensure the quality of results. This preprocessing includes identifying the spots for which the detectable characteristics indicate that data from those locations are very likely to be unreliable to exclude them from further analysis and normalization to remove the systematic variations in the data.

The goal of this work is to generate a more valid and reproducible measure of array-CGH microarray data. Microarrays were originally developed to measure relative gene expression levels, which vary greatly, in order to find genes that are differentially expressed in test and reference samples. Array-CGH microarrays are newly developed to measure relative copy number of DNA sequences. Array-CGH technology has more stringent performance requirements than Gene Expression microarrays in that the number of copies of specific DNA sequences within cells is highly regulated and reproducible. There performance requirements are: 1) To detect Single DNA copy number changes

within tumor cells with some degree of contamination from normal cells, and 2) Due to the high cost of the material and experiments, performing replicate experiments should be avoided as much as possible.

To achieve these requirements, we attempted to: 1) improve the removal of low quality spots, and 2) improve the normalization of the spot data to remove as much systematic variation as possible while preserving the real biological variations.

1.2 Overview of the thesis

Chapter 2 presents a summary of the basic biological background relevant to this work of this study as well as a summary of the various steps of microarray experiments in general. Chapter 3 deals with the particular type of array used in this work and explains the details of the array fabrication, probe labeling and hybridization and image acquisition applied at our center to obtain array-CGH data. The focus of this research is presented at the end of chapter 3.

The body of this thesis consists of two studies. The first study addresses the issue of removing the defect data points, i.e. spots of the image, that do not satisfy certain quality conditions, which correlate with the generation of unreliable data. The second study focuses on reducing the systematic variations in the data (through a process called normalization) to improve the consistency and accuracy of data across replicated experiments.

Chapter 4 addresses the issue of filtering out the low quality spots. It provides a literature review on this topic, the study hypothesis, methodology used to address the hypothesis and finally the results. Chapter 5 deals with the second study, optimization of the data normalization step. It contains a literature review of the topic, the study hypothesis and methodology, and the results. The conclusions and suggested future directions are given in chapter 6.

CHAPTER 2 BACKGROUND

In this chapter, a brief background on molecular biology concepts that are used in this thesis are presented. Microarray technology and experimental steps involved in microarray experiments are then explained.

2.1 Genes and Genomes

The blueprint of life is carried in the **genome**, an assembly of DNA bases organized into genes, and chromosomes. Cells pass an exact copy of the genome to newer cells during cell division, and the blueprint is inherited during reproduction. The genome is an organism's complete DNA sequence and encodes the genetic code required to create a particular organism with its own unique traits.

The structural arrangement of DNA looks like a ladder twisted into a helix, the sides of the ladder are formed by molecules of sugar and phosphate, while the rungs consist of pairs of nucleotide bases A (Adenine), T (Thymine), C (Cytosine) and G (Guanine) joined by hydrogen bonds. An important feature of the four bases is that they pair up with one another in a particular way. Nucleotide base A always pairs with T, and G always pairs with C. The two bases linked up in this way are called the "base pair". Each strand of the double helix consists of a sequence of nucleotides that are made of one of four bases A, T, G, C, a molecule of sugar and one of phosphate. The particular order of the bases arranged along the sugar-phosphate backbone is called the **DNA sequence**. Because of the way bases pair up with one another, the two strands of the DNA are said to be **complementary**. The size of DNA sequences are measured in a unit known as the **base pair** corresponding to one nucleotide pair of double stranded DNA.

All cells in a single human contain the same DNA but despite carrying the same set of instructions, cells are actually different. These differences are due to the fact that, stimulated by cell regulatory mechanisms or environmental factors, segments of DNA express the genetic code and provide instructions to the cells on when and in what quantity to produce specific proteins. These segments of the DNA are the **genes**. Each gene encodes a specific mRNA and protein, the latter of which imparts biological function in the cell. The process by which they become active is called their **expression**.

Gene expression takes place in two phases: transcription and translocation. During the transcription phase, one of the two complementary strands of the gene, transcribes base U (Uracil) for A, A for T, G for C and C for G into a strand of **mRNA** (messenger RNA). The mRNA transcript is moved from the nucleus to the cellular cytoplasm where it serves as a template on which tRNA molecules, carrying amino acids are lined up. The amino acids are then linked together to form a protein.

The **gene expression level** is a measure that provides a quantitative description of the gene expression by measuring the number of intermediary molecules produced during this process. Because the gene expression consists of copying its DNA code into mRNA molecules, a measure of the gene expression level is the abundance of mRNA produced during this process. This is the main assumption behind the large scale measurement of gene expression levels in gene expression microarrays [2].

2.2 Microarray technology

The microarray technology was first developed to simultaneously measure the relative expression level of thousands of genes within a particular cell population or tissue. In the newly developed array-CGH technology, the goal is to simultaneously measure the changes in the copy number of tens of thousands of DNA sequences within the genome of the tissue tested.

In the following section we define what the term "microarray" refers to and include an explanation of the gene expression microarrays and CGH microarrays.

2.2.1 What is a Microarray?

A microarray is an ordered array of microscopic elements on a planar substrate. "Microscopic" is defined as anything smaller than 1 mm (1000 μm). An ordered array is a collection of analytical elements configured in rows and columns. Analytical elements are also called **spots** in the microarray literature. Ordered elements must have a uniform size and spacing and a unique location on the microarray substrate.

Microarray elements are collections of **target** molecules that allow specific binding of **probe** molecules including genes and gene products. Microarray target material can be derived from whole genes or parts of genes and may include genomic DNA, cDNA, mRNA, protein, small molecules, tissues, or any other type of molecule that allows quantitative gene analysis. Target molecules include natural and synthetic derivatives obtained from a variety of sources such as cells, enzymatic reactions and processes that carry out chemical synthesis.

A planar substrate is a parallel and unbending support material such as glass, plastic, or silicon onto which a microarray is configured. Glass is the most widely used substrate material.

Specific binding refers to unique biochemical interactions between probe molecules in solution and their cognate target molecules on the microarray. Binding specificity allows a gene or gene product to be analyzed quantitatively with a single microarray target element [3].

Below we describe several technical concepts relevant to this measurement process:

Polymerase Chain Reaction: PCR is a method that allows selective amplification of any nucleic acid sequence from small quantities of starting material. PCR is used widely in microarray analysis for the amplification of DNA [3].

Cloned DNA: DNA Cloning consists of a number of molecular techniques that ultimately serve to place a defined segment of DNA within an organism, typically a different organism from which the DNA was originally derived, such that the DNA segment may be replicated repeatedly within the recipient organism. Several types of cloned DNA are used commonly in microarray analysis, including cDNA (complementary DNA) and BAC (Bacterial Artificial Chromosome). cDNA is a double stranded DNA molecule of ~0.2-5 kb (kilo base pair) length that is an exact replica of an

mRNA molecule. BAC DNA contains 50- to 250- kb segments of genomic DNA inserted into replicating bacteria. BAC clones are finding increasing use in microarray assays because they allow the representation of a large amount of genetic information on a single chip. The entire 3 billion bases of the human genome could be configured in a single microarray containing 30000 BAC clones with non-overlapping 100 kb inserts [3].

Reverse Transcription: the mRNA transcript of a gene can be experimentally isolated from a cell, and reverse-transcribed into a complementary DNA copy called cDNA [3].

Hybridization: is the process of base pairing of two single strands of DNA or RNA. DNA molecules are double-stranded and these two strands melt apart at a characteristic melting temperature, usually above 65°C. As the temperature is reduced and held below the melting temperature, single stranded molecules bind back to their counterparts. The process of binding back is again based on the principle of "base pairing", that is only two complementary strands of DNA can hybridize (bind) [3].

Oligonucleotides: are single stranded 15- to 70-nucleotide molecules made by chemical synthesis.

2.2.2 Gene Expression Microarrays

The use of expression microarrays is currently much more common than that of array-CGH microarrays so we begin by describing expression arrays.

In their most generic form, gene expression microarrays are ordered sets of DNA molecules attached to a solid surface. The DNA molecules are typically either oligonucletide or cDNAs. The matrix to which the DNA is attached is usually glass, silicon or nylon. Labeled (usually with a fluorescent nucleotide) cDNA representation of the cellular mRNA from a specimen (e.g. normal tissue, cell line, or tumor tissue) is hybridized to the array of DNA segments. Then the amount of cDNA at each DNA spot is detected and assumed to correspond to the transcript level (mRNA abundance) of the particular gene spotted at that location. Therefore the expression of thousands of genes can be analyzed in a single experiment [4].

2.2.3 Array-CGH

Microarrays have been exploited for gene expression studies but other applications can be envisioned and developed. One such application is the use of microarrays to study genomic DNA for detection of gains and losses of chromosomal regions [4]. Analysis of changes in DNA sequence copy numbers offers several advantages over the examination of the gene expression levels for the understanding of cancer and the process which leads to cancer.

Copy numbers of DNA sequences are more tightly controlled than the gene expression levels of the approximate 30000 human genes. Different genes are expressed at many different levels while the copy numbers of different sequences of normal DNA are always a constant. Another difference is that the range of copy number changes for different clones is small, from 50% to 100% decrease for loss and up to 100 fold increase for amplification. Thus the dynamic range of analysis is likely to be much more manageable than the gene expression changes which can be over 6 orders of magnitude in size [4].

Comparative Genomic Hybridization: In the past, the analysis of the genomes of the tumors has been accomplished with the process of Comparative Genomic hybridization (CGH) of fluorescently labeled DNA to metaphase spreads. In this technique, DNA from the tumor is labeled with a fluorescent dye in one color while a normal reference sample is labeled in a different color and these samples are co-hybridized to normal metaphase chromosomes (each chromosome representing a linear spatially localized target of ordered base pairs). Chromosomal imbalances across the genome in the test (tumor) DNA samples are quantified and positionally defined by analyzing the ratio of fluorescence of the two different colors along the target metaphase chromosomes. CGH has been successfully applied to analyze a variety of human tumors to detect chromosomal imbalances. However the resolution of CGH applied to metaphase spreads is limited to several mega base pairs. In addition, considerable cytogenetics expertise is required to apply this method. Given that chromosome distribution in every metaphase spread is unique it is very hard to develop automated methods for data capture and analysis. Moreover, the use of metaphase spreads limits the sensitivity of the method.

CGH performed on metaphase spreads is therefore not a high throughput technology and is limited to specialist research applications.

The limitations of CGH were resolved with the advent of microarray-based CGH. First described in 1997, matrix-CGH (also known as array-CGH) paved the way for higher resolution detection of DNA copy number aberrations. Array-CGH is based on the same principles as metaphase-CGH, except that the targets are mapped genomic clones instead of whole chromosomes [5, 6]. These genomic clones are spotted at specific locations on a glass slide in the form of a microarray.

The array format for CGH can provide a number of advantages over the use of chromosomes, including higher resolution and dynamic range, direct mapping of aberrations to the genome sequence and higher throughput. Furthermore, since the array format lends itself to automation, array CGH-based in vitro diagnostic devices are possible [1].

The performance goals of array CGH are more stringent than those of related array-based methods for measuring gene expression; the simplest array-CGH task is detection of large increases in copy numbers in DNA extracted from homogeneous cell lines. Achieving adequate performance is more difficult if one desires to reliably detect low level (single copy) gains and losses, especially as the size of the aberrant region decreases. Another dimension of challenge involves the use of tissue specimens, which may contain heterogeneous cell populations, (for example genomically normal cells within tumors), which makes the reliable detection of single copy changes in the tumor DNA subset relative to the normal diploid state even more difficult. In addition those precise measurements must be achieved for hybridizations involving the entire mammalian genome, a nucleic acid pool that has over ten times the complexity of usual expressed sequences, and which includes a significant quantity of repeated sequences. Moreover, the use of tissue from clinical specimens may impose constraints on the amount of DNA available for analysis. Finally, different applications have different tolerances for error, which substantially affect their performance requirements. For example, if one seeks composite information on the general characteristics of aberrations that occur in a set of specimens, the penalty for any single error is small. Indeed, missing a whole type of aberration is acceptable if other valuable information is obtained. However, it is much

more of a challenge to obtain specific information from an individual specimen with sufficient confidence for clinical use [1, 7].

2.3 Microarray Experiments

Microarray Analysis is the process of using microarrays for scientific exploration. Figure 2-1 summarizes the basic steps in a microarray experiment. These steps are described in more detail in the following sections. We mainly discuss the issues concerning "two-channel" DNA microarrays.

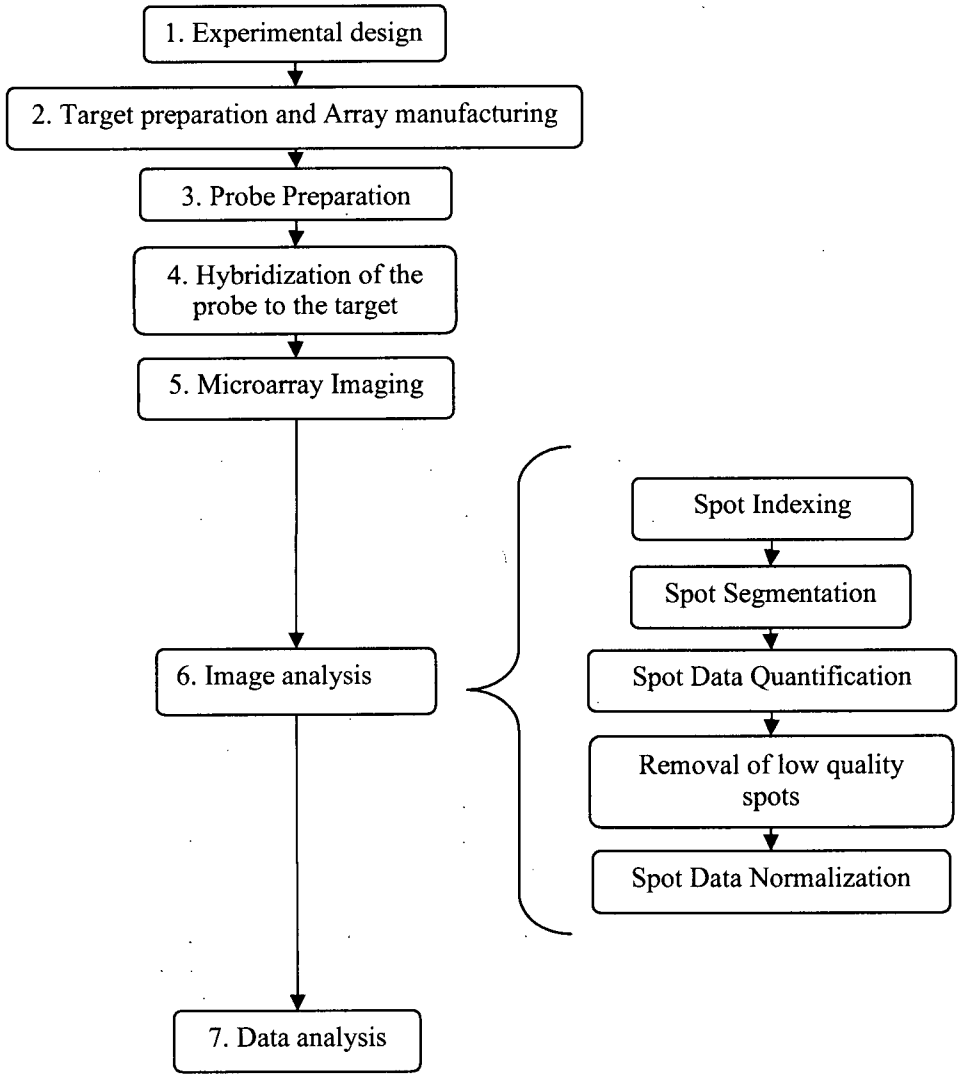

Figure 2-1 Microarray experimental steps

2.3.1 Experiment Design

Depending upon the biological question to be examined, a specific experiment is designed. The design of these experiments involves selecting the microarray probe, target, the number of biological and technical replicates to be performed, and the number of required replicate spots per target clone.

2.3.2 Target Preparation and Array Manufacturing

In this step, target DNA samples are prepared and the microarray is constructed. There are different types of microarrays which differ based on the DNA elements they are constructed with. These types of DNA elements include cDNAs, short oligonucleotides (e.g. 25bp) and long oligonucleotides (e.g. 70bp), and DNA drived from BAC clones.
Some microarray technologies, such as Short oligonucleotide arrays, synthesize the targets directly onto the array, which is also called a chip in this case.
The microarray technologies that do not synthesize the target DNA directly on the chip, first create the targets in a micro-plate format and then print these onto glass slides. Depositing the target DNA samples onto the glass slide, in order to create the high density arrays, is accomplished via microarray printing robots (also called **arrayers** or **spotters**) which come in many designs. The majority of spotters belong to the family of contact printers. These rely on a pin, which is loaded with sample, physically contacting the slide to deposit nano-liter scale volumes of printing solution [3].

2.3.3 Probe Preparation

The next step in a microarray experiment is the preparation of the test and reference DNA samples. The two samples will then be labeled by two different fluorescent dyes. The labeled samples are then mixed together. This mixture is called the **probe**.
Labeling: fluorescence dyes are typically used to label the DNA samples. Fluorescence can be defined as "the molecular absorption of light energy (photon) at one wavelength and its re-emission at another, usually longer, wavelength." Molecules that are able to absorb and emit light are known as **fluorochromes** or **fluorophores** [8].

Usually **CyDye fluorophores (fluors)** are used for labeling the probes before hybridization. CyDye fluors have well spectrally separated emission peaks that enable multiplexed detection. The relative probability that a fluorophore will be excited by a given wavelength of incident light is shown in its excitation spectrum. This spectrum is a plot of emitted fluorescence versus excitation wavelength. The relative probability that the emitted photon will have a particular wavelength is described in the fluorophore's emission spectrum, a plot of the relative intensity of emitted light as a function of the emission wavelength. Figure 2-2 shows the emission and absorption spectra of the two most commonly used CyDyes, Cy3 (green fluorescent dye) and Cy5 (red fluorescent dye) [8].

The **green** and **red** dyes have been interchangeably used for the **Cy3** and **Cy5** fluorescent dyes respectively throughout this thesis.

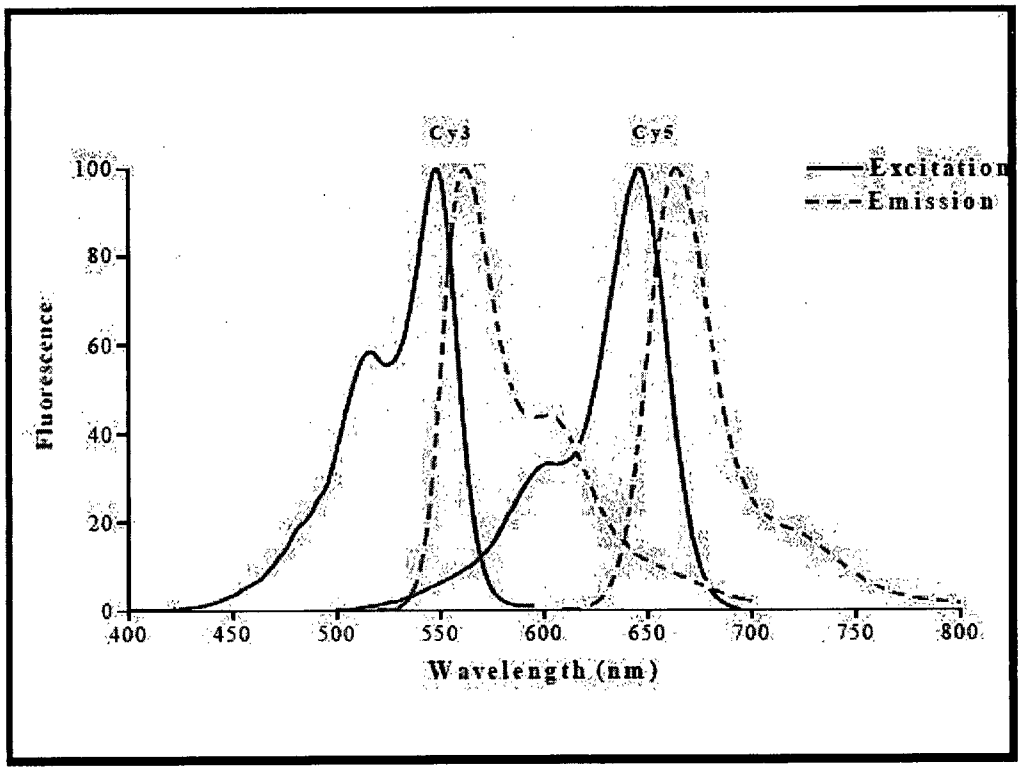

Figure 2-2 Emission and absorption spectra of Cy3 and Cy5 fluorescent dyes, figure from [8]

Fluorophores differ in the level of brightness they are capable of producing. Brightness depends on two properties of the fluorophore: its ability to absorb light (extinction

coefficient), and the efficiency with which it converts absorbed light into emitted fluorescent light (quantum efficiency) [8].

The quantum efficiency, excitation and emission spectra of a fluorophore can be affected by a number of environmental factors, including temperature, ionic strength, pH, excitation light intensity and duration, covalent coupling to another molecule, and non-covalent interactions (e.g. insertion into double-stranded DNA).

The intensity of the emitted fluorescent light varies with the intensity and wavelength of incident light and the brightness and concentration of the fluorophore. In general, when more intense light is used to illuminate a sample, more of the fluorophore molecules are excited, and the number of photons emitted increases. When the illumination wavelength and intensity are held constant, the number of photons emitted is a linear function of the number of fluorophore molecules. At very high fluorophore concentrations, the signal becomes non-linear because some of the emitted light is reabsorbed by other fluorophore molecules (**fluorescence quenching**). This is a common property of fluorescent compounds and requires close proximity of fluors. Different fluors have different quenching properties.

The amount of light emitted by a given number of fluorophore molecules can be increased by repeated cycles of excitation. In practice, however, if the excitation light intensity and fluorophore concentration are held constant, the total emitted light becomes a function of how long the excitation beam continues to illuminate those fluorophore molecules (this time is called the **exposure time**).

General requirements of labeled microarray samples are:

- Faithfully represent the relative abundance of each transcript in the mRNA population in both probes
- Sufficient amount of probe for efficient and even hybridization
- Sufficient degree of flour incorporation for required detection sensitivity, it should be noted that high labeling density can lead to quenching. [8]

The success of CyDye labeling is monitored through measuring the amount of CyDye in the probe DNA with **spectrophotometry**. A spectrophotometer is employed to measure the amount of light that a sample absorbs. The sample's absorbance spectrum is determined. The spectrum is a plot of absorbance versus wavelength and is characterized

by the wavelength (λ_{max}) at which the absorbance is the greatest. This wavelength is characteristic of each compound.

2.3.4 Hybridization of the Probe to Targets

After the probe is prepared, it is put on the surface of the microarray. Each DNA sequence in the probe will then find its complementary target sequence and hybridizes to it. Hybridization reactions between single stranded targets and probe molecules occur by hydrogen bond formation between the bases of complementary nucleic acid sequences [3].

The intensity of microarray hybridization signal is determined by:

- Amount of probe in hybridization reaction
- Number of molecules in target samples
- Labeling density
- Efficiency of hybridization of probe to targets (determined by hybridization conditions such as temperature, PH, etc., length of labeled molecules, structure of the target sequence, etc.)
- Detection set up (e.g. exposure time)

2.3.5 Imaging the Microarray

Imaging is the next step in microarray analysis. In this phase of the experimental process, the florescent intensity of the labeled probe molecules, bound to the target molecules, at each microarray location must be captured in a digital image by a scanning device. One image is generated for each **channel** of the microarray, i.e. for each different dye used.

Each scanner has the following components:

- Light source to activate the fluorescent molecules present on the array surface
- A strategy for exciting the fluorophores in a way that maximizes utility and minimizes the undesired effects such as light induced photo bleaching effects and cross talk caused by simultaneous excitation of different fluorophores.
- Optical design that provides a means of scanning the sample at the desired resolution

- Detectors for measuring the magnitude of the fluorescence signals
- A mechanism for orienting the slide and positioning during scanning [4]

In general microarray scanners fall into two main categories: those that use a white light source and charge coupled devices (CCDs) as detectors and others that use laser light with photomultiplier tubes (PMTs) [4]. Each category of scanners has its own advantages and disadvantages that are outside the scope of this study.

The scanning step results in one image file for each florescent label used in the microarray experiment. For example if two different fluorescent dyes are used for labeling the test and reference samples, then the microarray will be scanned at two different wavelengths corresponding to the emission spectra of the two dyes.

2.3.6 Microarray Image Analysis (Data Extraction)

The images obtained from the microarray slide in the previous step, are considered the raw data for the "image analysis" phase of the microarray experiment. One such image is shown in Figure 2-3. The process of converting the digital images into numerical measures of the amount of probe hybridized into each target spot is called "**Microarray Image Analysis**" (also known as data extraction).

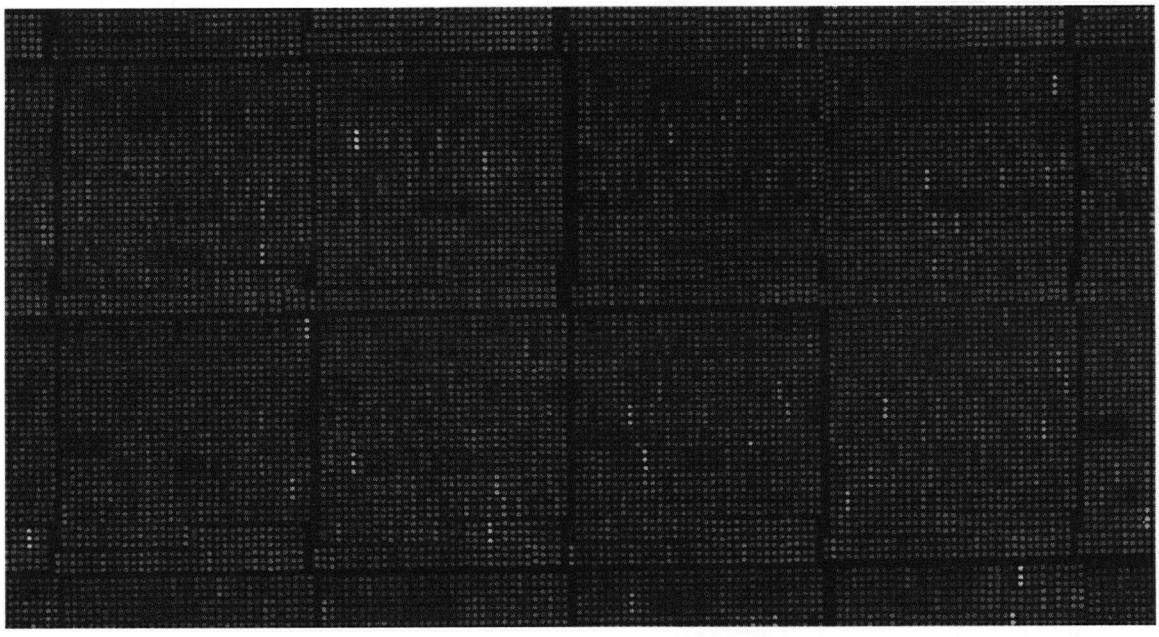

Figure 2-3 A sample microarray image

Since the introduction of microarrays, a number of microarray image analysis packages, both commercial software and freeware, have become available. The processing of scanned microarray images can be separated into three tasks:

1. **Indexing** or **gridding** is the process of assigning coordinates to each of the spots. Automating this part of the procedure permits high throughput analysis.
2. Segmentation allows the classification of pixels of the image either as **foreground**, or as **background**. Foreground pixels are pixels corresponding to a spot of interest. Background pixels are pixels outside of the boundary of the spot of interest.
3. In the spot data quantification step, for each spot in each microarray image (for a two-color microarray experiment, there are two microarray images, one for each channel), the foreground intensity (spot fluorescence), the background intensity (non-spot fluorescence) and possibly some spot quality measures are calculated [10].

Estimating the background intensity for each spot is generally considered necessary for the purpose of performing background correction. The motivation for background correction is that a spot's measured fluorescence intensity includes a contribution which is not specifically due to the hybridization of the probes to the target DNA. Background correction of the spot intensities is usually performed by subtracting background estimates from the foreground pixel intensities. This is done with the aim of improving the accuracy by reducing the bias in the spot data due to the background effects.

Spot quality measures may include measures of spot size or shape, or measures of background intensity relative to foreground intensity.

2.3.6.1 Indexing (Gridding)

The basic structure of a microarray image, i.e. the location of each spot within the array, is determined by the Arrayer (Spotter). However, to match an idealized model of the array with the image of the array, a number of parameters need to be estimated. These parameters include: separation between rows and columns of each block of spots, individual translation and/or rotation of blocks (caused by slight variations in arrayer's

print tip positions), separation between rows and column of spots within each block, small individual translation of spots, and overall position of the array in the image.
It is desirable for the addressing procedure to be as reliable as possible to ensure accuracy of the whole measurement process. Most software systems now provide both manual and automatic gridding procedures.

2.3.6.2 Segmentation

Segmentation classifies the image pixels as foreground or background. This allows fluorescent intensities to be calculated for each spotted DNA sequence as measures of the amount of DNA sequence hybridized to that spot. Existing segmentation methods for microarray images can be categorized into four groups according to the geometry of the spots they analyze [10]:

- **Fixed circle segmentation**: fits a circle with a constant diameter to all the spots in the image. This method is easy to implement and works nicely when all the spots are circular and of the same size.
- **Adaptive circle segmentation**: the circle's diameter is estimated separately for each spot.
- **Adaptive shape segmentation**: Two commonly used methods for adaptive segmentation in image analysis are the watershed and seeded region growing. These methods are beginning to be applied in microarray analysis, although not in the most widely-used software packages.
- **Histogram segmentation**: uses a target mask which is chosen to be larger than any spot. For each spot, foreground and background intensity estimates are determined in some fashion from the histogram of pixel values for pixels within the masked area. These methods therefore do not make use of any local spatial information.

2.3.6.3 Data Quantification

The key information that needs to be extracted from each spot of two-channel (two-color) microarrays is the relative amount of the DNA sequence corresponding to that spot in the test sample versus the amount of that sequence in the reference sample. Under idealized

conditions this translates to the ratio of the total fluorescent intensities of the spot in the two channel images. These idealized conditions are [4]:

- The incorporation of dyes into the probe during the labeling process is the same for both dyes used so that for the same amount of DNA samples the same amount of dye molecules are incorporated to the probes.
- The amount of DNA binding to the spot is proportional to the labeled DNA concentration in the probe.
- The detection efficiency of both dyes is the same (for the same number of fluorescent labels in the DNA, the quantum yield, photo bleaching and other physical chrematistics of two dyes are different but the camera exposure time should is usually adjusted to balance the intensity of both dyes.)
- There is no unbound or non-specifically bound probe attached to the spot.
- There is no auto-fluorescence of the glass slide and no other contaminant fluorescence.
- There is no signal contamination.
- The signal pixels are correctly identified by the image analysis software.

Traditionally, for two-color gene expression arrays, the "ratio" of fluorescent intensities of each spot in the two channels has been used to determine whether gene expression differs significantly for the red and green samples. Such an approach is intuitive since equal distributions for red and green values lead to a red/green ratio close to 1, and significantly unequal distributions lead to a red/green ratio significantly different from 1. This approach is typically applied by biologists developing microarrays [11].

The process of estimating the ratio of the amount of DNA sequences bound to the spots based on quantifying of the fluorescence intensities of the spot is called the "**data quantification**" phase.

In the most widely used method of data quantification, the pixel intensities of the areas in the image defined as foreground and background during the segmentation process, are averaged separately to give the foreground and background intensities, respectively. The median or other intensity extraction methods can be used when there are extreme values (outliers) in the spots that skew the distribution of pixel intensities. In this approach it is assumed that subtracting the background intensity of the spot from the foreground

intensity of the spot estimates the intensity of the probe DNA. The correctness of the whole approach of background subtraction in microarray images relies on the assumption that the local background is additive to the true signal. There are alternative methods of estimating the ratios that are not commonly used in image analysis software packages and therefore are not discussed here.

After extracting the ratio data for each spot, the first transformation that is commonly applied to microarray data is the log transformation. The raw ratio of intensities is not used. In stead the log transformed ratios are used. The log ratios are preferred to raw ratios because:

- Random error of the intensity measurement is approximately proportional to signal intensity; Moreover, most parametric statistical tests assume an additive rather than a proportional error model. Transforming expression data to a log scale (any base) removes much of the proportional relationship between random error and signal intensity (Low signals are often an exception. Random error of log-transformed data is often inversely proportional to the signal in the low signal range because of the proportionally nontrivial error associated with background correction). The Log transform has the advantage of transforming the error model from a proportional to an additive one since log (a/b) = log (a) – log (b).
- Distributions of replicated raw measured intensities and ratios tend to be asymmetric (skewed). This violates the normality assumption of many statistical tests. The central limit theorem affords little protection for most microarray studies because of the typically small sample sizes, which produce incorrect P-values associated with parametric analyses like the t-test and ANOVA.
- Summary statistics of replicated ratios yield different quantities, depending on the numerator/denominator assignment. In contrast, summary statistics of log ratios yield the same quantities, regardless of the numerator/denominator assignment. [29]

2.3.6.4 Removal of Low Quality Spots

It is crucial for any high throughput technology to have sufficient quality control for each step of the process. This is particularly true for microarray studies. Noise and irregularities of spot shape, size and position are common problems, especially in large-scale high-density microarrays. Therefore, users need to be able to acquire data quality measures to control for imperfections that happen during printing and hybridization. Without a good scheme to produce reliable, high quality data, analysis of data may lead to erroneous results.

The purpose of filtering out low quality spots is to identify those spotted samples that are likely to produce unreliable data therefore the experimenter wants to exclude from further analysis.

2.3.6.5 Data Normalization

As stated earlier, the key information that needs to be extracted from each spot of two-channel (two-color) microarrays is the relative amount of the DNA sequence corresponding to that spot, in one sample versus the amount of that sequence in the other sample. The goal of data normalization is to ensure that the ratio of intensity measurements made on the two images is equal to the DNA concentrations. For this to be true any systematic errors need to be removed. These systematic errors arise from background fluorescence within the DNA spots on the array, differences between the detection efficiency of the two dyes used for the test and control samples, difference in the incorporation efficiency of dyes into the DNA samples during the labeling process, fluorescence quenching effects that reduce the signal when the dyes fluoresce intensely, differences in hybridization efficiency, etc.

2.3.7 Microarray Data Analysis

After the image analysis step, the data extracted from the microarray slide are ready to be analyzed. Below we include a summary of microarray data analysis methods. Although the data analysis issue will not be addressed directly in this thesis but a summary of microarray data analysis methods is given below for both gene expression and CGH

microarrays. We believe this helps in understanding the differences the two studies described in this work would make on the final goal of microarray experiments.

2.3.7.1 Data Analysis for Gene Expression Arrays

This section describes the most popular techniques for the analysis of gene expression data in (possibly repeated) comparative experiments. The objective of this analysis is to identify the genes with significant expression change across two conditions (test and reference samples).

The choice of analysis method depends on the particular experiment design. A simple microarray experiment may be carried out to detect differential expression between two conditions. Using two color cDNA microarrays, samples can be compared directly on the same microarray or indirectly by hybridizing each together with a common reference sample. In the former case, the null hypothesis of no differential expression implies that the true log ratio should be zero and in the latter case that the log ratios (test sample to reference) should not differ between the two conditions. If a single color expression assay is used then we are again considering a null hypothesis of no expression level difference between the two conditions and methods described here can be applied directly.

A distinction should be made between RNA samples that represent replicate but independent biological samples, **biological replicates**, and those that represent repeated measurements of the same biological material, **technical replicates**. Ideally, each condition should be represented by multiple independent biological samples, biological replicates, in order to conduct statistical tests. If only technical replicates are available, statistical testing is still possible but the scope of the conclusions may be limited. If both technical and biological replicates are available, for example the same biological samples are each measured twice, the individual log ratios can be averaged to yield a single measurement for each biological unit in the experiment. More complicated settings that involve multiple layers of replication can be handled using the mixed model analysis of variance techniques.

Fold change is the simplest method for identifying differentially expressed genes. It is based on the observed ratio (or average of ratios) between two conditions. An arbitrary cut-off value (for example, 2 fold) is used to identify differentially expressed genes. This

is not a statistical test and there is no associated level of confidence. The fold change method is subject to bias if the data are not properly normalized and may also be sensitive to variance heterogeneity across genes. For example, an excess of low intensity genes may be identified as being differentially expressed due to an excess of variation relative to high intensity genes. Intensity specific thresholds have been proposed as a remedy for this problem. [12]

If replicate spots or experiments are available then one-sample t-test can be used on replicate measurements for the spot so as to test the hypothesis of no differential expression. The *t*-test is a simple, statistically based method for detecting differentially expressed genes. In replicated experiments, error variance can be estimated from the log ratios on a gene-by-gene basis and a standard *t*-test can be conducted for each gene. This gene-specific *t*-test is robust to variance heterogeneity across genes but it may have low power due to few degrees of freedom. It is possible to compute a global *t*-test using an estimate of error variance that is pooled across all genes under the assumption of homogeneous variation. This is effectively a fold change test and may suffer from the same biases if the error variance is not truly constant for all genes.

When there are more than two conditions in an experiment, we cannot simply compute ratios. A more general concept of relative expression is needed. One approach, which can be applied to cDNA microarray data from any experimental design, is to use ANOVA (Analysis of Variance) to obtain estimates of the relative expression for each gene in each sample. In the ANOVA model, the expression level of a gene in a given sample is computed relative to the weighted average expression of that gene over all samples in the experiment. We note that the ANOVA is not based on log ratios. Rather it is applied directly to intensity data. However the difference between two estimated expression values can be interpreted as the mean log ratio for comparing two samples [12].

2.3.7.2 Data Analysis for CGH Arrays

The experimental component of an array-CGH experiment, i.e. the new information that is sought in the experiment, is different from that of gene expression arrays. In gene expression microarrays the final goal is to find the genes that are expressed significantly

differentially in two samples or two conditions. The goal of array-CGH is to partition the clones of a given genomic probe into sets that have the same copy numbers.

Genomic rearrangements lead to gains or losses of sizable contiguous parts of the genome; therefore, in the analysis of the copy number changes of the clones, it is desirable to make use of the physical dependence of the nearby ordered clones.

The goal of the analysis of this type of data includes detection of locations of copy number changes, called **breakpoints**, and estimating the copy number value before and after the change. Knowing the exact locations of a breakpoint is important to identify possibly altered genes [13].

The problem of identifying the regions of gains and losses of DNA and finding the breakpoints is fairly new and ongoing research on this is being conducted while this report is being written. We include a summary of the articles addressing this issue.

Jong et al, [13, 14], define the problem as model fitting to search for most-likely-fit model for the given data. A model describes a number of breakpoints, a position for each, and parameters of the distribution of copy number for each.

Autio et al, [15], try to identify regions of amplifications and deletions, using k-means clustering and dynamic programming. The dynamic program utilizes the Markov property and identifies change points of the constant levels by minimizing sum of mean square errors for all combinations of CGH ratios.

Fridlyand et al, [16], use unsupervised Hidden Markov Models approach to utilize the spatial coherence between nearby clones and partition them into the states which represent underlying copy number for the group of clones.

CHAPTER 3 ARRAY-CGH EXPERIMENTS

In section 2.3, the steps of a microarray experiment were described. Detailed description of the specific steps of the array-CGH experiments of this study is presented in section 3.1.

In section 3.2, the specifications of the data of this study are presented.

After describing the experimental steps, in section 3.3, we specify the steps that are the focus of this study.

3.1 Steps of Array-CGH Experiments

The data of this study were obtained from array-CGH experiments performed in the BC Cancer Research Center with the SMRT (Sub Mega base Resolution Tiling) arrays [17] which are the first tiling resolution BAC arrays with complete coverage of the human genome using 32,433 fingerprint-verified individually amplified BAC clones.

3.1.1 Array Production from BAC DNA

The first step in a microarray experiment is to choose and prepare the DNA clones to be spotted on the array. The DNA samples to be spotted on the array are prepared by PCR. The PCR is done in Tetrad PCR machines. Each Tetrad machine accommodates four 96-well plates of PCR products. The DNA samples of the four plates are then re-arrayed in a 384-well ($96 \times 4 = 384$) microplate that is used by the Arrayer. Figure 3-1 shows how the re-arraying is done.

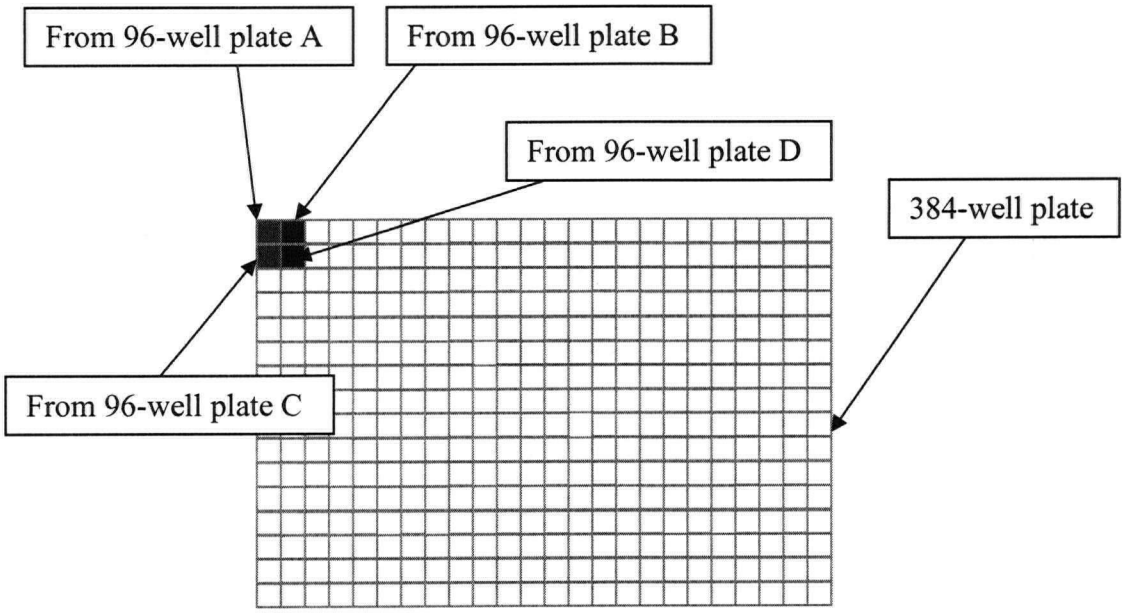

Figure 3-1 Rearraying of four 96_well plates into one 384-well microplate

The DNA samples from the 384-well microplate are then printed on the slides using a VersArray ChipWriter Pro (BioRad) Arrayer. This Arrayer uses an array of 4×12 spotting pins. The arrayer deposits the DNA samples in the specific locations on the array called **spots**.

The process of printing the slides takes place as follows:

The 4×12 spotting pins of the Arrayer are dipped in the 4×12 wells of the microplate and deposit 4×12 samples three times so each target clone is replicated in three spots. These three replicate spots are in a single column next to each other.

Next, the array of pins are dipped in the next 4×12 wells of the microplate and deposited onto the array. This "**dipping and depositing**" cycle is repeated until all the clones from the microplates are deposited onto the array (see figures 3-2 and 3-3).

The set of all of the spots that is printed by the same spotting pin is called a **subgrid**. Since the arrayer has 4×12 spotting pins, there are 4×12 subgrids on the microarrays.

The entire set of 32,433 solutions is spotted in triplicate onto two slides each containing 52272 spots. 2415 of these clones are spotted on both slides.

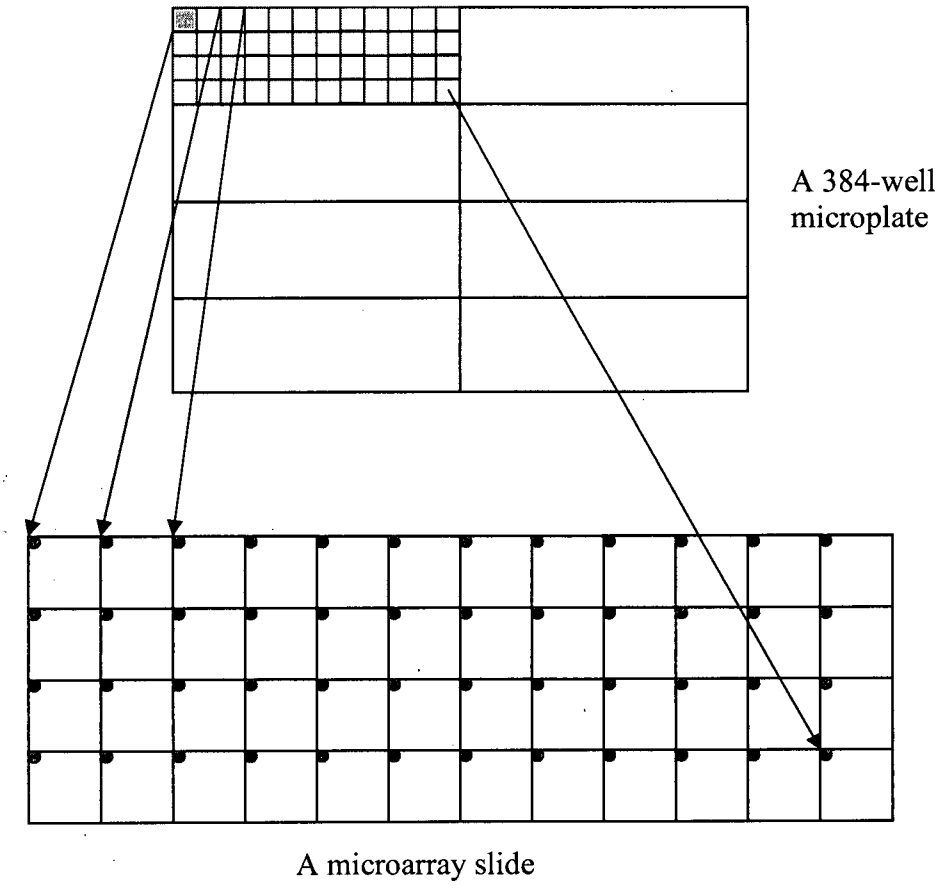

Figure 3-2 The first "dipping and depositing" cycle of printing the array, the pins are dipped in the first 48 wells of the microplate and spot the first 48 spots of the array, one spot from each subgrid is printed during each cycle.

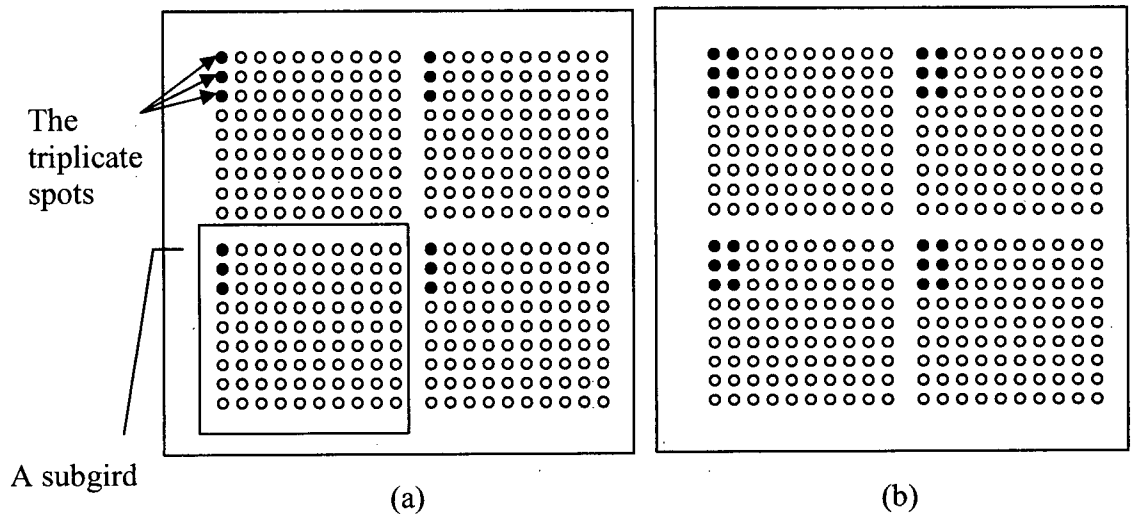

Figure 3-3 Grids and subgrids of a microarray, (a) printed spots after one cycle of "dipping and depositing" (b) after two cycles of "dipping and depositing", the printed spots are shown in full circles.

The volume of the DNA material that is deposited on each spot is 0.8 nano-liter at concentration of ~1 microgram/micro-liter. The spacing between the spots within a subgrid is133 micro meter.

3.1.2 DNA Labeling and Hybridization

The labeling of test and reference DNA has been done separately with Cy3 and Cy5 dCTPs (labeled C nucleotide) using a random priming protocol. Before hybridization unincorporated nucleotides are removed and the activity (the amount of each dye in the probe) of each dye is measured and human cot-1 DNA is added to the mixture of probes to block the repetitive sequences from hybridization with the spotted DNA.

3.1.3 Array Imaging

The hybridized slides were imaged using the CCD-based imaging system: ArrayWoRx Biochip Reader.

3.1.3.1 About ArrayWoRx Biochip Reader

This system is based on a CCD camera and filtered white light source.

The white light source is a high intensity metal halide bulb and associated optical components that deliver a beam of white light to the excitation filter.

The filter assembly holds up to four pairs of excitation and emission filters. One emission filter and one excitation filter are required to measure the intensity of a single fluorophore. Scanning software and motion control hardware place the appropriate filters in the light path for the fluorophore being measured.

The CCD camera, or Charge-Coupled Device camera, collects light which passes the emission filter and converts it into a digital signal that the scanner software processes to create an image file. The CCD chip is an array of semiconductor devices or camera pixels. Each camera pixel stores an electrical charge generated by the fluorescence light from the sample. This charge is proportional to the intensity of the light, or number of photons, that reaches the pixel. Electronic circuitry on the camera converts pixel electron

counts from the CCD chip into a digital signal that represents the intensity of light at each pixel.

Light emitted from the white light source is passed through an excitation filter, then distributed through 19 fiber optic strands and uniformly distributed to the slide/specimen. The emitted fluorescence travels through an objective lens, a designated emission filter, and is captured by the CCD camera (See figure 3-4).

Images are taken at multiple slide positions in order to assemble the high-resolution image "tiles" into a single, high-resolution image. [18]

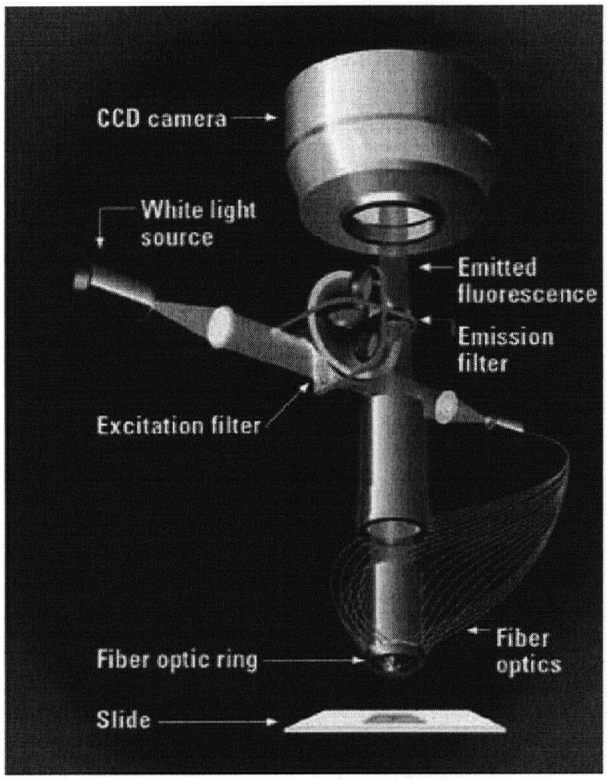

Figure 3-4 ArrayWoRx Biochip reader (from [18])

3.1.4 Image Analysis

The images were analyzed using SoftWoRx Tracker Spot Analysis software (Applied Precision). Below we describe how the software estimates the ratios:

First, the microarray geometry is specified by the user. This includes the specification of nominal values of spot diameter, spacing between spot centers, number of spots, and number of rows, column, and subgrids. The software will automatically adjust for

imperfections in the geometry. Usually the locations of some spots are not correctly identified by the software. Therefore, the user completes the spot finding process by manually adjusting the location of those spots.

After the spots are detected, the image pixels associated with the spots (foreground pixels) are determined. As a result, each spot is represented by a region of arbitrary shape.

Next, the pixels to be used in estimating the background are determined. To accomplish this, two circles are defined around the contour of the spot, one inner circle and one outer circle forming an annulus. The pixels within the annulus are used to estimate the background intensity. The radii of the inner and outer circles are 15% and 85% of the nominal edge to edge distance of two adjacent spots, respectively.

The "normalized intensity" of each spot in each of the two channels, is calculated as the mean of the intensities of the foreground pixels minus the median of the intensities of the background pixels.

The ratio for each spot is obtained by dividing the "normalized intensity" value of the spot in one channel by the "normalized intensity" value of that spot in the other channel. The ratios are further normalized so that the mean value of the log ratios is zero.

The output of the image analysis software, including the intensities of the spots, the estimated ratios, and some quality measures, is saved in a text file.

A custom analysis software (SeeGH, [19]) takes the generated text file as input and performs the removal of low quality spots as follows: Currently it averages the ratios of the triplicate spots and calculates their standard deviations. All spots with standard deviation higher than 0.075 or signal to noise ratio (defined as the intensity of the spot divided by the standard deviation of the spot's background) less than 20 are removed from further analysis. This software then visualizes all data as $\log_2$ ratios plotted versus the genomic order.

3.2 Data Description

Our study analysis is mainly based on data from four sets of experiments across 24 different microarray slides from the Cancer Genetics Department of BC Cancer Research Center.

The first experiment was a **self-self experiment**, i.e. both test and reference sample were from the same DNA sample and were only labeled separately with different dyes. Normal male genomic DNA was used for both test and reference material. The four slides making up this experiment are referred to as **MM-1** to **MM-4** through out this thesis.

Name	Cy3 DNA Sample	Cy3 Activity	Cy5 DNA Sample	Cy5 Activity	Cy3 Exposure time	Cy5 Exposure time	Resolution
H526-1	H526 400ng	9.1	Male Genomic 400ng	5.7	0.600 sec	1.700 sec	9.7560 um
H526-2	H526 600ng		Male Genomic 600ng		0.700 sec	1.300 sec	9.7560 um
H526-3	H526 400ng	8.5	Male Genomic 400ng	5.1	0.700 sec	2.000 sec	9.7560 um
H526-4	H526 400ng	8.5	Male Genomic 400ng	5.1	0.700 sec	2.000 sec	9.7560 um
H526-5	H526 200ng	9.1	Male Genomic 200ng	5	0.800 sec	1.600 sec	9.7560 um
H526-6	H526 200ng	9.1	Male Genomic 200ng	5	1.000 sec	2.000 sec	9.7560 um
H526-7	H526 400ng	11.5	Male Genomic 400ng	7.1			
H526-8	H526 400ng	11.5	Male Genomic 400ng	7.1			
MM-1	Male genomic 200ng	20.3	Male genomic 200ng	17			
MM-2	Male genomic 200ng	20.3	Male genomic 200ng	17			
MM-3	Male Genomic 250ng	8.6	Male Genomic 250ng	5.3	1.000 sec	2.000 sec	9.7560 um
MM-4	Male Genomic 300ng	8.3	Male Genomic 300ng	5.3	0.800 sec	2.000 sec	9.7560 um
MF-1	Female Genomic 300ng	6.3	Male Genomic 300ng	4	0.800 sec	2.500 sec	9.7560 um
MF-2	Male Genomic 300ng	9.1	Female Genomic 300ng	6.3	0.700 sec	1.100 sec	9.7560 um
T-1	400ng Male		400ng 0% Contamination (del)				
T-2	400ng Male		400ng 15% Contamination (del)				
T-3	400ng Male		400ng 30% Contamination (del)				
T-4	400ng Male		400ng 50% Contamination (del)				
T-5	400ng Male		400ng 75% Contamination (del)				

Name	Cy3 DNA Sample	Cy3 Activity	Cy5 DNA Sample	Cy5 Activity	Cy3 Exposure time	Cy5 Exposure time	Resolution
T-6	400ng Male	16.26	50% Contamination (amp)	9.03			
T-7	400ng Male	16.93	75% Contamination (amp)	9.39			
T-8	400ng Male	13.42	0% Contamination (amp)	7.29			
T-9	400ng Male	14.47	15% Contamination (amp)	7.73			
T-10	400ng Male	10.95	30% Contamination (amp)	6.4			

Table 3-1 Summary of slides names and descriptions

The second experiment uses H526 cell line DNA and compares it against normal male DNA. The 8 slides making up this experiment are named **H526-1** through **H526-8**.

The third experiment was normal male DNA versus normal female DNA. The two slides from this experiment are named **MF-1** and **MF-2**.

The Fourth set of experiments is a series of titration experiments comparing X chromosome loci to autosomal (non-sex) loci by comparison of male and female DNA [55]. A single copy deletion was simulated by hybridizing normal male versus normal female DNA, generating a 1:2 ratio of X chromosomes. Contamination from normal cells was then simulated by spiking varying amounts of female DNA into the male DNA sample (slides **T1** through **T5**). Single copy amplifications were modeled by comparing a 50/50 mixture of male and female DNA against a male DNA reference. In this model, contamination from normal cell was simulated by spiking varying amounts of female DNA into the male/female DNA mixture (Slides **T-6** through **T-10**).

The detailed description of the slides is given in table 3-1. In this table, **activity** of each dye is the amount of the incorporated dyes during the labeling process that is measured by the spectrophotometer. Also in table 3-1 the **exposure times** for scanning each channel of the microarray slide is reported. Since some information about some of the slides was missing, there are some empty cells in this table.

3.3 Aims

The aim of this thesis was to improve the "Image Analysis" step of the array-CGH experimental analysis in these two areas: 1) improving and optimizing the removal of spots likely to generate unreliable data (filter out low quality spots) and 2) normalization of the spot data to remove as much systematic variation as possible while preserving the real biological variations. The former issue will be addressed in chapter 4 and the latter will be addressed in chapter 5.

CHAPTER 4 FILTERING OUT THE LOW QUALITY SPOTS

This chapter addresses the issue of filtering out the low quality spots. It provides a literature review on this topic, study hypothesis, an examination of sources of artifacts in the data, methodology used to address the hypothesis and finally the results.

4.1 Background

It is crucial for any high throughput technology to have sufficient quality control for each step in the process to enable the collection of good quality data. The data acquisition step of microarrays analysis is not an exception. Noise and irregularities in spot shape, size and position are common problems which may affect the measurement accuracy, particularly in large scale high density microarrays. Therefore users need to be able to acquire a measure of the quality of data, to control for imperfections that happen during printing and hybridization. Without a good scheme to produce reliable, high quality data, any complex data mining tools one may use can lead to misleading results [20].

There are two different approaches to quality measurement of microarray data. First is through replicate spots. Spot replicates are considered to be a valuable source of information for data significance and confidence analysis of differentially expressed genes. They can also be successfully utilized for flagging of low quality spots, which are highly likely to produce unreliable results. The most common approach to quality control in this area has been based on the replicate outlier removal. The presence of an outlier replicate within measurements raises concern about quality of the measurements. The most significant drawback of this approach is the necessity for a fairly large number of replicates.

The other approach to quality assessment is the assessment of quality through confidence measures. These confidence measures are obtained during the image analysis phase. The choice of these measures mainly depends on the particular microarray design, equipment sophistication and measurement extraction procedures. The most widely used measures are the ratio of the standard deviation of the signal within the spot to its mean intensity, the offset of the spot from its expected position in the grid and spot circularity measures. These measures used separately or combined into some kind of a decision tree can be used to flag a spot as of low quality. However it is not obvious how to compose a unique confidence number from such set of quantities.

Sources that reduce the quality of a spot and increase the chance that it produces unreliable data can be separated into two groups. The first group consists of defects introduced during the slide printing and scanning process. The other group includes miscalculations as a consequence of the poor performance of the spot finding and image segmentation techniques applied to the image [4]. There are not a lot of publications that have done a systematic study on these quality assessment issues. The following represents a review of the few studies addressing this issue.

Wang et al, [20], define several quality scores for each spot on the array according to its size, signal to noise ratio, background level and uniformity, and saturation status. Based on these five individual scores, a composite score q_{com} is defined for each spot to give an overall assessment of its quality. They demonstrated that the variability in ratio measurements correlates closely with q_{com} in that high-quality spots are less variable and that q_{com} is better than intensity level or spot size used alone in a data filtering scheme.

Brown et al., [21], calculated the normalized standard deviation of the ratio distribution, a value they referred to as SRV (Spot Ratio Variability), as a measure of ratio non-homogeneity that summarizes the reliability of the expression ratio for a spot. They suggested using this metric in combination with other obvious problems to capture all anomalies including those not captured by the SRV alone.

Tseng et al, [22] used multiple spotting of each target sequence on a slide as a means to assess the quality of data for a spot on that slide. They assumed that the quality of data on the expression level of each gene is inversely related to the coefficient of variation (i.e. standard deviation divided by the mean) of the set of ratios of the corresponding multiple

spots. The measure was referred to ass CV. By a windowing procedure they marked all the genes having CV values larger than a threshold as poor quality data. For each gene they constructed a windowing subset by selecting the 50 genes whose mean intensities are closest to that of this gene. If the CV of this gene is among the top 10% among genes in its windowing subset then they regard the data on this gene as unreliable.

Ruosaari et al, [23] address the problem of detecting spots of low quality from the microarray images by extracting features describing the spatial characteristics of the spots on the microarray image and train a classifier using a set of labeled spots. They assess the results for classification of individual spots using region of ROC analysis and for a compound classification using a non-symmetric cost structure for misclassifications.

Hautaniemi et al, [24], suggested using Bayesian networks for computing the spot quality value from spot specific features. The features they used in their study were: spot intensity, size of the spot, roundness of the spot, alignment error, background intensity, background noise and bleeding of one spot's target into its neighboring spots.

In studies of Ruosaari et al [23] and Hautaniemi et al [24], the assessment of quality of the spots was done visually; however, Wang et al [20] and Brown et al [21] used the experimental variability of the ratios for the assessment of the quality of spots.

4.2 Hypothesis

The replicate filtering approach has a significant disadvantage: the need for a rather large number of replicate spots per target clone. This method is not cost effective. The fact that the amount of DNA available for experiment is usually limited makes it less practical especially for clinical experiments. When the number of replicate spots per target clone is not large there is no way to find the outlier measurements among the replicate measurements because the number of defect spots is not known. It can be one or more than one of the replicate spots. So the only solution would be to discard all the replicate measurements which eliminates all the measurements for a clone (hence the biological feature is excluded) and is not desirable.

The quality measure approach does not have the discussed disadvantage of the other approach. However the following two issues need to be addressed. First, the choice of the features of the spots that most significantly describe the validity or quality of the intensity

and/or ratio measurements and second how to combine the single selected features onto a unified quality score so that the final quality score conforms to the quality of the spots.
The few studies published on this issue have the following drawbacks: the selection of the features and the choice of the unified quality score have been done based on observations alone, the feature sets have not been comprehensive to contain all the possible quality features of the spots, and finally in cases that a classifier have been used for the purpose of identifying the defect spots, the performance of the classifier in terms of removing the outliers from the data have not been evaluated in a database that enables the assessment against truth .
Currently in our lab, our microarray images of array-CGH slides are analyzed with SoftWoRx Tracker microarray experiment management and analysis software. Using a custom in house viewing software (SeeGH, [9]) the standard deviation (**s.d.**) of triplicate spots are calculated and all spots with higher s.d. or lower signal to noise ratios (SNR), are excluded from further analysis. SNR is defined as the ratio of the background subtracted mean signal divided by the standard deviation of the background of the spot.
The hypothesis tested in this part of the thesis is: Can a binary decision tree consisting of linear discriminant functions with a comprehensive set of features calculated for each spot result in the same or better accuracy than the current approach for filtering out the low quality spots in array-CGH data?
If successful, the proposed method would reduce the number of replicate spots per target clone from three to two and the whole SMRT array set of clones would fit on one slide instead of two. The latter makes the experiments much more cost and material effective.

4.3 Usual Artifacts

In the following sections the types of artifacts observed in the microarray images in our database will be described.

4.3.1 Printing Artifacts

For understanding this type of artifacts, some knowledge of the microarray slide spotting system is necessary.

By following the pattern of spotting we can find all the spots that are printed from the same 96-well original plate. Figure 4-1 shows this grouping of spots on the array. We refer to these groups as the **plate groups**.

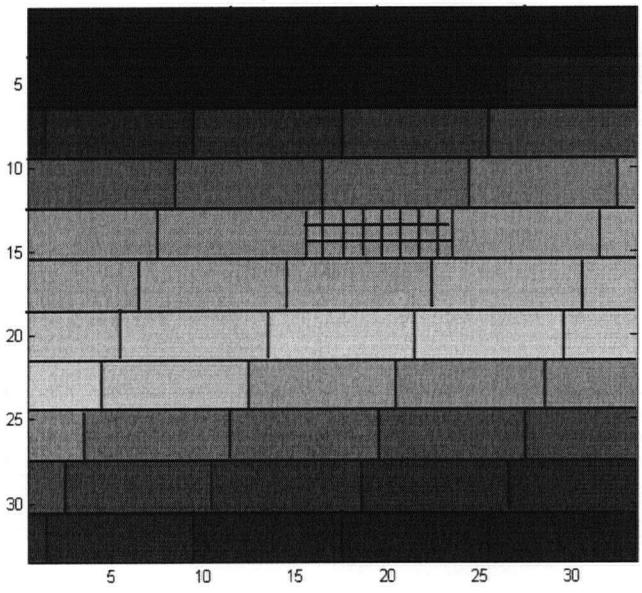

Figure 4-1 Grouping of the spots into plate groups on one subgrid of a slide, the same grouping repeats on each subgrid, the larger grid shows the plates and the smaller grid shows the spots

The reason that we are interested in this is that apparently sometimes during the preparation of probes one of the steps goes wrong in either one of four 96_well plates or all four of them in each round of PCR. The latter is the reason that some groups of spots are observed that all look very dim in the image. If the PCR was unsuccessful in all four of the plates of the tetrad, the pattern repeats in all subgrids, but if it is unsuccessful in just one of the plates, then the pattern repeats in every one out of four subgrids (figure 4-2).

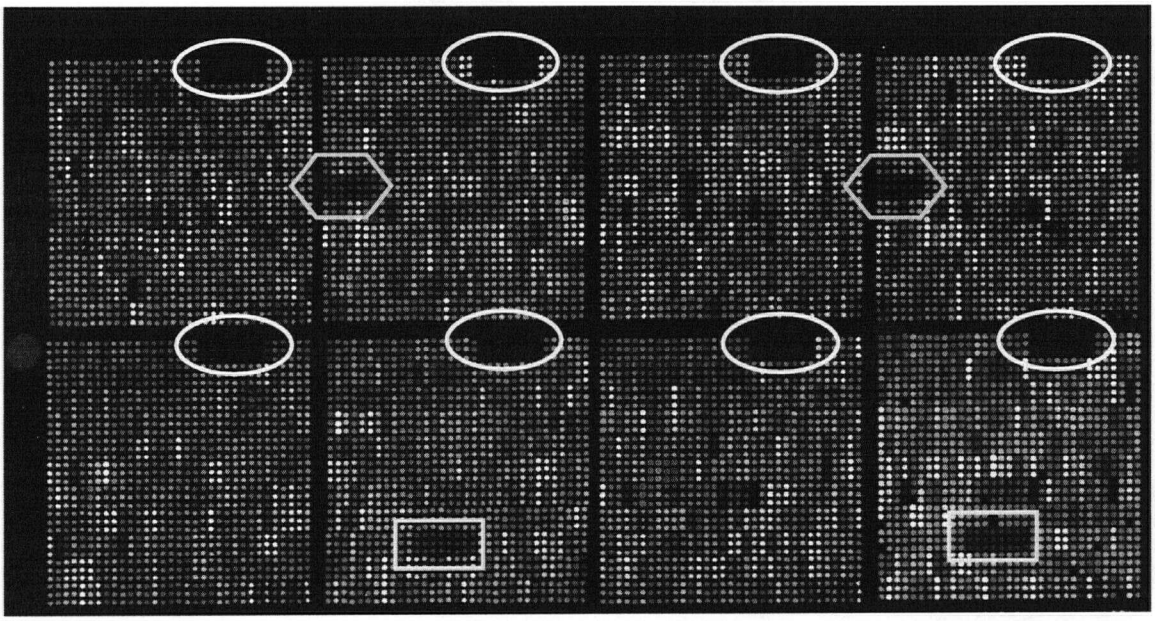

Figure 4-2 Printing artifacts, an example of "bad" plates

4.3.2 Background Contamination

Background defects can appear in various parts of an image due to a variety of reasons. Such an artifact can influence the signal level of a large number of spots located in the contaminated area (Figure 4-3).

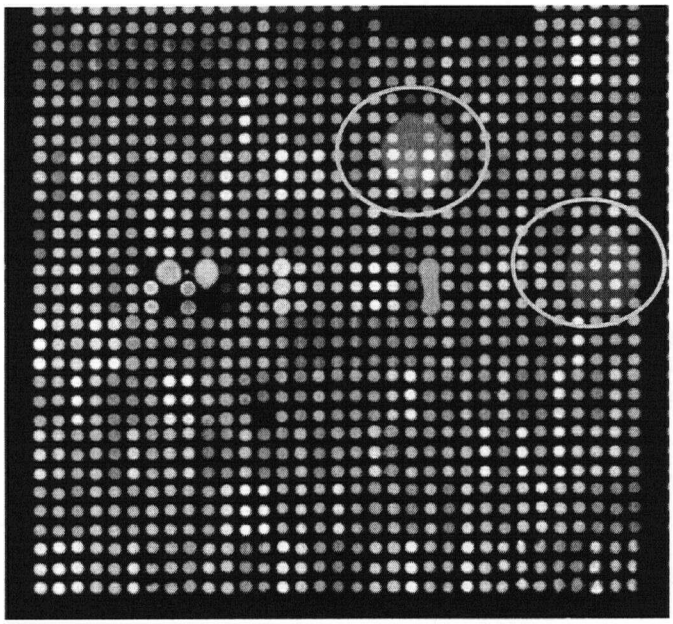

Figure 4-3 Example of background contamination

4.3.3 Signal Contamination

A non-homogenous distribution of material within the spot also indicates a problem which can be caused by of one of the following: an external object on the surface of the array, dye separation and/or clumping during hybridization, scratches on the surface of the slide that are formed at the time of washing the slides, etc. (figures 4-4 and 4-5)

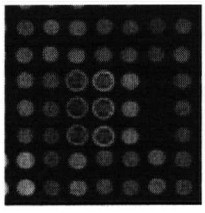

Figure 4-4 Dye separation, an example of signal contamination

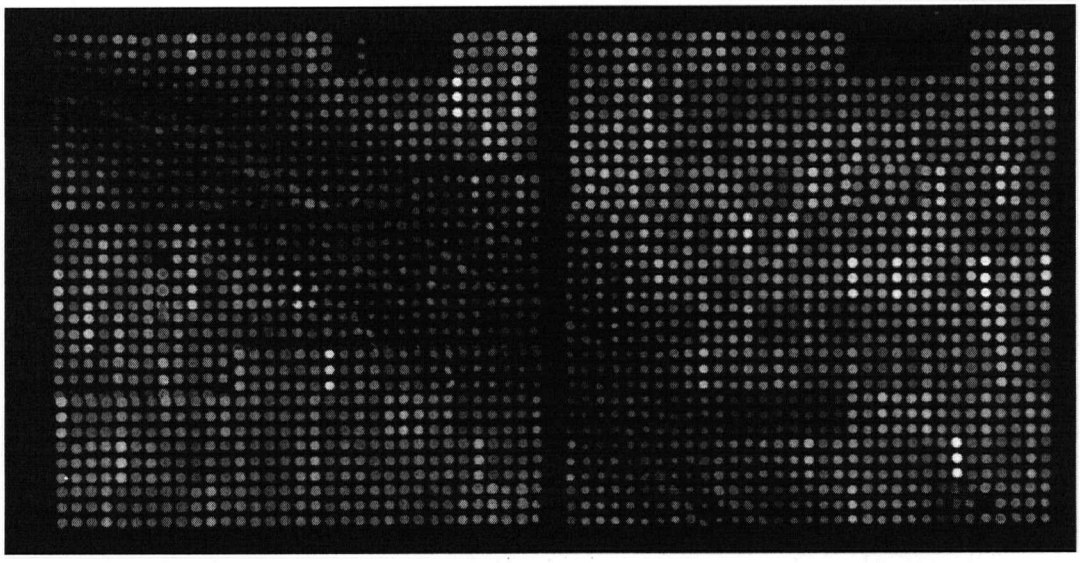

Figure 4-5 Example of scratch on the surface of the slide

4.3.4 Saturation

Saturation occurs when spot pixel intensity values exceed the detection range of the scanner, in our case the CCD camera detector. Saturation of bright spots should be avoided by properly setting the exposure time (the amount of time that a pixel of the CCD is exposed to the emission light from the excited sample). Yet saturation still may occur in some spots of amplified sequences or spots that contain contaminations in the

form of a bright artifact such as dye clumps and dust, since these typically result in a strong intensity value.

When saturation happens the measured signal is less than its true value and this in turn will result in incorrect signal ratios.

4.3.5 Spurious Artifacts

Even when the spot is not compromised during the slide preparation, hybridization and scanning, the data extraction phase can go wrong for several reasons and this makes the spot measurements unreliable. Since each slide contains tens of thousands of spots, it is not possible to manually check the results of the addressing and segmentation phases to see if the foreground and background are correctly identified by the automated image analysis part.

Some of the artifacts that can be introduced by the data extraction step are:

When there is a dust particle very close to the foreground of the spot patch or on the foreground, the segmentation can be fooled and misidentify the dust as the foreground.

If for some reason, during the printing of the slide, the spot's center location is shifted from its expected location, if the offset is too big, the addressing algorithms can not correctly locate the center of the spot and this effects the results of the segmentation process and as a result the wrong foreground and background values are found.

4.3.6 The Problem of Low Intensity Spots

The intensity of the spots has been found to be a significant factor affecting the accuracy and repeatability of the data and data variation increases for the low intensity spots.

This problem has been discussed in the literature from different points of view.

A weak spot, on a gene expression array, may represent an unexpressed gene. According to the linear model of measurement error for gene expression of Rocke and Durbin, [25], the intensity measurement y for each gene in each channel is modeled by the equation $y = \alpha + \mu e^{\eta} + \varepsilon$ where μ is the expression level in arbitrary units, α is the mean intensity of unexpressed genes, ε is the additive error prominent at low expression levels and η is the multiplicative error and is noticeable mainly for highly expressed genes. So, in this

context, weak spots should be treated in a way that the genes expressed at low but stable levels can be distinguished from those unexpressed [26]. Apparently this is not applicable to array-CGH data as there are no absent sequences in the control sample.

A spot whose intensity is below the detection limit is considered too weak to give reliable measurements. Microarray scanners are fluorescence imaging systems. The criterion for measuring the detection limit of a fluorescent imaging system is signal to noise ratio (SNR), which is the ability of the instrument to detect a signal (in this case, fluorescent dye bound to the arrayed biomolecules) above background (the microarray slide). SNR is typically expressed as the background subtracted signal divided by the variation in the background:

$$SNR = (signal - background)/s.d.\ \ of\ \ background$$

If SNR is accepted as a detection limit, the next step would be to choose a minimum acceptable SNR. According to [4] a commonly accepted criterion for the minimum signal that can be accurately identified is the sample value for which the signal is three times greater than the background noise, that is $SNR=3$.

The problem with this criterion is that the background inside each microarray spot might not be the same as the background around each spot and the true signal may start from a level lower or higher than the neighborhood background level. This makes the defined SNR incorrect.

The additive error from the background correction becomes significant for low intensity spots. The standard approach to background correction is to subtract an estimate of the background intensity from the intensity measured in the spot (the foreground intensity). This approach can cause problems when the foreground intensity is low, for example, if it is of the same magnitude as the background intensity. This situation will cause estimates of the expression ratio to become very noisy. Some methods have been proposed to improve the background correction [27] but they all depend on the assumption that the background estimated from around the spot is the same as the background inside the spot.

In inspecting the data from our database, the variation of the $\log_2$ ratios of the triplicate spots in each slide was found to increase as the intensity of the spots decreases (see figure 4-6).

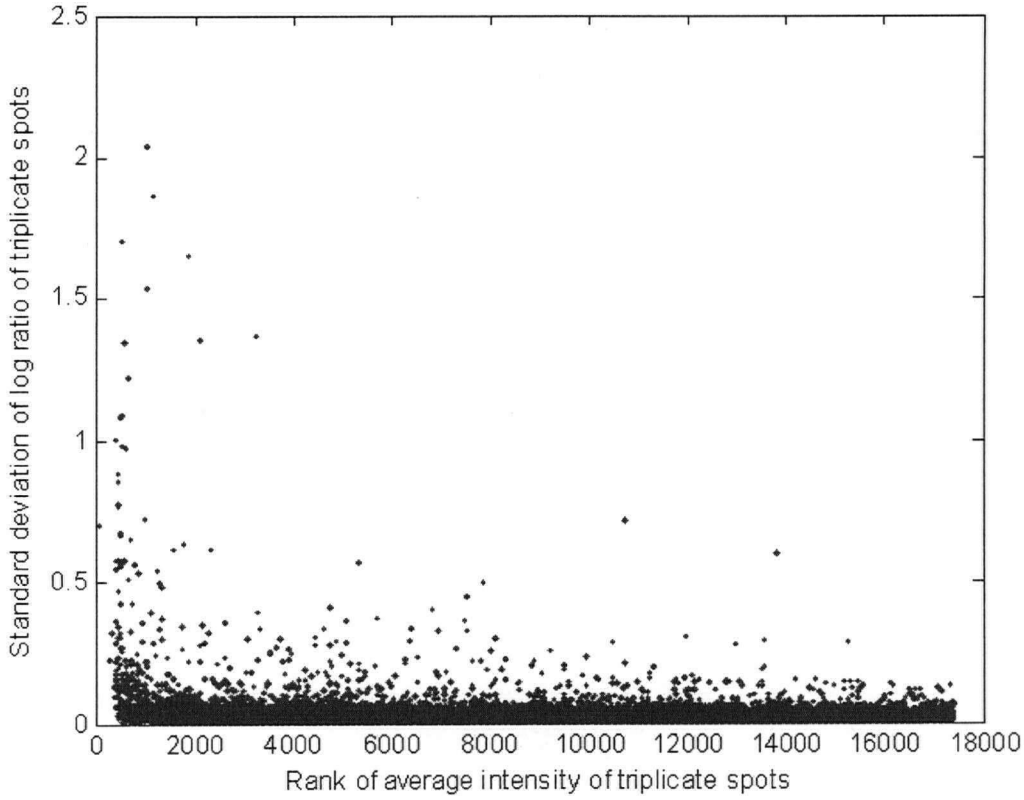

Figure 4-6 Plot of s.d. of the $\log_2$ ratios of triplicate spots sorted versus the average intensity of the triplicates

We believe that if the lower SNR or lower intensity of the spot affects the repeatability of the log ratio of the spot, it will be reflected in one of the quality features associated with the spot especially those who measure the consistency of pixel information from both channels. So, our approach to this problem was to include such features in the feature list and let the linear discriminant function analysis find the features that best discriminate the unstable low intensity spots. We believe this works better than choosing one arbitrary threshold for SNR or the intensity of the spot and filtering all the spots below the threshold.

4.4 Method of Quality Filtering

Our approach to the problem of filtering out the low quality spots is to extract, for each spot, a set of features indicating the characteristics of the spots. A binary decision tree consisting of linear discriminant functions is constructed to classify the spots into two groups of "good" and "bad" spots. In sections 4.4.1 through 4.4.3, the descriptions of the classifier, the feature extraction and the training set are given. In section 4.5, the constructed binary decision tree is presented and its performance in classifying the spots is evaluated. The discussions of the results are given at the end.

4.4.1 Classification

The classifier software program that was used in this study is a custom in house software, named CELL CLASSIFY, which was originally developed for classifying the images of cell nuclei for detection of abnormal cells.

CELL CLASSIFY is a general classification software program used to identify particular object types. When loaded with object features and optionally object images, it can be used to manually or automatically sort the objects into classification groups. The program can be used to explore the data graphically and process it by binary decision trees.

The CELL CLASSIFY requires an image and a feature file as input. The images from a microscope slide are recorded in a proprietary image file format (*.img). Image files constitute the primary data from which feature files are derived. Within the file, the entire information describing one image makes up a complete and independent record. This record contains: 1) The focused image, 2) The calculated mask (A segmentation process is applied to determine the object boundary and an ROI (region of interest) mask is created.), and 3) An image file header (comprising the file name, class, normalization coefficient, diagnosis code, and slide coordinates). The image file usually contains numerous records, each one corresponding to a different image.

Numerical values from each image file, known as features, are extracted and recorded in a format known as the binary feature file (*.fb?). The question mark in the file extension identifies which version of features was calculated.

The image file and its corresponding feature file are loaded in the CELL CLASSIFY program. The main windows and graphs that can be used for the purpose of visualization or classification are:

Groups: The user assigns individual images into categories, or Groups. The program assigns a unique identification color to each group which correlates with the information in other displays. The Groups window shows groups numbers from zero to nineteen and the number of images in each group.

Features: This window presents the feature names and feature values associated with the selected image. Users select the features to label the graphs and images during data display and analysis.

Image display: This window presents a gallery of the image records making up the file. It may display the images, or the masks in the cell's group color, according to user requests. Below each image an identification label describes the image that displays the assigned feature value. In addition to this image display, a user can view a magnified version of each individual image and its corresponding mask in a separate window.

Histogram display: this window displays the histogram of the selected feature. Dual and multiple histograms can also be used.

Scatter plot display: This window displays a two parameter scatter plot of two assigned features.

Operators: the binary operator formats used for object classification by CELL CLASSIFY. The selections include: 1) **Threshold operator**: allows the user to classify the images into two groups based on the value of the corresponding feature. 2) **Linear discriminant operator**: A Linear Discriminant Function is a binary operator used as an object classifier. From the defined list of features, Stepwise Selection only uses the features that provide the best discrimination. F tolerances (F-enter and F-remove) can be set by the user or the default values will be used. And 3) **Neural network operator**: this option was not used in this study.

Figure 4-7 shows a screen shot of the CELL CLASSIFY program.

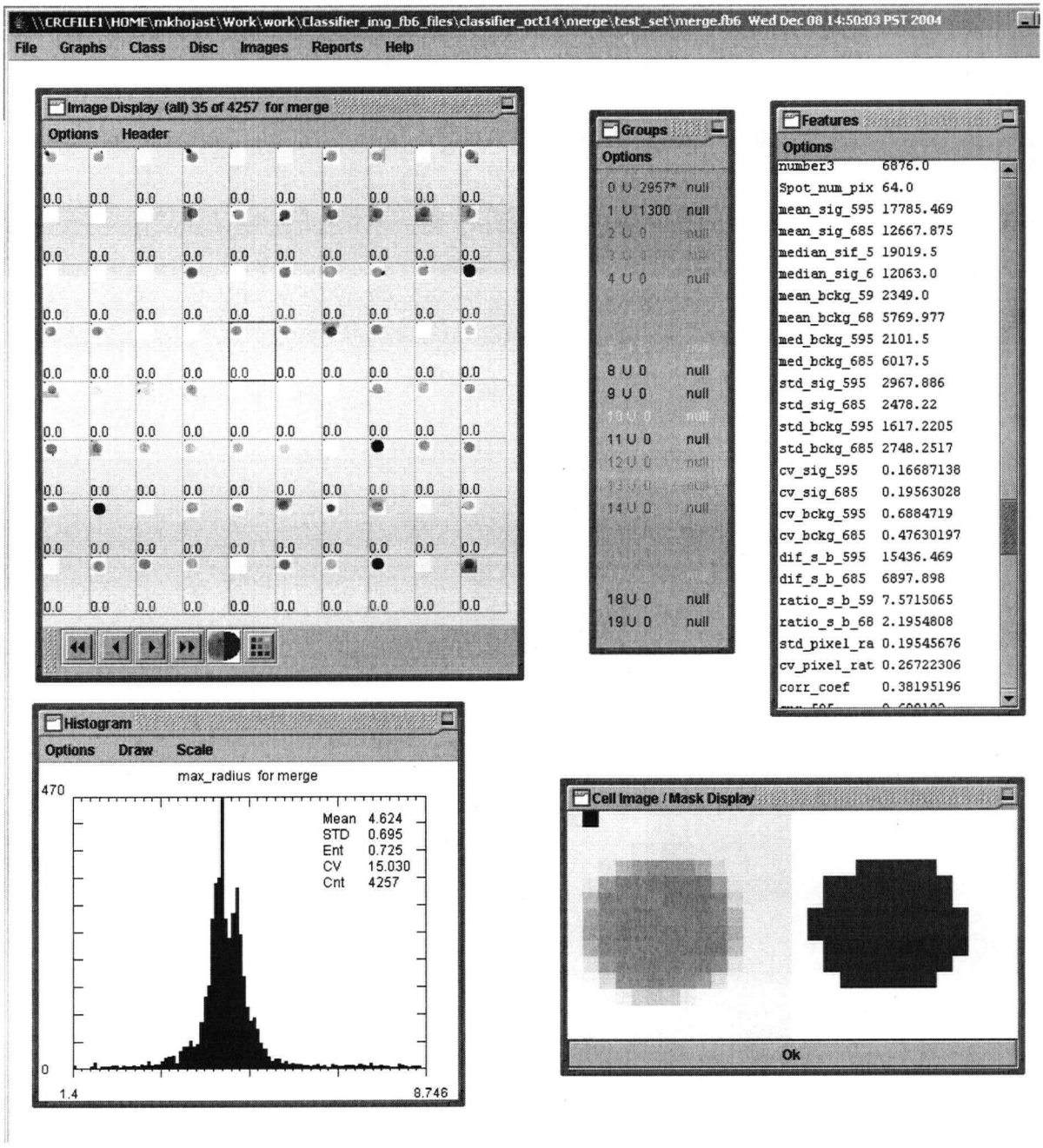

Figure 4-7 a screen shot of the "CELL CLASSIFY" program

4.4.1.1 Stepwise Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a commonly used technique for data classification and dimensionality reduction. Probably the most common application of LDA is to include many measures in the study, in order to determine the ones that best discriminate between groups. Put in another way, we want to build a model of how we

can best predict to which group a case belongs. In stepwise discriminant LDA, a model of discrimination is built step-by-step. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the discrimination between the group means. That variable will then be included in the model, and the process starts again. The stepwise procedure is "guided" by the respective F-enter and F-remove values. The F value for a variable indicates its statistical significance in the discrimination between groups, that is, it is a measure of the extent to which a variable makes a unique contribution to the prediction of group membership. A linear discriminant equation, $D_i = a + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$, where $X_1, X_2, \ldots X_p$ are the features, $b_1, b_2, \ldots, b_p$ are the corresponding coefficients, a is the a constant term, and $i = 1,2$ refers to each group, is constructed such that the two groups differ as much as possible on D_i. That is, the weights are chosen so that were you to compute a discriminant score (D_i) for each subject and then do an ANOVA (Analysis of Variance) on D, the ratio of the between groups sum of squares to the within groups sum of squares is as large as possible. In this analysis it is assumed that the data for the variables (features) represent a sample from a multivariate normal distribution [51].

4.4.2 Feature extraction

In addition to reporting the estimate of the foreground and background intensity of spots, most microarray image analysis software, generate some basic quality measures such as the size of the spot, signal to noise ratio (with different definitions), standard deviation of the foreground and background, etc. But if we want to define and use additional quality measures we will need to have access to the actual pixel values of the foreground and background regions of each spot. Unfortunately none of the software packages provide this information. This necessitated that we redo the segmentation of the microarray images into foreground and background of spots as it is required for the feature extraction step.

Before performing the segmentation, the microarray images were converted to the "image" file format (refer to section 4.4.1 for a description of image file format). Using the coordinates of the centroid of the spots reported in the output file of the SoftWorx

tracker software, a rectangular patch around each spot was taken and all the image pixel intensities in that patch were recorded in the image file in binary format. Each image data block is then followed by a binary image of the same size which is the mask of the foreground and background image obtained by segmentation. In addition each image block is preceded with a header that has some information about the block such as its top left corner position in the original microarray image and it also has some empty place. The classifier program fills the empty space with the information of the group and class of each "object" (figure 4-8).

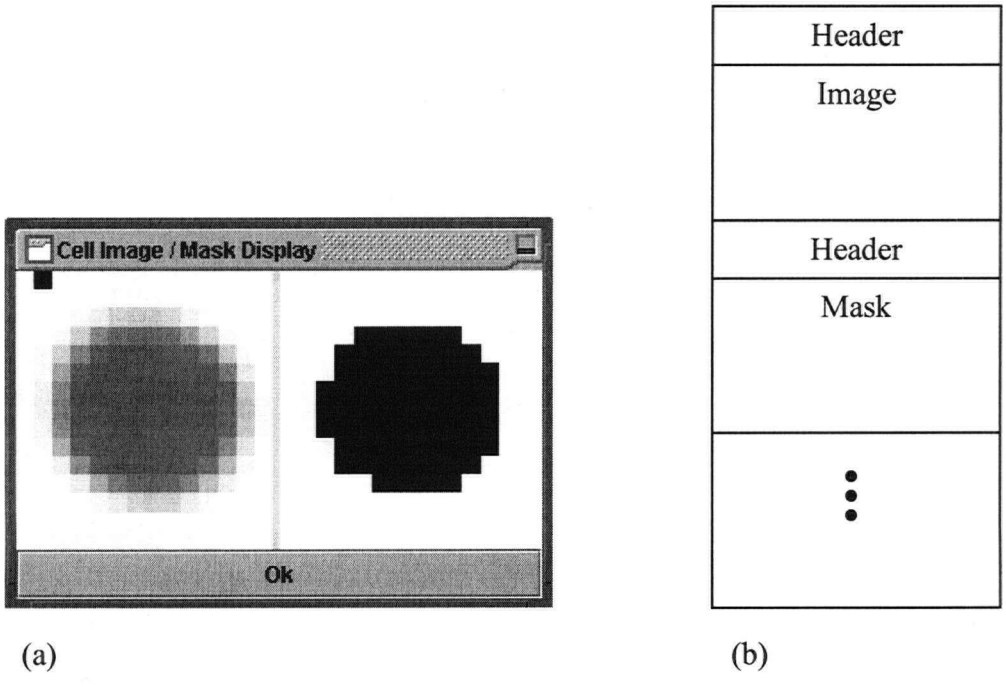

Figure 4-8 (a) a spot image and its mask, (b) format of data in the img file

4.4.2.1 Segmentation

As previously reported an adaptive segmentation approach has shown to work satisfactory for this purpose [10]. We chose "seeded region growing" segmentation [28]. SRG segmentation requires the specification of starting points, or seeds, i.e. the location and number of features should be determined beforehand. In microarray images the number of features (spots) is known exactly and the approximate locations of the spot centers are determined at the addressing stage. In this study, we didn't perform the

addressing stage instead we used the output of the image analysis software that reports the coordinates of the center of each spot.

The SRG algorithm

Segmentation of spots into foreground and background was carried out using the seeded region growing (SRG) algorithm of Adams and Bischof [28].

This method works as follows. A number of seeds are provided as input to the algorithm. These are groups of pixels which serve as starting points for a region growing process. Seeds may consist of only a single pixel or they can be of any size and do not need to form a connected set.

After specification of seeds, the algorithm proceeds by growing all the foreground and background regions simultaneously until all pixels in the image have been allocated to one of the regions. At each stage, all pixels which are as yet unallocated, but which have at least one neighbor which has already been allocated, are considered for allocation. Out of all these region-neighboring pixels, the algorithm selects the one whose pixel value is nearest (in terms of absolute grey-level difference) to the average of the pixel values in the neighboring region. The process repeats until all pixels have been allocated. Pixel queues are used to optimize the efficiency of the procedure.

For this application of SRG, the algorithm is applied to each image block in the image file.

The scanner generates two registered images for each microarray slide (one image for each channel). Before the segmentation can be performed, a combined image needs to be formed. One choice for the combined image can be the sum of the two images from two channels; these images are named R for the Cy5 channel and G for the Cy3 channel. But in order to prevent the combined image from being dominated by one of the channel images, a better way is to scale both of the images to the same scale as follows, provided that the majority of spots do not show any differential gene expression or copy number change.

So median pixel values are calculated for each image, m_R and m_G and the combined image is computed as:

$$(R + (m_R / m_G)G)/2$$

where R is the image of the cy5 channel and G is the image of the cy3 channel and m_R and m_G are the median pixel values of R and G respectively.

Then the segmentation algorithm is applied to each block of the combined image that corresponds to a block of the image in the "image" file so that the foreground mask is generated for each spot separately.

The next step is to construct the foreground and background seeds for each spot. The image analysis software has already adjusted the predetermined location of the spots from the template to the actual position of the spots. The coordinates of the centeroid of the spots are then taken from the output file of the SoftWoRx Tracker software. A patch around the center pixel including the 8-connected neighbors of the pixel is used as the foreground seed. As for the background seeds the pixels in the four corners of the rectangular patch around the spot and their immediate 8-connected pixels are used (figure 4-9).

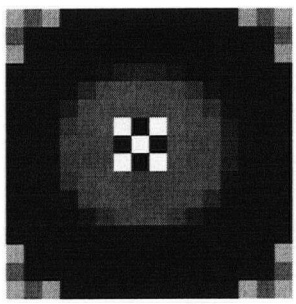

Figure 4-9 Spot image, the foreground seeds marked in the middle and the background seeds marked in the corners

4.4.2.2 Features

The reliability of the ratio measurement of a spot is affected by multiple characteristics or "features" of the spot images. In this study a variety of shape, texture, and foreground and background distribution features are generated for each spot so that the linear discriminant function analysis can find the features that best discriminate between the two groups of interest, "good" and "bad" spots.

Table 4-1 shows the name and description of the features specifically designed for the microarray spots in this study. In this table, for each spot, N is the total number of foreground pixels , M is the total number of background pixels, $\{r_i\}$ and $\{g_i\}$, i=1,2,…,N, are the foreground pixel intensities of cy5 and cy3 channels, respectively, and $\{br_i\}$ and

$\{bg_i\}$, i=1,2,…,M, are the background pixel intensities of cy5 and cy3 channels. Also in this table, s.d. is the standard deviation and c.v. is the coefficient of variation (standard deviation divided by the mean).

Feature name	Description
mean_sig_595	$mean\{g_i\}, i = 1,2,\cdots,N$
mean_sig_685	$mean\{r_i\}, i = 1,2,\cdots,N$
median_sig_595	$median\{g_i\}, i = 1,2,\cdots,N$
median_sig_685	$median\{r_i\}, i = 1,2,\cdots,N$
mean_bckg_595	$mean\{bg_i\}, i = 1,2,\cdots,M$
mean_bckg_685	$mean\{br_i\}_{i=1,2,\cdots,M}$
med_bckg_595	$median\{bg_i\}_{i=1,2,\cdots,M}$
med_bckg_685	$median\{br_i\}_{i=1,2,\cdots,M}$
std_sig_595	$s.d.\{g_i\}_{i=1,2,\cdots,N}$
std_sig_685	$s.d.\{r_i\}_{i=1,2,\cdots,N}$
std_bckg_595	$s.d.\{bg_i\}_{i=1,2,\cdots,M}$
std_bckg_685	$s.d.\{br_i\}_{i=1,2,\cdots,M}$
cv_sig_595	$c.v.\{g_i\}_{i=1,2,\cdots,N}$
cv_sig_685	$c.v.\{r_i\}_{i=1,2,\cdots,N}$
cv_bckg_595	$c.v.\{bg_i\}_{i=1,2,\cdots,M}$
cv_bckg_685	$c.v.\{br_i\}_{i=1,2,\cdots,M}$
dif_s_b_595	$mean\{g_i\}_{i=1,2,\cdots,N} - mean\{bg_j\}_{j=1,2,\cdots,M}$
dif_s_b_685	$mean\{r_i\}_{i=1,2,\cdots,N} - mean\{br_j\}_{j=1,2,\cdots,M}$
ratio_s_b_595	$mean\{g_i\}_{i=1,2,\cdots,N} / mean\{bg_j\}_{j=1,2,\cdots,M}$
ratio_s_b_685	$mean\{r_i\}_{i=1,2,\cdots,N} / mean\{br_j\}_{j=1,2,\cdots,M}$
std_pixel_ratio	$s.d.\left\{\frac{r_i}{g_i}\right\}_{i=1,2,\cdots,N}$
cv_pixel_ratio	$c.v.\left\{\frac{r_i}{g_i}\right\}_{i=1,2,\cdots,N}$
std_pixel_ratio_G	$s.d.\left\{\frac{g_i}{r_i}\right\}_{i=1,2,\cdots,N}$

Feature name	Description
cv_pixel_ratio_G	$c.v.\left\{\frac{g_i}{r_i}\right\}_{i=1,2,\cdots,N}$
std_pix_ratio_b	$s.d.\left\{\frac{r_i - median\{br_j\}_{j=1,2,\cdots,N}}{g_i - median\{bg_j\}_{j=1,2,\ldots,N}}\right\}_{i=1,2,\cdots,N}$
cv_pix_ratio_b	$c.v.\left\{\frac{r_i - median\{br_j\}_{j=1,2,\cdots,N}}{g_i - median\{bg_j\}_{j=1,2,\ldots,N}}\right\}_{i=1,2,\cdots,N}$
corr_coef	$Pearson\ correlation\ coefficient\ of\ \{r_i\}\ and\ \{g_i\}, i=1,2,\cdots,N$
snr_595	$\frac{mean\{g_i\}_{i=1,2,\cdots,N} - median\{bg_j\}_{j=1,2,\cdots,M}}{s.d.\{bg_i\}_{j=1,2,\cdots,M}}$
snr_685	$\frac{mean\{r_i\}_{i=1,2,\cdots,N} - median\{br_j\}_{j=1,2,\cdots,M}}{s.d.\{br_i\}_{j=1,2,\cdots,M}}$
sat_percent	$\frac{number\ of\ spot\ pixels\ that:\ r_i > 2^{16}-256\ or\ g_i > 2^{16}-256}{N}$
Spot_Peak_Int_G	$\max\{g_i\}_{i=1,2,\cdots,N}$
Spot_Peak_Int_R	$\max\{r_i\}_{i=1,2,\cdots,N}$

Table 4-1 Feature names and their descriptions

In addition to these features that have been specifically designed for microarray spots, a variety of morphological and texture features that have been originally developed for classifying cells of different types in the Cancer Imaging Department of BC Cancer research center are also used. The description of these features can be found in [52] and [53].

4.4.3 Training the classifier

After extracting all the features, an appropriate number of features that can best discriminate the "good" and "bad" spots are to be determined with the use of a stepwise LDA (Linear Discriminant Analysis) implemented in the "CELL CLASSIFY" program. In order to do that a training set and a testing set were assembled.

In order to build the training set and the test set, the data from 8 slides H526-1 through H526-8 were used. The standard deviation (s.d.) of the $\log_2$ ratios of triplicates was calculated for all the eight slides in this experiment and all the spots with s.d. higher than

0.075 were chosen in the first step. This was done in order to reduce the number of potential candidate spots for manual classification because with 52272 spots on each slide manual classification of all the spots would take a prohibitively long time. We believe that this would not affect the accuracy of classification because in the spots chosen in this way, we observed all types of defects that we would normally observe in spots from a microarray so by limiting the number of potential cases nothing is missed.

In the next step, the spots in this set were manually classified into two groups of "good" and "bad". Then the data from slides H526-1 to H526-4 were merged and used as the training set and the data from slides H526-5 to H526-8 were merged and used as a totally independent test set.

The total number of spots in the test set is 3303 from which 1840 spots are manually classified as bad and 1463 spots are classified as good. The total number of spots in the test set is 4257 from which 1300 spots are classified as bad and 2957 are classified as good.

4.5 Results

After examining the training set data, it was concluded that a single discriminate function was not able to discriminate between the two groups of objects, i.e. the "good" and "bad" spots. So, the classification was done in several steps:

It was observed that if from the three replicate spots on a slide; one or more of them have saturation percentage higher than zero, the standard deviation of the ratio measurement of the triplicates is higher than average. Therefore, in the manual classification of spots of the training set, spots with saturation percentage higher than zero were classified as "bad" spots.

The first step in classifying the spots into two groups of "good" and "bad" was therefore filtering away all the spots whose saturation percentage (percentage of the saturated pixels of the foreground) is higher than zero in either channel. A threshold operator was used for this purpose.

The segmentation method that was used to segment image pixels into foreground and background was an adaptive segmentation. So the masks of foreground regions of spots are connected objects of arbitrary shapes. Spots with masks that are non-circular are

obviously somehow unreliable as the irregularity in shape indicates the presence of an artifact. Spots with masks that are circular in shape, however, are not all "good" spots. The non-uniformities of the intensity of the spot surface might not affect the segmentation result if it is not strong enough. Spots with separated dyes or scratches on their foreground surface that have circular masks have been observed repeatedly in the training set.

Based on these observations, the classification of spots into two groups of "good" and "bad" was done in two steps. First step is to classify the spots into two groups based on the shape of their mask. The first group consists of spots with circular masks and the second consists of spots with masks of irregular shape.

The second step is to take the first group of the previous classification which consists of circular spots and classify the spots based on the texture of their foreground and background intensity. These two steps are explained in more detail below.

To train a classifier to find the spots with irregular shapes, a new training set was needed with "good spots" in one group and "bad spots" with irregular shapes in the other group. The training set spots were manually classified to assemble this new training set.

The linear discriminant function that performs this classification was constructed from morphological features only.

Throughout the experiment we found that a function with about 10 features is able to do the job. Table 4-2 shows the selected features and their coefficients in the linear discriminant function.

Feature	Coefficient	Feature Description
16 mean_radius	4.9892	mean value of the length of the objects's raidal vectors from the object centroid to its 8-connected edge pixels
17 max_radius	-5.9938	max values of the length of the objects's raidal vectors
18 var_radius	9.9991	variance of the length of the objects's raidal vectors
19 sphericity	48.1313	a shape measure, maximum equlas one for a circular object, mean_radius/max_radius.
20 eccentricity	-10.4536	a shape measure, estimate of the ratio of the major axis to a minor axis of the best fit ellipse which best describes the object and gives minimal value of 1 for circles.
22 inertia_shape	27.4646	a measure of roundness of an object, calculated as the moment of inertia of the object mask normalized by the area squared, to give the minimal value of 1 for circles
26 freq_high_fft	-0.0152	an estimate of fine boundary variation, measured as the energy of the high frequency Fourier spectrum (from 12th to 32nd harmonics) of the object's radial function
27 harmon01_fft	0.4556	estimate of boundary variation, calcualted as the magnitude of the Fourier transfrom coefficients of the objct radial function, for each harmonic
29 harmon03_fft	-0.2133	as above
35 harmon09_fft	0.5308	as above

Table 4-2 Morphological features chosen by the discriminant function analysis and their coefficient in the function

Next step was to find the circular spots that are somehow defected and unreliable. At this step, the morphological features are not included in the analysis as the spots to be classified have already passes the shape requirements.

The linear discriminate function analysis implemented in the CELL CLASSIFY program takes as input the "F-enter" and "F-remove" parameters of the stepwise linear discriminate analysis algorithm as well as the maximum number of features to be used. In order to find the best discriminant function the F-enter and F-remove parameters were set to their default values (F-enter of 4.0 and F-remove of 3.996). The number of features was varied from 1 to 30. Each number of features generates a linear discriminant function. The discriminant function was then applied to the test set to calculate the accuracy of the classifier.

It should be noted that the **overall accuracy** of the classifier is defined as the number of correctly classified objects divided by the total number of the objects. The **true positive** percentage is defined as the number of "good" spots that are classified as "good" spots

divided by the total number of good spots. The **true negative** percentage is the number of "bad" spots correctly classified as "bad" spots divided by the total number of bad spots. After 30 features, increasing the number of features allowed, did not change the number of selected features. So the experiment stopped at that point. Figure 4-10 shows the plot of accuracy of the classifier versus the number of features for both test set and training set. As the plot shows for the test set the accuracy first increases with increasing the number of features and then it gets to a constant level and after that it decreases which is because of over training of the classifier. So the best accuracy was obtained using 25 features and the corresponding function was chosen as the optimum classifier. Table 4-3 shows the features selected and their coefficients.

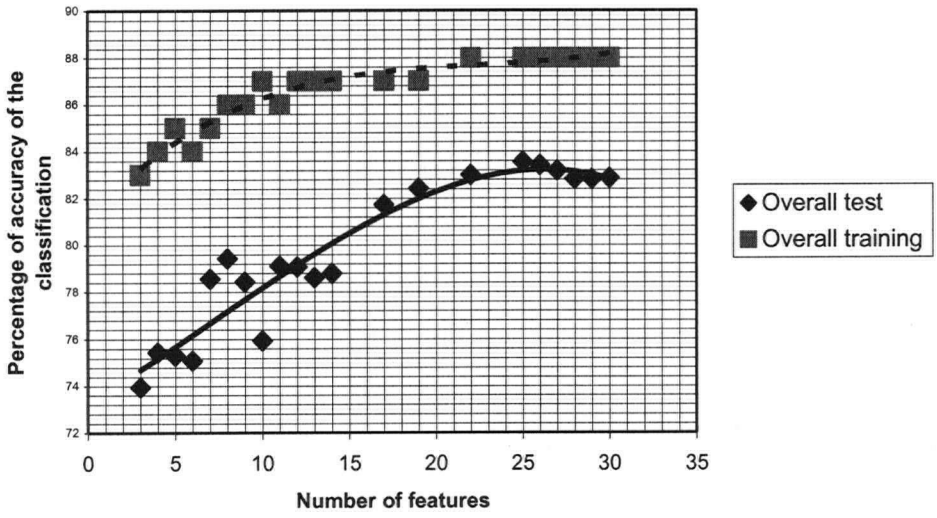

Figure 4-10 Accuracy of the classifier on the test set and training set using different numbers of features

Features	Coefficients	Feature Description
Medium DNA area	-1.6330	See references [52] and [53] for descriptions
Low DNA compactness	1.1410	
Entropy	-1.9339	
Energy	-5.2412	
Number of local maxima of the object intensity function	-26.8945	
Number of local minima of the object intensity function	7.4589	
Center of gravity	-4.7090	
Fractal dimension	-1.0495	
Average run length	-0.0777	
Minimum run length percent	-6.6135	

Features	Coefficients	Feature Description
std_sig_685	0.0005	Refer to Table 4-1
cv_sig_685	-14.1689	
cv_bckg_685	3.6076	
dif_s_b_685	0.0002	
ratio_s_b_595	0.6143	
std_pixel_ratio	10.8451	
cv_pixel_ratio	-30.9444	
corr_coef	3.8938	
std_pix_ratio_b	0.2860	
Spot_Peak_Int_R	-0.0003	
mean_bckg_595	-0.0003	
med_bckg_595	0.0004	
cv_sig_595	-0.0026	
std_pixel_ratio_G	-0.0042	
cv_pixel_ratio_G	-0.0039	

Table 4-3 Texture features chosen by the LDA algorithm and their corresponding coefficients in the function

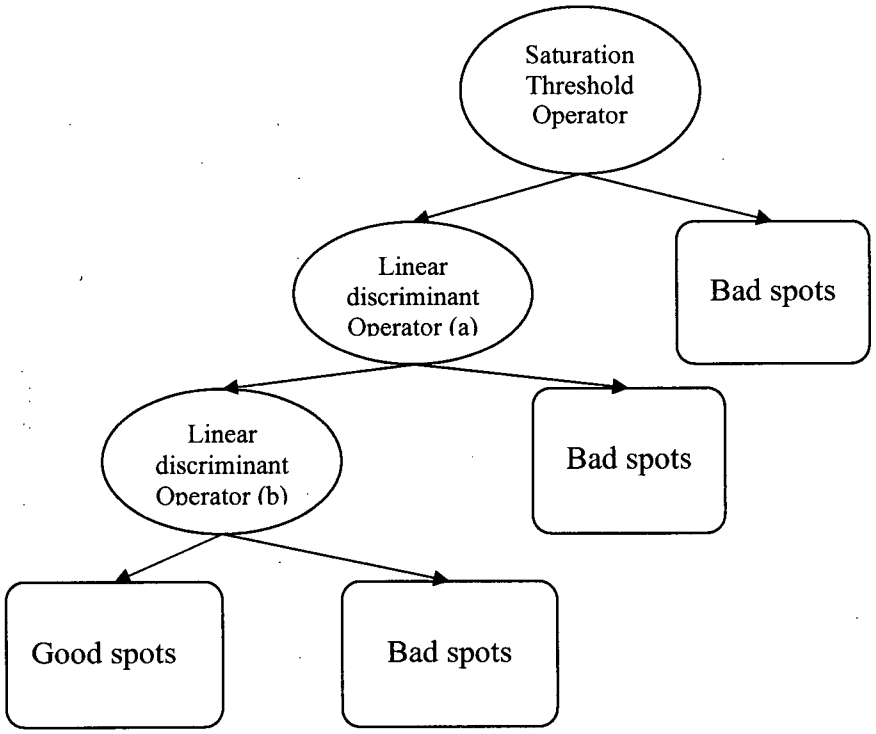

Figure 4-11 The designed binary decision tree

	true positive	true negative	over all
training set	79.50%	96.53%	89.31%
test set	91.08%	83.92%	87.50%

Table 4-4 Overall accuracy of the classifier in test and the training set

All of these steps were combined in a binary decision tree as shown in figure 4-11. Table 4-4 shows the overall results of the final classifier on the test and training set.

The classifier designed in this way, was then applied to data from each of the self-self experiments (slides MM-1 through MM-4, refer to section 3.2 for the description of these data). Use of the self-self experiments allows us to compare the $\log_2$ ratios with their expected values which are zeros. For comparison purposes, we filtered the data with the triplicate filtering method and with our classifier. For this experiment the $\log_2$ ratios of spots from each slide were normalized using the method that will be described in chapter 5.

When using the classifier, the constant term in the two discriminant functions needed to be adjusted for each data set. So we manually adjusted the constant term to get the appropriate separation of the good and bad spots. In fact, the values of the two discriminant functions can be considered as a quality score. By changing the constant term, the threshold of the quality score is changed. The threshold can not be set to a fixed value for all the slides as the distribution of the discriminant function values is different for different slides so the threshold should be adjusted according to that (when using the triplicate filtering method, the same threshold adjustment process is applied to the s.d. measure of the triplicates).

Table 4-5 shows the percentage of the standard deviation of the normalized $\log_2$ ratios before and after low quality spot filtering relative to the standard deviation of the $\log_2$ ratios before low quality spot filtering. Table 4-6 shows the variation, calculated as the sum of the absolute value of the $\log_2$ ratios, before and after quality filtering. As these two tables show removing the low quality spots decreases the total variation of the $\log_2$ ratios by a significant value.

In terms of reducing the variation, the classifier performs as well or some times better than the triplicate filtering method. But the fact that the classifier doesn't need replicate spots should also be considered when the two methods are compared. For the triplicate

filtering method if the s.d. of the $\log_2$ ratios of the triplicates is higher than a threshold, all of them are filtered as there is no robust way of finding the outlier measurement specially when the number of replicates is small (in this case three). So with only one or two of the spots defected, all of the triplicates are filtered and this removes all the measurements of a particular clone of the array. While the classifier removes only the defected spot and with one or two of the replicate spots filtered there will still be a measurement for that particular clone. Table 4-7 shows the number of excluded clones (clones for which there are no valid measurements) using each of the filtering methods. Table 4-8 shows the percentage of the number of excluded clones divided by the total number of clones of the slides. The results show that although the methods have comparable performance in reducing the variation, the number of clones excluded by the triplicate filtering method is always higher than the other method.

	MM-1	MM-2	MM-3	MM-4
Triplicate filtering	26.18%	36.83%	49.77%	49.97%
Discriminant function	24.93%	35.91%	42.39%	47.35%

Table 4-5 The s.d. of the $\log_2$ ratios after low quality spot filtering relative to their original s.d.

	MM-1	MM-2	MM-3	MM-4
Triplicate filtering	10.79%	4.38%	1.00%	5.64%
Discriminant function	10.23%	4.13%	2.27%	3.91%

Table 4-6 The percentage of reduction in the variation of $\log_2$ ratios after low quality spot filtering

	MM-1	MM-2	MM-3	MM-4
Triplicate filtering	2752	5691	3895	5055
Discriminant function	1643	2160	425	535

Table 4-7 Number of excluded clones after quality filtering

	MM-1	MM-2	MM-3	MM-4
Triplicate filtering	15.79%	32.66%	22.35%	29.01%
Discriminant function	9.43%	12.40%	2.44%	3.07%

Table 4-8 Numbers of excluded clones relative to the total number of clones

4.6 Conclusions and Discussions

We designed a binary decision tree with a threshold operator to identify the saturated spots and two linear discriminant functions to identify the spots with irregular shapes and spots with circular shapes but defected in some other way. The former uses a set of morphological features and the latter is constructed with texture features. The linear dicriminant functions are trained using a training set and an independent test set.

We then tested the performance of the classifier on identifying the low quality spots by applying the classifier to the data from four complete slides. We used the slides MM-1 through MM-4 as a model to investigate data variability as we know from the experimental design that the ratios should be uniform across the whole slide for these slides (i.e. we have an absolute definition of truth).

The constant terms in the two linear discriminant functions are to be adjusted interactively by the experimenter by looking at the chromosome plots of log ratios and/or the scatter plots of the spot intensities.

Throughout this experiment, we demonstrated that a large part (in this study as high as 10%) of variability of ratio measurements is due to the low quality of image spots and that the designed binary decision tree can identify the low quality spots therefore reducing the variation.

Use of the proposed method for quality filtering instead of filtering the spots based on the variation of the replicate spots in a slide which is the method that is currently used in our lab has a significant practical advantage. Using the new method enables us to reduce the number of replicates per target clone to two instead of three and the SMRT array clones that currently are spotted in triplicate across two slides can be fitted into one slide and this will make the array-CGH experiment more cost, material and time effective. We compared the results of triplicate filtering method and the new method in terms of reducing the variance and showed that the two methods have comparable performances while the new method doesn't require the replicates for quality filtering. This assures us that the SMRT array clones can be safely fitted into one slide without loss of accuracy at the filtering stage.

The main concept of using discriminant function analysis to select the features that best discriminate between the "good" and "bad" spots and use of a binary decision tree to

classify the spots is applicable to single channel microarrays too. However, a new set of features specific to single channel microarray spot data needs to be defined.

CHAPTER 5 NORMALIZATION

In this chapter the issue of removing the systematic variations from the microarray data will be addressed. The process of removing the systematic variations from the microarray data is called **Normalization**.

In this chapter, a review of normalization issues (including the systematic variations, normalization methods and methods of evaluation) is first given. Then systematic variations observed in the data from the data base of this study are discussed and a stepwise normalization strategy to remove those variations is proposed. The performance of this normalization strategy is then evaluated and compared to other existing methods. The conclusions are discussed at the end.

5.1 Background

The following sections present a review of the systematic variations known to be present in the microarray data. Existing normalization methods are then summarized. The existing models of the measurement system that relate the measured fluorescent intensity to the actual amount of probe hybridized to the targets are discussed. These models try to explain the biases in the measured intensities and/or ratios. The existing methods used for evaluating the performance of normalization on the data are then presented.

5.1.1 Review of Systematic Variations

The basic assumption underlying microarray analysis is that the measured intensities for each target represent the amount of the probe hybridized to the target. Before these intensities can be compared appropriately, a number of transformations and

normalizations must be carried out in the data to eliminate unreliable or low quality measurements (previously discussed) and to adjust the measured intensities to remove or minimize the systematic variations.

The ratio of the two fluorescent signals at each spot is commonly used to infer the ratio of the DNA concentrations in the two DNA samples compared on the array. The ratio of the fluorescent signals is influenced by systematic effects from non-biological sources that can introduce biases in estimated ratios. These biases should be removed before drawing conclusions about the relative levels of DNA. The process of removing systematic effects is often referred to as **normalization**.

The use of two samples and taking the ratio of their intensities instead of absolute intensity level, automatically removes the effect of variations in the size and amount of target DNA in each spot. However associating two samples that are labeled with two different fluorescent dyes introduces "**dye bias**" into the measurements. "**Dye bias**" is a systematic error that is caused because two dyes have different characteristics. Different physical characteristics such as molecule size make the efficiency of the incorporation of the labels into the probe DNA different. The efficiency of hybridization of labeled probes to the targets may also be affected by the characteristics of the dyes. Different quantum yields and different sensitivity to heat, light, pH, etc. make the efficiency of detection different for different dyes.

As a result of the dye bias, differences observed between red and green channel fluorescent intensities for a given transcript may be due to either a true biological difference or to a systematic bias resulting from individual transcript dependent differences in efficiencies of dye incorporation and sample hybridizations.

For each sequence that is present in both test and reference samples with the same amount, the number of dye molecules that incorporate to that sequence in the test sample will be different than the number of dye molecules that incorporate to the same sequence in the reference sample. The sequence is then hybridized to its corresponding target on the array. The hybridization efficiency may not be the same for the two differentially labeled sequences so the amount of hybridized probes labeled with different dyes may be different. Finally, when the spot is scanned and the fluorescent intensity is detected, the detection efficiency would not be the same for the two different fluorescent dyes. So, the

fluorescent intensity measured from the spot corresponding to that sequence won't be the same.

If all of the differences mentioned above (the difference in the efficiency of labeling, hybridization and detection) were the same for all of the sequences in the probe, then dye bias would be the same for all intensity measurements of the microarray. This is not usually the case.

A commonly observed bias in microarray data is the dependence of the log ratio (ratio of the florescent intensities of the spot in the two channels) for each spot on the average of the fluorescent intensities of that spot in the two channels [22, 30]. This is best observed in the plot of logarithm of the ratio of the spots intensities, against the average of the log intensities of the spots. This plot is called the **M-A plot**.

The dye-bias also generally varies with spatial position on the slide. Positions on a slide may differ because of differences between the print tips on the array printer, variation over the course of the print-run, non-uniformity in the hybridization, or from artifacts on the surface of the array which affect one channel more than the other.

Differences between spots may arise from differences in print quality, from differences in ambient conditions when the plates were processed or simply from changes in the scanner settings [33].

Another potential source of bias in two-color microarray experiments, which has been less considered in the literature, is the bias that may arise from **competitive hybridization** of differentially labeled probes [9]. In the two-color microarray hybridization experiments, differentially labeled DNA sequences from the test and reference that correspond to the same target spot, both **compete** for the same target. The rate of formation of the hybridized product for each spot depends on the hybridization rate constants, the amount of the unhybridized probe sequences, and the amount of the unhybridized target seqeunces. Therefore, the hybridization kinetics of the two sequences effectively determine the concentration of the final hybridized products. For example, if one of the two differentially labeled sequences hybridizes to the target sequence at a faster rate than the other labeled sequence, the ratio of the product signals will not be equal to the ratio of the initial amounts of the sequences in the probe.

5.1.2 Models of Measurement System

In this section different existing models that relate the intensity measurement to the actual amount of probe DNA hybridized to the target DNA are summarized. Note that there are no models specially developed for CGH microarrays and the following models have originally been developed for gene expression microarrays.

Consider an experiment with clones $i = 1,2,\ldots,I$ from DNA samples $c = 1,2,\ldots,C$. Let $X_{c,i}$ be the true DNA level for clone i in channel c. Let $Y_{c,i}$ be the corresponding observed DNA level. For spectrally well separated microarray experiments, the measurement functions for each channel are assumed to be separable equations. We then have that:

$$Y_{c,i} = f_{c,i}(X_{c,i})$$

where the measurement function $f_{c,i}(\cdot)$ is unknown. This is a general model that includes the measurement noise and systematic biases. The measurement function in general is dependent on the clone and on the channel.

For two-color microarray experiments ($c = 1,2$) we refer to two channels as R and G so we have:

$$Y_{R,i} = f_{R,i}(X_{R,i}) \text{ and } Y_{G,i} = f_{G,i}(X_{G,i})$$

A linear measurement function means that $f_{c,i}(\cdot) = b_c$ where b_c is the scale factor for channel c. If the measurement function was linear, then we would have:

$$Y_{R,i} = b_R \times X_{R,i} \text{ and } Y_{G,i} = b_G \times X_{G,i}$$

it then follows that the observed log ratios are:

$$\begin{aligned} M_i &= \log_2(Y_{R,i}/Y_{G,i}) = \log_2(X_{R,i}/X_{G,i}) + \log_2(b_R/b_G) \\ &= \log_2(r_i) + M_{bias} \end{aligned}$$

where r_i is the true ratio and M_{bias} is the constant bias. So if the measurement function was linear the bias of the log ratios would be additive and constant.

Obviously the linear measurement function doesn't explain the nonlinearities in M-A plots.

Rocke and Durbin [25] applied a two component model for analytical methods which was introduced by Rocke and Lorenzato [40] to gene expression microarray signals.

The following two component model was introduced by Rocke and Lorenzato [40]:

$$y = \alpha + \beta\mu e^{\eta} + \varepsilon$$

where y is the response of the measuring apparatus at concentration μ. $\eta \sim N(0, \sigma_\eta)$ and $\varepsilon \sim N(0, \sigma_\varepsilon)$. η represents the proportional error and ε represents the additive error. The normality of the error terms η and ε is assumed for convenience, but this is in practice often a reasonable assumption. This two-component model approximates a constant standard deviation for very low concentrations and approximates a constant relative standard deviation (coefficient of variation) for higher concentrations.

Rocke and Durbin [25] adapted that model to gene expression microarray signals. For gene expression arrays, this model is:

$$y = \alpha + \mu e^{\eta} + \varepsilon$$

where y is the intensity measurement and μ is the expression level in arbitrary units since the expression level can not actually be measured in molecular units. It can only be relatively measured because usually there is no calibration data (that is, samples of known expression levels). α is the mean intensity of unexpressed genes. ε and η are the additive error and multiplicative error terms respectively. $\varepsilon \sim N(0, \sigma_\varepsilon)$ and $\eta \sim N(0, \sigma_\eta)$. ε represents the standard deviation of the background (intensity of the unexpressed genes), and η represents the proportional error that always exists but is noticeable mainly for highly expressed genes.

The model of Cui et al, [35], is very similar to the model of Rocke and Durbin. It decomposes the measurement error terms into multiple components:

$$Y_{ik} = a_i + b_i X_{ik} e^{\eta_k + \zeta_{ik}} + \varepsilon_k + \delta_{ik}$$

The error associated with the fluorescent signal at spot k in channel i is decomposed into multiplicative and additive components. Each of these two components is in turn decomposed into a common component (ε_k and η_k) that is shared by both channels and a channel-specific component (δ_{ik} and ζ_{ik}). They assume that the distributions of all error components are symmetric with mean zero.

They attempted to simulate some of the microarray features by varying the parameters of their model. Here are their results from the simulations: "Excess variation at the low intensity end can be simulated using large channel-specific additive errors. Variation of

the log ratios at the high intensity end is controlled by the channel-specific multiplicative error. Curvature of RI plots can be generated by unequal mean backgrounds and/or unequal slopes between the two channels."

Based on the results of this simulation, the linear model (with an additive and a multiplicative error term) is promising.

5.1.3 Review of Normalization Methods

Depending on the experimental design, a decision must be made as to which set of genes to use for normalization. There are a number of considerations in this decision, such as the proportion of genes that are expected to be differentially expressed in the red and green samples and the availability of control DNA sequences.

All genes on the array may be used for normalization if only a relatively small proportion of the genes will vary significantly in expression between the two mRNA samples.

For Gene expression arrays, instead of using all genes on the array for normalization, one may use a smaller subset of genes, often called **housekeeping genes** (genes that are assumed to be expressed at a constant level al the time) that are believed to have constant expression across a variety of conditions. A limitation of housekeeping genes is that they tend to be highly expressed and hence may not be representative of other genes of interest.

An alternative to normalization by housekeeping genes is to use spiked controls or a titration series of control sequences. In the spiked controls method, synthetic DNA sequences or DNA sequences from an organism different from the one being studied are spotted (deposited) on the array (with possible replication) and included in the two different mRNA samples at equal amount. These spotted control sequences should thus have equal red and green intensities and could be used for normalization. In the titration series approach, spots consisting of different concentrations of the same gene are printed on the array. These spots are expected to have equal red and green intensities across the range of intensities [38].

Tseng et al, [22], proposed a rank invariant method to select the non-differentially expressed genes for normalization where all signals from both arrays are sorted and signals with ranks deviating by less than a threshold are included.

Yang et al, [30], correctly suggested that in case of the gene expression arrays, the use of all genes for normalization, in biological samples which show significant divergence, may not produce accurate normalized log ratios. In such instances, it would be more appropriate to normalize using the control spots alone.

After choosing the subset of genes for normalization, the appropriate normalization approach should be chosen. A number of normalization approaches have been introduced for gene expression microarray data analysis including:

1. Global normalization
2. Total intensity normalization
3. Normalization using ANOVA models
4. Shift transforms
5. Self-normalization
6. Adaptive normalization

Below we describe each of these methods.

Global Normalization: is based on two assumptions: firstly, that the center of the distribution of gene expression ratios (i.e. mean or median) on a log scale is zero. Secondly, that the systematic error makes the central line move vertically, but otherwise leaves the distribution invariant. Therefore a global normalization shifts the center of the log ratio distribution to zero. However the systematic error in microarray data is often not constant and the performance of the global normalization is not satisfactory.

Total Intensity Normalization: has been developed for the gene expression microarrays and relies on the assumption that the quantity of initial mRNA is the same for both labeled samples. Furthermore, one assumes that some genes are up-regulated in the query sample relative to the control and that others are down-regulated. For the hundreds or thousands of genes in the array, these changes should balance out so that the total quantity of RNA hybridizing to the array from each sample is the same. Consequently, the total integrated intensity computed for all the elements in the array should be the same in both the Cy3 and Cy5 channels. Under this assumption, a normalization factor is calculated by dividing the total intensity of all the spots in one channel by the total

intensity of the other channel. This factor is then used to re-scale the intensity for each spot on the array [39]. Since a constant scale factor is used for all spots data, this method is not able to remove the nonlinear biases. In addition, according to its assumptions, this method is only applicable to gene expression data.

The main difference of this approach and the global normalization approach is that based on its assumption, this method is specific to gene expression arrays; while global normalization approach can also be used for array-CGH microarray data.

Normalization using ANOVA models: An analysis of variance model for microarray data was proposed by Kerr et al. [36]. Similar models have been described elsewhere in the literature. The ANOVA (Analysis Of Variance) model is applied to transformed intensity data, for example, a logarithmic transform of raw intensity data. It allows one to account for sources of variation in the data that are attributable to factors other than differential expression of genes, thus it effectively normalizes the data. Main effects of the array and dye and interactions between array and dye, variety and gene, dye and gene, array and gene and variations of duplicated spots on one array are included in the model. The variety–by–gene terms capture variations in the expression levels of a gene across varieties and are the quantities of primary interest in the analysis and are referred to as relative expression values. The relative expression values are normalized data in the sense that effects due to the array, gene, spot, etc., have been removed. Although the use of relative expression values represents a departure from the customary analysis of ratios, differences in normalized expression values are in fact estimates of the log ratio of the relative expression between two samples (assuming the raw data have been log transformed).

The power of the ANOVA formulation is that it allows investigators to consider experiments that involve more than two samples and to combine information across multiple arrays that are hybridized with experimental samples in (almost) any arrangement.

Shift Transforms: Based on the linear model of Rocke and Durbin [25] (previously described) the so-called "shift-log" normalization method was proposed by Kerr et al.

[56]. This method adjusts the log ratios by adding a constant to the signal values of one channel and subtracting the same constant from the signals in the other channel prior to the logarithmic transformation:

$$Z_{R,i} = \log_2(Y_{R,i} + C) \text{ and } Z_{G,i} = \log_2(Y_{G,i} - C)$$

The constant C is estimated by minimizing the absolute deviation of each log ratio $(Z_{R,i} - Z_{G,i})$ from the median log ratio of the array. The shift-log transformation moves the origin on a scatter plot of Y_r versus Y_g along the line $Y_r = -Y_g$ to approach the regression line of Y_r versus Y_g. The curvature-causing background difference is, therefore, minimized. Shift-log does not specifically adjust the slope of the regression line of Y_r versus Y_g; therefore, it should be less effective on curvatures resulting from slope differences.

Newton et al., [37], proposed a similar shift transformation in the context of shrinkage estimation. Their method moves the origin along the line $Y_r = Y_g$ by adding the same positive constant to both channels.

$$Z_{R,i} = \log_2(Y_{R,i} + C) \text{ and } Z_{G,i} = \log_2(Y_{G,i} + C)$$

Although this was not the intended purpose of this transformation, it can decrease the curvature in an RI plot. However, when the slopes of the two channels are the same $(b_R = b_G)$ moving the origin along the $Y_r = Y_g$ line cannot bring the origin closer to the regression line [35].

It will be discussed later in this chapter why this method can't be applied to our data.

Self-Normalization (as named by Fang et al, [41]): The self-normalization method assumes that experimentally introduced error is multiplicative and that for corresponding spots in replicated measurements it is consistent. Based on this assumption, the error on a log scale is additive and a subtract operation applied to the data sets from two replicate experiments will remove this systematic error. This approach requires the association of a dye-flip technique. The dye-flip (also known as dye-swap or reverse labeling) technique generates paired slides where on the first slide one mRNA sample is labeled by Cy5 and the other mRNA sample is labeled by Cy3 while on the second slide the labels for the

two samples are exchanged. Based on self-normalization, the normalized result for a spot is half the difference between logged ratios measured from a pair of dye-flipped replicates for this spot. Therefore self normalization has the property that it corrects feature-specific (i.e. probe specific or spot-specific) differences [41].

Since this method is not applicable in cases that two dye-flip slides are not available, it was not considered in this thesis.

Adaptive Normalization involves regression techniques to estimate the bias from the data [39, 41]. This approach employs the assumption that the bias introduced in the experiment is dependent on a number of factors (intensity, print tips, spot position, etc.) and employs regression techniques to obtain a fit of the specific relationship and then makes the correction. Adaptive normalization performs differently for the different regression techniques employed.

The adaptive normalizations methods include linear or non-linear **intensity dependent** normalization and **spatial** normalization.

The intensity dependent normalization techniques consider the regression models in the form of

$$M = c(A)$$

Where $M = \log(R/G) = \log(R) - \log(G)$ is the log ratio and $A = 1/2(\log(R) + \log(G))$ is the average log intensity and R and G are the intensities of the Cy5 and Cy3 channels respectively and $c(\cdot)$ is a regression function.

For finding *c(A)* the promising approach of Yang et al. uses "LOESS" regression of log ratios (M) as a function of the average of log intensity (A) (the LOESS regression will be described in detail later in this chapter). LOESS was used for local linear regression of M versus A rather than regressing log(R) directly to log(G), to attribute uncertainty to both channels by regressing to the geometric mean of the intensity.

Workman et al, [42], proposed using splines for finding the fit. Their approach seeks to transform the signal distribution of one array to the signal distribution of a target array in order to make the signal distributions comparable. The method uses quantiles from array signals and target signals to fit smoothing B-splines. The splines are then used as signal-dependent normalization functions on the array signals.

Recently, some regression-based normalization techniques have been proposed to remove the location dependent systematic differences that result in spatial heterogeneity of signal. Variability in Cy3/Cy5 ratios has been shown to be generated, in part, by the specific print-tip used during the spotting of the cDNA probes by Yang et al, [30]. They introduced print tip-LOESS normalization which performs the intensity dependent LOESS over each print tip group.

Spatial effects are not only caused by the printing device but may also be related to temperature or humidity during the time of printing, the batch of cDNA represented by a specific plate, reagent flow during the washing procedure after hybridization, or from uneven or tilted glass surfaces during scanning. Workman et al, [42], suggested that signal gradients can be normalized by subtracting local signal estimates (log intensities or log-ratios) and devised a spatial gradient normalization using a two-dimensional Gaussian function. Local signal (log-ratio) was estimated for each probe using a weighted mean of neighboring probe signals. A sliding square window centered on the each probe (50×50 for oligonucleotide arrays, and 10×10 for cDNA) was used to define the local neighborhood. Weights were defined by their Euclidean distance to the center probe using a Gaussian function (standard deviation 19 for 50×50 neighborhood and 3 for the 10×10 neighborhood). For both oligonucleotide and cDNA array data, this adjustment was made after global quantile-spline normalization.

Wilson et al, [43], transformed the logged intensities to the mean versus difference scale (instead of the red versus green scale); fitted a single LOESS curve to the transformed data, computed the residuals from the curve fit and spatially smoothed the residuals with a median filter to estimate the spatial trend and computed the residuals from the spatial trend estimate. Spatial median filtering with a 3×3 block of spots (where each spot is represented by a pixel) was used to correct both streaky spatial effects, as well as picking up global trend while not being skewed by highly differentially expressed genes.

Colantuoni et al, [44], proposed the normalization of all array element signal intensities to a mean intensity that is estimated locally across the 2-D surface of each microarray. This mean intensity is estimated using the "LOESS" function in the R statistical language. The LOESS function is used to calculate the mean element signal intensity at

each point across the array surface using intensities at neighboring points. This method was originally proposed for one-channel microarray experiments.

Fang et al., [41], in order to remove the spatial biases, used a two dimensional regression technique that took the spot position as the independent variable and the log ratio as the dependent variable. They then correct the log ratios by subtracting the obtained fit from them. The regression is applied to each block separately.

5.1.4 Methods of Evaluating the Normalization Performance

While there are many publications on normalization for gene expression arrays only a few has been published on the performance evaluation of methods and validation of the assumptions. Below we include a summary of the few systematic studies that have addressed this important issue.

Usually the effect of normalization on reducing the bias and systematic variations is considered, but to what degree the biological variations are preserved is not tested. One way of evaluating the effect of normalization of gene expression data which has been used in most of the studies is to see how normalization affects the identification of differentially expressed genes. There is no unique way of identifying the differentially expressed genes and there is usually no true reference to compare the results to because the true expression level of genes is unknown.

In order to compare the different normalization procedures, Yang et al., [30], considered their effect on the location and scale of the log ratios M using box plots. A Gaussian kernel density estimator was also used to produce density plots of the log ratios for each of the normalization methods. They also considered the effect of the normalization procedures on the t-statistics used in the t-test to find differentially expressed genes.

Colantouni et al, [44], attempt to remove the spatially systematic artifacts and compare the spatial distribution of genes found to be differentially expressed before and after normalization. The point is that they want to remove the localization of differentially expressed genes. This localization results from the systematic bias.

Workman et al, [42], assess the efficiency of normalization in three different steps: global assessment in which they compare the distribution of signals before and after normalization, signal dependent assessment in which they visually compare the signal

dependent bias in the data before and after normalization, and finally biological assessment in which they take a set of genes known to be differentially expressed from other experiments (other than microarray) and compare the significance of a t-test to find the differentially expressed genes before and after normalization.

Park et al, [45], use the variability among the replicated slides to compare performance of global, linear and non-linear intensity dependent normalization methods. They also compare normalization methods with regard to bias and mean square error using simulated data.

5.2 Hypothesis

As seen above, there have been several studies on the normalization methods for gene expression microarray data and the effect of different normalization strategies on the reproducibility and accuracy of these data; however to the best of our knowledge no studies have been carried on the data from the newly developed CGH microarrays.

Although the microarray experimental steps are the same for both gene expression arrays and CGH arrays, the type of the quantity that is measured in each assay is different, in the case of the gene expression arrays the quantity of interest is the relative expression level of each gene in each sample and in case of CGH arrays, it is the copy number of the DNA sequences. Genes are expressed at many different levels in the genome and expression level of the same gene is not exactly the same in two different samples and that is what makes the true relative expression level of genes in an experiment unknown. **In contrast, in case of the CGH arrays, the copy number of DNA sequences is a known value for the normal cells, which make the majority of the probe in our experiments, so the "truth" about the relative copy number of the majority of clones is known and this fact can be used in the evaluation of the normalization methods**. Moreover, in gene expression assays a ratio of less than 2 is generally considered a non-significant change in the gene expression level; however for CGH arrays a single copy number amplification of a clone compared to a normal clone will result in a ratio of 3/2. Considering the fact that there is usually some degree of contamination of normal cells into the tumor (abnormal) cells, the fold change in the copy number can frequently be even smaller than

3/2. So the challenge would be to preserve the true copy number changes while removing the systematic variations.

There is ongoing research on normalization and preprocessing of gene expression microarray data and there is yet no best method to normalize the data. The normalization strategies that have been used for gene expression data need to be tested for accuracy and precision and may be adapted to the specific characteristics of the CGH microarray data.

We hypothesize that we can correct for systematic sources of variation while maintaining the true biological variations as small as single copy number changes in contaminated samples which makes the true fold change even smaller.

To test this hypothesis, after a thorough investigation of the systematic variations in the data from our array-CGH experiments, and based on the existing normalization methods for gene expression data, we propose a stepwise normalization framework. We develop several methods for evaluating the performance of this stepwise normalization using a variety of experiments. We show, through the results of our experiments, that stepwise normalization scheme that we propose is successful in terms of reaching the determined goals.

5.3 Systematic Variations

In the following sections we describe the different types of systematic variations that we observed in our data and attempt to identify their sources.

We start with a major source of bias, the background fluorescence and consider its effects on the log ratios.

We then continue the data investigation by looking at the histograms of data in our dataset.

Two significant sources of bias, intensity dependent dye bias and spatial dye bias are then examined and a measure for quantification of the strength of these biases is introduced.

A discussion about the non-linearities in the log ratios is then presented.

Finally another source of systematic variation, although not as significant as dye bias, is reported.

5.3.1 Background Fluorescence

The background fluorescence is caused by many sources including:

- Unbound or non-specifically bound probes
- Auto-fluorescence of DNA, glass slides and other material put on the spot (such as reagents, etc.)
- Contamination on the surface of the slide, etc.

The Background on the substrate presents a special case of bias. On the assumption of additive error, an estimate of background is usually subtracted from the measured fluorescent intensity of the spot foreground value before correcting other systematic errors.

The background inside a spot can not be measured but the background around the spot can be estimated. The conventional method for estimating the background around the spot is to select a region around every spot and compute the background from this area. Traditionally, it has been assumed that the background around the spot is the same as the background inside the spot. Recently, there have been some studies indicating that this assumption is not always true.

Tran et al, [46], published the results of a study in which they visualized the DNA spots before hybridization using unincorporated red dye staining and they found that DNA fluorescent intensity may begin at a level below the background signal intensity.

Martinez et al, [47], identified spot-localized, contaminating fluorescence in the Cy3 channel on several commercial and in-house printed microarray slides. They also determined the intensity of persistent spot-localized, contaminating fluorescence after hybridization could not be predicted from scanning microarray slides prior to hybridization trough mock hybridizations.

Inaccurate estimates of background fluorescence under the spot create a source of error, especially for low intensity spots. Some of the research groups working with microarrays have decided not to subtract the background intensity from the fluorescent intensity of the spots. If background intensity is not subtracted from the intensity of the spots, the log of red to green ratios shows a systematic bias towards the positive values for the low intensity spots, this happens because the background intensity in the red channel is higher than the background intensity in the green channel (which is in turn caused by the

detection setup). The ratio of red intensity to green intensity will then be artificially higher than what it should be because the measured intensity in the red channel, which is a combination of the true signal and the background signal, is higher than that of the green channel. As an example, Figure 5-1 shows the $\log_2$ ratio of the red to green intensity of each spot of slide H526-4 plotted versus the average of the $\log_2$ of red intensity and the green intensity for that spot. The increase in the $\log_2$ ratio for lower intensities that can be observed in this figure is a systematic bias. We refer to this bias as the **"background bias"**.

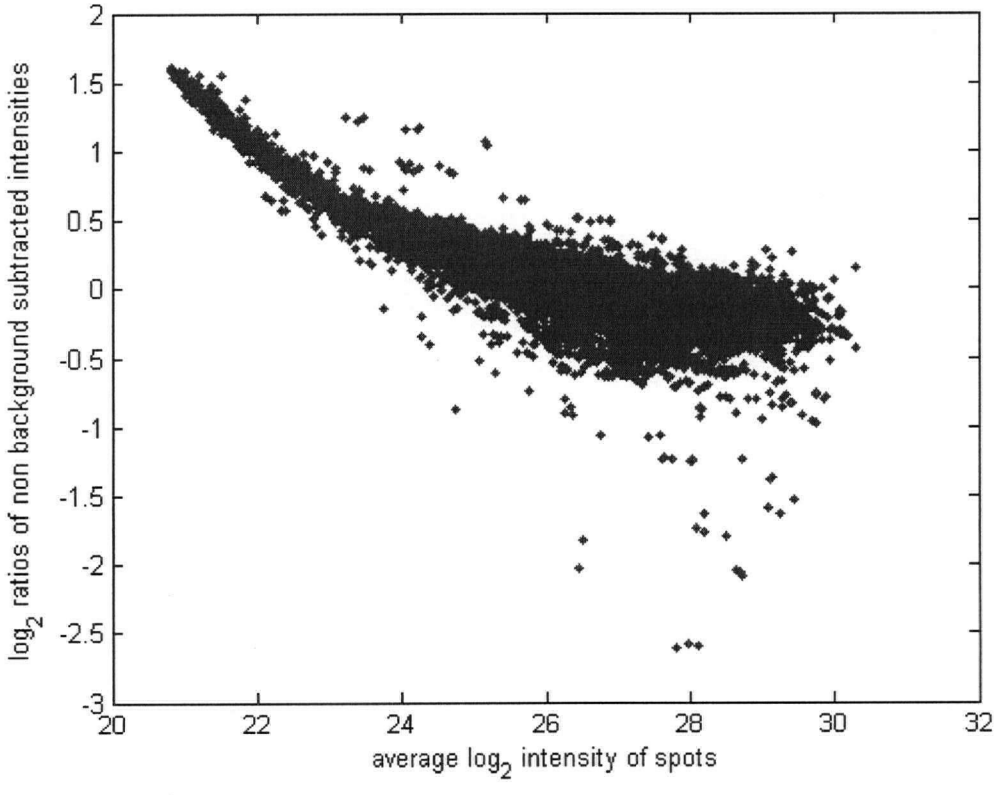

Figure 5-1 "Background bias"

Therefore not correcting for the background fluorescence has the disadvantage that introduces the explained bias into the data. In contrast, subtracting the background adds noise to the lower intensity spots. There seems to be a third option and that is to attempt to estimate the ratios without estimating the background and based only on the intensities of the spots pixels.

This motivates us to try a background-independent ratio estimation method.

5.3.1.1 The regression Method for Estimating the Ratios

According to Jain et al, [48], given the sets of foreground and background pixels for the test and reference channels, there are several ways to estimate the ratio. One method for ratio estimation is to compute the ratio of the foreground intensity less the estimated background intensity for the test and reference channels. The foreground and background intensities are the average of pixel intensities of the foreground and background regions. This corresponds to the ratio of the coordinates of the center of mass of the foreground pixels corrected for background. This method requires an estimate of background. If a line is fitted to the foreground pixel intensities of one channel versus the other, the slope of the line can be used as a background-independent estimate of the ratio (figure 5-2). We call this method of ratio estimation as the "**regression method**".

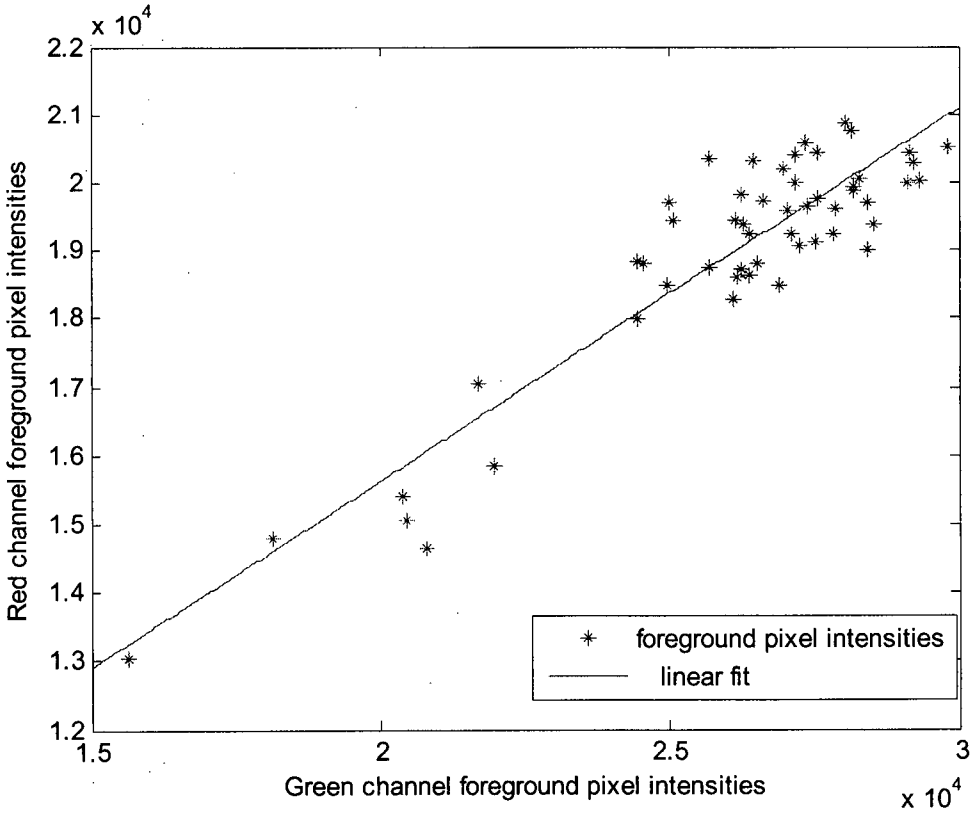

Figure 5-2 The linear regression method for estimating the ratio

There is no published study that evaluates the performance of the regression method in estimating the ratios. Since the idea behind this method is to estimate a background-independent ratio, we examined it in more depth in the following experiment.

Experiment: Each clone is spotted three times in our microarrays. The regression method and the conventional method of ratio estimation were compared based on the variability that they cause in the ratios of the triplicates. The best ratio estimation method is the one that generates the closest ratios for the triplicate spots.

Methods: the ratio for each spot was estimated using the "regression method". The ratio for each spot was also estimated using the conventional method as follows:

$$ratio = \frac{(mean\ \ of\ \ foreground\ \ pixel\ \ Intenistie\text{s} - estimate\ \ of\ \ background)_{Cy5\ \ channel}}{(mean\ \ of\ \ foreground\ \ pixel\ \ Intenistie\text{s} - estimate\ \ of\ \ background)_{Cy3\ \ channel}}$$

The median pixel intensity of the background pixels was used as the estimate of the background.

Using the ratios estimated by each method, we then calculated the coefficient of variation (standard deviation divided by the mean) of three $\log_2$ ratios obtained for three replicate spots for each triplicate set. The average of the coefficient of variations of the triplicate sets of each slide was used as a measure of closeness of the estimated ratios to the truth.

The fact that all three of the triplicates are printed with the same pin within the same dipping cycle and they are located in the same column of each subgrid next to each other makes them ideal for our comparison purpose. Since the triplicates are from the same clone and in about the same region, so no normalization is needed before this comparison can be performed.

The data from Slides H526-1 to H526-4 were used in this experiment.

Results: Table 5-1 shows the average of coefficient of variation of the triplicate $\log_2$ ratios of spots in each slide. The Regression method increased the variability in ratios of triplicates in all the cases.

	H526-1	H526-2	H526-3	H526-4
Conventional method	21.9%	4.4%	1.5%	0.4%
Regression method	54.1%	11.5%	23.7%	20.9%

Table 5-1 average coefficient of variation of triplicate $\log_2$ ratios

In an attempt to find the reason of this, we plotted for each spot the scatter plot of red intensities against green intensities. The intensities of foreground and background pixels are plotted with different markers. In order to test the effect of segmentation results on the regression method, the regression is performed with all the data points first and then the closest data point to the origin is removed and the regression is performed again and this is repeated until all the data points are removed. Then a plot of the slopes of the regression functions is prepared (figure 5-3).

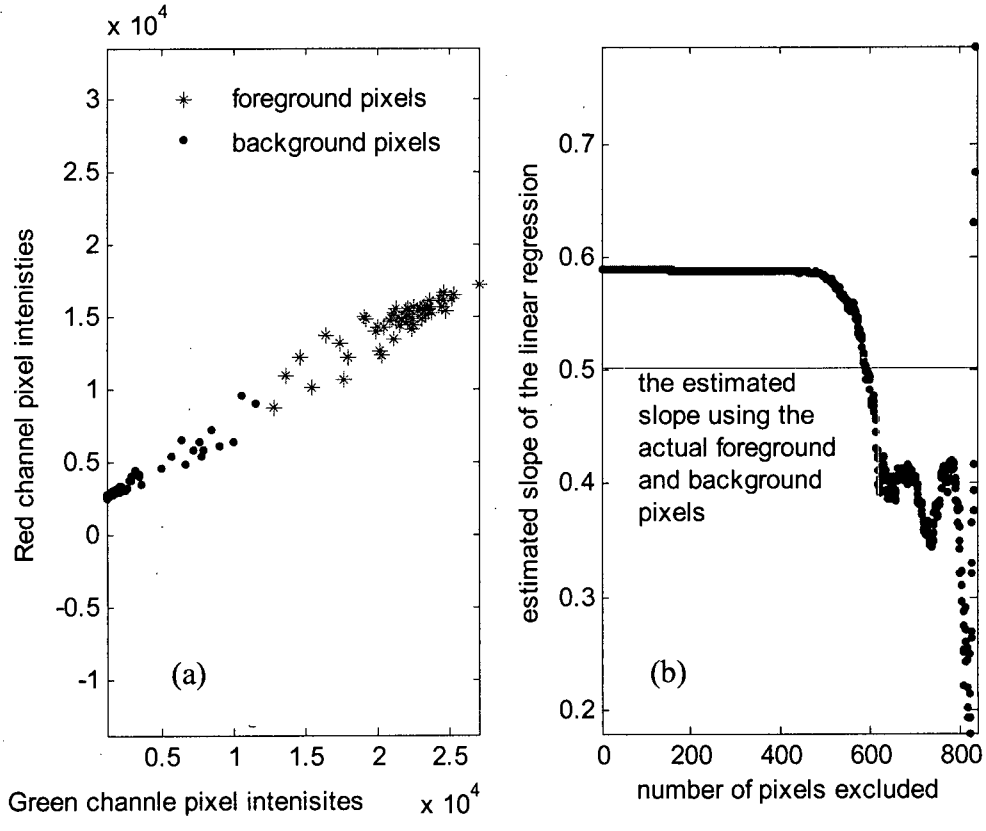

Figure 5-3 (a) Plot of red channel pixel intensities versus green channel pixel intensities with the foreground and background data points shown in different shapes, (b) Plot of the estimated ratio using the regression method for different combinations of the foreground and background pixels

In figure 5-3 (a) the points closer to the origin are the background pixels intensities and the points further from the origin are the foreground pixels. Figure 5-3 (b) shows that removal of the points that are very close to the origin does not change the slope of the regression function but as more and more of the points are removed and we get close to the pixels on the border of the foreground region then removal of each point changes the

slope considerably. In fact the border pixels are in the region of rapid changes in figure 5-3 (b).

According to this observation we believe that the reason that the variation of $\log_2$ ratios of the triplicate spots is increased by using the regression method is that this method is not robust to the misidentification of the pixels in the border of the foreground and background region so that removal or addition of each single pixel from or to the foreground affects the ratio a lot.

The conventional method of calculating the ratios is more robust to the misidentification of data points as foreground or background due to the fact that the pixel intensities are averaged to obtain the foreground intensity.

After a closer look, we found that the reason that the regression method is not robust is that the distribution of data points in the scatter plot is more like a cluster. In other words the range of variations of intensities of the foreground pixels is not wide enough to allow the robust estimate of the slope of the fitted line.

As a result the regression method can not be used as a ratio estimation method and we are still left with two options in dealing with the background issue: 1) To subtract the background so that the "background bias" (as described before) is not introduced to the log ratios. This approach has the disadvantage that it increases the variability of the log ratios of the lower intensity spots. 2) Not to subtract the background and to remove the background bias in the normalization phase. We will compare the performance of both of these approaches later in this chapter.

5.3.2 Histogram Inspection.

As the first and most basic tool of data inspection, we looked at the histograms of signals from our microarray database. Histograms are especially useful for getting an overall idea of the center (i.e., the location) of the data, spread (i.e., the scale) of the data, and skewness of the data. We examined the histograms of the raw intensities, background intensities and background-subtracted intensities.

Results:

- The centers of the distributions of the red channel intensities were found to be either about the same or lower than the green channel.

- The average of background intensity of the red channel was found to be higher than green channel.
- The center of the distribution of background-subtracted intensities was found to be lower for the red channel.
- The standard deviation of the intensities of the red channel was found to be lower than the green channel.

Figures 5-5 through 5-8 show the summary of the descriptive statistics of the intensities of the two channels.

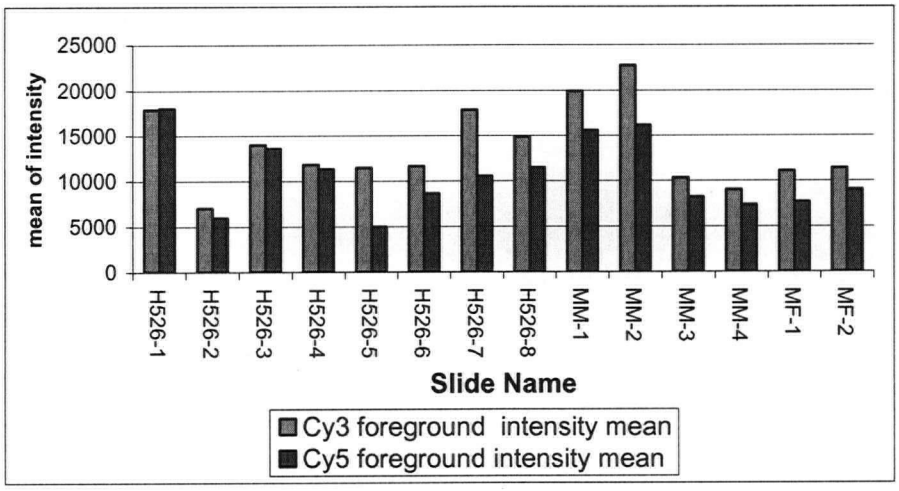

Figure 5-4 Average of foreground spot intensities (without background subtraction)

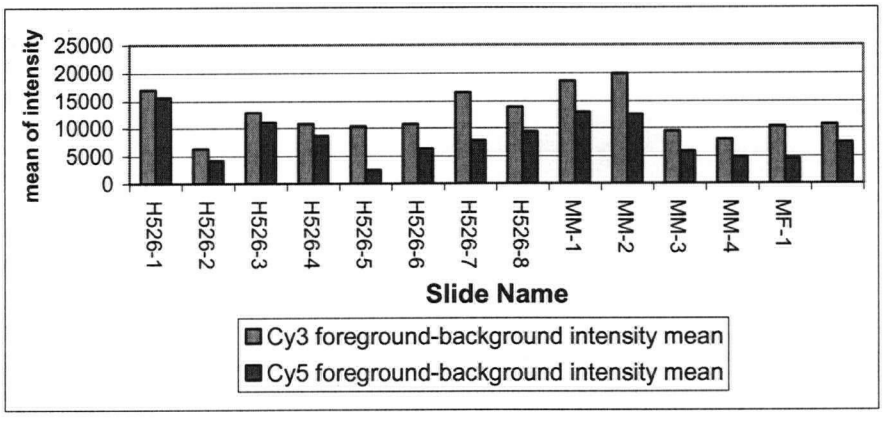

Figure 5-5 Average of foreground spot intensities (after background subtraction)

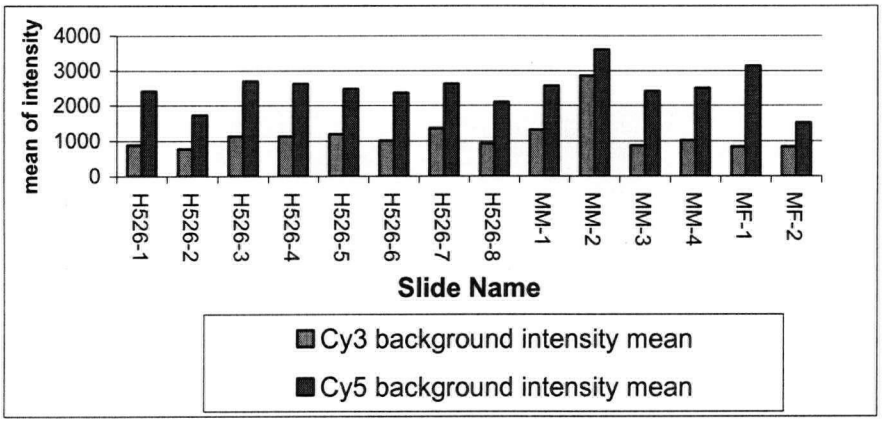

Figure 5-6 Average of background intensities of spots

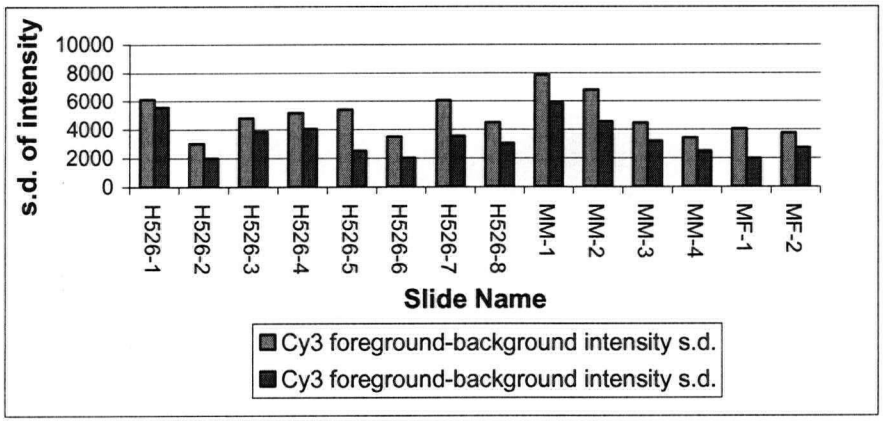

Figure 5-7 S.d. of foreground spot intensity (after background subtraction)

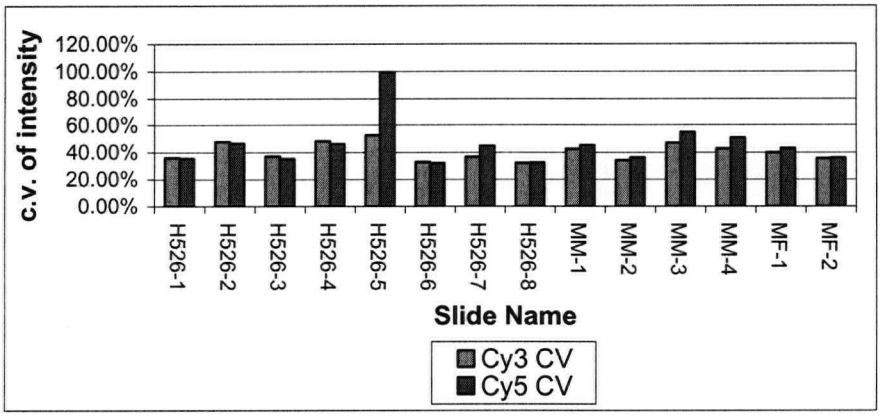

Figure 5-8 Coefficient of variation of the background corrected intensities

In order to compare the distributions of the ratios to the normal distribution, the quantile-quantile plots (QQ-plots) were used. In these plots, quantiles of input samples are plotted

versus the quantiles of a standard normal distribution. If the plot is linear then the distribution of the input samples is normal. Figure 5-9 shows the QQ-plot of $\log_2$ ratios from one of the slides in our database. This QQ-plot is typical of all of the datasets in our database. The plot shows that the distribution of the data is very close to normal except for very low and very high quantiles which correspond to the tails of the distribution. This assures us that the distribution of the $\log_2$ ratios can be confidently assumed to be normal (we will make use of this fact later in this chapter).

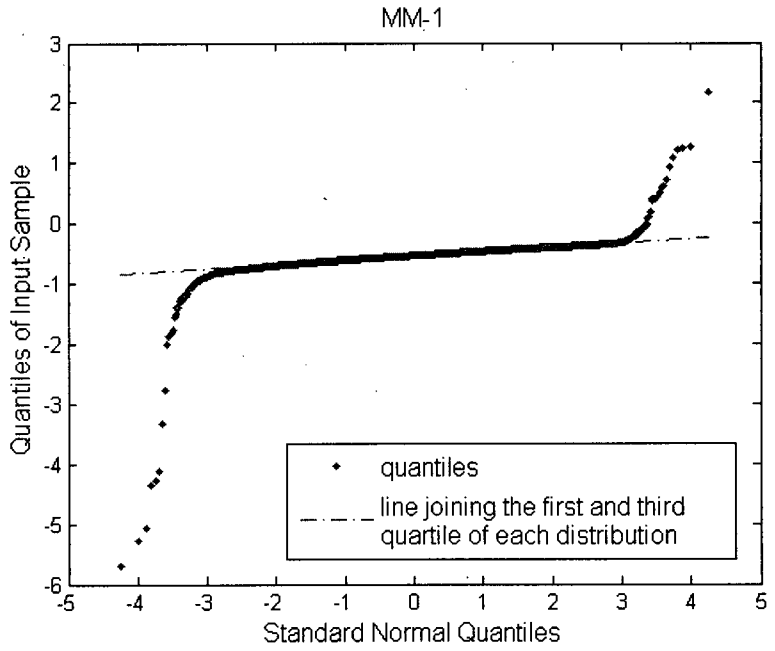

Figure 5-9 A typical QQ-plot of $\log_2$ ratios of the slides in our database

Discussion: When starting with the same amount of DNA for both probes (which is the case for all the slides in our database), under idealized conditions, the distributions of the measured fluorescent intensities is expected to be the same for both channels, however this was not the case.

Although the measures of repeatability of data will be discussed in detail later, but as a very simple measure of repeatability, the Pearson Correlation Coefficient of data from each pair of replicate H526 slides is used here to compare the repeatability of the data to see if there is a relationship between the closeness of distributions and the repeatability of the data. The correlation coefficients are shown in figure 5-10. The correlation coefficient of the data from slide H526-5 with other data is the least followed by the correlation

coefficient of the data from slide H526-7 with other slides. As figures 5-4 and 5-5 show, the difference between the average intensity of the two channels is the most for these two slides.

This preliminary analysis shows that the closeness of distributions of signals of the two channels affects the precision of the data.

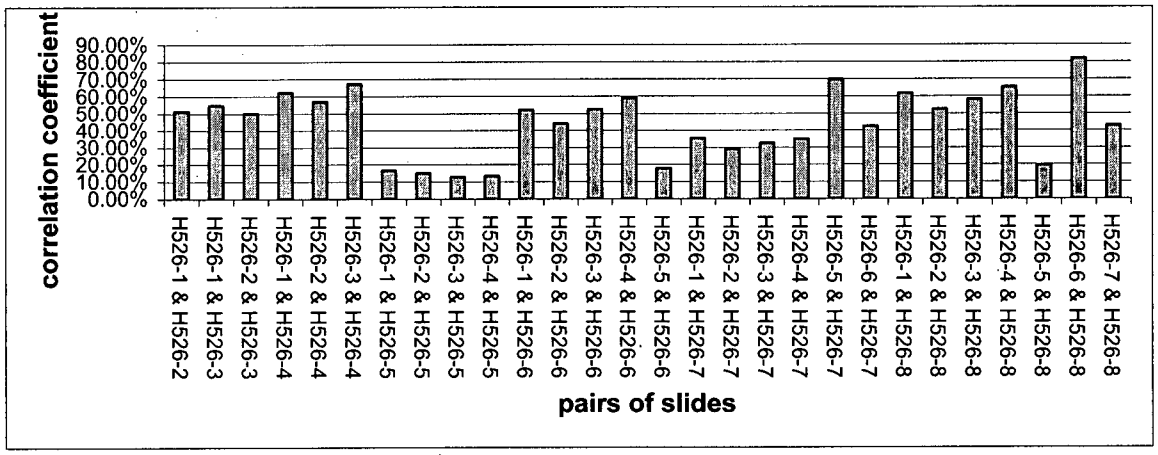

Figure 5-10 The correlation coefficient of the $\log_2$ ratios from each pair of replicate slides

A global linear normalization (i.e. global scaling of log ratios) that basically forces intensity distributions to have the same central tendency (arithmetic mean, geometric mean, median) by multiplying the ratios by a scaling factor is not able to improve the correlation because it only shifts the log ratios by a constant value, therefore has no effect on the correlation.

5.3.3 Intensity Dependent Dye-Bias

Perhaps the best way to show the intensity-dependent dye bias is through the **M-A plot**. In the so-called M-A plots the log ratio $M = \log_2(I_r/I_g) = \log_2(I_r) - \log_2(I_g)$ is plotted against the mean log intensities $A = 1/2(\log_2(I_r) + \log 2(I_g))$ where I_r and I_g are the intensity of the red and green channels respectively. Although M-A plots are basically only a 45° rotation of plot of I_r versus I_g with a subsequent scaling, they are usually

preferred to the I_r versus I_g plot. This is because, first, they reveal intensity-dependent patterns more clearly than the original plot because the patterns will be compared to the zero line (if ratio equals one then the log ratio is zero) and second, usually performing a regression follows the visual inspection of the M-A plot. In this way by regressing to the log of the geometric mean of the intensities ($A = 1/2\left(\log_2(I_r) + \log_2(I_g)\right) = \log_2\left(\sqrt{I_r \times I_g}\right)$), the error is attributed to both channels instead of just one of them.

Due to the extremely large number of data points, the M-A plot may not show the trend in the data clearly as the accumulation of the points is not seen clearly at each intensity interval. So it will be useful to smooth the M-A plot by a moving mean or a moving median filter or a nonlinear scatter plot smoother like LOESS, which will be described in detail later, along the intensity axis so that the dependence of M on A is shown more clearly (figure 5-11).

Some of the M-A plots clearly show a non-linear dependence of log ratios M on spot intensity A. For low intensities M is biased towards the negative values. Another point of interest in the plots is that the deviation around the smoothed curve is generally higher at lower intensities.

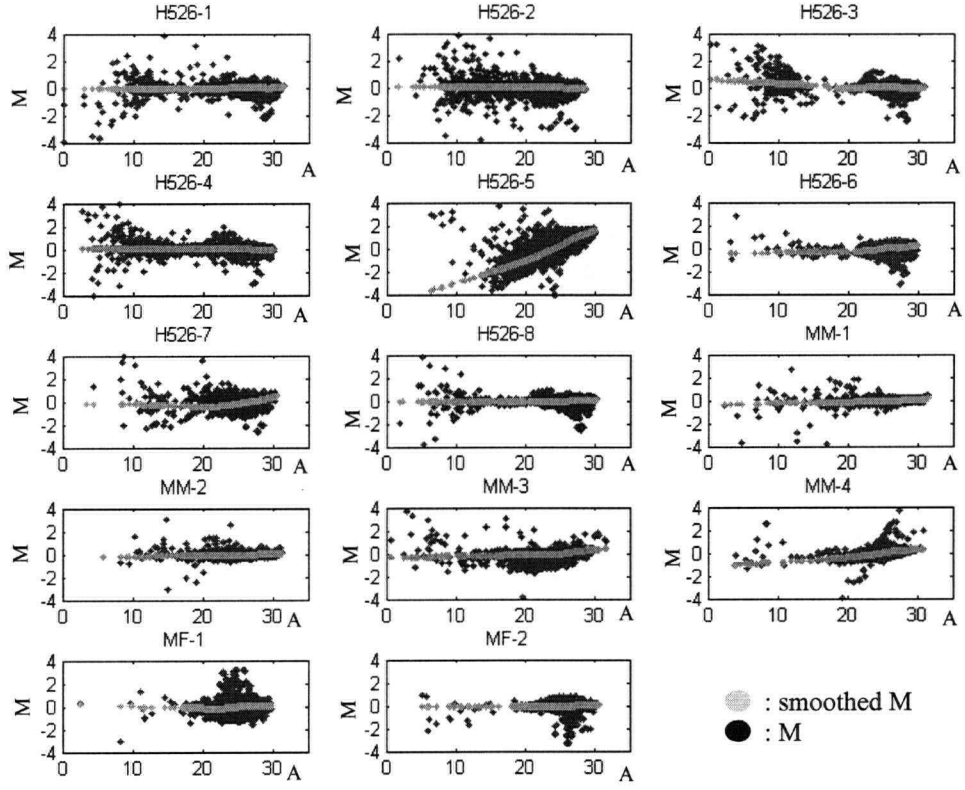

Figure 5-11 M-A plots of slides, the bright line shows the smoothed M

5.3.4 Spatial Dye-Bias

The representation of log ratios based on the corresponding spot location on the microarray slide is another type of plot which can be used to reveal the space-dependent dye bias. We refer to this plot as **M-XY plot**. As in the case of M-A plots, we can plot the spatially smoothed *M* to see the general trend of log ratios across the location on the array (figure 5-12)

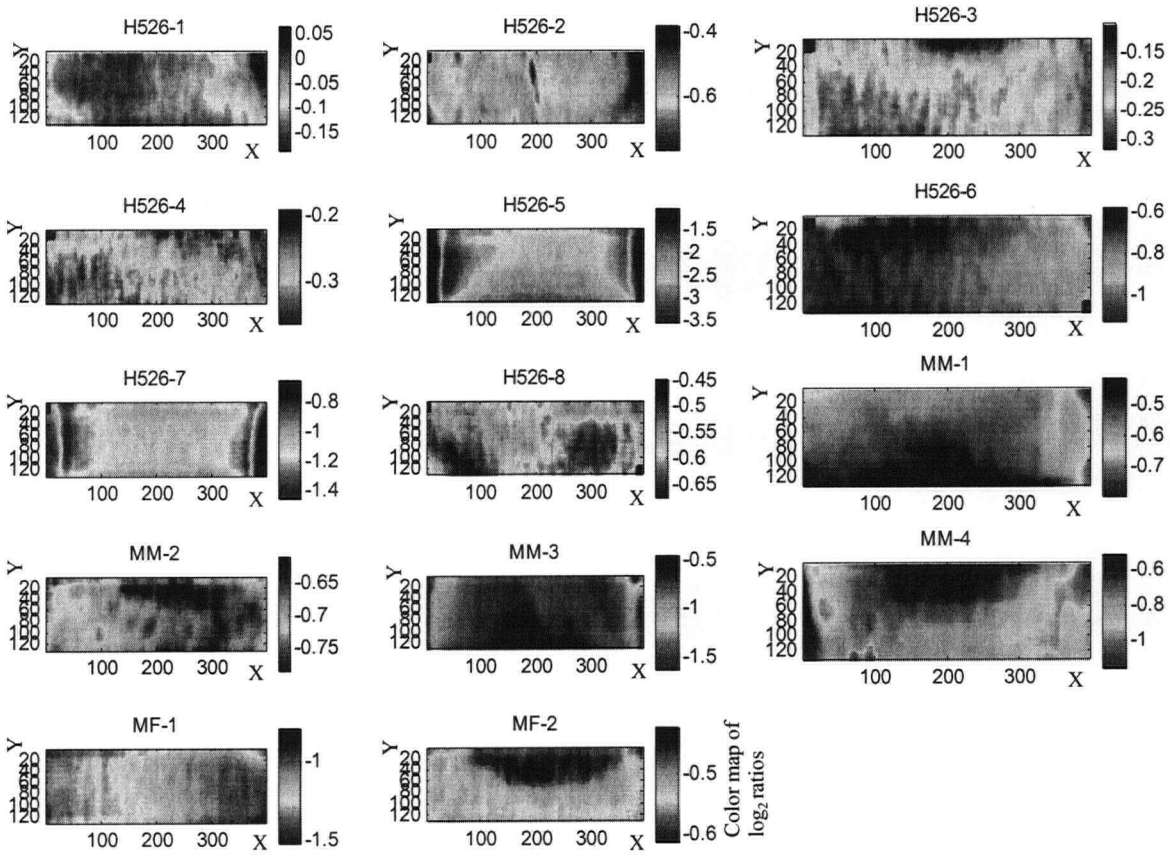

Figure 5-12 M-XY plots

In most of the slides, the spatial heterogeneity in log ratios is observed. In particular the patterns of the smoothed M-XY plot are interesting.

Some times the patterns start at the edges of the slide and it seems that it is related to the way the hybridization solution is put on the surface of the slide. There are usually two methods used for putting the probes on the surface of the slide. Sometimes the material is put in the middle of the slide and then the slide cover is put over the surface. Some times the material is put close to one edge of the slide. From the spatial patterns we guess that for example slides MM-3 and MF-3 are among the first group and slides MM-1 and MF-1 are among the second.

The spatial heterogeneity was originally believed to be caused by the different print tips used in printing the targets on the slides. For example, look at the box plot of log ratios of each print-tip group (figure 5-13). Now we know that at least for our arrays the spatial heterogeneity is not caused by the print tips for two reasons. First the spatial patterns are

not block wise and they seem to be continuous. Second if it was for the print tips then for all the arrays printed in the same print run, the patterns would be the same which is not the case.

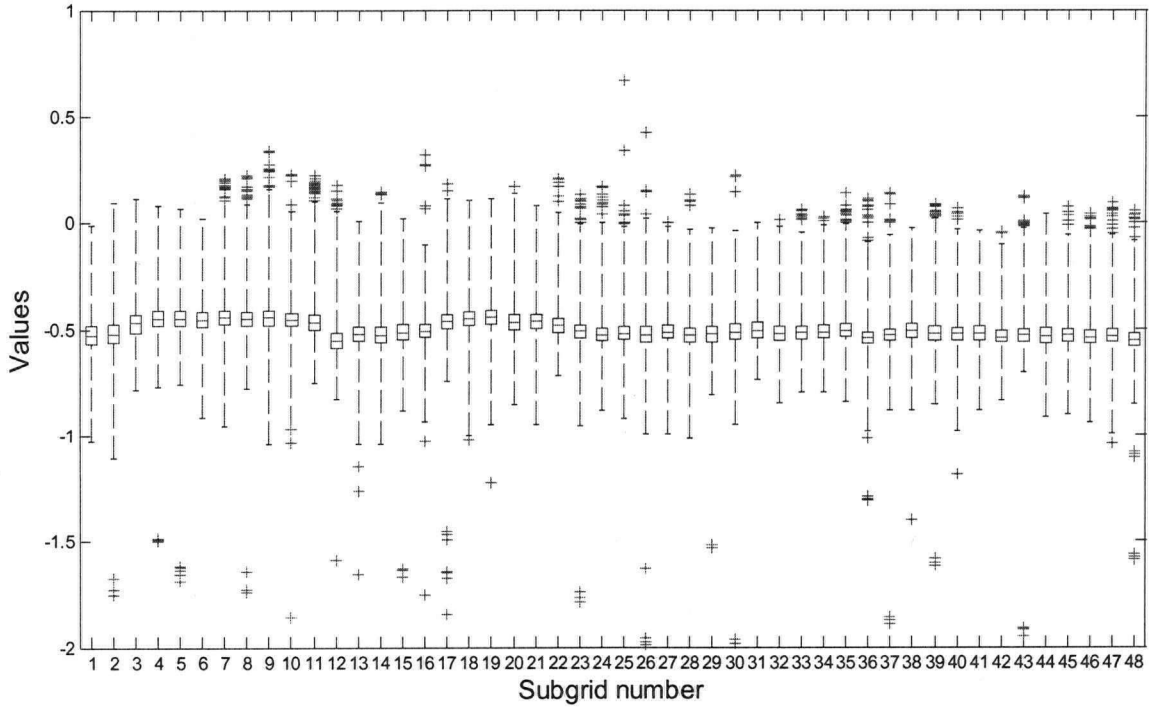

Figure 5-13 Box plot of the $\log_2$ ratios of each print tip group

Another important point is that there is the possibility that the spatial variability be caused by the true biological variations. We reject this possibility for two reasons; first, the print order of spots in our experiments did not follow a specific order so that groups of consecutive sequences were not all located close to each other, instead the location of sequences was random. Second, even if the patterns were representing biological variations by chance, then the same pattern should have been seen in all replicates of an experiment which is not the case.

5.3.5 Quantification of the Dye-Bias Effects

Visual inspection is an important first tool to detect the systematic biases but it is not useful by its own. To measure the degree of dependence of M ($\log_2$ ratio) on A (average $\log_2$ intenisty) and X-Y (spatial location), we need to have a measure to quantify these

effects. The importance of finding a measure that quantifies the strength of the intensity dependent and spatial biases is that by obtaining the bias before and after the normalization, we can get an estimate of how successful the normalization is in removing the bias.

Assume that the $\log_2$ ratios are plotted versus the average $\log_2$ intensities (M-A plot) or versus the spatial location (M-XY plot). The correlation of the $\log_2$ ratios of the spots and the smoothed $\log_2$ ratios (either versus the intensity or spatial location) can be used as a measure of the intensity dependent or spatial bias. We refer to this measure as "**Local correlation**". If the $\log_2$ ratios are not correlated, a "Local correlation" close to zero is expected.

For the intensity dependent bias, the moving averages of the $\log_2$ ratios as a function of intensities are calculated by sliding a window of appropriate size on the $\log_2$ ratios. The average of the $\log_2$ ratios inside the window is then taken. The $\log_2$ ratio of the spot in the center of the window is excluded in averaging to assure that the calculated moving average for each spot is independent of the $\log_2$ ratio of that spot. The calculation of the moving average was implemented using an FIR filter with symmetric padding at the edges, i.e. input array values outside the bounds of the array, were computed by mirror-reflecting the array across the array border. The filter coefficients were all one except for the center coefficient which is zero. The length of the window was 363.

For the spatial dye bias, we used a two-dimensional moving average filter which again excludes the point itself in calculating the average of data points in its neighborhood. A two-dimensional FIR filter was used for implementing it. Size of the window was 33 rows by 11 columns.

Figure 5-14 shows the local correlation values calculated for the slides in our database.

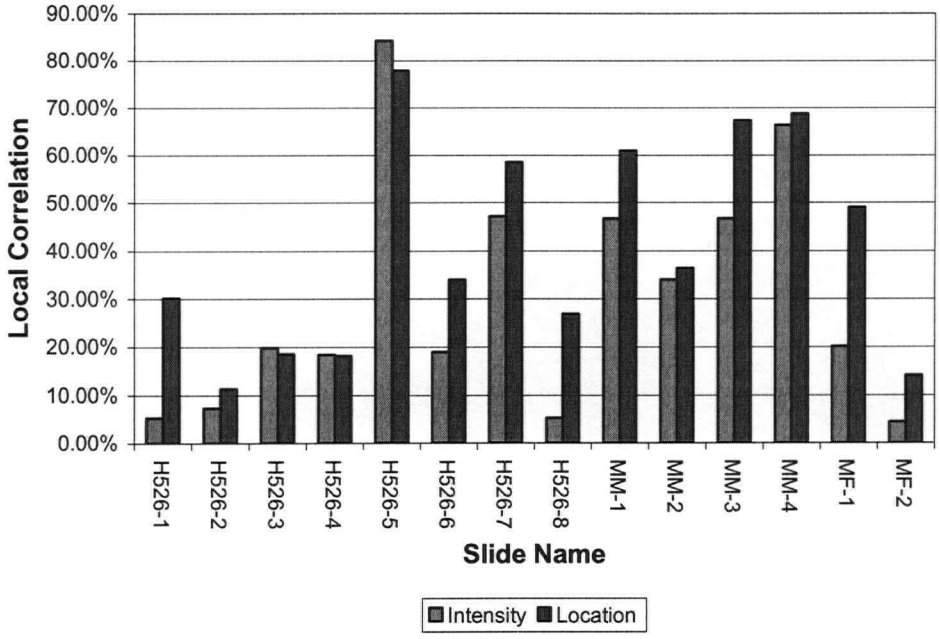

Figure 5-14 Local correlation of the $\log_2$ ratios with the smoothed $\log_2$ ratios

5.3.6 Non-Linearity in Log Ratios

The linear model of measured intensity was described in the Measurement function section. As the simulations of Cui et al, [35], showed the linear measurement function is able to explain the non-linearity in the M-A plot. However, there are some facts observed in our data that do not match with the characteristics of the linear function.

First, in some cases, the non-linearity not only exists in the M-A plot but also exists in the scatter plot of intensities of Red channel versus the intensities of the Green channel. Figure 5-15 shows an example of the non-linear trend in the scatter plot. For the linear measurement function to be true, the intensities of the two channels should be linearly related.

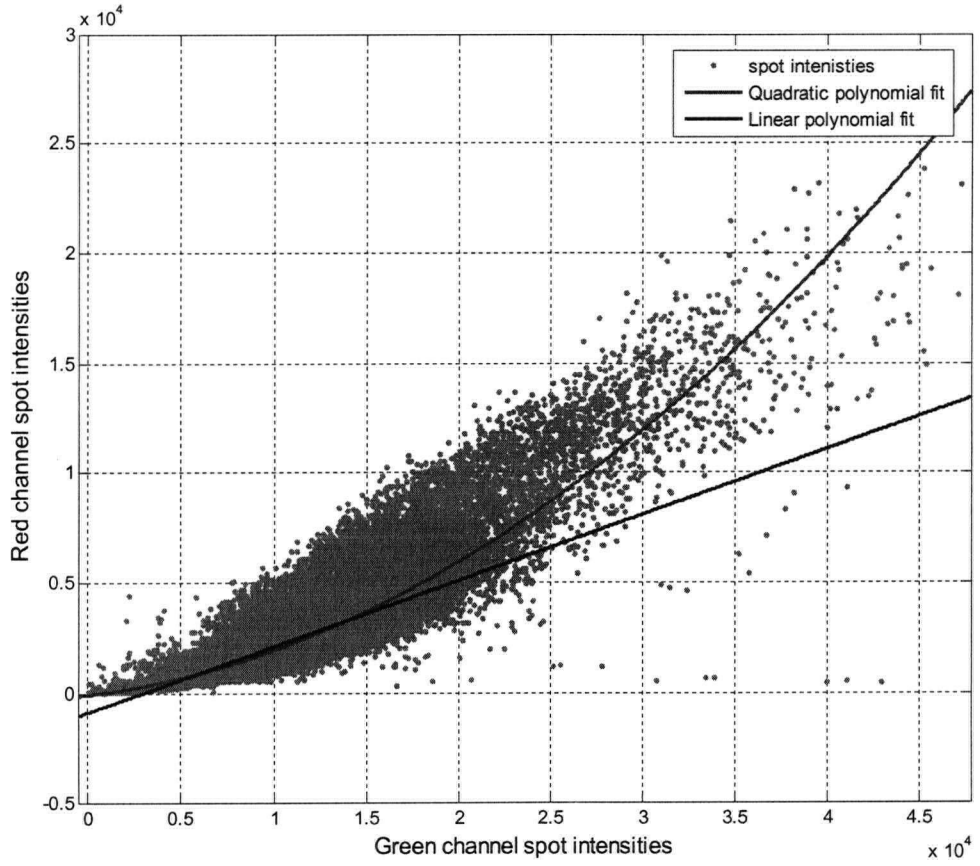

Figure 5-15 Plot of red channel intensities versus the green channel intensities that shows a nonlinear trend

Second, even if the nonlinearity in the scatter plot is quite low; the slope of the linear measurement function is dependent on the location of the spot on the array. To visualize this, we developed an interactive tool that allows us to select a region on the M-XY plot and then plots the scatter plot of the corresponding spots intensities in a new window. In this way one can compare the scatter plots of different regions to find the source of heterogeneity in the $\log_2$ ratios. A sample slide is shown in figure 5-16. Three different regions are chosen and shown on the M-XY plot and their corresponding scatter plots and the linear fit to the scatter plots are shown in figure 5-17. The slopes are quite different for three different regions.

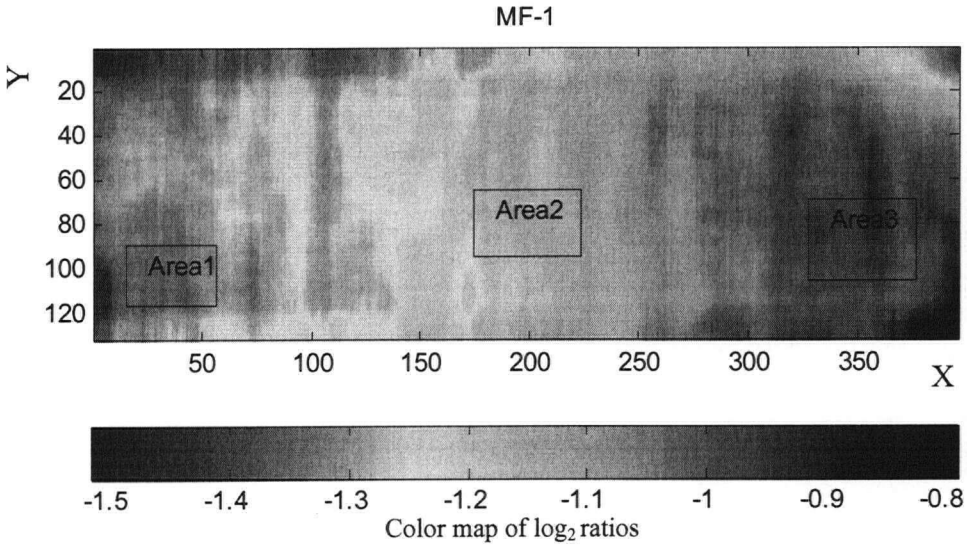

Figure 5-16 A sample smoothed M-XY plots with three different regions selected on it

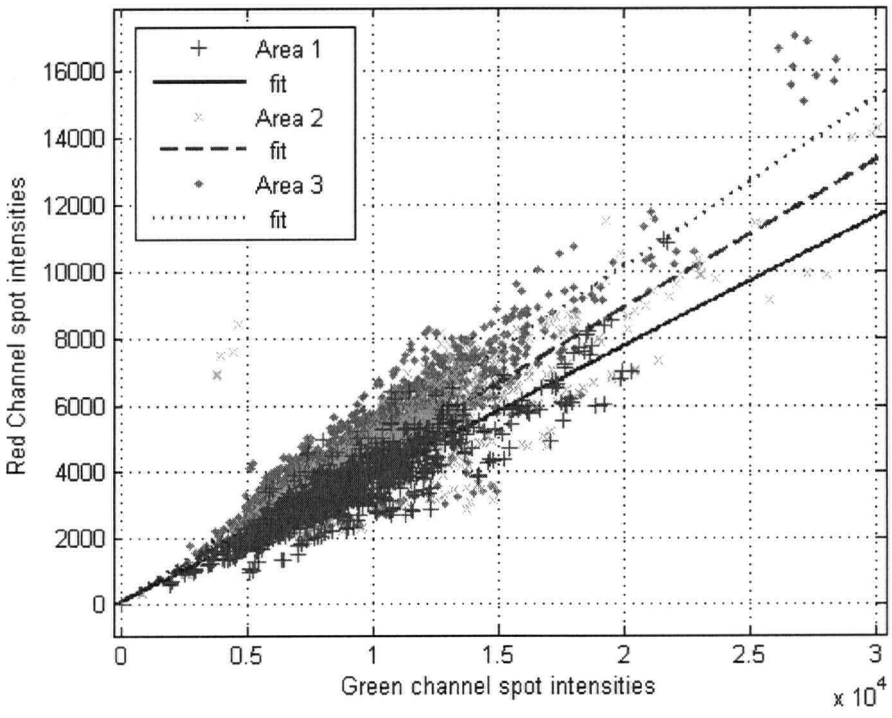

Figure 5-17 Plot of red channel intensities versus the green channel intensities for the spots in the three regions of figure 5-16

Based on these observations, we believe that first; the slope of the linear model of measured intensities varies across the surface of the array and generates the spatial heterogeneity. Second, the slope of the linear function depends on the actual fluorescent signal as well. Based on these conclusions, the "shift transform" normalization approach (previously described) that is based on the linearity of the measurement function would not perform satisfactory in this case.

5.3.7 Other Systematic Effects

For removing the spatial heterogeneity in the microarray data in the M-XY plots, we calculate the spatial trend (details in the methods section) and recalibrate the $\log_2$ ratios by subtracting the $\log_2$ of spatial trend from them. After removing the spatial trend we noticed another type of spatial pattern in the data.

The pattern seen in one subgrid is repeated in all subgrids and is the same as the plate groups. (refer to section 4.4.1 for the definition of plate groups).

As another diagnostic plot, we looked at the box-plot of $\log_2$ ratios from each group plate. The box plots show a systematic variation among different plate groups. The variation is systematic since when looking at all the samples of one plate group which are from different subgrids, we expect the median $\log_2$ ratio be near zero, i.e. positive and negative deviations should cancel out in each plate group but they do not.

This might be because of the fact that different clones that are produced in different plates have experienced different physical conditions during the PCR (Polymerase Chain Reaction) that has affected the efficiency of the PCR and resulted in this kind of artifact.

At this point, after inspecting the systematic variations in the data, we are ready to propose a frame-work for normalization that removes the observed biases from the data. These biases were intensity dependent dye bias, spatial dye bias, and plate effects.

5.4 Proposed Framework for Normalization

After inspecting the systematic biases introduced into the microarray data in our study and reviewing the normalization techniques currently used for normalizing the gene expression microarray data, we now describe our approach for normalizing the data.

According to the discussion of section 5.3.6 we decided to use a locally linear regression of $\log_2$ ratios as a function of $\log_2$ intensities as we found that the intensity dependent dye bias is nonlinear in nature. For the spatial dye bias we decided to use a since we observed that it only changes the slope of the linear model changes

Our proposed strategy consists of three steps:

1. Removal of spatial heterogeneity of ratios: first the ratios are spatially smoothed with a median filter to estimate the spatial trend. Then the residuals from the spatial trend estimate are computed,
2. Removal of plate specific variations: the ratios are grouped based on the plate batches. The ratios of each group are then scaled by the median ratio of the group.
3. Removal of intensity dependent dye bias by fitting a single LOESS curve to the results of the previous step and computing the residuals from the curve fit.

We assume that all the biases are additive to the $\log_2$ ratios (multiplicative to the raw data). This assumption has been shown to perform well in the literature.

Each of the steps stated above will be described in detail in the following sections.

5.4.1 Methods

As indicated earlier, the first step in choosing a proper normalization method is to choose a set of genes used for normalization. Our approach is to use all of the clones on the array for this purpose. The reason for choosing this approach is that we are dealing with experiments where the majority of clones are normal clones and only a small percentage of them are amplified or deleted. This is true especially because we measure the copy number of DNA sequences. Unlike the expression levels that are variable in different samples, the copy numbers are always fixed unless there is some kind of abnormality in the Genome.

In this section we describe the details of each step of our normalization procedure.

5.4.1.1 Intensity Dependent Normalization

According to the discussion of section 5.3.6 we decided to use a locally linear regression of $\log_2$ ratios as a function of $\log_2$ intensities as we found that the intensity dependent dye bias is nonlinear in nature.

We use the robust scatter plot smoother "LOESS", [57], implemented in the statistics toolbox of Matlab, to perform a local A-dependent normalization:

$$M = \log_2(R/G) \rightarrow \log_2(R/G) - c(A)$$

where *c(A)* is the LOESS fit to the M-A plot. The LOESS scatter plot smoother performs robust locally linear fits. In particular, it will not be affected by a small percentage of differentially expressed genes, which will appear as outliers in the M-A plot.

LOESS Description

The name "LOESS" is derived from the term "locally weighted scatter plot smooth," as the method uses locally weighted linear regression to smooth data. The smoothing process is considered local because, like the moving average method, each smoothed value is determined by neighboring data points defined within the span. The process is weighted because a regression weight function is defined for the data points contained within the span. In addition to the regression weight function, you can use a robust weight function, which makes the process resistant to outliers. Finally, it uses a linear or a quadratic polynomial for regression.

The local regression smoothing process follows these steps for each data point:

1. Compute the regression weights for each data point in the span. The weights are given by the function shown below.

$$\omega_i = \left(1 - \left|\frac{x - x_i}{d(x)}\right|^3\right)^3$$

   x is the predictor value associated with the response value to be smoothed, x_i are the nearest neighbors of x as defined by the span, and *d(x)* is the distance along the abscissa from x to the most distant predictor value within the span. The weights have these characteristics:

   a. The data point to be smoothed has the largest weight and the most influence on the fit.

b. Data points outside the span have zero weight and no influence on the fit.

2. A weighted linear least squares regression is performed. The regression uses a first or second degree polynomial (in this work only the first degree polynomial was used.
3. The smoothed value is given by the weighted regression at the predictor value of interest.

If the smoothing calculation involves the same number of neighboring data points on either side of the smoothed data point, the weight function is symmetric. However, if the number of neighboring points is not symmetric about the smoothed data point, then the weight function is not symmetric, therefore the span never changes.

Robust Smoothing Procedure

If the data contains outliers, the smoothed values can become distorted, and not reflect the behavior of the bulk of the neighboring data points. To overcome this problem, we can smooth the data using a robust procedure that is not influenced by a small fraction of outliers. The robust LOESS method includes an additional calculation of robust weights, which is resistant to outliers. The robust smoothing procedure follows these steps:

1. Calculate the residuals from the smoothing procedure described in the previous section.
2. Compute the robust weights for each data point in the span. The weights are given by the function shown below.

$$\omega_i = \begin{cases} \left(1-(r_i/6MAD)^2\right)^2 & |r_i| < 6MAD \\ 0 & |r_i| \geq 6MAD \end{cases}$$

r_i is the residual of the ith data point produced by the regression smoothing procedure, and MAD is the median absolute deviation of the residuals:

$$MAD = median(|r|)$$

The median absolute deviation is a measure of how much the residuals are spread out. If r_i is small compared to $6MAD$, then the robust weight is close to 1. If r_i is greater than $6MAD$, the robust weight is 0 and the associated data point is excluded from the smooth calculation.

3. Smooth the data again using the robust weights. Calculate the residuals from the smoothing procedure.
4. Repeat the previous two steps for a total of five iterations. [49]

5.4.1.2 About Window Size of LOESS

The span in the LOESS algorithm is a user-defined parameter and is the fraction of the data used for smoothing at each point; the larger the f value, the smoother the fit.
The span should be large enough to capture all the nonlinearity in the $\log_2$ ratios, at the same time it should be small enough to reduce the risk of over smoothing the data.
In our study, when all of the data from one slide (52272 spots) is used in LOESS smoothing, a span of more 10% means that for each data point more than 5000 neighboring data points participate in the local regression. The neighborhood size in this case is big enough to assume that spots with significant copy number changes won't affect the regression, especially since a robust version of LOESS is being used.
To further observe the effect of the span size on the normalized $\log_2$ ratios, the LOESS smoothing was performed with three different span sizes, 10%, 25% and 40% for slides H526-1 to H526-8. As you will see in the evaluations section, the choice of the span size does not have a significant effect on the smoothed curve. So a span size of 10% was chosen for normalizing the data using this method.

5.4.1.3 Spatial Normalization

Based on the observations described in section 5.3.6, the spatial dye bias is caused by changes in the slope of the linear function. Accordingly, we assume that the spatial bias is multiplicative and can be removed by estimating the bias at each spot and divide the ratio by the estimated bias. The spatial trend is estimated by computing, for each spot, the median ratio over its spatial neighborhood.
The size of the smoothing element in the spatial median filtering is an important choice to make. It should be small enough to be able to capture the general trend in the ratios and it should be large enough to reduce the risk of over fitting. We wanted to choose a window size that assures us that there are enough spots in the window that the outliers can not shift the median value and by outliers we mean both biological outliers and artifacts.

Since each consecutive three rows are replicates in our arrays, we also needed to choose an asymmetrical window which is longer in the columns than in the rows to make sure that the replicates are not being over emphasized.

In addition, we wanted to remove the "general" trend only, so that after removing it, the plate effects are still there. This is because we believe that plate normalization performs better for removing the plate effects. If the window size is chosen so small that enables the spatial normalization to remove the plate effects, there is a chance that some of the true biological variations are also removed by the normalization. But if the window size is bigger, the remaining plate effects after normalization can be removed by the plate normalization as will be described later. The plate normalization uses all the spots that are printed from the same plate group for normalization, so chances that the plate normalization removes the true biological variations are smaller.

We chose window size of 11×11 of "unique" spots which when taking the replications into account, corresponds to a window of $11 \times 3 = 33$ rows and 11 columns. We found the mentioned element size suited to pick up global trend while not being skewed by the altered clones.

A median filter was chosen rather than a mean filter so that amplified or deleted clones being outliers in each group of spots (pixels), would not affect the spatial trend estimates. Hence, altered clones will remain altered after this step as long as they are sufficiently isolated spatially, with just the spatial trend element of their values removed.

The median filtering is performed with symmetric padding at the edges, i.e. values outside the bounds of the input array are computed by mirror-reflecting the array across the array border.

Before performing median filtering, we filter the spots from the "bad" plates (as explained in the previous chapter) and interpolate the surface at the filtered points using bilinear interpolation.

After finding the spatial trend, the ratios are rescaled by the median filtered ratios i.e. each spot's ratio is divided by the median ratio of the spots in its neighborhood.

As stated earlier, the basic assumption for this step of normalization is that there is no predetermined order of spotting the clones with similar sequences or consecutive sequences in the same neighborhood.

5.4.1.4 Plate Effects

Assuming that the average of $\log_2$ ratios of spots from each plate group should be zero, we find the median $\log_2$ ratio of each plate group and subtract it from all the $\log_2$ ratios of the group so that the median $\log_2$ ratio of all plate groups equals zero after the normalization.

The assumption of median $\log_2$ ratio of zero for each plate can be violated if the copy numbers of the clones on a plate biologically differ between the test and the control samples. We do not believe this is the case in the experiments our studies here.

Before doing this, we filtered out the spots from those plate groups that have means that are "significantly" different from the other plate groups.

5.5 Evaluation of the Performance of the Proposed Normalization Strategy

To evaluate the performance of the proposed normalization strategy, we compare the accuracy (closeness of the measurements to the truth) and precision (closeness of the replicate measurements to each other) of the data obtained by different normalization methods for a given number of replicates.

There are two types of errors introduced to every measurement, systematic and random. Random error is a measure of uncertainty in the measurement and is therefore central to statistical inference. Random errors reflect inevitable uncertainties in all scientific measurements, making statistical procedures necessary. Systematic errors are biases; they result in a constant tendency to over or underestimate true values, thereby decreasing accuracy [29].

Accuracy depends on our ability to remove the systematic error contained in microarray data. For self-self experiments, the data points should center around the zero line on an M-A plot and this can be used for the assessment of the accuracy of normalized data. In general, for non-self-self data, it is not known around what line the data points should center. The precision of normalized data can be assessed by the data consistency; this is represented by the difference of normalized data from replicate experiments. The comparison of this difference between different normalization techniques applied to the

same data sets can answer the question of which normalization approach works better [41].

In gene expression arrays, the assumption that gene expression values obtained from biological replicate experiments are the same (provided that there is no error in the experiment) is not valid; however this assumption is true for CGH microarrays.

So, we perform our proposed stepwise normalization and some other normalization schemes (including the state-of-the-art techniques and more basic ones) and compare the consistency of the data obtained by performing each normalization method on the data from each replicate.

We would like to state that at the time of performing this study the state of the art in normalizing the two-channel gene-expression microarray data was print-tip LOESS normalization [20]. The normalization method used to normalize the data in our lab was global normalization.

As for the different methods, we started with simple one-step normalization methods and added to the complexity gradually. The methods that are compared are summarized in table 5-2. (In the following figures each of these methods will be referred to by its index number in table 5-2.)

In the following sections the detailed description of the four sets of experiments, which were done to assess the performance of the normalization, are presented. In this section, experiments are described and the quantities that were measured are introduced and their values are reported through out figures and tables. The results of each set of experiments will be discussed in section 5.6.

Index	Method Name	Description	Performed on
1	No normalization		ratio of background subtracted intenisties
2	Global mean scaling	ratios scaled by their mean	
3	Print tip mean scaling	ratios of each print-tip group scaled by the mean ratio of that group	
4	Global loess, span=10%		
5	Global loess, span=25%		
6	Global loess, span=40%		
7	Print tip loess, span=40%	loess performed on the ratios from each print-tip group	
8	Spatial	previuosly described	
9	Spatial + Plate	spatial followed by plate normalization	
10	Loess + Spatial	loess followed by spatial	
11	Loess + Spatial + Plate		
12	Spatial + Plate + Loess		
13	No Normalization		ratio of non-background subtracted intenisties
14	Global mean		
15	Print tip mean		
16	Global loess, span=10%		
17	Spatial		
18	Loess + Spatial		
19	Loess + Spatial + Plate		

Table 5-2 Summary of methods that are evaluated

5.5.1 Self-Self Experiments

The self-self experiments (slides MM-1 through MM-4) were used to study the effect of normalization on removing the bias from the data and increasing the accuracy.

In these experiments the samples to be compared are both from the same male genomic DNA pool and are labeled separately with cy3 and cy5 dyes. So the copy number of all the sequences is expected to be the same in both samples resulting in zero value for the $\log_2$ ratio of intensities. This is a special case in which we know what exactly the $\log_2$ ratios should be.

The effects of normalization in removing the bias were examined by calculating the following measures on the data normalized by each of the normalization methods:

- Standard deviation (s.d.) of the $\log_2$ ratios for each slide
- Correlation of $\log_2$ ratios and the smoothed $\log_2$ ratios as a function of the intensities (intensity dependent "local correlation")
- Correlation of $\log_2$ ratios and the smoothed $\log_2$ ratios as a function of the spatial location (spatial "local correlation")

Figure 5-18 shows the standard deviation of $\log_2$ ratios of each of the MM slides. Ten percent of the spots with the lowest average intensities were extracted and the standard devotions were calculated again. The new s.d. values are plotted in the same figure with a darker color.

Figure 5-19 shows the correlation of $\log_2$ ratios and $\log_2$ ratios smoothed against their corresponding spot's average intensities for each slide and for each method. Figure 5-20 compares the local correlation of $\log_2$ ratios and $\log_2$ ratios smoothed against their spatial location.

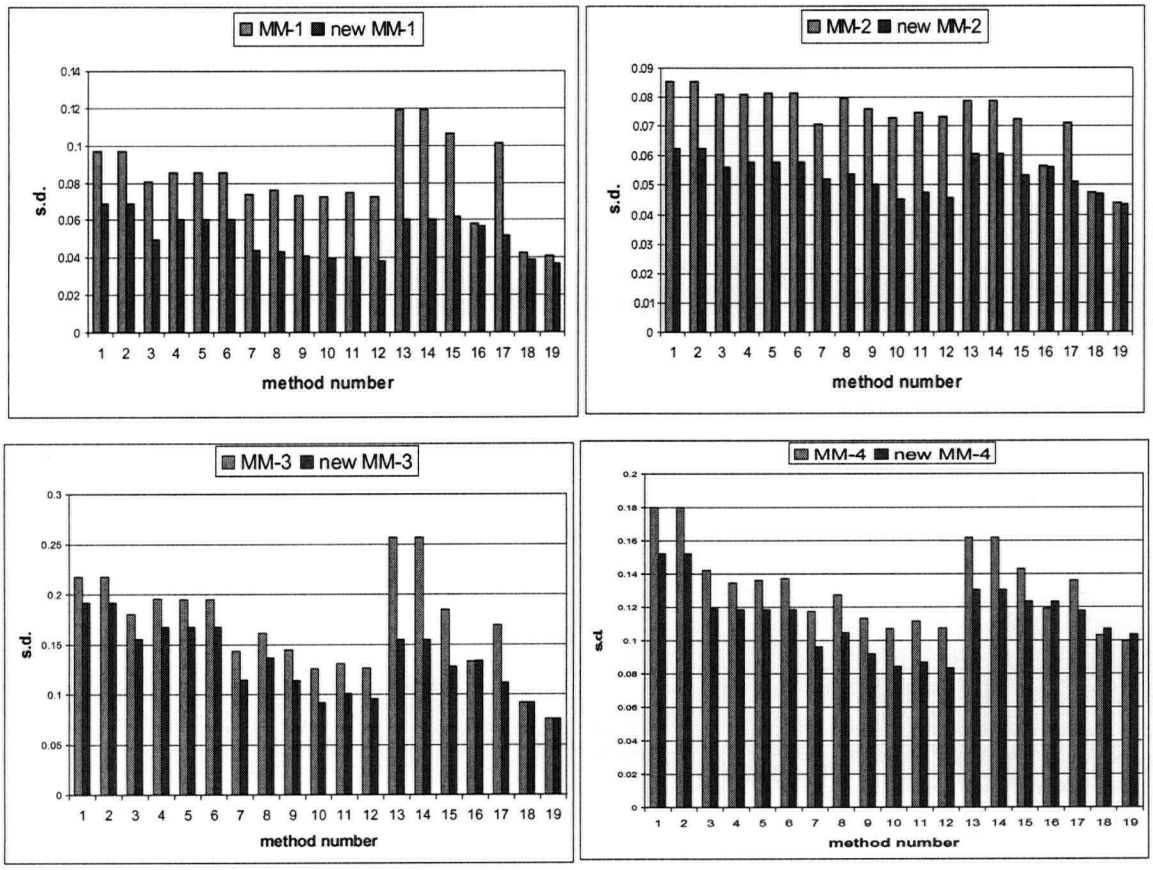

Figure 5-18 S.d. of $\log_2$ ratios after normalization for slides MM-1 through MM-4, (horizontal axis represents the method number (refer to table 5-2))

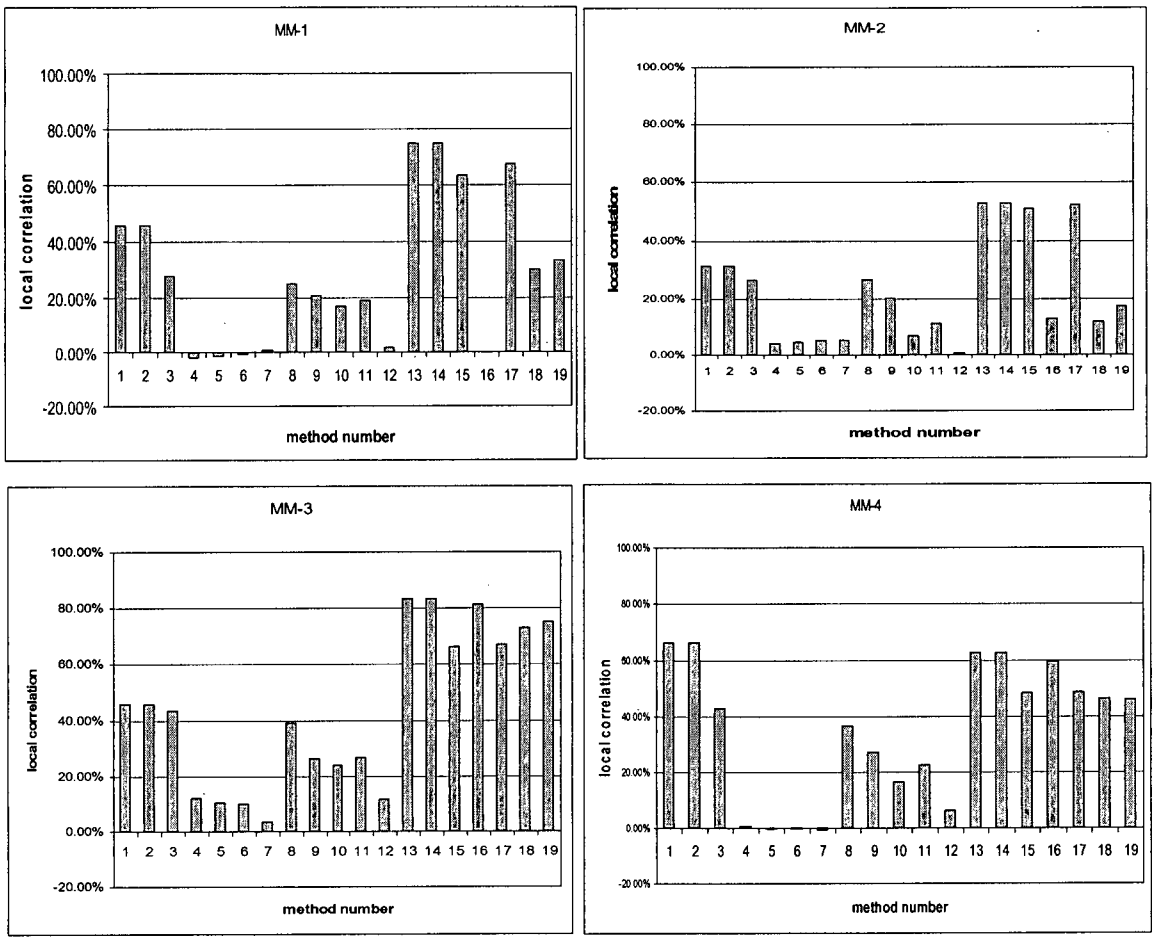

Figure 5-19 Local correlation of $\log_2$ ratios with the $\log_2$ ratios smoothed versus intensity after each normalization method (horizontal axis represents the method number (refer to table 5-2))

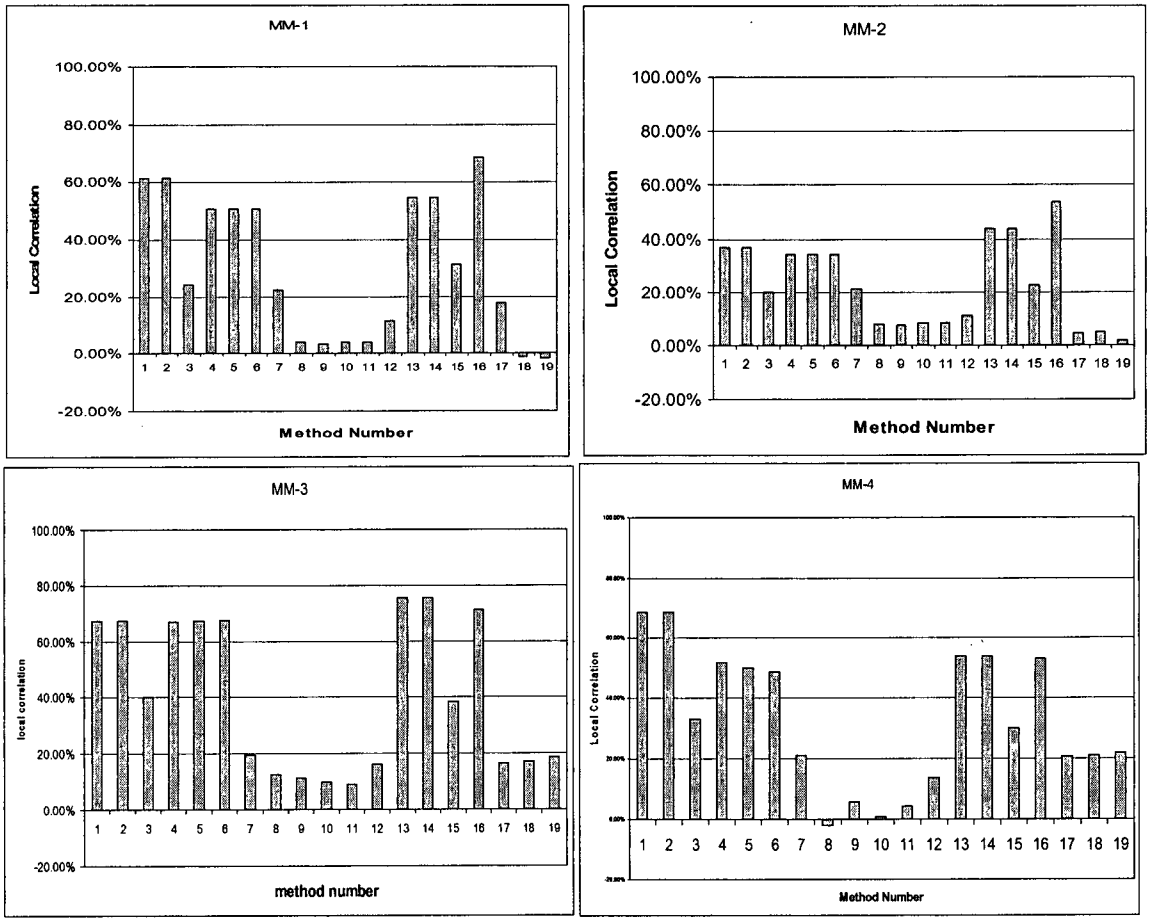

Figure 5-20 Local correlation of $\log_2$ ratios with the $\log_2$ ratios smoothed versus spatial location after each normalization method (horizontal axis represents the method number (refer to table 5-2))

5.5.2 Replicate Experiments

In order to see how the normalization affects the consistency of the data, 8 biological replicate experiments were performed (biological replicates in the sense that the hybridization involves DNA from different extractions, i.e. from different samples of cells from a particular cell line, this design involves a higher degree of variation in measurements compared to technical replicates who involve DNA from the same extraction).

The biological replication can not be used for gene expression microarrays as the actual expression values of genes are different in different biological samples. But due to the

different nature of array CGH technology and the fact that it measures changes in the DNA copy number, the same results are always expected from biological replicates. We used three different measures of repeatability of data which are as follows:

1. The **Standard deviations** of the $\log_2$ ratios of the same spot across the 8 replicate slides were calculated and averaged across all the spots for each normalization method. According to table 5-3, $M_{i,j}$, i=1,..,8 and j=1,..,52272 (52272 is the total number of spots in one slide.), refers to the $\log_2$ ratio of spot j in slide i. S_j is the s.d. of $M_{i,j}$ for i=1,..,8. The average of the standard deviations will then be:

$$\sum_{j=1}^{52272} S_j \Big/ 52272$$

The results are shown in figure 5-21.

	Spot 1	Spot 2	...	Spot 52272
Replicate slide1	$M_{1,1}$	$M_{1,2}$	...	$M_{1,52272}$
Replicate slide 2	$M_{2,1}$	$M_{2,2}$	...	$M_{2,52272}$
...	...	...		...
Replicate slide 8	$M_{8,1}$	$M_{8,2}$	...	$M_{8,52272}$
s.d. of $\log_2$ ratios	S_1	S_2	...	S_{52272}

Table 5-3 Calculating the average of s.d. of $\log_2$ ratios

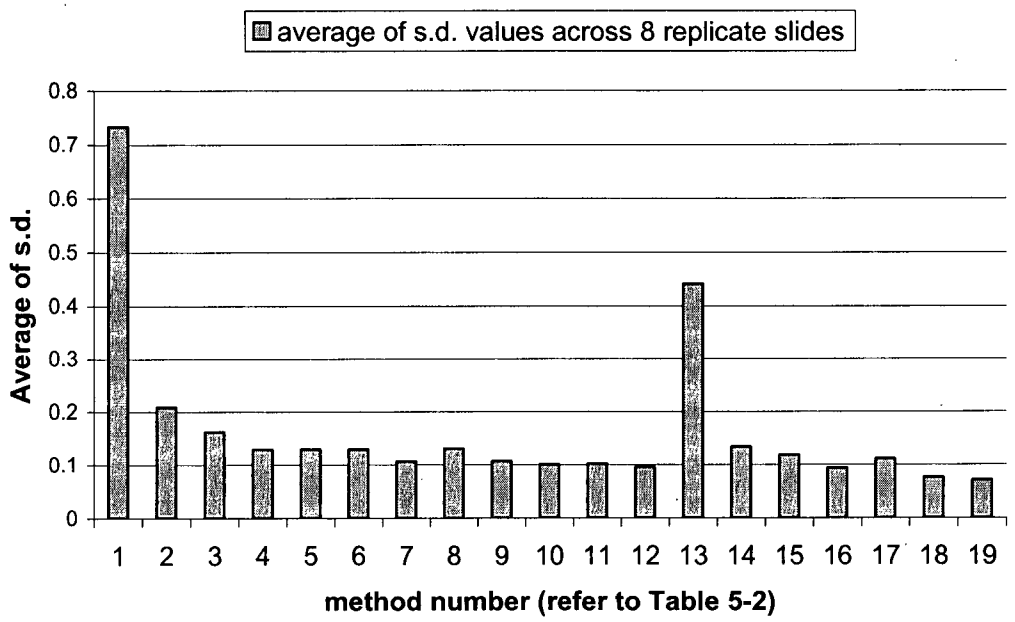

Figure 5-21 Average of the standard deviations of each spot's $\log_2$ ratio across replicate slides (horizontal axis represents the method number (refer to table 5-2))

2. **Pearson's Correlation Coefficient** which represents the covariance of normalized variables. Each variable is normalized by subtracting the variable mean and dividing by its standard deviation. When there is a linear relationship between the two variables, the correlation will have a value of 1, indicating perfectly matched order of the two variables. A value of -1 indicates a perfect negative covariation. A correlation value of 0 indicates a completely random relationship between the two variables.

The correlation coefficient was calculated for the data from each pair of the replicate slides. thus 28 different correlations were calculated for 28 pairs of slides. The average of the 28 correlation coefficients for each single method was calculated. The results are shown in figure 5-22.

3. **Intraclass Correlation Coefficient (ICC):** This is an ANOVA-based type of correlation. It measures the relative homogeneity within groups in ratio to the total variation:

$$r_{ICC} = \frac{\left(MS_{Between\ \ groups} - MS_{Within\ \ groups}\right)}{\left(MS_{Between\ \ groups} + (n-1) * MS_{Within\ \ groups}\right)}$$

where n is the number of cases in each category of the measured variable, $MS_{between\ groups}$ is the estimate of the between group variance, and $MS_{within\ groups}$ is the estimate of the within groups variance. These estimates are taken from the ANOVA tables.

The maximum value of the intraclass correlation coefficient is 1.0, but its maximum negative value is (-1/(n-1)). Intraclass correlation coefficient is large and positive when there is no variation within the groups, but the group means differ. It will be at its largest negative value when group means are the same but there is great variation within groups. A negative intraclass correlation occurs when between-group variation is less than within-group variation [50].

ICC was shown to be useful for the assessment of technical and biological variations in microarray experiments in [54].

The ICC was calculated for the set of data obtained from 8 replicate slides and normalized using each of the methods described above. The results are shown in figure 5-22.

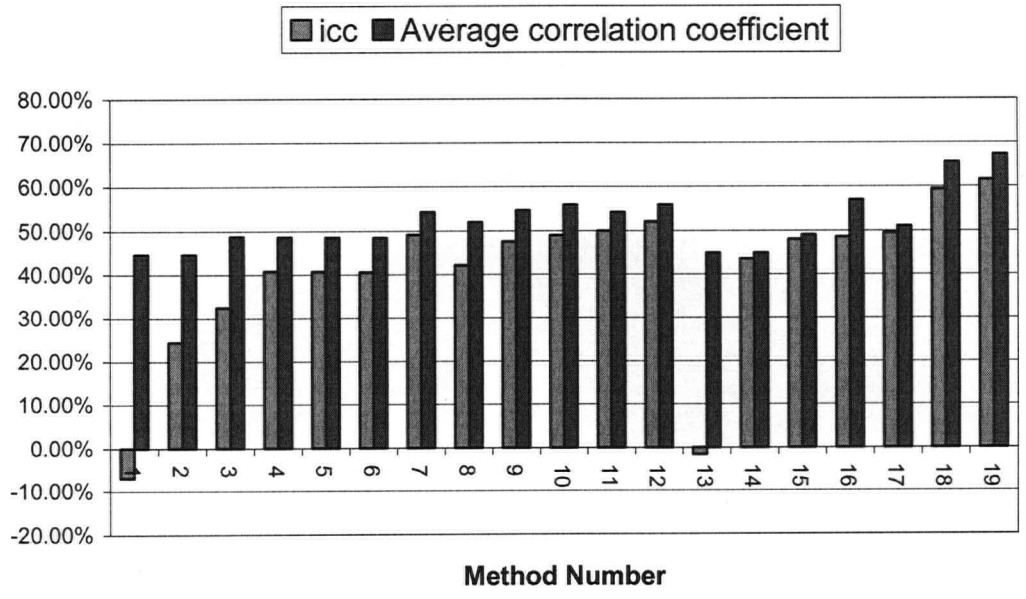

Figure 5-22 ICC and Average correlation coefficient of replicate slides (horizontal axis represents the method number (refer to table 5-2))

5.5.3 Male-Female Hybridizations

As another experiment in evaluating the effect of normalization on the microarray data, two experiments were conducted with male genomic DNA as one sample and female genomic DNA as the other sample. The first 22 chromosomes are in pairs for both males and females and as for the sex chromosomes, females have two X chromosomes and males have an X and a Y chromosome. This experiment simulated a single copy deletion by hybridizing normal male versus normal female DNA, generating a 1:2 ratio of X chromosomes and 2:2 ratio of chromosomes 1 to 22. So this experiment forms another case that the $\log_2$ ratios are known prior to the experiment.

The normalization methods described above were applied to the data from these two experiments and a t-test was performed on the $\log_2$ ratios of two groups of clones. The first group consists of clones from chromosomes 1 through 22 and the second group consists of clones from chromosome X.

For the t-test it is assumed that the two samples are from a normal distribution and the t-test is performed to test whether two samples could have the same mean when the standard deviations are unknown but assumed equal. The t-statistic is defined as:

$$T = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where s is the pooled sample standard deviation and n and m are the numbers of observations in the x and y samples.

The value of the T statistic is shown in figure 5-23 for both slides and for the results of each normalization method.

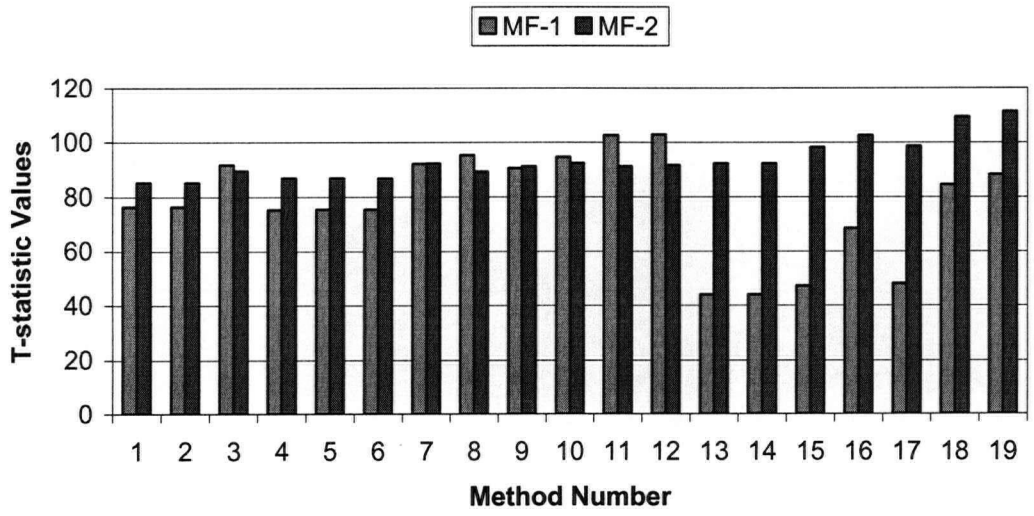

Figure 5-23 Values of T-statistic after each normalization method for slides MF-1 and MF-2 (for MF-2, the T-statistics are multiplied by -1 for comparison purposes) (horizontal axis represents the method number (refer to table 5-2))

5.5.4 Titration Experiments

Tumors contain a number of normal cells. Contamination from normal cells may affect the ability to detect copy number aberrations. In case of an amplification, contamination from normal cells (DNA sequences with copy numbers of 2) makes the average copy number of the sequences in the test sample smaller than what it really is. In case of a deletion, contamination from normal cells increases the average number of copies of DNA sequences in the test sample. In this experiment, we show that normalizing the spot data can allow us to handle more contamination in the tumor sample.

We used the data from a titration experiment that compares X chromosome loci (clones) to autosomal (non-sex) loci by comparing male and female DNA. A single copy deletion was simulated by hybridizing normal male versus normal female DNA, generating a 1:2 ratio of X chromosomes. Contamination from normal cells was then simulated by spiking varying amounts of female DNA into the male DNA sample (slides T1-T5). Single copy amplifications were modeled by comparing a 50/50 mixture of male and female DNA against a male DNA reference. In this model contamination from normal

cell was simulated by spiking varying amounts of female DNA into the male/female DNA mixture (Slides T-6 to T-10).

For these slides only the results of normalization with our proposed stepwise normalization (spatial + plate + LOESS normalization on the background subtracted data) were used. T-test was performed on two groups of data from each slide. The first group consists of clones from chromosomes 1 through 22 and the second group consists of clones from chromosome X. T-values are shown in Figure 5-24.

The purpose of this experiment was mainly to see how the normalization affects the single copy number amplification or deletion with some degree of contamination which makes the real copy number ratio even smaller. The main concern in this experiment was to find out whether or not normalization removes some of the biological variations i.e. whether or not it over normalizes the data.

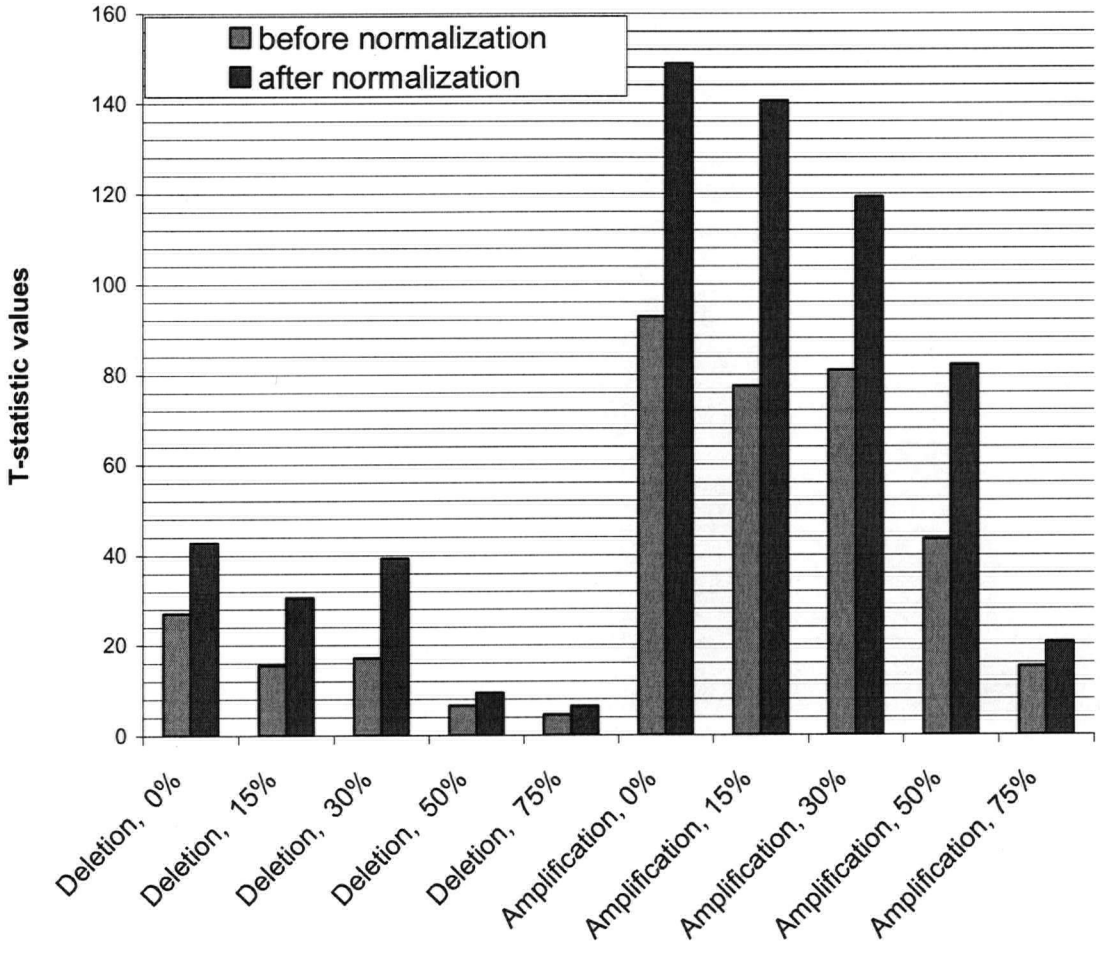

Figure 5-24 T-statistic values before and after normalization for the titration experiment slides (T1-T10)

5.6 Results and Discussions

A normalization scheme is expected to remove the systematic variations in the data and leave the true biological variations unchanged. So in evaluating the performance of the normalization methods, both of these issues should be considered.

In this section the results from the four sets of experiments that were performed and described in sections 5.5.1 through 5.5.4 will be discussed.

1. The self-self experiments were used as a model to investigate the accuracy of the normalized data. The standard deviation of the $\log_2$ ratios was used as a measure of accuracy. A lower standard deviation for the normalized data shows the better performance of the normalization in removing the bias.

As figure 5-18 shows, among the normalization methods that are performed on the ratios of background subtracted intensities, the three step normalization results in lowest s.d. in all four slides. Also, among the normalization methods that are performed on the ratios of non-background subtracted intensities, the three step proposed normalization scheme results in lowest s.d.

Now, let us compare the results of normalization methods that are performed on ratios of background-subtracted intensities and non-background subtracted intensities. It is observed that ratios of non-background subtracted intensities result in quite high s.d. values if they are not normalized or if only spatial normalization is performed. This result conforms to our observations of diagnostic plots of the ratios. If background subtraction is not performed on the intensities before taking the ratios, the ratios of red to green intensities are always higher for lower intensities. This is due to the inherent lower intensity of the cy5 channel intensities than the cy3 channel which is more pronounced for lower intensity spots. However, if LOESS normalization is performed on the data, the standard deviations become lower. This is because the curvature in the M-A plot caused by the "background bias" (for the definition, refer to section 5.3.1) is removed by the LOESS normalization. In fact, when the three-step proposed normalization is performed on the ratios of non-background subtracted intensities, it has better performance, in terms of reducing the s.d. of $\log_2$ ratios, than when it is applied to the ratios of background-subtracted intensities.

We suspect that this is not because some part of the variation of the $\log_2$ ratios, which we measure in form of the s.d., is not systematic; but is because of random variations of ratios. Background subtraction adds to the random variation of the intensities of the spots because of the uncertainties in estimating the background intensity. At lower spot intensities, this added noise becomes significant when compared to the intensity of the spot. As a result if background subtraction is not performed and the bias that it

introduces into the data is removed by the LOESS normalization, then the s.d. of ratios will decrease.

To test this idea, 10% of the lowest intensity spots were removed and standard deviations were calculated again for these four slides. As figure 5-18 shows not only the standard deviation of the new set of spots is lower than the original set, but also the s.d. of the ratios after normalization is now comparable for both cases of background subtracted and non-background subtracted intensities.

This suggests that the reduction of standard deviation of $\log_2$ ratios may not be the best indicator of reduction of the "systematic" variations. The other two measures, intensity dependent "Local correlation" and spatial "Local correlation", are more appropriate for the purpose of measuring the intensity dependent and spatial biases.

Figure 5-19 compares the intensity dependent bias obtained after each normalization method is performed on the data. From this figure it is observed that LOESS normalization is most effective in removing the intensity dependent bias when it is performed on ratios of background subtracted intensities (compare columns 11 and 19 in all four sub plots).

As figures 5-19 and 5-20 show the intensity dependent bias becomes smallest after global LOESS or print-tip LOESS normalization is performed on the data. Spatial normalization reduces the intensity dependency only slightly. Spatial normalization followed by LOESS is almost as effective as LOESS performed alone but when the LOESS normalization is followed by spatial normalization the local correlation is increased relative to when LOESS is performed alone or after spatial normalization. The same thing happens when spatial bias is measured. Spatial normalization has the best performance either alone or when performed after LOESS normalization but if it is followed by LOESS normalization, it is less effective in removing the spatial bias. This is an important result and indicates that if the LOESS normalization and spatial normalization are performed iteratively, their performance might improve in terms of reducing the bias.

2. To compare the effects of normalization on the precision (repeatability) of normalized data, 8 replicate slides were used. Three different measures of

repeatability, s.d., Pearson correlation, and intraclass correlation were used to compare the precisions.

The standard deviation becomes smallest after the three-step proposed normalization scheme is performed on the data. When the proposed normalization is performed on the ratios of non-background subtracted data, its performance is slightly better than when it is performed on ratios of background subtracted intensities.

The ICC and Pearson correlation coefficient conform to each other in almost all the methods except for non-normalized ratio. For methods 1 and 13, the ICC is quite low while the correlation coefficient is high. The reason is that without normalization, the averages of log ratios of different replicate slides are different. Correlation coefficient is not affected by the averages of the data; However ICC gets quite low in this case, because of the fact that between group variation is lower than within group variation.

Both ICC and Correlation coefficient are highest after the three-step normalization method. This applied for both the ratios of non-background subtracted intensities and ratios of background subtracted intensities. ICC and Correlation coefficient are slightly higher when background subtraction is not performed on intensities.

3. In the third study, the T-statistic values were calculated for the results of each normalization method for slides MF-1 and MF-2 and were compared to determine which method results in best separation of clones with no copy number change and those with a single copy number change (chromosome X clones). A larger value for the T-statistic shows higher separation between the means of the two samples.

For the data from slide MF-1, the largest T-statistic is obtained after spatial normalization followed by plate and LOESS normalizations are performed on ratios of background subtracted intensities. For this slide normalization methods performed on ratios of non-background subtracted intensity are not as effective. Especially if LOESS normalization is not performed on the ratios of non-background subtracted intensities, the T-statistic is even lower than that of non-normalized ratios of background subtracted intensities.

For MF-2 data, normalization methods do not significantly change the value of the T-statistic. The three-step normalization performed on the ratio of non-background

subtracted intensities slightly increases the T-statistic. As figure 5-14 shows the intensity dependent bias and the spatial bias are both quite low for this slide compared to the other slides (below 15%). This explains the fact that the T-statistic values do not change after normalization. Also the average of the background intensities for this slide is quite low compared to the other slides (see figure 5-6). This explains why performing and not performing background subtraction doesn't have a significant effect on the T-statistic of the ratios.

It should be noted that according to the results of previous experiments, the performances of methods 11 and 12 from table 5-2, which are "LOESS + spatial + plate" and "spatial + plate + LOESS" respectively, were "about" the same. In other words, the order of performing the spatial and intensity dependent normalizations did not affect the results.

4. The purpose of the last experiment (Titration experiment) was to find out how the normalization affects the ratio of single copy number changes with some degree of contamination of reference sample into the test sample. The normalization method was the three-step normalization performed on the ratios of background-subtracted intensities (method number12 in table 5-2).

The T-statistic values are higher after normalization in all cases which assures us that the separation of the groups is increased and the small copy number changes are maintained and even magnified. (As slides T1 to T5 simulate a single copy number deletion the T-statistics are negative.)

One significant result of this experiment is that the proposed normalization scheme enables us to deal with contamination of reference sample into the test sample by increasing the separation between the distributions of the normal and abnormal clones. In figure 5-24, for example, the T-statistic corresponding to slide T-9, which simulates a single copy amplification with 50% contamination, after normalization becomes quite close to the T-statistic of the slide T-6 that simulates a case with no contamination.

In figures 5-25 through 5-28, chromosome plots of the data from two of the replicate H526 slides, generated by SeeGH software [19], are shown. **Chromosome plots** show base two logarithm of ratios for each of the target DNA clones, as a function of the location of the clone in the chromosome. Figures 5-25 and 5-26 show the chromosome plots for chromosomes 1 and 2 of slide H526-5 respectively. Figures 5-27 and 5-28 show the chromosome plots for chromosomes 1 and 2 of slide H526-1 respectively. For each slide and each chromosome the $\log_2$ ratios are shown after global normalization and after three-step normalization.

The variability of $\log_2$ ratios from slide H526-5 is much higher than that of slide H526-1. In fact, data from slide H526-5 were the least reproducible data according to the Pearson correlation coefficients of each pair of H526 slides. In contrast, data from slide H526-1 were among the most reproducible datasets (refer to figure 5-10).

For the H526 cell line, the regions of copy number change are known [17]. The region of amplification on chromosome one and the micro-amplification on chromosome 2 are marked on the plots.

As the figures show, for data from slide H526-5 (low quality data), the normalization reduces the unwanted variations so that after the three-step normalization, due to the reduced variations, the altered regions are clearer. For data from slide H526-1 (high quality data), where the variation of the $\log_2$ ratios is quite low even before normalization, the important point to note is that normalization does not remove the true biological variations.

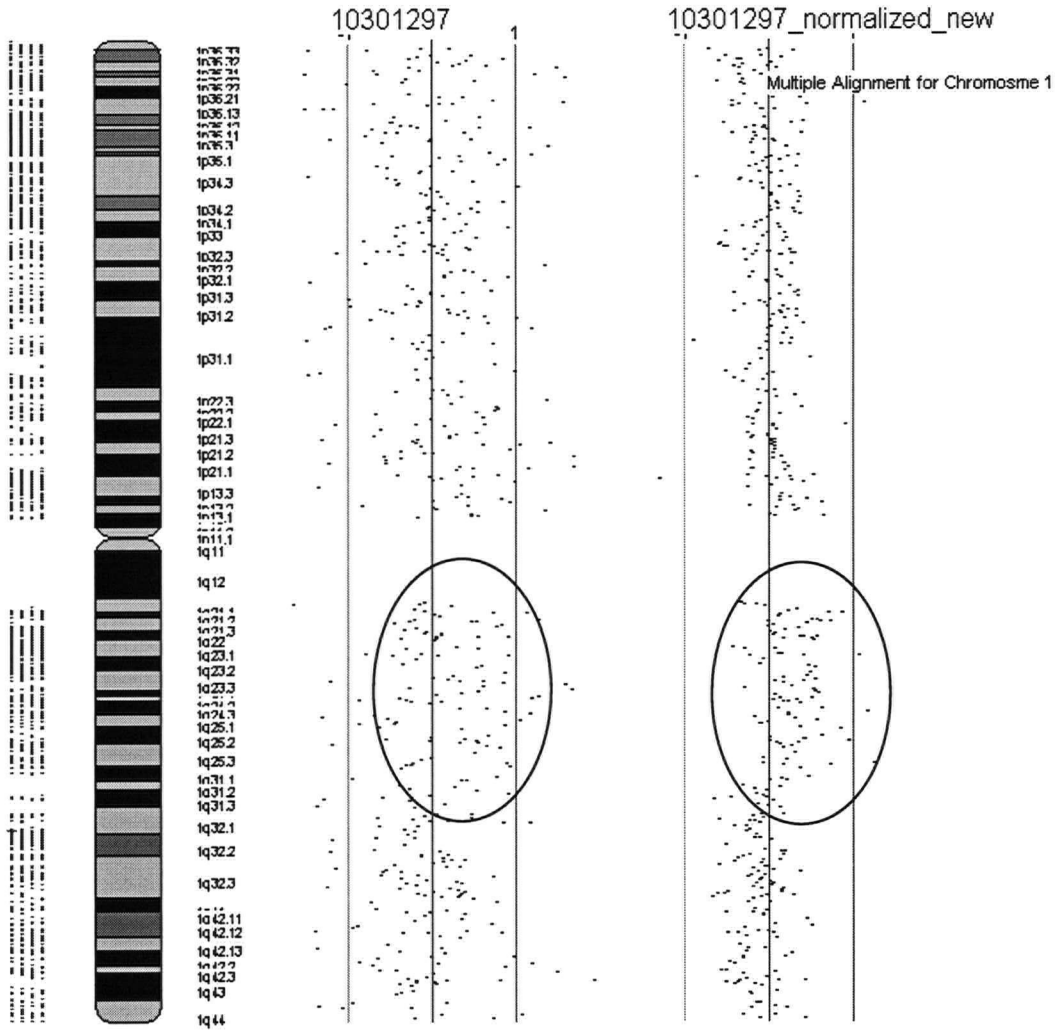

Figure 5-25 Plot of $\log_2$ ratios of clones from chromosome 1 versus their location across the chromosome, left: after global normalization, right: after the three-step proposed normalization, data from slide H526-5

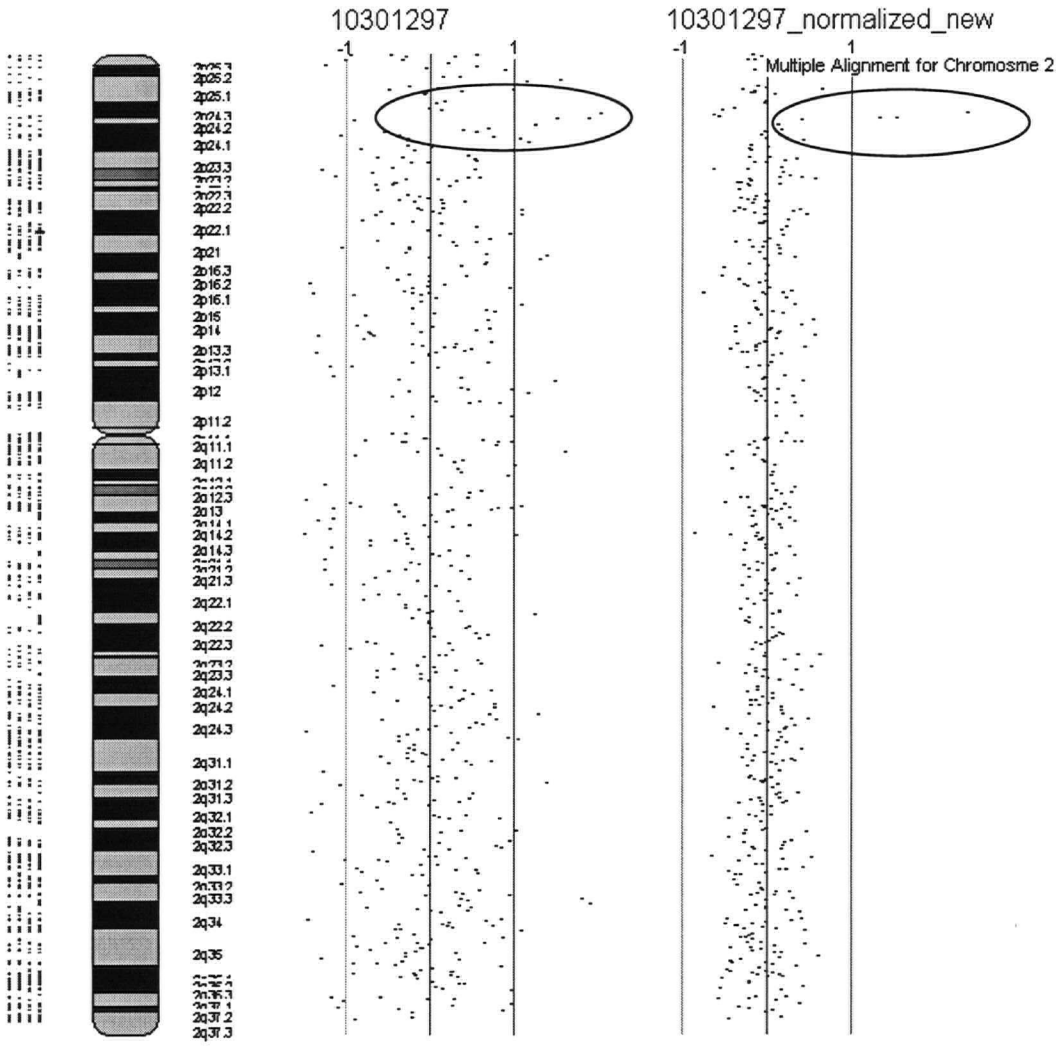

Figure 5-26 Plot of $\log_2$ ratios of clones from chromosome 2 versus their location across the chromosome, left: after global normalization, right: after the three-step proposed normalization, data from slide H526-5

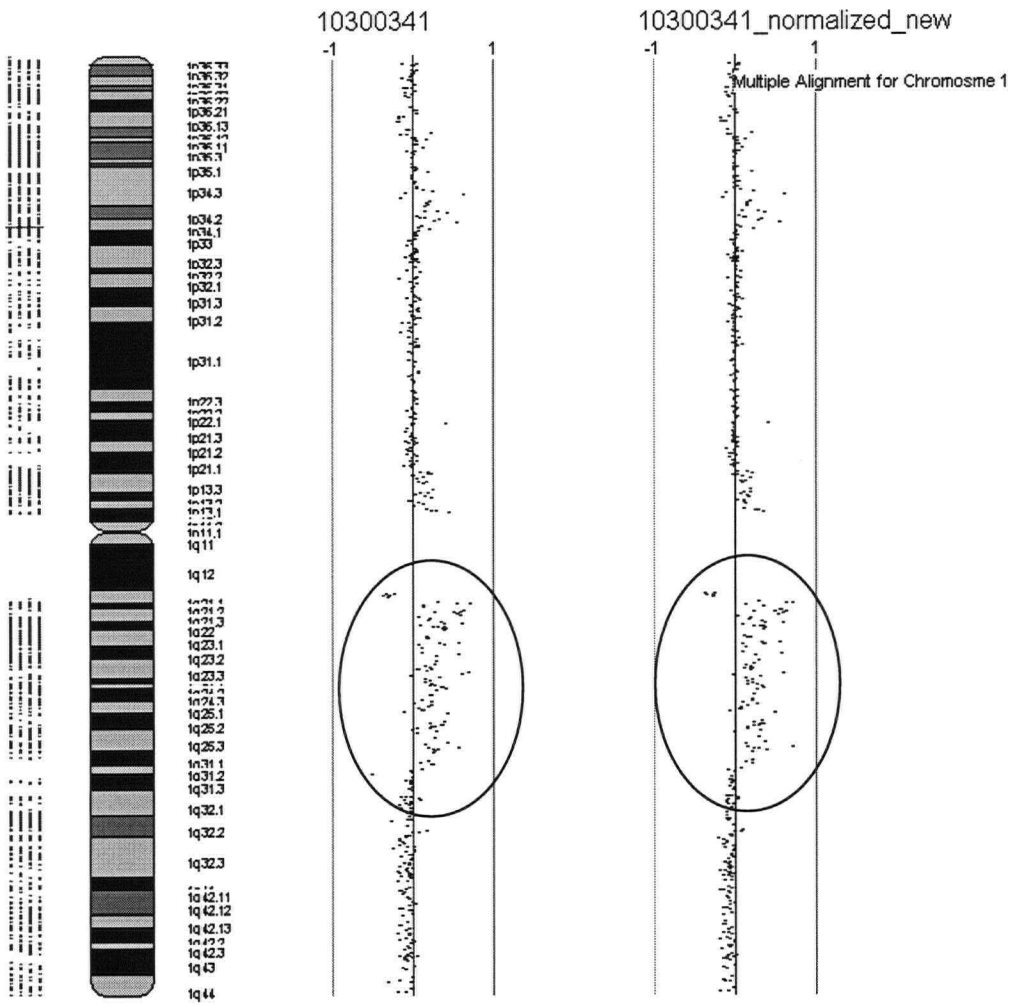

Figure 5-27 Plot of $\log_2$ ratios of clones from chromosome 1versus their location across the chromosome, left: after global normalization, right: after the three-step proposed normalization, data from slide H526-1. The three vertical lines for each plot are scale bars indicating of $\log_2$ ratios of -1, 0, and 1 from left to right.

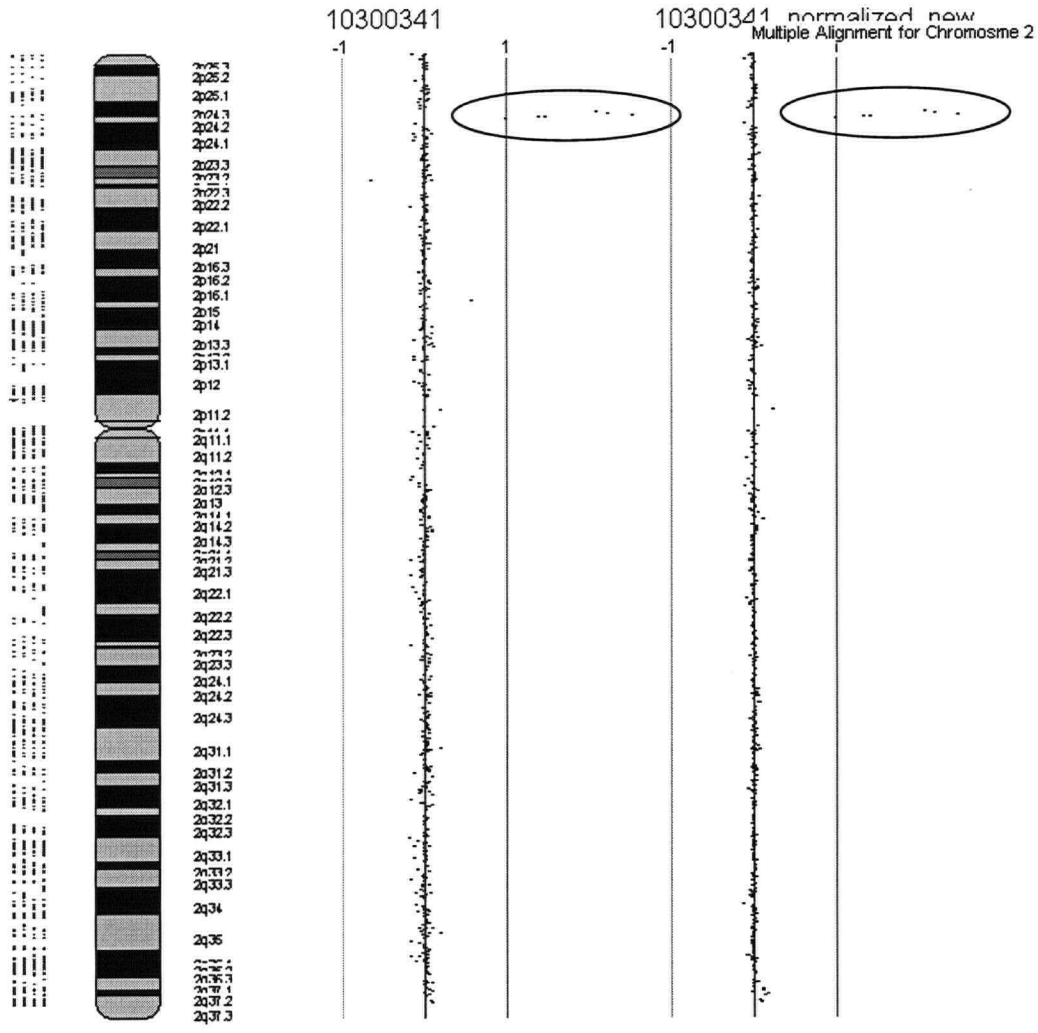

Figure 5-28 Plot of $\log_2$ ratios of clones from chromosome 2 versus their location across the chromosome, left: after global normalization, right: after the three-step proposed normalization, data from slide H526-1

5.7 Conclusions and Suggested Future Directions

Putting the results of the previous experiments all together, performing the proposed three step normalization increases the accuracy and precision of microarray data. It has better performance in terms of increasing the accuracy and precision of data than the print-tip LOESS normalization which is the most widely used normalization method for gene expression data and it of course has much better performance than the global normalization approach that is currently being used in our lab for normalizing the array-CGH data.

The proposed method was also tested for preserving the biological information while removing the systematic variations through a titration experiment and it was observed that even for copy number changes as low as single copy deletions with different degrees of contamination by normal DNA, the biological information are preserved (proven by the increased power of the T-test). This fact is the main difference of microarray CGH performance requirement compared to gene expression microarrays. As previously discussed, for gene expression microarrays, usually only genes with expression changes more than two fold are considered as differentially expressed.

The normalization methods were performed on the ratios of non-background subtracted intensities as well as the ratios of background subtracted intensities. It was observed that subtracting the background removes the "background bias", while not subtracting the background decreases the random variation especially for the lower intensity spots. The data from one slide of the male-female experiments which served as a model of single copy deletion showed that not subtracting the background may result in lower t-statistic values meaning the separation between the two group of normal and deleted clones has been decreased. Although the result is not quite conclusive as there are only two slides in this experiment but we believe that not subtracting the background may result in reduction of accuracy of the microarray data and suggest that further experiments be performed in this area.

Throughout this study, we found that the data from self-self experiments are correlated even after normalization. For self-self experiments, which simulate a case of no copy number change, the variability of the ratios is due to systematic or random errors. After the normalization, assuming that all the systematic errors are removed, the remaining variations are due to random error. Therefore the data from each self-self experiment is expected to be uncorrelated with the data from the other self-self experiments. However this was not the case. We smoothed the $\log_2$ ratios from slides MM-1 through MM-4 (the self-self experiments) as a function of the chromosomal location of the clones by a moving average function (with a window size of 64). When we compared the plots of the data before and after normalization, we observed that some of the correlations were removed; however, for some of the chromosomes the patterns of the normalized $\log_2$

ratios were still quite similar. Figure 5-29 shows two of the chromosome plots that show a strong correlation even after normalization.

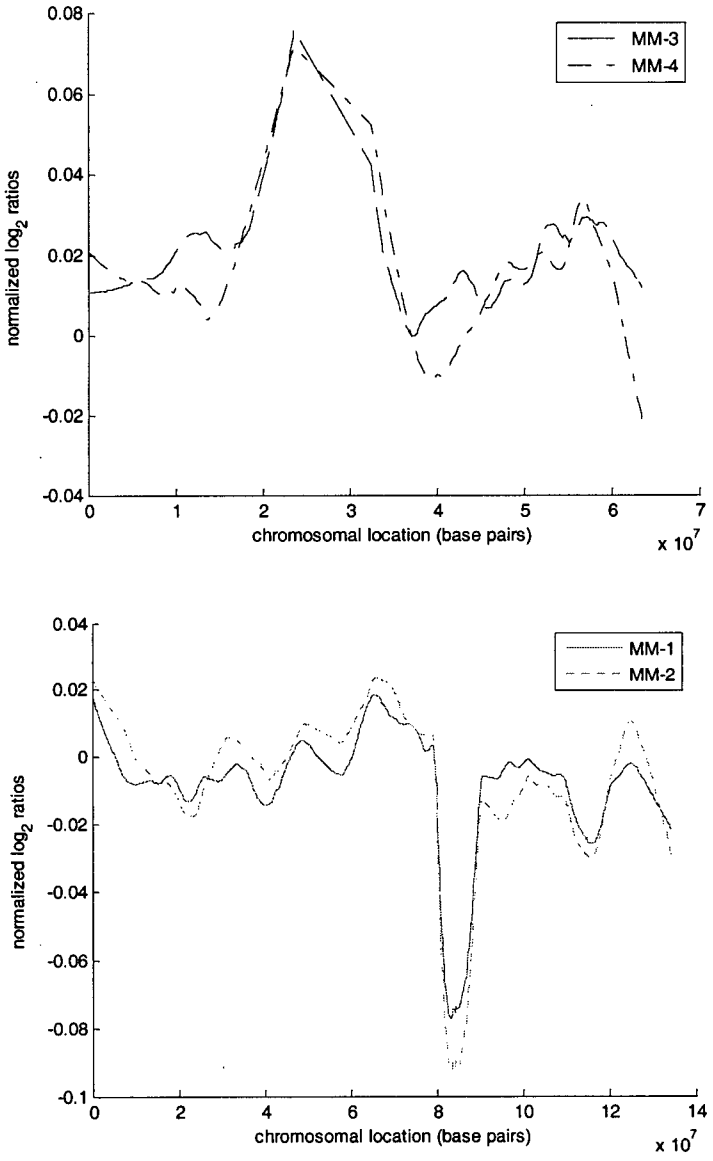

Figure 5-29 Log_2 ratios of clones plotted versus the chromosomal location shown for two chromosomes, upper: chromosome 11, lower: chromosome 19

This correlation indicates a systematic error that is dependent on the chromosomal location of the clones. We believe that more experiments need to be performed to further study this type of systematic error.

Throughout this study, we also observed that the **GC content** (the percent of G or C nucleotide bases in the sequence) of the target clone of each spot might have a systematic

effect on the intensities and the ratio corresponding to that clone. Patterns similar to the GC content of the clones plotted versus genomic order, are sometimes observed in the intensities and/or ratios corresponding to those clones. This might be due to the fact that the labeling of the sample DNA is done with dyes that are attached to the C nucleotide bases. We believe that this might be a source of systematic error and more investigation needs to be done in this area to find the source of this bias and ways to handle it. As an example, Figure 5-30 shows the $\log_2$ ratios and intensity of the clones of chromosome 7 plotted versus the genomic order along with the GC content of the clones.

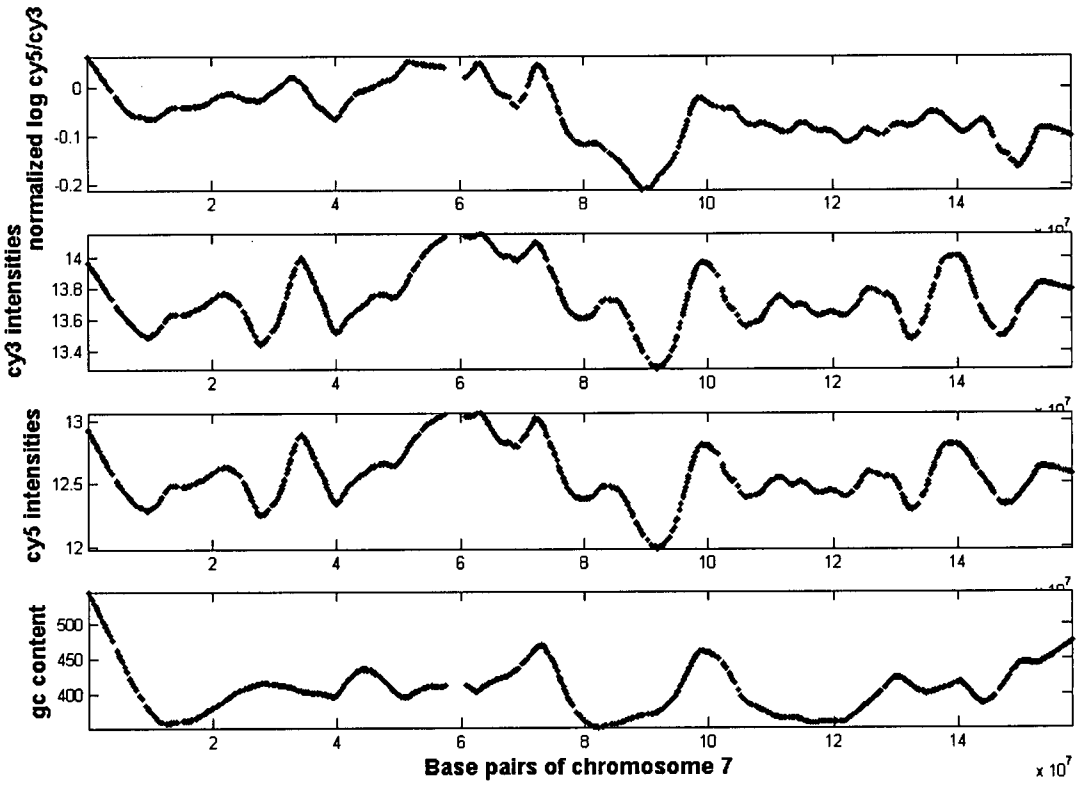

Figure 5-30 Log_2 ratios and intensity of the clones of chromosome 7 plotted versus the genomic order along with the GC content of the clones

One approach to a more complete study of normalization methods for microarray data is to perform the spatial and intensity dependent normalizations iteratively and to evaluate its performance in terms of accuracy and precision of data.

Finding better ways of dealing with the problem of background correction and its effect on the lower intensity spots will have a large impact on quality of microarray data. One approach will be to consider subtracting the background from the intensity of higher intensity spots and not subtracting the background for lower intensity spots.

Another open area for future work is to consider the effects of competitive hybridization in introducing bias into ratio measurements. One possible experiment for this study might be to design an array with the same target material but different target concentrations for the spots and investigate the effects of the target concentrations on the ratio of spots.

CHAPTER 6 CONCLUSIONS

This research focused on improving the "image analysis" step of array-CGH experiments. Specifically, two issues were addressed in this study: First, identifying the spots for which the detectable characteristics indicate that data from those locations are very likely to be unreliable, and second, normalization of the spot data to remove as much systematic variability (both experimental and device) as possible. These two tasks are both done with the aim of reducing the variability in the log ratios and increasing the validity of the results.

Regarding the first issue, filtering out the low quality spots, we aimed to identify the low quality spots in a more robust and efficient way than the current procedure. This was accomplished by the design of a binary decision tree with a threshold operator to identify the saturated spots, and two linear discriminant functions to identify the spots with irregular shapes and spots with circular shapes but defected in some other way.

The performance of the classifier in identifying the low quality spots was evaluated in an experiment by applying the classifier to the data from four microarray images and examining the reduction in variation that results from excluding low quality spots.

As part of this experiment, we demonstrated that a large part (in this study as high as 10%) of the variability of ratio measurements is due to low quality image spots. The binary decision tree was able to reduce the variability significantly by filtering out the low quality spots.

Use of our proposed method for filtering out low quality spots instead of filtering out the spots based on the variation of replicate spots on a slide (which is the current practice)

has a significant practical advantage. Using the new method enables us to reduce the number of replicate spots per target clone to two instead of three and this in turn enables the SMRT array clones that currently are spotted in triplicate across two slides to be fitted onto a single slide. This will make the array-CGH experiment significantly (almost a factor of two) more cost, material and time effective. We compared the results of triplicate filtering method and the new method in terms of reducing measurement variance and demonstrated that the two methods have comparable performances while the new method doesn't require the triplicates for quality filtering. This assures us that the SMRT array clones can be safely fitted into one slide without loss of measurement accuracy.

The second issue that was addressed in this research was the normalization of data. We investigated the systematic variations in the data in our microarray images. Based on our experimental observations of the systematic variations, we proposed a three-step normalization scheme to remove those systematic variations. We conducted four sets of experiments and developed a methodology to evaluate the performance of the normalization procedures. We showed that performing the proposed three-step normalization increases the accuracy and precision of microarray data. Our proposed method has better performance in terms of increasing the accuracy and precision of data than the global normalization approach, which is the approach currently taken in our lab for normalizing the array-CGH data. Our proposed scheme has also better performance than the print-tip LOESS normalization [38]. which is the most widely accepted and used normalization scheme for gene expression data, however to the best of our knowledge, its effect hasn't been evaluated on the CGH data previously.

The main difference in the performance requirement of CGH microarrays versus gene expression microarrays is the fact that for gene expression microarrays, genes with expression changes less than two fold are not considered differentially expressed and therefore are not of interest; however for CGH microarrays, detection of single copy number changes in contaminated samples that might be quite a bit smaller than two fold are of high interest. This requires the level of systematic variation removal to be significantly more stringent. The proposed method was tested such that it preserved the biological information while removing the systematic variations through a series of

titration experiments. It was observed that even for copy number changes as low as single copy deletions with varying degrees of contamination by normal DNA, the biological information could be preserved (demonstrated by the increased power of the T-test results from our normalization experiments).

6.1 Suggested Future Directions

With regards to the first study of this thesis, removal of low quality spots, it will be useful to track the reason for identifying a spot as of low quality to explain the cause of the low quality of the spots. This information can help in controlling the quality of microarray manufacturing and hybridization processes, and avoiding the artifacts by providing the experimenter with feedback.

With regards to the second study, normalization of spot data, our work strongly suggests that there would be significant benefit to a more complete study of normalization methods for microarray data such as to perform the spatial and intensity dependent normalizations iteratively and to evaluate the performance in terms of accuracy and precision of the resultant data.

Another useful study would involve developing improved methods of dealing with the problem of background correction and its effects on the lower intensity spots would have a large impact on the quality of microarray data. One approach to be considered for future work would be trying the subtraction of the background from the intensity of higher intensity spots and not subtracting the background for lower intensity spots.

Another open area for future work is to consider the role of competitive hybridization in introducing bias into ratio measurements. One possible experiment to examine this question would be to design an array with the same target material but different target concentrations for the spots and investigate the effects of the target concentrations on the ratio of spots.

REFERENCES

[1] Albertson D.G., Pinkel D. "Genomic microarrays in human genetic disease and cancer" *Hum Mol Genet,* 2003 Oct 15; 12 Spec No 2:R145-52.

[2] Sebastiani P.; Gussoni E.; Kohane I.S.; Ramoni M.F. "Statistical Challenges in Functional Genomics" *Statistical Science*, 2003; 18(1):33-70.

[3] Schena, M. "Microarray Analysis", Wiley-Liss, 2002.

[4] Kamberova G, Shishir S. "DNA Array Image Analysis Nuts & Bolts", DNA Press, 2002.

[5] Mantripragada K.K., Buckley P.G., de Stahl T.D., Dumanski J.P. "Genomic microarrays in the spotlight" *Trends Genet*, 2004 Feb; 20(2):87-94.

[6] Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P. "Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances" *Genes Chromosomes Cancer,* 1997 Dec;20(4):399-407.

[7] Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo W.L., Chen C, Zhai Y, Dairkee SH, Ljung B.M., Gray J.W., Albertson D.G. "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays" *Nat Genet*, 1998 Oct; 20(2):207-11.

[8] "Fluorescence Imaging, principles and methods" Amersham biosciences, 63-0035-28 Rev. AB, Oct 2002.

[9] Wang Y, Wang X, Guo S.W., Ghosh S. "Conditions to ensure competitive hybridization in two-color microarray: a theoretical and experimental analysis" *Biotechniques*, 2002 Jun; 32(6):1342-6.

[10] Yang Y.H., Buckley M.J., Dudoit S., and Speed T.P. "Comparison of methods for image analysis on cDNA microarray data" *Journal of Computational and Graphical Statistics,* 2002; 11:108-136.

[11] Chen Y, Dougherty E.R., and Bittner M.L. "Ratio-based decisions and the quantitative analysis of cDNA microarray images" *Journal of Biomedical Optics*, 1997 Oct; 2(4):364–374.

[12] Cui X, Churchill G.A. "Statistical tests for differential expression in cDNA microarray experiments" *Genome Biol.*, 2003; 4(4):210.

[13] Jong K., Marchiori E., van der Vaart A., Ylstra B., Meijer G., Weiss M. "Chromosomal breakpoint detection in human Cancer" *In Applications of Evolutionary Computing*. EvoBIO:Evolutionary Computation and Bioinformatics, 2003; 54–65.

[14] Jong K, Marchiori E, Meijer G, Van Der Vaart A, Ylstra B. "Breakpoint identification and smoothing of array comparative genomic hybridization data" *Bioinformatics,* 2004 Jun; [available online at www.pubmed.com]

[15] Autio R, Hautaniemi S, Kauraniemi P, Yli-Harja O, Astola J, Wolf M, Kallioniemi A. "CGH-Plotter: MATLAB toolbox for CGH-data analysis" *Bioinformatics*, 2003 Sep; 19(13):1714-1715.

[16] Fridlyand J., Snijders A., Pinkel D., Albertson D., Ajay Jain "Statistical issues in the analysis of the array CGH data" *Proceedings of the 2nd IEEE Computer Society Bioinformatics Conference (CSB 2003)*, 2003 Aug; 407-408.

[17] Ishkanian A.S., Malloff C.A., Watson S.K., DeLeeuw R.J., Chi B., Coe B.P., Snijders A., Albertson D.G., Pinkel D., Marra M.A., Ling V., MacAulay C., Lam W.L. "A tiling resolution DNA microarray with complete coverage of the human genome" *Nat Genet.,* 2004 Mar; 36(3):299-303.

[18] Available online at: http://www.api.com/lifescience/arrayworx-technology.html, on Nov 18th, 2003.

[19] Chi B, DeLeeuw R.J., Coe B.P., MacAulay C., Lam W.L. "SeeGH--a software tool for visualization of whole genome array comparative genomic hybridization data" *BMC Bioinformatics,* 2004 Feb; 5(1):13.

[20] Wang X, Hessner M.J., Wu Y., Pati N., Ghosh S. "Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction" *Bioinformatics*, 2003 Jul;19(11):1341-7.

[21] Brown C.S., Goodwin P.C., Sorger P.K. "Image metrics in the statistical analysis of DNA microarray data" *Proc Natl Acad Sci*, 2001 Jul 31; 98(16):8944-9.

[22] Tseng G.C., Oh M.K., Rohlin L., Liao J.C., Wong W.H. "Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects" *Nucleic Acids Res*, 2001 Jun 15; 29(12):2549-57.

[23] Hautaniemi S., Edgren H., Vesanen P., Wolf M., Jarvinen A.K., Yli-Harja O., Astola J., Kallioniemi O., Monni O. "A novel strategy for microarray quality control using Bayesian networks" *Bioinformatics*, 2003 Nov 1;19(16):2031-8.

[24] Salla Ruosaari S., Hollm´en J. "Image Analysis for Detecting Faulty Spots from Microarray Images" *Proceedings of the 5th International Conference on Discovery Science,* 2002; volume 2534 of Lecture Notes in Computer Science, pages 259–266. Springer-Verlag.

[25] Rocke D.M., Durbin B. "A model for measurement error for gene expression arrays" *J Comput Biol*, 2001; 8(6):557-69.

[26] Dozmorov I., Knowlton N., Tang Y., Centola M. "Statistical monitoring of weak spots for improvement of normalization and ratio estimates in microarrays" *BMC Bioinformatics*, 2004 May 05; 5(1):53.

[27] Kooperberg C, Fazzio TG, Delrow JJ, Tsukiyama T., Improved background correction for spotted DNA microarrays., J Comput Biol. 2002;9(1):55-66.

[28] Adams R., Bischof L. "Seeded Region Growing" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994 Jun; 16(6):641-647.

[29] Nadon R., Shoemaker J. "Statistical issues with microarrays: processing and analysis" *Trends Genet*, 2002 May; 18(5):265-71.

[30] Yang Y.H., Dudoit S., Luu P., Lin D.M., Peng V., Ngai J., Speed T.P. "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation" *Nucleic Acids Research*, 2002; 30(4):e15.

[31] Rosenzweig B.A., Pine P.S., Domon O.E., Morris S.M., Chen J.J., Sistare F.D. "Dye-Bias Correction in Dual-Labeled cDNA Microarray Gene Expression Measurements" *Environ Health Perspect.*, 2004 Mar; 112(4):480-7.

[32] Dombkowski A.A., Thibodeau B.J., Starcevic S.L., Novak R.F. "Gene-specific dye bias in microarray reference designs", *FEBS Lett.*, 2004 Feb 27; 560(1-3):120-4.

[33] Smyth G.K., Speed T. "Normalization of cDNA microarray data", *Methods*, 2003 Dec; 31(4):265-73.

[34] Bengtsson H., Hossjer O. "Methodological study of affine transformations of gene expression data with proposed normalization method" *Preprints in Mathematical Sciences,* 2003:38, Mathematical Statistics, Lund University, 2003.

[35] Cui X., Kerr M.K., Churchill G.A. "Transformations for cDNA Microarray Data", *Statistical Applications in Genetics and Molecular Biology,* 2003; 2(1):Article 4.

[36] Kerr M.K., Martin M., Churchill G.A. "Analysis of variance for gene expression microarray data." *J Comput Biol.,* 2000; 7(6):819-37.

[37] Newton M., Kendziorski C., Richmond C., Blattner F. "On differential variablility of expression ratios: improving statistical inference about gene expression changes from microarray data" *J. Comput. Biol.*, 2001; 8:37–52

[38] Yang Y.H., Dudoit S., Luu P., Speed T. "normalization for cdna microarray data" *Proceedings of the SPIE*, 4266, 141, 2001.

[39] Quackenbush, "Computational analysis of microarray data", *Nat Rev Genet.*, 2001 Jun; 2(6):418-27.

[40] Rocke, D.M., and Lorenzato, S. "A two-component model for measurement error in analytical chemistry" *Technometrics*, 1995; 37(2):176–184.

[41] Fang Y., Brass A., Hoyle D.C., Hayes A., Bashein A., Oliver S.G., Waddington D., Rattray M. "A model-based analysis of microarray experimental error and normalization" *Nucleic Acids Res.*, 2003 Aug 15; 31(16):e96.

[42] Workman C., Jensen L.J., Jarmer H., Berka R., Gautier L., Nielser H.B., Saxild H.H., Nielsen C., Brunak S., Knudsen S. "A new non-linear normalization method for reducing variability in DNA microarray experiments" *Genome Biol.*, 2002 Aug 30; 3(9):research0048.

[43] Wilson D.L., Buckley M.J., Helliwell C.A., Wilson I.W. "New normalization methods for cDNA microarray data" *Bioinformatics.*, 2003 Jul 22; 19(11):1325-32.

[44] Colantuoni C., Henry G., Zeger S., Pevsner J. "Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts" *Biotechniques*, 2002 Jun;32(6):1316-20.

[45] Park T., Yi S.G., Kang S.H., Lee S., Lee Y.S., Simon R. "Evaluation of normalization methods for microarray data" *BMC Bioinformatics*, 2003 Sep 02;4(1):33.

[46] Tran P.H., Peiffer D.A., Shin Y., Meek L.M., Brody J.P., Cho K.W. "Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals" *Nucleic Acids Res.*, 2002 Jun 15;30(12):e54.

[47] Martinez M.J., Aragon A.D., Rodriguez A.L., Weber J.M., Timlin J.A., Sinclair M.B., Haaland D.M., Werner-Washburne M. "Identification and removal of contaminating fluorescence from commercial and in-house printed DNA microarrays" *Nucleic Acids Res.*, 2003 Feb 15; 31(4):e18.

[48] Jain A.N., Tokuyasu T.A., Snijders A.M., Segraves R., Albertson D.G., Pinkel D. "Fully automatic quantification of microarray image data" *Genome Res.*, 2002 Feb;12(2):325-32.

[49] Documentation of Matlab curve fitting toolbox, available online at http://www.mathworks.com/access/helpdesk/help/toolbox/curvefit on Dec 8th, 2004.

[50] Available online at: http://www2.chass.ncsu.edu/garson/pa765/statnote.htm, on Nov 18th, 2003

[51] Available online at: http://www.statsoftinc.com/textbook/stdiscan.html, on Nov 18th, 2003

[52] Doudkine A, Macaulay C, Poulin N, Palcic B. "Nuclear texture measurements in image cytometry." Pathologica, 1995 Jun; 87(3):286-99.

[53] Guillaud M, Cox D, Adler-Storthz K, Malpica A, Staerkel G, Matisic J, Van Niekerk D, Poulin N, Follen M, MacAulay C., "Exploratory analysis of quantitative histopathology of cervical intraepithelial neoplasia: objectivity, reproducibility, malignancy-associated changes, and human papillomavirus." Cytometry. 2004 Jul; 60A (1):81-9.

[54] Pellis L, Franssen-van Hal NL, Burema J, Keijer J. "The intraclass correlation coefficient applied for evaluation of data correction, labeling methods, and rectal biopsy sampling in DNA microarray experiments." Physiol Genomics. 2003 Dec 16; 16(1):99-106.

[55] Garnis C, Coe BP, Lam S, MacAulay C, Lam WL, "Improved array resolution reduces tissue microdissection requirement for array CGH." Submitted to Genomics, 2004.

[56] Kerr MK, Afshari CA, Bennett L, Bushel B, Martinez J,Walker NJ, Churchill GA. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica, 2002,* 12:203–217.

[57] Cleveland, WS, "Robust Locally Weighted Regression and Smoothing Scatter plots." Journal of the American Statistical Association, Vol. 74, pp. 829-836, 1979.