ABSTRACT

STATISTICAL IMAGE ANALYSIS

OF SPOTTED ARRAYS

by

Filippo Posta

There is a lot of systematic and specific variability in microarray experiments, this variability affects measured gene expression levels, leading to unreliable gene profiling or an heavy load of extra experiment to statistically confirm the data observed in one experiment.

The aim of this work is to systematically analyze, using statistics, the image derived from a cDNA microarray experiment to have a better understanding of this variability and thus a better confidence over the data obtained from an experiment.

Using technologies available at the Center for Applied Genomics, Newark, New Jersey. Selected images derived from different type of microarray experiments have been analyzed in statistical fashion to find answers about the variability of biological data. Statistical methods such regression have been applied to the whole image, print-tips and single spots; leading to answers, confirmations and new ideas about issues regarding analysis and reliability of microarrays experiments.

STATISTICAL IMAGE ANALYSIS

OF SPOTTED ARRAYS

by

Filippo Posta

A Thesis

Submitted to the Faculty of

New Jersey Institute of Technology

In Partial Fullfillment of the Requirements for the Degree of

Master of Science in Computational Biology

Federated Biological Sciences Department

August 2002

APPROVAL PAGE

STATISTICAL IMAGE ANALYSIS

OF SPOTTED ARRAYS

by

Filippo Posta

Dr. Michael Recce, Thesis Advisor — Date
Director of Center for Computational Biology,
NJIT

Dr. Peter Tolias, Committee Member — Date
Graduate Biology Faculty, Rutgers, Newark, NJ

Dr. Ron Hart, Committee Member — Date
Graduate Biology Faculty, Rutgers, Newark, NJ

BIOGRAPHICAL SKETCH

Author: Filippo Posta

Degree: Master of Science

Date: August 2002

Undergraduate and Graduate Education:

- Master of Science in Computational Biology
  New Jersey Institute of Technology, Newark, NJ, 2002

- B.S./M.S. in Mathematics
  Universita degli Studi di Siena, Siena, Italy, 2000

Major: Computational Biology

This thesis is dedicated to

my family and Philana Diem.

Thanks for your support.

ACKNOWLEDGMENT

The author wishes to express his sincere gratitude to his advisor, Professor Michael Recce, for his guidance, friendship, and moral support throughout this research.

Special thanks to Dr.Ron Hart and Dr.Peter Tolias, for serving as members of the committee.

The author appreciates the timely help and suggestion from Donna Wilson, Mark Albano and all the members of the Center from Applied Genomics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

There is a lot of systematic and specific variability in microarray* experiments. This variability affects measured gene expression levels, leading to unreliable gene profiling or a heavy load of replicates to statistically confirm the data observed in one experiment.

The main objective of this thesis is to analyze the variability of the data obtained from a microarray experiment.

The use of microarray technology for gene expression profiling has become widespread. The two most popular platforms are manufactured short-oligo arrays (Affymetrix) and spotted arrays. The increasing popularity of spotted arrays is primarily due to affordability and flexibility. While it is easy to obtain updated probes of sequenced genomes and create experiment-specific slides, pre-spotted arrays and labeling kits are readily available. Unfortunately, cDNA microarrays are not as reliable as their more expensive alternatives. This is mainly due to the high degree of variability that is generated during the preparation and the execution of a spotted-array experiment (i.e. hybridization conditions, sample concentration, dye quality and chemistry, laser alignment etc.).

Such variability has forced researchers to replicate experiments to establish statistical significance for gene expression. Unfortunately, comparing replicates of the same experiment is not straightforward and has raised issues, like normalization across slides.

---

* In this Thesis, the terms microarray and spotted-array are used interchangeably.

This thesis tries to assess the variability of spotted arrays by statistically analyzing the scanned image obtained at the end of a microarray experiment. The analysis is done systematically, pixel by pixel, unlike accepted procedures in which only a few specific parameters are extracted from the image. The median intensity values among the different pixels identifying a spot are commonly used for the final analysis and interpreted by the experimenter. While there is a lot of information that might be useful, much of data is overlooked.

Another objective of the thesis is to utilize this information for data validation. Statistical methods are used to give a measure of the quality of the parameters that are normally used in gene profiling. In addition, the use of an original parameter is proposed for data analysis. This new value is directly inferred by the pixel distribution of every single spot, providing a measure of both fold change and confidence.

Overall, the thesis is divided in four chapters and two appendices.

Chapter 1 describes the process of creating a microarray experiment to help the reader understand the source of variability that will be studied in the following chapters.

Chapter 2 presents a global analysis of the image to identify biases that can be imputed to the fabrication of a microarray: laser alignment, saturated pixels, and gradient problem.

Chapter 3 presents some applications of statistical analysis to find new ways to asses for spot quality, to correctly select a spot, and few words are spent over the issue of normalization.

Chapter 4 is dedicated to fold change analysis, a brief review of standard fold change analysis is proposed followed by some statistical consideration that will lead to

the introduction of a new parameter that can be used as an alternative to the generally used median ratio.

Appendix A describes materials and methods used for the thesis, but only in terms of the type of machines and software used for this work.

Appendix B describes the R code that has been used for the data analysis.

CHAPTER 2

SPOTTED ARRAYS

A microarray is a biochemical technology based on the principle of hybridization between complementary strands of nucleic acids, that allows gene expression to be assessed on a genomic scale, giving researchers the opportunity to assess in parallel the expression of hundreds of genes in a single experiment. The end product of a microarray experiment (Figure 2.1) can be seen as a matrix of microscopic spots, with each feature representing a gene. The intensity of the feature is a relative measure of gene expression.

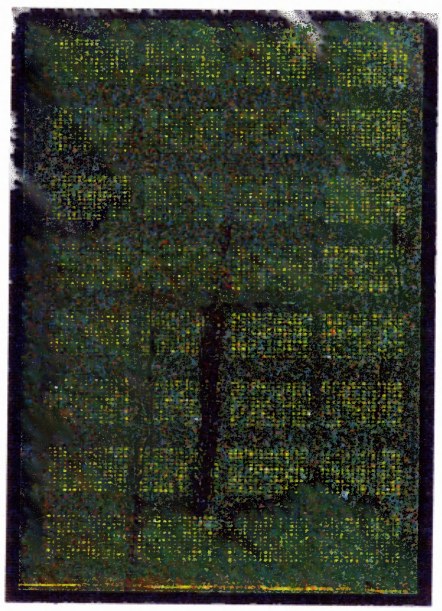

Figure 2.1 Image obtained from a microarray experiment.

To reach the final result there are a series of required steps prior to analysis. It is during the execution of each step that error can be introduced and the reliability of the data can be altered. A schema [1] of these steps is represented in Figure 2.2.

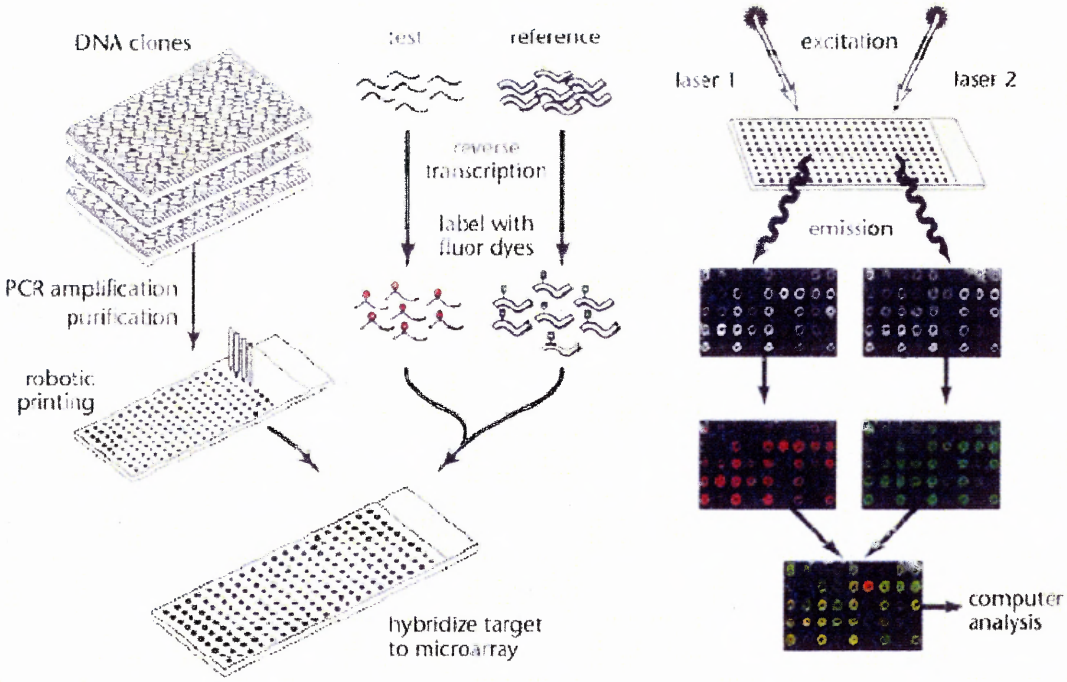

Figure 2.2 Schema for a typical Microarray Experiment.

The schema can be divided into three main steps:

1. Array Fabrication
2. Sample Preparation and Hybridization
3. Data Extraction

2.1 Array Fabrication

Microarrays are constructed by arraying purified PCR products or oligonucleotides at high density, on coated[†] glass slides. A growing number of companies offer synthesized oligo sets representing entire genomes. Experiment-specific oligos are also available for custom arrays. PCR products are typically generated from cDNA clones using universal primers, or specific primers. In addition to known gene products, expressed sequence

[†] Coating is necessary to enhance hydrophobicity and adherence of target DNA.

tags (ESTs) are spotted for gene discovery and gene mapping in many organisms. The target DNA must be purified to eliminate elements (like salts, detergents, primers etc.) that may compete during hybridization, or add to background fluorescence. A sufficient amount of the purified DNA is then suspended in a buffer compatible with the slide surface it is applied to.

The arrayer, is the robotic printing device that deposits a known amount of target DNA onto the slide surface. While the architecture for arrayers may vary, the result is an array of microscopic gene products in well-defined spots across the surface of the coated slide. A popular printing method involves "quill" pens. Fluid is drawn into the spotting pen via capillary action, and surface tension interactions to dispense solution into the slide.

The arrayer proceeds in a systematic fashion, with each pen printing a well-defined area of the slide. Each section of the entire array printed is the collective work of one "print-tip," or the equivalent of one "block." For example in Figure 2.1, there are 32 blocks resulting from 32 different print-tip operations.

After the target DNA has been spotted, the slide is post-processed with a compatible blocking chemistry, and ready for hybridization with fluorescently labeled cDNA.

2.2 Sample Preparation and Hybridization

In microarray experiments, differential comparisons begin with two different samples of mRNA, representing two distinct conditions assayed one against the other. Each RNA

sample is reverse transcribed (RT) to incorporate fluorescent molecules that are later visualized with a laser scanner. These representations of cellular mRNAs are obtained in the RT reaction using an oligo-dT primer, or random primers when the transcripts lack poly-adenylated tails. The products are labeled from the 3′ end, allowing for hybridization to any complementary sequence printed on the slide.

Cyanine-3 (Cy-3) and Cyanine-5 (Cy-5) are fluorescent molecules most frequently used, as they are readily incorporated by the RT enzyme. The excitation and emission spectra of the two dyes also allows for discriminating optical filtration, as they don't overlap.

The final step in microarray preparation is hybridization of the labeled cDNA to the immobilized DNA printed on the slide. Hybridization conditions must be stringent enough to minimize cross hybridization and other undesirable effects. Slides are then placed in a temperature-controlled environment (hybridization machine, water bath, etc.) until hybridization reaches equilibrium. Slides are then carefully washed and spun dry, to remove residuals from the hybridization solution that could contribute to noise. At this point, the microarray is scanned and the image acquired for data extraction and analysis.

2.3 Data Extraction

Due to its highly regular arrangement of detector elements (the spots) and crisply delineated signals (the labels), microarrays can be digitally processed for data extraction. A laser scanner is used to perform this operation; it uses two lasers to excite the fluorescent labels in the slide, then the fluorescence for each channel is converted to

electrical signals that will create the digital image. The image will then be processed pixel by pixel and a statistical overview of the data extracted from the image is given. The statistical data obtained is then the value that is used to define the expression of each specific spot for the microarray experiment. In general, the value that is used to identify the state of a gene in one of the two conditions is the median pixel intensity (R_g and G_g for the red and green channel of a gene g).

The median intensity can be considered as an approximate value to assess for gene expression. This approximation is not as reliable as the one obtained with Affy chips. However, using spotted arrays it is possible to compare two different samples in the same hybridization conditions, while using Affy chips two different experiments are needed.

For this reason, the most relevant (and reliable) value that can be extracted from a spotted array experiment is not a measure of the absolute expression of a gene in any of two different conditions (i.e. median intensity values for red and green channels). Instead, a measurement of relative expression (i.e. red versus green) for each spot is used.

To measure the relative level of expression between two conditions, the value that is used is the median ratio of the intensity:

$$medianratio = \frac{R_g}{G_g} \qquad \textbf{(2.1)}$$

This value is also called fold change, since it asses how many times the red (or green) channel is up-regulated (i.e. more "expressed") if compared to the green (or red) channel.

The median ratio obtained from the scanned image is a raw value that is rarely used as it is. Instead, different factors are applied to it to correct for the variation that occurs in any slide experiment. Moreover, after the median ratio has been corrected, it goes under another transformation, it is taken to log space. In fact, by using the logarithm of median ratio[‡] a fold change in any of the two directions will give the same absolute value but different sign (i.e. $\log(2) = 0.301$ and $\log(0.5) = -0.301$). This type of result will be easier to be interpreted. Since it gives a straightforward comparison among fold changes in opposite directions.

In the next chapter is presented a global study of the pixel distribution of a microarray slide and how is it possible to infer biases from it.

[‡] From now on, the word ratio will imply ratio of medians, unless stated otherwise.

CHAPTER 3

OVERALL ANALYSIS

After a slide has been scanned, there are some steps that need to be completed before the spotted array experiment can be used for comparison among other similar data. These steps are named in table 3.1, together with a short description of them and the section of the thesis where they are analyzed.

Table 3.1 Analys Steps Required to Complete a Microarray Experiment

Step	Description	Thesis's Section
Slide Scanning	A scanning device scans the microarray. The scanning is done channel by channel.	Chapter 2
Gridding	Every spot in the scanned image is associated with what it represents (genes,ESTs,....). Then the area of each spot is accurately delimited.	Section 4.1
Flagging	Every spot within a slide is marked with a flag if it contains defects.	Section 4.2
Noise Model	This part of the slide preparation it is not done yet. Ideally it will allow to separate good data (representing a spot) from bad data (not representing the spot) within each spot of the slide.	Under Study
Background Correction	The software that does the analysis also collects the background data for each spot and channel. Background values are then extracted for each spot and then used to correct analogous spot's values. For example, the median intensity of a spot is corrected by subtracting the median background intensity for that same spot (in either channel).	Not covered. It will be after the noise model will be defined.
Normalization	Normalization is applied throughout the slide to correct different biases.	Chapter 3 and Section 4.3
Gene Ranking	Genes are ranked basedon their fold change (within a single slide)	Chapter 5

In this chapter, some of the biases that are introduced during the slide's

fabrication process are analyzed. For some of the biases, a simple look at the image gives the possibility to identify them. However, it is not straightforward how to deal with them during the analysis step. In this thesis three particular biases are analyzed:

- Intensity Gradient
- Laser Alignment
- Saturated Pixels

3.1 Intensity Gradient

The intensity gradient problem relates to the fact that often in microarray slides one of the two channels has a discontinuous higher intensity than the other one. For example, a slide may be greener toward one side while red and green are more balanced toward the other side of the slide. This effect can clearly be seen in Figure 3.1, which pictures the image of a same on same experiment, i.e. an experiment where the same sample is labeled with two different dyes and hybridized. In this type of experiment the image is expected to be all yellow because the red and green dye, label the same sample in the slide and the juxtaposition of green and red results in yellow.

There are different opinions regarding the origins of this problem and consequently the way it should be tackled. The gradient may be due to the material that the slide is made of, the print-tip, different dye properties (like incorporation), or the cover slip used to cover the slide while sitting for hybridization. The last reason is considered to be the cause as noticed by biologists in the CAG lab but a standard procedure has not been implemented to correct the problem during the analysis step.

However, an optimal use of the cover slip appears to prevent the intensity gradient, or at least minimize it.

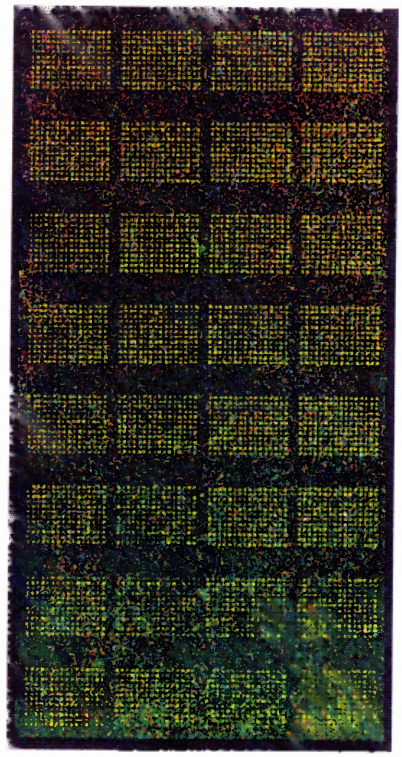

Figure 3.1 Same on same slide with green gradient.

In general, there are two methods that are applied to normalize[§] the data in respect of the gradient [5]. The first assumes that the difference between the values for the green and red channel is constant. This constant factor is evaluated and the spot intensity values are adjusted by multiplying them by the factor for the resulting value[**]. This method is simple but as we can see from Figure 3.1 it doesn't reflect the fact that the gradient is not evenly distributed across the slide. This method is still useful to correct the differences in

[§] Normalization is the term used to describe the process of removing the variation among the different biases that are present inside a microarray experiment.

[**] The are various ways to evaluate this constant; the most used requires creation of a set of housekeeping genes that are used to generate the normalization factor t used to scale the two channels to match each other.

dye incorporation between the two channels, assuming that this bias is constant throughout the slide.

The other method is based on a smoothing function called loess [5]. The loess is a local smoothing function whose main feature is the ability to smooth only data that are close to each other, minimizing the effect of outliers. The normalization method based on loess assumes that the constant use of the same print-tips during the printing phase alters the print-tips themselves, contributing to the gradient. During the printing phase, the robot doesn't work row by row, but instead each print-tip identifies a single block, confirming that the print-tip effect, if any, is not the cause of the gradient since is block related. However, loess gives really good results since the smoothing function acts locally and is able to compensate the gradient effect, but it is not completely clear if the smoothing function alters the biological meaning of the data or not.

As previously stated, it has been discovered that the intensity gradient effect is mainly due to the cover slip. The cover slip is a glass surface that is put over the slide during hybridization. Sometime the cover slip doesn't sit properly (i.e. in parallel with the microarray) over the slide thus creating differences in the availability of hybridization solution to different region of the slide. An example is given in Figure 3.2.

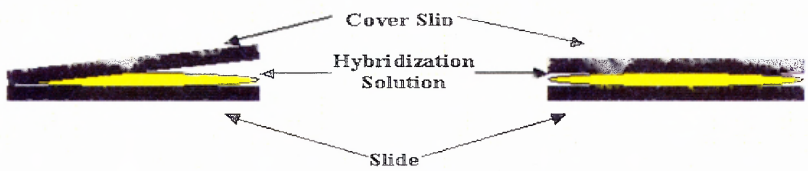

Figure 3.2 Correct (right) and incorrect (left) positioning of cover slip.

From Figure 3.2 is possible to see that the amount of solution available in the slide on the left, decreases while moving from right to left on the slide, and assuming that there is no angle on the other axis of the surface, a gradient from left to right can be expected.

This is the type of result that can be observed by calculating the pixel distribution of every row of blocks within a slide, and then plotting them against each other. Figure 3.3 shows the distributions of every row of blocks for the red and green channel for the slide of Figure 3.2. Each color in the two graphs represents a different row of blocks; the slide is 8 x 4 (blocks). There are eight lines, the top row is colored in yellow, while the bottom one is in green; as it is possible to observe, the lines smoothly move from yellow to green, constantly raising the top of the bell-shaped curve representing the distribution, and reducing the wideness of it. In fact at the top (yellow line) the intensities are higher and there is a broader range of them, while at the bottom line (green line) the peak is at a lower level, meaning that the overall intensities are smaller and the peak on the curve is much tighter, implying that there is less variation among the data.

The reason a gradient is observed in some slides, resides in a decrease of intensity level in the same direction as the angle formed by the cover slip used during the hybridization step, and by the different incorporation properties among the dyes, which is more obvious at lower intensities. How to tackle this problem using the pixel data is not clear yet, since a proper normalization procedure has not been found. However, there are some ideas under study that include the use of a local smoothing function (like loess) to be applied in the direction of the gradient (in this example by row of blocks).

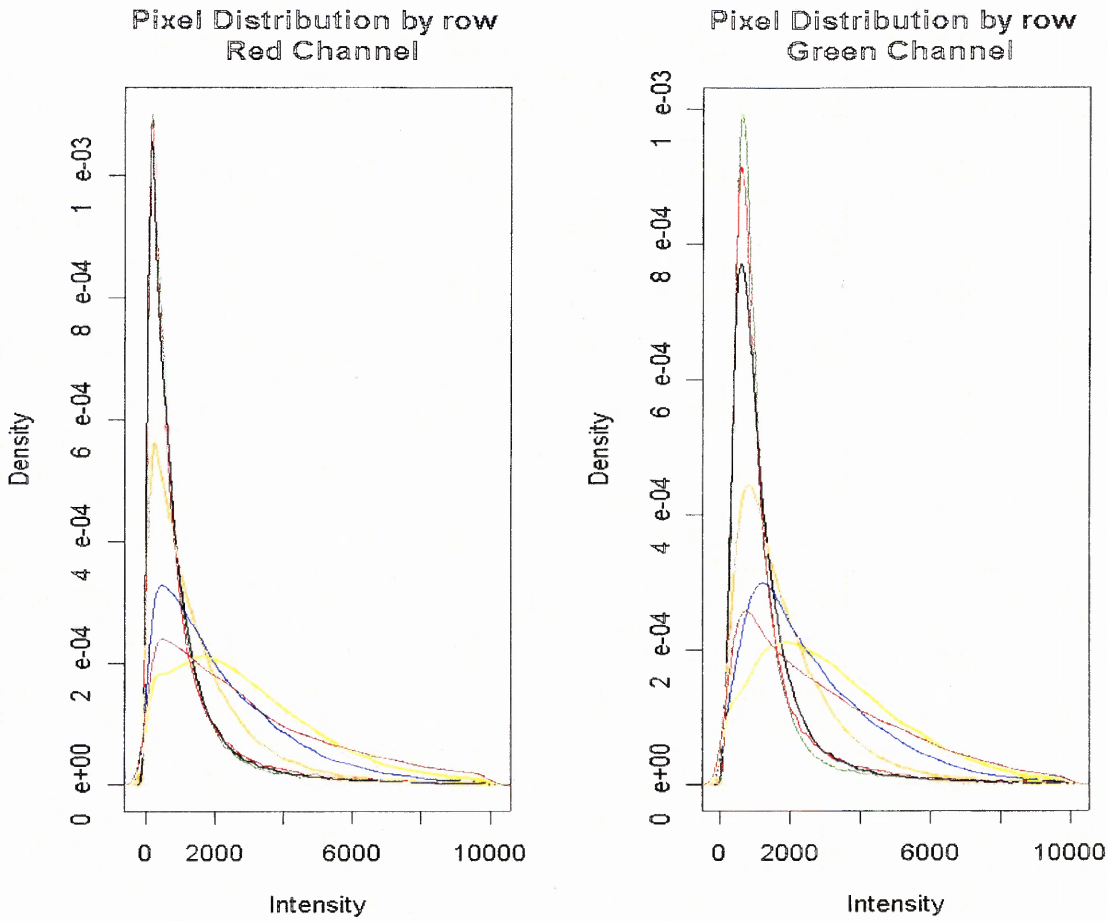

Figure 3.3 Row by row (of blocks) pixel distribution.

3.2 Laser Alignment

To extract data from the arrayed samples, the fluorescent labels are excited using two lasers (if there are two different labels), and the images created by the two lasers are superimposed to create the final image. Sometimes the alignment between the lasers is not correct, and as a result it is possible to observe that on the edges of the spot one color is more present than the other color. The spots in Figure 3.4 are obtained from a scanner robot with the lasers improperly aligned.

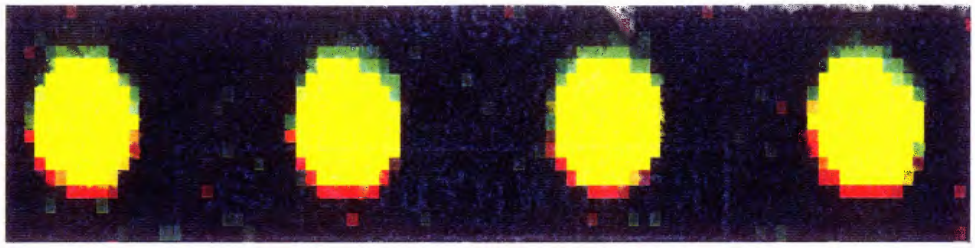

Figure 3.4 Spots resulting from not properly aligned lasers.

It is possible to notice that the bottom of each spot in the figure is red, and the top is green, while the rest of the spot is bright yellow, as it should be, since the image is from a same on same microarray slide. In the case of the picture is clear that there is an alignment problem, but sometimes it is not that clear and if only a broad look at the slide is taken, the error might be completely missed as shown in Figure 3.5. In this picture, a broad view of the same slide (and same spots) is given; a view that does not clarify if the red edges that are observed in some of the spots are just random effects or a global bias.

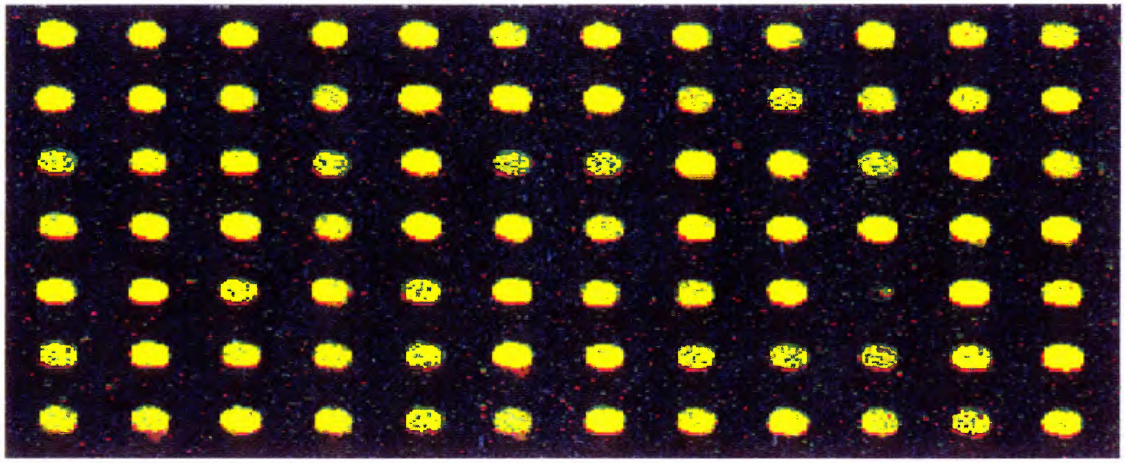

Figure 3.5 Broader view of spots with wrong laser alignment.

To assess for statistical evidence of wrong laser alignment, the row distribution of

every pixel in a given spot has been plotted, leading to the graphs depicted in Figure 3.6.

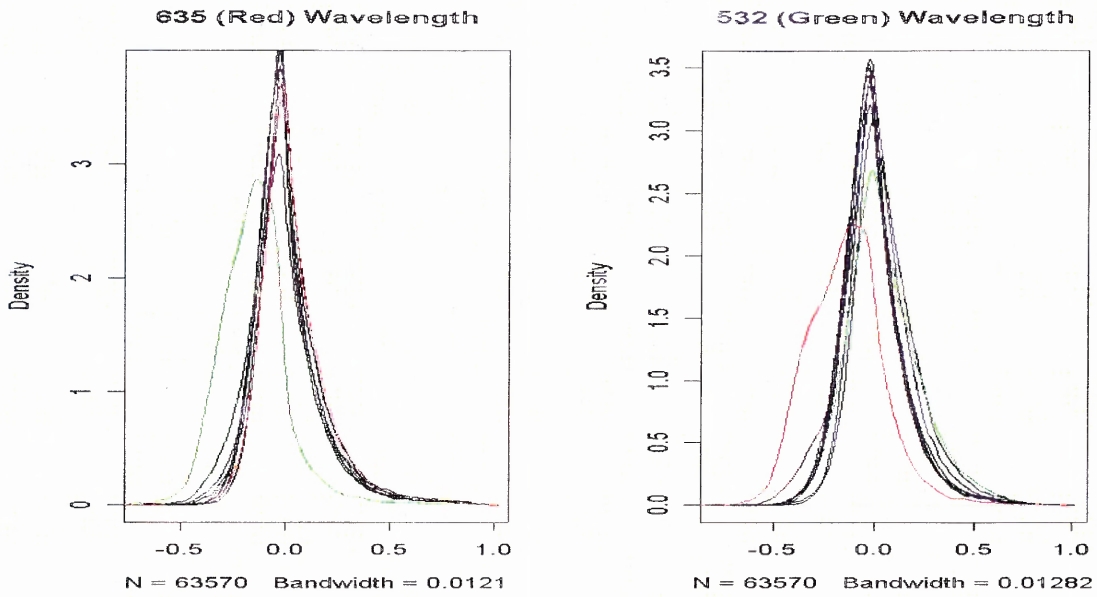

Figure 3.6 Row by row distribution of pixels within a spot.

This figure consists of two graphs, one for each of the two dye intensities. Within the graph, each line represents the pixel distribution of a particular row inside a spot. The line colored in green represents the pixel distribution of the first row (at the top of a spot) for all the spots in the slide, while the line colored in red represents the pixel distribution of the bottom row for every spot on the slide.

In the graph on the left of Figure 3.6, it is possible to observe that the green line representing the top row pixel distribution for the red channel has a different pattern, in particular the mean (i.e. the peak of the curve) is lower and the distribution is shifted toward the left if compared to the other distributions. This means that the pixels in the top row have lower intensities compared to the ones in the rest of the spot, included the bottom row that is highlighted in red and confuses itself with the other lines. The same

result can be observed for the green channel pixel distribution, but the lines are swapped as expected, with the bottom row (red line) shifted more to the left than the others, green line (top row) included.

The same result had been obtained for every microarray slide that had been produced from that same scanner (before a technician solved the problem). Moreover, the same analysis had been done plotting the distribution of the columns of the spot. As a result, all the column distributions look the same in both channels. In fact, the alignment problem was x-axis related, while the lasers were properly aligned along the y-axis.

The laser alignment problem introduces a bias that extends itself to the analysis of the data. The standard value that is taken to represent the state of a spot is the median pixel intensity, and as we will see later in the thesis (Figure 5.2), the median's position density distribution is approximately uniform across the whole spot area, meaning that the median pixel's intensity has equal probability of being in any position within a spot. In particular, if the median lies in the top row or in the bottom one of a single spot, the resulting value cannot be considered as a good evaluation of the median intensity for that particular spot. For this reason, it is recommended to assay for laser alignment by checking the pixel distribution by row and column and discard from the analysis all the rows (or columns) that behave differently form the overall behavior of the spot.

3.3 Saturated Pixels

Another issue regarding microarray slides concerns the limited ability of scanners to detect high intensities, leading to the phenomenon of saturation.

Under the saturation condition, the value of the intensity for one or more pixels

has the maximum value that the scanning hardware can detect. In this situation, it is impossible to discriminate among saturated pixels since they all appear to have the same value even if that is not the case. Luckily, the phenomenon of saturation is not common since it only involves a limited number of pixels within a few spots. On the other hand, spots are more likely to have more than one saturated pixel leading to a shift toward higher intensities of the pixel distribution and the impossibility of using them for analysis.

Across a single slide, the effect of saturation, though small, is still visible when the distribution of pixels for the slide is plotted as in Figure 3.7, where a bump at the end of the distribution's curve can be clearly seen. This bump is located at the far end of the distribution and specifically at the value 65535 on the intensity axis. This value is the maximum intensity recognized by the robot scanner.

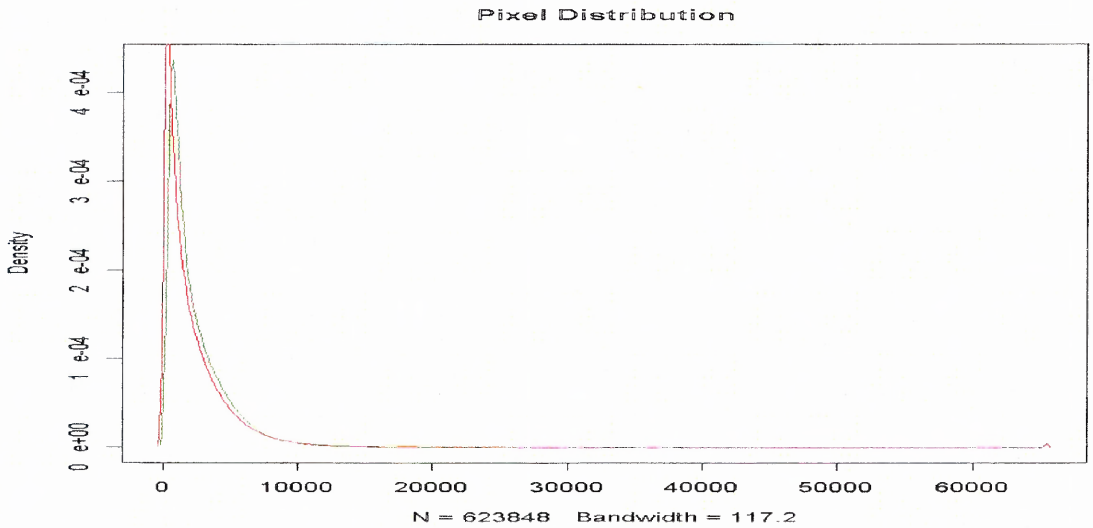

Figure 3.7 Pixel Distribution for a single slide.

The same plot is obtained for every slide containing high-level intensities.

Generally is rare to find a big number of isolated saturated pixels. Saturation can be found as big clusters of pixels in spots representing genes that are usually highly expressed (like actin), as we can see in Figure 3.8, which displays the ribosomal protein L35 from a Hela cell same on same experiment.

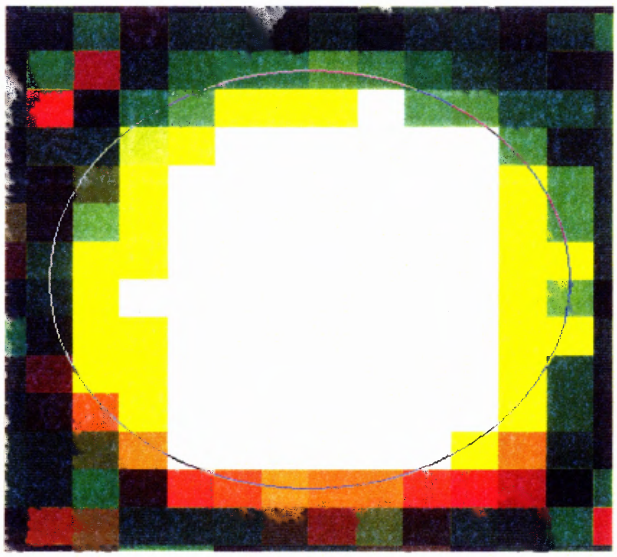

Figure 3.8 Example of a saturated spot.

Unfortunately, because of the limitation of the hardware it is preferred to discard these spots from the analysis since it is not possible to discriminate among them.

CHAPTER 4

GRIDDING, SPOT SELECTION AND NORMALIZATION

In this chapter the process of analysis and tune-up of a single microarray slide is divided into three different steps and revisited using the same point of view that has been used throughout the thesis: pixel analysis. The goal is to provide fast and automated statistical analyses that will substitute the slow and manual analyses that are required to tune-up a slide after the scanning is done. Unfortunately, due to the huge amount of data and the proportional amount of experimentation needed, the presented ideas have not been confirmed experimentally, even though the assumptions upon which they are based seem to hold throughout the few slides that have been used for this thesis.

The Chapter is divided into three sections representing three different issues involved in data analysis of a single microarray slide:

- Gridding
- Spot Quality
- Normalization

4.1 Gridding

After the image has been created from the scanner, a software program is used to associate every spot in the image to the gene it represents and to extract information from the image itself. During the identification process, every spot is automatically delimited (usually by a circle) and every pixel within the delimited region belongs to the gene

associated with the spot (in Figure 3.8 it is possible to see the delimiting circle for that spot). Unfortunately, the software is not perfect, and the experimenter has to check the gridding made by the software, and make corrections for every spot that has not been identified in the proper way. The process of checking that all the spots are properly delimited and identified is called gridding and is extremely time (and eye) consuming. Moreover, it adds some degree of variability since different experimenters tend to grid in different ways, leading to different values for evaluation of the expression of a spot.

Evidence of the variability that can be introduced by gridding spots is given in Figure 4.1, where three different circles are used to grid the same spot. On the left, the spot is gridded in such a way that it maximizes the amount of yellow (the image comes from a same on same slide), the spot in the middle is gridded to maximize the amount of information contained in the spot, while minimizing the background, and the spot on the right is gridded so that only the core of the spot is used and the blurriness near the borders of the spot is discarded.

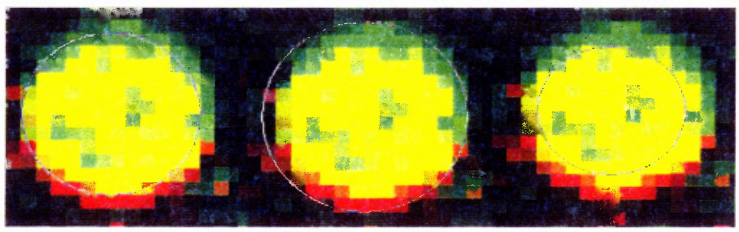

Figure 4.1 Different alternatives to grid the same spot.

All of these gridding options have a logical explanation, but they all lead to a different value of the median (or any other statistical value related to the spot). Usually, the fluctuations introduced through gridding are minimal (and do not alter too much the analysis), but they exist and if the gridding variability is added to the time wasted to grid

each slide, it is easy to figure out how much there is to gain with an efficient and reliable method that will carry out the gridding procedure automatically. To fulfill this purpose, it has been assumed that by using the same circle for each spot, and by making this "universal" circle big enough to contain an approximately equal amount of pixel from the spot and from the background around the spot, then every spot would have had a density distribution of pixels with two main peaks, one relative to the presence of background pixels and the other generated by the spot's pixels. If this assumption is true, then all the pixels that are real data (i.e. belonging to a spot) would be selected just by looking at the distribution of the pixels, and thus an algorithm can be created that will discriminate among background and feature pixels.

The first step to be done in this direction, it is to select an appropriate size to be used for every spot. This size has to be big enough to discriminate between feature and background pixels, but cannot be too big since throughout the whole slide there are sparse bright pixels (as shown in Figure 4.2) that might create a smooth density function, that does not discriminate among different pixels.

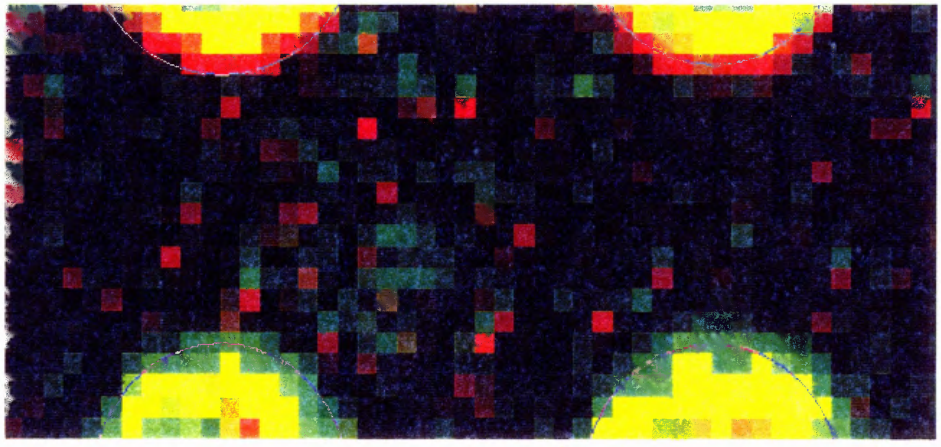

Figure 4.2 In the space between features can be noticed sparse bright spots.

A set of different spot sizes has been selected ranging from 156 pixels to 392[††] and the distribution of the pixels for every size has been plot. The result of this approach is shown in figure 4.3, where only the results for the more relevant sizes are shown.

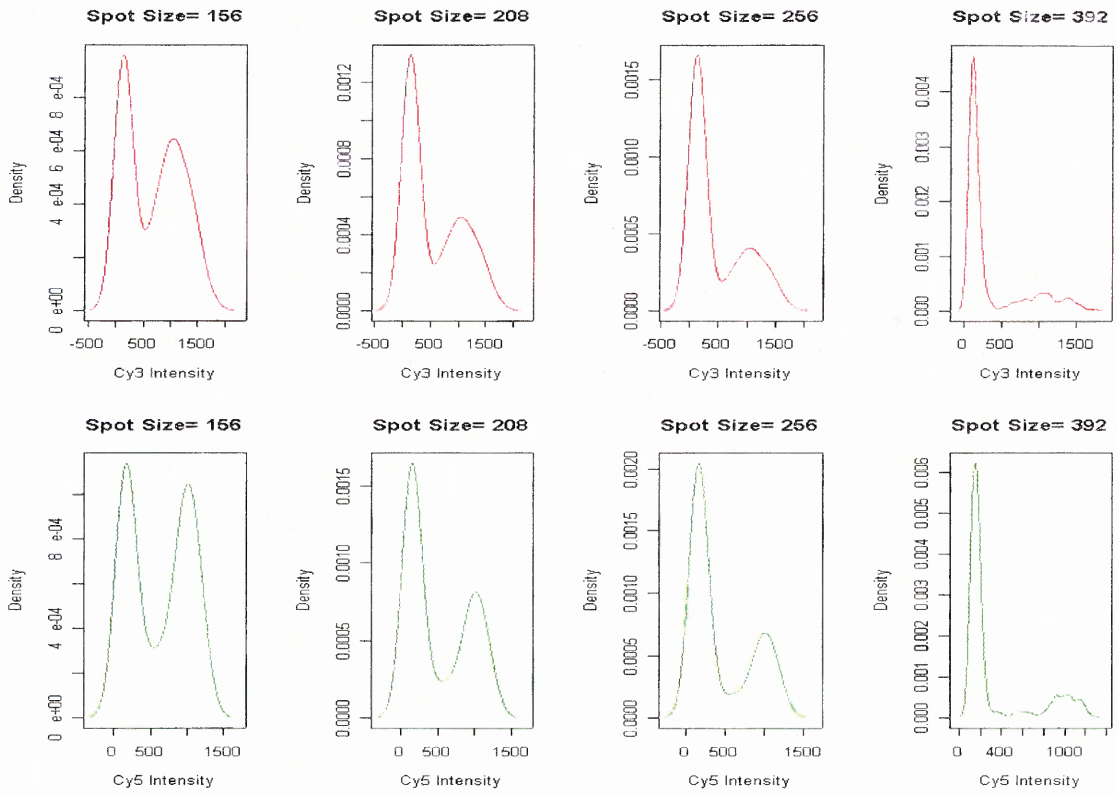

Figure 4.3 Spot pixel distribution for different sizes of spots.

In this picture are shown the distributions of pixel intensities of the same spot across four different pixel sizes (156,208,256,392). The graphs above represent the distribution for the red channel. Below, the green channel is analyzed. From the different graphs it is possible to notice that for higher sizes the effect of noise in the data is too big to identify pixels that belong to a spot, so the size of choice should be one in between 156

[††] 392 is usually the biggest size a circle can have without intersecting neighbor circles.

and 208. To choose between the two of them, a set of low intensity spots had been selected to check if the same discrimination can be observed.

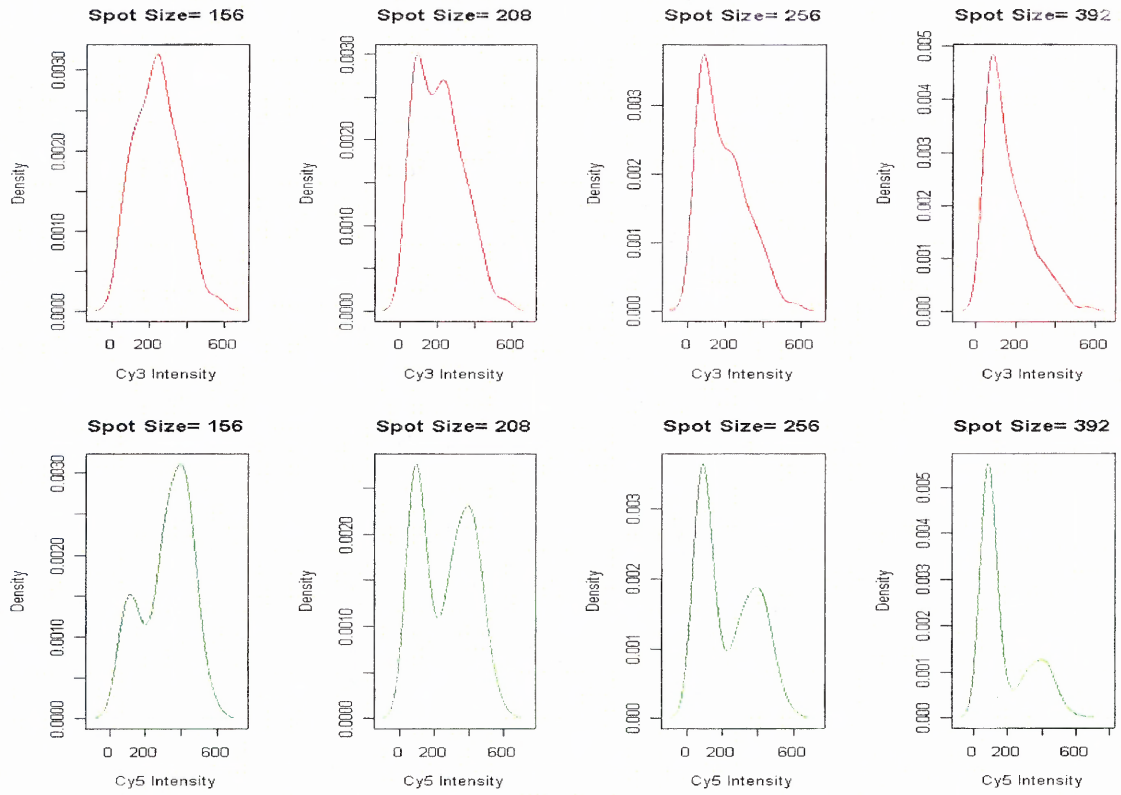

Figure 4.4 Pixel distribtuion for spots with low intensity.

The result of this analysis is represented in Figure 4.4 and shows how for a spot size of 156 pixels. The background and feature pixels distribution for the red channel form a unique line from which it is impossible to discriminate, but if a spot size of 208 pixels is taken then a small discrimination tends to appear. This discrimination disappears whenever the size of the spot becomes higher.

The results obtained at lower intensities are not as neat as for the rest of the spots, but it is encouraging that is still possible to graphically see two peaks in the distribution.

At present, an automated method that implements this idea is still under study.

4.2 Spot Quality

Another major issue regarding microarray slides is the presence of defects that can completely alter a feature, making it unacceptable for the analysis. An example of defected spots is represented in Figure 4.5.

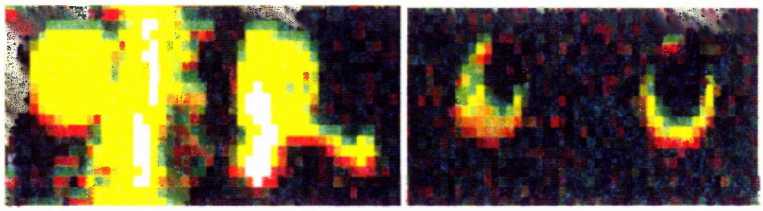

Figure 4.5 Spots presenting defects.

This kind of spot is usually flagged during the gridding step (a procedure that is less time consuming than the gridding itself). The flag helps the software that does the analysis to eliminate bad spots from calculation involving the whole microarray (like the evaluation of a normalization constant to scale the two channels).

The origins of spot's defects are various: dust, scratches over the slide surface, print-tip defects, unwanted deposits created during printing or hybridization, or a change in spot morphology due to post-processing the slide. Depending on the various sources that generate spot's defects, the graphical outcome of those same spots can vary: from extremely bright and not rounded spots (like the ones on the left of Figure 4.5), to spots presenting big clusters of black within themselves (right side of Figure 4.5).

To assess for a spot's quality, the probability distribution of single spots has been analyzed; the idea in this case is to first verify the assumption that the frequency

distribution for the pixels contained in a microarray feature is normal and then to determine the quality of a spot by assessing for departure from normality of its pixel frequency distribution.

To verify the normality hypothesis, graphical and statistical methods have been used. The graphical method consists on plotting the cumulative frequency distribution of randomly selected spots to visually check if the resulting graph would approximately fall on a straight line. Some of the analyzed spots are shown on Figure 4.6. In Figure 4.6 there are two rows of different plots: on the top row, a cumulative frequency distribution for the pixels is plotted, together with the line where the distribution should lie if it is normal. While in the bottom row, there are represented the density distributions of the same spots as the ones in the top row.

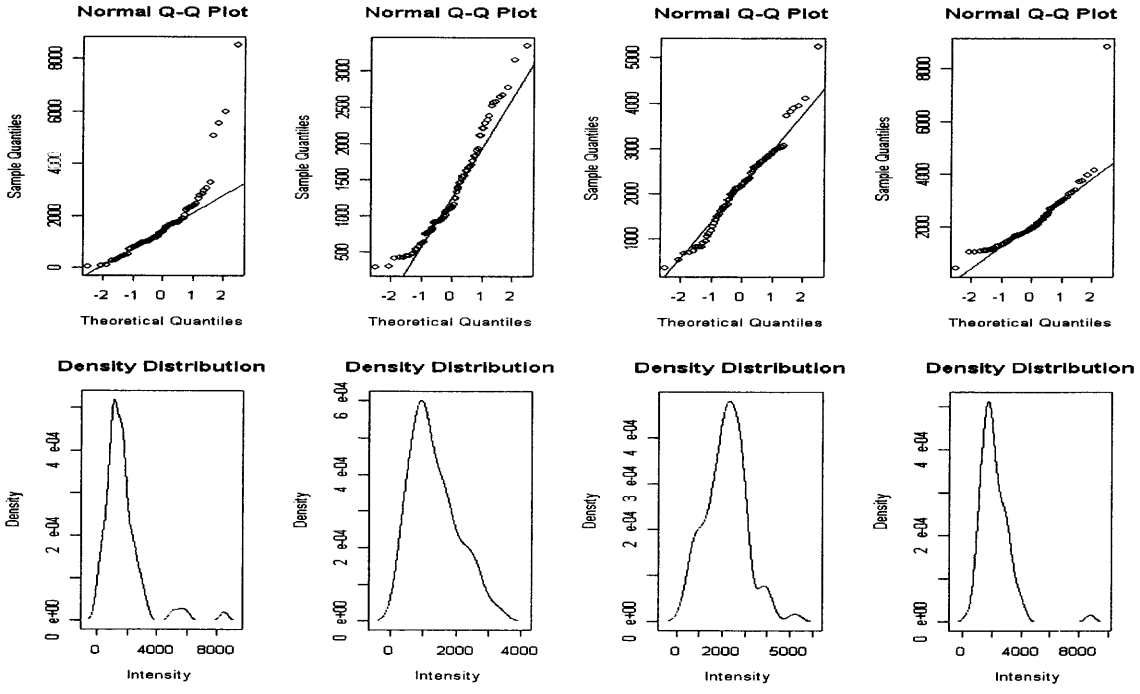

Figure 4.6 Cumulative and Density distribution for random spots.

The normality hypothesis seems to be confirmed by this raw graphical analysis. In fact, it is possible to infer from the figure that the cumulative function approximately lies on a line, and the points that lie far away from it seem to represent noise in the data. This explanation seems to be supported by the density function too, where bumps at the end of the distribution can be related to the points lying far away on the corresponding cumulative distribution plot.

To further confirm the graphical analysis, the Shapiro-Wilk test was applied to a set of different microarray slides. The Shapiro-Wilk test is a calculation of a statistic (called W) that will confirm (or deny) the assumption of normality; the higher the value obtained for W, the higher the probability that the tested data have a normal density distribution. The Shapiro-Wilk test has been chosen because it is very powerful if compared to other available normality tests (like the D'Agostino test). The test's only drawback is that it performs poorly for datasets containing identical data, but it is hardly the case for the dataset that have been used.

The results obtained by applying the Shapiro-Wilk test were extremely good. In all the analyzed slides an extremely high percentage of genes had a W value of .7 or higher thus confirming the normality assumption to be correct. The overall results are shown in Table 4.1 for seven of the used slides. Each value in a cell represents the least Shapiro-Wilk test value that a gene in the corresponding slide (and channel) has to have to be contained in the corresponding percentage, i.e. the cell on the top corner states that in Slide1[‡‡], 90% of the genes in the red channel have a Shapiro-Wilk test value of at least 0.65.

[‡‡] The full name and specification of the slides is not relevant and is omitted.

Table 4.1 Shapiro-Wilk values by slide and percentile

Percentages	Slide 1	Slide 2	Slide 3	Slide 4	Slide 5	Slide 6	Slide 7
Red 90%	0.65	0.77	0.7	0.78	0.77	0.79	0.8
Green 90%	0.75	0.81	0.8	0.8	0.75	0.75	0.78
Red 99%	0.28	0.37	0.44	0.57	0.63	0.63	0.5
Green 99%	0.35	0.63	0.67	0.64	0.51	0.51	0.41

From Table 4.1, it is also possible to notice the large variance from the smallest value having a 99 percentile of genes compared to the 90 percentile. This big difference, suggests that perhaps the spots that are more distant from normality are spots with defects or spots with a high level of background noise.

Unfortunately, this has been shown not to be the case, because it is not possible to determine the quality of a spot by simply analyzing its departure from normality. This result is shown in Figure 4.7, where the cumulative frequency distributions for eight spots from the same slide are plotted. The spots have been arranged in such a way that the first two graphs (moving from left to right) on the top row (of graphs) represent spots with the lowest Shapiro-Wilk test result, and have not been flagged. The last two spots on the first row have the lowest two Shapiro-Wilk test scores and had been flagged by the experimenter. The second row of graphs contains plots of the distribution of the two spots with highest test score and no flag (on the bottom left of Figure 4.7), and the distribution of the two features with highest Shapiro-Wilk test that had been flagged (on the bottom right of Figure 4.7).

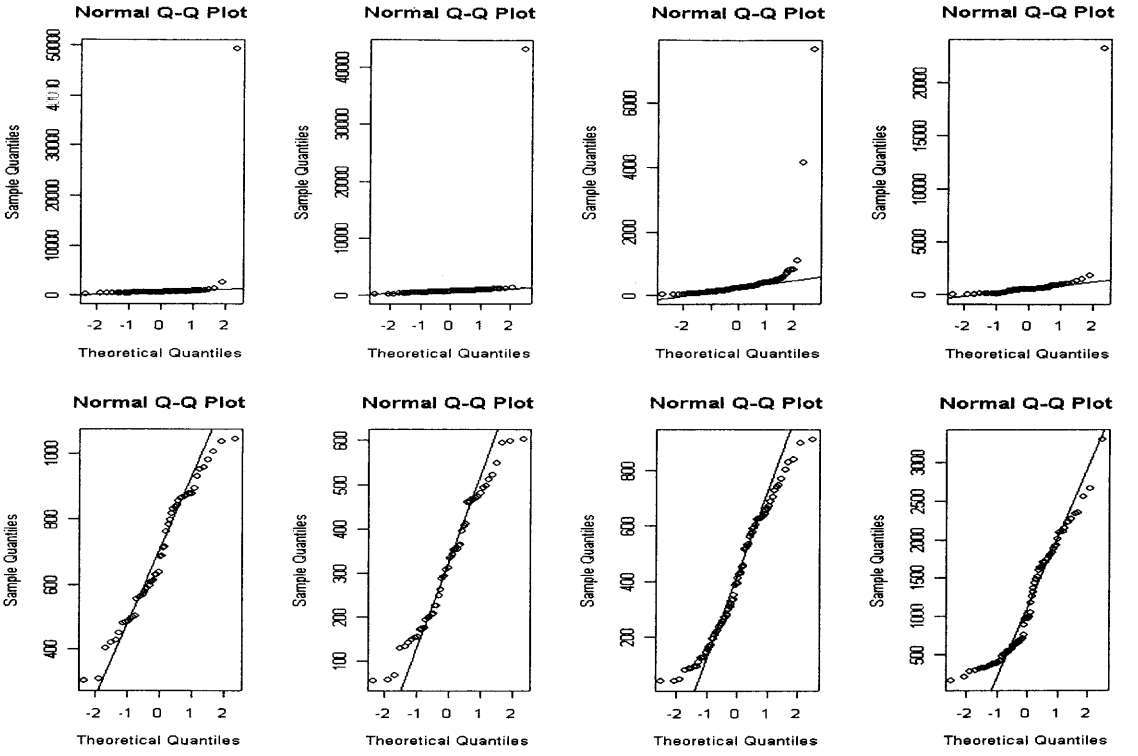

Figure 4.7 Good and Flagged Spots cumulative frequency distributions.

It is impossible to visually discriminate among the flagged and not flagged spots (the same result is obtained using density distributions instead of cumulative).

Another attempt has been made in this direction (departure form normality of spots with defects) by using as discriminating parameter, the kurtosis value of the distribution. The kurtosis of a distribution is a measure of how much the shape of the distribution reflects normality. If the kurtosis value of a distribution is zero, or close to zero, then the distribution has a normal shape. If the value is less than zero, then the distribution is said to be platykurtic meaning that it can be the composite of one or more populations with the same variance and different means. Finally, if the kurtosis of the distribution is greater than zero, then the distribution is said to be mesokurtic, meaning

that the distribution can be the composite of two distributions with same mean but different variances.

The idea in this case is to be able to assess for normality and also have a guess at the reason for departure from normality, but unfortunately the separation of good and bad spots, though better than the one obtained using the Shapiro-Wilk test, it is still not discriminative enough, resulting in a high percentage of false positives (i.e. good spots with kurtosis different from zero) and false negatives (i.e. bad spots with kurtosis close to zero).

Overall, the result is that it is not possible to discriminate between good and bad spots solely by looking at departure from normality. The reasons are mainly two: every spot, good or bad, has a normal-like distribution (at least in one channel), and the presence of noise through the data makes it impossible for the available techniques to clearly evaluate the normality of a spot's pixel distribution (top row of graphs in Figure 4.7).

At the time of writing this thesis, the efforts to solve the problem of spot selection are directed toward the definition of a set of spots that well describe a slide. This set will then be used to determine a "general" spot characteristic of the slide; this spot will have a specific normal distribution, and will be compared to every other spot within the slide. The spots with distributions similar to the one of the "general" spot will be used for analysis and the remaining ones will be discarded. This idea seems very promising, but it is strictly related to the creation of a good set of training data, and the need for the "general" spots for different slides to look the same.

4.3 Normalization

In microarrays studies, normalization is the term used to describe the process of removing the variation among the different biases that are present inside a microarray experiment.

At the present time, there are two main techniques used for normalization (as described in Chapter 3). One evaluates a constant factor that is globally (sometimes locally) applied to the microarray slide, while the other makes use of a local smoothing function called loess. Both methods are not applied straightforwardly and require a lot of thinking to decide whether to apply them locally (print-tip, rows of block, cluster of spots) or globally. The choice of the genes to include in the normalization is a topic for discussion, together with the issue of utilizing values with or without subtraction of the local (or global) background.

This thesis takes an inside look at these biases and then removes them from the source as noise, whenever possible. Thus, the problem of normalization at this stage of the work is not a major issue. In fact, whenever the proposed ideas will be fully implemented, the final result will be a microarray consisting only of good features, and each feature will consist only of those pixels that specifically represent the spot (i.e. the noise will be filtered out) and thus, no normalization will be needed within the spot. However, across the slide some scaling will still be needed to compensate for green or red or print-tip biases.

At this time, it is not realistic to infer a proper method for normalization, but the overall idea is that the techniques that have been applied so far to evaluate scaling factors would work fine, whenever the source of variation is statistically assessed (i.e. laser alignment section Chapter 3).

CHAPTER 5

FOLD CHANGE ANALYSIS

In this chapter the pixel prospective that characterize the thesis, will be used to get a better understanding of what happens at the analysis step of a microarray experiment.

The chapter is divided in two sections: the first one is about the standard fold change analysis that is applied to assay for gene expression, and the reliability of this kind of analysis. The second section introduces a new idea to assess gene expression, an idea that does not change what has been used so far in microarray gene profiling, but extends it in a natural way so that the amount of information obtained from the image is maximized.

5.1 Fold Change Analysis

The median ratio is the traditional value used to assess for gene expression between the two samples of a microarray experiment. The software used to scan the slide automatically obtains the median ratio, which is calculated by taking the ratio between the median of the intensities of a spot in red channel, and the median of the intensities of the same spot, but in the green channel. The use of the median instead of the mean is due to the fact that the median is not as much influenced by outliers as the mean.

In Figure 5.1, is given a graphical explanation of the better approximation obtained by using the median instead of the mean. In this figure the pixel distribution of a spot is plotted, and two points on the graph are highlighted, the one represented with a diamond indicates the location of the median, while the one represented by a filled circle

indicates the location of the mean. Clearly, the median represents the population better than the mean[§§].

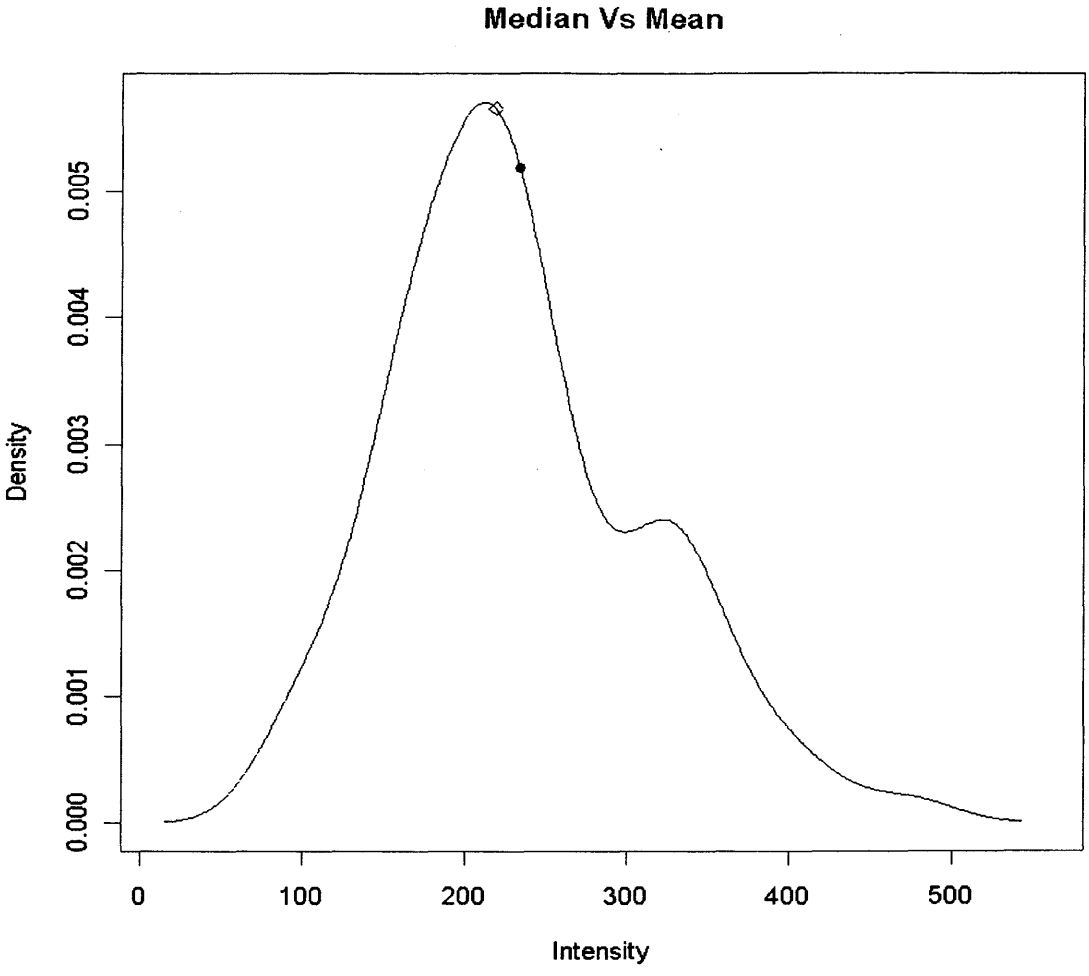

Figure 5.1 Graphical explanation for the choice of median over mean.

After a raw value of median ratio had been evaluated, different factors are applied to it, to correct the biases that had been studied in the previous chapters, like print-tip effect, and different dyes incorporation.

The final tune up of the median ratio is given by moving it to log space, i.e. by using the logarithm of the ratio instead of its plain value. In fact, whenever an

[§§] The same result had been obtained by randomly select genes in different slides and different channels

experimenter wants to profile a gene expression, he (she) does so in terms of fold of change. Unfortunately, if the plain value is used, a three fold change up (red channel three times bigger than green channel) will give a result of 3, while a three fold change down (green channel three times bigger than red) will result in a value of 0.333. This disparity among the values does not help when the genes need to be ranked based on their fold change. In fact, while it is easy to assess for up regulation, the same cannot be said for down regulation. It will be much easier if the median ratio is expressed in such a way that a two-fold change has the same absolute value, but it is positive in case of up regulation and negative otherwise.

This result can be achieved by using the logarithm of median ratio; in fact the logarithm of 3 is 0.5849, while the logarithm of 0.333 equals –0.5849. The only problem that can be encountered with the use of the logarithm is the base of choice, since by changing the base the value is changed too; for this reason, the standard convention is to use a base of two whenever the use of the logarithm is used.

Once the median log ratio has been evaluated for every spot in the slide, the spots are ranked by fold change. The ones with a fold change higher than a certain threshold (usually no less than two-fold change) are then further investigated.

How reliable is this fold change? Is it truly the most representative value for the whole spot? To try to answer these questions, different analyses had been done over the distribution of pixels among a single spot.

In Chapter three, it has been seen that if the lasers are not properly aligned, then the pixels at the edges of the spot should not be used for the analysis. For this reason, it makes sense to check the position of the median for each spot, in different slides and for

both channels, to see if the median's position can be characterized in some way. A median's position analysis is shown in Figure 5.2, where the three graphs represent the median density distribution (by position) for three different microarray experiment, which are different in time, typology, and scanner used to obtain the image. All the six curves (three for the red channel and three for the green one) are almost uniform between 1 and 80 (for uniformity reasons, only spot with pixel size 80 had been used, since more than 90% of the spots have a pixel size of 80), meaning that the median can randomly fall in any position within a spot. The peaks and drops that can be seen on the graphs are not as statistically significant as they look, even though it is still possible to see the lasers misalignment for the graph on the left. Overall, it is possible to assume that the median uniformly appears in any pixel position of a spot.

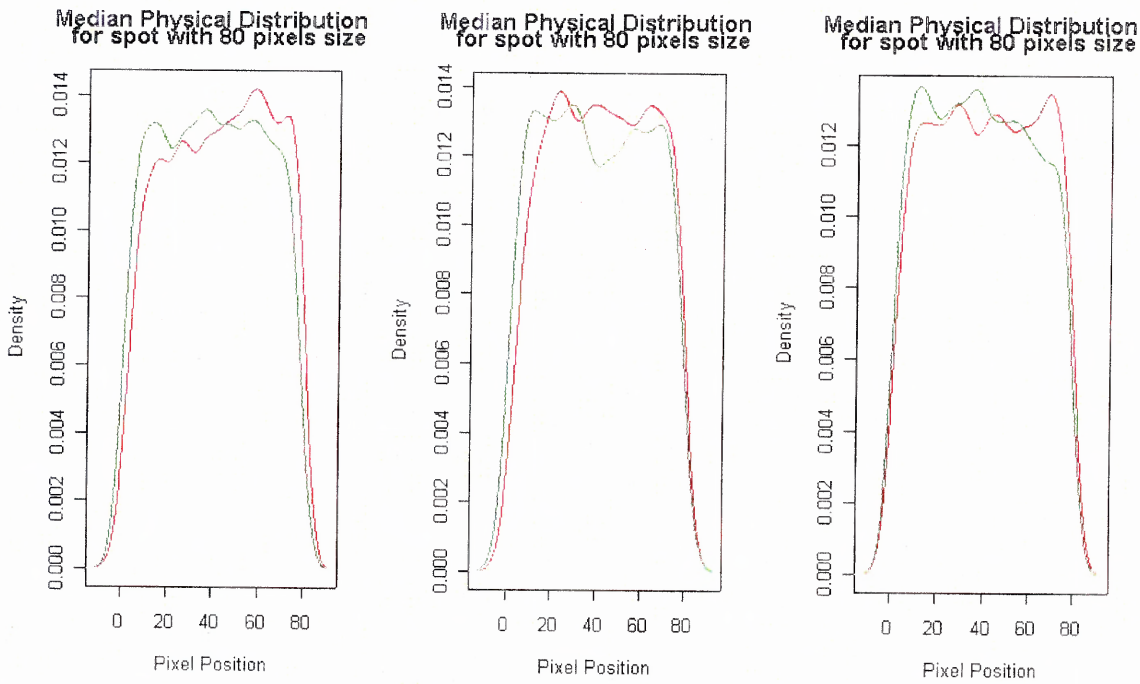

Figure 5.2 Median position distribution.

Another question that can be raised about the median ratio is, how representative of the whole spot distribution it is, i.e. is there enough correlation among the pixels, such like the calculated fold change can be observed among the majority of the pixels?

To answer this question, randomly selected spots from different slides where visually analyzed by plotting the pixel of the green channel versus the pixel of red channel, ordered by position. The same analysis was done for the whole slides by creating spreadsheets containing the correlation among red and green channel for each spot. As a result, no correlation whatsoever was found, and a lot of spots are actually anti-correlated (of course with small absolute values).

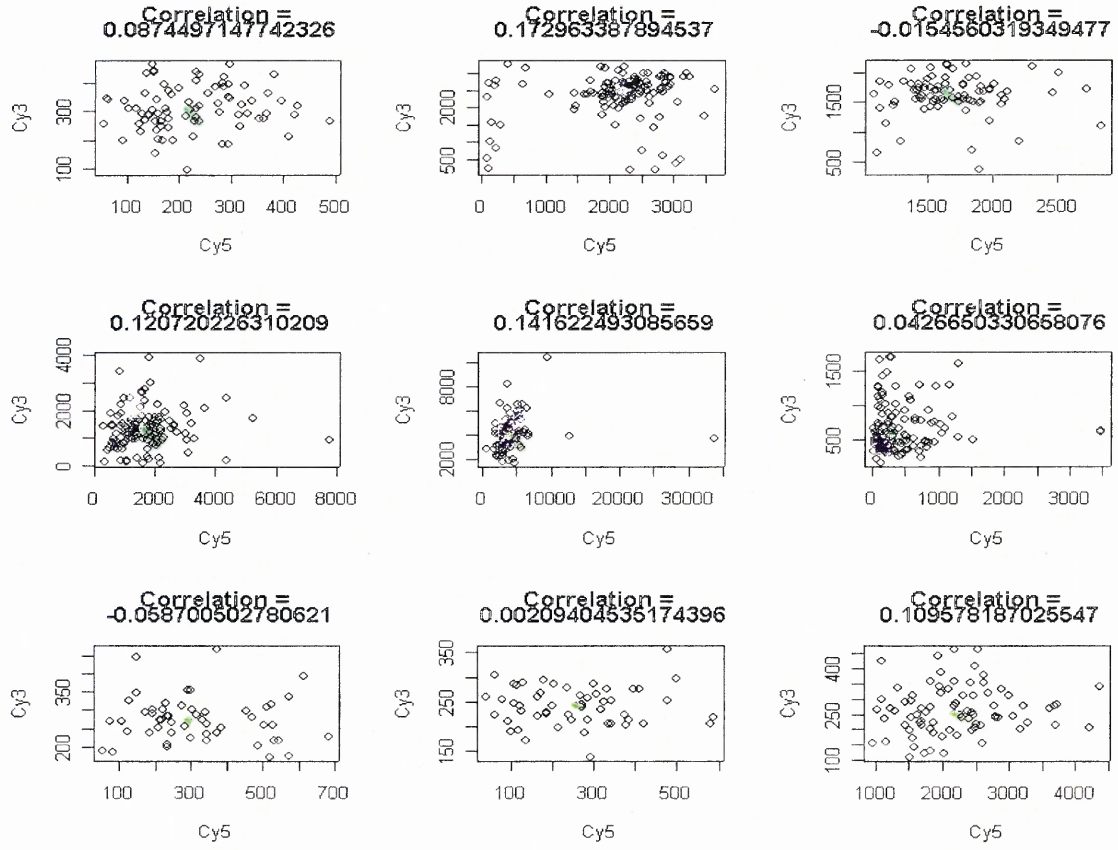

Figure 5.3 Cy5 and Cy3 correlation by position.

Figure 5.3 gives a visual representation of the results. The plotted pixel values spread all over the graph without any specific correlation, the green dot that can be noticed in the graph is the position of the median for the two channels.

The cause of these extremely discouraging results, resides in two implicit assumptions: that the arrayer will print each pixel's area uniformly, and that hybridization will happen proportionally in each channel, pixel area by pixel area. This is not true, mainly because of the extremely small dimensions of a microarray and the composition of the hybridization solution.

However, it is fair to assume high correlation if, instead of ordering the pixel by position within the spot, we rank the pixels by their intensity's values. This idea is implicitly assumed whenever the median is used. In fact, the median of a set of numbers is evaluated by first ranking the numbers in order of magnitude, and then the value that is in the middle position of this ranking is taken as the median value.

The same analysis as the one in Figure 5.3 (and for the same spots) is executed after ordering the spots by pixel intensity. The results (Figure 5.4) validate the use of the median and the previous assumption. High and positive correlation can be observed for every spot, with the median (represented by the green dot in Figure 5.4) that seems to capture the overall behavior of the feature.

The use of the verb seem is not casual. While the great majority of the points approximately lie on a line, this line is not always the main diagonal or a parallel to it as it possible to notice in Figure 5.5.

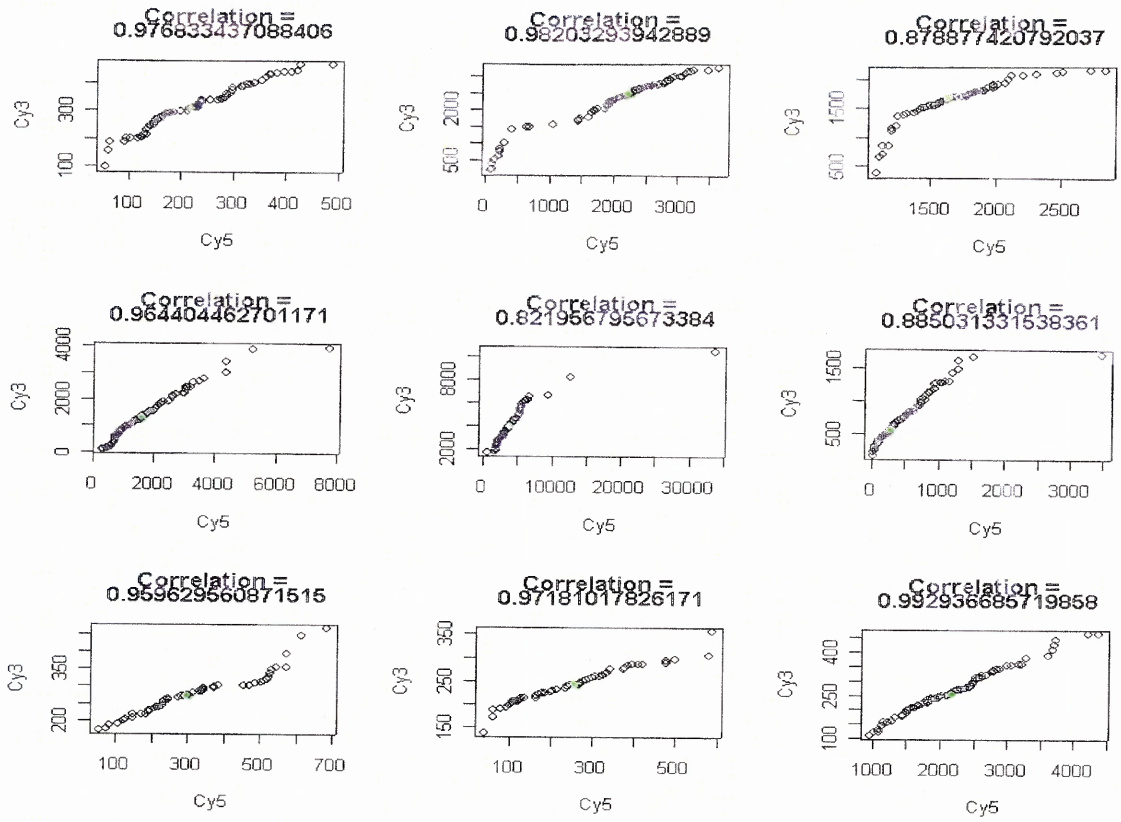

Figure 5.4 Cy5 and Cy3 correlation after being ranked by intensity.

In Figure 5.5 the same data as Figure 5.4 are represented. However, together with the pixels scatter plot, two lines are plotted: the black one is the main diagonal, where all the points should lie if there is no fold change; while the green line*** is the parallel to the main diagonal that intersects the distribution in the median. The reason the correlation line should be the main diagonal, or one of its parallel, it is based on the fact that the median fold change should extend to as many pixels within a spot as possible. In this ideal condition, if there is no fold change the plot will give a green line that matches the main diagonal (black line). If there is fold change, the green line will be a parallel to the

*** This line will sometimes be referred to as median line.

main diagonal which intersects the y-axis at the value given by the difference among the median pixel intensity of the red and the green channel[†††].

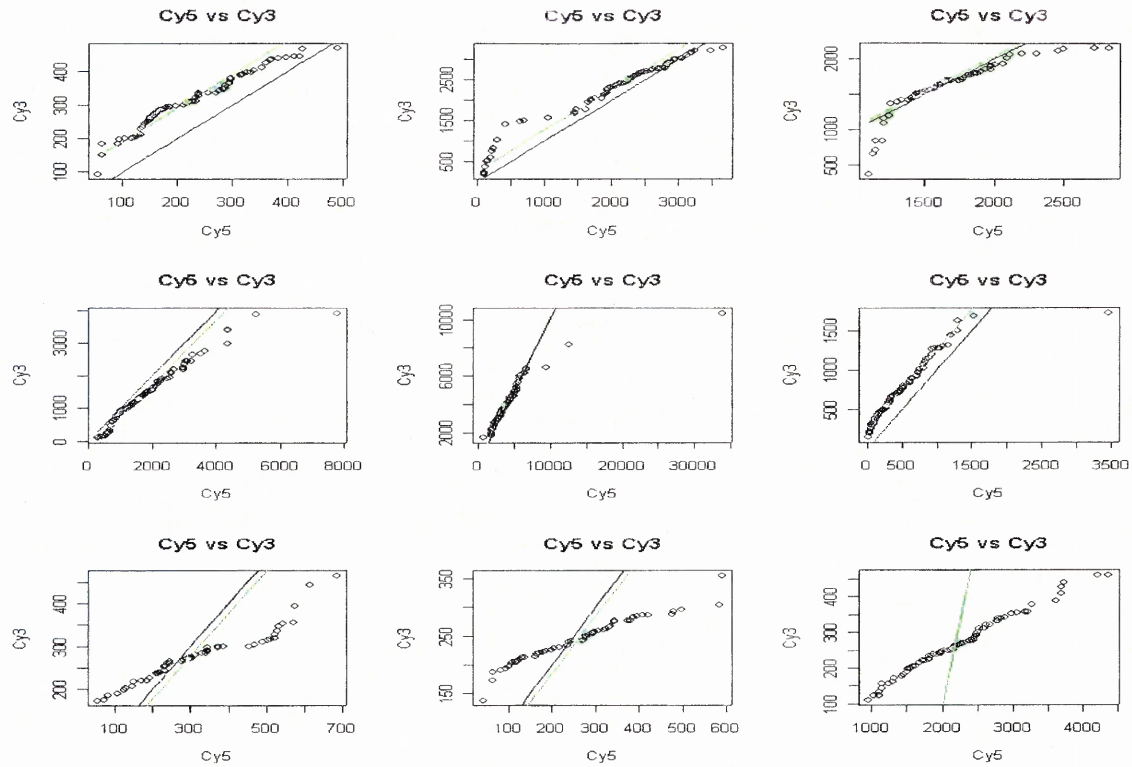

Figure 5.5 Cy5 and Cy3 plotted against main diagonal (black) and a line through the median but parallel to main diagonal (green).

Giving a close look at Figure 5.5, it is possible to note that while the first six spots (from left to right) seem to behave ideally, the pixel distribution for the last three spots is really far away from the main diagonal or a parallel to it. Another observation is the last spot has a big difference among the green and red channel intensities. While the penultimate two have low intensities in both channels.

These results suggest that this type of analysis could be used to check for the quality of a spot and for spot selection.

[†††] The analytical proof of this fact is not relevant for the thesis and then it is not included in the work.

The idea of using a graphical analysis like the one showed before to assess for quality has been implemented in this thesis only to give a confidence measurement of the evaluated median ratio. In this thesis we estimate the confidence on the ratio of median for a spot, by first ordering by intensity the pixels in the green and red channels (as in Figures 5.4 and 5.5). In this case, the line that is parallel to the main diagonal and that passes through the median of both the red and green channels (green line in Figure 5.5) is estimated. Finally the obtained line is fitted to the green and red data. To assess for confidence the sum of the residuals obtained from the fitting process is taken as parameter. The smaller this sum is, the closer the pixels of the red and green channel are to the median line (green line in Figure 5.5).

By ranking the spots by the sum of residuals, it is possible to have a measure of how good the median ratio represents the fold change across a whole spot. The obtained result gives a good discrimination among spots. It is possible to note this discrimination in Figure 5.6, where on the right are plotted spots with low sum of residuals and on the left are plotted spots with high sum or residuals. Clearly, the median ratios evaluated for the spots on the right can be trusted more than the ones on the left.

A drawback for this method is that it works fine only for spots with similar pixel intensities. Spots with high intensities have usually higher residuals than the ones with low intensities. This is an expected obstacle, and to go past it, all the sum of residuals for a single spot had been divided by the median intensity for that same spot. This form of normalization is not sufficient and currently a better way to normalize the data is under study.

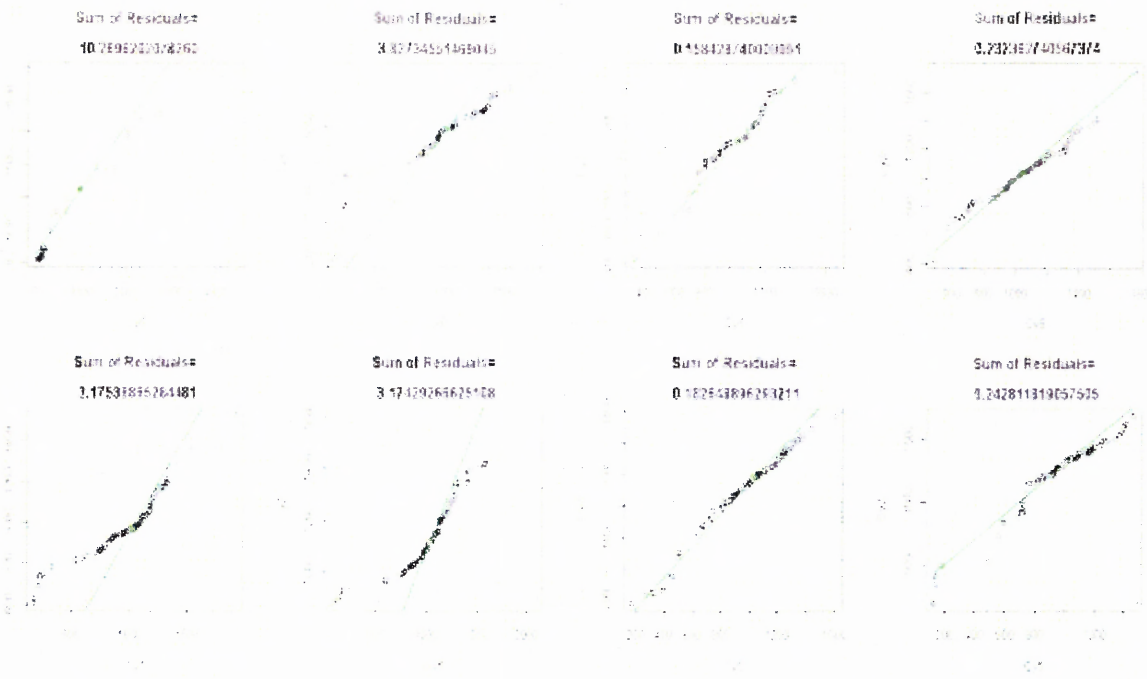

Figure 5.6 Sum of residuals analysis.

5.2 A New Parameter for Analysis

In this section, an idea for a new parameter to be used to analyze the data is given. This idea is based on what have been discovered so far, especially regarding the results of the previous section. In that section it has been shown how the median fold of change might not be representative of the whole chip. To assess the goodness of the median ratio a confidence measure based on a fit of the data has been evaluated. The new parameter that is going to be introduced is an extension of the work of the previous section.

The objective is to find a parameter that is as representative of the whole pixel spot distribution as possible, and relate it to the concept of fold of change. To achieve this goal, the data for every single spot are transformed as in the previous section. The pixels

for the green and red channel are ordered by intensity and then plotted against each other. Then, for each spot, the best line that is parallel to the main diagonal and minimizes the sum of residuals is chosen to be the median line for its spot. How far away this line is from the main diagonal will determine the absolute value of fold of change. While the position of the line above or below the main diagonal will determine if the spot is down (above the main diagonal) or up (below the main diagonal) regulated.

By using this parameter, researchers will be assured that the observed fold change reflects the distribution of every pixel within a spot. Unfortunately, from its definition, this parameter will be heavily altered by the presence of outliers. To expect better results form this type of parameter, a noise model should be introduced first, so that all the bad data are thrown away.

CHAPTER 6

CONCLUSIONS

The statistical analysis of the scanned image from a microarray experiment has revealed the existence of a big amount of information that has never been used before. This new information can be used in various stages of a microarray experiment, from quality control, to spot selection, to the final analysis.

This study has led to good results toward the identification of different biases by analyzing the pixel distribution for different slide's regions (print-tips, row of print-tips, spots). However, it has failed to identify the best way to tackle the problem. The same results have been obtained for other issues related to microarray experiments, like spot finding, spot quality, confidence on median ratio analysis. New ideas had been introduced to solve these issues, together with a theoretical validation of the assumptions behind them, through the pixel analysis of the image. Unfortunately, these ideas have not been tested accurately enough to become facts.

The main reason of this lack of practical application is due to the absence of a well-defined set of training data, which can lead to a proper definition of noise model for microarray experiments. The noise model can then be used to separate between good and bad data, thus allowing to accurately testing the ideas proposed in this thesis.

Another issue that is related to the validity of the results of this thesis concerns normalization. For many of the ideas of the thesis (i.e. sum of residuals to determine the confidence on the median ratio of a spot) the results obtained where not globally applicable to the slide. Instead, similar data (meaning data with similar intensity) confirm

the validity of the presented ideas.

Overall, the thesis fulfills the objective of extracting new meaningful information from the amount of data created by a scanned microarray slide. These data are never completely used for analysis, but only to extract statistical values (like the median) from the image. However, it seems that an accurate statistical (and biological) analysis of the image can lead to a better validation of the data contained in a single spotted arrays experiment and thus a better confidence on the biological information that can be inferred from the data.

APPENDIX A

MATERIALS AND METHODS

This Appendix contains the detailed specifications of the technologies used to create, scan and extract data from the image. The nine slides used for the analysis are described too, but not in detail since sample preparation and hybridization (among other steps) are not relevant for the material of this thesis.

Microarray Printing

Microarrays were printed on aged poly-L-lysine slides, using the OmniGrid microarrayer (GeneMachines) and quill type printing pins (Majer Precision). Relative humidity in the arrayer was held between 40% to 50%, with the room temperature at 24°C.

Hybridization

Hybridization buffer was heated for 2 minutes at 100°C then centrifuged at room temperature and 14,000 Xg for 2 minutes to pellet any particulate matter and facilitate cooling. After blocking, lifter slips (Erie Scientific) were placed over arrays and all but 2µl of the hybridization solution loaded, to avoid any precipitant carryover. Arrays were placed in hybridization chambers (GeneMachines) with 30µl of dH_2O pipetted at far ends to maintain humidity. Slides were incubated for 12 to16 hours at 65°C.

Scanning

For excitation of Cy3 and Cy5, both wavelengths, 532nm and 650nm were scanned simultaneously using the GenePix 4000 Microarray Scanner from Axon Instruments. The Center for Applied Genomics owns two of them and only one had the laser alignment problem described in chapter 2 of this thesis.

Data Extraction

The software used to extract the pixel data from the scanned image is GenePix Pro 3.0 from Axon Instruments. This software creates .txt file that consist of seven tab-delimited columns. The first three of them are used to identify the spot (block, row and column positions). Each of the last four columns consists of a vector of pixel intensities (each intensity is separated from the other by a comma) obtained from the image. There are four columns because there are four main pixel types: pixel obtained from red or green channels of a spot; and pixel obtained from red or green channels of a spot's background.

APPENDIX B

R CODE

This Appendix contains the R code used to do the analysis and create the figures of the thesis. Not all the programs used in the thesis are reported, but only the ones related with a specific figure and that require some programming challenge. Each program is presented through the referenced figure in order of appearance within the thesis.

Figure 3.3

This figure had been generated using the following two R functions, the first one is called **blockIndex**, takes as input the block position of a gene and returns the row of block to which it belongs, assuming that the microarray slide is 8 x 4 (blocks).

```
blockIndex <- function(x){
 x <- x/4
 if ( x <= 1) return(1)
 else if ( x <= 2) return(2)
 else if ( x <= 3) return(3)
 else if ( x <= 4) return(4)
 else if ( x <= 5) return(5)
 else if ( x <= 6) return(6)
 else if ( x <= 7) return(7)
 else if ( x <= 8) return(8)
}
```

The second function, called printTipDis, takes as input a list of files containing the pixel data, and evaluates the pixel density distribution of each row of block within the slide. Then the function plots each distribution with a different color. The same run is done for both the green and the red channel.

```
printtipDis<- function(x){
 par(mfrow=c(1,2))
 for( w in 1:(length(x))){
     dati <- read.table(x[w], header=TRUE, sep="\t")
     rowW1Ls <-list(row1W1=c(),row2W1=c(),row3W1=c(),
        row4W1=c(),row5W1=c(),row6W1=c(),row7W1=c(),
        row8W1=c())
     rowW2Ls <-list(row1W2=c(),row2W2=c(),row3W2=c(),
        row4W2=c(),row5W2=c(),row6W2=c(),row7W2=c(),
        row8W2=c())
     j <- 1
     while ( j<= length(dati[[1]])) {
        gene <- dati[[4]][j]
        g <- unlist(strsplit(gene, ","))
        g <- as.numeric(g)
        g <- blockIndex(g)
        gene <- dati[[5]][j]
        g2 <- unlist(strsplit(gene, ","))
        g2 <- as.numeric(g2)
        g2 <- nD(g2)
        index <- blockIndex(dati[[1]][j])
        rowW1Ls[[index]] <- c(rowW1Ls[[index]],g)
        rowW2Ls[[index]] <- c(rowW2Ls[[index]],g2)
        j <- j+1
     }
     yax <- paste("Density")
     xax <- paste("Intensity")
     print(paste(w," file Done!!!"))
     plot(density(rowW1Ls[[8]]),main=paste("Pixel
        Distribution by row \n Red Channel"),
        xlab="Intensity",col="green")
     lines(density(rowW1Ls[[7]]),col="red")
     lines(density(rowW1Ls[[6]]),col="black")
     lines(density(rowW1Ls[[5]]),col="orange")
     lines(density(rowW1Ls[[4]]),col="blue")
     lines(density(rowW1Ls[[3]]),col="pink")
     lines(density(rowW1Ls[[2]]),col="yellow")
     lines(density(rowW1Ls[[1]]),col="brown")
     plot(density(rowW2Ls[[8]]),main=paste("Pixel
        Distribution by row \n Green Channel"),
        xlab="Intensity",col="green")
     lines(density(rowW2Ls[[7]]),col="red")
     lines(density(rowW2Ls[[6]]),col="black")
     lines(density(rowW2Ls[[5]]),col="orange")
     lines(density(rowW2Ls[[4]]),col="blue")
     lines(density(rowW2Ls[[3]]),col="pink")
```

```
      lines(density(rowW2Ls[[2]]),col="yellow")
      lines(density(rowW2Ls[[1]]),col="brown")
      }
}
```

Figure 4.6

Figure 4.6 has been generated by the R function **cumAndDenPlot**, this function takes as input a list of spots indexes and the name of the file containing a scanned image from a microarray experiment, and outputs two plot for each spot, one representing the cumulative density function for the spot (together with a normality line) and the other representing the density distribution for the pixels within the spot.

```
cumAndDenPlot <- function(listOfspots,imageFile){
   par(mfrow=c(2,length(listOfspots))
   dati <- read.table(imageFile, header=TRUE, sep="\t")
   for( i in 1:length(listOfspots)){
      spot <-as.numeric(unlist(strsplit
          (dati[[4]][listOfspots[i]],",")))
      qqnorm(spot)
      qqline(spot)
   }
   for( i in 1:length(listOfspots)){
      spot <- as.numeric(unlist(strsplit
          (dati[[4]][listOfspots[i]],",")))
      plot(density(spot),main=paste("Density
          Distribution"), xlab="Intensity")
   }
}
```

Figure 5.1

Figure 5.1 has been generated by using two R functions: **findY** and **spotMedianMeanPlot.** The first one takes as input a number (that in this case will be

either the median or the mean) and a vector. Then retrieves the subscript of the element of the vector that is the closest (by value) to the input number.

```
findY <- function(el,vect){
   y1 <- f[1]
   y2 <- f[1]
   k <- 2
   ind <- 1
   while(y1 < x){
      y1 <- f[k]
      y2 <- f[k-1]
      ind <- k
      k <- k+1
   }
   y1 <- (y1-x)
   y2 <- (x-y2)
   if(y1 < y2) y <- ind
   else y <- (ind-1)
   return(y)
}
```

The function **spotMedianMeanPlot**, takes as input a spot index and a file image. Then it calculates the density distribution for the spot identified by the input index and it plots the distribution, together with the position within the distribution curve of the median (represented with a diamond) and the mean (represented with a full circle).

```
spotMedianMeanPlot <- function(spotIndex,imageFile){
   dati <- read.table(imageFile, header=TRUE, sep="\t")
   spot <- as.numeric(unlist(strsplit(dati[[4]][x],",")))
   d <- density(spot)
   plot(d,main=paste("Median Vs Mean"),xlab="Intensity")
   yMedian <- findY(median(g),d$x)
   yMedian <- d$y[yMedian]
   points(median(spot),yMedian,pch=23)
   yMean <- findY(mean(g),d$x)
   yMean <- d$y[yMean]
   points(mean(spot),yMean,pch=19)
}
```

Figure 5.2

This Figure 5.2 has been generated using the R function **medianPosition**. This function takes as input a spot size and a vector of string, with each string representing an image file name. The output of the function is a graph representing the distribution of the median by position within the pixel. This type of graph is plotted for every image contained in the list.

```
medianPosition <- function(spotSize,imageList){
   par(mfrow=c(1,length(imageList)))
   for( w in 1:(length(imageList))){
      medianCy5 <- c()
      medianCy3 <- c()
      dati <-read.table(imageList[w],header=TRUE, sep="\t")
      for(j in 1:length(dati[[1]])){
         spot <- dati[[4]][j]
         spot <- as.numeric(unlist(strsplit(spot, ",")))
         o <- order(spot)
         if(length(spot)==spotSize){
            medianCy5 <- c(medianCy5,o[(spotSize/2)])
         }
         spot <- dati[[5]][j]
         spot <- as.numeric(unlist(strsplit(spot, ",")))
         o <- order(spot)
         if(length(g)==s){
            medianCy3 <- c(medianCy3,o[(spotSize/2)])
         }
      }
      plot(density(medianCy5),main="Median Physical
          Distribution\n for spot with ",spotSize,"
          pixels size", xlab="Pixel Position", col="red")
      lines(density(medianCy3),col="green")
   }
}
```

Figures 5.3, 5.4, 5.5

These figures have been generated with the same three functions: **createMedianLine**,

orderSpot, **redVsGreenPlot**. Only few changes, due for display reasons, had been made to the **redVsGreen** function to create the four different figures.

The first function, **createLine**, takes as input two vectors of number and creates a parallel to the main diagonal the passes through the point with x-axis coordinate equal to the median of the first input vector and y-axis coordinate equal to the median of the second input vector.

```
createLine <- function(x,y){
    k <- (median(x)-median(y))
    l <- c()
    for(i in 1:length(x)){
       l <- c(l,x[i]-k)
    }
    return(l)
}
```

The function **orderSpot** takes a vector of pixel values from a spot and orders it in ascending order.

```
orderspot <- function(x){
   o <- order(x)
   l <- length(x)
   spot <- c()
   for(i in 1:l){
      spot <- c(spot,x[o[i]])
      i <- i+1
   }
   return(spot)
}
```

The function **redVsGreenPlot** takes as input a list of spots indexes and a string referring to a scanned image from a microarray experiment. For each spot in the list, the green and red channel pixels are ordered by intensity in ascending fashion and then plotted against each other. After, the line through the median pixel intensities and parallel

to the main diagonal is computed and plotted. Finally the residual errors from using the median line to fit the spot's pixels are evaluated, and the sum of residuals is computed and printed.

```
redVsGreenPlot <- function(spotList,imageFile){
   par(mfrow=c(2,(length(spotList)/2)))
   dati <- read.table(imageFile, header=TRUE, sep="\t")
   for(j in 1:length(spotList)){
      Cy5 <- c()
      Cy3 <- c()
      resid <- c()
      spot <- dati[[4]][spotList[j]]
      spot <- unlist(strsplit(spot, ","))
      Cy5 <- as.numeric(spot)
      Cy5 <- orderspot(Cy5)
      spot <- dati[[5]][spotList[j]]
      spot <- unlist(strsplit(spot, ","))
      Cy3 <- as.numeric(spot)
      Cy3 <- orderspot(Cy3)
      line <- createLine(Cy5,Cy3)
      for(i in 1:length(line)){
         tmp <- (Cy3[i]-line[i])*(Cy3[i]-line[i])
         resid <- c(resid,tmp)
      }
      sumOfResid <-sum(resid)/((median(Cy3))*(median(Cy3)))
      plot(Cy5,Cy3, main=paste("Sum of Residuals=\n
          \n",sumOfResid), xlab="Cy5",ylab="Cy3")
      points(median(Cy5),median(Cy3), pch=19, col="green")
      lines(Cy5,line, col="green")
   }
}
```

REFERENCES

1. Brown, P., and Botstein D. "Exploring the New World of the Genome with DNA Microarrays" Nature Genetics Supplement 21 (January 1999): 33-37.

2. Cheung, et al. "Making and Reading Microarrays" Nature Genetics Supplement 21 (January 1999): 15-19.

3. Bowtell, D. "Options Available-from Start to Finish- for Obtaining Expression Data by Microarray" Nature Genetics Supplement 21(January 1999):25-32.

4. Sawitzki, G. "Quality Control and Early Diagnostics for cDNA Microarrays." R News 2/1 (March 2002): 6-10.

5. Yang, Y., Dudoit, M., Luu, P., et al. "Normalization for cDNA Microarray Data: a Robust Composite Method Addressing Single Slide and Multiple Slide Systematic Variation" Nucleic Acids Res. 30/4 (February 2002): e15.

6. Dudoit, S., Yang, Y., and Bolstad B. "Using R for the Analysis of DNA Microarray Data" R News 2/1 (March 2002): 24-32.

7. Loader C. "Local Regression and Likelihood" Springer 1999.