CHAPTER II

cDNA MICROARRAY

Any sufficiently advanced technology is indistinguishable from magic.

- Arthur C. Clarke

In ancient the scientists had the genomic vexed problem that they could not solve even though every problem has a solution and the best answer depending on the best analysis method. The inspiration in puzzle of human genetic stimulated the scientists to propose a new innovation. Microarray is a novel technology containing a magic for data analysis. The magic of microarray analysis is broadly thought to the biological, agricultural and medical science, replacing tradition biology assay. This chapter provides an introduction to microarray including the history, definition, types and the procedure to analyze microarray data.

2.1 History of microarray

Early twentieth century, the scientists aimed to study the characteristics genes. They discovered the method to describe genes such as northern blotting, germ-line mutation, microsatellite but these methods are laboratorious time consuming. Each method is useful to study a few genes at a time. But living things have a lot of genes such as human have around 20,000-25,000 genes, therefore, it would take a very long time to investigate. Patrick O. Brown and David Botstein pioneer from Stanford University discovered the technique for measure mRNA transcription (gene expression) with single expression in 1999 so called "Microarray" (1). The

microarray was emerged as a powerful technique to study the broad biological question and assess difference of mRNA which is abundance in difference biological samples. Moreover, microarray technology is used to explain the molecular characterization in point of monitoring gene expression levels that may be differentially expressed by chance on a genomic scale.

2.2 Definition of microarray

Microarray has the initial concept of the hybridization properties of a nucleic acid perfectly complement a supplement molecule. Microarray is a solid substrate (such as glass slide, nylon membrane and silicon thin-film cell) that consists of DNA sequences of know genes deposited in a regular grid like array (Figure 2.1) (2). An array has many thousands of spots, each position on array containing the mRNA molecules. Sometime a spot can be considered as microarray element. The target sequence may be in the form of cDNA or oligonucleotides, while other materials for example genomic DNA clones may be deposited as well. Typically, target DNA sequences on the surface of an array are immobilized (Figure 2.2) which taken from of mRNA converted to cDNA. The target spots on array are located orderly and the DNA sequence of each spot is precise recorded in a computer database. Whereas, DNA is free on the surface of grid is called "probe" (3). Probes are unfixed nucleic acid extracted from biological sources of interest, such as tissue from normal cell or disease cell or mutant organism. Typically, probes derivatived from RNA, often use mRNA, is converted to cDNA and labeled with opposite fluorescence dyes such as cyanine dyes Cy3 or green dyes and Cy5 red dyes or radioactivity later hybridized with mRNA from two independent tissues on to the array surface (Figure 2.2) (2).

This process is according to the Watson and Crick theorem which described a complementary of single strand that allow to study the position of hybridization. Moreover, cDNA complementary reflects gene expression or the activity of gene.

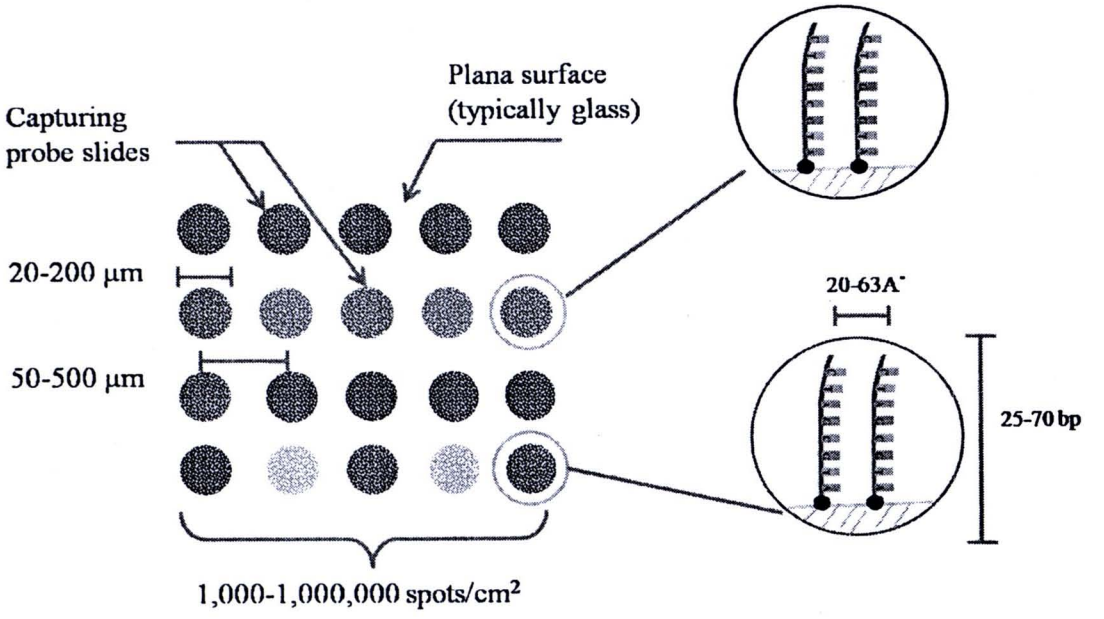


Figure 2.1 Anatomy of DNA microarray.

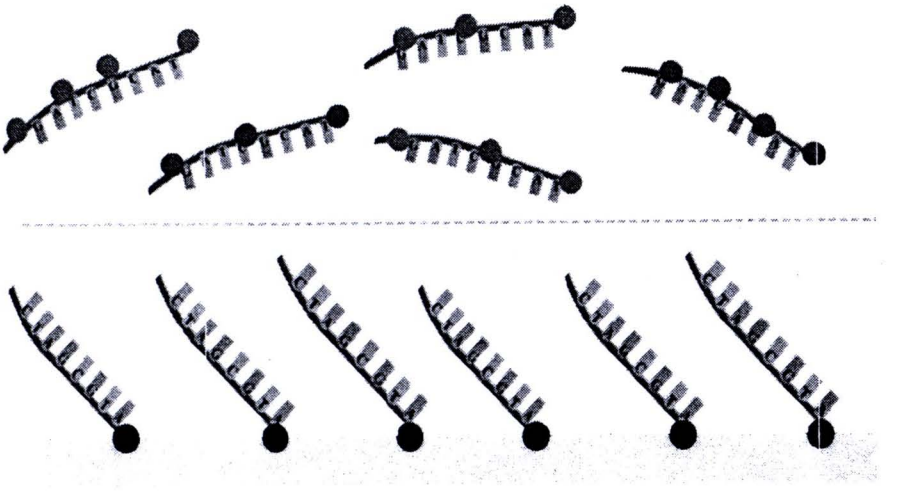


Figure 2.2 The immobile mRNA on glass slide is called target. The mRNA is labeled with fluorescent dyes that flow on solid surface called probe.

2.3 Type of microarray

DNA Microarrays can rapidly detect genetic variants, provide a genome-wide and genotype analysis according to their types. Types of microarray depend upon the kind of immobilized sample onto microarray support and the information obtained. The microarray experiments can be categorized in three classes, microarray for expression analysis, mutation analysis and comparative genomic hybridization. For microarray expression analysis, the cDNA immobilization on solid substrate is usually derived from the known gene. Gene expression is studied by comparing two different cells condition. The solutions represent an over-expression or down-expression of an individual gene. This experiment is the famous of clinical study such as compare between normal vs. cancer, cell line with vs. without drug treatment. Microarrays for mutation analysis; a mutation is a permanent change in a DNA sequence and can be transmitted to offspring. For this type, the immobilization is cDNA and usually the probe is a single nucleotide base that gene may be difference from the parental by exchange of a single base as also known as Single Nucleotide Polymorphism (SNP). The analysis of this method is detecting a single base difference between two sequences that so called SNP detection. For comparative genomic hybridization this method is the scanning for signals (sufficiently intense and specific) entire genomes for variations in DNA copy number changes. It is usually used for identification in the increase or decrease intensity ratio. The well-known applicant for this type is detection chromosomal fragments harboring genes involved in a disease (4).

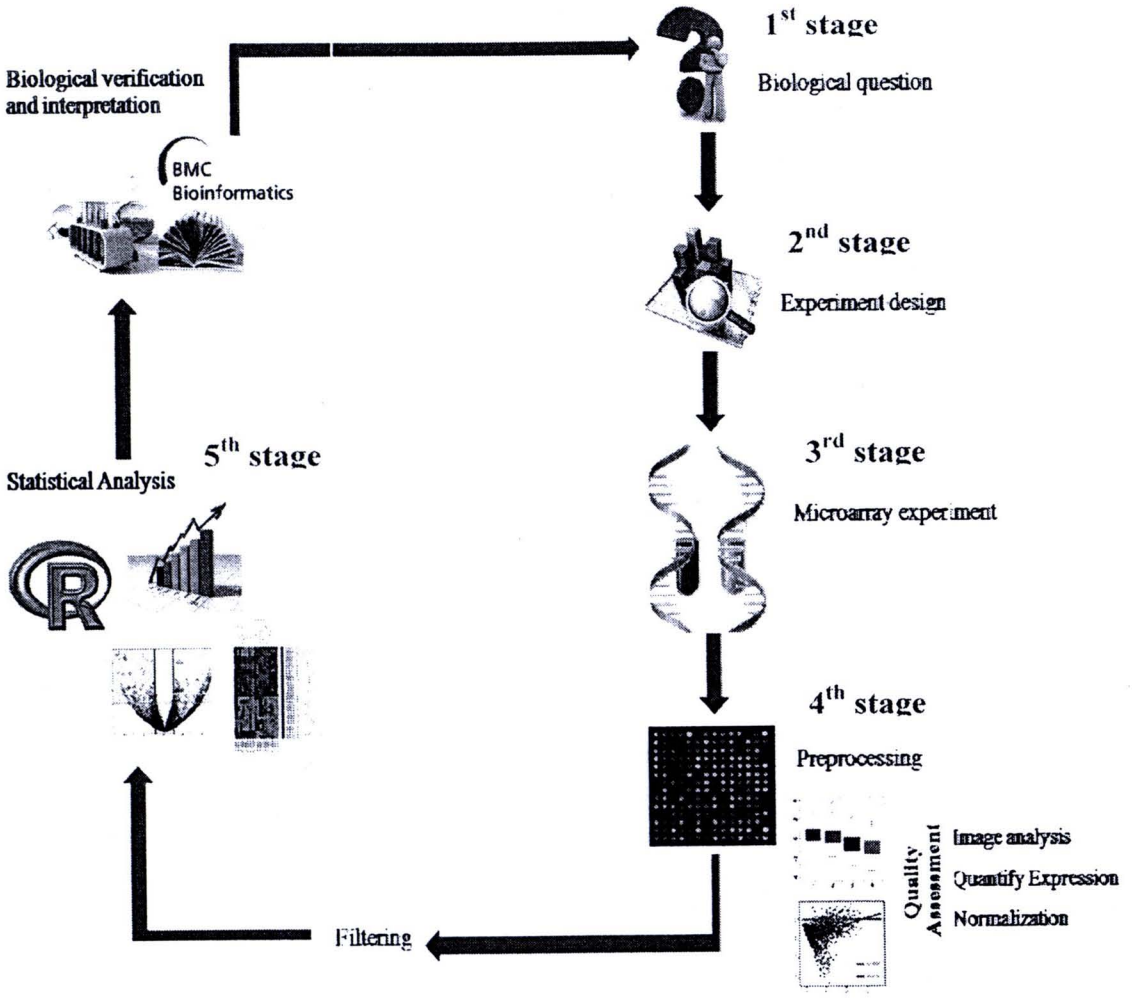


Figure 2.3 The six steps of microarray analysis cycle.

2.4 The procedure to analyze microarray data

The dramatically increase of the genomic data, abundance genetic for study their functions. This problem can be resolved when microarray was introduced in mid-1999. Figure 2.3 shows a microarray concept consisting of five stages, biological question, experiment design, microarray experiment, pre-processing and statistical analysis.

2.4.1 Biological question

Basically, every experimental work starts in principle with a question. In the bioinformatics field, the biology question applies to the molecular biology. Bioinformatics is a novel informative science developed to manage the accelerated growth of biological data to serve the goal of Human Genome Project in analyzing the sequence of human genome. The major problem is how to analyze and manage the current high-quality biological data. The bio-informatics is the conceptual rising of biology in term of molecules applying to the informatics technology, a multidiscipline to understand and organize the information derived from mathematical, statistical and computer. Therefore, bioinformatics becomes essential for biological experiment that needs data management and data analysis as human ability to record or analyze the large scale data is limited. In twenty century, plenty of bioinformatics including either free software or commercial software are available. The biological data source is allowed for public access to learn and investigate from any place in this global. Figure 2.4 is a cartoon showing the concept of bioinformatics. The popular biological question is gene expression that vividly described as the flow of genetic information.

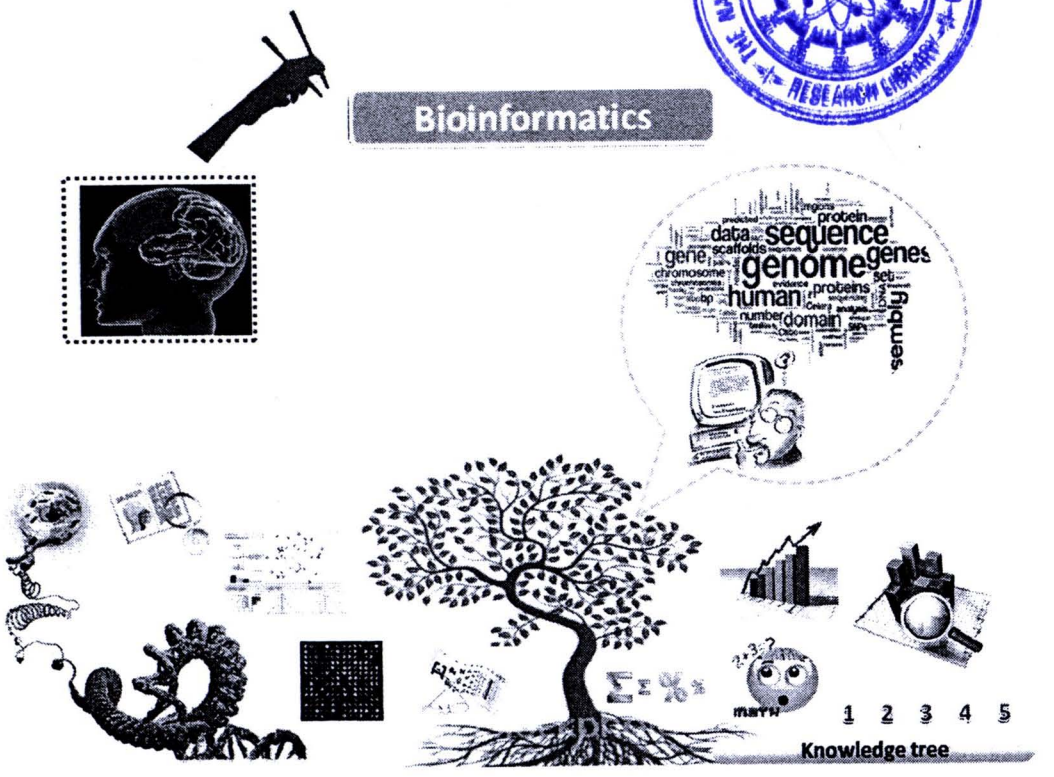


Figure 2.4 The bioinformatics is a novel multidiscipline of molecular biological, mathematics, statistics and computation.

Francis Crick coined the key of the flow of genetic information in living organism cell termed "central dogma" in 1958 (5). The theory of central dogma includes three main processes: DNA synthesis (replication), RNA synthesis (transcription), and protein synthesis (translation) (Figure 2.5).

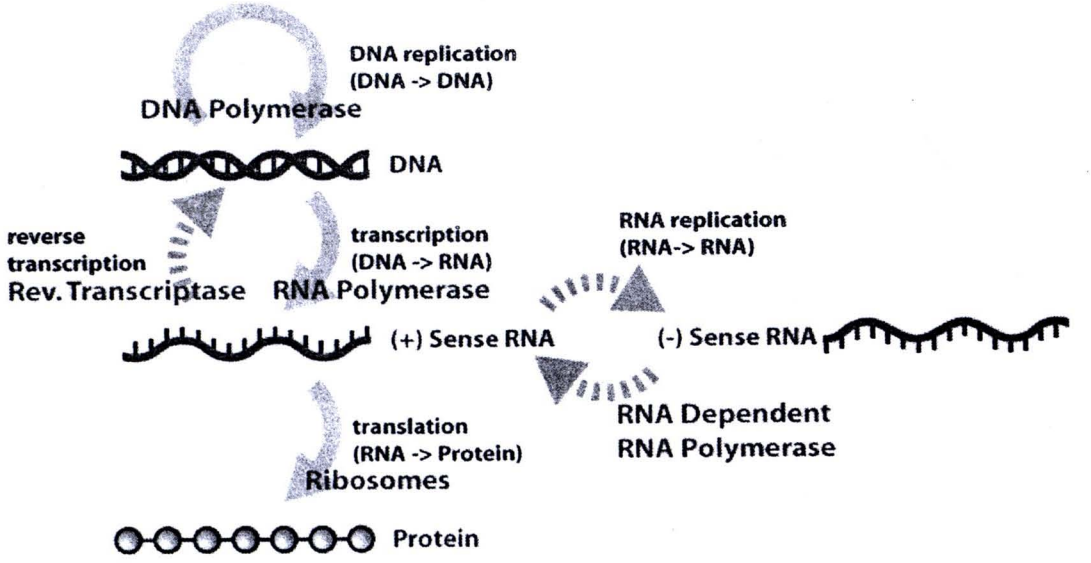


Figure 2.5 Central dogma (6).

DNA synthesis (replication) is semi-conservative in which each DNA stands serve as a template. For each new DNA copy, the double-stranded of its synthesizes the molecule one: from parent strand and the other: new daughter strand. The new DNA copy strand complementary to parental bases. Sometime the DNA replication process is defined as a process to copy the genetic information.

The transcription is a process of copying the genetic information from DNA to RNA. RNA carries the genetic information as a single-stranded form, which is impermanent, less stable and easy to degenerate from the chemical substance or environment. However, the RNA is necessary in the flow of genetic information since DNA is never directly translated to protein. In organism cells, RNA usually acts as messenger molecules that carry the information copy from DNA. The translation is the process described how mRNA molecules on RNA strand direct to synthesize the proteins. The mRNA is translated into protein by elaborate pieces of cellular machinery called ribosome. If mRNA is not made from a gene, then no protein will

be produced and the gene is said to be silent or inactive. In contrary, if the gene is transcribed into RNA copy, the mRNA is said to be expressed. The caution of the state of information is that protein never back-translated to RNA. But except for the retroviruses, DNA is never created from RNA. The synthesis of DNA from RNA by reverse transcriptase is reverse transcription process. Many tools and technologies are provided for gene expression study including microarray technique.

2.4.2 Experiment design

Following biological question, the experimental design is necessary for selection of the samples suitable for the objective. Thus before experimental design, the objective should be cleared. The hypothesis is really important to design the experiment. The microarray prominent hypothesis is that there are genes whose expression is down-regulated or up-regulated in the system comparing between tumor and healthy samples. The best experiment should be suitable for the objective and hypothesis that lead to be successful of microarray analysis. Basically, the experimental design apply to microarray are two broad aspects; designing the array and allocation of sample to the slides (7). The designing of the array grasps meaning in which cDNA sequence to print, what library to spots and what quality controls to include. The later, allocation of samples to the slides, refers to the assignment of dye labels to the samples and to determine which samples should be paired and hybridized on the same slides. The main problem is the use of reference sample, typical the references are labeled with green dry. Many experiment designs are provided for single-color or two-color microarray experiment such as reference design, loop design as shown in Figure 2.6.

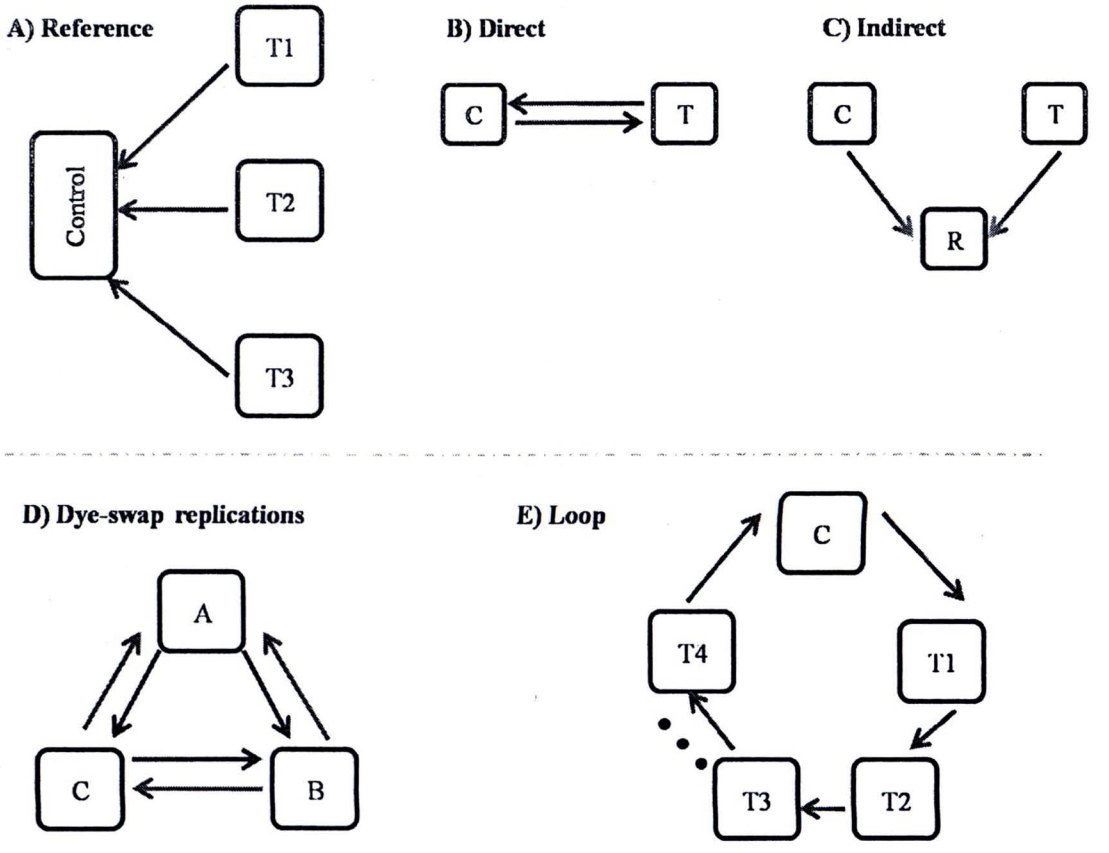


Figure 2.6 The most common microarray design. A) Reference design B) Direct comparison C) Indirect comparison D) Dye-swap designs E) Loop design.

The reference design (Figure 2.6A) or natural design choice is the most common design for two-color. This design depends on the material available and biology condition. The reference designs can be separated into two categories: direct or indirect comparison (8). For example, in comparing gene expression in two samples T and C with direct comparison design, sample T and C is measured directly in the single slide while indirectly comparison T and C are measured separately with difference slide design (Figure 2.6B-D). Dye-swap experiments or dry-swap replicate is a design that need to be done during the hybridization processing in twice, and second hybridization labeling with the reverse dye. It uses two array and two samples

design (Figure 2.6D). The first array sample C is labeled with green dye and C is labeled with the other while second array the dyes are interchanged. The simplest of loop design is described in each sample compare directly with other samples in a designed cycle (Figure 2.6E) (9). A drawback of loop design is that each sample must be labeled with both red and green dyes. The variance error of all data estimates is double and difficult to estimate. Time-course experiment depends on the comparison of interested sample. This design needs multiple-slide experiment according to the transcript abundance of gene over time points or after treatment or stimulation. The factorial design as suggested by Fisher R.A in 1926 (16), is very interesting in clinical trials. It allows to study the effect of two or more than one experimental factor simultaneously. For example, which genes respond differently to stimulation in tumor compared to normal cells. The factorial designs for microarray are used to study both the gene expression differences caused by single factors alone. The results of the integration effect of two or more factors, especially when it differs from that might be predicted in the basis of factors separately, also called interaction, quoted by Speed T. (2003) (10).

2.4.3 Microarray experiment

The microarray experiment process is shown in Figure 2.7. The procedure after collecting sample is RNA extraction from both normal or untreated cells and diseased or treated cells by using reagents. After that, the RNA should be purified. The purity and quality of RNA can be confirmed by Northern analysis or PCR. To collect mRNA, aRNA (Poly [A] RNA) is used for RNA isolation. The mRNA is then converted to a colored cDNA by the labeling procedure. The mRNA molecules are

labeled by two difference dyes. The traditional two-color microarray are green-fluorescine cyanine dyes (Cy3) and red-fluorescine cyanine dyes (Cy5). The mRNA from normal and disease is labeled by Cy3-dyes and Cy5-dyes respectively.

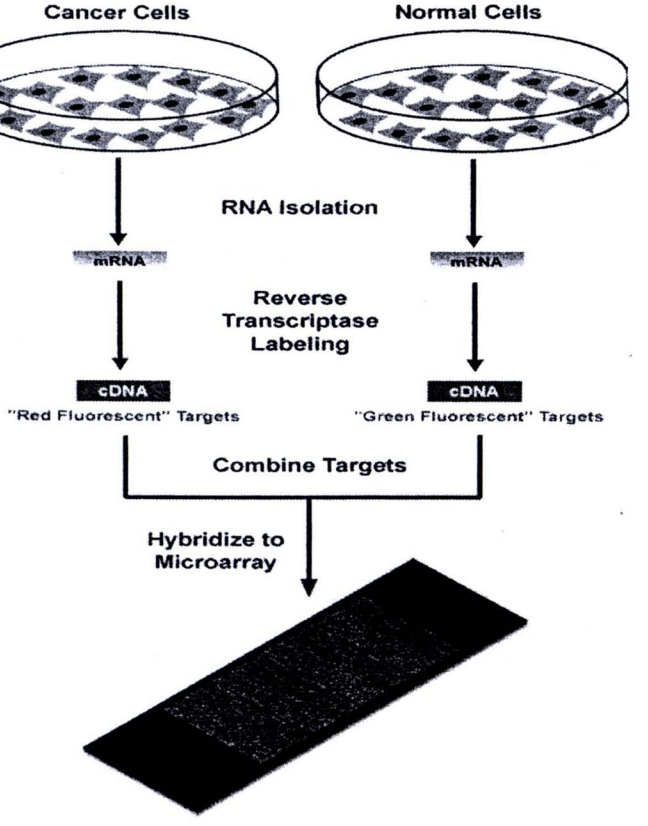


Figure 2.7 The procedure of microarray experiment (11).

The fluorescent molecules incorporate into mRNA molecule at the time of conversion to cDNA. The two labeled cDNA are mixed and hybridized to the microarray slide. The populations of two probes labeled are competitively hybrid with complementary targets. The candidates will find to join and disjoint some base pair for reforming a double-stranded DNA molecule. Most of the complementary cDNA stands can hybridize but few complementary cDNA stands stray without hybridization which will be washed off by washing solution. The slide is scanned by the microarray

scanner. The intensity of dyes for each spot is detected by laser scanning. The scanning process consists of three steps, first, scanning throughout the microarray to detect green-labeled cDNA signal from normal tissues, second, scanning in the same array to detect red cDNA signal from disease. Finally, the red image is merged with the green image and the composite picture is printed. A spot with equal green-labeled and red-labeled is presented as yellow color if none is presented as black.

2.4.4 Pre-processing data

Basically, preprocessing is a process of preparation the intensity for the arrayed probes to eliminate the effect of background noise. This method is to control microarray data effecting from systematic error while retaining biological variation (RNA isolation, probe labeling, hybridize and scanning) leading critically to the validity of the results obtained. Preprocessing method includes three main steps: image analysis, expression quantification, and normalization.

Open-source software for cDNA microarrays: the Bioconductor project and R

R is a freely available computer language associated with statistical field, usually using statistical computing and graphic. It can be accessed everywhere and run on laptop with Window, Unix, MacOs, etc, and permit everyone to learn with a lot of R tutorial. It is the most computer language used for analysis an enormous data for example biological, genetic, medical data. In addition, R has many contributors, invest the time and concentrate on newly developed methods for interactive data analysis, especially in Bioinformatics applications, including microarray data. The software has been developed rapidly, and extended by a large collection of the

functionality provided by additional units of software called packages. The bioinformatics field uses R packages based on Bioconductor project – the development project aimed to offer tools for an analysis of genomic data (e.g. sequence, microarray, annotation and many other data types). Limma is a term of linear models for microarray data. It is a well-known software package of R language. Limma package is usually used for analysis of gene expression from microarray data, detection genes with a significant difference expression in microarray experiments (2).

2.4.4.1 Image analysis

Generally, image generation is occurred after hybridization of two different cDNA labeled and capture their signals as laser stimulates the fluorescent dyes to emit protons whether or not the proton emissions is emit with different wavelength depend on type of fluorescent dyes. Moreover, proton emission can be selective filtered to quantification of the amount emitted by each dyes. The emitted light is called intensity of fluorescent, measured by using dense grid of pixel located on array (12). The implicit assumption for microarray image is that signal intensity referred to the mRNA level. Microarray image analysis aims to ensure that the intensity indicator from each spot represents the expression levels (Figure 2.8). This procedure is also used to select the feature for further analysis. The procedure of microarray analysis consists of gridding, segmentation, and information extraction.

The useful additional aspect associated with image processing is the visualization of microarray image. The input data of microarray images is defined

with two variables; "R" and "G", R is red color corresponding to Cy5 dry and G is green corresponding to Cy3 dry.

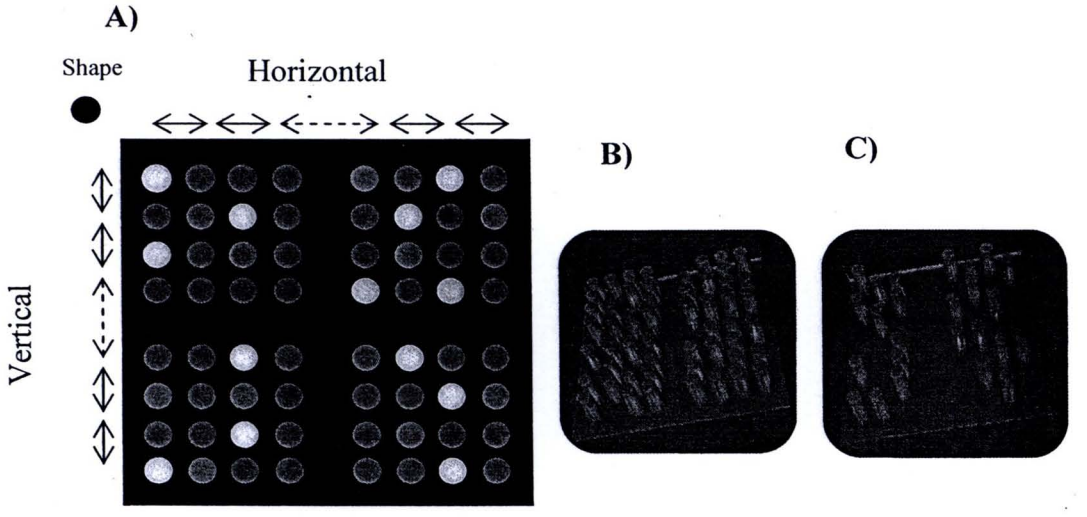


Figure 2.8 Ideal image processing ; each spot should has only intensity distribution. A) A single array composes of different spots intensity. Obviously, three colors are summarized microarray data after scanning process with microarray scanner.
B) Spots on an array are detected by red laser excitation. (C) Spots are detected by green laser excitation. Yellow spots are found in both of laser scanning (3).

Gridding

Basically, gridding is a mathematic method of interpolation data from an arbitrary 2D or 3D pattern (surface, vector and image) to a uniform grid. Gridding method for image is a process to get point on the sphere corresponding to points on a flat image. The spots on microarray are arranged in a rectangular pattern thus microarray image gridding method usually uses a rectangular grid mask to patch the pixel (Figure 2.9) (9, 10 and 12). In other word gridding is used for locating each spot. The technique for gridding, before patching the grid mask area, whole slide

should be scanned to observe the spot intensity followed by the use of the right first mask spot as the control spot which usually mask for the housekeeping gene intensity, and after that the grid mask can be moved to cover most of expressed spots subject to the grid size (Figure 2.10).

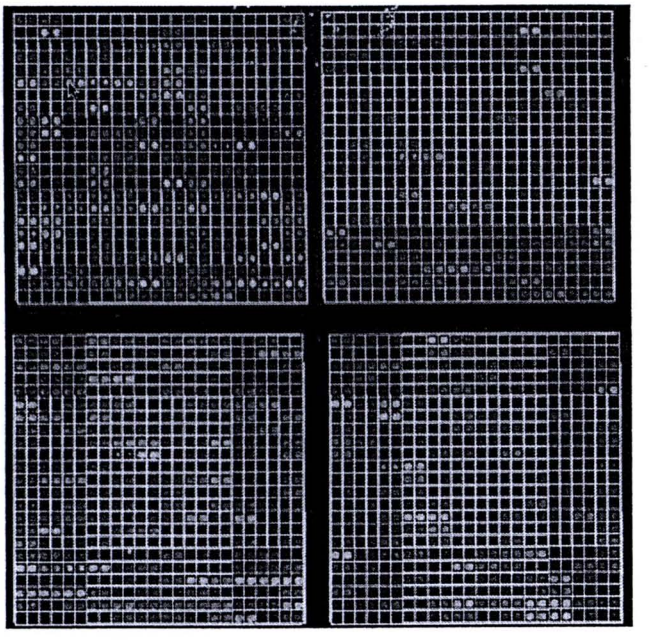


Figure 2.9 Grid array patch on the array (13).

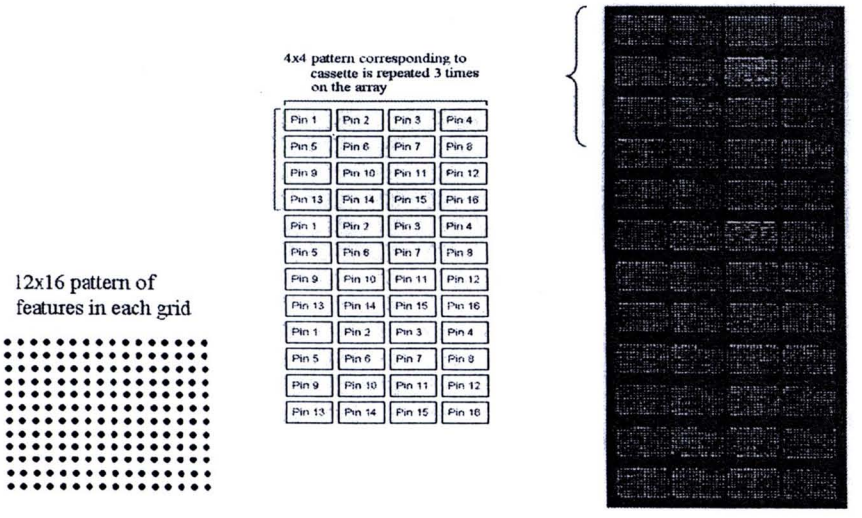


Figure 2.10 An example image of microarray. The spotting robot has 16 pins arranged in a 4×4 pattern. The 16 pin is repeated three times on array. Each grid consists of 192 features arranged in a 12×16 patterns (14).

Segmentation

Followed the gridding process is a segmentation to identify accurately a single overall intensity value for each spot. Generally, segmentation method is a process of partitioning the image into different regions. The other meaning segmentation is a method to classify pixels into either foreground or background. Foreground is set of pixels, within printed cDNA spots. Thus the fluorescent intensity for each spot of cDNA sequence can be computed as the transcription abundance. Segmentation method for microarray can be categorized into four groups depending on the geometry of the spots which they produce: fixed circle segmentation, adaptive circle, adaptive shape segmentation and histogram segmentation (Figure 2.11).

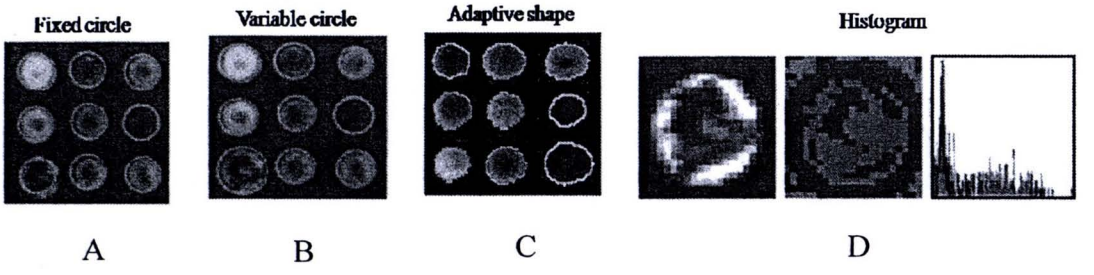


Figure 2.11 The segmentation results for (A) fixed circle, (B) Adaptive circle, (C) Adaptive shape, (D) Histogram segmentation (15).

Fixed circle segmentation fit the circle diameter to all spots in the image. Theoretically, if background effects foreground thus foreground values additively background value extensively intensity cannot be reliably estimate but one could use a large of cycle diameter for segmentation such that entire spot covered of all spots. Occasionally, method of fixed cycle diameter is too large can yield perfect good (unbiased) estimate is background distribution is remove. So, fixed circle segmentation should be adjusted by improve the circle segmentation. Adaptive circle segmentation is a method of using intensity data to estimate the best cycle diameter spot by spot that method are divided into three methods; adaptive circle, shape and histogram circle segmentation. The adaptive circle segmentation depends on the circularity constraint by finding the brightest ring cover the pixel with high value. The adaptive shape segmentation requires the specification of starting point and adaptive circular collect the pixels classified as foreground. The histogram segmentation is a method to detect each spot, foreground and background intensity are estimated and determined in some fashion from the histogram values for each

pixel in the square target area (10, 12 and 16). The segmentation methods are provided by commercial software as shown in Table 1.

Table 2.1 Segmentation methods and examples of software implementation (16).

Fixed circle	**ScanAlyze, GenePix, QuantArray**
Adaptive circle	GenePix, Dapple
Adaptive shape	Spot, region growing and watershed
Histogram method	ImaGene, QuantArray, DeArray and adaptive thresholding

Information extraction

Segmentation used to identify the pixel as either foreground or background region. Next, the feature intensity extraction uses to ensure the signal intensity for each spot that including three steps: foreground fluorescence intensity pairs (R, G), background intensities, and quality determination. The first step, foreground fluorescence intensity pairs (R, G) aims to detect the level of hybridization at the particular area on the slide. Foreground intensities are calculated according to the total amount of pixel intensities within the spot mask. The second step is extraction background intensities by assuming that the spot intensity includes a contribution from non-specific hybridization and other chemical on the glass. Background intensity is the signal intensity of spot feature adjustment by measuring and removing the perturbation intensity. The estimation of background intensity methods are implemented in four differences software packages; local background, morphological opening, constant background and no background adjustment. Local background intensities estimate the background intensities by focusing on small regions

surrounding the spot mask. The local background usually estimates the median of pixel values within specific local regions. Morphological opening approaches to background adjustment by using the non-linear filter. Constant background is a global method which subtracts a constant background for all spots. No background adjustment is a method to calculate log-ratios with no background adjustment (15, 16).

Quality assessment

Before proceeding to normalize or any high level analysis of the microarray data, the quality assessment is an important step to ensure that the quality of the data is suitable for analysis. The components of quality assessment before normalization include spot quality, array quality, and Ratio (2 spots combined). These three quality assessments can be further separated. Brightness method from spot quality determines the quality by calculating the ratio intensity between foregrounds and background intensities. The array quality assesses variance of replicated control spots. While the Ratio (2 spots combined) considers the signal to noise ratio;

$$log_2(\frac{fg.median}{bg.median})$$

The other methods of quality assessment see Stekel D (14).

2.4.4.2 Expression quantification

The problem for microarray image is a various distribution of the spot intensity before using image data for further analysis in expression quantification. Figure 2.12 shows the distribution for each spot consisting of the foreground intensity and

background intensity. Before further analysis, the background intensity should be subtracted from these data. The data should contain only foreground intensity, and transform data into the same scale.

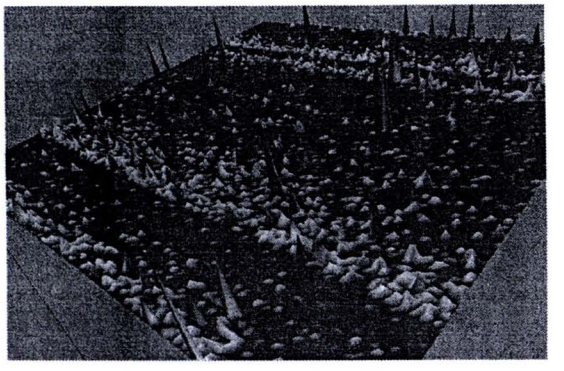


Figure 2.12 The various intensity distributions of whole spot on microarray glass slide (17).

Background correction

Background correction is the image transformation. Background correction is the process to subtract the background intensity from the foreground intensity of each spots based on background estimation. In Figure 2.12, the mRNA expression can be represented various differentially expressed distribution. This procedure aims to collect only foreground intensity (Figure 2.13). The estimation of background intensity should consider all variables that are sensitive to the image. The variable can be occurred from difference source including mRNA preparation, probe labeling, the incomplete washing after hybridization, feature of the slide that bind dye or RNA, and imprecision in spot location (segmentation) during image acquisition. For more comprehensive list of variable source in cDNA microarray see Schuchhardt *et al.*,

(2000) (18). Various algorithms to adjust adaptively foreground intensity in microarray background correction are subtraction, minimize method, and model.

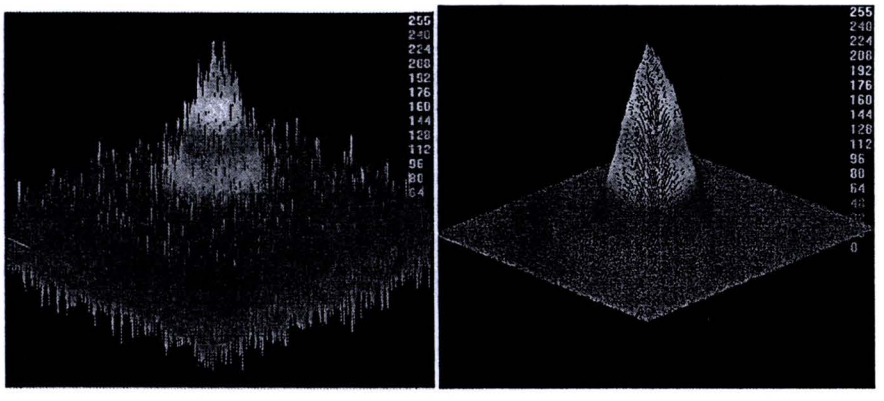


Figure 2.13 The idea of background correction for one spot in 3D

a) a spot intensity contain variation of the intensity (foreground + background intensity)

b) only foreground intensity (17).

Background correction defined R_b and G_b as red and green background intensities respectively. R and G is true foreground signal intensity determined from the raw data while R_f and G_f are defined as red and green foreground intensities with background The foreground intensities can be calculated by the simple mathematical formula; $R = R_f - R_b$ and $G = G_f - G_b$. The corrected intensities R and G are usually expressed in log-scale; M is the standard abbreviation for the log base 2 of different ratio: $M = log_2(\frac{R}{G})$ and A is an abbreviation of the average log signal: $A = \frac{1}{2}(log_2(RG)$ for each spot. The traditional method for background correction is subtraction. The R_b and G_b value are estimated from local background and subtract them from the foreground intensity. Kooperberg *et al.*, (2002) (19) suggested the

novel method, modified from empirical Bayes model, involving a convolution of a normal distributions to background adaptively adjust the foreground intensity from each spot (19-21). The Kooperberg model (19) is shown in Equation 2.1

$$p(\mu|\sigma_b,\sigma_f,X_b,X_f)=\frac{\emptyset\left(\frac{X_{f-\mu}-X_b}{\sigma_d}\right)\Phi\left(\frac{(X_{f-\mu})\sigma_b^2+X_b\sigma_f^2}{\sigma_f\sigma_b\sigma_d}\right)}{\sigma_d\int_0^{\infty}\Phi\left(\frac{X_f-v}{\sigma_f}\right)\emptyset\left(\frac{X_b-v}{\sigma_b}\right)dv} \tag{2.1}$$

where X_f, X_b is the observed foreground and background mean intensity respectively and $\Phi(\cdot)$ is the density of standard normal distribution, $\Phi(x)=\int_{-\infty}^{x}\emptyset(y)dy$ is cumulative normal standard distribution, $\sigma_f=\frac{aSD_f}{\sqrt{n_f}}$, $\sigma_b=\frac{aSD_b}{\sqrt{n_b}}$, $\sigma_d=\sqrt{\sigma_b^2+\sigma_f^2}$ and a is a scale factor. The expected values for true signal, $E(\mu|X_b,X_f,\sigma_b,\sigma_f)$, are calculated in each color channel for each spot. The $a-values$ are estimated separately for each channel of background intensity defined by $(\frac{SD_b}{\sqrt{n_b}})$ (20-21). Edwards (2003) (21) suggested the approach to avoid the negative intensity value after foreground intensity adjustment by subtraction method. The subtraction is replaced by a smooth monotonic function model as foreground and background intensity difference is less than a threshold value δ. The smooth monotonic functions of red and green intensities are shown in Equation 2.2 and 2.3 respectively.

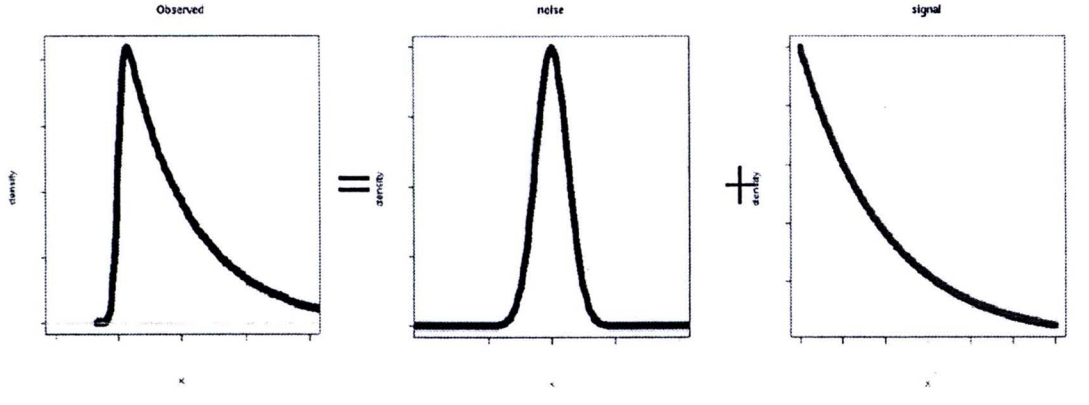
$$R=\begin{cases}R_f-R_b & if\ R_f-R_b>\delta\\ \delta\exp\left[1-(\frac{R_b+\delta)}{R_f}\right] & otherwise\end{cases} \tag{2.2}$$

$$G=\begin{cases}G_f-G_b & if\ G_f-G_b>\delta\\ \delta\exp\left[1-(\frac{G_b+\delta)}{G_f}\right] & otherwise\end{cases} \tag{2.3}$$

The popular algorithm usually used to fully assess the differential gene expression is normal-exponential convolution model (normexp method) proposed by Silver JD

et al., (2009) (22). Figure 2.14 shows a spot that compose of two distributions: one from signal intensities and the other from background intensities.

The normexp method for red channel assumes the model as $R_f = R_b + B + S$, where B is the residual background not captured by R_b and S is true expression intensity signal, similar for green channel model. The S is assumed as the exponential distribution with mean (α). The exponential distribution is usually defined by $E(\alpha)$ and B is normal distribution with mean (μ) and variance (σ^2) that defined by $N(\mu, \sigma^2)$. These parameters μ, σ^2 and α are independently and differentially assumed between each channel on each array. The normexp model is written clearly as $X = B + S$ and defined $X = R_f - R_b$ for a background subtraction from the foreground intensity thus $X = B + S$, where $B \sim N(\mu, \sigma^2)$ and $S \sim E(\alpha)$.



Spot intensity(X) Background intensity(B) Signal intensity(S)

Figure 2.14 The ideal distribution of spot intensity of normexp convolution model (22).

The joint density of B and S is described as the product of density (Equation 2.4):

$$f_{B,S}(b,s;\mu,\tau,\alpha) = \frac{1}{\alpha}\exp\left(\frac{-s}{\alpha}\right)\emptyset(b;\mu,\sigma^2) \tag{2.4}$$

where $s > 0$ and $\emptyset(\cdot)$ is a Gaussian density function. The simple transformation to give the density of X and S is

$$f_{X,S}(x,s;\mu,\sigma^2,\alpha) = \frac{1}{\alpha}\exp\left(\frac{\sigma^2}{2\alpha^2} - \frac{x-\mu}{\alpha}\right)\emptyset(s;\mu_{S\cdot X},\sigma^2) \tag{2.5}$$

where $\mu_{S\cdot X} = x - \mu - \sigma^2/\alpha$. Integrating over s gives the marginal density of X:

$$f_X(x;\mu,\sigma,\alpha) = \frac{1}{\alpha}exp\left(\frac{\sigma^2}{2\alpha^2} - \frac{x-\mu}{\alpha}\right)[1 - \Phi(0;\mu_{S\cdot X},\sigma^2)] \tag{2.6}$$

where $\Phi(\cdot)$ is the Gaussian distribution function. Dividing the joint by marginal gives the conditional density of S given X as

$$f_{S|X}(s|x;\mu,\sigma,\alpha) = \frac{\emptyset(s;\mu_{S\cdot X},\sigma^2)}{1-\Phi(0;\mu_{S\cdot X},\sigma^2)} \quad , \text{for } s > 0 \tag{2.7}$$

Estimated signal with given observed intensities is the conditional; expectation function is shown in Equation 2.8.

$$\mathbb{E}(S|X = x) = \mu_{S\cdot X} + \frac{\sigma^2\emptyset(0;\mu_{S\cdot X},\sigma^2)}{1-\Phi(0;\mu_{S\cdot X},\sigma^2)} \tag{2.8}$$

The expectation of true signal intensities be able to perform strictly positive that adjust each spot for each array (Figure 2.15) (21).

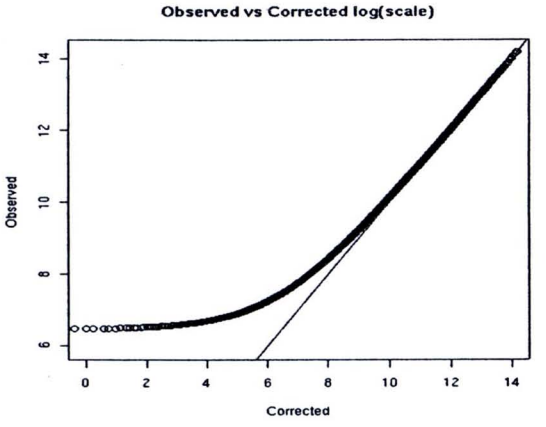


Figure 2.15 Adaptive background correction produces positive signal (22).

Diagnostic plot

After the quality assessment of microarray data from image processing, it can be ensured that the data is high quality suitable to further normalization. The diagnostic data usually uses the graphical visualization of red and green foreground and background as log-scale, log intensities or intensity log2 ratio (M-value) or average log2 intensities (A-value). Layout parameters as print-tip group or plate is stratified microarray spots. The purpose of stratifying method is to identify signal printing, hybridization, riding scanning artifacts. The diagnostic plot often uses box plot, histogram, density plot, scatter plot, etc. Box plot proposed by Tukey J. (1997) (10), is a plot consists of box and whisker. It is an excellent tools attempt to summarize the information in the data set as the conveying location and variation changing between information groups. Figure 2.16 shows a box plot with median, upper inner fence (Maximum) and lower inner fence (Minimum), inter-quartile rang (IQR) that represented the different between first quartile (25%) and third quartile (75%) and the middle of box is second quartile (50%) or median of the information. The outside

values are defined values greater than 1.5 IQR above 75th percentile and less than the 25th percentile.

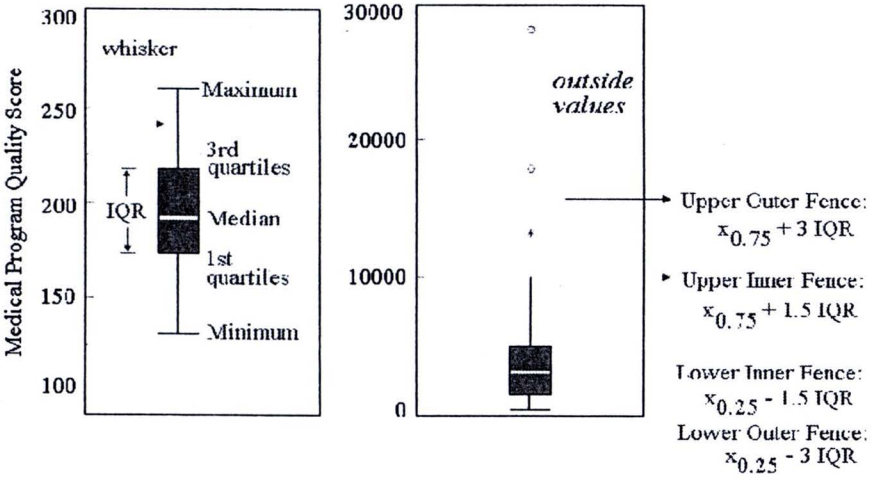


Figure 2.16 Anatomy of box plot (15).

The scatter plot is a visualization of microarray spot statistics. The goal of scatter plot is to identify genes that are differentially expressed between two experimental conditions. The points on the plot represent the intensity-dependent ratio of raw microarray data. The typically scatter plot represents the microarray data using a MA-plot, it first appeared in Luu *et al.*, (2001) (10). MA-plot has two dimensions of which intensity log-ratio $M = log_2(\frac{R}{G})$ on the y-axis and the average log intensity $A = \frac{1}{2}(log_2 R + log_2 G)$ on x-axis (Figure 2.17).

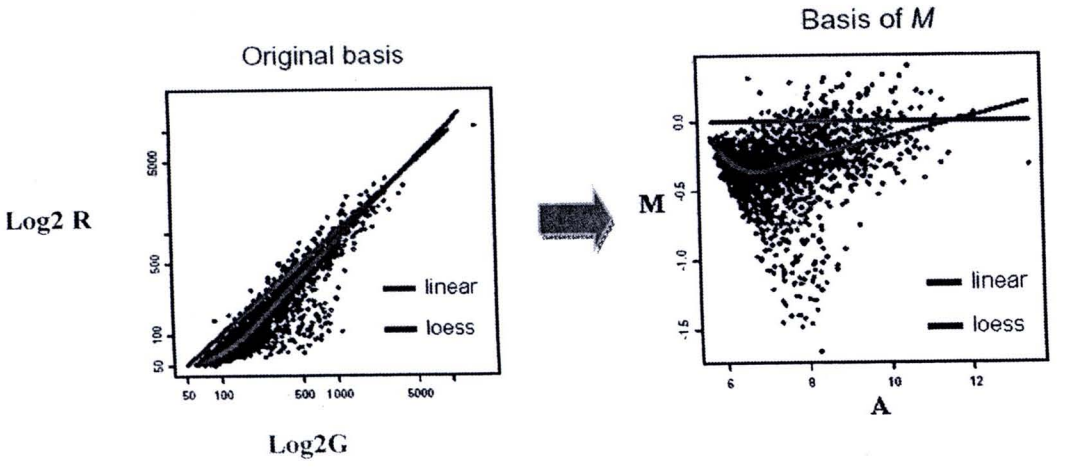


Figure 2.17 The concept of MA-plot (15).

Background correct increases the accuracy of spot intensity. However, after the background correction the microarray data have some effects remain. Thus, the normalization method is necessary before further analysis.

2.4.4.3 Normalization

The variations in the microarray experiment can be separated into two categories; biological and non-biological variation. Basically, the biological variations come from many sources such as amount of RNA in the biopsy, RNA extraction, reverse transcription, labeling, DNA quality, spotting efficient. These variations are managed on pre-processing procedure. The non-biological variations are often called systematic variation and the management of them usually uses the normalization method (14). The normalization is procedure to indentify and remove the systematic variation, perturbation biological signal. The aim of normalization is to ensure that the observed intensity does not include the artifactual biases (systematical variation). In addition, there are two possible assumptions in the

normalization process. First, most of the genes on the array are not differentially regulated under the experiment. In other word, the intensity ratio R/G should be close to one, therefore, the value of intensity log-ratio corresponds to zero, $log_2 \frac{R}{G} \cong log_2 1 \cong 0$. Second, the small proportions of gene differentially expressed could be up-regulated and down-regulated. The normalization is awareness of reflects for further analysis usually mention to analysis differential gene expression (9). For two-color normalization can be separated into two class; within-array and between-array normalization (24).

Within-array normalization

Within-array normalization is a method attempt to adjust the microarray data within-array normalized M-values for each array. The most common normalization for two-color is loess normalization method. Loess normalization on cDNA microarray data is based on the M-A plot. For each spot i on an array represents a value of average log intensity and log ratio respectively (A_i, M_i). Loess is a term of locally weight linear regression, performing the regression technique to estimate a regression curve. The aim of loess normalization is to find the best-fit curve through the data (Figure 2.18). The fit is based on the normalization curve as M-value. M is adjusted by the normalized model (25, 26). Thus, loess model is written as Equation 2.9.

$$M' = M - c(A) \tag{2.9}$$

where M is raw log-ratio intensity, M' is the resulting normalized log-ratio and $c(A)$ is a smoothing loess curve. The $c_i(A)$ value is called the fitted value of curve through points A_i. Loess function was introduced by Cleveland (1979) (27). The function

estimates a locally weighted polynomial regression through a set of points (A_k, M_k) in the neighborhood of every data point (A_i, M_i). The smoothing procedure for each (A_i, M_i) in data set is designed to accommodate data with local model as

$$M = f(A) + \varepsilon \tag{2.10}$$

where f is a smooth function as polynomial regression function:

$$f(A) = M_k = \beta_0 + \beta_1 A_k + \beta_2 A_k^2 \ldots + \beta_d A_k^n \tag{2.11}$$

and ε is residual of loess regression (28). The normalized is computed as the ε by using the weight function that performed by Cleveland (1979) (27). The assumption of smoothness allows point in a neighborhood of data points to be used in forming $c_i(A)$ values. For each A_i, weights, $w_k(A_i)$ is defined for all A_k,$k = 1,2,3, \ldots, n$ using the weight function W. The weight function is minimized by sum of square of the first order polynomial (28):

$$\sum_{k=1}^{n} w_k(A_i) \, [M_k - \beta_0 - \beta_i A_k]^2 \tag{2.12}$$

where i is a data point and k is a neighborhood point. The $w_k(A_i)$ is decreased as the distance of A_k from A_i increased. If the distance of A is more than one thus $w_k(A_i)$ is zero that relies on the forth properties of weight function in Cleveland (1979) (26). The overall regressions are computed as $M = f(A) + \varepsilon$ where $f(A_i) = f_i(A_i)$ (29). The $c(A)$curve is difference that based on the local region. Thus the loess model is $M' = M - c_i(A)$. The best fit-curve $c(A)$ through the data is achieved of the distance of each point converge of central-weighted means that is equal to zero (Figure 2.19).

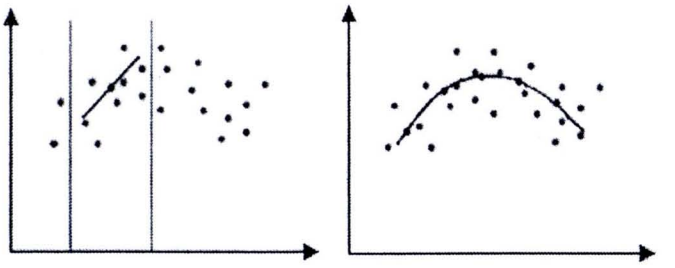


Figure 2.18 Loess regression (Locally weight polynomial regression) (14).

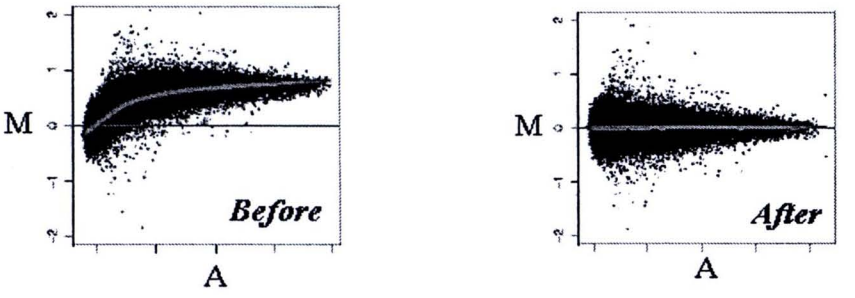


Figure 2.19 The MA-plot of all data before and after normalized by loess method (14).

Between-array Normalization

Between-array normalization method is used to adjust the original intensities in log-scale without using A values in the normalization options. The popular normalization methods of between-array are quantile method and Aquantile method. The quantile normalization method proposed by Bolstand *et al.*, (2003) (30). Quantile normalization assumes that most genes on the array are expressed at relation to mRNA level across arrays and the bulk of genes are constant. Quantile normalization equalizes the intensity distribution of probe for every array. The normalization distribution is chosen by averaging each quantile across arrays. Aquantile

normalization ensures two channels have the same empirical distribution of the A-values across arrays leaving the M-values unchanged (31).

2.4.5 Statistical analysis

The goal of microarray data analysis is to indentify differentially expressed genes under two different conditions. The statistics is essential to determine whether the change in gene expression is experimentally significance. The various statistical methods are contributed to identify different genes. The common hypothesis being tested is that no difference in gene expression between condition; the log-intensity ratio between the expression of each gene in the two samples should be zero. The simplest statistical test method to compare genes expression in two conditions is the t-test. The microarray raw data from the experiment is M-value of each gene i on array j that can be written M_{ij}, where $i = 1,2, \ldots N$ and $j - 1,2, \ldots . n$. Data analysis process assumes that the data have already been normalized by the best method. Thus, most of the M-value should be zero but in fact some genes are not. Let M_g is the M-value of gene that true M-value of gene is different from zero and R_g be the average M_g of one gene. Let SE be the standard error that computed combination data across all gene. Therefore, the standard t test can be conduct for each gene is so-called global t-test (Equation 2.13)

$$t = \frac{R_g}{SE} \tag{2.13}$$

and the gene-specific t-test is shown in Equation 2.14

$$t = \frac{R_g}{SE_g} \tag{2.14}$$

Gene-specific t-test uses the information from one gene at a time (10, 32 and 34). Occasionally, the term of standard error of the ordinary t-test is difficult to estimate and subject to erratic fluctuations when sample is small. Thus the moderatedt-statistic is suggested. The moderated t-test is similar to ordinary t-test except that the standard error is moderated across gene by using Bayesian model or else. Moderated t-statistic is computed for each gene and for each contrast (33-34). Tusher *et al.*, (2001) (32) proposed a refinement of t-statistics which avoid the difficulty just mentioned by adding a constant term suggested by Efron *et al.*, (2000) (35). The t-test of Tusher (32) is well-known as significance analysis of microarray (named SAM).

$$\text{SAM} = \frac{R_g}{c+SE_g} \tag{2.15}$$

where the constant c can be taken to be equal to 90^{th} percentile of SE of all the genes (33). By adding the SE , the standard deviation for gene sample $(M_{ij}), j = 1,2, \ldots n.$ Baldi P, *et al.*, (2001) (36) proposed the regularized t-test by developing a Bayesian probabilistic framework for analysis of microarray expression data as

$$t = \frac{R_g}{\sqrt{\frac{v_0 SE^2 + (n-1)SE_g^2}{v_0+n-2}}} \tag{2.16}$$

where v_0 is a tunable parameter that determine the relative contributions of gene-specific and global variances and n is the number of replicate measurement for each condition. Lönnstedt I, Speed T. (2001) (34) proposed the empirical Bayes log posterior odds ratio of differential expression versus non-differential expression well-known as B-statistics.

After a test is computed, it is converted to a p-value, measuring of evidence against hypothesis in statistic test. The p-value of gene is decided under the condition

of the significance level (α) often defined as $\alpha = 0.005$. Gene with p-value falling below a prescribed significant level may be regarded as significant. The interpretation of the statistical test of each gene as p-value is provided the simple dichotomy of "significant" or "not significant" at the predefined significant level. The p-value of moderated t-test has the same meaning as the ordinary t-test except the increasing of the degrees of freedom that reflecting the greater reliability as smooth standard errors. The p-value and B-test ranks genes according to their significant with staring at the top of list. Occasionally, p-value and B-statistics are calculated under the assumption that M_{ij} of all genes as a random variables from normal distribution. Following this assumption, the p-values of all genes depend on normality. But in fact, there are never exactly true of microarray data, if not regard and use p-value for ranking genes difference usually presented of large deviations of normal distribution also that error should be modified. The statistical method often used to adjust is Benjamini and Hochberg (37) method that first time suggested by Benjamini Y. and Hochberg Y. (1995) (37). Benjamini and Hochberg method controls false discovery rate at a specified level. The p-value is adjusted with $p \leq \frac{i}{m}q$, where i is the number of genes that selected previously, m is the total number of genes tested and q is the desired FDR (35).

Linear Model and Fitted Model

Two-color microarray data prepared for input to linear model analysis are calculated in log-ratio. The intensity of gene g on j array is expected that these have been appropriately normalized. Let define $y_g = (y_{g1}, y_{g2}, \ldots y_{gj})$ as a vector of expression value of gene g for each j arrays. Thus the linear model is defined as

$$E[y_g] = X\alpha_g \tag{2.17}$$

where X is designed matrix with j columns that represents the different samples hybridized to the j array and α_g is a vector of coefficient value for each genes. The α_g is estimated by least-square error under the condition of gene with residual variance σ_g^2 with sample value s_g and residual degree of freedom f_g.

Linear model is fitted to log-ratios derived from experiment so that the coefficient can be estimated by the testing differential expression. According to Dunning MJ (38):

$$\beta_g = C^T\alpha_g \tag{2.18}$$

where C is a contrast matrix for each single-channel, α_g is a coefficient of gene g and β_g is a contrast value of gene interesting. Let μ_{gk} is un-scaled standard deviations for each k th contrast of gene g The t-test are modified by Bayes approach.

The moderated t-statistics for each gene.

$$\tilde{t}_{gk} = \frac{\hat{\beta}_{gk}}{\mu_{gk}\cdot\tilde{s}_g} \tag{2.19}$$

The moderated t-statistics values is associated log-odds statistics computed from posterior log-odds of a given gene with differentially expressed. Thus, the moderated t-test of model fitting is commonly used to rank genes according to the evidence for difference genes (20, 37).

Fold-Change

Fold-change suggested by Tusher *et al.,* (2001) (32) is a common method for indentifying differentially expressed genes that evaluated the expression in log ratio between two difference samples or two conditions. The fold-change value is used to rank the order of gene and the threshold value for cut-off genes which difference expression is chosen, i.e., two-fold difference. The two-fold difference is used for cut-off differentially expressed genes with regarding to average expression under one condition to the other. Those genes with greater than two-fold difference are reported as up or down regulation.

Let x_g and y_g denoted the log_2 expression levels of gene g in two difference samples: usually defined x for the reference or control samples and y for the others such as tumor. Let $\overline{x_g}$ and $\overline{y_g}$ are the mean of log_2 expression of gene g in the control and tumor, respectively.

$$FC = \frac{\overline{x_g}}{\overline{y_g}} \tag{2.10}$$

Typically, identification of differentially expressed genes is the transformation of fold-change value into logarithm fold-change.

$$logFC = log\,\overline{x_g} - log\overline{y_g} \tag{2.11}$$

The logarithm fold-change can be used to determine whether a gene is up-regulated ($logFC > 0$) or down-regulated ($logFC < 0$) between two difference samples compared.