Microarray Gene Expression: Towards Integration and Between-Platform Association of Affymetrix and cDNA arrays

A thesis
submitted in partial fulfilment
of the requirements for the Degree of
Doctor of Philosophy
at
Lincoln University
by
Chintanu Kumar Sarmah

Lincoln University
2010

Abstract of a thesis submitted in partial fulfilment of the
requirements for the Degree of Doctor of Philosophy.

Microarray Gene Expression: Towards Integration and Between-Platform Association of Affymetrix and cDNA arrays

by

Chintanu Kumar Sarmah

Microarrays technology reveals an unprecedented view into the biology of DNA. Information science is moulding this revolution in gene expression profiling with its distinctive skilfulness to transform it into a technologically-advanced and perpetually rejuvenating branch of science while simultaneously contributing to further streamlining the processes involved.

With the advancement of the technology along with the increase of popularity, microarrays afford the luxury that gene expressions can be measured in any of its multiple platforms, which include arrays from commercial vendors like *Affymetrix*$^{\circledR}$ (Santa Clara, CA, USA), *Agilent*$^{\circledR}$ (Palo Alto, CA, USA), and other proprietorial arrays of various laboratories. The technology is expanding rapidly providing an extensive as well as promising source of data for better addressing complex questions involving biological processes. The ever increasing number and publicly available gene expression studies of human and other organisms provide strong motivation to carry out cross-study analyses. Integration of multiple studies that are based on the same technological platform, or, combining data from different array platforms carries the potential towards higher accuracy, consistency and robust information mining. The integrated result often allows constructing a more complete and broader picture.

Various comparison studies have been published over the years, and the overall observation on accuracy, reliability and reproducibility of microarray investigations can be summarized as cautious optimism. In the midst of all the relentless chase in finding suitable remedies for the issues of microarray data integration, this project is an attempt of cross-platform data integration belonging to chilhood leukaemia patients tested on microarray platforms, Affymetrix and cDNA. Keeping in mind the nature of the resultant microarray data from the two platforms, a new ratio-transformation method has been proposed, and is applied to the cancer data. The approach, subsequently, highlights that its usage can address the issue of incomparability of the expression measures of Affymetrix and cDNA platforms. The method is, later, tested against two established approaches, and is found to produce comparative results.

The encouraging cross-platform outcome leads to focus attention on examining further in the direction of defining the association between the two platforms. With this motivation, a wide range of statistical as well as machine learning approaches is applied to the microarray data. Specifically, the modelling of the data is elaborately explored using – regression models (linear, cubic-polynomial, loess, bootstrap aggregating) and artificial neural networks (self-organizing maps and feedforward networks). In the end, the existing relationship between the data from the two platforms is found to be nonlinear, which can be well-delineated by feedforward network with relatively more precision than the rest of the methods tested.

Keywords: microarray technology, gene expression, Affymetrix, cDNA, DNA, cross-platform, data integration, childhood leukamia, cancer, ratio-transformation, machine learning, artificial neural networks, regression, linear, nonlinear, polynomial, loess, bootstrap aggregating, self-organizing maps, feedforward networks.

Acknowledgements

This thesis is a culmination of a period of work with Associate Professor Sandhya Samarasinghe. I owe her a great debt of gratitude for her critiques, suggestions, and guidance. Professor Don Kulasiri and Dr. Daniel Catchpoole assisted during this span of time with advice and suggestions whilst in need. Daniel kindly provided the microarray data, which were used in this research. I thank you all for being there as a robust support system for the entire period.

I appreciate senior lecturer Dr. Wynand Verwoerd for always being forthcoming to me with his ideas and knowledge. A sincere acknowledgement of appreciation to Dr. Takayoshi Ikeda and Adrien Souquet for a few useful discussions. My heartfelt thanks also goes to everyone at Lincoln University library – you are undoubtedly a great team!

On a personal note, I would never have embarked on so ambitious a move without the unshakable support of each member of my family. You are the strength that makes everything possible, and also makes me moving hemispheres on a whimsical decision to follow science. Dedicating this work to my parents, I also extend a special tribute to loving Ma, who had to untimely leave for certain divinely mission somewhere up in the clouds. Ma, if you wish to go for a rebirth in future (refuting atheism-related theories such as the one renowned scientist, Professor Stephen Hawking presents in his recent book, *The Grand Design*), hope you would find a world with lesser irresponsible drivers on the streets. Further, I am unreservedly grateful to my wife, Santana for your love, patience, support, and encouragement.

Last but not least, I warmly thank all others whose names I did not mention, but who contributed in any form towards the successful completion of this work.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

The new millennium is currently witnessing a high-paced information revolution that was initiated in the latter part of the 20th century. This has gifted the common people to realise that the dreams that were seemed distant not a too long ago are indeed possible to see under the broad daylight. Computer technology and internet have catalysed and continually been adding a fuel to this ongoing renaissance. With regards to the promise of our better health through its huge impact on the bioscientific, bioengineering and medical fields, the pair has ushered *Bioinformatics*, 'the combination of biology and information technology, dealing with the computer-based analysis of large biological data sets' (Fogel & Corne, 2003). The applications of bioinformatics in gene expression profiling help disease diagnosis, prognosis, and therapy. Particularly, microarray-based methods are conferring the freedom to conduct large-scale gene expression profiling measurements; and in conjunction with bioinformatics, it has unleashed a wealth of powerful and previously unattainable prognostic information on cell growth and survival. This availability, versatility as well as integration of new technologies have eliminated many previously existing obstacles and boundaries to march towards unravelling the complex mechanisms hidden beneath complex diseases and networks that regulate gene expression.

The methods to measure gene expression were revolutionized by Kary Banks Mullis's invention of the *in vitro* technique, *polymerase chain reaction* (PCR) in 1985 that awarded him Nobel prize for Chemistry in 1993. PCR (Mullis et al., 1986; Saiki et al., 1985) exponentially amplifies and synthesises new DNA molecules via enzymatic replication. While the variants of PCR, such as RT-PCR (reverse transcription polymerase chain reaction) or Q-PCR (real time quantitative polymerase chain reaction, or qrt-PCR) can detect the expression of one gene within one reaction or to a maximum of a few genes in optimised state, high-throughput analysis of higher number of genes is very time consuming, and requires a lot of technical and personal power. In 1995, two seminal publications, namely Schena et al. (1995) and Smith et al. (1995), led by investigator, Patric O. Brown of the *Howard Hughes Medical Institute* and his colleagues, launched the era of low cost gene-expression microarray analysis. From 1995, the technique of microarrays, which started off with simultaneous gene expression analysis of 45 genes within one experiment, has been improved dramatically and has become a widely used tool for studying global gene expression of cells in culture or complex tissues in different organisms. This technology has indeed transformed the classical paradigm of studying '*one gene at a time*', and provided technological and conceptual advancement through its high-throughput capability of simultaneously interrogating the RNA expression of the whole genomes on a single chip. From the late nineties, researchers have started conducting microarray experiments using either of the two distinct techniques - *cDNA-microarrays* and *Oligonucleotide microarrays*. With the development of this field, different labs have begun to routinely fabricate customized arrays.

As gene expression microarrays gradually became widely applied for addressing increasingly complex biological questions, an unprecedented amount of data have started been generated. This catalyzes contributions from various interdisciplinary fields, which constitute integral components of the technology. The knowledge of different fields soon becomes a necessity while studying microarray technology, as depicted in Figure 1.1. It has also liberated the researchers to employ microarray technology in a much wider range of applications, including experimental annotation of the human genome, discovery of gene functions, analysis of complex diseases, biological-pathway dissection, tumour profiling, diagnostic and prognostic predictions for various cancers, drug-target identification and validation, biomarker identification, and compound-toxicity studies (Imbeaud & Auffray, 2005).

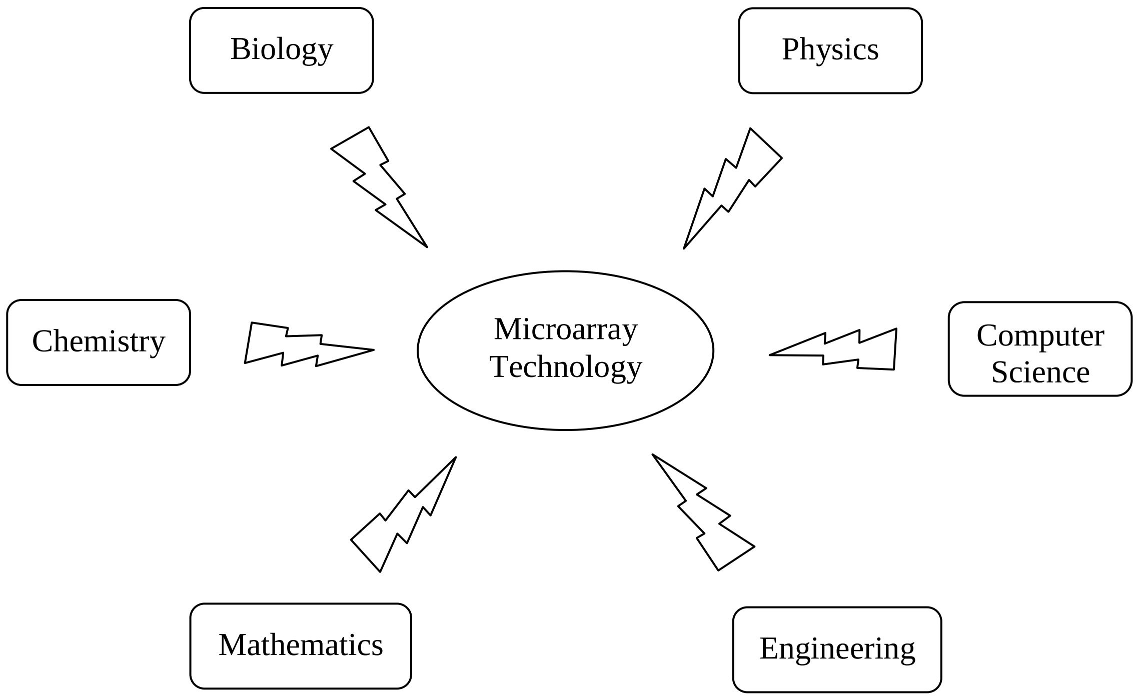


Figure 1.1 Microarray technology requires interdisciplinary knowledge

Over a short time, microarray technology has indeed positioned itself in the scientific world as a reliable approach for gene expression analysis. There are, however, still issues that are not yet unanimously resolved, such as reliability and reproducibility (S Draghici, Khatri, Eklund, & Szallasi, 2006; P. J. Park et al., 2004), experimental design (Yee Hwa Yang & Speed, 2002), statistical issues (Nadon & Shoemaker, 2002; Gordon K. Smyth, Yang, & Speed, 2003), image processing (Jouenne, 2001), and others (Imbeaud & Auffray, 2005; Murphy, 2002; P. J. Park et al., 2004). One such critically unresolved niche of microarray technology lies in the integration of data from different microarray experiments.

The freedom of having multiple platforms to conduct microarray investigations as well as the ever increasing number and publicly available gene expression studies of human and other organisms provide the researchers with strong motivations to carry out cross-study analyses. Integration of multiple microarray experiments carries enormous potential towards obtaining higher accuracy, consistency and robust information mining. Moreover, the integrated results can help in constructing a broader picture crystallizing the biological mechanisms.

The goal intended to be attained in this research work remains within the vicinity of intersection of two specific platforms - *cDNA* (or, *spotted arrays*) and *Affymetrix*®. Firstly, a novel approach is to be designed and implemented that integrates the data from the two platforms. This method is then required to be validated as well as evaluated to examine where it stands in the midst of methods available from microarray literature. Further, investigation needs to be carried out with the merged data towards analysing whether there is any association between the two platforms; and if the answer is positive, then carry out investigations and find out how best this association could be defined.

The overall thesis, including this introductory segment, is comprised of seven chapters. A glimpse of the layout follows.

Chapter 2: Microarray Technology and Cancer

The 2nd chapter provides a broad overview of microarray technology. Starting with an introductory overview, it explains the various microarray types and the process of microarray data analysis along with the challenges and applications of the technology. The chapter also highlights the fact that cancer has become a perfect candidate for evaluation by microarray technology, being the disease both dreadful and challenging because of its polygenic nature. It appraises the application of microarray technology in cancer research. Besides, this chapter provides an overview of this disease in general, and leukaemia in particular as the data used in this project belong to a group of childhood leukemia patients.

Chapter 3: Microarray Data Integration: A Review

Integration of data from different microarray experiments is a challenging problem. This chapter carries out a review on several important experiments conducted and published with regards to microarray data integration.

Chapter 4: Data Assessment and Normalization

Assessing the quality of data is critical prior to carrying out any analytical investigations. This chapter begins with introducing the data, which would be used for carrying out the investigations, and then conducts an elaborate assessment of the quality of these data.

Normalization is a transformation method applied to expression data that appropriately adjusts the individual hybridization intensities so that meaningful biological comparisons can be made. After data quality assessment, the focus remains on the application of normalization on the datasets along with the effects. Finally, the chapter conducts a post-normalization quality check on the data.

Chapter 5: Transformation of Expression Data

Microarray experiments are often conducted using two of the most commonly used platforms - *Affymetrix*® and *spotted arrays*. However, there is always an issue of incomparability between the expression data from these two microarray platforms. This chapter attempts to address this issue by structuring a new approach, which is subsequently validated as well as evaluated.

Chapter 6: Formation of a Crossover

The 6th chapter explores in the direction of seeking an association between the two platforms, Affymetrix and spotted arrays. In this regard, a wide range of statistical and machine learning approaches are applied to the microarray data to probe into this possibility. Finally, the chapter compares all the methods, and highlights the ones that stand out in this investigation.

Chapter 7: Closing Remarks

This chapter presents the concluding segment of the thesis that contains the final remarks on the work including the advantages and limitations, and potential lines of future investigations.

References

The final section furnishes the list of citations used in this research.

Chapter 2
Microarray Technology and Cancer

2.1 Microarray Technology

2.1.1 Microarrays – An Overview

All living organisms contain DNA, a molecule that holds all the information required for development and functioning of any organism. *Deoxyribo nucleic acid*, or DNA encodes for genes, and through the process of gene expression, the information from a gene is used in the synthesis of a functional gene product - either protein or RNA. The process usually starts in the nucleus of a cell when the genetic information of DNA flows to messenger RNA (mRNA) by a process called *transcription*. The mRNA then goes out of the nucleus to the cytoplasm of the cell, and interacts with *ribosome*, a specialized complex. Ribosome decodes the information to *amino acids*, the building blocks of proteins, through another process known as *translation*. A type of RNA called *transfer RNA* (tRNA) assembles the protein, one amino acid at a time. This flow of information from DNA to RNA to proteins is so fundamentally important in molecular biology that it is called the *central dogma*. A portrayal of the process, as given by *US National Library of Medicine*, is in Figure 2.1. In brief, this process of turning the genetic information present in the DNA into proteins is known as *gene expression*.

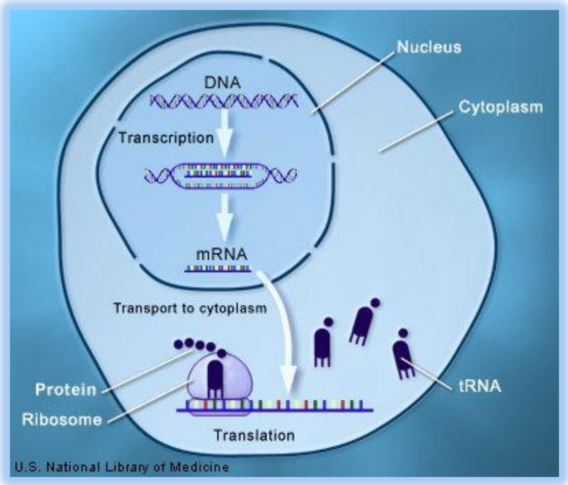


Figure 2.1 Formation of proteins from genes

The human body contains different types of cells, and all the cells contain the same DNA. However, each type of cell expresses a unique configuration of genes. This is assured by the control of the regulatory elements, which switch the genes to either on- or off- state. Microarrays are a tool used to record such states of DNA.

Microarrays provide a way to gain information on the deepest biological mysteries encoded in the informationally complex DNA. Cellular DNA is structurally helical often with two antiparallel strands made up of a combination of four *nucleotides*, or *bases*: adenine, cytosine, guanosine, and thymidine (abbreviated respectively as A, C, G, or T). The nucleotides are covalently linked to a sugar phosphate backbone of each strand. According to a set of pairing rules, the nucleotides of one strand remain hydrogen-bonded with the nucleotides of the other strand. For the cells to express genes, the strands are opened by gene expression machinery so that complementary RNA-copies of a gene can be synthesised. Two complementary single-stranded nucleic acid molecules tend to come together and reanneal to form a double helix complex (Marmur & Doty, 1961). Two single-stranded nucleic acid molecules that are not fully complementary can also bind, but as the complementarity increases, the binding becomes stronger. Overall, this binding process is called *hybridization*. Hybridization is at the centre of many biological as well as *in vitro* analytical processes. Even if molecules come from different sources, they will hybridize if they match.

Hybridization-based approaches have been used for decades to measure nucleic acid sequences (Amos, 2005). Developed at Stanford University, *northern blot* technique (Alwine, Kemp, & Stark, 1977) is the most widely accepted standard for hybridization-based assay of gene expression where the size and abundance of RNA transcribed from a gene is measured. Microarrays are developed from blotting assays, the techniques that are used in molecular biology and clinical research to identify unique nucleic acid (or, protein) sequences in a highly specific and sensitive way (Hayes, Wolf, & Hayes, 1989).

In a microarray framework, there is a *substrate*, or an *array* made of nylon membrane, plastic or glass on which various fragments of *single stranded DNA*, or ssDNA are attached in localised features while arranging in regular grid-like pattern. The substrate is then used to answer a specific query regarding the ssDNA on its surface. The term, *probe* is used to refer the ssDNA. The *target* is a solution of ssDNA that is applied for hybridization with the probes on the substrate. This hybridization between the targets and the probes on the surface of the substrate is essential to conduct the required interrogation. During the hybridization process, the target formes heteroduplexes[1] via base-pairing with the probes. Subsequently, as the

[1] A heteroduplex is a double-stranded molecule of nucleic acid where each complementary strand is derived from different parent molecules.

hybridization completes, the amount of gene expression is computed, probed into and quantified.

2.1.2 Microarray Types

There are mainly two commonly used microarrays that fall into a broader category known as *nucleic acid microarrays*: *cDNA microarrays* and *Oligonucleotide microarrays*. Each effectively serves as a genomic readout while possessing unique characteristics along with advantages as well as disadvantages in a given context.

2.1.2.1 Spotted Microarrays

Spotted, or *cDNA microarrays* were the first available platform that originated in Pat Brown's laboratory, and continue to enjoy broad application. These are primarily a comparative technology where relative concentrations between two samples are examined.

In spotted microarrays, the *probes* are either libraries of PCR (polymerase chain reaction) products that correspond to mRNAs, cDNAs[2] or oligonucleotides[3]. Once synthesised, these are transferred to the substrate, usually glass microscope slide. The probes are printed in an orderly manner at specific locations called *spots* or, *features* using a robot equipped with nibs capable of wicking up DNA from microtiter plates and depositing it onto the glass surface with micron precision (M Schena et al., 1995). Babu (2004) explains it with a schematic, which is given in Figure 2.2.

Samples to be compared are labelled with uniquely coloured fluorescent tags before being mixed together. The fluorescent labelling is done with the fluorophores *Cy3* and *Cy5*, represented by the pseudo-colours green and red respectively, using either of the two common approaches – *direct* or *indirect* labelling. In direct fluorescent labelling, the fluorescent tags are attached in a covalent manner to the target molecules using enzymatic or chemical means, while in indirect labelling, the tags are attached in a non-covalent and indirect way to the target molecules using dendrimers, antibodies or other reagent (Mark Schena, 2003). Some investigators believe that all arrays should be performed both forward- and reverse labelled. That is, for an array with sample A labelled with Cy3 and sample B with Cy5, there should be another array where sample A is labelled with Cy5 and sample B with Cy3. However, Dobbin and his colleagues (Kevin Dobbin, Joanna H. Shih, & Richard Simon, 2003; K. Dobbin, J. H. Shih, & R. Simon, 2003) recommend against this reverse labelling, also known as *dye-swap*.

[2] mRNA is very unstable outside of a cell, and converted in the laboratory to complementary DNA (cDNA), which only contains expressed DNA sequences, or exons. In the process, often incomplete sequences, called expressed sequence tags (ESTs) result from each mRNA molecule due to certain technical aspects.

[3] A short sequence of nucleotides.

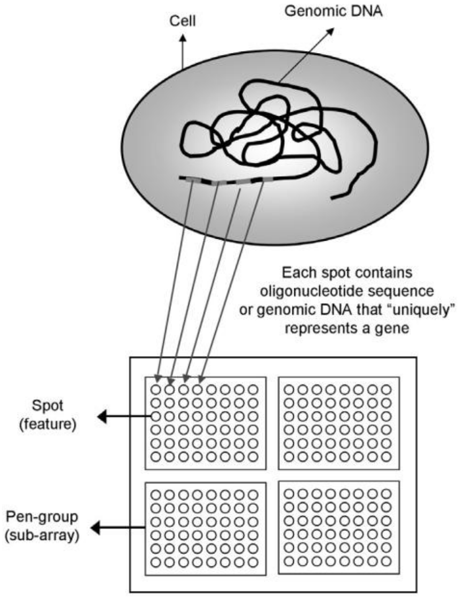

Cell
Genomic DNA
Each spot contains oligonucleotide sequence or genomic DNA that "uniquely" represents a gene
Spot (feature)
Pen-group (sub-array)

Figure 2.2 Spotting in cDNA microarrays

The labelled cDNAs are allowed to hybridize with the probes on the substrate under stringent conditions. Hybridization process continues for several hours, which provides a way of comparing the relative differences between the two samples on a per spot basis depending on the fractional occupancy of the spot hybridized by each sample. At the end of hybridization, excess of the labelled samples is removed by washing, and the slide is dried. Laser scanning is the next and final experimental stage. Here, the slide is excited using a laser at different wavelengths, one for each of the fluorophores used, and the respective fluorescence is captured as two independent, 16-bit, black-and-white TIFF[4] images (Causton, Quackenbush, & Brazma, 2003). The intensity of each spot on these two images is theoretically proportional to the amount of mRNA transcripts of the query (or, test) and control (or, reference) sample. Image recognition software processes the two images, and produces the gene expression levels by converting the gene expression pixel-level intensities into numeric values. An overview[5] of a typical experiment is provided in Figure 2.3.

[4] *Tagged Image File Format* (abbreviated, TIFF) is a file format for storing images.

[5] Modified image. Original source: *University of Wisconsin*, USA (http://tinyurl.com/27gh2ez)

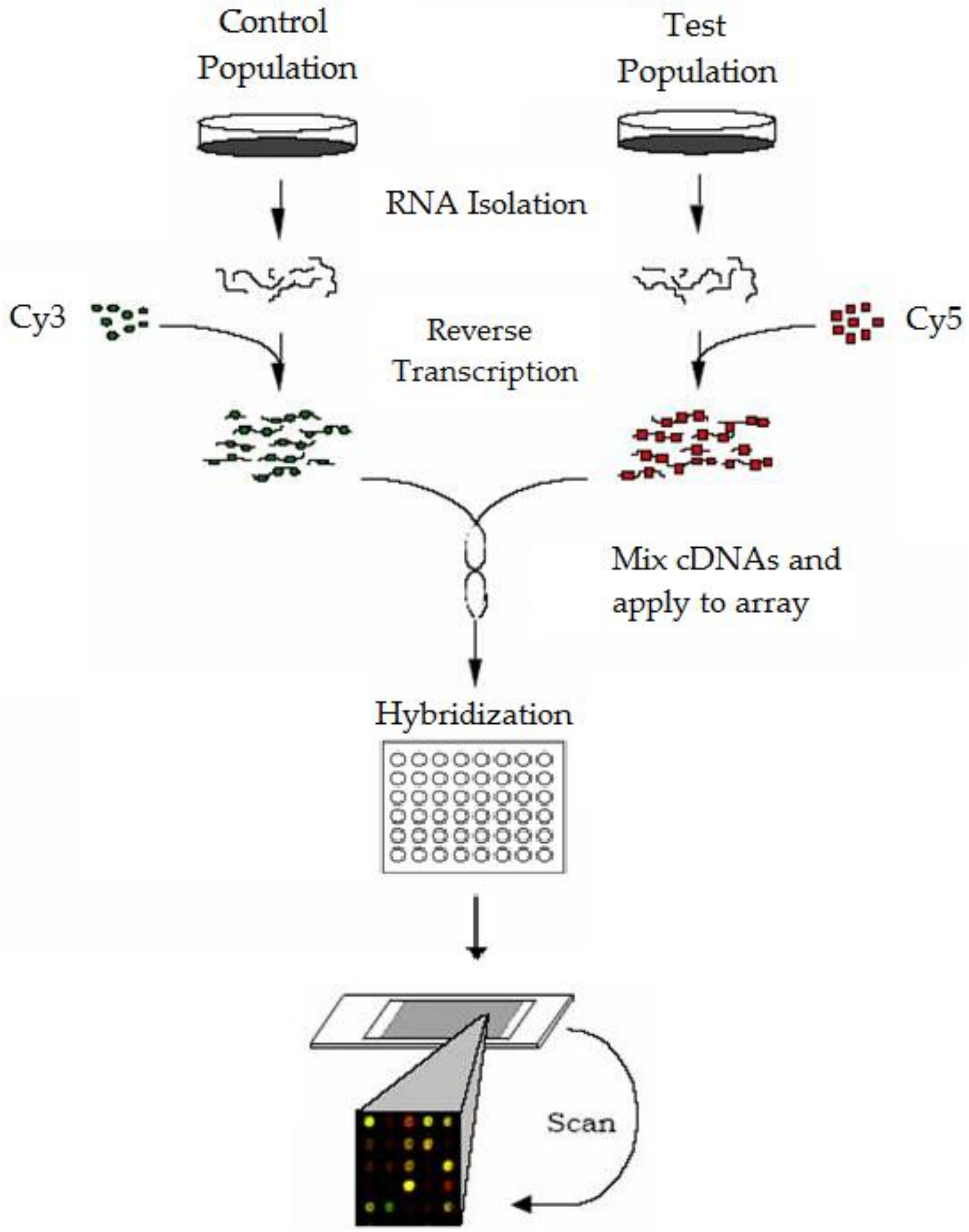


Figure 2.3 Overview of a typical spotted microarray experiment

For the purpose of visually displaying the information, both the images of raw intensities are compressed into 8-bit images, using a *square root transformation*, from which the image processing software creates a composite image (usually 24 bit) that exhibits artificial florescence colours for Cy3- and Cy5- channels ranging from green through yellow to red for the spots (Y. H. Yang, Buckley, & Speed, 2001). Therefore, in the absence of *dye-swap*, the decisions or comments can be made based on the spot-colours: a) Red spot: genes prevalently expressed (*upregulated*) in the tumour sample; b) Green spot: genes prevalently expressed in the normal sample (*downregulated* in tumour); c) Yellow spot: Genes equally expressed in both normal and tumour tissue; d) Black spot: Genes not detected in any of the samples. This is summarised in Table 2.1, and is also shown in Figure 2.4.

Table 2.1 Significance of the spot-colours

Spot color	*Signal strength*	*Gene expression*
Yellow	Healthy = Diseased	Unchanged
Red	Healthy < Diseased	Induced
Green	Healthy > Diseased	Repressed
Black	Unknown/no expression	Unknown/no expression

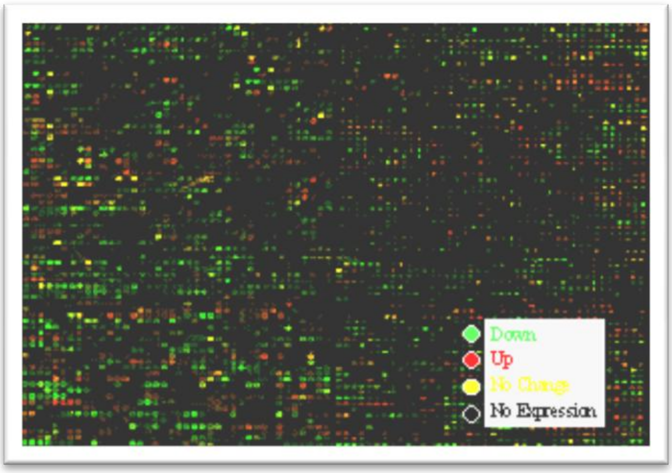


Figure 2.4 Scanned cDNA image

2.1.2.2 Oligonucletide Arrays

Oligonucleotide arrays are fundamentally different from spotted arrays. Unlike cDNA arrays which can use long DNA sequences, oligo arrays can ensure the required precision only for short sequences. Therefore, these arrays represent a gene using several short ssDNA sequences, called oligonucleotides, or oligos. Three approaches represent the in-situ process of microarray fabrication:

- The *photolithographic* approach is based on the same technique as used in the semi-conductor industry to make the microprocessors. *Affymetrix Inc.* (Santa Clara, California) has commercialised the photolithographic method, pioneered by Fodor et al. (1991). Affymetrix refers their technology as *GeneChip*$^{\text{TM}}$ microarrays, where

GeneChips are the probe-holding devices, and are also generally referred to as *biochips*. At Affymetrix, GeneChips are manufactured by a proprietary, light-directed chemical synthesis process, which combines solid-phase chemical synthesis with photolithographic fabrication techniques.

- The *ink jet* approach employs the technology used in the ink jet colour printers. Nucleotides (A, T, G and C) are loaded in the four cartridges. As the print head with the cartridges moves over the array-substrate, specific nucleotides are deposited where required. Several companies such as *Protogene* (Menlo Park, CA) and *Agilent Technologies* (Palo Alto, CA) in collaboration with *Rosetta Inpharmatics* (Kirkland, WA) have developed methods of in situ synthesis of oligonucleotides on glass arrays using ink jet technology.
- The *electrochemical synthesis* approach is introduced by *CombiMatrix Corporation*[6] (Washington, USA). The process uses small electrodes embedded into the substrate. After solutions containing specific bases are washed over the substrate, electrodes are activated on required positions in a predetermined sequence allowing them to be constructed base-by-base.

Here, the focus would remain on Affymetrix *GeneChips,* which are the most ubiquitous and long-standing commercial microarray platform in use (Seidel, 2008).

Affymetrix represents a gene through multiple probe-pairs which are contained in a silicon chip, *GeneChip*. Typically 16–20 of these probe-pairs, each interrogating a different part of the sequence for a gene, make up what is also known as a *probeset*; and some more recent arrays, such as the HG-U133 arrays, use as few as 11 probes in a probeset (B. M. Bolstad, Irizarry, Astrand, & Speed, 2003). The size of a standard GeneChip is 1.28 cm × 1.28 cm; and over 6.5 million squares, or *features* are present on each chip. In each feature, there are millions of identical probes. The design of Affymetrix probes is not usually in the hands of the researchers. A probe consists of a short oligonucleotide sequence containing 25 nucleotides, called a *25-mer*; and all the probes are synthesised on the chip one base at a time, and in parallel at all locations. A paired probe is composed of: a) a perfect match (PM), which is the exact sequence of the chosen fragment of the gene, b) a mismatch (MM), which is same as PM but contains a mismatch nucleotide in the middle of the fragment. Affymetrix anticipates that the MM probe does not hybridize well to the target transcript, but hybridizes to many transcripts to which the PM probe cross-hybridizes (Simon et al., 2004). Therefore,

[6] http://www.combimatrix.com/index.htm

the intensity difference between PM and MM paired probe is considered to be a better estimate of the hybridization intensity to the true target transcript.

A single sample is usually hybridized to GeneChips. For using as target, the total mature, spliced, poly-A tail added RNA isolated from the cell being studied is turned into a double stranded cDNA through *reverse transcription*. At the time of running the array, the cDNA is allowed to go through *in vitro* transcription back to RNA (now known as cRNA), and labelled with *biotin*. The labelled cRNA is then randomly fragmented in to pieces anywhere from 20 to 400 nucleotides in length, and the cRNA fragments are added to GeneChip for hybridization.

The hybridization occurs at a critical temperature. After hybridization, the difference in hybridization signals between PM and MM, as well as their intensity ratios, detected by scanning the array with a laser serves as indicators of specific target abundance. The value that is usually taken as representative for each gene's expression level is the average difference between PM and MM. Ideally, this average value is expected to be positive because the hybridization of the PM is expected to be stronger than the hybridization of the MM. However, many factors, including non-specific hybridizations and a less than optimal choice of the oligonucleotide sequences representative of the gene, might result in an MM hybridization stronger than the PM hybridization for certain probes. The calculated average difference might be negative in such cases, and these negative values introduce noise into the dataset. The overall principle behind Affymetrix technology is summarised in Figure 2.5 (S. Draghici, 2002).

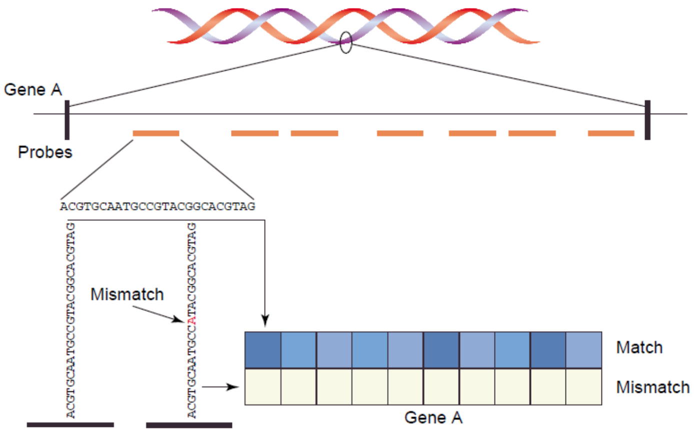

Gene A
Probes
ACGTGCAATGCCGTACGGCACGTAG
ACGTGCAATGCCGTACGGCACGTAG
Mismatch
ACGTGCAATGCCATACGGCACGTAG
Match
Mismatch
Gene A

Figure 2.5 The principle behind Affymetrix technology

The expression data from both types of microarrays are finally obtained in the form of a matrix with genes as rows and conditions as columns, and subsequently biologically meaningful information is extracted and added. Accordingly, Figure 2.6 presents the final fate of a microarray image[7].

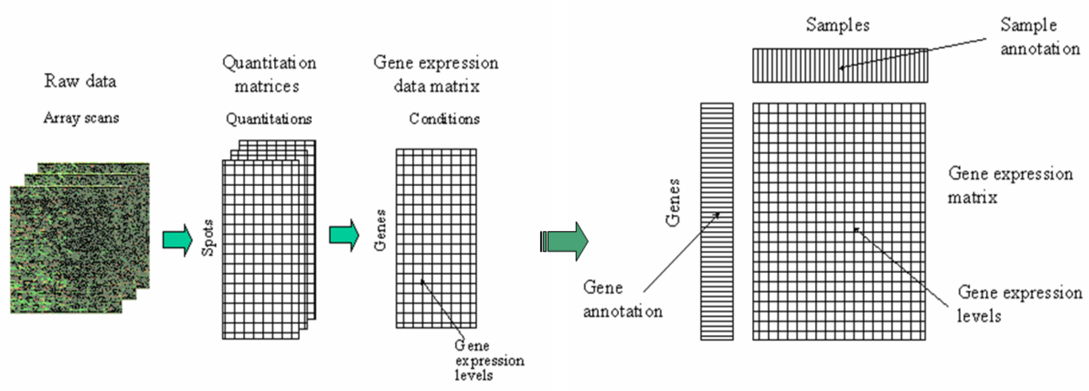


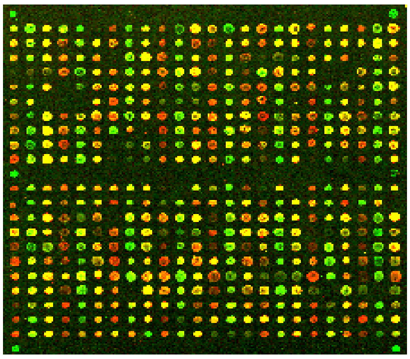
Figure 2.6 A theoretical account of the fate of a microarray image

2.1.3 Processing of Array Output

The outputs of microarray experiments require processing before they can be used for extracting meaningful information. Image processing and normalization are the two preliminary microarray data processing stages.

2.1.3.1 Image Processing

Regardless of the technology, the arrays are scanned after hybridization and independent, 16 bit, digital, grey-scale TIFF images are generated for query and control samples (Causton et al., 2003). Figure 2.7 presents two typical pseudo-coloured images from Affymetrix and cDNA platforms. The process of image processing for the two platforms is different, and is briefly given below.



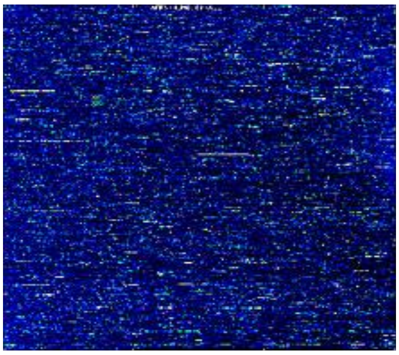


Figure 2.7 Typical cDNA (left) and Affymetrix image (right)

[7] Image Source: *European Bioinformatics Institute* (http://tinyurl.com/5uc5bg)

2.1.3.1.1 Image of cDNA Microarrays

Analysis of a cDNA image seeks to extract intensity for each spot or feature on the array, and it involves various image processing stages that can be carried out through different microarray image analysis software. The analysis is done mainly using the following steps –

A. **Gridding.**

This is usually a semi or fully-automated measure based on Bayesian statistics to locate each spot on the slide. In the process of *gridding*, a grid is placed over the hybrid compound fluorescence in the image so that each fluorescence is contained within a patch. This is shown in the image[8] of Figure 2.8.

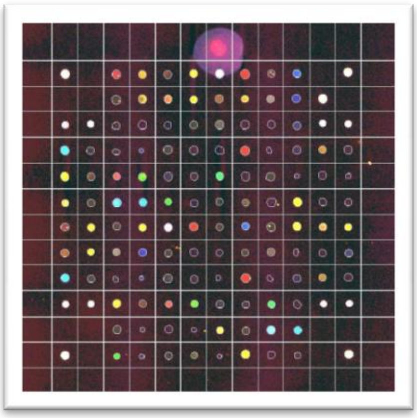


Figure 2.8 Aligning a grid for identification of each spot

B. **Segmentation.**

A microarray spot contains two components – *signal* and *background*. Signal corresponds to the true intensity values of the foreground, and the background, or noise is the unwanted intensity values associated with events like spurious biochemical processes and substrate reflection. It is depicted in the image[9] of Figure 2.9. Once the signals are identified, they need to be separated from the background. *Segmentation* performs the task of partitioning the image into foreground (spot) and background.

[8] Image Source: *The University of British Columbia*, Canada (http://tinyurl.com/2amwsxe)
[9] Image source: *Stanford Microarray Database* (http://smd.stanford.edu/)

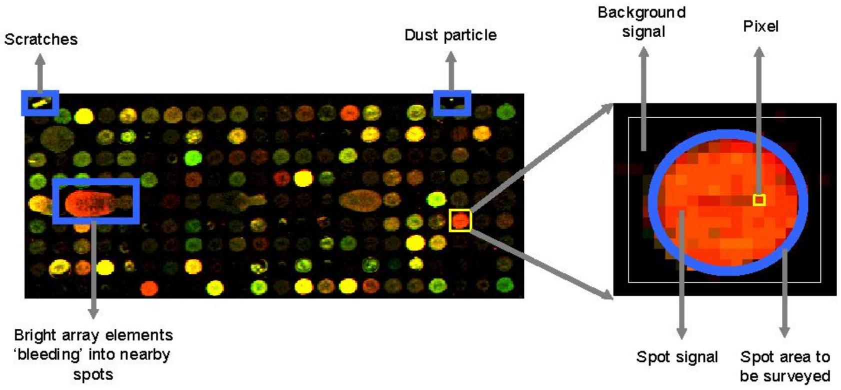


Figure 2.9 A microarray slide and a spot

Several algorithms are in use for segmentation process. Yang et al. (2002) categorises the various existing segmentation schemes into four groups: (1) fixed circle segmentation, (2) adaptive circle segmentation, (3) adaptive shape segmentation, and (4) histogram segmentation.

Fixed Circle Segmentation sets a round region of constant diameter in the middle of each spot as the target site, and is provided in most existing software packages including *ScanAlyze*[10], *GenePix* (Axon Instruments, Redwood City, CA) and *QuantArray* (GSI Lumonics, Inc., Watertown, MA). This is the most straightforward method which assumes that all spots are circular with constant diameter, and everything inside the circle is the signal and everything outside is the background. But this assumption rarely holds, and so most image analysis software includes some more advanced segmentation methods. ***Adaptive circle segmentation***, used by tools like *GenePix* and *Dapple*[11](Buhler, Ideker, & Haynor, 2000), estimates circle diameter separately for each spot. The circular spot signals are quite rare, and therefore, ***adaptive shape segmentation*** tries to find the best shape to describe a spot. ***Histogram method***, used by tools like *ImaGene* (BioDiscovery, Inc., Los Angeles, CA) and *QuantArray*, analyses the signal distribution in and around each spot to determine which pixels belong to the spot and which pixels belong to the background.

C. Foreground Intensity Extraction and Background Correction.

Once the spot and background areas are defined, each pixel within the area is taken into account; and, the mean, median, and total value of the intensity over all the pixels in the defined area are reported for both the spot and background. The signal and background intensity is computed in several different ways, the most common being the mean and the

[10] Available at: http://rana.lbl.gov/EisenSoftware.htm
[11] Available at: http://www.cs.wustl.edu/~jbuhler/dapple/

median. Background subtraction is the process where the intensity corresponding to the background is subtracted from the spot intensity to obtain more accurate quantitation representing a spot.

D. Expression Ratio and its Transformation.

The relative expression level for a gene can be measured as the amount of red or green light emitted after laser excitation. The common measurement used to relate this information is called *Expression Ratio*, T_k, which is denoted by:

$$\mathrm{T_k} = \frac{R_k}{G_k} \qquad (1)$$

where for each gene k on the array, R_k and G_k represent the spot intensity metric for the tumour sample and the healthy sample, respectively. The spot intensity metric for each gene can be represented as a total intensity value or a background subtracted median value.

It is common practice to transform the raw counts into a different scale that is more convenient and statistically sound. There are two kinds of transformation reported for the expression ratio - *inverse transformation* and *logarithmic transformation*. The latter takes the logarithm base 2 value of the expression ratio [i.e., $\log_2$ (expression ratio)]. It is considered a better transformation procedure because it treats differential up-regulation and down-regulation equally, makes the distribution more symmetrical and the variation less dependent on absolute signal magnitude (Babu, 2004; Simon et al., 2004). The $\log_2$–ratio for each spot can be written as given in equation 2, where $R_{Foreground}$ and $G_{Foreground}$ represents the foreground (the patch of a spot) mean or median intensities of red and green channels, and $R_{Background}$ and $G_{Background}$ denotes the corresponding background mean or median intensities.

$$\log_2 \frac{R_{Foreground} - R_{Background}}{G_{Foreground} - G_{Background}} \qquad (2)$$

2.1.3.1.2 Image of Affymetrix GeneChip^{TM}

Affymetrix has integrated its image processing algorithms into the experimental process of *GeneChip* software, and thus, there are no decisions to make for the end users (Stekel, 2006).

Affymetrix GeneChip experiments are managed with the *Affymetrix GeneChip Operating Software* (GCOS) or *Affymetrix Microarray Suite* (MAS). Once the fluorescent-tagged nucleic acid sample is injected into the hybridization chamber, and hybridization takes place to the complementary oglionucleotides on the chip, the hybridized chip is scanned and the laser excited fluorescence across the chip is converted to a 2D image. This image data file (.DAT) can be exported as a .TIFF image. The image data file is used by the software to generate a .CEL file that gives the position and intensity information of each probe for one GeneChip, in addition to the position of masks and outliers.

The Affymetrix output result file is the .CHP file, where the average signal intensities are linked to gene identities. The report file (.RPT) is generated from the .chip file, and it summarizes the quality control information about expression analysis settings and probe set hybridization intensity data. Besides, there are two more files that are used in the actual analysis process - *Experiment File* (.EXP) and *Chip Description file* (.CDF). The former contains parameters of the experiment such as probe array type, experiment name, equipment parameters and sample description. The .CDF file is provided by Affymetrix and describes the layout of the chip. According to the overall Affymetrix file types summarised in Figure 2.10, the .DAT files are analysed and the intensity data, thus generated, are saved as .CEL files. The .TXT file is a .CHP file in text format.

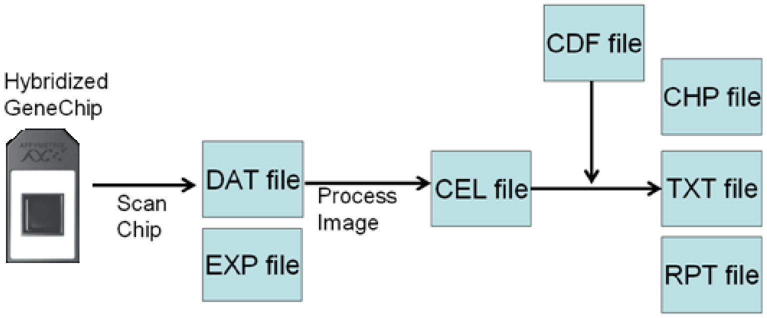


Figure 2.10 Affymetrix data files

A typical Affymetrix probe set contains 11 perfect match probes and 11 mismatch probes. Although Affymetrix has a standard method for summarizing 22 readouts to obtain a single number for gene expression (Affymetrix, 2002), many approaches are available (Rafael A. Irizarry, Wu, & Jaffee, 2006). Usually, the final expression of a gene is the average difference between all the PM and MM probes of a gene, and is considered proportional to the actual expression level of the gene. It is given in equation 3, where n represents the total number of probe pairs for the gene, and PM_i and MM_i indicate the corresponding PM and MM probe intensities after background correction for the i^{th} probe pair of the gene.

$$Difference_{probepair} = PM_i - MM_i$$

$$Average\ difference_{probepair} \cong \frac{1}{n} \sum_i (PM_i - MM_i) \qquad (3)$$

2.1.3.2 Data Normalisation

Data normalisation is an important aspect, and plays an important role in the early stage of microarray data analysis as the subsequent analytical results are very much dependent on it. The normalization methods rely on the fact that gene expression data can follow a normal distribution, and the entire distribution can be transformed about the population mean and median without affecting the standard deviation. The objective of normalization is to eliminate the measurement variations and measurement errors, and to allow appropriate comparison of data obtained from the expression levels of genes so that the genes that are not really differentially expressed have similar values across the arrays. Normalization is also used to identify and eliminate questionable and low quality data.

Normalization approaches typically use either a control set of genes or the entire genes from an array. The use of a control set requires only one assumption, i.e., the control genes are detected at constant levels in all of the samples being compared.

Housekeeping genes constitute a type of control gene set, and are considered to be used in normalization as they are expressed in most, if not all cells. As the cells need these genes for cell maintenance and survival, such genes are expected to be similarly expressed in all samples of experiment. However, it is difficult to identify these genes as the genes regarded to be housekeeping for one tissue type may not be the same for another type of tissue. To ensure that a gene can be considered as a housekeeping gene, carefully controlled experiments are performed. A number of techniques are used to identify housekeeping genes based on the observed data, such as the rank invariant selection method of Schadt et al. (2001), and the iterative method of Wang et al. (2002). For GeneChips, Affymetrix Inc. claims to have integrated the housekeeping genes in the chips after supposedly testing them on a large number of various tissue types with the resultant low variability in those samples.

Spiked-ins, or *spiked controls*, are another set of control genes, which are exogenous RNA added proportionately to both query and reference samples, otherwise not found in either sample. The need of these exogenous control genes arises as there is accumulating evidence to suggest that many housekeeping genes change in expression under some circumstances (P. D. Lee, Sladek, Greenwood, & Hudson, 2002; Thellin et al., 1999).

Besides, there is one more alternative for selecting a dataset for applying normalization. It is to order the genes or signal from each spot based on expression level, and using only those within a fixed window centred within the dataset (e.g., those between the 30th and 70th percentile) or those within a fixed number of standard deviations of the mean (Eric E. Schadt, Cheng Li, Byron Ellis, & Wing H. Wong, 2001; Tseng, Oh, Rohlin, Liao, & Wong, 2001).

Once a gene set for normalization is selected, normalization process can be conducted.

2.1.3.2.1 cDNA Normalization

For cDNA microarrays, normalization involves determining the amount by which the genes of the red channel are over- or under expressed relative to the green channel. This bias is known as *normalization factor* or *scaling factor*, and is different for different arrays. The normalisation factor, C_{jk} is subtracted from the log-ratio of the background-corrected red and green signals as shown in the equation 4 below to find the normalised signal intensity, X_{jk} for a gene, *k* on array, *j*. Here, R_{jk} and G_{jk} represent background-corrected red and green signals, respectively.

$$x_{jk} = \log\left(\frac{R_{jk}}{G_{jk}}\right) - C_{jk} \qquad (4)$$

Approaches to calculate the normalization factor can be divided into three categories: *global normalization*, *intensity-based normalization* and *location-based normalization* as well as a hybrid of intensity- and location-based normalization.

i) ***Global***, or ***Linear Normalization.***

Global normalization applies the same normalization factor to all the genes on the array, but the value varies from array to array. It assumes that the red and green intensities possess an approximately linear relation. Global normalization uses the global median of log intensity ratios as median is less likely to be influenced by the outliers. Moreover, as it is assumed that the over-expressed proportion of the genes in a given sample is approximately equal to the under-expressed proportion, so by using median, focus remains on those genes which are not differentially expressed in the red and green channels and are expected to be at the centre of the log-ratio distribution.

Global normalization is the simplest and widely used normalization method that works well for most applications including in situations where a relatively small number (example, 50-100) of normalization genes are normalized. The expression can be formulated as below, where *S* represents the set of normalization genes. Here instead of median, mean can also be used, but it is to be noted that mean is affected by outliers.

$$C_{jk} \equiv C_j = \underset{k \in S}{median}\,(\log \frac{R_{jk}}{G_{jk}}) \qquad (5)$$

ii) ***Intensity-Based Normalization.***

Intensity-based normalization is described in Yang et al. (2002), and it is necessary that there be normalization genes across all intensity values in order to perform this normalization. Again, even if all genes are being used in normalization, there is the implicit assumption that at each intensity level, there are equal numbers of up- and down-regulated genes. However, it is possible that this assumption could be violated, if all the high- (or low-) intensity genes share similar biology. While using intensity-based normalization at intensities for which there are few spots, the normalization could be based on a rather small number of points that may result overfitting to those particular values.

Dudoit et al. (2002) demonstrates a version of representation of intensity whereby a plot becomes more revealing in terms of identifying spot artefacts and detecting intensity dependent patterns in the log ratios. This representation plots log intensity ratio, $M(=\log_2 \frac{R}{G})$ on the y-axis against the mean log intensity, $A\ (=\log_2 \sqrt{R \times G}\ )$ on the x-axis (R and G represents background adjusted intensity levels for a given spot). This *M vs. A plot* (*MA*, or *RI plot*) shows whether log ratio, M is dependent on the overall spot intensity (which is RNA abundance over all normalization genes), A. In other words, the plot helps to detect intensity dependent patterns in the log-ratios. When it is so found, then it would suggest that an intensity (*A)* dependent normalization method may be preferable than global methods (such as normalization by the mean or median of *M* values).

MA plots are interpreted as follows: The array requires –

- No normalization: The graph-points appears symmetrically scattered around the horizontal line, M=0.
- Global normalization: The graph-points appears symmetrically scattered around a horizontal line, and the line will be shifted up or down, away from M=0 by an amount equal to the required normalization.
- Intensity-based normalization: The graph-points follow a line with non-horizontal slope or a non-linear curve.

Yang et al. (2002) suggest a normalization method for gene expression data that uses smoothing of the MA plot, and this approach is referred to as *intensity-based normalization*. If intensity-based normalization is decided to apply, a curve is fitted to the *MA plot* for the normalization genes. *Loess* curves are more commonly used compared to other smoothing functions. Then, normalization factor, C_{jk} is defined as in equation 6, where f_j is the smoothing function fitted to j^{th} array, and A_{jk} is the average intensity of gene, k on the j^{th} array.

$$C_{jk} = f_j\,(A_{jk}) = f_j\,(\log_2 \sqrt{R_{jk} \times G_{jk}}) \qquad (6)$$

iii) *Location-Based Normalization.*

Many times, due to even subtle differences on the degree of wear of the print-tips used to create a slide, the spots on the array vary. Location-based normalization refers to this aspect which deals with normalizing with respect to the print-tip.

Each print-tip generates a grid that is located at a separate place on the array. Yang et al. (2002) suggest performing normalization separately for each print-tip. For normalization within a grid, the same formula is used (i.e., with median) as mentioned under, *Global Normalization* on page 20.

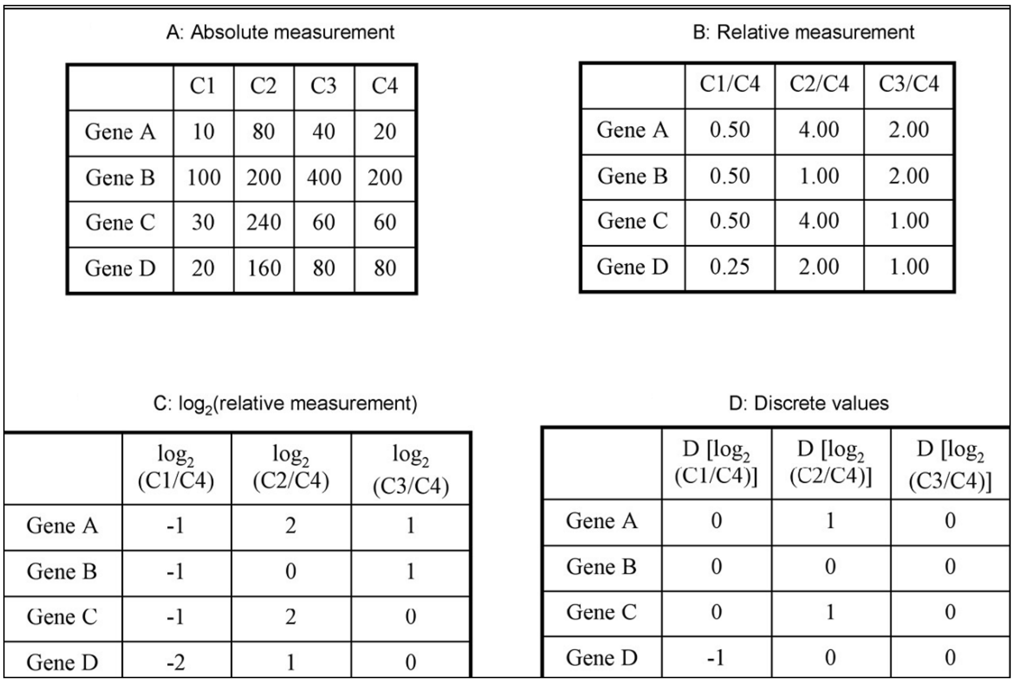
For location-based normalization, there should be significant numbers of normalization genes within each grid as well as on the entire array, and thus, the method is not applicable to a small number of housekeeping or spiked control genes. Moreover, to account for all location effects, estimation methods based on several parameters exist which look beyond the print-tip effect.

iv) *Merging of Location and Intensity Normalization.*

It is possible to combine both location- and intensity-based normalizations for better results. Two possible actions can be taken in this regard. One option is to apply global normalization to each grid of the array and then, to apply intensity-based normalization to the entire array. According to Yang et al. (2002), a better alternative is to use intensity-based normalization separately within each grid. However, it is not suitable at intensities where the data are sparse.

After normalization the processed data can be represented in the form of a matrix, *gene expression matrix*. Babu (2004) shows it figuratively as in Table 2.2, where each row corresponds to a particular gene, and each column either corresponds to an experimental condition or a specific time point at which expression of the genes has been measured. The expression levels of a gene across different experimental conditions are together termed as the *gene expression profile*, while that of all genes under an experimental condition are together termed as the *sample expression profile*.

Table 2.2 Gene expression matrix



A: Absolute measurement

	C1	C2	C3	C4
Gene A	10	80	40	20
Gene B	100	200	400	200
Gene C	30	240	60	60
Gene D	20	160	80	80

B: Relative measurement

	C1/C4	C2/C4	C3/C4
Gene A	0.50	4.00	2.00
Gene B	0.50	1.00	2.00
Gene C	0.50	4.00	1.00
Gene D	0.25	2.00	1.00

C: $\log_2$(relative measurement)

	$\log_2$ (C1/C4)	$\log_2$ (C2/C4)	$\log_2$ (C3/C4)
Gene A	-1	2	1
Gene B	-1	0	1
Gene C	-1	2	0
Gene D	-2	1	0

D: Discrete values

	D [$\log_2$ (C1/C4)]	D [$\log_2$ (C2/C4)]	D [$\log_2$ (C3/C4)]
Gene A	0	1	0
Gene B	0	0	0
Gene C	0	1	0
Gene D	-1	0	0

[A: The value of each matrix-cell, in arbitrary units, reflects the expression level of a gene under a condition. B: Condition C4 is used as a reference and expression ratios are obtained by normalizing all other conditions with respect to C4. C: In this table, all expression ratios were converted into the $\log_2$ values. D: Discrete values for the elements in C are obtained by converting $\log_2$ values > 1 to 1, < –1 to –1, and a value between –1 and 1 to 0. (Babu, 2004)]

2.1.3.2.2 Normalization of Affymetrix Arrays

Affymetrix *GeneChip* arrays have single channel (and colour), and use the same normalization methods for all the arrays, unlike the two-colour cDNA microarrays. Location-based normalization is not used for these arrays as location-specific intensity imbalances, even if they may appear, are less severe having smaller degree of impact on the mean differences of the individual genes. Normalization of Affymetrix arrays is done mainly to account for variations associated with technological reasons. Like cDNA microarrays, normalization factor should be calculated separately for these arrays too.

i) Global or Linear Normalization.

It is a straight-forward method of normalisation, as used in cDNA microarrays, where one normalization factor is used for all the genes on the array. Affymetrix makes use of *average intensity* (different from *cell average intensity*) of an array which is defined as the mean of all the average difference values except the lowest and highest 2% of the data which is not included in the averaging calculation. The idea of this procedure is to find the normalization factor by making the average intensity of the experimental array numerically equivalent to the average intensity of the baseline array[12], as given in equation 7.

$$\begin{aligned} &\Rightarrow C_{jk} = \frac{1}{n} \times \sum_{k=1}^{n} PM_{kj} - MM_{kj} \\ &k \in S,\ S = 96\ percentile \end{aligned} \tag{7}$$

ii) Intensity Based Normalization.

Like cDNA arrays, *MA plots* can also be generated for GeneChip arrays to determine whether intensity-based normalisation is required. In such a plot, a pair of arrays is compared, prior to which a choice needs to be made as to which array to normalize against. If X_k and Y_k denote the normalised signal log value for gene k on two arrays, X and Y, respectively, then M vs. A can be plotted based on equation 8.

$$\begin{aligned} M_k &= X_k - Y_k \\ A_k &= \frac{1}{2}(X_k + Y_k) \end{aligned} \tag{8}$$

[12] Baseline Array**:** An array designated as the baseline when used in comparison analysis with which the experimental array is compared to detect changes in expression.

The result of *MA plots* can be interpreted as follows:

- No normalization: The graph-points appears to be scattered around the horizontal line at M=0. Many times, this is the case as the genes do not vary considerably from array to array.
- Global normalization: The graph-points appear symmetrically scattered around a horizontal line, and the line will be shifted up or down, away from M=0 by an amount equal to the required normalization.
- Intensity-based normalization: The graph-points follow a line with non-horizontal slope or a non-linear curve.

  There is another method of intensity-based normalization as recommended by Simon et al. (2004). Here, a baseline array is chosen whose scaling factor is closest to the median of the scaling factors of the arrays being analysed. Then *MA plots* are generated considering the signal for the array being normalised as the query channel and that for the baseline-array as the reference. If *MA plot* suggests intensity-based normalization, then quantile normalization or loess smoother-based normalisation can be applied using the baseline-array as the reference.

Bolstad et al. (2003), based on a study on the methods of intensity-based normalization of Affymetrix data, recommends quantile normalization method. The method is based on the assumption that the distribution of the expression values does not change dramatically between arrays and that there is a monotone relationship between the gene expression level and probe value within a single array.

Overall, for Affymetrix, there are dozens of methods - as of 2006, more than 30 methods have been identified (Rafael A. Irizarry et al., 2006). Many such methods are popular, namely MAS5 (Affymetrix, 2002), RMA (R. A. Irizarry et al., 2003), GCRMA, dCHIP (C. Li & Wong, 2001), GLA (Zhou & Rocke, 2005); however, no method is clearly the best (Qin et al., 2006).

2.1.4 Applications of Microarrays

The development and use of microarrays are expanding rapidly. It was initially developed for *DNA-mapping* (Carig, Nizetic, Hoheisel, Zehetner, & Lehrach, 1990) and *sequencing-by-hybridization*, or *SBH* (Bains & Smith, 1988; Drmanac, Labat, Brukner, & Crkvenjakov, 1989; Khrapko et al., 1989) applications. Over time, microarray technology has been used in varied applications.

Commonly, microarrays are used in gene expression measurements – ranging from characterizing cells and processes (J. DeRisi et al., 1996; J. L. DeRisi, Iyer, & Brown, 1997; Hughes, Marton et al., 2000) to clinical applications such as tumour classification (Alizadeh et al., 2000; Golub et al., 1999). The technology is also very commonly used in *genotyping* and the measurement of genetic variation (Magi et al., 2007; Winzeler et al., 1998).

Microarray technology can characterize different molecular complexes of DNA or RNA shedding light on their biological mechanisms. For example, *P-bodies* are such identified complexes of protein and RNA, which are believed to take part in gene expression by regulating mRNA in the cytoplasm (Parker & Sheth, 2007), and microarrays could be used to monitor and characterize the trafficking of cellular RNA through this complex.

The position of a gene or a DNA sequence on a chromosome is location-specific, and any change in the positions is implicated in tumorigenesis and cancer. Using comparative genomic hybridization, microarrays have been used to examine this as well as *aneuploidy*[13] in a variety of cell types (Pollack et al., 1999; Shadeo & Lam, 2006). As Khodursky et al. (2000) have examined, microarrays can be used to probe into the progress of replication forks, the structure that forms within the nucleus when two DNA strands start separating into two single-stranded DNA during the process of DNA replication. Microarray technology is also used for genome-wide screening of RNA modifying enzymes (Hiley et al., 2005), and increasing our understanding of *gene regulatory circuitry* (Boyer et al., 2005; T. I. Lee et al., 2002). Hoheisel (2006) and Stears et al. (2003) are two useful reviews that highlights several other useful scientific applications of microarray technology.

There is notable applications of microarray technology in pharmaceutical industry (Crowther, 2002). The technology is intelligently applied in drug discovery (Debouck & Goodfellow, 1999; Sauter, Simon, & Hillan, 2003) on the basis of obtained gene expression information. It gives rise to the production of preventive or curative drugs that impart their therapeutic activity by binding to specific cellular targets, inhibiting protein function and altering the expression of cellular genes. One could also envision an improved and reduced cost of health care, drugs with no or fewer side effects, patient genotyping, personalised medicine, besides efficient treatment and cure of patients of genetic diseases in time to come.

[13] *Aneuploidy* is a type of chromosome abnormality having an abnormal number of chromosomes.

Microarrays can be used for computational purposes as in *DNA computing* (Kari, 1997, 2001; Kari & Landweber, 2000; Tanaka, Kameda, Yamamoto, & Ohuchi, 2005). While being used in this form, microarrays merely turn into simple tools for parallel and efficient manipulation of a large number of symbolic strings to solve computationally intractable problems such as performing efficient searches in large dimensional spaces.

In a nutshell, applications of microarray technology have completely diversified, and penetrated into a long list of varied scientific areas, which also includes domains such as genetic diseases and oncology (Albertson & Pinkel, 2003; Macgregor, 2003; Pusztai, Ayers, Stec, & Hortobagyi, 2003), proteomics (MacBeath, 2002), microbiology (Lucchini, Thompson, & Hinton, 2001), toxicology (Nuwaysir, Bittner, Trent, Barrett, & Afshari, 1999), physiology (Gracey & Cossins, 2003), parasitology (Boothroyd, Blader, Cleary, & Singh, 2003), psychiatry (Bunney et al., 2003), forensic science (L. Li, Li, & Li, 2005), and agriculture and crop science (Galbraith & Edwards, 2010). The full range of applications is too numerous to document, besides there are improvements and adaptations that are continually being made. Nevertheless, the technology in general permits the novice users to adopt it readily, and more experienced users to push the boundaries of discovery.

2.1.5 Challenges in Microarrays

Microarray Technology is relatively new as compared to other molecular biology techniques, and as such it has a number of challenges that its users often come across. A few are given below:

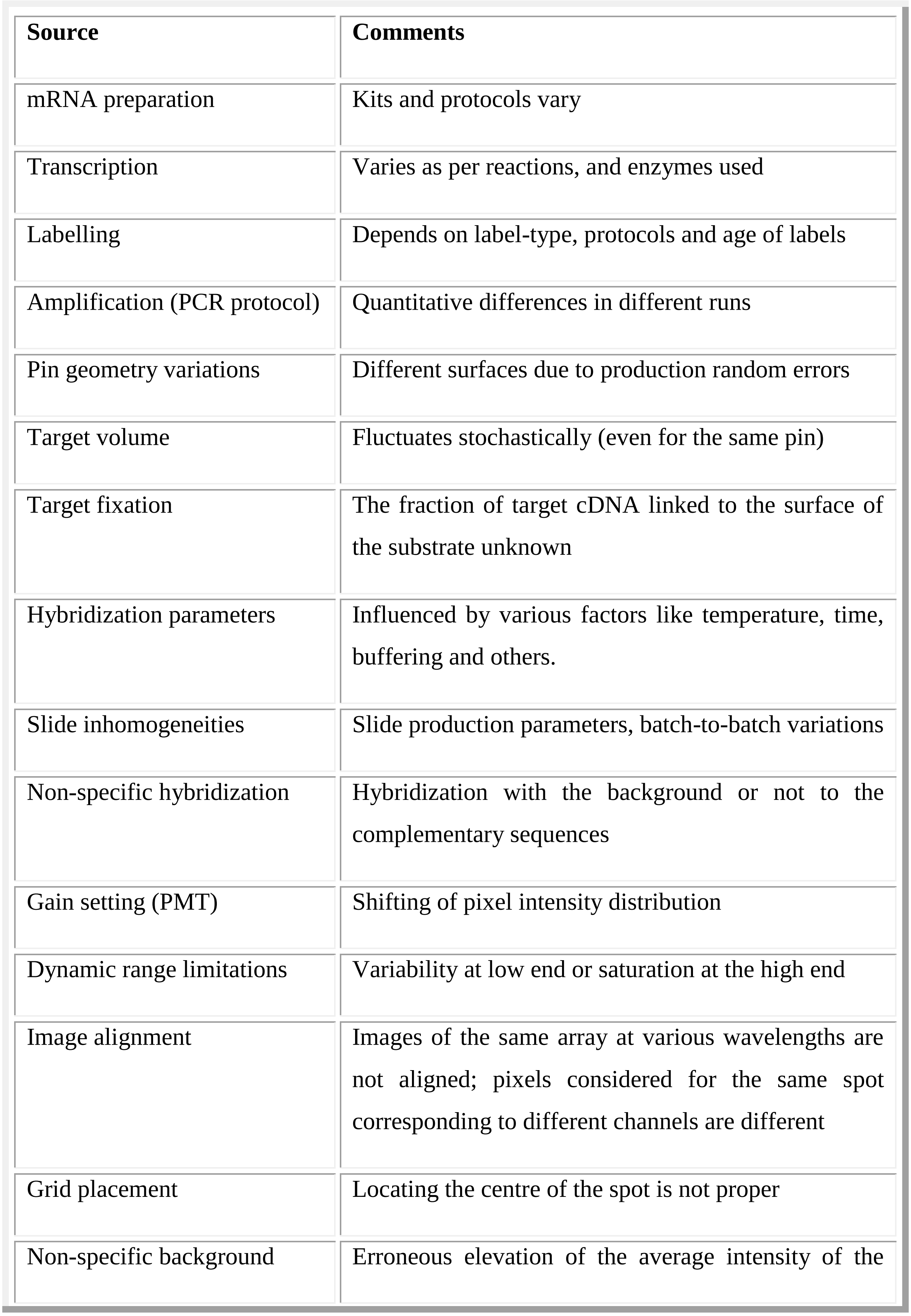
A. Platforms.

There are several microarray platforms. Various laboratories make their own arrays in addition to the popular commercial vendors such as Affymetrix, Agilent, Illumina (San Diego, US). Stears et al. (2003) provide a list of microarray vendors. With the increasing number and accessibility of gene expression studies of various organisms, each platform of this technology serves as a genomic readout along with unique characteristics that offer advantages or disadvantages in a given context.

B. Noise.

Noise is a major challenge in microarray technology. It is very unlikely that two experiments carried out separately but under the same conditions will give the same results. Due to the nature of the technology, noise is an inescapable phenomenon, and can infiltrate at any stage during the process. Draghici (2005) compiles a list of major sources of noise, and it is presented in Table 2.3.

Table 2.3   List of major noise sources



Source	**Comments**
mRNA preparation	Kits and protocols vary
Transcription	Varies as per reactions, and enzymes used
Labelling	Depends on label-type, protocols and age of labels
Amplification (PCR protocol)	Quantitative differences in different runs
Pin geometry variations	Different surfaces due to production random errors
Target volume	Fluctuates stochastically (even for the same pin)
Target fixation	The fraction of target cDNA linked to the surface of the substrate unknown
Hybridization parameters	Influenced by various factors like temperature, time, buffering and others.
Slide inhomogeneities	Slide production parameters, batch-to-batch variations
Non-specific hybridization	Hybridization with the background or not to the complementary sequences
Gain setting (PMT)	Shifting of pixel intensity distribution
Dynamic range limitations	Variability at low end or saturation at the high end
Image alignment	Images of the same array at various wavelengths are not aligned; pixels considered for the same spot corresponding to different channels are different
Grid placement	Locating the centre of the spot is not proper
Non-specific background	Erroneous elevation of the average intensity of the

	background
Spot shape	Hard to segment irregular spots from background.
Segmentation	Contaminants may seem like true signal
Spot quantification	Pixel mean, median.

C. Blind trust can be treacherous.

Result of microarray experiments cannot be trusted entirely. This is a technology, which works at the mRNA level in most cases, and thus, remains distanced from many underlying mechanisms. For example, in most cases, the microarrays measure the amount of mRNA specific to a particular gene as it is based on the premise that the expression level of the gene is directly proportional to its amount of mRNA. However, it is not always true that the amount of mRNA accurately reflects the amount of protein. And, even if it is assumed that it does, a protein may require post-translational modification(s) to become active and perform its role in a cell. Therefore, validation of microarray results through investigation using other techniques and perspectives is an important aspect.

D. Sheer Number of Genes.

Microarrays interrogate thousands of genes in parallel. The classical metaphor, *needle in a haystack* is an accurate description of the task, which brings error in statistical inferences when the number of variables, usually genes, is much greater than that of the experiments. There are several statistical techniques that have been trialled and tested; however, this problem, termed the issue of multiple comparisons, remains to be one of the most challenging topics in *life sciences*.

E. Analytical Methodology.

There is no consensus regarding the standard process of analytical methodology. In conducting microarray analysis, there can be a large number of possible combinations involving background correction methods, summarization methods, normalization methods, and comparison strategy (e.g., ANOVA, SAM, *t*-test). All these contribute to variation in the process.

F. Gene Nomenclature.

Besides the numerous combinations for microarray analytical methodology, results of any microarray study can be reported in different gene nomenclatures, such as that of *Genbank* (Benson et al., 1999), *Entrez Gene* (Maglott, Ostell, Pruitt, & Tatusova, 2006), *The European Molecular Biology Laboratory* or, *EMBL* (Stoesser, Tuli, Lopez, & Sterk, 1999), *Unigene* (Pontius, Wagner, & Schuler, 2003), *RefSeq* (Pruitt, Tatusova, & Maglott, 2006), *Online Mendelian Inheritance in Man* or, *OMIM* (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005) and Affymetrix gene identifiers. Although translation tools such as *DAVID* (Dennis et al., 2003), *GoMiner* (Zeeberg et al., 2003), *RESOURCERER* (Tsai et al., 2001), *L2L* (Newman & Weiner, 2005), *List of Lists-Annotated*, or *LOLA* (Cahan et al., 2005) are available, this disparity may act as associative impediment in microarray technology in general.

G. Varied Repositories.

Various repositories have been established in the name of sharing microarray data. Many journals also require that data be made public in order to be published. Most of the repositories focus on either a particular technology or, an organism or, both, and these are either commercial or non-commercial. Examples of commercial databases include *Merck & Co. Inc.* (http://www.merck.com/)-subsidiary *Rosetta Inpharmatics* and *Gene Logic* (http://www.genelogic.com/). A few non-commercial databases of primary importance include *ArrayExpress* (A. Brazma et al., 2003), *Gene Expression Omnibus* or GEO (Edgar, Domrachev, & Lash, 2002), *Center for Information Biology gene EXpression database* or *CIBEX* (Ikeo, Ishi-i, Tamura, Gojobori, & Tateno, 2003), *ExpressDB* (Aach, Rindone, & Church, 2000) and *GeneX* (Mangalam et al., 2001). A list of microarray databases is also given by Gardiner-Garden & Littlejohn (2001). The overall reliability of data quality, however, in these repositories is not secured – some repositories are undoubtedly of high quality, but it is doubtful whether the same applies to all available repositories. Therefore, it is necessary that one verifies data quality beforehand, if the data come from public repositories, so that the output of the analysis gives accurate as well as meaningful results.

H. Use of Different Splice-Variants as Probes.

Various tools help in the translation between different nomenclatures. However, different microarray platforms use different splice forms of the transcripts. Ideally, we must know all the relevant splice forms of transcripts along with quantification of the sensitivity and specificity of the probes to different splice variants for effective nomenclature translation between different platforms. In practice, it is so far not completely achieved making these irresolvable disparities inescapable; and the effect would pass on to the subsequent results. The *MicroArray Quality Control* (MAQC) project (L. Shi et al., 2006) also shows that it could only cross-reference 12,091 transcripts between all of the major platforms, although some array platforms interrogate over 54,000 transcripts.

I. Software Tools

There are several software tools available for microarray data analysis. Information on various software tools is available from sources like *SMD*[14], Mark Fontenot's *Microarray Software List*[15] at *Southern Methodist University*, Texas, and Dresen et al. (2003). Also, Dudoit et al. (2003) reviews three of the most widely used and comprehensive open source systems - the statistical analysis tools written in R (Ihaka & Gentleman, 1996) through the Bioconductor project (R. C. Gentleman et al., 2004), the TM4 software system (Saeed et al., 2003) available from *The Institute for Genomic Research* (TIGR; Rockville, MD, USA), and the *BioArray Software Environment* (Saal et al., 2002) developed at Lund University, Sweden (http://base.thep.lu.se). Overall, there are several options available for a user to investigate microarray data. The downside of it, however, is that different software packages or tools may generate different results for essentially the same analysis.

2.2 Cancer

Microarray technology has delivered a compelling approach that allows for simultaneous investigations of all cellular processes at once. Being both dreadful and challenging due to its polygenic nature, cancer becomes a perfect candidate for evaluation by this process.

Cancer, or malignant neoplasm, is a disease of cells, and is used to describe about 200 different diseases affecting organs or systems throughout our bodies. These malignant tumours have two features: they can spread into or infiltrate nearby organs or tissues; and cancer cells can break off the original tumour, and be carried in the bloodstream or lymphatic system (which normally fights infection) to distant sites in the body where they may form new tumours called *metastases* or *secondaries*. As cancer cells can spread to the vital organs and affect their normal function, cancer anywhere in the body is a potentially life-threatening disease. According to WHO (*World Health Organization*), cancer accounts for 7.6 million (or 13%) of all deaths from a total of 58 million deaths worldwide in 2005; and the main types of

[14] SMD or *Stanford Micoarray Database* (http://tinyurl.com/24skjy2)

[15] http://tinyurl.com/3xk83jn

cancer leading to overall cancer-mortality (deaths per year) are: lung (1.3 million), stomach (1 million), liver (662,000), Colon (655,000), breast (502,000).

In particular to this project, the data has been obtained from childhood leukaemia patients. Childhood leukaemia is a type of cancer; however, to understand leukaemia, it helps to know about our blood.

2.2.1 Blood, the Life-sustaining Fluid

Blood is a specialized form of liquid connective tissue that performs various important functions within the body including transportation of oxygen, nutrients and hormones and removal of waste products, regulation of body pH[16] and core body temperature. It is a life-sustaining fluid for humans, and are composed of cells and cell fragments suspended within a liquid, called *blood plasma*. The cells and the cells types are of seven types:

- *Erythrocytes*, or *red blood cells* (RBCs)
- *Thrombocytes*, or *platelets*
- 5-kinds of *leukocytes*, or *white blood cells* (WBCs)
    - Three kinds of *granulocytes* (granulocytes are a category of white blood cells characterised by the presence of granules in their cytoplasm.)
        - *Neutrophils*
        - *Eosinophils*
        - *Basophils*
    - Two kinds of leukocytes without granules in their cytoplasm
        - *Lymphocytes*
        - *Monocytes*

The production of blood cells is called *hemopoiesis*. Prior to birth, hemopoiesis occurs primarily in the liver and spleen, but some cells develop in the thymus, lymph nodes, and red bone marrow. After birth, most production is limited to red bone marrow in specific regions, but some white blood cells are produced in lymphoid tissue.

Bone marrow is the soft material in the center of most bones, and it is where new blood cells are made. Active bone marrow is found in almost all bones of the infants; however, by the

[16] pH is a measure of the acidity or basicity of a solution, and is defined as the cologarithm of the activity of dissolved hydrogen ions (H^+).

teenage years, it is found mainly in the flat bones (skull, shoulder blades, ribs, and pelvis) and vertebrae (the bones that make up the spine). The bone marrow is made up of a small number of blood stem cells, more mature blood forming cells, fat cells, and supporting cells that help cells grow. Blood stem cells, also known as *pleuripotential cells* or, *hemocytoblasts*, go through a series of changes to make new blood cells. While a stem cell divides, one of the daughter-cells remains as a stem cell, and the rest becomes a precursor cell, either a *myeloid cell* or a *lymphoid cell*. The myeloid and lymphoid cells continue to mature into various blood cells. The picture[17] in Figure 2.11 depicts the process of maturing a stem cell into either a myeloid cell or a lymphoid cell, where:

- The myeloid stem cell matures into a myeloid blast. The blast can form a red blood cell, platelets, or one of several types of white blood cells.
- A lymphoid stem cell matures into a lymphoid blast. The lymphocytes develop from these lymphoid blasts to become mature and infection fighting cells. There are mainly two main types of lymphocytes: B-lymphocytes (B-cells) and T-lymphocytes (T-cells). Although both B-cells and T-cells can develop into leukaemia, B-cell leukaemia is more common than T-cell leukaemia.

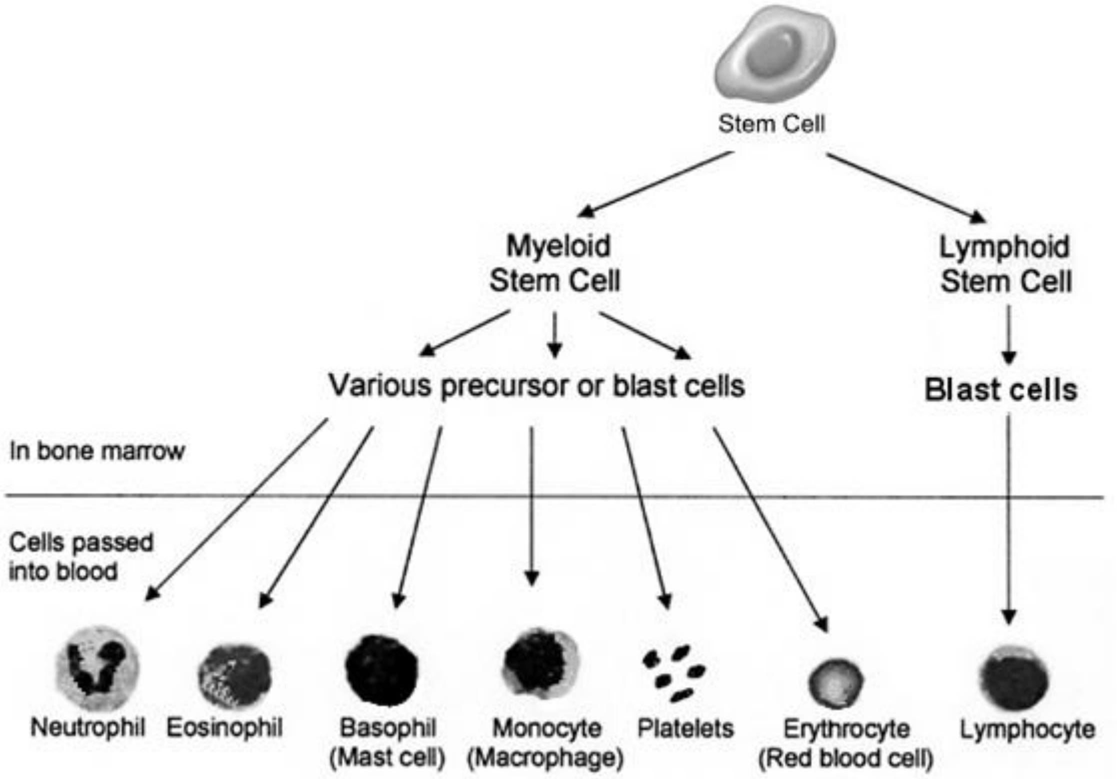


Figure 2.11 Stem cell maturing into different blood types

[17] A modified image. Original source: *The National Health Service*, UK (http://tinyurl.com/2c3xqr6)

2.2.2 Leukemia – the Cancer of Blood

The word, *Leukemia* or, *Leukaemia* comes from the Greek word *leukos* which means ‘white’ and *aima* which means ‘blood’. Leukemia is a part of the broad group of diseases, called *hematological neoplasms*, and can develop at any point in cell differentiation. The disease represents a number of cancers in the blood cells, usually white blood cells (WBC), and starts in the *bone marrow*, the soft tissue inside most bones where the blood cells are made. As mentioned above, bone marrow of a healthy person makes:

- White blood cells, which mainly help the body fight infection.
- Red blood cells, which carry oxygen to all parts of the body.
- Platelets, which help blood clot.

Leukaemia makes the bone marrow to produce a large number of abnormal white blood cells, called leukaemia cells. As these WBCs multiply in an uncontrolled and abnormal way, it leaves little room in the *bone marrow* for the other types of blood cells and for the new blood cells to be produced while making it hard for the normal blood cells to do their work. This process leads to a shortage of red blood cells (RBC) causing severe bleeding (as regular blood-clotting doesn't occur) or serious infection. This can lead to serious problems such as anemia, poor blood clotting, infections; in addition to various other health issues including nausea, fever, chills, night sweats, flu-like symptoms, headache, tiredness and weight-loss. Leukemia cells can also spread to other organs (metastasize) where they can keep other cells in the body from functioning normally and causing swelling or pain. Both children and adults can develop leukemia, and currently, there is no real means of prevention of the disease. Researchers believe that the following are a few likely causes of leukaemia - radiation exposure, viruses (HTLV-1 and HIV), certain chemicals like benzene and alkylating chemotherapy agents used in previous cancers, use of tobacco, genetic predisposition, maternal-fetal transmission.

Leukaemias are subdivided in two ways – one is based on the rate of progression of the disease, and the other is the type of affected blood cell. The former classification gives two types, *acute* and *chronic*. Acute leukemia crowds out the healthy blood cells more quickly than chronic leukemia; and hence, it is a rapidly progressing disease. The classification based on affected blood cell subdivides leukaemia into either *lymphocytic (*or, *lymphoblastic)*, or *myelogenous* (or, *myelocytic* or, *non-lymphocytic*) *leukemia*. As given in Figure 2.11 (on page 33), if the cancerous transformation occurs in the type of marrow that makes lymphocytes, the disease is called *lymphocytic leukemia*. Again, the disease is called *myelogenous leukemia* if the cancerous change occurs in the type of marrow cells that go on to produce red blood cells, other types of white cells and platelets. Thus, combining the two groups from both type of classification, a total of four types of leukaemia are present as shown in Figure 2.12. Both ALL (*Acute Lymphocytic Leukaemia*) and AML (*Acute Myelogenous Leukaemia*) can further be divided into different subtypes.

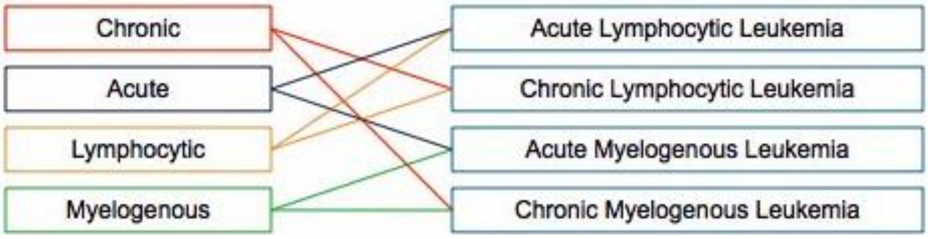

Chronic
Acute
Lymphocytic
Myelogenous
Acute Lymphocytic Leukemia
Chronic Lymphocytic Leukemia
Acute Myelogenous Leukemia
Chronic Myelogenous Leukemia

Figure 2.12 Types of leukaemia

2.2.2.1 Leukaemia in Children

Leukaemia is the most common cancer in children and adolescents; overall, however, it is a rare disease. Childhood leukaemia accounts for 1 out of 3 cancers in children[18]. Any of the blood forming or lymphoid cells from the bone marrow can turn into a leukaemia cell. In children, acute leukaemia is much more common while chronic leukaemia is common in adults. Besides ALL and AML, *Juvenile Myelomonocytic Leukaemia* (JMML) is a rare type of leukaemia that occurs most often in young children under the age of 4 years. This cancer begins from myeloid cells, and its progression is unlike the conventional pace of either acute or chronic leukaemia.

According to *National Cancer Institute*, USA[19], ALL (Acute Lymphocytic Leukaemia) is common in early childhood, between 2 and 4 years of age, and is slightly more common in boys of European descent. AML occurs equally among boys and girls of all races; and the cases are more spread out across the childhood years, although it is slightly more common during the first 2 years in infants and during the teenage years.

[18] *American Cancer Society* (http://www.cancer.org)
[19] *National Cancer Institute*, USA (http://www.cancer.gov)

2.3 Microarrays in Cancer Research

Rapid advancements in Biotechnology and completion of *Human Genome Project* (Bentley, 2000; Venter et al., 2001) has gifted the new technology – DNA microarray technology, which has presented us with a compelling approach that allows for simultaneous evaluation of all cellular processes at once; and cancer, being one of the most challenging diseases, presents itself as a perfect candidate for evaluation by this approach. The ultimate goal of cancer research is to improve the diagnosis as well as treatment of cancer through accurate disease classification and patient stratification, which allows for the design of therapies that are more targeted to specific cancer subtypes and potentially improves the effectiveness of existing regimens based on therapeutic response and adverse events.

Until this century, the study of cancer and its clinical behaviour has been on the basis of histopathologic examination using microscopy. The process often cannot reflect the complexity of causation or production of tumours (*oncogenesis*) because of its major limitation that it can only predict the general categories of cancer, and is unable to achieve high sensitivity and specificity of prediction in clinical practice (Liotta & Petricoin, 2000). As histologically similar cancer patients may have a different clinical outcome, there was a persistent need to find new tools complementary to the conventional histopathologic evaluation for increasing the sensitivity and specificity of cancer diagnosis and prognosis. Microarray technology provides a fitting response to this need. Besides being able to analyse expression of thousands of genes together, the investigators are now able to relate gene expression patterns to clinical phenotypes. The technology offers significant potential to identify molecular signatures capable of differentiating cancer from normal tissues, predicting and prognosis, detecting recurrence and monitoring response to cancer treatment, besides improving our understanding of causes and progression of cancer for the discovery of new drug targets.

In cancer biology, microarrays are used for several applications. The remainder of this section provides a glimpse of some of the applications.

Microarrays have been used for tumour classification, which may have therapeutic implementations. Golub et al. (1999), being among the first to demonstrate the use of gene expression profiling for cancer diagnosis, were able to identify two genetic profiles that distinguished, otherwise histologically similar, acute myeloid (AML) and acute lymphoblastic (ALL) leukaemia. Until then, the two types of blood cancers were diagnosed based solely on histopathology, immunotyping and cytogenetic analysis – that were not completely error free. Following this work, several groups have used DNA microarrays for classifying tumours.

Various microarray methods have been used effectively as tools for identifying the downstream targets and functions of tumour-suppressor genes. Microarray-based expression profiling can be used in identifying target genes for several gene products that directly or indirectly regulate transcription. There are reports of identifying targets of tumor suppressor genes such as *p53* (Fortin et al., 2001; Mori et al., 2002; Zhao et al., 2000), *BRCA1* (Harkin et al., 1999; MacLachlan et al., 2000), *β*-catenin and Plakoglobin (Shtutman, Zhurinsky, Oren, Levina, & Ben-Ze'ev, 2002), *Myc* (Coller et al., 2000) etc.

There is a substantial interest in understanding the association between disease and mutation, including single-nucleotide polymorphism (SNP)[20]. For mutation detection, there are several conventional methods, like *Chemical Mismatch Cleavage* or CMC (Cotton, Rodrigues, & Campbell, 1988), *Denaturing Gradient Gel Electrophoresis* or DGGE (Myers, Maniatis, & Lerman, 1987), and *Single-Strand Conformational Polymorphism* or SSCP (Orita, Suzuki, Sekiya, & Hayashi, 1989). However, such methods have several disadvantages including their time-consuming procedure, less cost-effectiveness. Microarray based approaches have reportedly been carried out for mutation studies (Favis & Barany, 2000; Hacia, Brody, Chee, Fodor, & Collins, 1996; Wen et al., 2000); and the mutation detection is found to be fast with higher accuracy and sensitivity compared to the conventional methods.

Metastasis, spread of cancer from one organ or tissue to another, is another area where microarray-based expression profiling has been used. A few such examples are – studying metastasis in osteosarcoma (Khanna et al., 2001), colorectal tumor (Yanagawa et al., 2001) and brain metastasis (Nishizuka et al., 2002).

Use of microarrays is expected to yield insights into the mechanisms of drug resistance and suggesting alternative treatment methods. Cancers either remain resistant to chemotherapy or after responding initially to chemotherapy, recur later becoming a multi-drug resistant tumour. This stubborn drug-resistance is a significant obstacle to treating cancer patients using chemotherapy. Several groups, for example Kudoh et al. (2000) and Sakamoto et al. (2001), have demonstrated the feasibility of applying microarrays in identifying this resistance mechanism of cancer cells.

[20] SNP is a DNA sequence variation when a single nucleotide (out of A, T, C, and G) in the genome differs between members of a species or, between paired chromosomes in an individual.

In contrast to the conventional methods, microarray-based methods in drug-discovery process have tremendous potential. It would simplify as well as hasten the entire, currently lengthy and complicated, process of drug discovery. In their review, Debouck & Goodfellow (1999) discussed certain ways in which microarrays would likely to affect the process of drug discovery. Further, certain other approaches have also been made towards using microarray-based methods in the process of drug-discovery (Hughes, Roberts et al., 2000; Ross et al., 2000; Scherf et al., 2000).

Overall, microarray-based gene expression profiling is unearthing the concealed information in cancer biology. This would hopefully lead further to provide better and refined diagnostic methods and therapeutic strategies.

Chapter 3

Microarray Data Integration: A Review

With the increase of the collection of microarray data, especially in MIAME (Alvis Brazma et al., 2001)-compliance public repositories such as *ArrayExpress*[21] (A. Brazma et al., 2003), *Gene Expression Omnibus*[22] or *GEO* (Edgar et al., 2002), *Center for Information Biology gene EXpression database*[23] or *CIBEX* (Ikeo et al., 2003), a growing number of investigators are looking at meaningful extraction of information by integration of various microarray experiments. As microarray studies tend to explore specific areas of biological function, integration of data from multiple microarray experiments is considered to allow construction of a more complete as well as a broader biological picture. Integrated microarray data is potentially beneficial in several other ways including that it can compensate for the possible errors of individual experiments, amplify the sample-size, and may lead to higher accuracy, consistency and robust information mining.

Integration of microarray investigations can include integration of studies that are based on the same technological platform. Researchers around the world also combine data from different array platforms based on their needs. However, integration of data from different microarray studies still remains a challenging problem as microarray datasets do not become readily comparable due to factors that can be attributed to biological and technical causes associated with the generation of these data (R. A. Irizarry et al., 2005; W. P. Kuo, Jenssen, Butte, Ohno-Machado, & Kohane, 2002). Nevertheless, with the accumulating amount of important microarray data generated from various microarray experiments, many investigators have taken up the challenging task of meaningful integration of microarray data as well as overcoming the barriers of microarray platform-dependency, in order to improve our understanding of biological processes, medical conditions, and diseases. Here, some of these efforts of microarray data integration are reviewed.

3.1 Data Integration in Microarrays

Microarray technology has become an indispensable tool for monitoring genome wide expression levels of genes in a given organism. From the Patric Brown’s lab, the technology has evolved representing both a technological and a conceptual advancement of the field, and

[21] http://www.ebi.ac.uk/arrayexpress
[22] http://www.ncbi.nlm.nih.gov/geo/
[23] http://cibex.nig.ac.jp/index.jsp

has expanded worldwide, where many laboratories are now making their own arrays, in addition to the availability of commercial vendors. With the increasing number and availability of gene expression studies of various organisms, there has been a pressing need to develop approaches for integrating results across multiple studies.

In a cross-study analysis, the data, relevant results and statistics of several studies are combined. There are different practical advantages in such studies. Cross-study analysis has the potential to strengthen and extend the results gathered from the individual studies. This can turn an investigation towards higher accuracy and consistency, and thus, help in robust information mining. Moreover, output of such a study can provide a broader picture of gene-expression as the final 'integrated'-result emerges based on a set of individual studies. Cross-study analysis can also compensate for the possible data-errors in individual studies. The cost of such a study can be kept low by using the exiting studies, as otherwise the setting up of each microarray investigation is not inexpensive. However, while attempting to actualize integration of microarray studies, there are much higher challenges and difficulties as genetic expressions of different studies are neither readily comparable nor can directly be combined. There are several approaches to cross-study analysis, and they somewhat broadly fall into two categories – A. studies where integration occurs at the interpretative level, B. studies where integration takes place with rescaling of the expression values.

3.1.1 Integration at the Interpretative Level

Meta-analysis is emerging as a standard way for the comparison of microarray studies at interpretative level. It involves comprehensive reanalysis of the primary data by merging data from multiple studies. Certain general reviews on meta-analysis include Hedges & Olkin (1985), Cook et al. (1995), Normand (1999), Ghosh et al. (2003) and Moreau et al. (2003). As broadly defined by Normand (1999), meta-analysis is the quantitative review and synthesis of the results of related but independent studies. Despite having certain demerits of merged primary dataset as reviewed by Larsson et al. (2006), the method is becoming useful in microarray studies with the expansion of the sheer volume of microarray data. The success of meta-analysis is dependent on the quality of the underlying data. When accuracy of one or more concerned microarray platforms is questionable, the outcome may become influenced. Nevertheless, browsing through the various studies, where the observation on accuracy, reliability and reproducibility of microarray platforms clearly ranges from relatively discouraging (Severgnini et al., 2006; Tan et al., 2003) through cautiously optimistic (R. A. Irizarry et al., 2005; Larkin, Frank, Gavras, Sultana, & Quackenbush, 2005) to impressive

(Canales et al., 2006; Leming Shi et al., 2006), the overall assessment of the usefulness of meta-analysis of similar microarray studies is cautious optimism. Moreover, the major sources that contribute to the discordance in this regard are mainly – random noise, biological and experimental variations in the samples being analysed, and the variation due to the technical methodology used in the platforms. It is possible to overcome the discordance to a greater extent with judicious and robust application of relevant statistical methods, standard reporting methods, as well as careful application of meta-analysis techniques.

The core objectives of meta-analysis are to increase efficiency in detecting an overall treatment effect, to estimate degree of benefit associated with a particular study, and to assess the amount of variability between studies etc. In the recent past, several statistical methods aiming at detecting differentially expressed genes among multiple conditions have been proposed in individual experiments (Breitling, Armengaud, Amtmann, & Herzyk, 2004; Efron, Tibshirani, Storey, & Tusher, 2001; Newton, Noueiry, Sarkar, & Ahlquist, 2004; Tusher, Tibshirani, & Chu, 2001). Pan (2002) has published a comparative review on these statistical methods in replicated microarray experiments. However, most standard meta-analysis methods cannot be applied directly to microarray experiments as microarray technology is unique with its slew of issues, including its diverse experimental platforms, complicated data structures, presence of duplicate spots as well as often having a large number of genes tested for differential expression.

In 1925, a simple application of meta-analysis was implemented as Fisher's Inverse χ^2 test (Fisher, 1925). The method computes a combined statistic from the *P*-values obtained from the analysis of the individual datasets, S = -2 log(Π_i P$_i$). Here, S follows a Chi-square distribution with 2*l* degrees of freedom under the joint null-hypothesis. The approach does not require additional analysis, and is easy to use; however, it cannot estimate the average magnitude of differential expression in microarrays just by working with the p-values. The approach also remains highly dependent on the method used in the individual analysis.

Meta-analysis based on the t-statistic was reviewed by Normand (1999) in the context of biostatistical applications. Choi et al. (2003) adopted the classic biostatistical meta-analysis framework for microarray analysis, and implemented their methods as a Bioconductor (R. C. Gentleman et al., 2004)-package, *GeneMeta*[24]. The approach of Choi et al. (2003) was a model-based systematic integration of microarray datasets, where a hierarchical modeling approach to assess intra- and inter-study variation was used. The method estimated an overall

[24] *www.bioconductor.org/packages/bioc/html/GeneMeta.html*

effect size as the measure of differential expression for each gene through parameter estimation and model fitting. The effect size was a *t*-like statistic, which was the summary statistic for each gene from each individual dataset, and was defined to be a standardized mean difference between cancer and normal samples in a microarray data set. Integration of data using this meta-analysis method promoted the discovery of small but consistent expression changes and increased the sensitivity and reliability of analysis. Later, Hong and Breitling (2008) found that this t-based meta-analysis method greatly improved over the individual analysis, however it suffered from potentially large amount of false positives when P-values served as threshold.

Based on the traditional effect size model (Choi et al., 2003), Hu et al. (2005) proposed a model for implementing an efficient methodology for identifying genes that are differentially expressed between lung *adenocarcinoma* samples and normal samples by modeling the effect size and integrating information from two Affymetrix oligonucleotide studies. In this study, they presented a measure to quantify Affymetrix gene chip data quality for each gene in each study where the quality index measured the performance of each probeset in detecting its intended target. They extended the traditional effect size model by using the quality index as a weight for combining information from different Affymetrix chip types, and incorporating this weight into a random-effects meta-analysis model.

Rhodes et al. (2002) proposed a statistical model for performing meta-analysis in their four prostate cancer microarray datasets, two of which were cDNA (also known as, *spotted arrays*) data and the remainder Affymetrix microarray data. The model was based on the statistical confidence measure rather than the expression levels, while avoiding direct comparisons of data sets and related cross-platform normalization issues. Each gene in each study was treated as an independent hypothesis, and significance was assigned based on random permutations. Then a meta-analysis model was implemented to assess the similarity of the findings between studies based on multiple inference statistical test for each possible combination of studies. This ultimately identified statistically reliable sets of over- and under-expressed genes in prostate cancer. A cohort of genes were found to be consistently and significantly dysregulated in prostate cancer. The approach of Rhodes et al. is highly conservative because of the choice of null hypothesis; and therefore, the approach may not be recommendable. The data used by Rhodes et al. (2002) were later used by Choi et al. (2003), and they demonstrated that their method could lead to the discovery of small but consistent expression changes with increased sensitivity and reliability.

A Bayesian mixture model transformation of microarray data was proposed by Parmigiani et al. (2002). The modeling framework was used for molecular classification, and it provided both a statistical definition of differential expression and a precise, experiment-independent, definition of a molecular profile. It also generated natural similarity measures for traditional clustering and gave probabilistic statements about the assignment of tumors to molecular profiles.

The rank product is a non-parametric statistic, and was first proposed to detect differentially expressed genes in a single dataset (Breitling et al., 2004). To integrate multiple microarray studies from different platforms and/or different laboratories, a rank product meta-analysis algorithm was implemented as a Bioconductor package, *RankProd* (F. Hong et al., 2006). The algorithm computed pairwise fold change (FC) with replicates for each gene between treatment and control in both directions, respectively. Then, it transformed FC into rank among all genes under study, searched for genes that were consistently top ranked across replicates, and finally generated a single significance measurement for each gene in the combined study. In this approach, converting FC into ranks increased robustness against noise and heterogeneity across studies.

Grutzmann et al. (2005) performed a meta-analysis of four independent studies that applied high-density arrays for expression profiling of pancreatic cancer. They used a consensus set of UniGene clusters measured in all four studies, and applied a random effect model described by Whitehead & Whitehead (1991), whereby expected values of individual study effects were assumed to be normally distributed. With the random effect model, an unbiased estimator for the PDAC (*Pancreatic ductal adenocarcinoma*) effect across all studies was measured, and was used to measure joint differential expression of a gene across all studies.

With three publically available breast cancer datasets having information on lymph node status, Garrett-Mayer et al. (2008) compared the strength of evidence of gene–phenotype associations as well as combined effects across studies. For this, the three studies were first analyzed for reliability, and then, the comparability of results with regards to the genes associated with lymph node status was assessed. Instead of actually combining the data across studies, they mainly performed a comparative analysis making inferences based on the genes consistently measured in all studies, and finally estimated combined inferential statistics. Their proposed methods were implemented in the R (Ihaka & Gentleman, 1996)-library, *MergeMaid*[25] (Cope, Zhong, Garrett, & Parmigiani, 2004). The novel addition in this work

[25] http://astor.som.jhmi.edu/MergeMaid/

was the use of a reliability measure, which was extended to be applied for more than two studies.

Meta-analyses methods are useful; however, as Eysenck (1995) mentioned, they require careful selection of inclusion criteria for participating studies and sound statistical models to avoid misleading conclusions. To date, broader comparisons across various integration approaches have not been conducted. However, Hong and Breitling (2008) compared performance of three widely used methods - Fisher's inverse Chi-square approach, t-like statistic of Choi et al. (2003) and rank product method (Breitling et al., 2004; F. Hong et al., 2006), and found that among the three methods, the non-parametric rank-product method outperformed in terms of sensitivity and specificity.

In general, the overall framework used in all the above studies, where data integration occurs at the interpretative level can be outlined as shown in Figure 3.1.

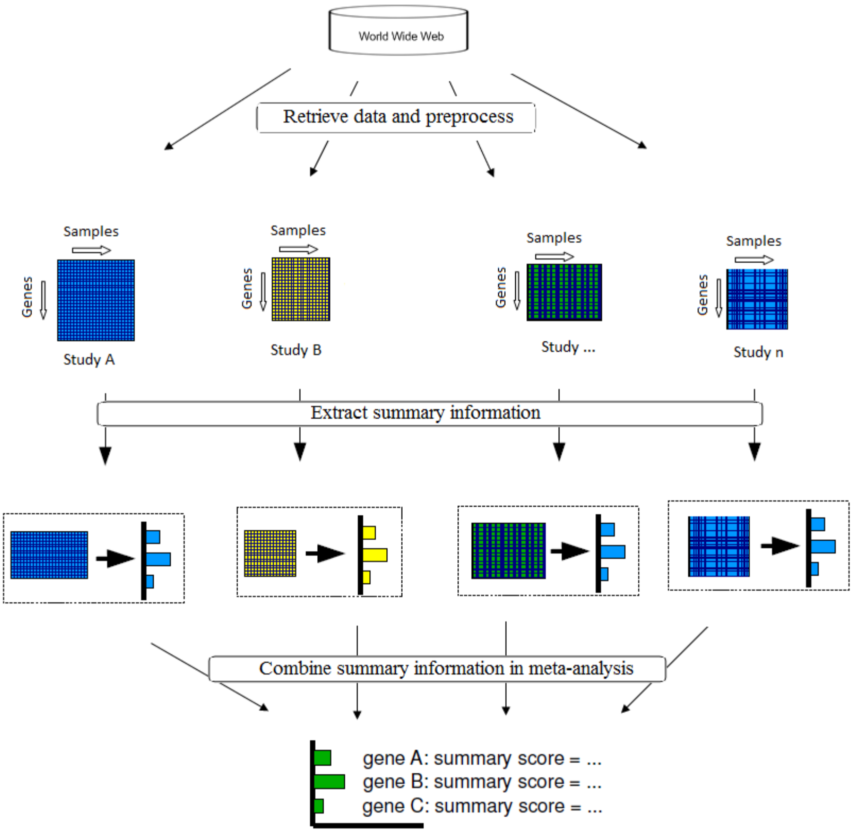

World Wide Web
Retrieve data and preprocess
Samples
Genes
Study A
Samples
Genes
Study B
Samples
Genes
Study ...
Samples
Genes
Study n
Extract summary information
Combine summary information in meta-analysis
gene A: summary score = ...
gene B: summary score = ...
gene C: summary score = ...

Figure 3.1 Microarray data integration at interpretative level

3.1.2 Integration with Rescaling of the Expression Values

Contrary to the meta-analysis approaches, where the results of the individual studies are combined at an interpretative level, there are published researches where microarray expression data from various studies are integrated after transforming the expression values to numerically comparable measures. This is attained by deriving the genetic expression values from the individual platforms, and then, applying specific data transformation and normalization methods. The derived data from the individual studies are subsequently combined, which enlarges the sample size. Any further analysis, as required, is carried out on the new merged dataset. The cross-referencing of the genes between the platforms is usually achieved using UniGene database (Wheeler et al., 2000).

Ramaswamy et al. (2003) reported rescaling of gene-expression values of a common set of genes. The set of the common genes were from five microarray datasets generated by individual labs on different microarray platforms. The rescaled common genes were combined to produce a larger set of data. From the combined dataset, a gene expression signature was identified, which distinguished primary from metastatic tumors.

A standard normalization scheme can be used to combining cDNA and Affymetrix data. Hwang et al. (2004) normalized the expression values of each gene across the samples for each platform so that the mean of each gene equals to zero and the standard deviation equals to unity, respectively. The normalized data were, then, combined to form a large dataset. Earlier, Cheadle et al. (2003) proposed normalization and standardization of cDNA microarray intensity values within datasets using a *Z-score transformation* method. The method converted the raw intensity data from each experiment into $\log_{10}$, and then, Z-scores were calculated by the classical method, i.e., by subtracting the overall average gene intensity (within a single experiment) from the raw intensity data for each gene, and dividing that result by the standard deviation of all of the measured intensities. The application of this classical method in microarray normalization provided a way of standardizing data across a wide range of experiments, while allowing comparison of microarray data independent of the original hybridization intensities.

Based on the *distance weighted discrimination* (DWD) method of Marron & Todd (2002), Benito et al. (2004) integrated cDNA data with Agilent oligonucleotide data. DWD, which was basically an improvement method for *Support Vector Machines* in HDLSS (*High Dimension*, *Low Sample Size*) contexts, was used as an approach for removing systematic bias effects and then, merging the different data sets.

A gene-specific scaling factor was calculated in Bloom et al. (2004), and was used to integrate microarray data from Affymetrix and cDNA platforms. Here, for each gene common to both platforms, expression levels for a reference RNA sample on the spotted arrays was averaged and compared to expression measured for the reference RNA sample on the appropriate Affymetrix *GeneChip* to calculate the scaling factor. This scaling factor was used to adjust the remaining data towards integrating the platforms.

Shen et al. (2004) used a *two-stage Bayesian mixture modeling strategy* based method proposed by Parmigiani et al. (2002). This model was to integrate multiple independent studies addressing similar questions while considering different platforms – Affymetrix and *inkjet* oligonucleotides. The mixture modeling approach reportedly unified disparate gene expression data based on a probability scale of differential expression, the *poe*-scale (Parmigiani et al., 2002), and derived an inter-study validated 90-gene 'meta-signature' that predicted relapse-free survival in breast cancer patients.

In addition to common data transformation and normalization procedures, Jiang et al. (2004) added a distribution transformation (*disTran*) step in their study. The method transformed two microarray datasets belonging to two Affymetrix chip types so that the empirical distributions of two lung cancer datasets could become identical and be combined. The *disTran* method reportedly provided improved consistency in the expression patterns of the multiple datasets.

Two data integration methods, namely quantile discretization (QD) and median rank scores (MRS) were used in Warnat et al. (2005) for direct integration of raw microarray data from six publicly available cancer microarray gene expression studies conducted by means of cDNA and oligonucleotide microarrays. In this study, comparable measures of gene expression from the independent data sets of the varied microarray platforms were numerically derived such that the different microarray data adhere to a common numerical range. These derived data were then integrated, and used to build SVM (support vector machine) classifiers for cancer classification. Similar to *disTran*, the quantile normalization technique, i.e., MRS, and QD of Warnat et al. (2005) were used to transform the microarray data from diverse platforms so that their empirical distributions are identical. The approaches (*disTran*, MRS and QD) can significantly improve the comparability of cross-platform microarray data. These methods work well for classification tasks, but can suffer from information reduction, limiting their applicabilities other than classification.

Stafford & Brun (2007) presented a calibration process for cross-laboratory and cross-platform microarray expression data. Using Agilent and Affymetrix expression platforms, they employed precision and sensitivity measurements along with biological interpretation for better selection of genes with respect to a particular outcome. Precision and sensitivity measurements were useful in finding the minimal detectable fold-change and raw performance values for a microarray platform. Gene Ontology and pathway analyses were considered in the study as a valuable way of examining and comparing the actual biological interpretation.

Xu et al. (2008) used four independent breast cancer datasets, and identified a structured prognostic signature consisting of 112 genes organized into 80 pair-wise expression comparisons. They extended a previously proposed method (Geman, d'Avignon, Naiman, & Winslow, 2004), validated on a prostate cancer study, to predict distant metastases in breast cancer. The method of integration was based on the ranks of the expression values within each sample. Since the ranks of the features were invariant to all types of within-array preprocessing, there was no need to prepare the data for integration, in particular there was no need for data normalization.

XPN (Shabalin, Tjelmeland, Fan, Perou, & Nobel, 2008) is another method that deals with the problem of cross-study normalization: how to combine microarray datasets in order to produce a single, unified dataset to which standard statistical procedures can be applied. The method was based on a block linear model, and used three existing breast cancer datasets from Agilent oligonucleotide platform and Affymetrix GeneChip. The model assumed that the samples of each available study fell roughly into one of the statistically homogenous sample groups identified, and that each group was defined by an associated gene profile that was constant within each of the estimated gene groups. The proposed method applied sample standardization and gene median centering before combining the data from the studies. To identify blocks (or, clusters) in the data, *k-means clustering* was applied independently to genes and samples of the combined data. Each gene expression value subsequently became a scaled and shifted block mean plus noise. XPN was reportedly preserved biological information according to ER (error rate) prediction error rates while removing systematic differences between platforms.

NLT or Normalized Linear Transform (Xiong, Zhang, Chen, & Yu, 2010) is a method in which the samples of two microarray platform were linearly mapped such that the numerical range of the expression values of each gene became identical. The mapped data were, then, combined and normalized across samples to zero mean and unity standard deviation. Apparently, the approach avoids information reduction as it preserves the relative ranking order of the expression values for each gene.

The methods highlighted above pose important examples of integration of microarray datasets with rescaling of the gene expression values. Each of the approaches is unique; however, the overall organization of the methods follows a general framework, which is outlined in Figure 3.2.

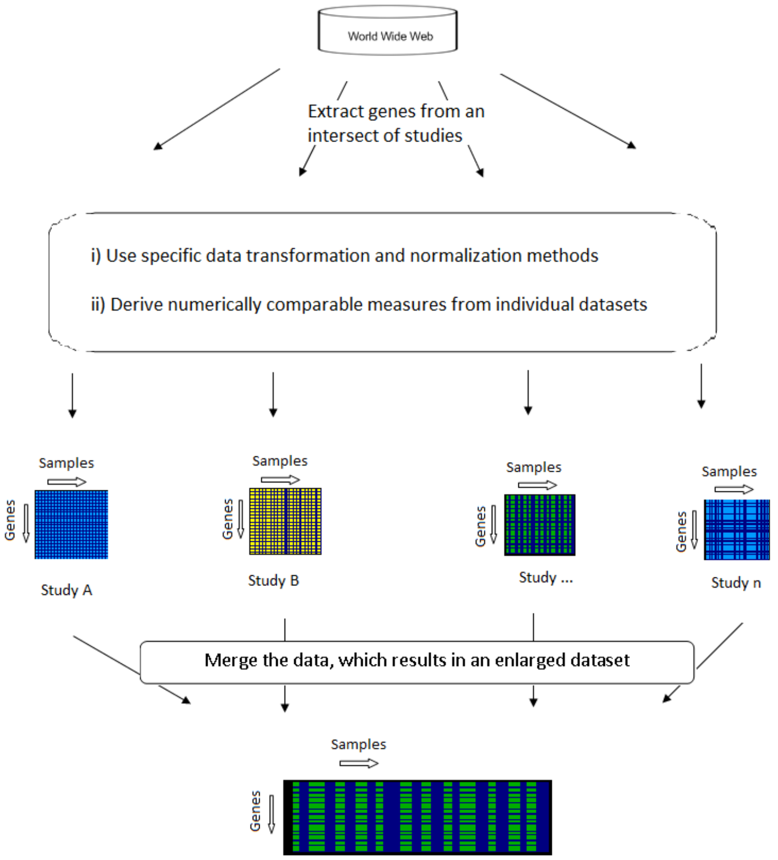

World Wide Web
Extract genes from an intersect of studies
i) Use specific data transformation and normalization methods
ii) Derive numerically comparable measures from individual datasets
Samples
Genes
Study A
Study B
Study ...
Study n
Merge the data, which results in an enlarged dataset

Figure 3.2 Microarray data integration with rescaling of expression values

Note :

- Aspects of this chapter have been published:

Sarmah, C. K., & Samarasinghe, S. (2010) Microarray data integration: frameworks and a list of underlying issues. *Current Bioinformatics*, *5*(4), 280-289.

Chapter 4
Data Assessment and Normalization

4.1 Data Collection

Affymetrix *GeneChip*® and *GenePix*® cDNA data were obtained from the Tumour Bank, The Children's Hospital at Westmead, Australia. The data belonged to childhood leukemia patients. Seven of these children were analysed both on Affymetrix (*HGU-133A* chip) as well as on cDNA platforms. Additionally, there are ten Affymetrix HGU-133A chips obtained from 10 healthy children. This research project is based on these datasets, while emphasising on the important consideration that these data were generated from an ideal experimental setup.

Certain data-quality assessments, followed by data normalization process are carried out with the help of open-source statistical software, *R* (Ihaka & Gentleman, 1996) and *Bioconductor* (R. C. Gentleman et al., 2004). Assessing the quality of data is given its due importance as it ensures that the homogeneity of the data remains, and that the data adhere to the minimal data quality standards, although it may not conclusively indicate flawlessness in the original microarray data generation pipeline.

Unless stated otherwise, the sources used here to help illustrate the processes and their outcomes are: Gentleman et al. (2005), Hahne et al. (2008), Kauffmann et al. (2009), Wilson & Miller (2005).

4.2 Affymetrix Data

Assessing the quality of data is critical prior to carrying out any analytical investigations. A list of assessments is made using the available dataset. Subsequently, normalization is carried out, which is followed by assessing the quality of the normalized arrays.

4.2.1 Assessment of Raw Affymetrix Data

4.2.1.1 Inspection for Hybridization Artefacts

A simple look at the images of the scanned arrays can pick up hybridisation artefacts arising from factors including scratches, air bubbles, and problems with staining, mixing or washing. Appendix A.1 displays pseudo-images of the intensities from all features on each array on the basis of how they are physically arranged on the arrays. It does not indicate detection of any notable artefact.

4.2.1.2 MA Plots

Appendix A.2 presents MA plots for each array against a pseudo-array, which contains the median values of all the arrays. Accordingly, *M* and *A* is defined as:

$$\begin{aligned} M &= \log_2(I_1) - \log_2(I_2) \\ A &= \frac{1}{2}[\log_2(I_1)+\log_2(I_2)] \end{aligned} \qquad (9)$$

where, I_1 is the intensity of the array studied, and I_2 is the intensity of the pseudo-array. Typically, the mass of the distribution in an MA plot should be about the M = 0 axis, without having any trend in the mean of M as a function of A. A trend, shown as a horizontal red line in a plot, in the lower range of *A* usually indicates that the arrays have different background intensities, whereas that in the upper range of A refers to saturation of the measurements. However, both of these can be addressed to a certain extent by background correction and non-linear normalization (e.g., quantile normalization), respectively.

4.2.1.3 Array Intensity Distributions

Systematic bias and related anomalies across the arrays can be identified by plotting the array intensity distributions. Figure 4.1 gives the distribution of the raw, $\log_2$-transformed probe intensities across all 17-GeneChip arrays, which include 7 sick and 10 healthy children. A box in the boxplots or a line in the plot of smoothed histograms corresponds to one array. Ideally, one expects the boxes to have similar size (IQR) and *y*-position (median); and similar shapes and ranges in the smoothed histograms. With regards to the distribution of the untreated arrays in the figure, it does not highlight any alarming variations.

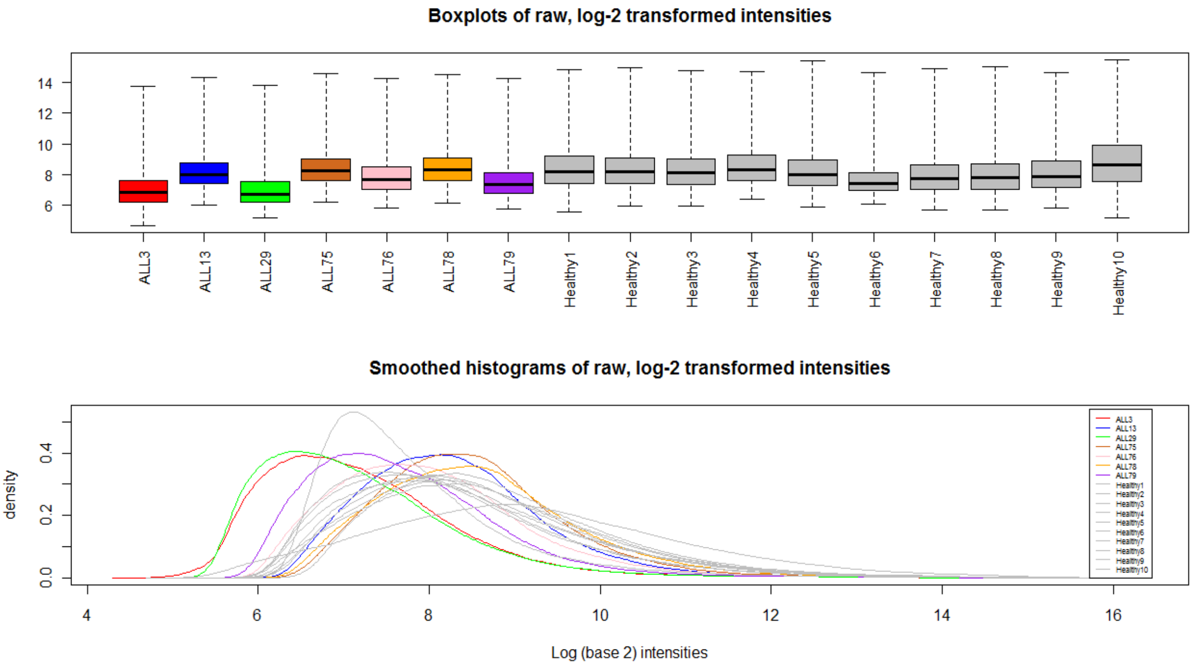


Figure 4.1 Intensity distributions of raw, $\log_2$-transformed Affymetrix arrays

4.2.1.4 Between-Array Comparison

A heatmap serves to assess whether one or more arrays are different from the rest; and thereby, detect the outlier arrays. It is also at times used to check whether arrays cluster according to certain biological meaning. A heatmap is, thus, a representation of distances between the arrays, where the median of the absolute values of the difference between each array-pair is considered as a measure of distance. It is shown in equation 10, where M_{xi} and M_{yi} represents the M-value of the i^{th} probe on the x and y array. M_{xi} (similarly, M_{yi}) can be decomposed as : $M_{xi} = z_i + \beta_{xi} + \varepsilon_{xi}$, where z_i is the probe effect for probe i (the same across all arrays), ε_{xi} are independent and identically distributed (*i.i.d.*) random variables with mean zero, and β_{xi} represents differential expression effects and is such that for any array x, the majority of values β_{xi} are negligibly small, i.e., close to zero.

$$d_{xy} = median\,|M_{xi}\text{-}M_{yi}| \qquad (10)$$

Arrays whose distance matrix entries, i.e., d_{xy} values, are way different should bring reason for suspicion. Figure 4.2 is a false colour heatmap of between-array distances, computed as the median absolute difference (L_1-distance) of the M-values for each pair of arrays on every probes without any filtering. The colour scale is chosen to cover the range of L_1-distances encountered in the dataset. Using the principles of Kauffmann et al. (2009), arrays for which the sum of the distances to the others is much different from the others, can be considered as outlier arrays. In expectation, all values of d_{xy} are the same, namely 2-times the standard deviation of ε_{xi}. Arrays whose distance matrix entries are way different from the rest remain spaced apart in the accompanying tree-diagram (in the upper side) of a heatmap. Accordingly, in Figure 4.2, none of the arrays seem to lie as outlier.

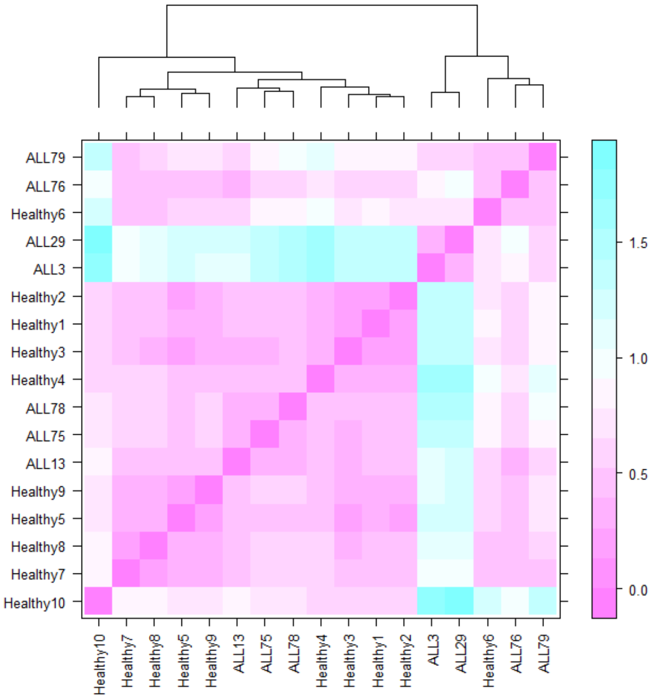
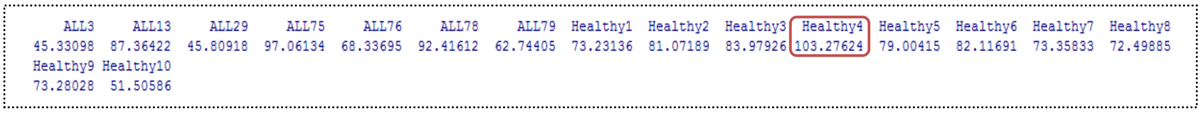


Figure 4.2 Heatmap for between-array distances for raw Affymetrix arrays

4.2.1.5 GeneChip-Specific Assessments

4.2.1.5.1 Average Background

Signal-to-noise ratio can be affected by background intensity of Affymetrix arrays. The typical average background values range from 20 to 100 for arrays scanned with the *GeneChip*® Scanner (Affymetrix, 2008). Extreme background intensity values of arrays outside of this range may be indicative of problems. There is only one array, *Healthy4*, which is found to be not extreme but falls just outside of this range, as shown in the snapshot below. The average background array intensities are also listed to left of Figure 4.3 (on page 55) and Figure 4.4 (on page 56). (More details on these figures will be provided in section 4.2.1.5.5: GAPDH and β-actin ratios).

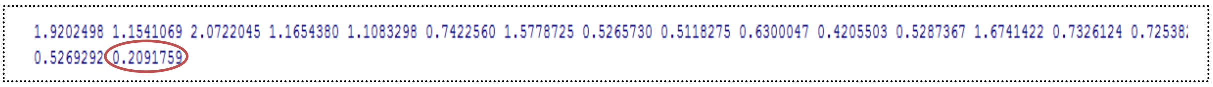


```
    ALL3     ALL13     ALL29     ALL75     ALL76     ALL78     ALL79  Healthy1  Healthy2  Healthy3  Healthy4  Healthy5  Healthy6  Healthy7  Healthy8
45.33098  87.36422  45.80918  97.06134  68.33695  92.41612  62.74405  73.23136  81.07189  83.97926 103.27624  79.00415  82.11691  73.35833  72.49885
Healthy9 Healthy10
73.28028  51.50586
```

4.2.1.5.2 *Scale Factors*

Scaling, the simplest type of normalization, makes the assumption that the distribution of signal intensities on an array is normal (or, Gaussian); and it merely shifts the distribution to be centred to a particular point. Affymetrix's MAS 5.0 expression summary algorithm scales the mean of the signals to a certain value, the default being 500, while discarding the top and bottom 2% of an array as outliers. To determine whether an array is of poor quality, Affymetrix suggests that the scale factors should be similar among samples and not vary more than about 2 to 3-fold from each other.

Scale factors using MAS 5 algorithm can be viewed in Figure 4.3 and Figure 4.4 where the blue stripe in the image represents the range where scale factors are within 3-fold of the mean for all chips. The scale factors are plotted as a horizontal line from the centre line of the image. A horizontal line to the left from the centre line corresponds to a down-scaling, and to the right represents an up-scaling. Scale factors that fall within this 3-fold region are coloured blue, while the rest remaining outside this area are coloured red. Among the untreated Affymetrix arrays, *Healthy10* is just falling outside the area, as shown below in the box.

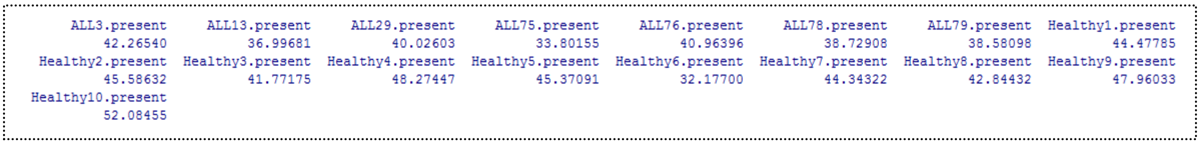


```
1.9202498 1.1541069 2.0722045 1.1654380 1.1083298 0.7422560 1.5778725 0.5265730 0.5118275 0.6300047 0.4205503 0.5287367 1.6741422 0.7326124 0.72538
0.5269292 0.2091759
```

4.2.1.5.3 *Detection Calls*

Detection calls provide an overall measure of quality. They are used for flagging genes as having been reliably detected, and are given by '% Present'-call that represents the percentage of probesets called 'present' on an array (B. M. Bolstad et al., 2005). Probesets are flagged *Marginal*, or *Absent* when the PM values for that probeset are not considered to be significantly above the MM values for the same probeset.

High variations in present calls between similar samples give cause for suspicion as it means varying amounts of labelled RNA have been successfully hybridized to the chips because of certain noise or interference in the array processing pipeline. However, the *% present*-scores vary considerably with tissue type, and the type of experiment condition under study; and consequently, no absolute quality cut-offs is recommended. Percent present scores are listed to the left of Figure 4.3 and Figure 4.4, and also are presented below.

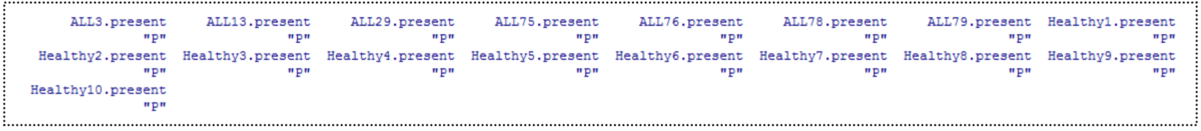


```
    ALL3.present     ALL13.present     ALL29.present     ALL75.present     ALL76.present     ALL78.present     ALL79.present  Healthy1.present
        42.26540          36.99681          40.02603          33.80155          40.96396          38.72908          38.58098          44.47785
Healthy2.present  Healthy3.present  Healthy4.present  Healthy5.present  Healthy6.present  Healthy7.present  Healthy8.present  Healthy9.present
        45.58632          41.77175          48.27447          45.37091          32.17700          44.34322          42.84432          47.96033
Healthy10.present
        52.08455
```

4.2.1.5.4 Hybridisation Controls

Into the hybridisation cocktail just prior to it being placed on a GeneChip, a number of control oligonucleotides are added to subsequently verify the efficiency of hybridization performance. These additional, labelled cRNAs (*BioB, BioC, BioD* and *CreX*) are also known as *Spike-in probesets*, and are derived from *Bacillus subtiliis*, a bacterium commonly found in soil. The intensity of these transcripts is examined later, along with the consideration of the fact that nothing should bind to their probesets.

BioB should ideally be called present on every array. Another acceptable level for it to be called ‘present’ is their presence in 70% of the chips in an experiment. If BioB is routinely absent, it indicates that the assay is performing with suboptimal sensitivity. Results for the 17 chips are listed below, which indicate that all the chips have performed well in this respect.



```
     ALL3.present     ALL13.present     ALL29.present     ALL75.present     ALL76.present     ALL78.present     ALL79.present  Healthy1.present
              "p"               "p"               "p"               "p"               "p"               "p"               "p"               "p"
 Healthy2.present  Healthy3.present  Healthy4.present  Healthy5.present  Healthy6.present  Healthy7.present  Healthy8.present  Healthy9.present
              "p"               "p"               "p"               "p"               "p"               "p"               "p"               "p"
Healthy10.present
              "p"
```

4.2.1.5.5 GAPDH and β-actin Ratios

Affymetrix probesets are designed to hybridize to either end of certain quality control genes, most notably GAPDH and *β*-actin. GAPDH and *β*-actin are relatively long genes, and most cell types ubiquitously express them. Majority of Affymetrix chips contain separate probesets targeting the 5′, mid and 3′ regions of these genes.

Typically, transcription starts from the 3′ end of a gene. Therefore, a measure of the quality of the RNA hybridised to a chip is possible to obtain by comparing the signal from the 3′ probeset to either the mid or 5′ probesets. A high 3′ : 5′ signal ratio indicates the presence of truncated transcripts, which may be either due to the under-performance in the *in vitro* transcription stage or because there is a general degradation of the RNA.

Often RNA to be hybridized to a chip is also prepared using the *Affymetrix small-sample protocol*, instead of the *Affymetrix standard protocol*. The former uses an extra amplification step that may increase the frequency of short transcripts in solution, and unavoidably introduce some 3′ bias into the population of labelled transcripts. In such cases, 3′ to mid ratios is recommended for quality measurement (Affymetrix, 2008).

In Figure 4.3 and Figure 4.4, GAPDH and *β*-actin ratios for 3′:5′ and 3′: mid are shown respectively. GAPDH ratios are plotted as circles, and β-actin ratios are as triangles. GAPDH values that are considered potential outliers (ratio > 1.25) are coloured red, otherwise they are blue. On the other hand, β-actins are longer genes, and the recommended value for the ratio is below 3. The 3′:5′ plot of Figure 4.3 presents the measures of GAPDH and β-actin in *Healthy6* and *ALL29* outside of the recommended value. However, only the GAPDH measure of *Healthy6* is found higher than the recommended value in 3′: mid plot of Figure 4.4.

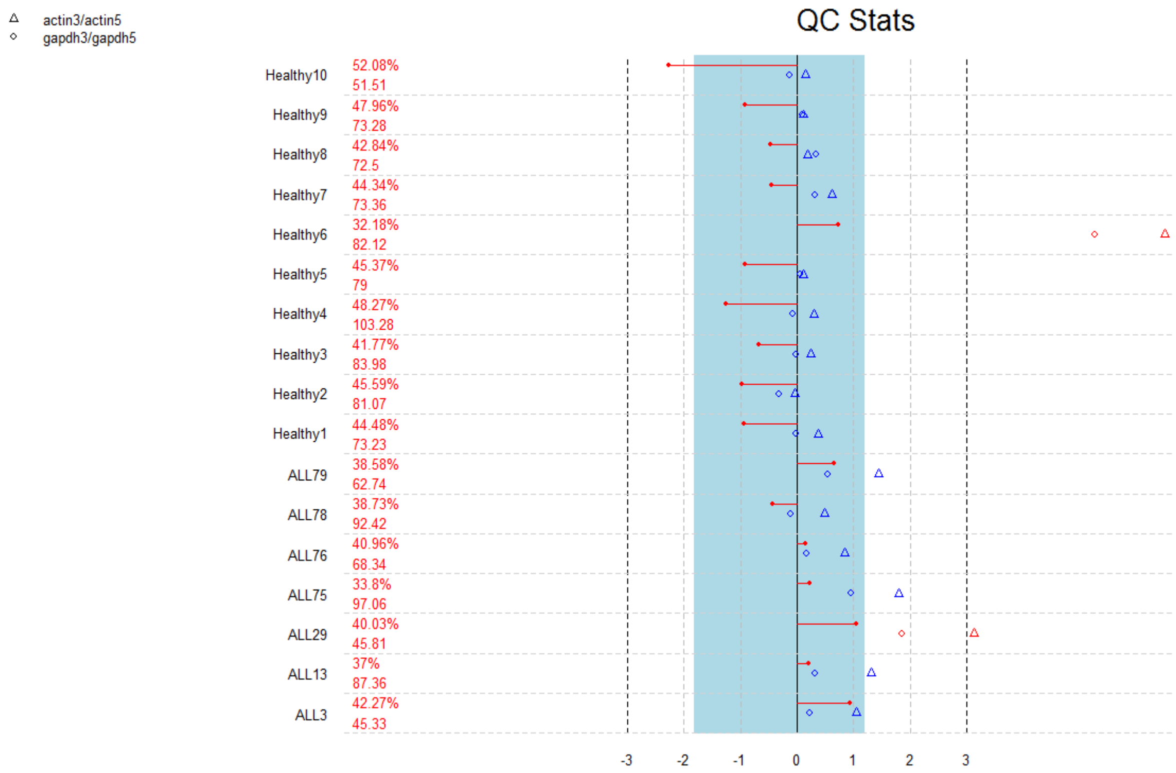


Figure 4.3 β-Actin and GAPDH (3′:5′ ratios)

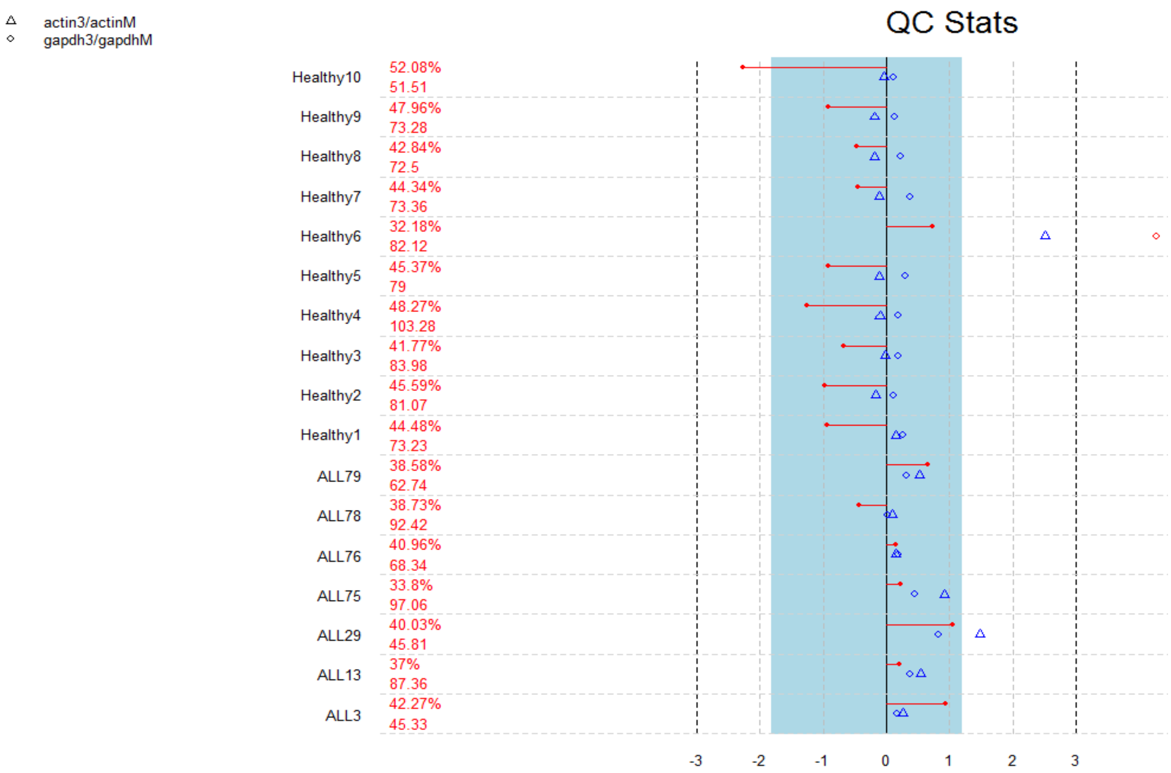


Figure 4.4 β-Actin and GAPDH (3′: mid-ratios)

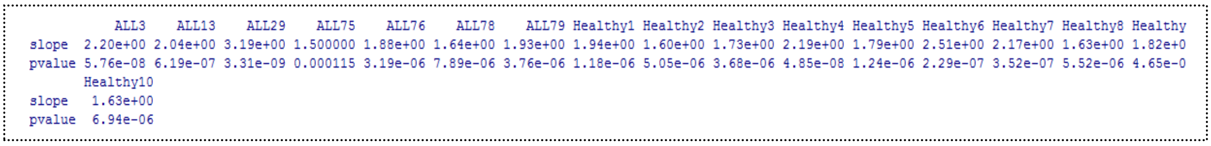
4.2.1.5.6 RNA Degradation

For assessing chip-quality, a more global indicator is often desirable than using the expression measures of only a few control genes such as β-Actin and GAPDH. Analysis of RNA degradation compensates this requirement.

As Gautier et al. (2004) explains, RNA molecules are unstable, and subject to degradation that characteristically starts from the 5′ end of each transcript. This also causes the intensities of the probes at the 3′ end of a probeset to remain systematically higher than those at the 5′ end. Individual probes in each probeset are numbered from the 5′ end of the transcript, so relative position within the transcript is known. The mean expression of the individual probes as a function of their relative positions is represented in a RNA degradation plot, which detects poor quality RNA. An array is represented by a single line in such plots, and an array with a slope very different from the rest indicates that RNA used for that array has potentially been handled quite differently from other arrays. Again, high slopes refer to degradation; however it is more important to have agreement between the arrays.

The degradation plot shown in Figure 4.5 is based on ordering the probes within a probeset according to their 3′ position, and then combining the signal from similarly located probes across the array. Each line represents one of 17 HG-U133A chips, and plotted on the *Y*-axis is the mean intensity by probe position.

There is no standard value that tells about how large a slope must be to consider the array to have too much degradation. Different chip-types have different characteristic slopes due to the differences in probeset architecture. According to Bomstad et al. (2005), a slope of 1.7 is typical for HG-U133A chips, and the slopes that exceed it by a factor of 2 or more might indicate degradation.



```
          ALL3    ALL13    ALL29    ALL75    ALL76    ALL78    ALL79 Healthy1 Healthy2 Healthy3 Healthy4 Healthy5 Healthy6 Healthy7 Healthy8 Healthy
slope  2.20e+00 2.04e+00 3.19e+00 1.500000 1.88e+00 1.64e+00 1.93e+00 1.94e+00 1.60e+00 1.73e+00 2.19e+00 1.79e+00 2.51e+00 2.17e+00 1.63e+00 1.82e+0
pvalue 5.76e-08 6.19e-07 3.31e-09 0.000115 3.19e-06 7.89e-06 3.76e-06 1.18e-06 5.05e-06 3.68e-06 4.85e-08 1.24e-06 2.29e-07 3.52e-07 5.52e-06 4.65e-0
       Healthy10
slope   1.63e+00
pvalue  6.94e-06
```

The retrieved degradation summary for the arrays is presented in the box above. For high quality RNA, a slope of 1.7 is typical for HG-U133A chips; and the slopes that are 2 fold or higher than this number may indicate RNA degradation (B. M. Bolstad et al., 2005). However, in general, agreement between the chips is more important than the actual value. None of the HG-U133A chips currently being assessed is found to have a slope outside this recommended value. The RNA degradation plot in Figure 4.5 does not indicate any disagreement between the chips either.

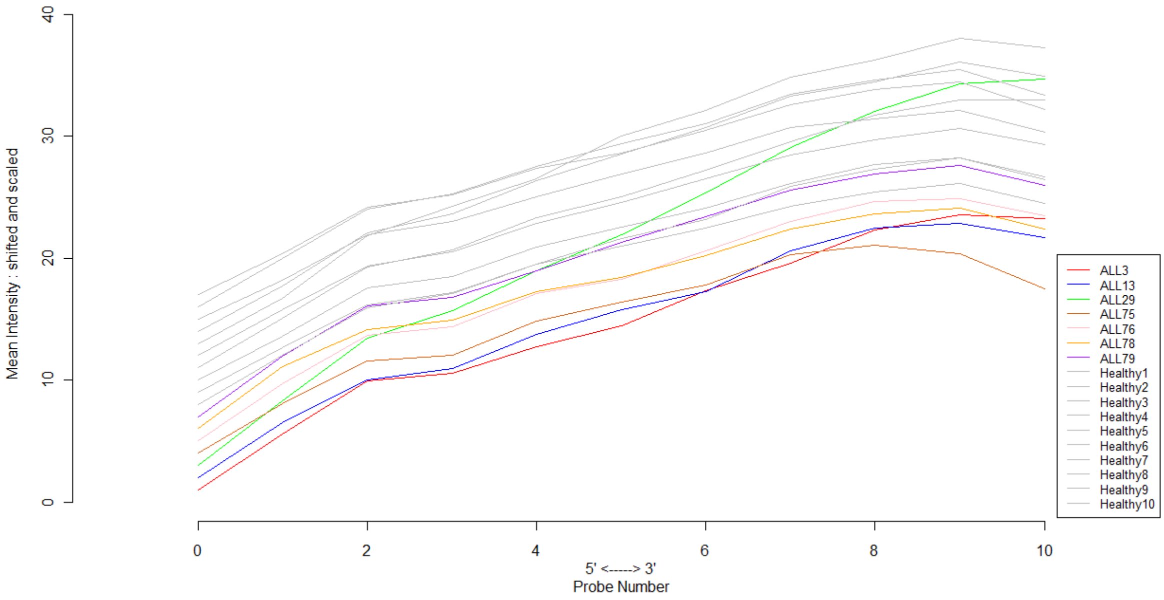


Figure 4.5 RNA degradation plot

4.2.1.5.7 Relative Log Expression (RLE) Plot

Relative expression can be defined as the difference between the log scale estimates of expression $\hat{\theta}_{gi}$ (for each gene, *g* on each array, *i*) and the median value across arrays for each gene, m_g. This can be expressed as in equation 11.

$$M_{gi} = \hat{\theta}_{gi} - m_g \qquad (11)$$

In a RLE plot, problematic arrays are indicated by larger spread, or by a center location different from relative expression, y=0, or both. This means that ideally, the boxes of RLE plot would have small spread, and be centred at y=0. The RLE plot constructed for the 17 HG-U133A chips is given on Figure 4.6, and shows that the ideal spread and y=0 axis is absent in many of these untreated arrays.

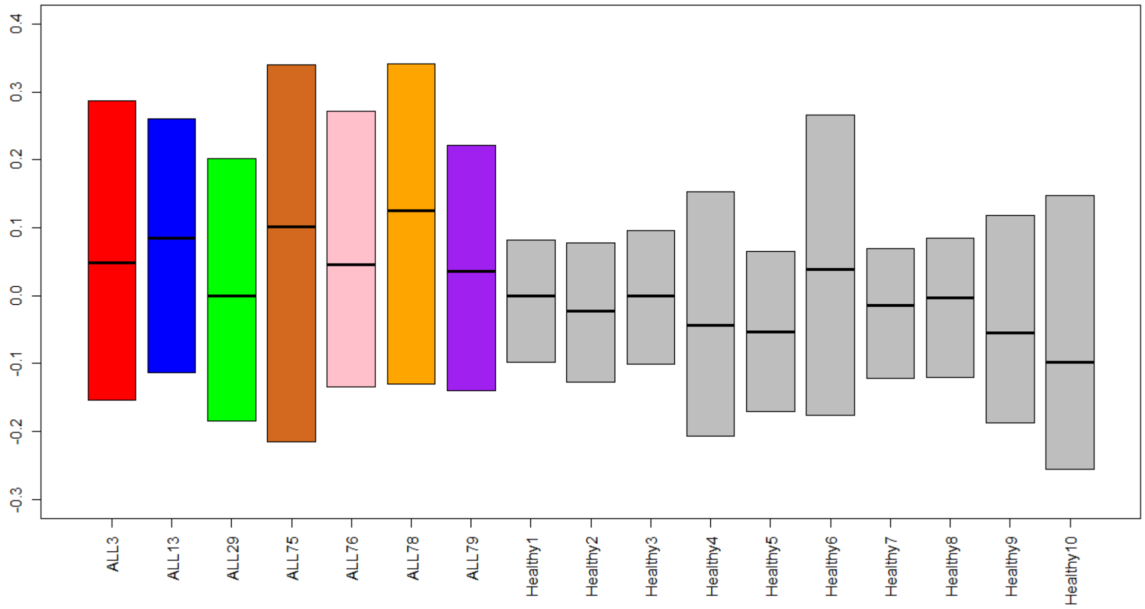


Figure 4.6 RLE (Relative Log Expression) plot

4.2.2 Affymetrix Data Normalization

The general purpose of normalization is to make the results from each of the arrays comparable with the rest. There are various ways as well as combinations proposed for normalization.

There are numerous approaches for normalizing Affymetrix arrays, more than 30 methods have been identified as of 2006 (Rafael A. Irizarry et al., 2006). However, none of the methods is clearly the best (Qin et al., 2006) - each having own trade-offs and making different assumptions about the data. Nevertheless, based on the overall favourable comments and performance in various studies including Bolstad et al. (2003), Grewal et al. (2007), Mar et al. (2009) and web-information[26], *quantile normalization* method using *Robust Multichip Average* (RMA) algorithm is accepted for normalizing the group of 17 Affymetrix (*HG-U133A*) chips.

The *Robust Multi-array (or Multi-chip) Average* or RMA (R. A. Irizarry et al., 2003) uses quantile normalization, and is used here for normalizing the chips. RMA is largely the work of Terry Speed's group at University of California at Berkeley, and only uses PM probes as the method assumes that including the MM probes introduces more variability than the correction is worth. In RMA, the expression measure is obtained using three steps : convolution background correction, quantile normalization, and a summarization method based on a multi-array model fit that uses the median polish algorithm (Tukey, 1977). Starting with the raw probe-level data from a set of GeneChips, the perfect-match (PM) values are background-corrected, quantile normalized, and then finally the linear model is fit to the normalized data to obtain an expression measure for each probe set on each array.

Background correction used in RMA is aimed at correcting only PM values, and is a non-linear correction using a probabilistic model, done on a per-chip basis. It involves a convolution of an exponentially distributed (with mean, α) signal, X and normally distributed (with mean, μ and standard deviation, σ) noise, Y caused by optical noise and non-specific binding. Therefore, the observed PM intensity, S = X + Y. Under this model, the background corrected model is given by E(X|S=s). Benjamin Milo Bolstad (2004) presents the background correction in equation 12, where a=s-μ-σ^2α, b=σ, Φ represents the standard normal distribution function, ϕ is the density function of the normal distribution.

$$E(X \mid S = s) = a + b \frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{s-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{s-a}{b}\right) - 1} \qquad (12)$$

Quantile normalization (B. M. Bolstad et al., 2003), also introduced by Terry Speed's group at University of California at Berkeley, is a robust, routinely used and fast normalization method, which aims to make the distribution of probe intensities the same for every Affymetrix chip. For this, the arrays of signal intensities are sorted in a way that the highest

[26] *U.S. National Cancer Institute* (http://tinyurl.com/27bv7f3)

signal from each array is replaced by the average of all of the highest signals, and the second highest on each array is replaced by the average of all the second highest, and so on. The resultant data do not heavily skew, and the variability of expression measures across chips reduced.

Quantile normalization method forces the values of the quantiles to be equal, and projecting each point of the two vectors' quantiles onto a 45^0 diagonal line produces a transformation that gives the same distribution to both the vectors. Transformation of an intensity is done, as given in equation 13 (Benjamin M. Bolstad, 2006), where x_{ij} is intensity i of a probeset on array j; G_j is the distribution function for the j^{th} array and is estimated in practice using the empirical distribution function; F is the empirical distribution of the averaged sample quantiles across all arrays; and, x^*_{ij} is the normalized intensity.

$$x^*_{ij} = F^{-1} (G_j (x_{ij})) \qquad (13)$$

Expression summarization is the final component of RMA normalization. From a set of background-corrected and quantile-normalized PM probe intensities for each probeset, the process computes a single number to represent the expression level of the targeted gene. The summarization method for RMA is median polish algorithm (Tukey, 1977), which is a robust method that iteratively fits the linear model of equation 14 with constraints median (θ_j)=median (α_i)=0 and median$_i$ (ε_{ij})=0. In the equation, the superscript (n) represents the nth probe set on array j, y_{ij} refers to the observed intensity of the ith probe, α_i represents a probe effect, θ_j is an array effect, and ε_{ij} is measurement error. The $\log_2$ expression values are given by $\hat{\beta}_j^{(n)} = \hat{\mu}^{(n)} + \hat{\theta}_j^{(n)}$.

$$\log_2\left(\mathrm{y_{ij}}^{(n)}\right)=\mu^{(n)}+\theta_i^{(n)}+\alpha_i^{(n)}+\varepsilon_{ij}^{(n)} \qquad (14)$$

The median polish fits the model iteratively, successively removing row and column medians, and accumulating the terms, until the process stabilizes. The residuals obtained at the end give rise to the summarized value for each probe set.

4.2.3 Post-Normalization Assessment

To evaluate the effect of normalization on the Affymetrix arrays, an assessment is carried out of which a few important results are reported here.

4.2.3.1 MA Plots

MA plots for the 17 post-normalized Affymetrix-arrays are given in Appendix A.3. Unlike the earlier MA plots of the raw Affymetrix arrays, these plots show that the mass of the distribution remains about the M = 0 axis, besides having no serious trend in the mean of M as a function of A. The issues present in the untreated arrays have apparently been addressed by the normalization process.

4.2.3.2 Array Intensity Distributions

RMA-normalized Affymetrix chips are shown in Figure 4.7. Comparing the intensity distributions of the raw Affymetrix chips (Figure 4.1, page 51), it appears that normalization has been able to bring about homogeneity in the array intensity distributions.

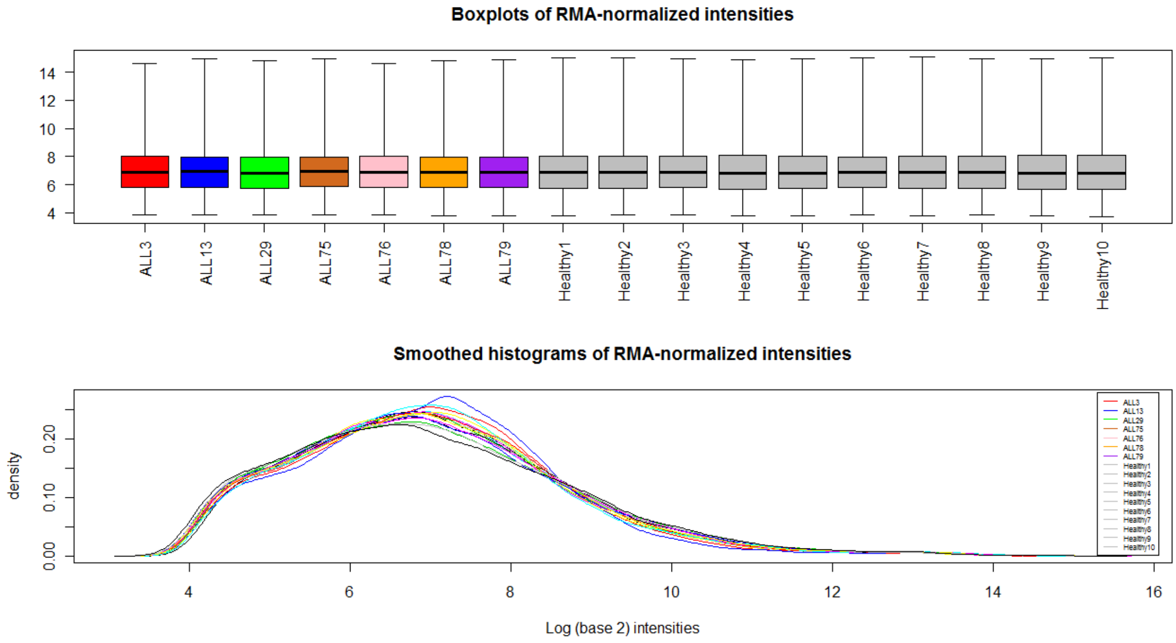


Figure 4.7 Boxplots and smoothed histograms of RMA normalized intensities

4.2.3.3 Normalized Unscaled Standard Error (NUSE) Plot

NUSE values are useful for comparing arrays within one dataset, although their magnitudes are not comparable across different datasets. NUSE plot allows identification of arrays where the standard errors for the gene expression estimates are generally larger relative to the other arrays. The low-quality arrays in a NUSE plot are those that are significantly elevated or more spread out than others. The NUSE plot represents standard error estimates from the *PLM* (probe-level model) fit.

PLM is a model that is fit to probe-intensity data. Specifically, a PLM provides parameter estimates for probe sets and arrays on a probe-set by probe-set (i.e. gene by gene) basis. It is a model of the form: $y_{ij}^{(k)} = f\left(X_{ij}^{(k)}\right) + \varepsilon_{ij}^{(k)}$, where $X_{ij}^{(k)}$ are measured factors, for example probe-effects and treatment specific effects, and covariates for a particular probe and *f* is an arbitrary function. The indices *i*, *j*, and *k* refer to probe array and probeset respectively. A type of PLM is a linear array effect model, which has a parameter for each array. For each probeset k = 1, 2, ..., *K* with *i* = 1, 2, ..., I_k probes each on *j* = 1, 2, ..., *J* arrays, the model (also used by RMA-method), $\mathrm{y}_{\mathrm{ij}}{}^{(k)} = \alpha_i^{(k)} + \beta_j^{(k)} + \varepsilon_{ij}^{(k)}$ is fit, where $\mathrm{y}_{\mathrm{ij}}{}^{(k)}$ are pre-processed $\log_2$ PM intensities, $\alpha_i^{(k)}$ are probe effects and $\beta_j^{(k)}$ are array effects ($\log_2$ expression values). Also, it is assumed that $E\left(\varepsilon_{ij}^{(k)}\right) = 0$, $Var\left(\varepsilon_{ij}^{(k)}\right) = 0$ and $\sum_{i=1}^{I_k} \alpha_i^{(k)} = 0.$If $\hat{\sigma}$ is the estimated residual standard deviation of a probeset in PLM model and $W_i = \sum_j w_{ij}$ is the total probe weight of the probeset in chip *i*, the expression value estimate $(\hat{\mu}_i)$ for the fixed probeset on chip *i*, and its standard error (SE) are given by:

$$\hat{\mu}_i = \sum_j y_{ij} \cdot \frac{w_{ij}}{W_i} \quad \text{and} \quad SE(\hat{\mu}_i) = \frac{\hat{\sigma}}{\sqrt{W_i}} \qquad (16)$$

Replacing $\hat{\sigma}$ by 1 gives *Unscaled Standard Error* (*USE*) of the expression estimate, and to compensate for heterogeneity caused by probes with high variability, low affinity, or a tendency to cross-hybridize, the USE is divided by its median over all chips. This measure is called as *Normalized Unscaled Standard Error* (*NUSE*), and is given by equation 16.

$$\begin{aligned} NUSE(\hat{\mu}_i) \approx \frac{USE(\hat{\mu}_i)}{Median_l\left\{USE(\hat{\mu}_l)\right\}} &= \frac{1}{\sqrt{W_i}} \Big/ Median_l\left\{\frac{1}{\sqrt{W_l}}\right\} \\ &= \frac{Median_l\left\{\sqrt{W_l}\right\}}{\sqrt{W_i}} \end{aligned} \qquad (15)$$

Typically, the arrays should centre around the median NUSE=1, with approximately equal box sizes (i.e. IQRs). Figure 4.8 gives a NUSE plot. The distribution of the chips in the plot is acceptable (though it appears otherwise because of the use of a 'zoom-in' scale ranging from 0.95 to 1.15), and the arrays do not appear to present any quality control problems.

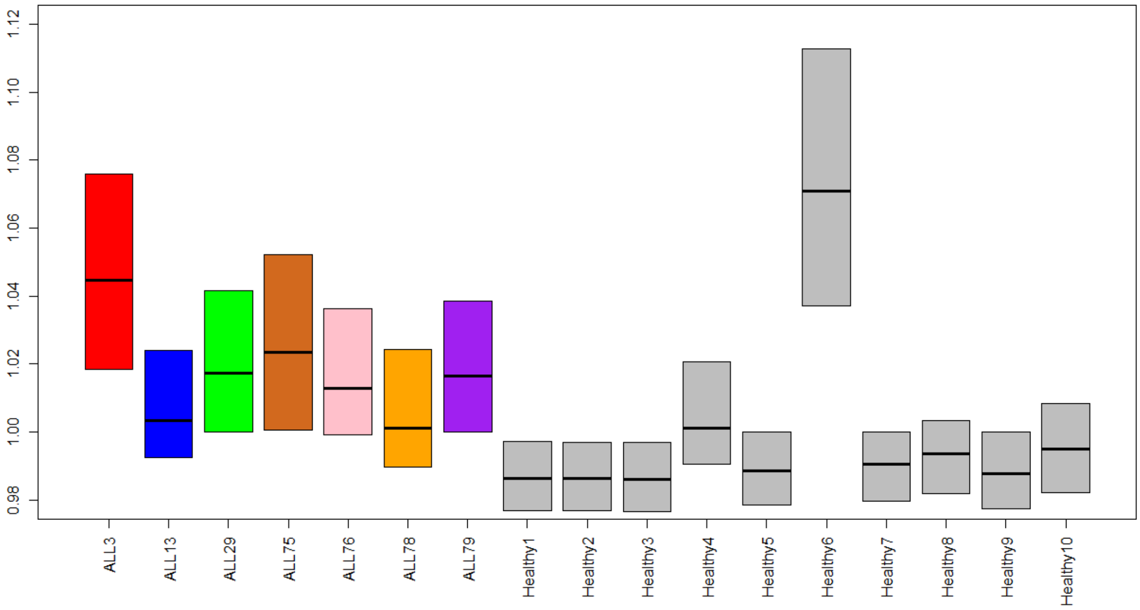


Figure 4.8 NUSE (Normalized Unscaled Standard Error) plot

In the overall assessment of the Affymetrix arrays above, it may be argued that one or two specific arrays tend to give reason for suspicion about quality in certain occasions. However, nothing has unanimously revealed; and thus, it is still premature to decide on either inclusion or exclusion of any array from the downstream analysis pipeline unless normalization is conducted, and post-normalization quality check is done on the arrays.

4.2.3.4 Between-Array Comparison

A heatmap plot is rendered in Figure 4.9, which records post-normalization between-array distances measured by their absolute median difference. In comparison to the earlier between-array test, this heatmap provides a re-organization of the arrays based on between-array distances computed through the arrays' median absolute difference after the process of RMA normalization. The figure does not reflect any potential issue with the normalized arrays.

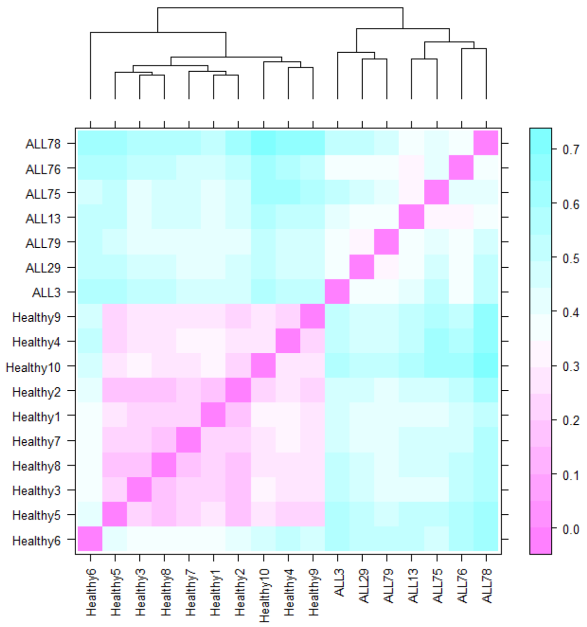


Figure 4.9 Heatmap of normalized Affymetrix data

Based on the overall post-normalization assessments carried out above, all the Affymetrix chips are found to be usable in the downstream analysis.

4.3 cDNA Data

Similar to Affymetrix data, exploratory data-quality analysis is also conducted on cDNA data at both pre- and post-normalization stage. Through this, anomalous array(s) would be identified while assessing the raw arrays; and later, after normalization, the array(s) that continues to behave as outliers would be dropped from the downstream analysis.

4.3.1 Assessment of Raw cDNA Data

4.3.1.1 MA Plots

To examine the imbalance between the red and green intensities in the data, a scatter plot of M and A values can be used. Such MA plot displays the log-ratio of red intensities, *R*, and green intensities, *G*, on the *y*-axis versus the overall intensity of each spot on the *x*-axis. The log-ratio, M is:

$$\begin{aligned} M &= \log_2 \mathrm{R} - \log_2 \mathrm{G} \\ &= \log_2 (\mathrm{R/G}) \end{aligned} \tag{17}$$

The average intensity, A, is measured by -

$$\begin{aligned} A &= \frac{1}{2}(\log_2 R + \log_2 G) \\ &= \log_2 \sqrt{RG} \end{aligned} \qquad (18)$$

The MA plot amounts to a 45^o rotation of the ($\log_2$G, $\log_2$R) coordinate system followed by scaling of the coordinates. Therefore, it is a representation of ($\log_2$G, $\log_2$R) data in terms of the log ratio, M. As any regression performed on the log-ratio (M) against average intensity (A) treats the two dyes equally, such regressions are more robust than regressions of logR on logG or logG on logR. MA plots also reveal more than normal scatter plots in identifying whether the red and green dyes respond differentially, and in a linear or non-linear fashion; and, based on that, a normalization method can be selected.

The two dyes ideally should behave in a similar fashion where the spots are symmetrically scattered about a horizontal line through zero, i.e., M=0; and in that case, no normalization is required. If the line is shifted up or down away from 0, a linear normalization by an amount equal to the shift away from the line, M=0, is required. Presence of a trend in the lower range of A usually indicates that the arrays have different background intensities, which may be addressed by background correction. A trend in the upper range of A usually comes from a systematic difference arising in the process of the microarray experiment. An overall non-linear scatter of data in an MA plot is often dealt with intensity dependent, non-linear normalization methods, such as the much advocated and Cleveland (1979)-proposed *robust locally weighted regression*.

Appendix B.1 shows the MA plots obtained from the cDNA arrays. The arrays are clearly not ideally scattered.

4.3.1.2 Array Intensity Distributions

A simple summary of the distribution of the probe intensities across all cDNA arrays is shown in Figure 4.10. Note here that a few of the arrays are repeats (non dye-swaps), and a patient's subsequent gene expression level would be an average of that patient’s available repeats. The in Figure 4.10 shows boxplot-distribution of green and red channel, along with their combined ratio-measures on $\log_2$-scale. Typically, one expects the boxes to have similar IQR (size) and median (*y*-position). The existing variations are expected to be minimised once the process of normalization completes.

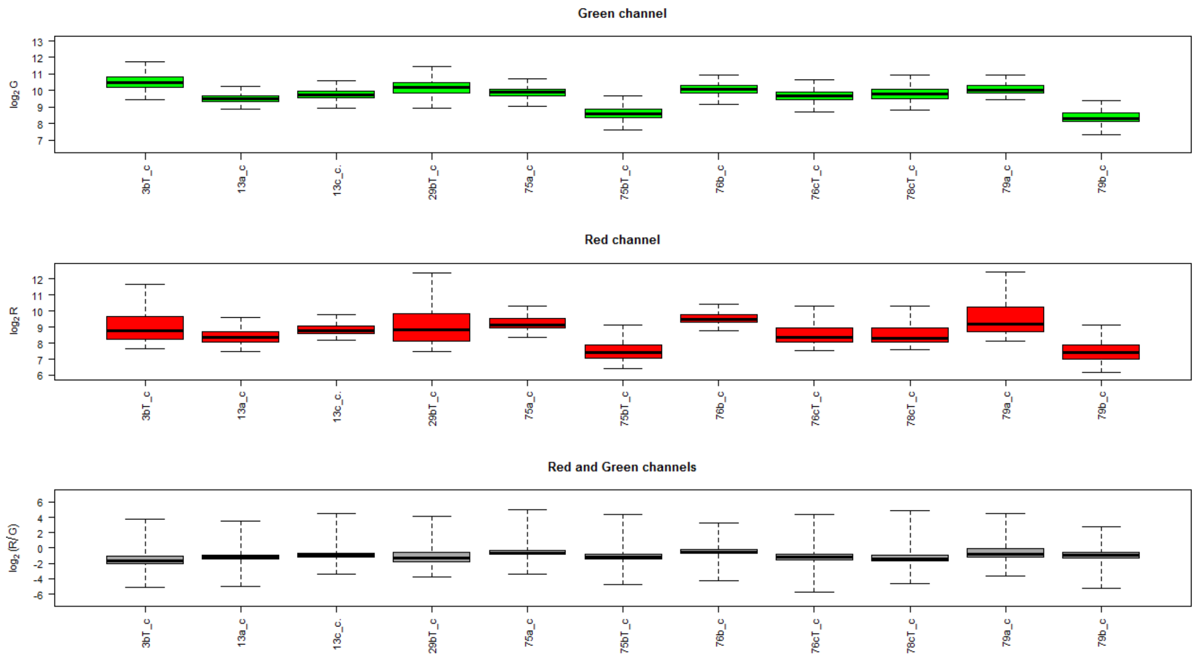


Figure 4.10 Untreated expression measures of green and red channels

4.3.1.3 Between-Array Comparison

Figure 4.11 presents a false colour heatmap of between-array distances of the raw, cDNA data. Table 4.1 provides the array-names corresponding to the array-numbers shown in the figure.

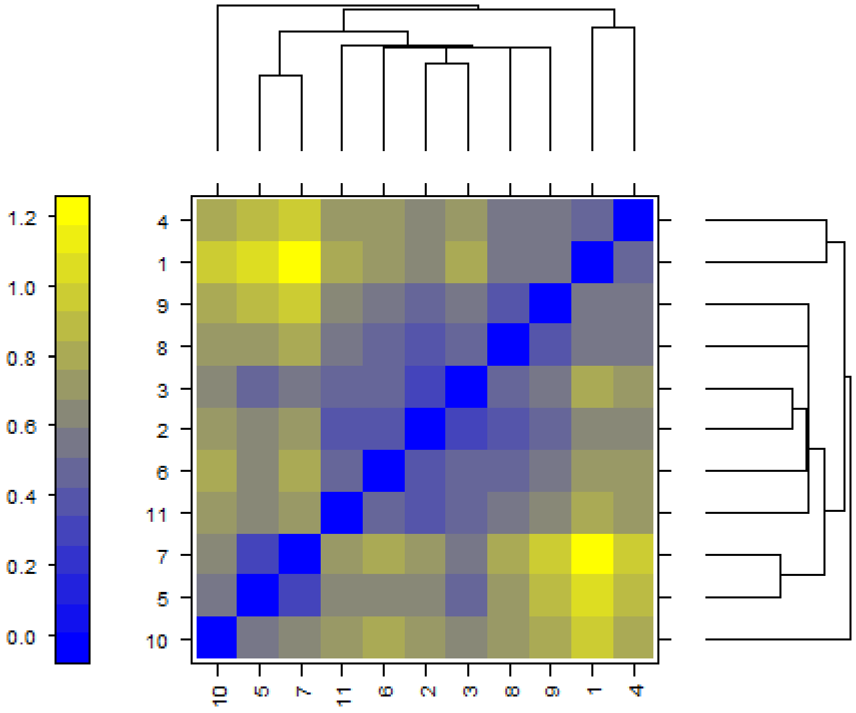
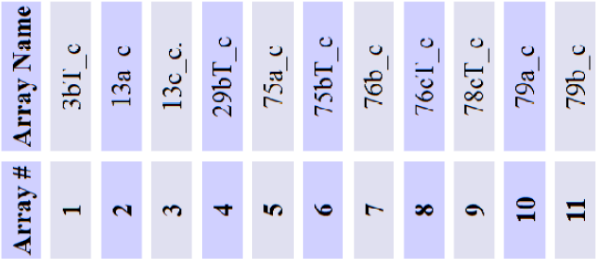


Figure 4.11 Heatmap of distances between the raw cDNA arrays

The heatmap is computed as the median absolute difference of the vector of M-values. The figure helps in deducing through visualizing that none of the arrays is an obvious outlier.

Table 4.1 cDNA array-numbers corresponding to the names



Array #	Array Name
1	3bT_c
2	13a_c
3	13c_c
4	29bT_c
5	75a_c
6	75bT_c
7	76b_c
8	76cT_c
9	78cT_c
10	79a_c
11	79b_c

4.3.2 cDNA Data Normalization

Prior to the normalization method, an adaptive background correction, *Normexp+offset* is used for the current *GenePix*-generated arrays, as recommended by Ritchie et al. (2007). It is an usual assumption in background correction of cDNA arrays that given the observed foreground intensities, R_f and G_f, background correction for two-colour microarray data allows the true signal to be estimated by subtracting the background from the foreground values, such that $R = R_f - R_b$ and $G = G_f – G_b$. The corrected intensities are then used to form the log-ratio of each dye's intensity, $M = \log_2 (R/G)$, and average log intensity, $A = ½ (\log_2 R + \log_2 G) = ½ (\log_2 RG)$, for each spot. The *normexp+offset* method of background correction is based on the normal and exponential convolution model previously used to background correct Affymetrix data as part of the RMA algorithm (R. A. Irizarry et al., 2003; McGee & Chen, 2006). Using this method, a convolution of normal and exponential distributions is fitted to the foreground intensities using the background intensities as a covariate, and the expected signal given the observed foreground becomes the corrected intensity. The corrected intensities, thus obtained, are positive, but may be close to zero. Therefore, a small positive offset is added to effectively move the corrected intensities away from zero. This should also reduce the variation of the low intensity M-values since $\log_2$ [(R+offset)/(G+offset)] will be close to 0 for R and G, both small relative to the offset. Based on the findings of Ritchie et al. (2007), an offset value of 50 is used here for background correction. The effect of background-correction for the cDNA arrays are shown in Figure 4.12. Comparing the two plots of the figure, the horizontal fanning-out of the red and green channels appears to have reduced by the background correction, besides shifting the corrected intensities away from zero.

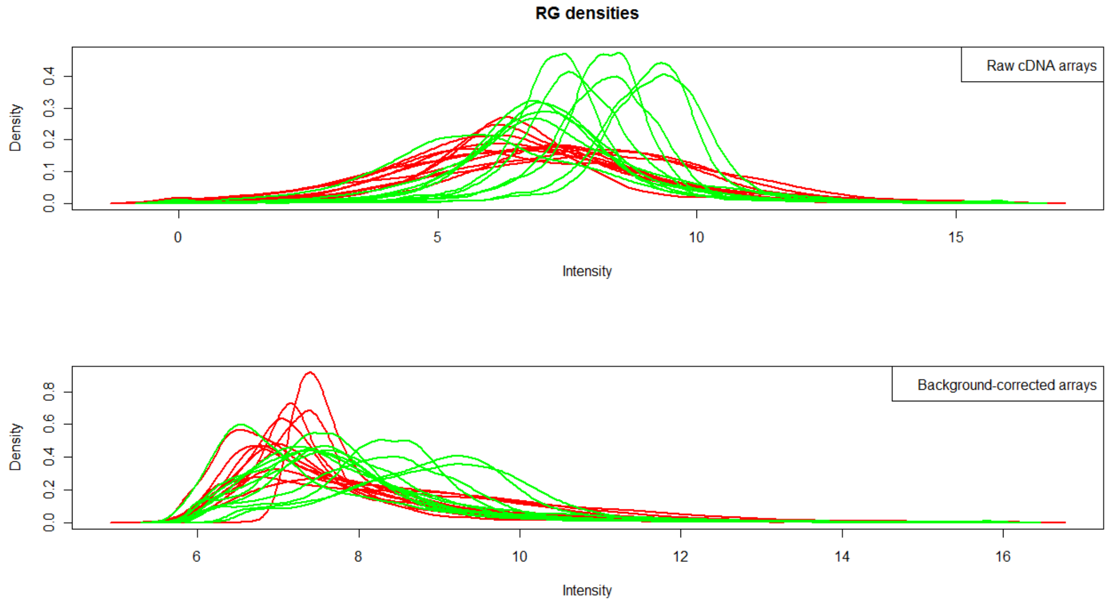


Figure 4.12 Effect of background-correction on red and green channels

As illustrated by Smyth and Speed (2003), there is a range of normalization methods for spotted microarrays, and these methods may be broadly classified into *within-array* normalization and *between-array* normalization. The former group includes those methods that normalize the M-values for each array separately, while the latter normalizes the intensities or log-ratios to be comparable across arrays.

Between-array normalization is only done when there are substantial differences between the cDNA arrays, giving them different spreads of M-values, usually for reasons including differences in print quality, differences in ambient conditions when the plates were processed or simply from changes in the scanner settings (Gordon K. Smyth & Speed, 2003). This method of normalization is usually, but not necessarily, applied after normalization within-arrays (Gordon K. Smyth, 2005). As it is not routinely done for two-colour microarray data, it will also not be attempted in this analysis unless there is good evidence of its requirement after within-array normalization.

A variety of methods have been developed for the normalization of two colour array data (Baird, Johnstone, & Wilson, 2004; Dabney & Storey, 2007; Tseng et al., 2001; D. L. Wilson, Buckley, Helliwell, & Wilson, 2003; Wit & McClure, 2004). The methods assume that the population to be normalized are roughly equally distributed, the number of genes differentially expressed is small, and the direction of expression is symmetric. The most popular method is lowess, aka loess, normalization utilizing local regression to fit each population (Seidel, 2008), and it has been found to be robust in simulated experiments even when 20% of the genes show differential expression in just one direction (Oshlack, Emslie, Corcoran, & Smyth, 2007). *Printtiploess* (Cleveland, 1979; Y. H. Yang, Dudoit et al., 2002) is a loess normalization method, and is reportedly found to perform best in studies such as (Hua, Tu, Tang, Li, & Xiao, 2008). This method is also regarded as an effective method because of its ability to adjust for systematic differences between different print-tips (T. Park et al., 2003; Tseng et al., 2001). Printtiploess normalization is selected here for normalizing the cDNA arrays.

Printtiploess is an average intensity, *A* [i.e., combined intensity of each dye, A = ½ ($\log_2 R$ + $\log_2 G$)]-dependent normalization, which is applied to the individual subgrids, the area of the cDNA array where all the spots were deposited by a single spotting-pin. It is regarded as an effective method for its ability to adjust for systematic differences between different print-tips (Insuk, Sujong, Changha, & Jae Won, 2008; T. Park et al., 2003; Tseng et al., 2001). It assumes that the printtip groups have the same distributions on each of the arrays, the red and green intensities are related by a constant factor, i.e. $R = kG$, and the center of the distribution of log ratios is shifted to zero. It is given in equation 19, where c = $\log_2$k is the median or mean of M (i.e., log-ratio of R and G) for a gene set.

$$\log_2 \frac{R}{G} \;\rightarrow\; \log_2 \frac{R}{G} - c \;=\; \log_2 \frac{R}{kG} \qquad (19)$$

As discussed by Yang et al. (2002), the lowess scatter plot smoother performs robust locally linear fits to the MA plots for the subgrids. This can be represented by equation 20, where *ci*(*A*) is the lowess fit to the MA-plot for the *i*th grid (i.e. for the *i*th print tip group), $i = 1, ..., I$, and *I* denotes the number of print tips.

$$\log_2 \frac{R}{G} \;\rightarrow\; \log_2 \frac{R}{G} - c_i(A) \;=\; \log_2 \frac{R}{k_i(A)G} \qquad (20)$$

The state of the arrays before and after printtiploess normalization (with background correction) is shown in Figure 4.13, where the plot with *Normalization: None* indicates that normalization as well as background correction is yet to be conducted.

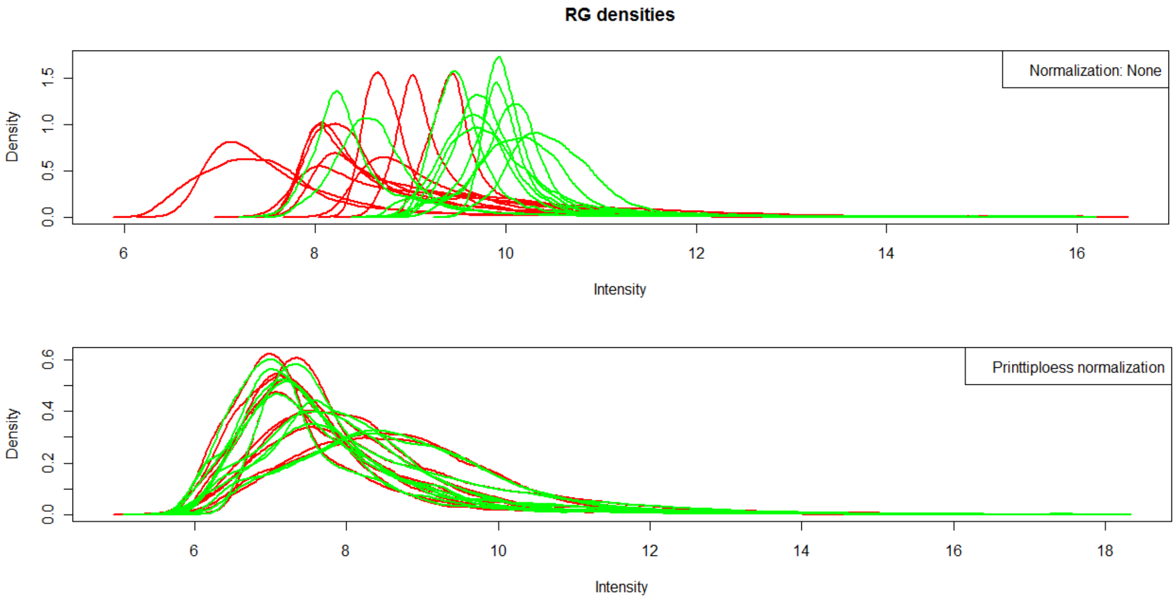


Figure 4.13 Printtiploess normalization on cDNA arrays

4.3.3 Between-Array Normalization

The intensity distributions across arrays are assumed to be the same, which is not always true. For the arrays to be comparable, the intensity distributions need to be similar. Printtiploess normalization conducted above does not affect the A values, and it normalizes the M-values for each array. This makes the red and green distributions essentially the same for each array. The next question is whether normalization is required between the arrays because there may still be considerable variation between the arrays. For this, Figure 4.14 is generated, which provides the distributions of the normalized M-values of the arrays. The figure indicates that between-array normalization may be required as different arrays are showing different spreads of M-values rather than an expected similar spread.

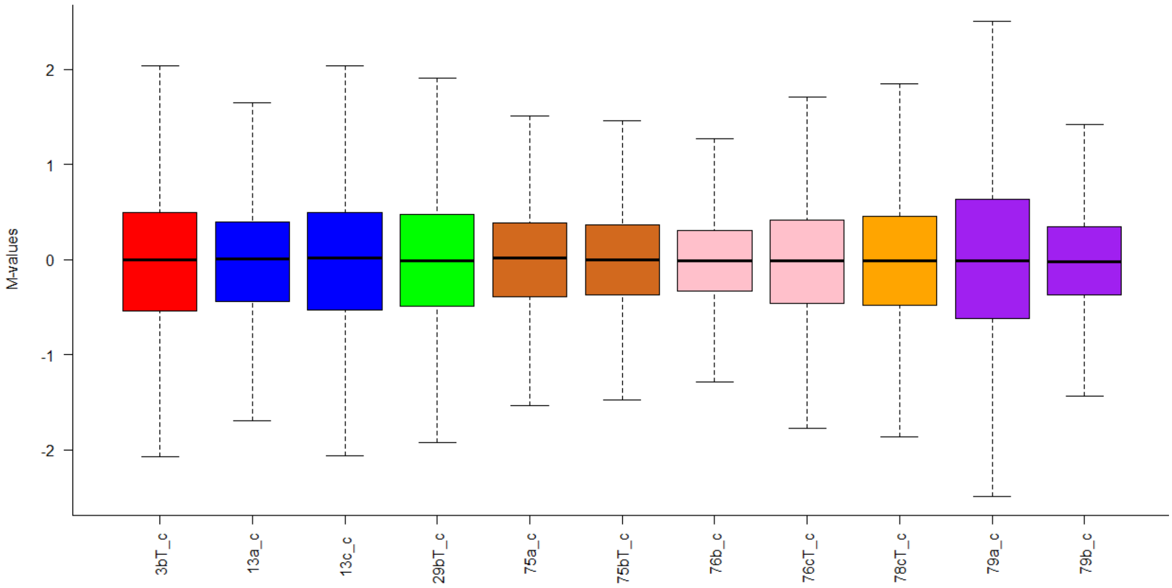


Figure 4.14 *M*-value distribution before between-array normalization

There are several between-array normalization methods including *scale, quantile* and *vsn*. The scale normalization method, proposed by (Y. H. Yang, Dudoit et al., 2002; Y. H. Yang, Dudoit, Luu, & Speed, 2001), and further explained by Smyth and Speed (2003), has rendered better result producing similar spread of the M-values across the cDNA arrays, as shown in Figure 4.15. The basic idea here in this normalization is to simply scale the log-ratios to have the same median-absolute-deviation (MAD) across arrays.

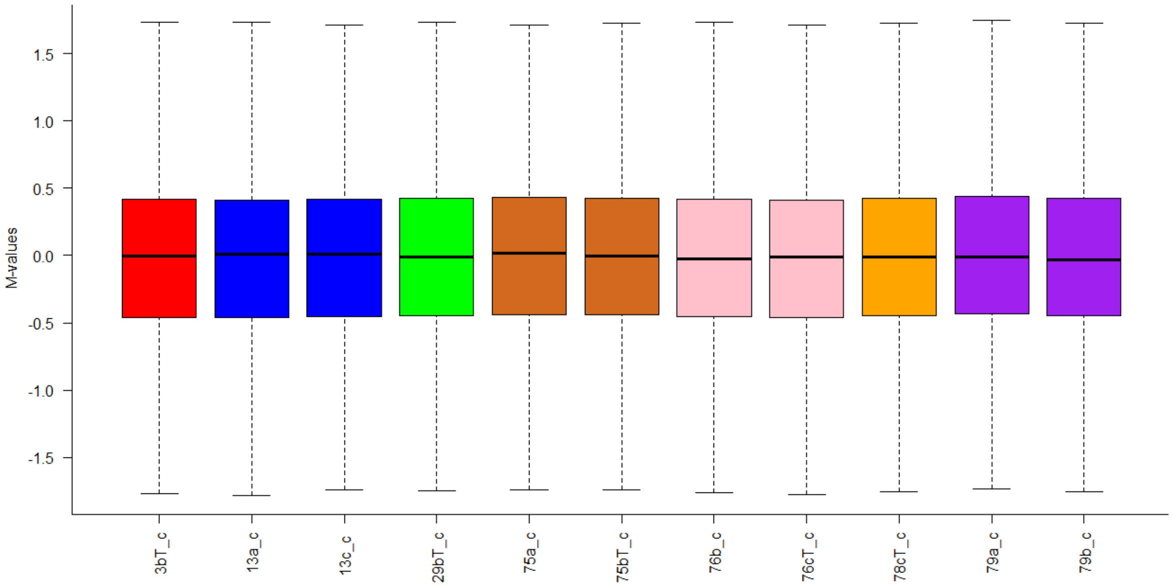


Figure 4.15 M-value distribution after between-array normalization

4.3.4 Post-Normalization Assessment

The overall effects of normalization on the spotted arrays have been assessed, and a few important ones are reported below.

4.3.4.1 MA Plots

Contrary to the MA plots of the raw cDNA data in Appendix B.1, MA plots of normalized arrays render better plots in Appendix B.2, where the mass of the data are desirably seen to be about the M=0 axis.

4.3.4.2 Array Intensity Distributions

Figure 4.16 reports post-normalization smoothed histograms of the spotted arrays. Comparing the arrays in the earlier states, the arrays tend to lack varying distributions with lesser fanning-out of the red and green channels.

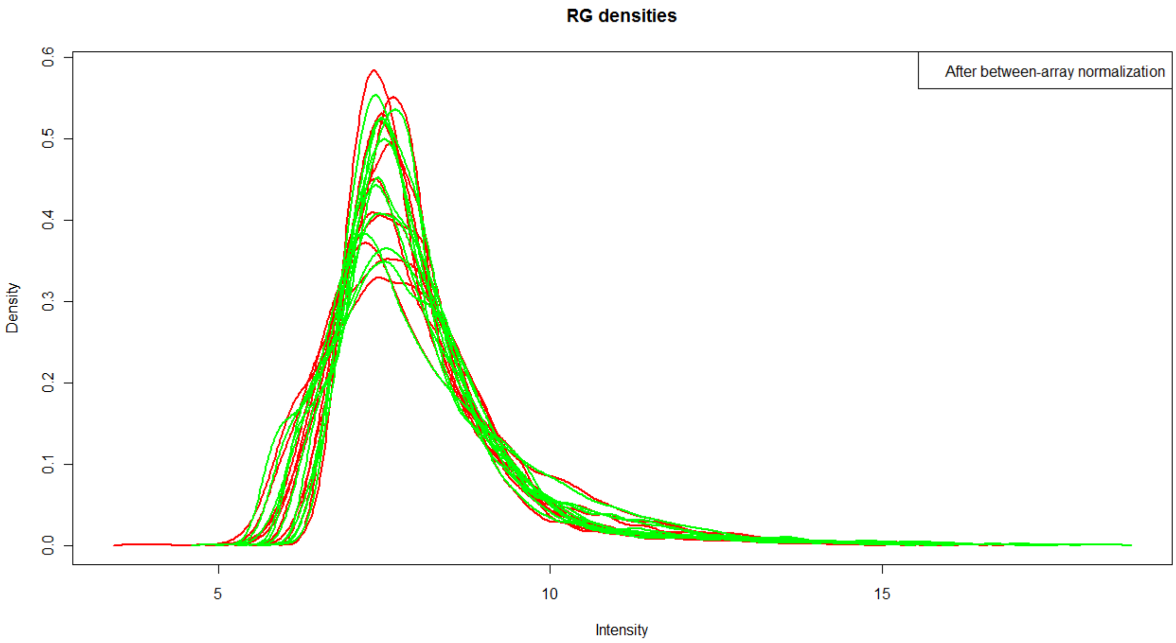


Figure 4.16 Post-normalization density estimates of cDNA arrays

4.3.4.3 Between-Array Comparison

Figure 4.17 gives a heatmap of between-array distances. The distances between the arrays are found to have reduced in this plot, besides there seems to have no outlier array as none of the arrays has an exceedingly large distance from the rest.

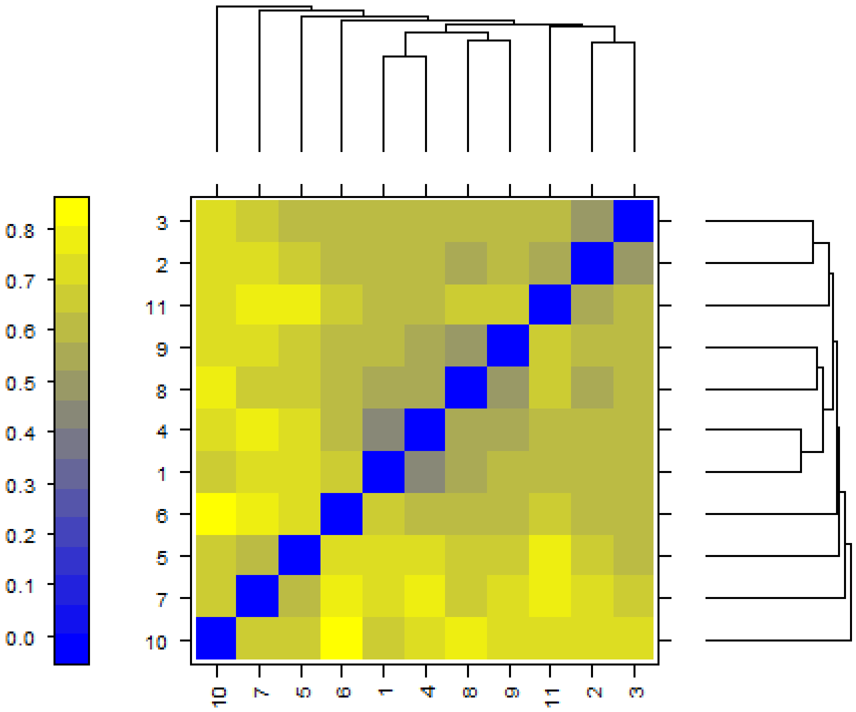


Figure 4.17 Post-normalization heatmap of cDNA arrays

Overall, it is seen from this investigation relating to Affymetrix and cDNA data that there are arrays that tend to behave undesirably at the pre-normalization stage. However, the respective normalization method has removed the bias making the concerned arrays homogeneous.

Chapter 5
Transformation of Expression Data

5.1 Finding Differentially Expressed Genes

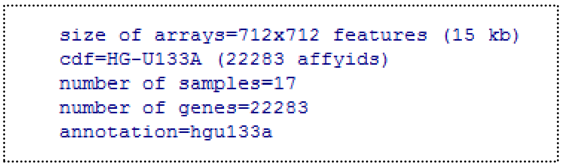
Genes that show little variation between samples are very unlikely to hold useful information. The differentially regulated genes tend to vary between the conditions specified, and are considered important towards revealing information. The specified conditions with regards to this project are the children with leukaemia and the healthy children as the reference condition.

Ideally, the set of differentially expressed (DE) genes should remain the same for investigations conducted in multiple microarray platforms. However, this does not happen in practice as the intensity values are generally affected by various sources of noise and fluctuations. In cDNA platform, the problem of noise is higher than the platforms like Affymetrix because the former has more scope for noise to be introduced from the stage of array-construction upto scanning of the images. It is also reported by Lee et al. (2000) that in cDNA, the probability that a single spot will display as a signal even if the mRNA is not present is as large as 10%, whereas non-displaying of a signal while a spot does contain complementary DNA remains at a non-negligible probability of about 5%. Moreover, in comparison to the oligonucleotide libraries, there are concerns involving the probe contents of cDNA libraries about annotation, clone identity, and probe performance (Woo et al., 2004). However, this does not mean that Affymetrix platform is free from flaws. It too has issues such as non-specific hybridization and less than optimal choice of the oligonucleotide sequences representative of a gene. Nevertheless, the concerns with cDNA arrays often come up more predominantly contributing to the fact that they have issues with reliability and that the DE genes do not necessarily match in identical microarray investigations.

In this context, it is decided to rely more on the normalized Affymetrix arrays to select the list of differentially expressed genes. The same set of genes from the cDNA platform will then be extracted, and be considered as the genes of interest for this platform.

The normalized gene expression data from Chapter 4 are used in the current process of retrieving the DE genes.

With 17 Affymetrix chips (of 7 leukemic and 10 healthy children), the number of available genes found is 22,283.



```
size of arrays=712x712 features (15 kb)
cdf=HG-U133A (22283 affyids)
number of samples=17
number of genes=22283
annotation=hgu133a
```

Like most other array manufacturers, Affymetrix includes a number of control probes on their arrays. A set of 68 such control probes is removed reducing the total number of available genes to 22,215. From these genes, differently expressed genes are to be retrieved.

There is a plethora of approaches for finding DE genes. Fold change method is one of the simple and intuitive methods, where at least two- to three-fold difference between the conditions - control and experiment, is considered significant (J. DeRisi et al., 1996; J. L. DeRisi et al., 1997; C. H. Jiang, Tsien, Schultz, & Hu, 2001; Wellmann et al., 2000). However, this highly used method has serious drawbacks, including the fact that the arbitrarily-chosen fold-threshold can often be inappropriate. Further, applying constant threshold for the fold change of all genes, false-positives are generated at low-intensities reducing the specificity while sensitivity is reduced at high intensities by missing the true positives.

Alternatively, the second widely used method, called *unusual ratios*, considers the distribution of measurements within the data. Used in many studies such as Schena et al. (1995), Schena et al. (1996), Tao et al. (1999), this method involves selecting those genes with experiment-to-control ratios at a specified distance, usually ±2 standard deviations away from the mean experiment-to-control ratio. The intrinsic drawback of this method is that it always reports a fixed proportion threshold, i.e., 4.6% of the genes as differentially expressed even if the set actually contains a greater or lesser proportion of truly-regulated genes (S. Draghici, 2002; Sorin Draghici, 2005; Zhang, 2006).

To estimate variability of the normalized dataset of this project, a sample-to-sample comparison is considered a relatively unbiased method. Again, instead of simple standard deviation across all samples, which can potentially introduce intensity-dependent bias, *relative standard deviation* (also, known as the *coefficient of variability*) is accepted here as a better option. Along with this, a statistical *false-discovery rate*-component is also integrated, which will be subsequently followed through in the succeeding description.

The coefficient of variability, CV-filter measures the variability of a gene across all experiments. It is calculated as the gene's standard deviation across all samples divided by the mean. High CV-value reflects high variability of genes among the samples and between the conditions - control and experiment.

To filter out the least variable genes out of the remaining genes that are free of control probes, 90th percentile of the distribution of CV-values are selected. Figure 5.1 shows the chosen cut-off that picked the highest ranked 10% of CV-values.

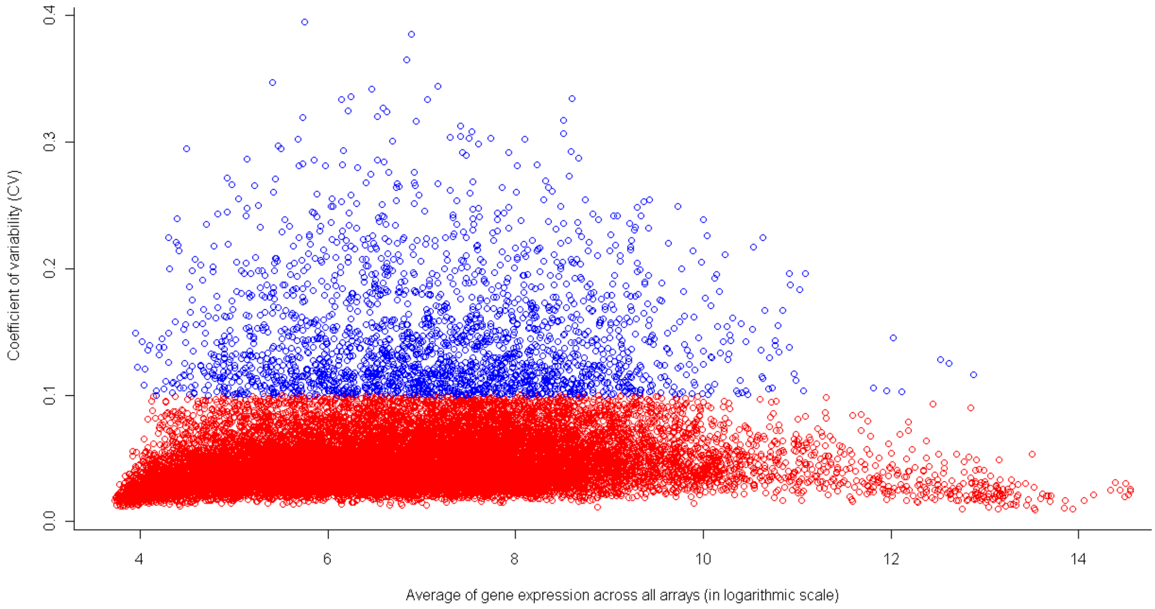


Figure 5.1 CV as a function of average gene expression across Affymetrix arrays

A histogram is used in Figure 5.2 to show the distribution of the overall data prior to filtering of the least variable genes. It indicates a highly skewed distribution, which is adjusted upon log-transformation, with the cut-off clearly separating the bulk from the highest CV-values. Judging by the relatively even distribution of high CV-values across the expression range, there should not be any significant bias introduced by the filtering.

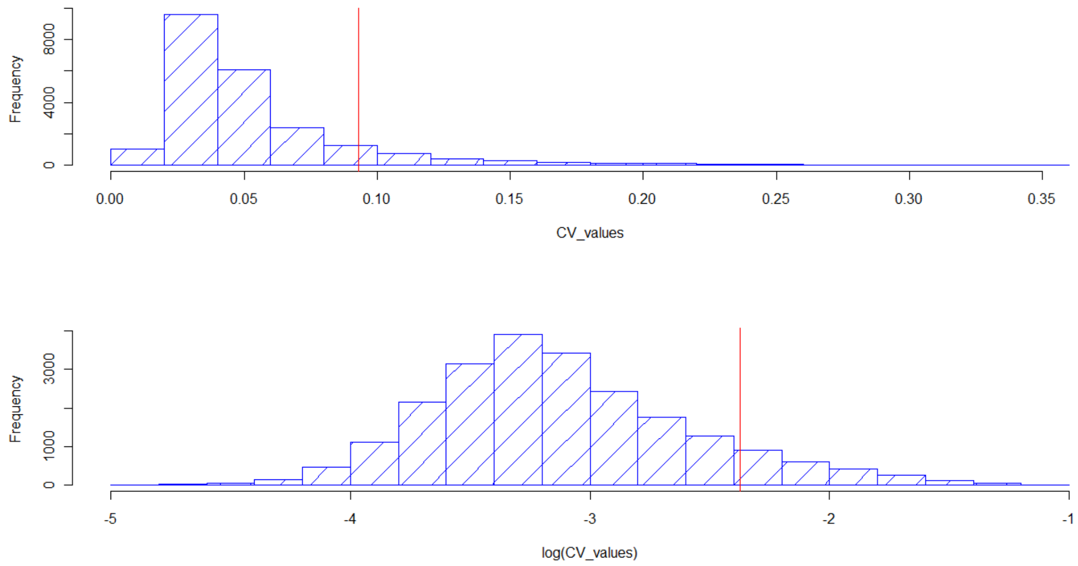


Figure 5.2 Linear and logarithmic CV-values with filtering cut-off

Finally, the filtering out of the uninteresting genes reduced the total number of highly variable genes to 2,222.

To the shortened list of genes belonging to the two experimental conditions (healthy and leukemic), an empirical Bayes method (G. K. Smyth, 2004) is applied. This is an adaptive strategy towards increasing statistical power, and simultaneously reducing the risk of false positives. The method stabilizes the variance estimates in such a way that if the estimated sample variances are not very different, the empirical Bayesian (EB) model arrives at essentially a pooled estimate; and if the variances are very different, the model shrinks the dispersions to a lesser amount. As Robinson & Smyth (2007) describes, the EB rule works well in practice and renders increased precision in estimating dispersion, which leads to gain in power for testing between experimental conditions. For the microarray dataset of this project, the EB method is expected to improve on the accuracy of estimating variability for individual genes through shrinking of the standard deviation by including genes expressed at similar levels of expression in both patients and controls. The p-values, subsequently obtained, need to be adjusted to account for the *multiple testing* (or, *multiple comparisons*) *problem*.

As Miller (1981) illustrates, multiple testing problems bring in error in inferences when a set of statistical inferences are considered simultaneously; and, loss of statistical power in inference imposed by the multiple testing is common during simultaneous analysis of thousands of genes. The popular method of Benjamini & Hochberg (1995) is used here that adjusts p-values for multiple comparisons; however, there are other methods on offer including Hommel (1988), Holm (1979), Hochberg (1988) and Benjamini & Yekutieli (2001). Benjamini & Hochberg (1995) method controls the *false discovery rate* (FDR), the expected proportion of the significant results that are in fact type I errors ('false discoveries') amongst the rejected hypotheses in multiple comparisons. The false discovery rate is a relaxed condition; and the Benjamini & Hochberg's method is a better compromise between sensitivity and specificity as it controls the proportion of false significant results instead of controlling the chance of making even a single type I error. For the current data set, FDR control is set to a conservative value of 0.05.

Figure 5.3 presents a histogram of the raw, unadjusted p-values, and compares the distribution to that observed after adjustment to account for multiple testing correction. It also shows how the distribution would be if there were no experiment effect (i.e., a uniform distribution), besides indicating the cut-off for the statistical significance, i.e., FDR control=0.05. The clear deviation from the uniform distribution indicates that there is indeed a strong experiment effect, and that the p-values of the genes vary. Although adjusting for multiple testing substantially shifts the lowest p-values to less significant levels, there are still a sizeable proportion of p-values that fall below the significance cut-off of 0.05.

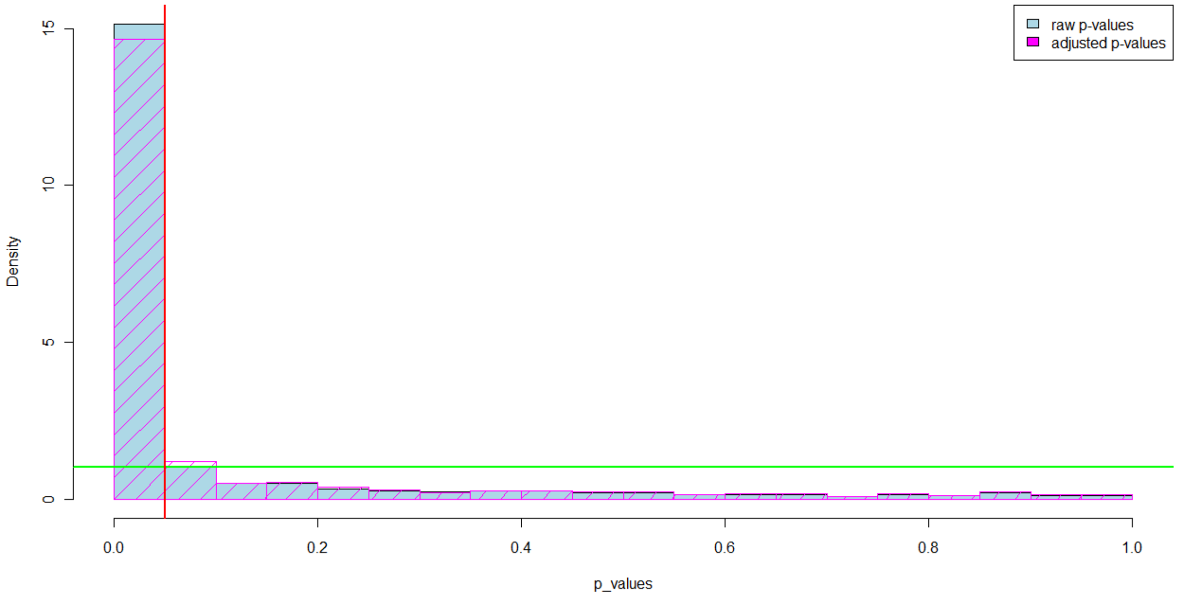


Figure 5.3 Distribution of raw and adjusted p-values

[The horizontal and the vertical line is the theoretical uniform distribution and the false discovery rate cut-off at 0.05, respectively]

In Figure 5.4, an MA-plot displays the log fold change between leukemic and normal samples as a function of the average expression level across all samples, where the two-fold limits are indicated by horizontal lines. Similar information is also displayed using a Volcano plot in Figure 5.5, which is constructed by plotting the negative logarithm of the p-values as a function of the base 2 log-transformed fold changes. Here, the statistically significant genes are highlighted with sharp blue circles and 2-fold limits are symbolized by vertical lines. The statistical significance cut-off (0.05) is overlaid as a horizontal line.

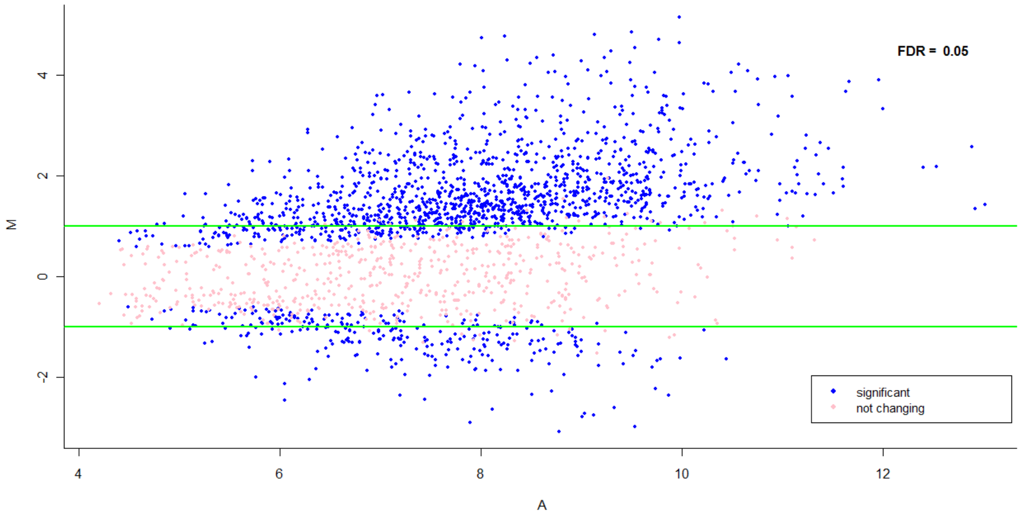


Figure 5.4 MA-plot comparing healthy and leukaemic samples

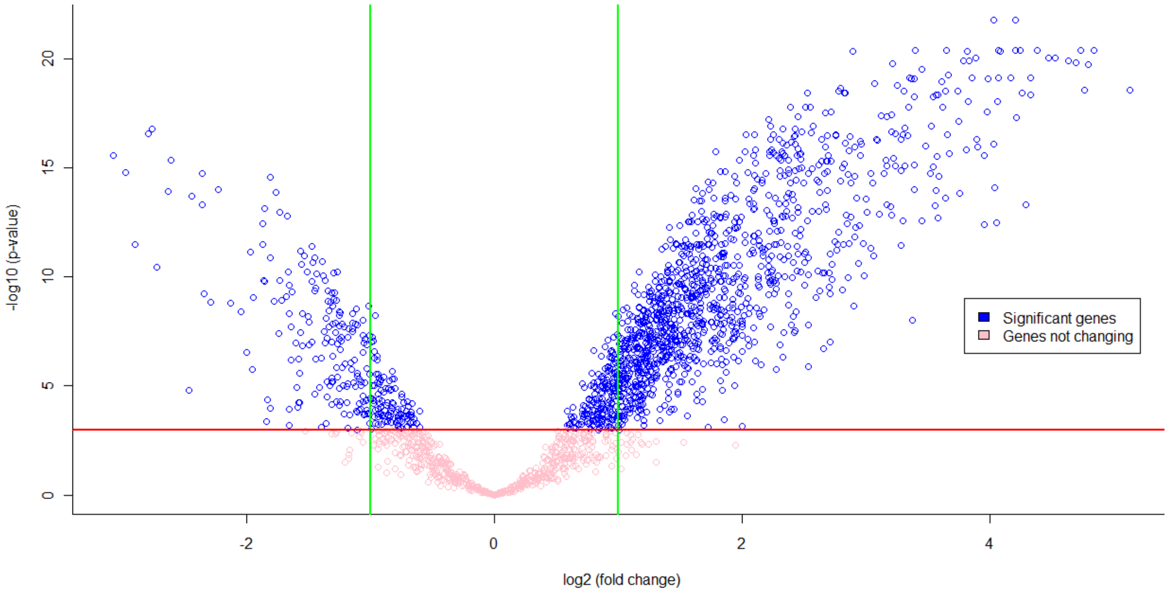


Figure 5.5 Volcano-plot of the comparison between healthy and diseased samples

Finally, it is found that the overall procedure on Affymetrix chips has picked a total of 822 genes as differentially expressed. These genes belonging to the 7-patients are overlaid, and shown as a scatter plot in Figure 5.6.

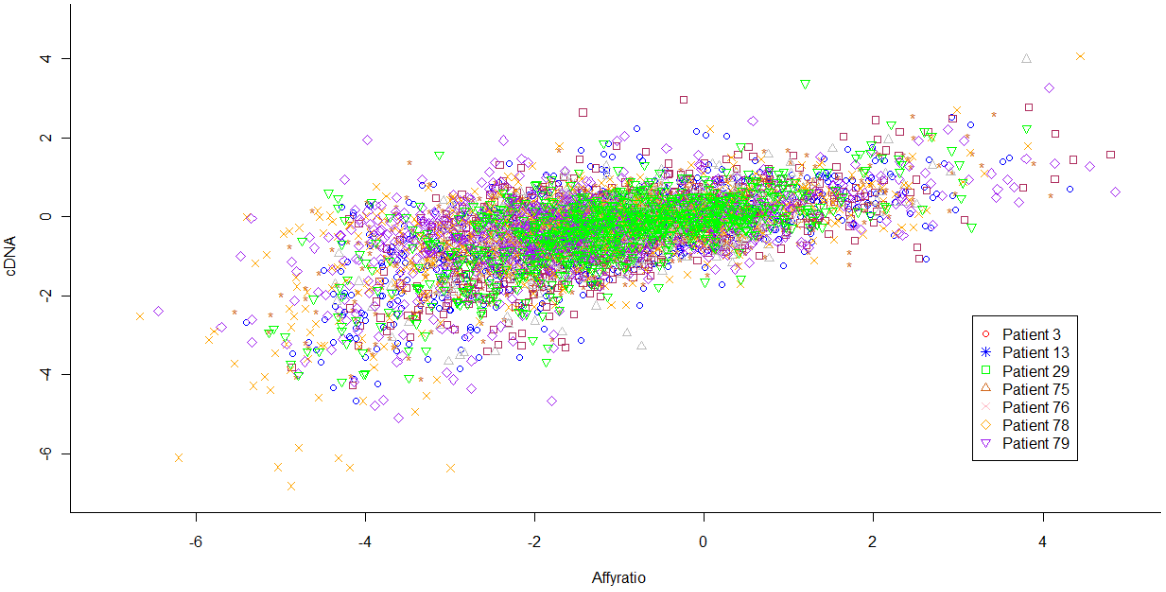


Figure 5.6 Scatterplot of significant genes from 7-patients

5.2 Ratio-Transformation

UniGene database (Wheeler et al., 2000) is used to annotate the retrieved 822 differentially expressed genes.

Affymetrix data contain relatively lesser noise than cDNA, and various issues affecting the cDNA platform have been discussed earlier. Considering this fact, the same set of 822 genes from our cDNA data is also retrieved to use in the downstream analysis. It is assumed here that as the arrays in both platforms belonged to the same 7-childhood leukaemia patients, the same set of genes would ideally be expressed differentially in either platform.

A known fact for Affymetrix and cDNA data is that they invariably do not hold any relationship between them at all. This once again proves to be true with regards to the original microarray datasets of these 7-childhood leukaemia patients. The data obtained for these patients from Affymetrix and cDNA platform bears absolutely no relationship.

Once the DE genes are obtained, the correlation (Pearson product-moment correlation coefficient, *r*) between the data from both platforms is again tested, and the result is found to be 0.13. This indicates that there is still no relation between them; however, this value shows certain improvement over the result obtained in the earlier test with regards to the whole dataset of both platforms.

Fundamentally, Affymetrix and cDNA data have difference in their data structure. cDNA gene expression data is represented using a measure of relative expression, which is expressed in terms of *expression ratio*. As shown in Equation 1 of Chapter 2, it is a ratio between the expression intensity metric for any tumour sample to the respective healthy sample. However, the value that is usually taken as representative for the expression level of a gene in Affymetrix platform is the average difference between all the PM and MM probes (Equation 3, Chapter 2). This apparently differentiates the nature of the generated data from the two platforms in the sense that - while cDNA produces expression ratios for its genes, Affymetrix renders actual expression measures of the genes. This basic difference in the nature of the generated data is neither new nor has this been unknown to the users since the launching of these two platforms. However, hardly any information could be gathered from the literature to suggest either exploration has been carried out based on this primary difference in the nature of the data or any attempt has been made to check whether investigating on this difference could lead to addressing the relationship between the two platforms. Adhering to this lack of information as a motivation in the backdrop, the task aimed ahead is to mitigate the difference between the 7-lieukaemic patients' data obtained from the cDNA and Affymetrix platform and to examine whether it brings any improvement. To do this along these lines, the datasets from the diverse platforms must be transformed in some way so that both find a common and comparable ground.

As cDNA and Affymetrix data are expression ratios and actual expression measures respectively, the rational way of transformation would be either to convert the cDNA dataset to actual expression measures similar to Affymetrix data, or ratio-convert the Affymetrix dataset.

As mentioned earlier, there are 10 Affymetrix arrays available, which belong to the same number of healthy children. The set of 822 DE genes found in the leukemic children are also identified in each of these healthy arrays. Previously in section 4.2.2: *Affymetrix Data Normalization*, the RMA normalization produced $\log_2$ expression measures for all Affymetrix arrays, i.e., for both healthy and leukemic patients. The $\log_2$ expression values of 822 DE genes belonging to the healthy Affymetrix arrays are now converted to their respective anti-logs, and each gene's expression value is averaged across these 10 healthy arrays. It gives rise to a single, averaged and log-free expression value for each of the 822 genes. Simultaneously, the expression antilogs of each of the 7-leaukaemic patients' DE genes are calculated. Then, the *Affy_{ratio}* for a gene of a patient can be found by dividing the calculated expression value by the corresponding gene's average antilog value from the healthy Affymetrix arrays, and subsequent $\log_2$-conversion of the obtained value, as shown earlier in equation (1). This assures that similar to cDNA, where the expression level of a gene remains in the form of a tumour-to-healthy ratio, this transformation converts the Affymetrix expression data into tumour-to-healthy ratios.

The overall formulation of $Affy_{ratio}$ can be presented by equation (21) if expression level of a gene, *x* from one of the diseased Affymetrix chips is *D* and the average of this gene's expression from the set of 10 healthy Affymetrix chips is *H*.

$$Affy_{\text{ratio}} = \log_2 \frac{Anti\log\,(D_{x_i})}{\frac{\sum_{x=1}^{10} Anti\log\,(H_{x_i})}{10}} \qquad (21)$$

With this changeover implemented, both cDNA and $\text{Affy}_{\text{ratio}}$ data can be, in theory, considered to have reached a mutually comparable level. However, it is necessary to check what practical impact this has caused on the overall relationship with regards to the pair of datasets (i.e., Affymetrix and cDNA-pair) before and after the transformation.

It is already known that both datasets initially had no correlation between them; and with DE genes, it increased to 0.13. Therefore, keeping 0.13 as a benchmark to evaluate whether the process has caused any change in the relationship between the datasets, the correlation between $\text{Affy}_{\text{ratio}}$ and cDNA is tested. In results, it is found that the correlation between the $\text{Affy}_{\text{ratio}}$ and cDNA has increased considerably to 0.6, which is, in effect, an approximately 6-fold improvement from the previous result. So, from the viewpoint of correlation, this change is substantially positive as it has catalysed the earlier relation to attain a six-fold increment. However, important questions are simultaneously raised such as how the overall distribution of the data is affected by the process, and whether the induced transformation has caused any unwanted alterations within the dataset.

Towards answering the questions, distribution of the original Affymetrix (contains the prefix, *ALL*) and cDNA (with the prefix, *cDNA*) data, along with the $\text{Affy}_{\text{ratio}}$ data for the seven different leukemic children are plotted, as shown in Figure 5.7. In comparison to the original

Affymetrix (Affy$_{original}$) data, the plot indicates that the transformed Affymetrix data (Affy$_{ratio}$) align more closely with the cDNA than the Affy$_{original}$.

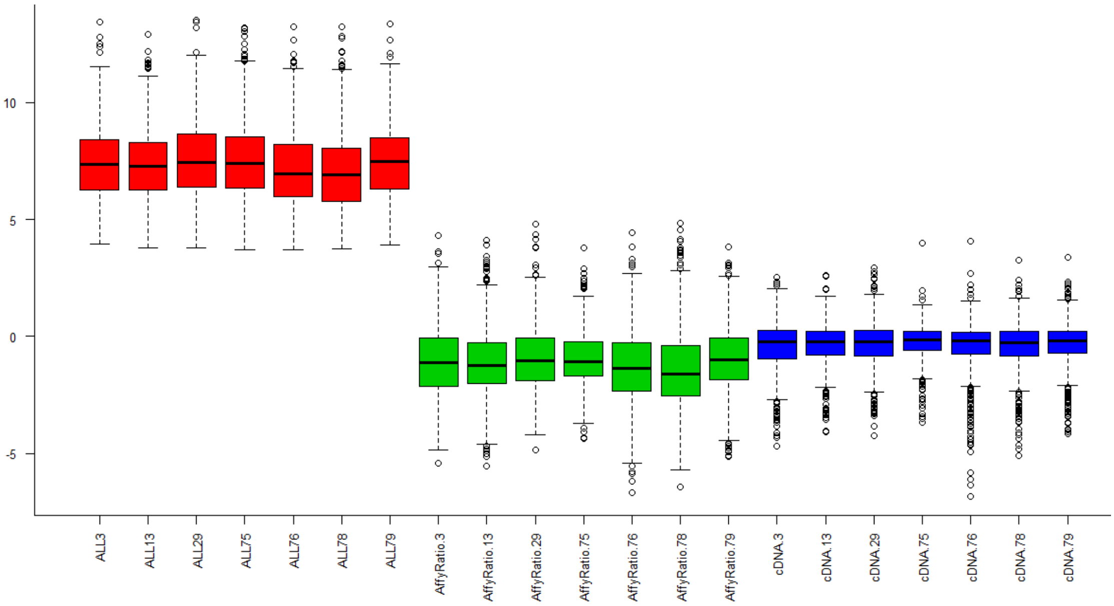


Figure 5.7 Array distribution before and after ratio-transformation

At this stage, it is intuitive to ponder on whether the changes introduced into the original microarray data have brought in any unwanted alteration to the overall state of the dataset. Further to that, if the integrity of the data is found to be unviolated, then the next important query that comes up is where this current approach stands in the midst of other microarray data merging methods. To examine these aspects, it becomes necessary now to carry out certain validation as well as evaluation tests.

5.3 Method Validation and Evaluation

Hierarchical clustering (S. C. Johnson, 1967) is useful to find the closest associations among gene profiles under evaluation where it seeks unsupervisedly to build a hierarchy of clusters based on relatedness. Whether any unwanted change has been caused to the microarray data through the process of ratio-transformation can be evaluated through hierarchical clustering. The method when applied to the pre- and post- transformed microarray data would highlight if any change has occurred to the overall state of the data.

With Euclidean distance and Ward's agglomerative procedure (Joe H. Ward, 1963), a divisive hierarchical clustering is conducted on the Affymetrix genes before transformation and another on the transformed data. The result is unable to present any unwanted variation as shown in Figure 5.8.

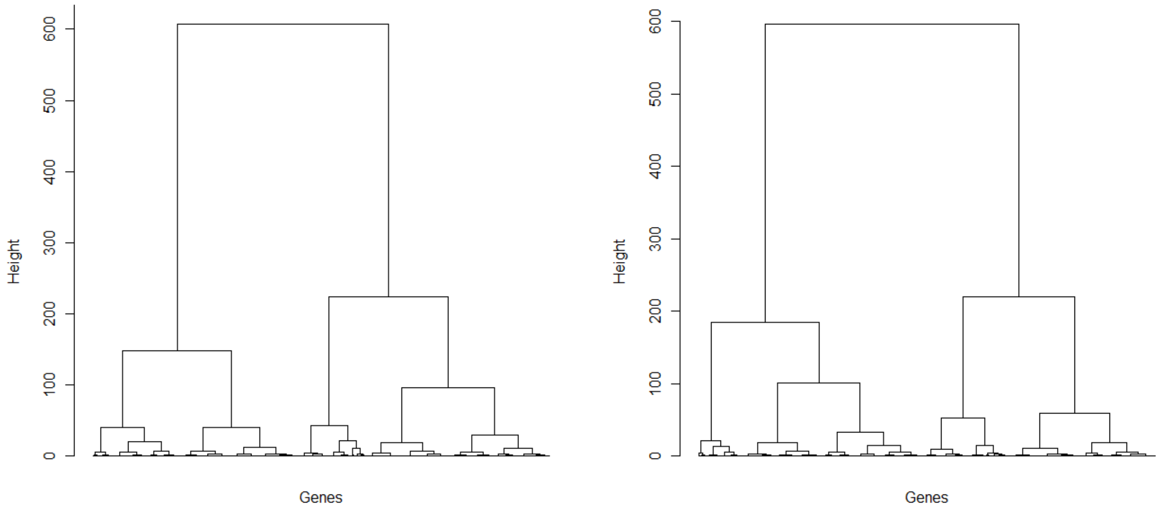


Figure 5.8 Hierarchical gene clustering of Affy$_{\text{original}}$ (left) and Affy$_{\text{ratio}}$ (right)

A similar hierarchical clustering is also applied to the patients to check whether the method has caused any change in the relationship among the patients. The outcome of this test also fails to substantiate that the change caused to the data has altered any relative relationship between the patients. Divisive hierarchical clustering of the patients is shown in Figure 5.9.

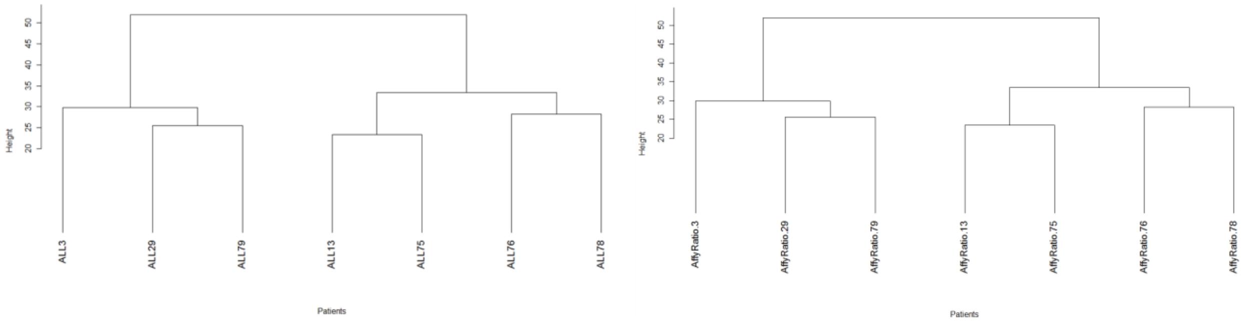


Figure 5.9 Hierarchical patient clustering of Affy$_{\text{original}}$ (left) and Affy$_{\text{ratio}}$ (right)

Both gene- and patient-clustering conducted above can be used to confirm that the overall relationship in the microarray data has not been violated due to the transformation method.

Next, as the consistency of the data is found to be unviolated, it becomes intriguing to evaluate where the current approach stands in the midst of other microarray data merging methods. The process of sample standardization and gene centering is an approach which in practice reportedly performs as well as a data merging approach (W.P. Kuo et al., 2009; Shabalin et al., 2008; Simon et al., 2004). The ability of the data-transformation method can be evaluated with this approach.

Using the method, each microarray sample is first standardized; and, if there is variation in the range of data between the samples from both the platforms, then gene-centred. However, it is difficult to judge how much variation is considered appropriate; and therefore, gene-centering is done once with sample standardization, and once without it.

In classical statistics, one of the fundamental distributions is the *normal distribution* or the *Gaussian distribution*. The probability density function for the normal distribution having mean, μ and standard deviation, σ is given by equation 22.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad (22)$$

Each microarray samples from either platform can be standardized by making $\mu = 0$, and σ = 1 in the probability density function. This gives the probability density function for the standard normal distribution as shown in the equation 23.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \qquad (23)$$

Once the samples are standardized, each gene belonging to each study is centred. As the genes are arranged in the rows of the dataset while the columns contain the various samples, the gene centering is done by subtracting the row-wise mean from the values in each row of data, so that the mean value of each row becomes zero. The samples from multiple platforms can subsequently be merged as sample standardization followed by centering of each gene in each study is done.

The method is applied to the normalized Affymetrix and cDNA data. The Pearson correlation coefficient is found to be -0.02615, which explains that the method does not improve the correlation between the two datasets.

Further, only gene centering is applied to the dataset of each platform. This time, however, the correlation coefficient is found to have increased to 0.46. This implies that for the microarray data, the sample standardization is not required, instead only gene centering improves the relation. However, the value of this correlation still remains below the result obtained from the ratio-transformation method.

Distance Weighted Discrimination or *DWD* (Marron, Todd, & Ahn, 2007) is a method, which is used by Benito et al. (2004) for batch correction and adjustments in biases including across microarray platform effects. It is based on modern statistical discrimination methods and has reportedly been effective in removing biases present in a breast tumour microarray data set. The method progresses by finding a direction, *DWD direction*, in which the sample-vectors from two studies are well-separated. It then translates the samples from each study along that direction until their respective families of vectors have significant overlap. This shifting each study's samples in DWD direction helps to remove the biases. To evaluate the relative standing of ratio-transformation method, the DWD statistical correction algorithm is applied to the normalized datasets of Affymetrix (HG-U133) and cDNA belonging to the seven leukemic patients. The resultant data is found to have a correlation of 0.77. The post-ratio transformed microarray data gave a correlation of 0.6. Although, unlike ratio-transformation, DWD method uses distance measures, there is an improvement in the latter method of merging the two sets of microarray data.

To compare further with other methods, approaches including *XPN* (Shabalin et al., 2008) and *Probability of Expression method* (Parmigiani et al., 2002; Shen et al., 2004) have been explored. However, it is experienced that such methods are not suitable for relatively smaller sample size. In a personal communication, Andrey Shabalin confirms in this regard that his team's XPN method does not work for smaller sample size. This issue may again be considered as a negative aspect for such methods that they can only consider data with large sample size.

With regards to the gene-centering and DWD methods, the method of ratio-transformation can be ranked in between DWD and gene centering method.

In summary, the ratio-transformation process highlights that its usage can address the issue of incomparability of expression data from Affymetrix and cDNA platform. The outcome of the above method is encouraging considering the fact that Affymetrix and cDNA expression data otherwise always remain incomparable. The encouraging outcome inspires to focus attention towards examining further in the direction of possible association between the two platforms. With this motivation, downstream analyses are taken up that are described as well as probed into in the following chapter.

Note :

- Aspects of this chapter have been oral-presented and published -

Sarmah, C. K., Samarasinghe, S., Kulasiri, D., & Catchpoole, D. (2010). *A simple Affymetrix ratio-transformation method yields comparable expression level quantifications with cDNA data*, in: C. Ardil (Ed.) International Conference on Bioinformatics and Bioengineering, World Academy of Science, Engineering and Technology, Cape Town, South Africa, vol 61, pp. 78-83.

Chapter 6
Formation of a Crossover

While studying microarray literature, it is often observed that a study of merging cross-platform data excludes the scope of exploring how various statistical and/or machine learning approaches would tend to contribute in defining the relationship of the data of the diverse platforms. Introducing an approach to merge Affymetrix and cDNA data in Chapter 5, here the aspect of using and comparing a wide range of statistical as well as machine learning methods are attempted in this direction. The succeeding sections would focus on examining these attempts and their relative effectiveness in the hope that it also would overall contribute subsequently to broadening the usual scope of such cross platform studies.

Each of the seven leukemic patients' data from either platform is examined here to be modelled and tested for their ability in predicting the outcome for the remaining patients. These entire data are also concatenated in two variables, viz. $\text{Affy}_{\text{ratio}}$ and cDNA, each having 5754 genes (i.e., a patient's 822 DE genes × 7 patients). Out of 5754 DE gene expression data, a set of 4504 genes' expressions are randomly picked, which would be applied as a separate training dataset to be used by each of the methods. The remaining 1000 DE (i.e., 5754 – 4504 = 1000) gene expression data would be used for testing a trained framework, wherever possible.

The expression levels of the individual patients are considered for modelling only to represent each patient's ability to predict for others had there been no other patient's data available to form either the large 'global' set or the random set. It is expected that this, in a way, would help to judge the impact of each patient's contribution towards the model building from the larger set.

As performance indicators of the retrieved models, *mean square error* (MSE) and *Pearson product-moment* c*orrelation coefficient* (simply abbreviated as *corr. coef.*), symbolised by *r*, would be used. They are expanded in equation 24, where x, y, $\hat{y}$ and n represent the independent variable, dependent variable, predicted variable and the total number of data, respectively. In the results, it is desirable to have lower MSE-values. Corr. coef. represents the strength of the linear relationship between the variables, and the value of r is such that $-1 \leq r \leq +1$. In case of a strong positive correlation, r remains close to +1, whereas r-value close to -1 represents strong negative correlation. An r-value of zero means there is random, non-linear relationship, while r= ±1 means that all the data points lie exactly on a straight line with either positive or negative slope.

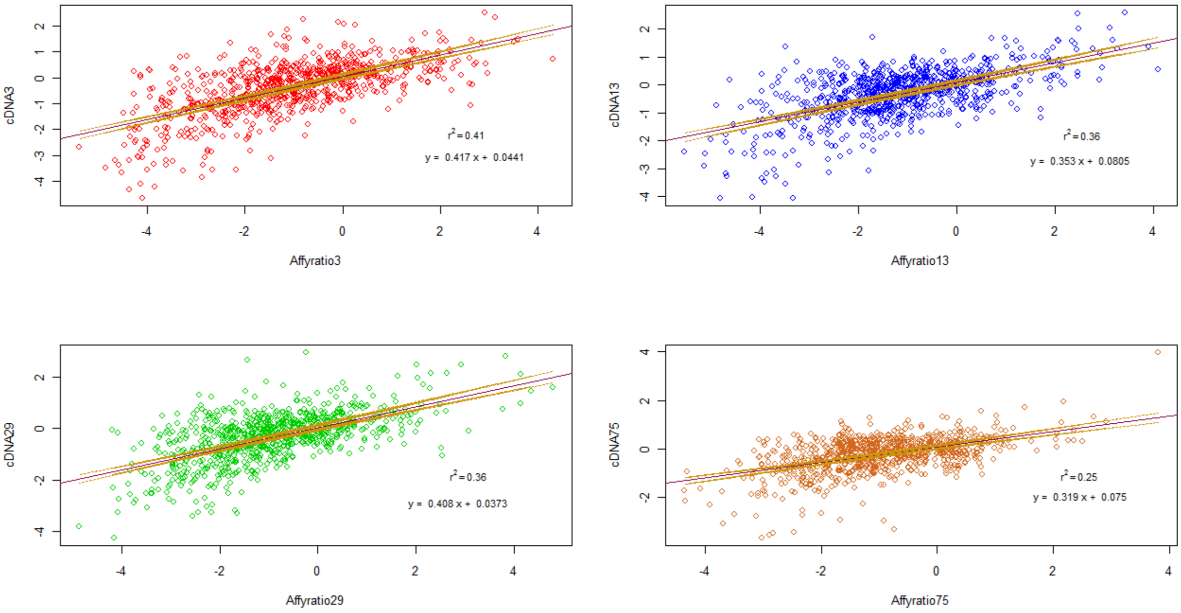
$$
\begin{aligned}
MSE &= \frac{\sum (y-\hat{y})^2}{n} \\
r &= \frac{\sum (x-\overline{x})(y-\overline{y})}{\sqrt{\sum (x-\overline{x})^2 \sum (y-\overline{y})^2}}
\end{aligned}
\qquad (24)
$$

6.1 Modelling the data

6.1.1 Linear model

6.1.1.1 Linear regression

To begin with, bivariate linear regression is decided to apply to test the strength and predictability of the linear model(s). In Figure 5.1, distributions of each patient's data in the form of scatter plots are presented along with regression equation, coefficient of determination (r^2) and 95% prediction confidence interval (CI). In all these figures, $Affy_{ratio}$ and cDNA data are considered predictor and response variable, respectively. The figures indicate that two patients, viz., patients 75 and 78, apparently have relatively low r^2-value. Further, all the linear fits are shown overlaid in a separate plot in Figure 6.2, which shows that the fits do not vary much from each other.



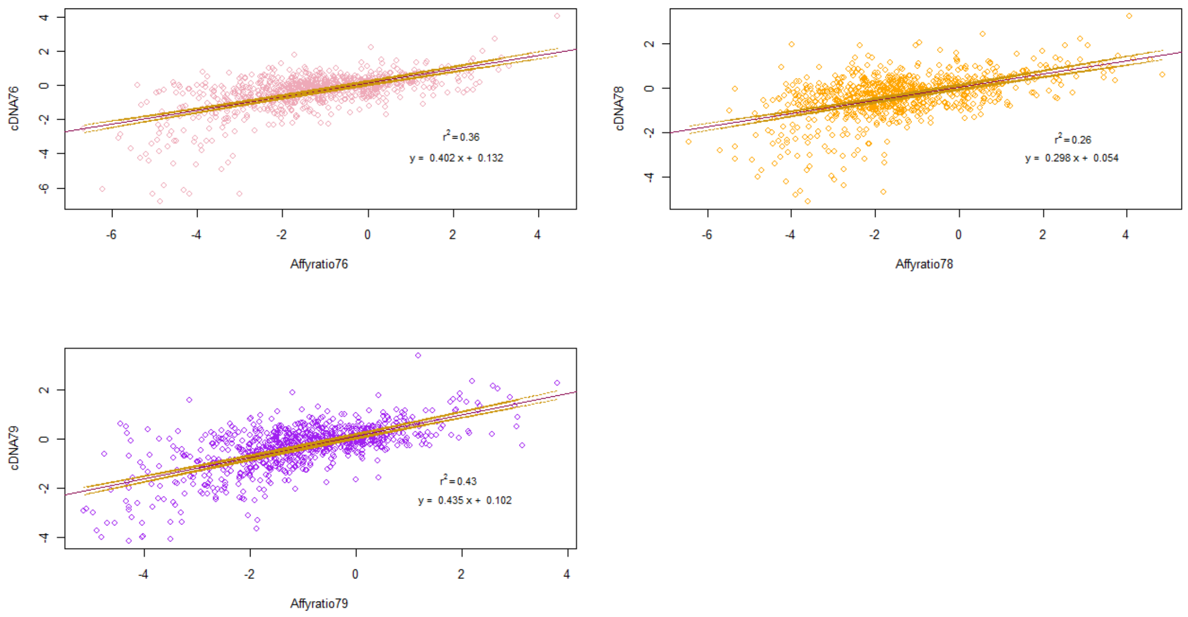


Figure 6.1 Scatterplot of individual patient’s data

[95% prediction CI, regression equation and r^2 value are overlaid]

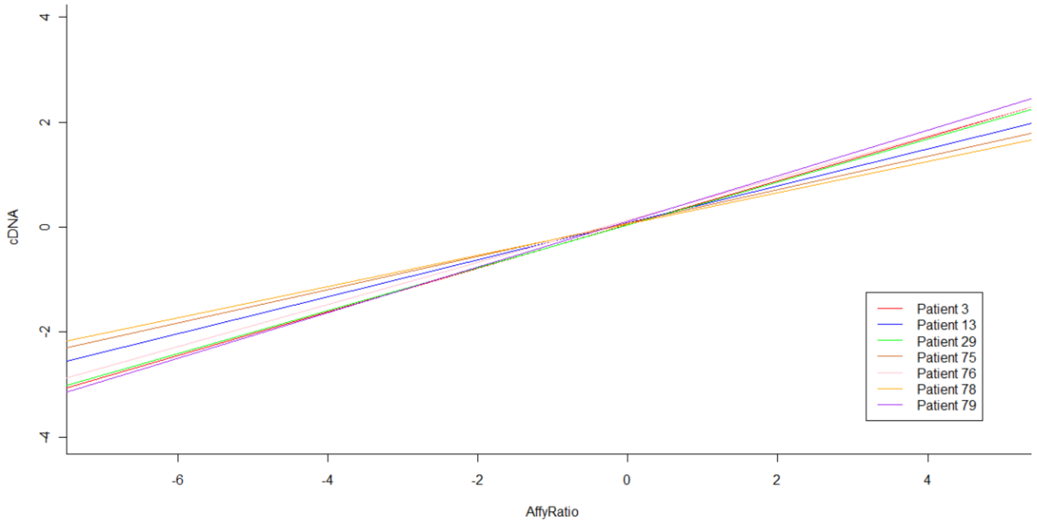


Figure 6.2 Overlaying of linear fits for each patient

Table 6.1 gives the model outputs of – (i) the whole dataset, (ii) individual patient tested against the remaining patients, and (iii) the random data. It presents regressional output with MSE and r-values. The tilde sign (~) between two variables indicates that the variable succeeding this sign is independent, and is a function of the first variable. *Coefficient of Variation* (CV) is also computed, and is the ratio of the standard deviation (σ) to the mean (μ) expressed in percentage, i.e., σ×100/μ. CV it is a useful statistic for comparison as it reveals the degree of variation from one data series to another.

Table 6.1 Linear regression

Data	Equation (y=mx+c)	Model used to test against	Corr. Coef. (r)	MSE (Mean Sq. Error)	CV of r
Whole dataset (cDNA ~ Affyratio)	0.373x + 0.075	Itself	0.5886	0.6013066	-
cDNA3 ~ Affyratio3	0.417x + 0.044	Itself	0.6423	0.6399171	-
		13	0.5728	0.5287462	16.48
		29	0.5959	0.5828090	
		75	0.4386	0.4279020	
		76	0.5956	0.7987785	
		78	0.4351	0.8017380	
		79	0.6526	0.5060829	
cDNA13 ~ Affyratio13	0.353x + 0.081	Itself	0.5962	0.5071630	-
		3	0.6264	0.6617987	10.88
		29	0.5821	0.5975301	
		75	0.4921	0.4015075	
		76	0.5992	0.7933181	
		78	0.5029	0.7389151	
		79	0.6393	0.5211792	
cDNA29 ~ AffyRatio29	0.408x + 0.037	Itself	0.5960	0.5826557	-
		3	0.6421	0.6401278	16.69
		13	0.5763	0.5255871	
		75	0.4443	0.4252660	
		76	0.5967	0.7971025	
		78	0.4444	0.7936326	
		79	0.6521	0.5067014	
cDNA75 ~ AffyRatio75	0.319x + 0.075	Itself	0.4966	0.3991997	-
		3	0.6100	0.6839783	6.68
		13	0.5921	0.5109993	
		29	0.5674	0.6128594	
		76	0.5886	0.8090309	
		78	0.5131	0.7285678	
		79	0.6230	0.5393163	
cDNA76 ~ AffyRatio76	0.402x + 0.132	Itself	0.6039	0.7863858	-
		3	0.6100	0.6839783	9.16
		13	0.5921	0.5109993	
		29	0.5674	0.6128594	
		75	0.4966	0.3991997	
		78	0.5131	0.7285678	
		79	0.6231	0.5393163	

Data	Equation (y=mx+c)	Model used to test against	Corr. Coef. (r)	MSE (Mean Sq. Error)	CV of r
cDNA78 ~ AffyRatio78	0.298x + 0.054	Itself	0.5145	0.7272214	-
		3	0.6008	0.6960965	7.34
		13	0.5874	0.5153991	
		29	0.5598	0.6204843	
		75	0.4955	0.3997533	
		76	0.5804	0.8208011	
		79	0.6134	0.5498618	
cDNA79 ~ AffyRatio79	0.435x + 0.102	Itself	0.6545	0.5038388	-
		3	0.6406	0.6422600	15.87
		13	0.5738	0.5278420	
		29	0.5933	0.5856468	
		75	0.4464	0.4242430	
		76	0.5981	0.7951187	
		78	0.4315	0.8048679	
Training set: 4504 data	0.380x + 0.081	Itself	0.5892	0.6172286	-
		1000 test data	0.5771	0.5298651	-

It is an impediment that there is no information found in the literature that can prescribe benchmark-values for such type of investigations. Thus, it is not pragmatic to comment at this stage on how good or bad the obtained linear models are, unless some other methods are tested and the results are compared. Moreover, it is important to note that such bivariate linear regression cannot address potential non-linear hidden patterns in a seemingly linear data. Therefore, adequate consideration for applying non-linear models is required.

6.1.2 Consideration for non-linear models

Attempting to use non-linear models on the microarray dataset is futile if just a linear model can adequately represent the data. Therefore, it is necessary to check the need for non-linear methods. However, it is often difficult to determine such necessity just based on simple visualization as even an apparently linear-looking data can contain underlying non-linear patterns undetectable to the eyes. As a way out to find an answer, two statistical tests are conducted in which a linear and a cubic polynomial model are used where the latter would query based on the non-linearity of the data.

6.1.2.1 F-test using ANOVA

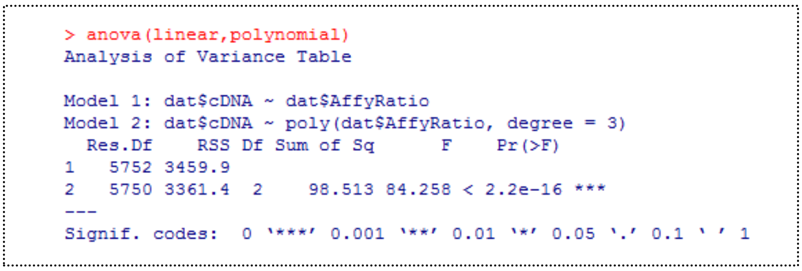
This test is also called *extra sum-of-squares* test, and is based on statistical *hypothesis testing* and ANOVA (analysis of variance).

The idea here is that once the data are fit to the two models, goodness-of-fit is quantified as the sum of squares of deviations of the data points from the model. Then, the complexity of the models is measured with the degrees of freedom (*df*), which equal the number of data points minus the number of parameters fit by regression. If the simpler model (the null hypothesis) is correct, the relative increase in the sum of squares approximately equals the relative increase in degrees of freedom. If the more complicated (alternative hypothesis) model is correct, then the relative increase in sum-of-squares (going from complicated to simple model) becomes greater than the relative increase in degrees of freedom. The F-ratio equals the relative difference in sum-of-squares divided by the relative difference in degrees of freedom. The equation along with its common form is shown in equation 25.

$$F = \frac{\dfrac{SS_{null} - SS_{alt}}{SS_{alt}}}{\dfrac{df_{null} - df_{alt}}{df_{alt}}} = \frac{\dfrac{SS_{null} - SS_{alt}}{df_{null} - df_{alt}}}{\dfrac{SS_{alt}}{df_{alt}}} \qquad (25)$$

F-ratios are always associated with degrees of freedom for the numerator and that for the denominator. The F-ratio in the equation has df_{alt} degrees of freedom for the denominator, and $df_{null} – df_{alt}$ degrees of freedom for the numerator. ANOVA computes an F-ratio from which it calculates a probability (P)-value. If the obtained P-value is less than the set statistical significance level, usually $\alpha = 0.05$, the alternative (complicated) model fits the data better than the null hypothesis (simpler) model. Otherwise, there is no compelling evidence supporting the alternative model, and so the simpler null model can be accepted.

The *extra sum-of-squares* test is computed for the 5754 DE microarray genes. As the snapshot below shows, the output renders a probability less than $2.2e^{-16}$. This suggests that the probability of obtaining a calculated F-value of 84.258 by chance is $2.2e^{-16}$ or smaller. This is highly unlikely; and hence, it is likely that the nonlinear model would provide improvement over the linear model.



```
> anova(linear,polynomial)
Analysis of Variance Table

Model 1: dat$cDNA ~ dat$AffyRatio
Model 2: dat$cDNA ~ poly(dat$AffyRatio, degree = 3)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   5752 3459.9
2   5750 3361.4  2    98.513 84.258 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.1.2.2 Akaike's Information Criterion

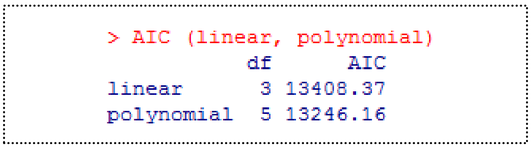
As an alternative approach to F-test and choosing a model with the use of statistical hypothesis testing, Hirotugu Akaike developed an approach for comparing models based on information theory. This method is called *Akaike's information criterion* or AIC (Akaike, 1974), which does not rely on P-values or the concept of statistical significance. Unlike the F test, which can only be used to compare nested[27] models, Akaike's method can be used to compare both nested and non-nested models. Moreover, as AIC is a different as well as a distinctly independent approach than the F-test, it is decided to test this method with the microarray dataset.

AIC method combines maximum likelihood theory, information theory, and the concept of the entropy of information (Burnham & Anderson, 2002). It is known in statistics as a penalized log-likelihood, and can be written as shown in equation 26.

$$
\begin{aligned}
&AIC = -2l + 2(p+1) \\
&l(\mu,\sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{\sum (y_i-\mu)^2}{2\sigma^2}
\end{aligned}
\tag{26}
$$

In the equation, p is the estimated coefficients in the model, and 1 is added here for the estimated variance. Log-likelihood, a measure of comparing the fit of two models, is denoted by *l*, and the value of which gets higher with better model. A somewhat similar structure of equation as given above is used by several statistical software. However, in simple terms, AIC can be defined as a method of comparing alternative specifications by adjusting the error sum of squares for the sample size and the number of coefficients in the model (p), i.e., AIC = log(SSE) + 2(p).

While using for comparison, AIC can be computed exactly as ANOVA to determine how well the data supports each model. The model with the lowest AIC score is most likely to be a better fit. When applied to the 5754 DE microarray genes, polynomial is found to have the lower AIC as shown in the box below, and therefore, can be preferred over linear regression.



```
> AIC (linear, polynomial)
           df      AIC
linear      3 13408.37
polynomial  5 13246.16
```

[27] When a model is a simpler case of the other, the models are said to be *nested*.

Both the statistical tests above present an indication that non-linear methods may potentially bring improved outcomes with regards to the microarray data. This confers a trust upon exploring the non-linear methods further.

6.1.3 Non-linear models

6.1.3.1 Polynomial regression

Polynomial models are useful to investigate the presence of possible curvilinear effects in the response function. Such regression fits a nonlinear relationship to the data where the dependent variable is modelled as an nth order function of the dependent variable. Every polynomial corresponds to a polynomial function, and can be represented as shown in equation 27, where n is a non-negative integer and a_0, a_1, a_2, ... a_n are constant coefficients.

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_2 x^2 + a_1 x + a_0 \tag{27}$$

The results of polynomial regression applied to the microarray datasets are given in Table 6.2. For comparison of these results with the linear regression of Table 6.1, the MSE values can be used such that relative decrease of MSE along with no change or an increase of correlation (positive or negative), r is an indication of better representation of the relationship by a model. The obtained polynomial results are found to be relatively improved compared to the values of linear regression. This has also been already confirmed when statistical tests were carried out to probe the presence of non-linearity in the data.

Table 6.2 Cubic polynomial

Data	Equation $y = b_1x^3 + b_2x^2 + b_3x + C$	Model used to test against	Corr. coef. (r)	MSE	CV of r
Whole dataset: (cDNA ~ Affyratio)	$y = 0.0084x^3 - 0.0058x^2 + 0.2669x + 0.0494$	Itself	0.6042	0.58419	-
cDNA3~Affyratio3	$y = 0.0058x^3 - 0.0323x^2 + 0.2938x + 0.0848$	Itself	0.6629	0.61054	-
		13	0.5771	0.52488	17.22
		29	0.6044	0.57357	
		75	0.4854	0.40499	
		76	0.6197	0.76247	
		78	0.4021	0.82914	
		79	0.6593	0.49827	

Data	Equation $y = b_1x^3 + b_2x^2 + b_3x + C$	Model used to test against	Corr. coef. (r)	MSE	CV of r
cDNA13 ~ Affyratio13	$y = 0.011x^3 + 0.0098x^2 + 0.2526x + 0.0241$	Itself	0.6107	0.49340	-
		3	0.6388	0.64475	11.30
		29	0.5768	0.60313	
		75	0.5166	0.38841	
		76	0.6431	0.72593	
		78	0.5001	0.74164	
		79	0.6465	0.51307	
cDNA29 ~ Affyratio29	$y = 0.0081x^3 - 0.024x^2 + 0.3138x + 0.0566$	Itself	0.5638	0.61651	-
		3	0.6170	0.67459	10.39
		13	0.6022	0.50151	
		75	0.5083	0.39296	
		76	0.6226	0.75797	
		78	0.5009	0.74085	
		79	0.6362	0.52475	
cDNA75 ~ Affyratio75	$y = 0.0264x^3 + 0.0339x^2 + 0.2008x - 0.0007$	Itself	0.5266	0.38289	-
		3	0.6365	0.64802	21.19
		13	0.5773	0.52462	
		29	0.5529	0.62745	
		76	0.6116	0.77480	
		78	0.3270	0.88322	
		79	0.6423	0.51787	
cDNA76 ~ Affyratio76	$y = 0.0151x^3 + 0.0115x^2 + 0.2105x + 0.0294$	Itself	0.6470	0.71971	-
		3	0.6360	0.64866	10.96
		13	0.6077	0.49630	
		29	0.5646	0.61561	
		75	0.5193	0.38698	
		78	0.4877	0.75375	
		79	0.6407	0.51964	
cDNA78 ~ Affyratio78	$y = 0.006x^3 - 0.0003x^2 + 0.221x + 0.0231$	Itself	0.5241	0.71736	-
		3	0.6074	0.68737	7.93
		13	0.5963	0.50710	
		29	0.5502	0.63021	
		75	0.5017	0.39646	
		76	0.6177	0.76562	
		79	0.6144	0.54869	
cDNA79 ~ Affyratio79	$y = 0.008x^3 - 0.0041x^2 + 0.3431x + 0.0807$	Itself	0.6638	0.49304	-
		3	0.6570	0.61902	16.20
		13	0.5865	0.51622	
		29	0.6016	0.57667	
		75	0.4885	0.40342	
		76	0.6256	0.75336	
		78	0.4174	0.81668	
Training set: 4504 data	$y = 0.009x^3 - 0.0049x^2 + 0.2672x + 0.0495$	Itself	0.6064	0.59786	-
		1000 test data	0.5835	0.51400	-

6.1.3.2 Locally weighted regression

In the methodology of time series, there is an old idea deeply buried where the data measured at equally spaced points in time were smoothed by local fitting of polynomials (Macaulay, 1931). Then, the era of contributions came where chronologically Watson (1964), Stone (1977), Cleveland (1979), Hastie & Tibshirani (1986) and Cleveland & Devlin (1988) introduced as well as streamlined the local fitting methods into the more general case of regression analysis. Professor William S. Cleveland (1979) proposed and further developed by him and Susan Devlin (1988), it is the specific local fitting method, *locally weighted regression*, which is the subject of this section.

The curve fitting regression technique introduced by William S. Cleveland is called LOWESS, which stands for *locally weighted regression scatter plot smoothing*. Its derivative, LOESS stands more generally for a local regression, and differs from LOWESS based on the model used in the regression: LOWESS uses a linear polynomial whereas LOESS uses a quadratic polynomial (Saeed et al., 2006). Many researchers consider LOWESS and LOESS as synonyms.

More descriptively, the method of *locally weighted regression* or *Loess* (aka *Lowess*) can be considered as locally weighted polynomial regression. The method combines much of the simplicity of linear least square regression with the flexibility of nonlinear regression. To achieve this, it uses a nearest neighbour algorithm and determines localized subsets of data. Local polynomials of usually first or second degree are fit to these subsets of data using weighted least squares. A user specified *smoothing parameter* (*f*) gives the flexibility to the Loess function, and it is approximately the fraction of points to be used in the computation of each fitted values. There is no single correct value of *f*, and the values can range from 0 to 1. However, different *f* values give different summaries. As Chambers et al. explains (1983), a small value of *f* gives a very local summary of the middle of the distribution of *y* in the neighbourhood of x. Such value tends to force the function to excessively conform to the data, and only points whose abscissas are relatively close to x_i determine y_i. This produces high resolution, but a lot of noise. For large values of *f*, the summary is much less local. In this case, there is low resolution with less noise. With respect to the smoother-line in the scatter plot, the larger the *f*-value gets, the lesser becomes the wiggle in response to the fluctuations in the data, or vice versa.

The subset of data used in each weighted least squares fit is comprised of the data whose explanatory variables are closest to the point at which the response is being estimated. Based on the weight function, closer a data remains to the point of estimation, higher the weight it attains. Therefore, a local model can be considered to have the most influence by the nearby data than the points that are further apart. Any weight function can be used in this purpose as long as it satisfies the properties listed in Cleveland (1979).

Application of loess method to the DE genes of the microarray data is quite possible. However, on the basis of the principles involved in loess, it is found that any attempt of finding goodness of its fit through measures such as r and MSE is rather practically meaningless. The reason lies in the explanation of the loess method given above. In loess, a locally weighted estimate of a specified degree over a given fraction of the data is computed, where the region over which the fit is performed slides to the right in each iteration. The combination of all these individual results produces the final fit. Again, this makes little practical sense to determine the form of the loess model; and because of that, measures such as r and MSE is rather pointless for loess models. It may be possible to estimate some r- like measures for the loess model by carefully deriving from its definition, and MSE-like estimate by extension, but it may not actually be meaningful as unlike regression, which produces pre-specified, parametric model for which the parameters are calculated from the data, loess lacks any such analogue, and the entire loess fit is estimated solely from the data without producing a single coherent model: with the change of either the span of the data or the degree of the local fit or both, there would be change in the r- and MSE-like estimates.

Loess has been considered critically for applying in the DE genes of the microarray data. It is, however, subsequently avoided being used to its full potential because of its data-driven attribute - as none of the outcomes can be considered to be in line with the results of the investigations using the other methods. Nevertheless, to examine how the method contributes varying from the linear and polynomial distribution, the algorithm given in the box below is used for the 5504 DE genes, and the output is graphically presented in Figure 6.3 using *ggplot2* (Wickham, 2009), an implementation based on the *Grammar of Graphics* (Wilkinson, 2005). In applying the algorithm, the smoothing parameter and the degree of the local polynomial used is 0.75 and 2, respectively. The comparative graphics shows that the loess and the polynomial fits are close to each other and are relatively better fits than the linear model.

LOESS algorithm:

- The data has *n* data points, (x_i, y_i), *i* = 1, 2, 3, ..., n.
- User supplies the smoothing parameter (*f*), the fraction of points to be used in the computation of each fitted vales. Let *q* be *fn* rounded to the nearest neighbour.
- Computation of neighbourhood weight function:
  - Let *T(u)* be a *tricube weight function*:

$$\mathrm{T}(u) = \begin{cases} (1-|u|^3)^3 & for \quad |u|<1 \\ 0 & \quad\;\; |u|\geq 1 \end{cases}$$

  - Weight given to point, (x_k, y_k) while computing a smoothed value at x_i is:

$$t_i(x_k) = T\left(\frac{(x_i - x_k)}{d_i}\right)$$

    [d_i is the distance from x_i to its q^{th} nearest neighbour along the x-axis. x_i is counted as a neighbour of itself.]
  - Neighbourhood weights are obtained for all neighbourhood points.
- A line is fitted to a strip of the scatter plot that has the points, (x_i, y_i) using weighted least squares with weights, $t_i(x_i)$. That is, values of *a* (intercept) and *b* (slope) are found, which minimize $\sum_{k=1}^{n} t_i(x_k)(y_k - a - bx_k)^2$.
- Further, to prevent distortion by a small fraction of outlying points, an additional stage of robustness procedure can be used:
  - Find residuals (*r*) for all the fitted values and *m*, the median of the absolute values of the residuals: $r_i = y_i - \hat{y}_i$, and *m* = *median*|r_k|.
  - Based on the sizes of the residuals, define a set of robustness weights. The robustness weight for the point (x_k, y_k) is: $w(x_k) = B(r_k / 6m)$. It uses bisquare weight function, *B*(*u*), which is -

$$\mathrm{B}(u) = \begin{cases} (1-u^2)^2 & for \quad |u|<1 \\ 0 & \quad\;\; |u|\geq 1 \end{cases}$$

  - The robustness weight for the point (x_k, y_k) is: $w(x_k) = B(r_k / 6m)$
  - To re-fit a line to the strip's each point in the scatter plot, the new smoothed value at x_i is calculated using the original neighbourhood weight multiplied by the robustness weight for that point.

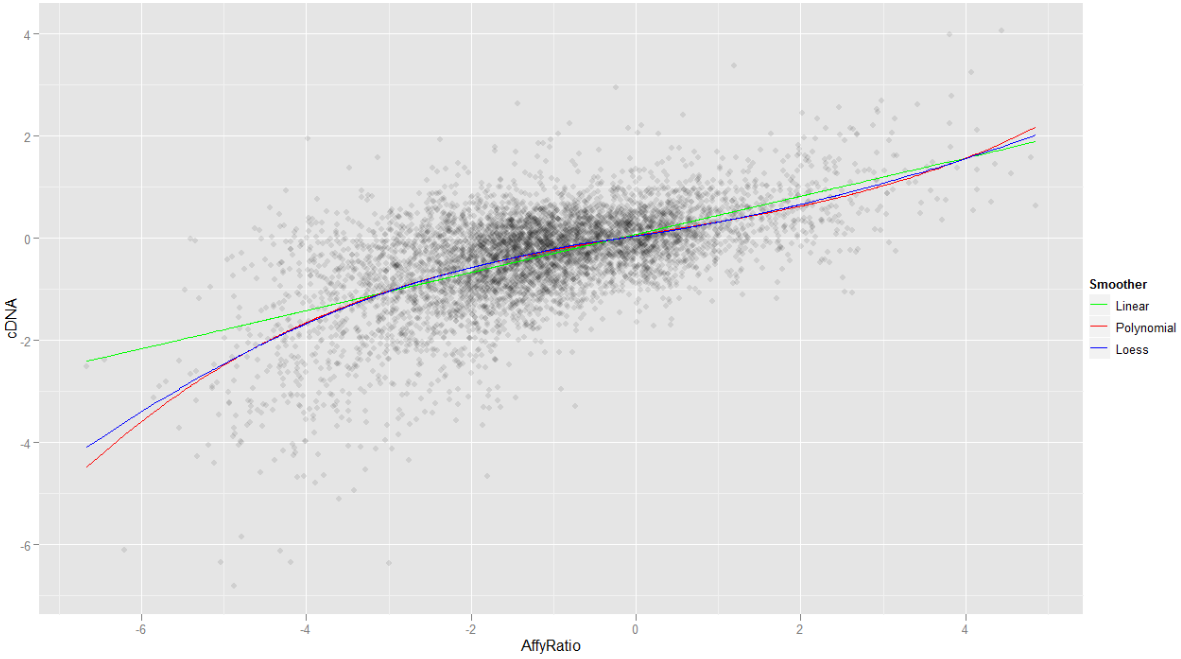


Figure 6.3 Microarray data with linear, polynomial and loess fit

6.1.3.3 Bootstrap Aggregating

Bootstrap aggregating is a method useful for avoiding model overfitting to data with variance reduction. It has been in use for a varied range of microarray studies (Sandrine Dudoit & Fridlyand, 2003; Lu, Devos, Suykens, Arus, & Huffel, 2007; Politis, 2008), and is known to provide stability and accuracy to a model. It comes from the concept of *bootstrapping*. The method of bootstrapping is briefly introduced here prior to addressing bootstrap aggregating.

Bradley Efron invented the concept, *bootstrapping*, in 1979 through his paper - Efron (1979). The word, *bootstrapping* refers to a group of metaphors that generally mean: a self-sustaining process that proceeds unaided. The term is believed to have originated from the German scientist and librarian, Rudolf Erich Raspe's classic collection of tall stories published in 1785, *The Surprising Adventures of Baron Munchausen*, where the main character escapes from a swamp by pulling himself up by his bootstraps. Bootstrapping is a well-known method for estimating standard errors, bias, and constructing confidence intervals for the parameters, and has been popularised from 1980s due to the introduction of computers in statistical practice.

Bootstrap is the most recently developed, computer-intensive approach to retrieve statistical inference. In traditional statistical techniques, it is reasonably a common practice to consider the distribution of a dataset based on certain assumptions. For example, assuming that a dataset is normally distributed is quite acceptable. However, this clearly cannot be true always; besides, there is decidedly no consensus on what distribution would be believable. In such cases, bootstrapping can be used to go around, and let the data reveal its true self. This is achieved by sampling from the empirical distribution of the data without replacing or adding to the data.

Usually, a statistic is computed on a dataset and the investigator knows that one statistic while being unable to see the possible variability present in that statistic. Bootstrap draws a large number of samples using random sampling with replacement from the dataset that the investigator is working with, and computes the statistic on each of these samples. Just like multiple samples give sampling distribution, bootstrap samples provide bootstrap distribution, and thereby presents a way to explore variability as well as to estimate standard errors, bias and constructing confidence intervals for the parameters. A schematic of bootstrapping is given in Figure 6.4, where the bootstrap statistics are used to evaluate the original sample statistics.

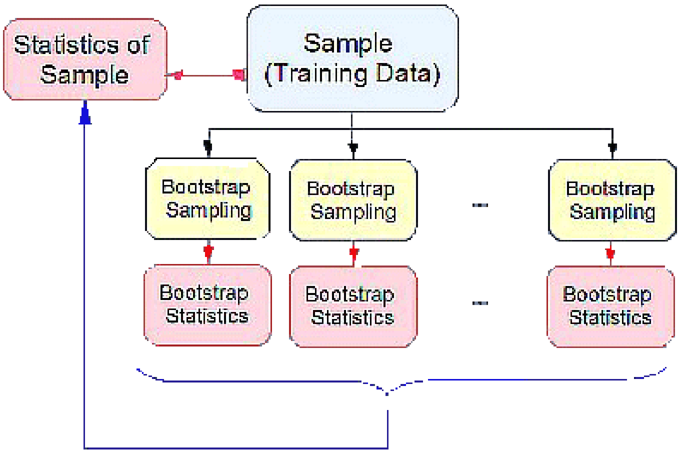


Figure 6.4 A schematic of bootstrapping process

The computational algorithm involved in bootstrapping is probed into and presented in the box below. The assumptions on which the overall approach is based on are: a) the sample from where the bootstrapping is carried out is a valid representation of the population; b) the sub-samples obtained from bootstrapping come from the same distribution of the population; and c) each of the sub-samples is drawn independently from the rest.

Bootstrap algorithm:

- Let the original sample be L = (x_1, x_2, ... , x_n), where x_i is drawn from an empirical population distribution, $\hat{F}$.
- Repeat *B* times :
  - Generate a sample L_k of size *n* from *L* by sampling with replacement.
  - Compute $\hat{\theta}^*$ for x^*
- The corresponding bootstrap values are : $\hat{\theta}^* = \left(\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_B^*\right)$
- Use the values of $\hat{\theta}^*$ to calculate the parameters of interest.

Notations:

θ = Parameter; * = Data generated from bootstrapping; ^ = An estimate

__B__ootstrap __agg__regat__ing__, or *bagging* is a machine learning meta-algorithm introduced by Leo Breiman (1996); and it is used here to investigate the microarray data. Bagging is an ensemble method, i.e., a method of combining multiple predictors. To apply bagging to the microarray data, a computational algorithm is constructed and is given in the following box.

Bagging algorithm:

- Let the original sample be L = (x_1, x_2, ... , x_n) where x_i is drawn from an empirical population distribution, $\hat{F}$.
- Repeat *B* times :
  - Generate sample, L_k of size $n' \leq n$ from *L* by sampling from *L* randomly and with replacement. If n' = n, then 63.2% of unique values of *L* is expected to remain in L_k, the rest being duplicates, i.e., 36.8% of the data that is not used.
- Develop k-models by fitting samples of L_k.
- Combine the predictors of the models by either averaging the output for regression (or, voting for classification).

The results of implementing the algorithm to the microarray data are provided in Table 6.3. The table shows that although the obtained r-values are relatively comparable to the earlier results of linear and polynomial regression, the values of MSE are found to be much higher, with low standard deviation. This may be an indication that even though the method of bootstrap aggregating is a useful method in a number of published microarray studies, it may not be suitable for applying in the current context.

Table 6.3 Bootstrap aggregating

Data		Coefficients					Corr. coef.	MSE	Standard deviation		CV of r
		Intercept	x	x^2	x^3	x^4	(r)	(Mean Sq. Error)	r	MSE	
Whole dataset:	(cDNA ~ Affyratio)	-0.82944	1.02726	0.03492	-0.00201	0.00071	0.5938	1.48701	0.10663	0.03994	-
Arrays:	cDNA3 ~ Affyratio3	-0.62953	1.08522	-0.08158	-0.03548	0.00064	0.6496	1.55120	0.16793	0.11713	9.45
	cDNA13 ~ Affyratio13	-0.89934	0.92715	0.09243	0.09037	0.01867	0.6073	1.48132	0.17450	0.12871	
	cDNA29 ~ Affyratio29	-0.64357	0.90678	0.00929	-0.00172	0.00203	0.5998	1.30646	0.16882	0.10989	
	cDNA75 ~ Affyratio75	-0.79819	0.75063	-0.03018	0.05600	0.02537	0.5084	1.55282	0.17496	0.97268	
	cDNA76 ~ Affyratio76	-0.98388	1.13081	0.04338	0.00292	0.00165	0.6253	1.76413	0.17607	0.13322	
	cDNA78 ~ Affyratio78	-1.11651	1.13272	0.05289	-0.02176	-0.00084	0.5345	2.20767	0.16959	0.16094	
	cDNA79 ~ Affyratio79	-0.68878	1.02781	0.01207	0.00858	0.00419	0.6577	1.21044	0.17928	0.12432	
Training set	(with 4504 data)	-0.84227	1.02154	0.04031	-0.00127	0.00062	0.5956	1.48062	0.11380	0.04681	-
Test set	(1000 data)	-0.80233	1.05375	0.03864	0.00694	0.00258	0.5872	1.55877	0.17094	0.11155	-

6.1.3.4 Self-Organizing Maps

Machine learning approaches, such as *Artificial Neural Networks* (ANN) are considered to be effective computational methods that enable efficient capture of the trends potentially available in the data.

Pioneered by Rosenblatt (1962), Widrow & Hoff (1960) and Widrow & Stearns (1985), ANN represent a computational tool, based on the properties of biological neural systems. These neural networks are useful in the sense that they incrementally learn from their data-environment, and efficiently reveal the inherent complexity present in the data. This helps in providing reliable predictions for new situations containing noisy and partial information. ANN are especially powerful while fitting arbitrarily complex non-linear models to data. This task is carried out by the neurons, which are units that locally process data with nonlinear data processing capabilities similar to the concept of learning in the brain. Neurons possess dynamic weights that remain as free parameters in the architecture making the entire network flexible. This flexibility in the network enables ANN to freely follow the pattern in the input data to map with the output, and to solve a variety of problems. A simple neural model and its components are elaborated below under the section: 6.1.3.5 *Feedforward Neural Network*.

Self-organizing map (SOM) is the most widely used unsupervised neural networks. Introduced by Teuvo Kohonen (1982a, 1982b, 1998), it uses only the input data and projects it onto one- or two-dimensional grid for meaningful interpretation of its inherent structure and patterns as well as for visual validation (Kohonen, 2001). As Figure 6.5 shows, the input layer of a SOM represents the input variables, while the output layer consists of either a one-dimensional (1D) or a two-dimensional (2D) layer of neurons. The weights are free parameters that link the input data to the output neurons, and own the same dimension as the inputs.

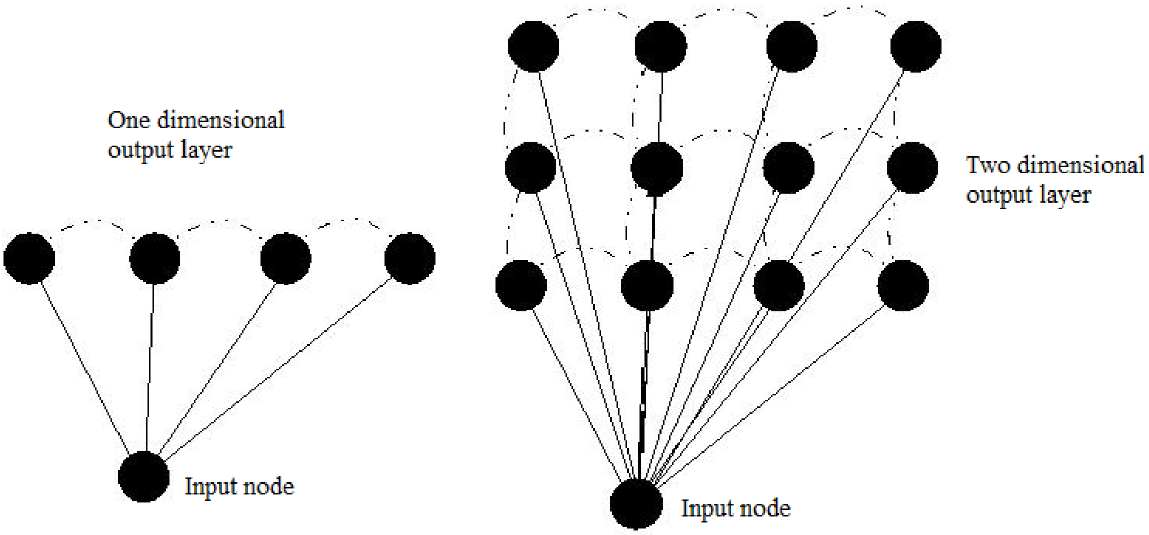


Figure 6.5 Self Organizing Map (1D and 2D)

Adhering to the characteristic attributes of ANN, the neurons in SOM learn unsupervisedly through a *competitive learning* scheme to specialize in responding to a specific set of inputs. As the weights evolve through such learning, each weight finally assumes the centre position of a cluster. Each neuron with its final weight becomes the winner for inputs from its cluster. The final weight vector becomes the representative of the cluster, and the corresponding inputs remain closer to this weight vector than to the rest. Thus, the competitive learning plays a vital role by facilitating evolution of weights and their movements to respective cluster centres.

As an input is presented to the SOM network, the process of competitive learning starts, and the winner is selected based on either neuron activation or distance to the input vector. In neuron activation, each neuron calculates its weighted sum of inputs, i.e., $\sum_{i=1}^{n} w_{ij} x_i$, where x_i is the i^{th} input variable and w_{ij} is the weight of input x_i to j^{th} output neuron. A neuron drops out if this neuron activation value is below a threshold (or, zero). Finally, one neuron emerges as winner which has the highest activation, and it represents the input vector. This winner selection can also be done using the distance between an input and a weight vector, and can be explained using equation 28, which is a representation of neuron activation.

$$\sum_{j=1}^{n} w_{ij} x_i = \| x_i \| \| w_{ij} \| \cos\theta \qquad (28)$$

In the equation, relative lengths of input and output neuron's weight vector are represented as ||x|| and ||w|| respectively, and θ is the angle between the input vector, ***x*** and the weight vector, ***w***. A smaller angle results in higher value of cosθ*,* producing a higher net input. Therefore, a weight closer to an input vector causes a larger activation. The closeness of a weight to an input vector can be found using various distance measures, including *Euclidian*, *correlation*, *direction cosine*, and *city block distance*. As the distance is obtained between an input vector and the weight vectors of all the output neurons, the neuron with the smallest distance becomes the winner. Using equation 29, these weights are updated so that it moves closer to the input vector, while all the other weights remain unchanged. The *β* shown in the equation is called *learning rate* (or, *step length*), which indicates proportion of movement the winning weight vectors make towards the input vector.

$$\triangle w_j = \beta d_j = \beta(x - w_j) \tag{29}$$

Peltarion Synapse®, v1.3.6 (Peltarion Corporation, Stockholm, Sweden) is used to generate Figure 6.6, which presents a set of adaptive self-organizing maps that uses the DE genes of Affy$_{\text{ratio}}$ and cDNA from the entire 5754 DE microarray genes. In the background, it is made possible by the neighbourhood feature in SOM.

Topology preservation is a unique characteristic of human brain, whereby it organizes the tasks of similar nature, such as vision and speech, to be controlled by regions having spatial proximity to each other (Samarasinghe, 2007). It was incorporated into SOMs as neighbourhood feature, which helps to preserve topological characteristics of inputs. The inputs spatially closer together must be represented in close proximity in the output layer or map of a network. Therefore, besides the winner, the neighbouring neurons also adjust their weights during the process of learning. For an n-dimensional input vector **x** with components {$\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$}, the weights of the winner and neighbours are adjusted to $\mathbf{w}'_j$ as given in equation 30, where *β* is the learning rate and NS is the *neighbourhood strength*. Neighbourhood strength determines how the weight adjustment decays with distance from the winner, and its commonly used functions are linear, Gaussian, and exponential.

$$\mathbf{w}'_j = \mathbf{w}_j + \beta \text{ NS } [\mathbf{x} - \mathbf{w}_j] \tag{30}$$

In a maplet of Peltarion Synapse, each hexagonal cell represents the processing elements, neurons or nodes. Each neuron represents none or many input data points to which it is closest to in the feature space (or, the value range). Onto the same node or to the neighbouring nodes of the map, similar data are mapped. This grouping leads to spatial clustering of similar input patterns of the microarray data in neighbouring parts of the SOM, and the clusters appearing on the map become organized themselves unsupervisedly. The final arrangement of the clusters on the map tends to reveal the relationships of the variables of the input space. In the figure, the number of DE genes associated with a node is shown as a black dot in the hexagon. The size of the dot is approximately proportional to the number of genes associated with the node in question. The maplets have the same topological mapping, so a node (and, implicitly a group of genes) in one maplet has the same position in the other. The maplets in the figure indicate that there is more or less a proportional variation in the expression levels of Affy$_{\text{ratios}}$ and cDNAs in the feature space.

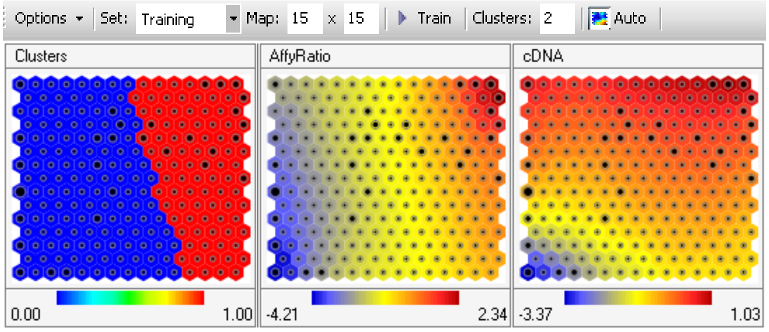

Set: Training
Map: 15 x 15
Clusters: 2

Figure 6.6 Self-organizing map

SOMs are considered highly efficient techniques for exploratory data analysis. This exploratory technique is explored here further to investigate whether it can be used for defining relation between microarray data from Affymetrix and cDNA platform. It is an attempt based solely on the principles of SOM as well as on its inherent properties to broaden its usage towards employing it as a prediction tool.

Each neuron of a trained SOM includes a specific set of datapoints. In a 2D space, such a neuron holds a final weight and the weight bears two components, one in *x*- and the other in *y*-direction. With this as a preface, a computational algorithm is constructed for SOM, and is given below in the box.

Algorithm used for SOM:

- Let the training dataset for microarray be L = (x_1y_1, x_2y_2, ….. , x_ny_n) where x_i and y_i is Affy$_{\text{ratio}}$ and cDNA respectively.
- Train the data using the regular SOM algorithm (Kohonen, 1982a, 1982b, 1998, 2001).
- Use the test dataset, T = (a_1b_1, a_2b_2, ….. , a_nb_n) where a_i and b_i is test data from Affy$_{\text{ratio}}$ and cDNA respectively.
- For each a_i :
    - Advance in x-direction by the value, a_i
    - in y-space, search for the closest neuron, N_c
    - Average the cluster of y_i -values that come under N_c. This $\bar{y}_i$ represents the corresponding SOM-output of the a_i value.

Matlab®, 2010a (The MathWorks Inc., Massachusetts, USA) is used to implement the algorithm. An instance of implementation is given in Figure 6.7 where the final positions of the neurons are shown when SOM-training is completed. The training and test data used belong to the random drawn 4504 and 1000 datasets respectively. In the figure, the neuron positions are demarcated by rectangles while the positions of the training and test data are shown as dots (.) and crosses (×), respectively.

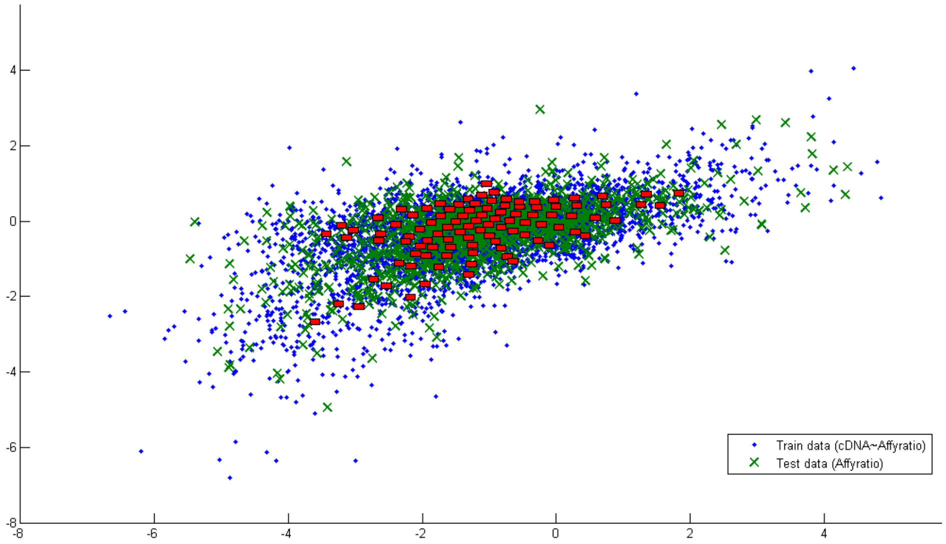


Figure 6.7 Final neuron-positions along with training- and test-data

The obtained outputs are given in Table 6.4. It is evident from the table that the results are better than that of bootstrap aggregating. However, they are not as good as those of either the linear or polynomial regression.

Table 6.4 Output of self-organizing maps

Data	Used to test against	Corr. Coef. (r)	MSE (Mean Sq. Error)	CV of R
Whole dataset (cDNA ~ Affyratio)	Itself	0.4650	0.8994	-
cDNA3 ~ Affyratio3	Itself	0.5329	0.9495	-
	13	0.4414	0.9082	16.39
	29	0.4188	0.9070	
	75	0.3029	0.7486	
	76	0.4948	1.1448	
	78	0.4090	1.2287	
	79	0.4883	0.8315	
cDNA13 ~ Affyratio13	Itself	0.4854	0.7155	-
	3	0.4868	0.9132	18.34
	29	0.4059	0.8321	
	75	0.3015	0.6214	
	76	0.4988	1.0069	
	78	0.3749	1.0906	
	79	0.4789	0.7396	
cDNA29 ~ Affyratio29	Itself	0.4326	0.9234	-
	3	0.5284	0.9589	11.01
	13	0.4496	0.9063	
	75	0.3897	0.6843	
	76	0.5002	1.1481	
	78	0.4402	1.1990	
	79	0.5055	0.8035	
cDNA75 ~ Affyratio75	Itself	0.3270	0.6556	-
	3	0.5015	0.9019	9.92
	13	0.4406	0.7832	
	29	0.4322	0.8292	
	76	0.4911	1.0460	
	78	0.4003	1.0847	
	79	0.5171	0.7019	
cDNA76 ~ Affyratio76	Itself	0.5382	1.0810	-
	3	0.4981	0.9606	15.43
	13	0.4422	0.7935	
	29	0.3700	0.8891	
	75	0.3254	0.5908	
	78	0.3760	1.1938	
	79	0.4456	0.8758	

Data	Used to test against	Corr. Coef. (r)	MSE (Mean Sq. Error)	CV of R
cDNA78 ~ Affyratio78	Itself	0.2178	1.1450	-
	3	0.3044	1.1472	17.06
	13	0.2346	0.9310	
	29	0.2477	1.0101	
	75	0.1760	0.7333	
	76	0.2589	1.3010	
	79	0.2640	0.9712	
cDNA79 ~ Affyratio79	Itself	0.5462	0.7085	-
	3	0.5380	0.9215	15.24
	13	0.4799	0.8240	
	29	0.4256	0.8526	
	75	0.3455	0.6373	
	76	0.5025	1.1024	
	78	0.4209	1.2065	
Training set (with 4504 data)	Itself	0.4643	0.9316	-
Test set	(using training set of 1000 data)	0.4405	0.8586	-

6.1.3.5 Feedforward Neural Network

Supervised neural networks are the mainstream of neural network development, and the *feedforward neural networks* fall in the category of supervised networks. The concept of these networks starts with the idea of a simple neuron model. The first and the simplest type of it, called *perceptron model* was invented by Frank Rosenblatt (1962). The perceptron model is rarely used now-a-days, and its significance is only left with its historical contribution to neural networks.

In a simple neuron model (without feedback or competition), the neuron receives inputs (x_1, x_2 ..... x_n) from multiple sources. Each input has an associated *weight*, which is initialised with random value. Both inputs and weights can typically be real values, i.e., positives or negatives. *Bias* is an additional input supplied to the neuron, and it incorporates the effects that are not accounted for by the inputs. This overall architecture is called *neuron model*, which learns until it properly performs the task of mapping a given input dataset to output through iterative modification of the initial random weights.

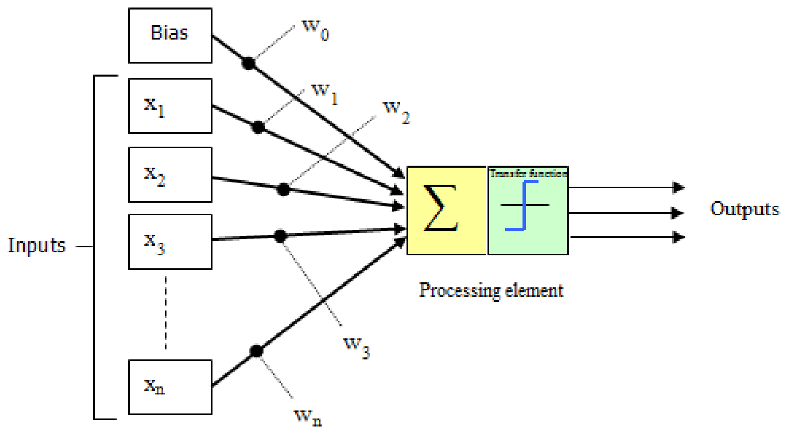


Figure 6.8 A model of a neuron

Figure 6.8 depicts a regular neuron model that consists of the weights, bias, summation processor, and a transfer function. The *summation processor* sums all the weighted inputs, and modifies the signals through the *transfer* (or, *activation*) *function*. The transfer function is usually non-linear, and it transforms the weighted input non-linearly to an output. The transfer function can be a threshold function allowing only those signals that reach a certain threshold level, or is a continuous function of the combined input. The final output of a neuron model can be presented as in equation 31.

$$\sigma\left(\sum_{j=1}^{n} w_j x_j + b\right) \qquad (31)$$

In the equation, σ represents a non-linear function, and w_j is the weight associated with the j^{th} x_j, while b is the bias weight. There is a wide range of options for non-linear functions, including *Sigmoid*, *Gaussian*, *sine*, *arc tangent*, and their different variants.

Using a linear function, a neuron model becomes analogous to a multiple linear regression model in statistics where the bias, b_0 becomes the intercept of statistical terminology. As in statistics, here too the intercept represents the factors that are not accounted for by the inputs. The output of a linear neuron model is given in equation 32.

$$y = \sum_{i=1}^{n} w_i x_i + w_0 = w_1 x_1 + w_2 x_2 + \text{........} + w_n x_n + w_0 \qquad (32)$$

A neural network often used in practical applications can consist of an input layer, a single- or multi-layer of neurons, and an output layer. Accordingly, the terms, *hidden layers* and *hidden neurons* are used to indicate respectively the layers and the neurons between the input and the output layer. This is depicted through Figure 6.9, which presents a framework of a feedforward network. In feedforward networks, all the connections remain unidirectional from input to output layers.

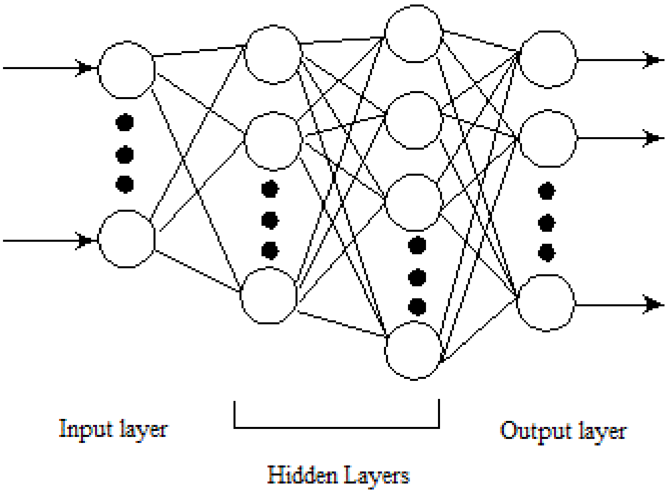


Figure 6.9 A feedforward neural network

In multi-layer feedforward networks, it is said that high number of neurons with multiple layers often tends to create undesirable complexity. The same is empirically experienced here too while working with our microarray data. Therefore, a simple feedforward network is finally preferred, which consists of one neuron in the middle layer, besides the input- and the output-layer. Again, Matlab® is used to do the required computations. Various parameters used in these calculations are given below:

i) Training function: *Levenberg-Marquardt* method (More, 1977) is used here as a learning method. It is a second-order method, and relies on both first and second derivative of error (slope and curvature) while searching for the optimum weights. The method is considered as a hybrid algorithm as it combines the advantages of *steepest descent* and *Gauss-Newton* methods. Levenberg-Marquardt algorithm is a fast method, and it primarily makes use of the Gauss-Newton method; but encountering situations where the 2nd derivative is negative, it reverts to the steepest descent method, and uses only the first derivative.

ii) Transfer function: Transfer functions calculate a neural layer's output from its net input. Here, hyperbolic tangent sigmoid function (Vogl, Mangis, Rigler, Zink, &

Alkon, 1988) is used. It can be mathematically represented as given in equation 33 where *la* stands for *linear activation* of a neuron as shown earlier in equation 32.

$$\tanh(la) = \frac{2}{1+e^{-2\times la}} - 1 \qquad (33)$$

The final outputs obtained from the feedforward network are given in Table 6.5. These results are indeed better than all the other approaches examined so far.

Table 6.5 Results from feedforward network

Data	Used to test against	Corr. coef. (r)	Best validation performance at (MSE)	CV of r
Whole dataset (cDNA ~ Affyratio)	-	0.6253 (TO, i.e., Training output)	0.5187	-
cDNA3 ~ Affyratio3	-	0.6632 (TO)	0.4728	-
	13	0.5801	0.5114	16.45
	29	0.6156	0.5693	
	75	0.4891	0.3943	
	76	0.6229	0.7576	
	78	0.4188	0.8155	
	79	0.6646	0.4829	
cDNA13 ~ Affyratio13	-	0.6287 (TO)	0.4854	-
	3	0.6410	0.6383	10.77
	29	0.5783	0.5970	
	75	0.5200	0.3793	
	76	0.6506	0.7136	
	78	0.5157	0.7260	
	79	0.6497	0.5056	
cDNA29 ~ Affyratio29	-	0.5884 (TO)	0.4694	-
	3	0.6579	0.6179	11.61
	13	0.6130	0.4974	
	75	0.5106	0.3797	
	76	0.6294	0.7411	
	78	0.5081	0.7359	
	79	0.6567	0.5014	
cDNA75 ~ Affyratio75	-	0.5407 (TO)	0.3672	-
	3	0.6421	0.6403	10.12
	13	0.5823	0.5156	
	29	0.5530	0.6122	
	76	0.6274	0.7626	
	78	0.4943	0.7303	
	79	0.6467	0.5120	

Data	Used to test against	Corr. coef. (r)	Best validation performance at (MSE)	CV of r
cDNA76 ~ Affyratio76	-	0.6510 (TO)	0.4259	-
	3	0.6421	0.6402	10.18
	13	0.6091	0.4917	
	29	0.5751	0.6049	
	75	0.5266	0.3793	
	78	0.5013	0.7405	
	79	0.6421	0.5181	
cDNA78 ~ Affyratio78	-	0.5431 (TO)	0.6561	-
	3	0.6138	0.6821	7.69
	13	0.6020	0.4958	
	29	0.5547	0.6257	
	75	0.5101	0.3890	
	76	0.6222	0.7532	
	79	0.6193	0.5459	
cDNA79 ~ Affyratio79	-	0.6721 (TO)	0.45201	-
	3	0.6596	0.6078	15.47
	13	0.5909	0.5161	
	29	0.6122	0.5673	
	75	0.4969	0.3954	
	76	0.6341	0.7505	
	78	0.4311	0.7702	
Training set (with 4504 data)	-	0.6267 (TO)	0.5400	-
	Test set: 1000 data	0.6042	0.4962	-

6.2 Summary of results

All the various types of statistical and machine learning approaches have been rigorously applied above. The idea behind is based on exploring whether and how useful the methods would be when applied to microarray data in a situation when they come from two separate platforms, and when they have passed through a data transformation phase.

Broadly, the available results provided in Table 6.1 to Table 6.5 can be studied by comparing the model outputs concerning the whole and the random dataset. Table 6.6 summarises these results. With the simple neural architecture, the feedforward network is able to present the best results, while cubic-polynomial delivers the next best set of results. The summary table also shows that despite its enormous potential, bootstrap aggregating method has failed to deliver a comparable outcome than the rest of the methods. The self-organizing maps (SOM) are used by various researchers to constitute a very powerful and unsupervised data visualization technique. This technique has been probed into and redesigned to make it operational to address the current task, which otherwise falls outside of its usual application environment. This redesigning of SOM's application makes it capable of bringing better outcomes than the bagging method, but comes out to be relatively less effective than the remaining approaches.

Table 6.6 Summary of results

Model	**Whole dataset**		**Random dataset**			
			Training set		Test set	
	MSE	*r*	*MSE*	*r*	*MSE*	*r*
Linear	0.6013	0.5886	0.6172	0.5892	0.5299	0.5771
Polynomial (cubic)	0.5842	0.6042	0.5979	0.6064	0.5140	0.5835
Bootstrap aggregating (bagging)	1.4870	0.5938	1.4806	0.5956	1.5588	0.5872
Self Organizing Maps (SOMs)	0.8994	0.4650	0.9316	0.4643	0.8586	0.4405
Feedforward network	0.5187	0.6253	0.5400	0.6267	0.4962	0.6042

A look at the tables in Table 6.1 to Table 6.5 also suggests that at the level of individual patients, the predictive gene expression of atleast one patient, viz., patient number 78, produces at times results that tend to exceed the range of the outputs obtained by the others. However, it is difficult to question its data quality as while carrying out the elaborate data quality assessment in Chapter 3, no indication could be deduced regarding the presence of any serious faults in any of these arrays. The predicted results of the remaining patients are found to be more or less similar.

Further, in the final segment ahead, Chapter 7 delivers the closing remarks.

Note :

- "*Writing Scholarship, 2010*" awarded on merit by Lincoln University, Christchurch, New Zealand based on a proposal for a research article from aspects of this chapter.

Chapter 7
Closing Remarks

7.1 Summary

DNA is the magic molecule that encodes all the information required for the development and functioning of an organism; and microarrays are a tool used to reveal an unprecedented view into the biology of DNA. With the advent of individual experiments generating thousands of data or observations, a hypothesis-driven endeavour has turned into hypothesis-generating endeavour that flashes light across an entire terrain of gene expressions. Joining what used to be primarily wet science, information science moulded it skilfully into an ever rejuvenating branch of science while incessantly contributing to further streamlining the processes involved.

Microarrays afford the luxury that gene expressions can be measured in any of its multiple platforms. The impediment, however, appears as user tries to jointly study multiple platforms. Various comparison studies have been published presenting completely contradictory results - some have observed agreement in results obtained with different platforms (Barczak et al., 2003; Carter et al., 2003; Hughes et al., 2001; Kane et al., 2000; H. Y. Wang et al., 2003; Yuen, Wurmbach, Pfeffer, Ebersole, & Sealfon, 2002), others have not at all (Kothapalli, Yoder, Mane, & Loughran, 2002; W. P. Kuo et al., 2002; J. Li, Pankratz, & Johnson, 2002; Tan et al., 2003). A review on various notable works in the direction of cross-platform integration of microarray data is presented in Chapter 3. However, all these published methods have their own advantages as well as disadvantages.

In the midst of the relentless chase to find remedies for the issues of microarray data integration, is there a chance that an answer is lying underneath the nature of the microarray data itself ? This was the question set for answering while commencing the attempt of cross-platform integration of data from Affymetrix and cDNA microarray platforms. Data provided by *The Children's Hospital* at Westmead, Australia contained the much-needed cancer patients' data, where the patients were reportedly tested on both the platforms.

Keeping in mind the nature of the resultant microarray data from these platforms, a new ratio-transformation method has been proposed and applied to the data. It subsequently highlights that its application can address the issue of incomparability of the expression measures of

Affymetrix and cDNA platforms. The method is later tested against two established approaches, and is found to produce comparative results.

The encouraging outcome from the above method led to focus attention on examining further in the direction of defining the association between the two platforms. With this motivation, a wide range of statistical as well as machine learning approaches is applied to the microarray data. Finally, the existing relationship between the data from the two platforms is found to be nonlinear, which can be well-delineated by feedforward network with relatively more precision than the rest of the methods tested.

7.2 Conclusions

The focus of the work carried out in this research remains in the gene expression levels of two specific platforms, Affymetrix and cDNA. Summarily, the work presents a novel as well as an alternative way of integrating expression levels from these two platforms. The approach is relatively uncomplicated compared to its counterparts; and while taking a different standpoint to the problem of data integration across microarray platforms, it delivers better results compared to the conventional ways of integration of gene expression levels. It also produces close results when tested with a popular method, *DWD* (Benito et al., 2004; Marron et al., 2007). Further, another major highlight of this work is its distinctively extensive exploration implementing a wide range of statistical as well as machine learning approaches towards drawing the closest association between the two platforms. The resultant output from this segment of the study suggests that the relation between the two microarray platforms is non-linear; and given a gene’s expression level in one platform, there is a possibility that a feedforward neural network would provide more accurate expression value of the gene in the other platform compared to the rest of the approaches trialled.

7.3 Advantages and Limitations

There are methods available for microarray data integration for large sample data, such as the *Probability of Expression* method (Parmigiani et al., 2002; Shen et al., 2004) and *XPN* (Shabalin et al., 2008). However, these methods are many times found to be unusable for set ups involving small microarray sample size. Besides being a non-complex exploit, the ratio-transformation approach can be applied to both small and large sample data. Further, it works on the true expression measures unlike several other methods, where the core component in the data integration methodology involves transforming the data using measures, such as distance (Benito et al., 2004; Marron & Todd, 2002), probability scale (Parmigiani et al., 2002; Shen et al., 2004), ranking of fold change (Breitling et al., 2004; F. Hong et al., 2006) etc. as discussed in Chapter 3. While comparing the ratio-transformation approach with gene-centering and DWD-method in Chapter 5, DWD provided a slight improvement over ratio-transformation method. However, there are a few virtues of the ratio-transformation method that deserve highlighting.

The ratio-transformation method is an attempt to view the problem of cross platform data-integration from a different perspective of concentrating the focus of investigation on the nature of the generated data from the platforms. The approach is crafted based on the fundamental characteristics of the two platforms as well as on the prominent distinguishing features of their relationship; and therefore, has evolved from a sound base providing the required rigor. It also furnishes greater transparency as well as simpler applicability enabling a prospective user to relate to it depending on the basic knowledge about microarray technology in general while attaining similar or higher level of accuracy delivered by a variety of available complex statistical and machine learning approaches. From this view point, this approach can counterbalance any apparent advantages of other available methods, specifically DWD. DWD method finds a separating hyperplane between the two microarray batches, and adjusts the data by projecting the different batches on the DWD plane, finds the batch mean, and then subtracting out the DWD plane multiplied by this mean. With regards to the DWD-approach, Johnson & Li (2007) confirms that researchers face difficulties while trying to implement this method, and a few of the difficulties include that the method is "fairly complicated", and can be applied to only two batches at a time. In an example of DWD, a stepwise approach is used by Benito et al. (2004) - first adjusting the two most similar batches, and then comparing the third against the previous (adjusted) two. This stepwise method provides reasonable results in their three-batch case, but this could potentially break down in cases where there are many more batches or when batches are not very similar. Further, the DWD approach may also be considered as a black-box method, which tends to fall short of providing much insight into the process underneath.

Further, given an expression level of a differentially expressed gene of one platform, the investigations on between-platform association intends to provide a framework which presents an estimate of the possible expression value in the other platform. However, it is possible to critique this attempt to be a prototypical rather than a method of global generalization as it has been conducted on a relatively small sample space. However, it is unlikely that investigating with a larger set of data would present a greatly exceeding outcome because the current sample space can also be assumed as a random sample from a larger dataset. Thus, the overall output of the large hypothetical dataset can be expected to follow a trend similar to that of the current output.

7.4 The Road Ahead

Microarray platform integration study conducted here can be considered as a foundation in an attempt of exploration based on the nature of the resulting data of Affymetrix and cDNA platforms. Further consolidation of information on the basis of various aspects on the background of the data, gene-wise information and relevant key facts are expected to provide finer predictions with higher accuracy. This is a promising road of investigations ahead, although without contesting the fact that the task would involve substantial information warehousing, increased computing power as well as high-end computational skills. However, this line of interrogations can potentially contribute towards bringing down the curtain on the differences between the Affymetrix and cDNA platforms.

7.5 Final Remarks

Microarray technology has strongly emerged due to the fact that it can provide a rapid snapshot of gene expression pattern of a tissue. It also helps in our understanding of global networks of bio-molecular interactions. Scientific areas including diagnosis, drug development, functional genomics, and comparative genomics are stimulated with the development of this high throughput technique resulting in avalanche of data from innumerable number of experiments.

With the emergence of microarray technology from the shadows of being ‘cautionary tale’ (Sherlock, 2005), the steps towards the growth in the area of microarray data integration have already been initiated. This thesis is a further exploration in this direction, however, viewing the domain as well as the question from a distinctly separate perspective. The conducted work maintains the highest housekeeping standards, besides carrying out a series of trials and testings with the use of a wide range of applications, methods and algorithms. Subsequently, the process is believed to have put its own contribution in the parade of unlocking the hidden treasures of biological knowledge.

References

Aach, J., Rindone, W., & Church, G. M. (2000). Systematic management and analysis of yeast gene expression data. *Genome Res, 10*(4), 431-445.

Affymetrix. (2002). Statistical algorithms description document. *Affymetrix technical manual*.

Affymetrix. (2008). GeneChip expression analysis technical manual. *Affymetrix technical manual*.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Albertson, D. G., & Pinkel, D. (2003). Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet, 12 Spec No 2,* R145-152.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature, 403*(6769), 503-511.

Alwine, J. C., Kemp, D. J., & Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci USA, 74*(12), 5350-5354.

Amos, M. (2005). *Theoretical and experimental DNA computation* (1st ed.). New Jersey: Springer.

Babu, M. M. (2004). An introduction to microarray data analysis. In R. P. Grant (Ed.), *Computational Genomics: Theory and Application* (pp. 225-249). Norwich: Horizon Bioscience.

Bains, W., & Smith, G. C. (1988). A novel method for nucleic acid sequence determination. *J Theor Biol, 135*(3), 303-307.

Baird, D., Johnstone, P., & Wilson, T. (2004). Normalization of microarray data using a spatial mixed model analysis which includes splines. *Bioinformatics, 20*(17), 3196-3205.

Barczak, A., Rodriguez, M. W., Hanspers, K., Koth, L. L., Tai, Y. C., Bolstad, B. M., et al. (2003). Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Research, 13*(7), 1775-1785.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., et al. (2004). Adjustment of systematic microarray data biases. *Bioinformatics, 20*(1), 105-114.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc., 57*, 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist., 29*(4), 1165-1188.

Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F. F., Rapp, B. A., et al. (1999). GenBank. *Nucleic Acids Research, 27*(1), 12-17.

Bentley, D. R. (2000). The Human Genome Project - an overview. *Medicinal Research Reviews, 20*(3), 189-196.

Bloom, G., Yang, I. V., Boulware, D., Kwong, K. Y., Coppola, D., Eschrich, S., et al. (2004). Multi-platform, multi-site, microarray-based human tumor classification. *American Journal of Pathology, 164*(1), 9-16.

Bolstad, B. M. (2004). *Low level analysis of high-density oligonucleotide array data: background, normalization and summarization.* University of California at Berkeley, Berkeley.

Bolstad, B. M. (2006). Pre-processing DNA microarray data. In W. Dubitzky, M. Granzow & D. P. Berrar (Eds.), *Fundamentals of data mining in genomics and proteomics* (pp. 51 - 76). NY, USA: Springer.

Bolstad, B. M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R. A., et al. (2005). Quality assessment of Affymetrix GeneChip data. In W. Wong, M. Gail, K.

Krickeberg, A. Tsiatis & J. Samet (Series Eds.), R. Gentleman, V. Carey, W. Huber, R. Irizarry & S. Dudoit (Eds.), *Bioinformatics and computational biology solutions using R and Bioconductor* (pp. 33-47). NY, USA: Springer.

Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics, 19*(2), 185-193.

Boothroyd, J. C., Blader, I., Cleary, M., & Singh, U. (2003). DNA microarrays in parasitology: strengths and limitations. *Trends Parasitol, 19*(10), 470-476.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell, 122*(6), 947-956.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., et al. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. . *Nature Genetics, 29*(4), 365–371.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., et al. (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res, 31*(1), 68-71.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123 - 140.

Breitling, R., Armengaud, P., Amtmann, A., & Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters, 573*(1-3), 83-92.

Buhler, J., Ideker, T., & Haynor, D. (2000). Dapple: improved techniques for finding spots on DNA microarrays. In *CSE Technical Report*. Missouri: Washington University in St. Louis.

Bunney, W. E., Bunney, B. G., Vawter, M. P., Tomita, H., Li, J., Evans, S. J., et al. (2003). Microarray technology: a review of new strategies to discover candidate vulnerability genes in psychiatric disorders. *Am J Psychiatry, 160*(4), 657-666.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: a practical information-theoretic approach* (2nd ed.). New York: Springer-Verlag.

Cahan, P., Ahmad, A. M., Burke, H., Fu, S., Lai, Y., Florea, L., et al. (2005). List of lists-annotated (LOLA): a database for annotation and comparison of published microarray gene lists. *Gene, 360*(1), 78-82.

Canales, R. D., Luo, Y., Willey, J. C., Austermiller, B., Barbacioru, C. C., Boysen, C., et al. (2006). Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol, 24*(9), 1115-1122.

Carig, A. G., Nizetic, D., Hoheisel, J. D., Zehetner, G., & Lehrach, H. (1990). Ordering of cosmid clones covering the Herpes Simplex virus type I (HSV-I) genome: a test case for fingerprinting by hybridisation. *Nucleic Acids Res, 18*(9), 2653-2660.

Carter, M. G., Hamatani, T., Sharov, A. A., Carmack, C. E., Qian, Y., Aiba, K., et al. (2003). In situ-synthesized novel microarray optimized for mouse stem cell and early developmental expression profiling. *Genome Res, 13*(5), 1011-1021.

Causton, H., Quackenbush, J., & Brazma, A. (2003). *Microarray gene expression data analysis: a beginner's guide*. Massachusetts, USA: Blackwell. (2003)

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). Studying two-dimensional data. In *Graphical methods for data analysis* (pp. 75-127). California: Wadsworth & Brooks/Cole.

Cheadle, C., Vawter, M. P., Freed, W. J., & Becker, K. G. (2003). Analysis of microarray data using Z score transformation. *J Mol Diagn, 5*(2), 73-81.

Choi, J. K., Yu, U., Kim, S., & Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics, Vol. 19* (Suppl. 1), 84-90.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*(368), 829-836.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83*(403), 596-610.

Coller, H. A., Grandori, C., Tamayo, P., Colbert, T., Lander, E. S., Eisenman, R. N., et al. (2000). Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc Natl Acad Sci USA, 97*(7), 3260-3265.

Cook, D. J., Sackett, D. L., & Spitzer, W. O. (1995). Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam consultation onmeta-analysis. *Journal of Clinical Epidemiology, 48*(1), 167–171.

Cope, L., Zhong, X., Garrett, E., & Parmigiani, G. (2004). MergeMaid: R tools for merging and cross-study validation of gene expression data. *Stat Appl Genet Mol Biol, 3,* Article29.

Cotton, R. G., Rodrigues, N. R., & Campbell, R. D. (1988). Reactivity of cytosine and thymine in single-base-pair mismatches with hydroxylamine and osmium tetroxide and its application to the study of mutations. *Proc Natl Acad Sci USA, 85*(12), 4397–4401.

Crowther, D. J. (2002). Applications of microarrays in the pharmaceutical industry. *Curr Opin Pharmacol, 2*(5), 551-554.

Dabney, A. R., & Storey, J. D. (2007). Normalization of two-channel microarrays accounting for experimental design and intensity-dependent relationships. *Genome Biol, 8*(3), R44.

Debouck, C., & Goodfellow, P. N. (1999). DNA microarrays in drug discovery and development. *Nat Genet, 21*(1 Suppl), 48-50.

Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol, 4*(5), P3.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., et al. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet, 14*(4), 457-460.

DeRisi, J. L., Iyer, V. R., & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science, 278*(5338), 680-686.

Dobbin, K., Shih, J. H., & Simon, R. (2003). Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *J Natl Cancer Inst, 95*(18), 1362-1369.

Dobbin, K., Shih, J. H., & Simon, R. (2003). Statistical design of reverse dye microarrays. *Bioinformatics, 19*(7), 803-810.

Draghici, S. (2002). Statistical intelligence: effective analysis of high-density microarray data. *Drug Discov Today, 7*(11 Suppl), S55-63.

Draghici, S. (2005). *Data analysis tools for DNA microarrays*. Boca Raton, USA: Chapman & Hall/CRC. (2003)

Draghici, S., Khatri, P., Eklund, A. C., & Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet., 22* (2), 101-109.

Dresen, I. M., Husing, J., Kruse, E., Boes, T., & Jockel, K. H. (2003). Software packages for quantitative microarray-based gene expression analysis. *Curr Pharm Biotechnol, 4*(6), 417-437.

Drmanac, R., Labat, I., Brukner, I., & Crkvenjakov, R. (1989). Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics, 4*(2), 114-128.

Dudoit, S., & Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics, 19*(9), 1090-1099.

Dudoit, S., Gentleman, R. C., & Quackenbush, J. (2003). Open source software for the analysis of microarray data. *Biotechniques, Suppl*, 45-51.

Dudoit, S., Yang, Y. H., Callow, M., & Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sinica, 12*, 111–140.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res, 30*(1), 207-210.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics, 7*(1), 1 – 26.

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Emperical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association, 96*(456), 1151-1160.

Eysenck, H. J. (1995). Problems with meta-analysis. In I. Chalmers & D. G. Altman (Eds.), *Systematic Reviews* (pp. 64-74). London: BMJ Publishing Group.

Favis, R., & Barany, F. (2000). Mutation detection in K-ras, BRCA1, BRCA2, and p53 using PCR/LDR and a universal DNA microarray. *Ann N Y Acad Sci, 906*, 39-43.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., & Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science, 251*(4995), 767-773.

Fortin, A., Cregan, S. P., MacLaurin, J. G., Kushwaha, N., Hickman, E. S., Thompson, C. S., et al. (2001). APAF1 is a key transcriptional target for p53 in the regulation of neuronal cell death. *J Cell Biol, 155*(2), 207-216.

Galbraith, D. W., & Edwards, J. (2010). Applications of microarrays for crop improvement: here, there, and everywhere. *Bioscience, 60*(5), 337-348.

Gardiner-Garden, M., & Littlejohn, T. G. (2001). A comparison of microarray databases. *Brief Bioinform, 2*(2), 143-158.

Garrett-Mayer, E., Parmigiani, G., Zhong, X., Cope, L., & Gabrielson, E. (2008). Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics, 9*(2), 333-354.

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). Affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics, 20*(3), 307-315.

Geman, D., d'Avignon, C., Naiman, D. Q., & Winslow, R. L. (2004). Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol, 3*, Article19.

Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A., & Dudoit, S. (2005). *Bioconductor and computational biology solutions using R and Bioconductor*. NY: Springer.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol, 5*(10), R80.

Ghosh, D., Barette, T. R., Rhodes, D., & Chinnaiyan, A. M. (2003). Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Functional and Integrative Genomics, 3*(4), 180-188.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science, 286*(5439), 531-537.

Gracey, A. Y., & Cossins, A. R. (2003). Application of microarray technology in environmental and comparative physiology. *Annu Rev Physiol, 65*, 231-259.

Grewal, A., Lambert, P., & Stockton, J. (2007). Analysis of expression data: an overview. *Curr Protoc Bioinformatics, Chapter 7*, Unit 7.1.

Grutzmann, R., Boriss, H., Ammerpohl, O., Luttges, J., Kalthoff, H., Schackert, H. K., et al. (2005). Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene, 24*(32), 5079–5088.

Hacia, J. G., Brody, L. C., Chee, M. S., Fodor, S. P. A., & Collins, F. S. (1996). Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two‒colour fluorescence analysis. *Nature Genetics, 14*(4), 441 - 447.

Hahne, F., Huber, W., Gentleman, R., & Falcom, S. (2008). *Bioconductor case studies*. NY: Springer.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase ofhumangenes and genetic disorders. *Nucleic Acids Research, 33*(Database issue), D514–D517.

Harkin, D. P., Bean, J. M., Miklos, D., Song, Y. H., Truong, V. B., Englert, C., et al. (1999). Induction of GADD45 and JNK/SAPK-dependent apoptosis following inducible expression of BRCA1. *Cell, 97*(5), 575-586.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science, Vol. 1, No. 3 (Aug., 1986), pp. 297-310, 1*(3), 297-310.

Hayes, P. C., Wolf, C. R., & Hayes, J. D. (1989). Blotting techniques for the study of DNA, RNA, and proteins. *BMJ, 299*, 965–968.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hiley, S. L., Jackman, J., Babak, T., Trochesset, M., Morris, Q. D., Phizicky, E., et al. (2005). Detection and discovery of RNA modifications using microarrays. *Nucleic Acids Res, 33*(1), e2.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802.

Hoheisel, J. D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet, 7*(3), 200-210.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*(2), 383-386.

Hong, F., & Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics, 24*(3), 374-382.

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., & Chory, J. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics, 22*(22), 2825-2827.

Hu, P., Greenwood, C. M., & Beyene, J. (2005). Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *Bmc Bioinformatics, 6*(1), 128-138.

Hua, Y. J., Tu, K., Tang, Z. Y., Li, Y. X., & Xiao, H. S. (2008). Comparison of normalization methods with microRNA microarray. *Genomics, 92*(2), 122-128.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., et al. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol, 19*(4), 342-347.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., et al. (2000). Functional discovery via a compendium of expression profiles. *Cell, 102*(1), 109-126.

Hughes, T. R., Roberts, C. J., Dai, H., Jones, A. R., Meyer, M. R., Slade, D., et al. (2000). Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat Genet, 25*(3), 333-337.

Hwang, K. B., Kong, S. W., Greenberg, S. A., & Park, P. J. (2004). Combining gene expression data from different generations of oligonucleotide arrays. *Bmc Bioinformatics, 5*, 159.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *J. Comput. Graph. Statist., 5*, 299–314.

Ikeo, K., Ishi-i, J., Tamura, T., Gojobori, T., & Tateno, Y. (2003). CIBEX: center for information biology gene expression database. *Comptes rendus Biologies, 326*(10-11), 1079-1082.

Imbeaud, S., & Auffray, C. (2005). 'The 39 steps' in gene expression profiling: critical issues and proposed best practices for microarray experiments. *Drug Discov. Today, 10*(17), 1175-1182.

Insuk, S., Sujong, K., Changha, H., & Jae Won, L. (2008). New normalization methods using support vector machine quantile regression approach in microarray analysis. *Comput. Stat. Data Anal., 52*(8), 4104 - 4115.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics, 4*(2), 249-264.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., et al. (2005). Multiple-laboratory comparison of microarray platforms. *Nat Methods, 2*(5), 345-350.

Irizarry, R. A., Wu, Z., & Jaffee, H. A. (2006). Comparison of Affymetrix GeneChip expression measures. *Bioinformatics, 22*(7), 789-794.

Jiang, C. H., Tsien, J. Z., Schultz, P. G., & Hu, Y. (2001). The effects of aging on gene expression in the hypothalamus and cortex of mice. *Proc Natl Acad Sci USA, 98*(4), 1930-1934.

Jiang, H., Deng, Y., Chen, H.-S., Tao, L., Sha, Q., Chen, J., et al. (2004). Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *Bmc Bioinformatics, 5*, 81.

Joe H. Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236-244.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241-254.

Johnson, W. E., & Li, C. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*(1), 118-127.

Jouenne, V. Y. (2001). *Critical issues in the processing of cDNA microarray images.* Unpublished Masters Thesis, Virginia Polytechnic Institute and State University, Virginia.

Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D., & Madore, S. J. (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res, 28*(22), 4552-4557.

Kari, L. (1997). DNA computing: Arrival of biological mathematics. *The Mathematical Intelligencer, 19*(2), 9-22.

Kari, L. (2001). DNA computing in vitro and in vivo. *Future Generation Computer Systems, 17*(7), 823-834.

Kari, L., & Landweber, L. F. (2000). Computing with DNA. In S. Misener & S. A. Krawetz (Eds.), *Bioinformatics: Methods and Protocols* (1 ed., Vol. 132, pp. 413-431). New Jersey: Humana Press.

Kauffmann, A., Gentleman, R., & Huber, W. (2009). arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. *Bioinformatics, 25*(3), 415-416.

Khanna, C., Khan, J., Nguyen, P., Prehn, J., Caylor, J., Yeung, C., et al. (2001). Metastasis-associated differences in gene expression in a murine model of osteosarcoma. *Cancer Res, 61*(9), 3750-3759.

Khodursky, A. B., Peter, B. J., Schmid, M. B., DeRisi, J., Botstein, D., Brown, P. O., et al. (2000). Analysis of topoisomerase function in bacterial replication fork movement: use of DNA microarrays. *Proc Natl Acad Sci USA, 97*(17), 9419-9424.

Khrapko, K. R., Lysov Yu, P., Khorlyn, A. A., Shick, V. V., Florentiev, V. L., & Mirzabekov, A. D. (1989). An oligonucleotide hybridization approach to DNA sequencing. *FEBS Lett*, *256*(1-2), 118-122.

Kohonen, T. (1982a). Clustering, taxonomy and topological maps of patterns. In *Proceedings of the 6th International Conference on Pattern Recognition* (pp. 114-128). Munich, Germany.

Kohonen, T. (1982b). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*(1), 59-69.

Kohonen, T. (1998). The self-organizing maps. *Neurocomputing, 21*(1-3), 1-6.

Kohonen, T. (2001). *Self-Organizing Maps* (3rd ed.). NY: Springer.

Kothapalli, R., Yoder, S. J., Mane, S., & Loughran, T. P. (2002). Microarray results: how accurate are they? *Bmc Bioinformatics, 3*(22).

Kudoh, K., Ramanna, M., Ravatn, R., Elkahloun, A. G., Bittner, M. L., Meltzer, P. S., et al. (2000). Monitoring the expression profiles of doxorubicin-induced and doxorubicin-resistant cancer cells by cDNA microarray. *Cancer Res, 60*(15), 4161-4166.

Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L., & Kohane, I. S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics, 18*(3), 405-412.

Kuo, W. P., Liu, F., Jensen, T.-K., Benson, S. L., Cepko, C. L., Hovig, E., et al. (2009). A systematic comparison of gene expression measurements across different hybridization-based technologies. In G. Hardiman (Ed.), *Microarray innovations: technology and experimentation* (1st ed.). Boca Raton: CRC Press.

Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R., & Quackenbush, J. (2005). Independence and reproducibility across microarray platforms. *Nat Methods, 2*(5), 337-344.

Larsson, O., Wennmalm, K., & Sandberg, R. (2006). Comparative microarray analysis. *Omics, 10*(3), 381–397.

Lee, M.-L. T., Kuo, F. C., Whitmorei, G. A., & Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci USA, 97*(18), 9834-9839.

Lee, P. D., Sladek, R., Greenwood, C. M. T., & Hudson, T. J. (2002). Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res, 12*(2), 292-297.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science, 298*(5594), 799-804.

Li, C., & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA, 98*(1), 31-36.

Li, J., Pankratz, M., & Johnson, J. A. (2002). Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. *Toxicol Sci, 69*(2), 383-390.

Li, L., Li, R. Y., & Li, C. T. (2005). [SNP genotyping by multiplex amplification and microarrays assay and forensic application]. *Fa Yi Xue Za Zhi, 21*(2), 90-95.

Liotta, L., & Petricoin, E. (2000). Molecular profiling of human cancer. *Nat Rev Genet, 1*(1), 48-56.

Lu, C., Devos, A., Suykens, J. A. K., Arus, C., & Huffel, S. V. (2007). Bagging linear sparse Bayesian learning models for variable selection in cancer diagnosis. *IEEE Trans Inf Technol Biomed, 11*(3), 338-347.

Lucchini, S., Thompson, A., & Hinton, J. C. (2001). Microarrays for microbiologists. *Microbiology, 147*(Pt 6), 1403-1414.

Macaulay, F. R. (1931). *The smoothing of time series*. California: National Bureau of Economic Research, Inc.

MacBeath, G. (2002). Protein microarrays and proteomics. *Nat Genet, 32 Suppl*, 526-532.

Macgregor, P. F. (2003). Gene expression in cancer: the application of microarrays. *Expert Rev Mol Diagn, 3*(2), 185-200.

MacLachlan, T. K., Somasundaram, K., Sgagias, M., Shifman, Y., Muschel, R. J., Cowan, K. H., et al. (2000). BRCA1 effects on the cell cycle and the DNA damage response are linked to altered gene expression. *J Biol Chem, 275*(4), 2777-2785.

Magi, R., Pfeufer, A., Nelis, M., Montpetit, A., Metspalu, A., & Remm, M. (2007). Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation. *BMC Genomics, 8*, 159.

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2006). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research, 00*(Database issue), D1–D6.

Mangalam, H., Stewart, J., Zhou, J., Schlauch, K., Waugh, M., Chen, G., et al. (2001). GeneX: an open source gene expression database and integrated tool set. *IBM Systems Journal, 40*(2), 552-569.

Mar, J. C., Kimura, Y., Schroder, K., Irvine, K. M., Hayashizaki, Y., Suzuki, H., et al. (2009). Data-driven normalization strategies for high-throughput quantitative RT-PCR. *Bmc Bioinformatics, 10*, 110.

Marmur, J., & Doty, P. (1961). Thermal renaturation of deoxyribonucleic acids. *Journal of Molecular Biology, 3*(5), 585-594.

Marron, J. S., & Todd, M. J. (2002). *Distance Weighted Discrimination*. NY: Cornell University.

Marron, J. S., Todd, M. J., & Ahn, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association, 102*(480), 1267-1271.

McGee, M., & Chen, Z. (2006). Parameter estimation for the exponential-normal convolution model for background correction of affymetrix GeneChip data. *Stat Appl Genet Mol Biol, 5*(1), Article24.

Miller, R. G. J. (1981). *Simultaneous statistical inference* (2 ed.). NY: Springer.

More, J. J. (1977). The Levenberg-Marquardt algorithm: implementation and theory. In G. A. Watson (Ed.), *Numerical Analysis* (Vol. 630, pp. 105-116). Berlin, Heidelberg, New York: Springer-Verlag.

Moreau, Y., Aerts, S., De Moor, B., De Strooper, B., & Dabrowski, M. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends in Genetics, 19*(10), 570-577.

Mori, T., Anazawa, Y., Matsui, K., Fukuda, S., Nakamura, Y., & Arakawa, H. (2002). Cyclin K as a direct transcriptional target of the p53 tumor suppressor. *Neoplasia, 4*(3), 268-274.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol, 51 Pt 1*, 263-273.

Murphy, D. (2002). Gene expression studies using microarrays: principles, problems, and prospects. *Adv Physiol Educ., 26*(1-4), 256-270.

Myers, R. M., Maniatis, T., & Lerman, L. S. (1987). Detection and localization of single base changes by denaturing gradient gel electrophoresis. *Methods Enzymol, 155*, 501-527.

Nadon, R., & Shoemaker, J. (2002). Statistical issues with microarrays: processing and analysis. *Trends in Genetics, 18*(5), 265-271.

Newman, J. C., & Weiner, A. M. (2005). L2L: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biol, 6*(9), R81.

Newton, M. A., Noueiry, A., Sarkar, D., & Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics, 5*(2), 155-176.

Nishizuka, I., Ishikawa, T., Hamaguchi, Y., Kamiyama, M., Ichikawa, Y., Kadota, K., et al. (2002). Analysis of gene expression involved in brain metastasis from breast cancer using cDNA microarray. *Breast Cancer, 9*(1), 26-32.

Normand, S. L. (1999). Tutorial in biostatistics-meta-analysis: formulating, evaluating, combining, and reporting. *Stat. Med*, *18*, 321–359.

Nuwaysir, E. F., Bittner, M., Trent, J., Barrett, J. C., & Afshari, C. A. (1999). Microarrays and toxicology: the advent of toxicogenomics. *Mol Carcinog*, *24*(3), 153-159.

Orita, M., Suzuki, Y., Sekiya, T., & Hayashi, K. (1989). Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics*, *5*(4), 874-879.

Oshlack, A., Emslie, D., Corcoran, L. M., & Smyth, G. K. (2007). Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. *Genome Biol*, *8*(1), R2.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, *12*, 546–554.

Park, P. J., Cao, Y. A., Lee, S. Y., Kim, J. W., Chang, M. S., Hart, R., et al. (2004). Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. *J Biotechnol*, *112*(3), 225-245.

Park, T., Yi, S. G., Kang, S. H., Lee, S., Lee, Y. S., & Simon, R. (2003). Evaluation of normalization methods for microarray data. *Bmc Bioinformatics*, *4*, 33.

Parker, R., & Sheth, U. (2007). P bodies and the control of mRNA translation and degradation. *Mol Cell*, *25*(5), 635-646.

Parmigiani, G., Garrett, E. S., Anbazhagan, R., & Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *J. R. Statist. Soc. B*, *64*(4), 717-736.

Politis, D. N. (2008). Bagging multiple comparisons from microarray data. In I. Mandoiu, R. Sunderraman & A. Zelikovsky (Eds.), *Bioinformatics Research and Applications* (Vol. 4983, pp. 492-503). Berlin, Heidelberg: Springer.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., et al. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*, *23*(1), 41-46.

Pontius, J. U., Wagner, L., & Schuler, G. D. (2003). UniGene: a unified view of the transcriptome. In J. McEntyre & J. Ostell (Eds.), *The NCBI Handbook*. Bethesda (MD): National Library of Medicine (US).

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2006). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *00*(Database issue), D1–D5.

Pusztai, L., Ayers, M., Stec, J., & Hortobagyi, G. N. (2003). Clinical application of cDNA microarrays in oncology. *Oncologist*, *8*(3), 252-258.

Qin, L. X., Beyer, R. P., Hudson, F. N., Linford, N. J., Morris, D. E., & Kerr, K. F. (2006). Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *Bmc Bioinformatics*, *7*, 23.

Ramaswamy, S., Ross, K. N., Lander, E. S., & Golub, T. R. (2003). A molecular signature of metastasis in primary solid tumors. *Nat Genet*, *33*(1), 49-54.

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., & Chinnaiyan, A. M. (2002). Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, *62*(15), 4427-4433.

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., et al. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, *23*(20), 2700–2707.

Robinson, M. D., & Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, *23*(21), 2881-2887.

Rosenblatt, F. (1962). *Principles of neurodynamics: perceptrons and the theory of brain mechanism*. Washington, DC: Spartan Books.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet, 24*(3), 227-235.

Saal, L. H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A., & Peterson, C. (2002). BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol, 3*(8), SOFTWARE0003.

Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A., et al. (2006). TM4 microarray software suite. *Methods Enzymol, 411*, 134-193.

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., et al. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques, 34*(2), 374-378.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., et al. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science, 230*(4732), 1350-1354.

Sakamoto, M., Kondo, A., Kawasaki, K., Goto, T., Sakamoto, H., Miyake, K., et al. (2001). Analysis of gene expression profiles associated with cisplatin resistance in human ovarian cancer cell lines and tissues using cDNA microarray. *Hum Cell, 14*(4), 305-315.

Samarasinghe, S. (2007). *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition*. New York: Auerbach.

Sauter, G., Simon, R., & Hillan, K. (2003). Tissue microarrays in drug discovery. *Nat Rev Drug Discov, 2*(12), 962-972.

Schadt, E. E., Li, C., Ellis, B., & Wong, W. H. (2001). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl, Suppl 37*, 120-125.

Schadt, E. E., Li, C., Ellis, B., & Wong, W. H. (2001). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data *Journal of Cellular Biochemistry, 84*(S37), 120-125.

Schena, M. (2003). *Microarray Analysis*. New Jersey, USA: Wiley-Liss.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270*(5235), 467–470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., & Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA, 93*(20), 10614-10619.

Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., et al. (2000). A gene expression database for the molecular pharmacology of cancer. *Nat Genet, 24*(3), 236-244.

Seidel, C. (2008). Introduction to DNA microarrays. In F. Emmert-Streib & M. Dehmer (Eds.), *Analysis of Microarray Data: A Network-Based Approach* (1st ed., pp. 1-26). Weinheim: Wiley-VCH.

Severgnini, M., Bicciato, S., Mangano, E., Scarlatti, F., Mezzelani, A., Mattioli, M., et al. (2006). Strategies for comparing gene expression profiles from different microarray platforms: application to a case-control experiment. *Anal Biochem, 353*(1), 43-56.

Shabalin, A. A., Tjelmeland, H., Fan, C., Perou, C. M., & Nobel, A. B. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics, 24*(9), 1154-1160.

Shadeo, A., & Lam, W. L. (2006). Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Res, 8*(1), R9.

Shen, R., Ghosh, D., & Chinnaiyan, A. M. (2004). Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics, 5*(1), 94.

Sherlock, G. (2005). Of fish and chips. *Nature Methods, 2*(5), 329-330.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., et al. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol, 24*(9), 1151-1161.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., et al. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology, 24*(9), 1151-1161.

Shtutman, M., Zhurinsky, J., Oren, M., Levina, E., & Ben-Ze'ev, A. (2002). PML is a target gene of beta-catenin and plakoglobin, and coactivates beta-catenin-mediated transcription. *Cancer Res, 62*(20), 5947-5954.

Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., & Zhao, Y. (2004). *Design and analysis of DNA microarray investigations*. New York: Springer.

Smith, V., Botstein, D., & Brown, P. O. (1995). Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proc Natl Acad Sci USA, 92*(14), 6479-6483.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol, 3*, Article3.

Smyth, G. K. (2005). Limma: linear models for microarray data. In R. Gentleman, V. Carey, W. Huber, R. Irizarry & S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (pp. 397-420). New York: Springer.

Smyth, G. K., & Speed, T. (2003). Normalization of cDNA microarray data. *Methods, 31*(4), 265-273.

Smyth, G. K., Yang, Y. H., & Speed, T. (2003). Statistical issues in cDNA microarray data analysis. *Methods Mol Biol., 224*, 111-136.

Stafford, P., & Brun, M. (2007). Three methods for optimization of cross-laboratory and cross-platform microarray expression data. *Nucleic Acids Res, 35*(10), e72.

Stears, R. L., Martinsky, T., & Schena, M. (2003). Trends in microarray analysis. *Nat Med, 9*(1), 140-145.

Stekel, D. (2006). *Microarray bioinformatics* Cambridge: Cambridge University Press. (2003)

Stoesser, G., Tuli, M. A., Lopez, R., & Sterk, P. (1999). The EMBL Nucleotide Sequence Database. *Nucleic Acids Research, 27*(1), 18-24.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics, 5*(4), 595-620.

Tan, P. K., Downey, T. J., Spitznagel, E. L., Jr., Xu, P., Fu, D., Dimitrov, D. S., et al. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res, 31*(19), 5676-5684.

Tanaka, F., Kameda, A., Yamamoto, M., & Ohuchi, A. (2005). Design of nucleic acid sequences for DNA computing based on a thermodynamic approach. *Nucleic Acids Res, 33*(3), 903-911.

Tao, H., Bausch, C., Richmond, C., Blattner, F. R., & Conway, T. (1999). Functional genomics: expression analysis of Escherichia coli growing on minimal and rich media. *J Bacteriol, 181*(20), 6425-6440.

Thellin, O., Zorzi, W., Lakaye, B., Borman, B. D., Coumans, B., Hennen, G., et al. (1999). Housekeeping genes as internal standards: use and limits. *J Biotechnol, 75*(2-3), 291-295.

Tsai, J., Sultana, R., Lee, Y., Pertea, G., Karamycheva, S., Antonescu, V., et al. (2001). RESOURCERER: a database for annotating and linking microarray resources within and across species. *Genome Biol, 2*(11), SOFTWARE0002.

Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., & Wong, W. H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res, 29*(12), 2549-2557.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Massachusetts: Addison Wesley.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA, 98*(9), 5116–5121.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351.

Vogl, T. P., Mangis, J. K., Rigler, A. K., Zink, W. T., & Alkon, D. L. (1988). Accelerating the convergence of the back-propagation method. *Biological Cybernetics, 59*(4-5), 257-263.

Wang, H. Y., Malek, R. L., Kwitek, A. E., Greene, A. S., Luu, T. V., Behbahani, B., et al. (2003). Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. *Genome Biol, 4*(1), R5.

Wang, Y., Lu, J., Lee, R., Gu, Z., & Clarke, R. (2002). Iterative normalization of cDNA microarray data. *IEEE Trans Inf Technol Biomed, 6*(1), 29-37.

Warnat, P., Eils, R., & Brors, B. (2005). Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *Bmc Bioinformatics, 6*, 265.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A, 26*(4), 359-372.

Wellmann, A., Thieblemont, C., Pittaluga, S., Sakai, A., Jaffe, E. S., Siebert, P., et al. (2000). Detection of differentially expressed genes in lymphomas using cDNA arrays: identification of clusterin as a new diagnostic marker for anaplastic large-cell lymphomas. *Blood, 96*(2), 398-404.

Wen, W.-H., Bernstein, L., Lescallett, J., Beazer-Barclay, Y., Sullivan-Halley, J., White, M., et al. (2000). Comparison of TP53 mutations identified by oligonucleotide microarray and conventional DNA sequence analysis. *Cancer Research, 60*(10), 2716-2722.

Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., et al. (2000). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res, 28*(1), 10-14.

Whitehead, A., & Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine, 10*(11), 1665-1677.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis* (2nd ed.). New York: Springer.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *IRE WESCON Convention Record, 4*, 96-104.

Widrow, B., & Stearns, S. (1985). *Adaptive signal processing*. Englewood Cliffs, NJ: Prentice Hall.

Wilkinson, L. (2005). *The grammar of graphics* (2nd ed.). New York: Springer-Verlag.

Wilson, C. L., & Miller, C. J. (2005). Simpleaffy: a BioConductor package for Affymetrix quality control and data analysis. *Bioinformatics, 21*(18), 3683-3685.

Wilson, D. L., Buckley, M. J., Helliwell, C. A., & Wilson, I. W. (2003). New normalization methods for cDNA microarray data. *Bioinformatics, 19*(11), 1325-1332.

Winzeler, E. A., Richards, D. R., Conway, A. R., Goldstein, A. L., Kalman, S., McCullough, M. J., et al. (1998). Direct allelic variation scanning of the yeast genome. *Science, 281*(5380), 1194-1197.

Wit, E., & McClure, J. (2004). Normalization. In *Statistics for Microarrays: Design, Analysis and Inference* (1st ed., pp. 57-101). Chichester, England: John Wiley & Sons, Ltd.

Woo, Y., Affourtit, J., Daigle, S., Viale, A., Johnson, K., Naggert, J., et al. (2004). A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *J Biomol Tech, 15*(4), 276-284.

Xiong, H., Zhang, Y., Chen, X. W., & Yu, J. (2010). Cross-platform microarray data integration using the normalised linear transform. *Int J Data Min Bioinform, 4*(2), 142-157.

Xu, L., Tan, A. C., Winslow, R. L., & Geman, D. (2008). Merging microarray data from separate breast cancer studies provides a robust prognostic test. *Bmc Bioinformatics, 9*, 125.

Yanagawa, R., Furukawa, Y., Tsunoda, T., Kitahara, O., Kameyama, M., Murata, K., et al. (2001). Genome-wide screening of genes showing altered expression in liver metastases of human colorectal cancers by cDNA microarray. *Neoplasia, 3*(5), 395-401.

Yang, Y. H., Buckley, M. J., Dudoit, S., & Speed, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics, 11*(1), 108–136.

Yang, Y. H., Buckley, M. J., & Speed, T. P. (2001). Analysis of cDNA microarray images. *Brief Bioinform, 2*(4), 341-349.

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

Yang, Y. H., Dudoit, S., Luu, P., & Speed, T. P. (2001). Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel & E. R. Dougherty (Eds.), *Microarrays: optical technologies and informatics (proceedings of SPIE)* (Vol. 4266, pp. 141-152). San Jose, California: SPIE-International Society for Optical Engineering.

Yang, Y. H., & Speed, T. (2002). Design issues for cDNA microarray experiments. *Nat Rev Genet., 3*(8), 579-588.

Yuen, T., Wurmbach, E., Pfeffer, R. L., Ebersole, B. J., & Sealfon, S. C. (2002). Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Research, 30*(10).

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., et al. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol, 4*(4), R28.

Zhang, A. (2006). *Advanced analysis of gene expression microarray data*. New Jersey: World Scientific.

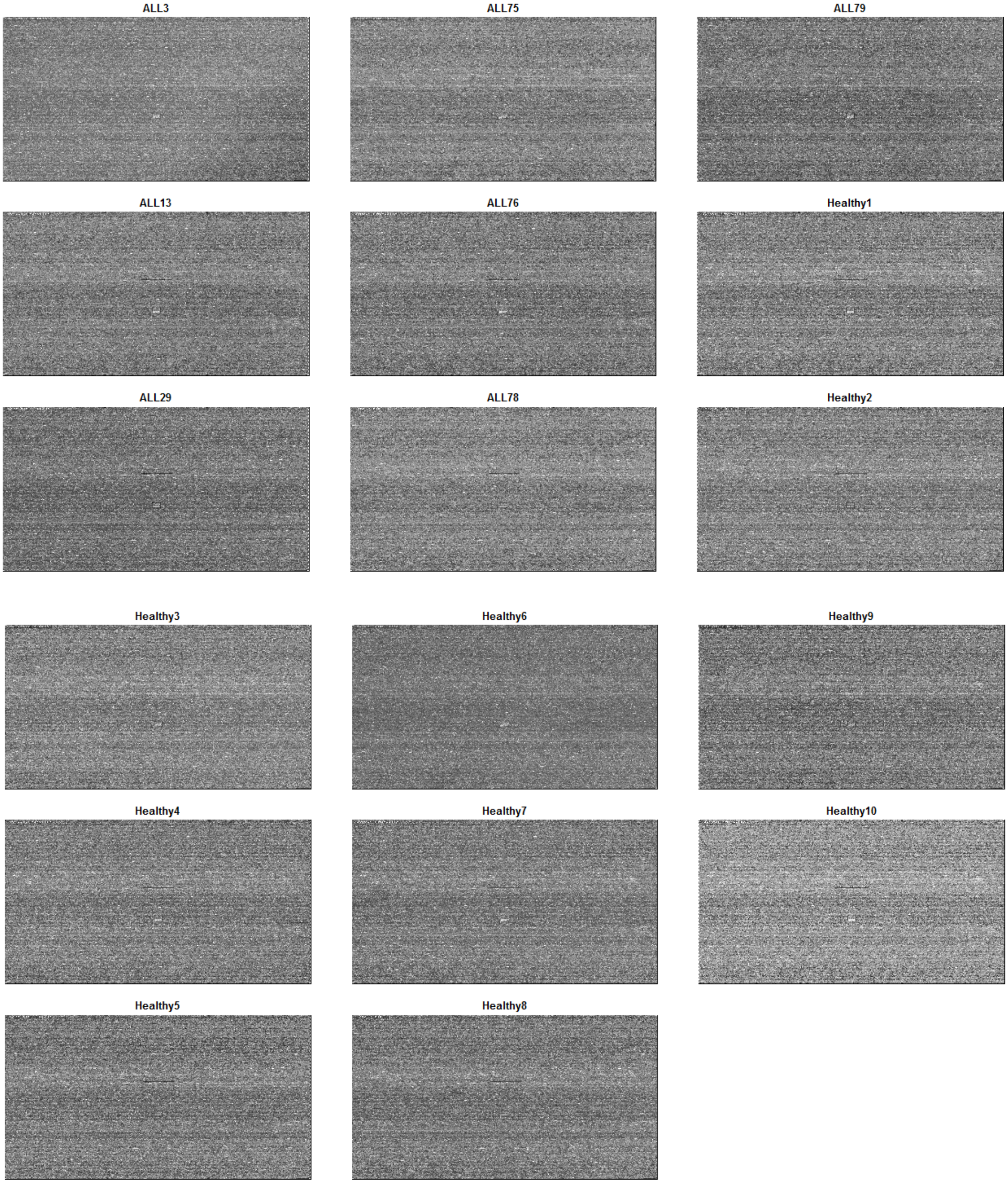
Zhao, R., Gish, K., Murphy, M., Yin, Y., Notterman, D., Hoffman, W. H., et al. (2000). Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev, 14*(8), 981-993.

Zhou, L., & Rocke, D. M. (2005). An expression index for Affymetrix GeneChips based on the generalized logarithm. *Bioinformatics, 21*(21), 3983-3989.

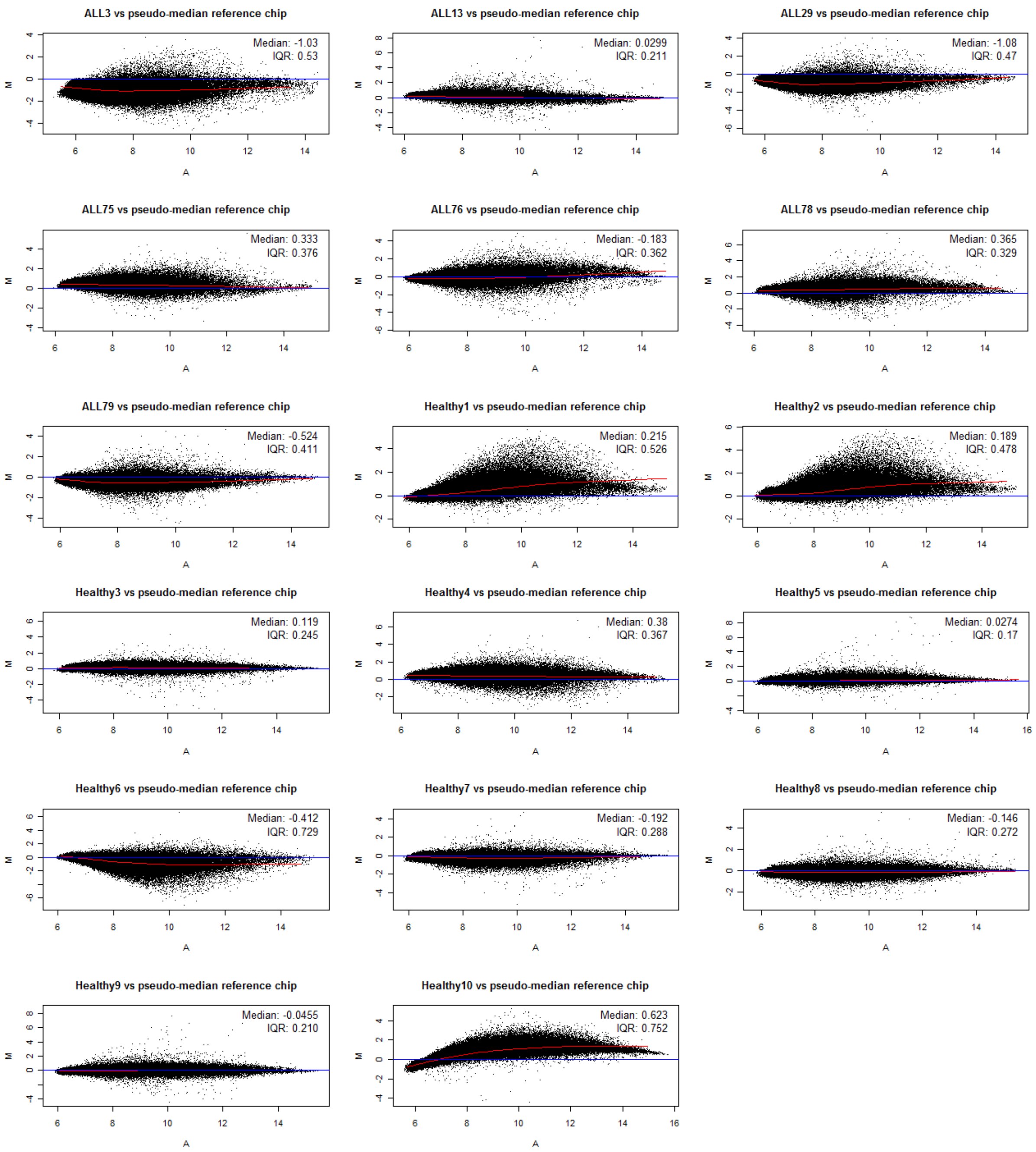
Appendix A

Assessment of Affymetrix Arrays

A.1 Reconstruction of Original Scanner Image

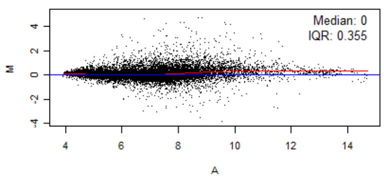


A.2 MA Plots of Raw Affymetrix Arrays

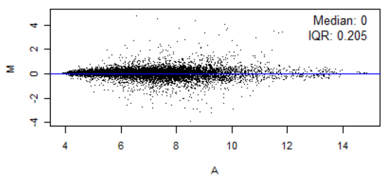


A.3 MA Plots of Normalized Affymetrix Arrays

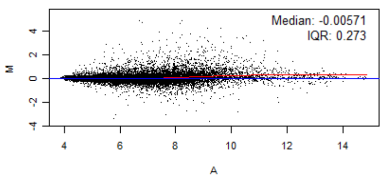
ALL3 vs pseudo-median reference chip



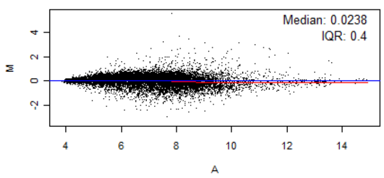
ALL13 vs pseudo-median reference chip



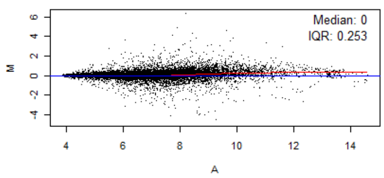
ALL29 vs pseudo-median reference chip



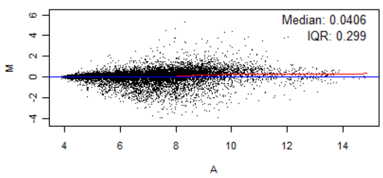
ALL75 vs pseudo-median reference chip



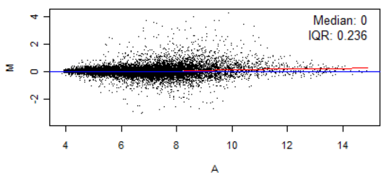
ALL76 vs pseudo-median reference chip



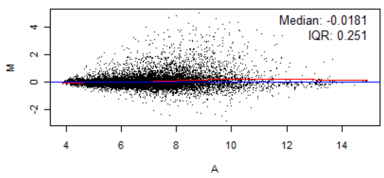
ALL78 vs pseudo-median reference chip



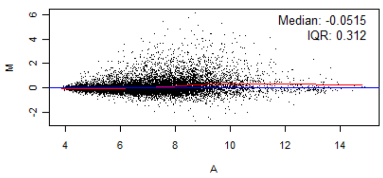
ALL79 vs pseudo-median reference chip



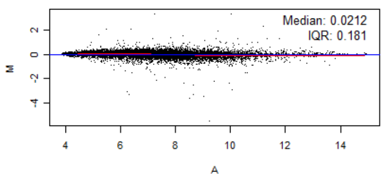
Healthy1 vs pseudo-median reference chip



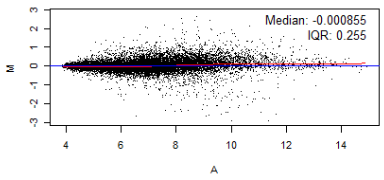
Healthy2 vs pseudo-median reference chip



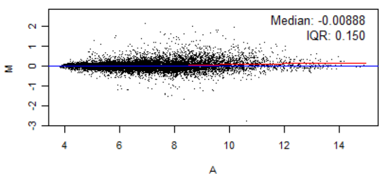
Healthy3 vs pseudo-median reference chip



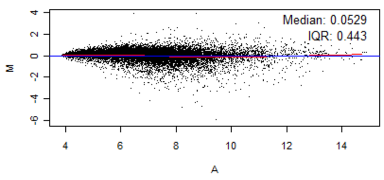
Healthy4 vs pseudo-median reference chip



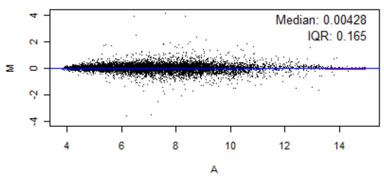
Healthy5 vs pseudo-median reference chip



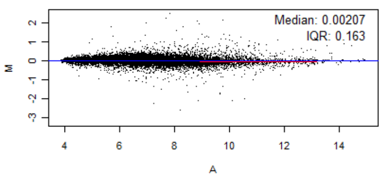
Healthy6 vs pseudo-median reference chip



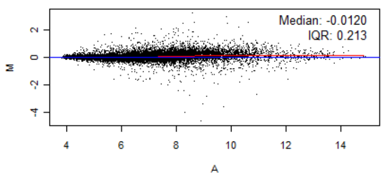
Healthy7 vs pseudo-median reference chip



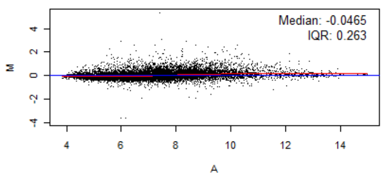
Healthy8 vs pseudo-median reference chip



Healthy9 vs pseudo-median reference chip



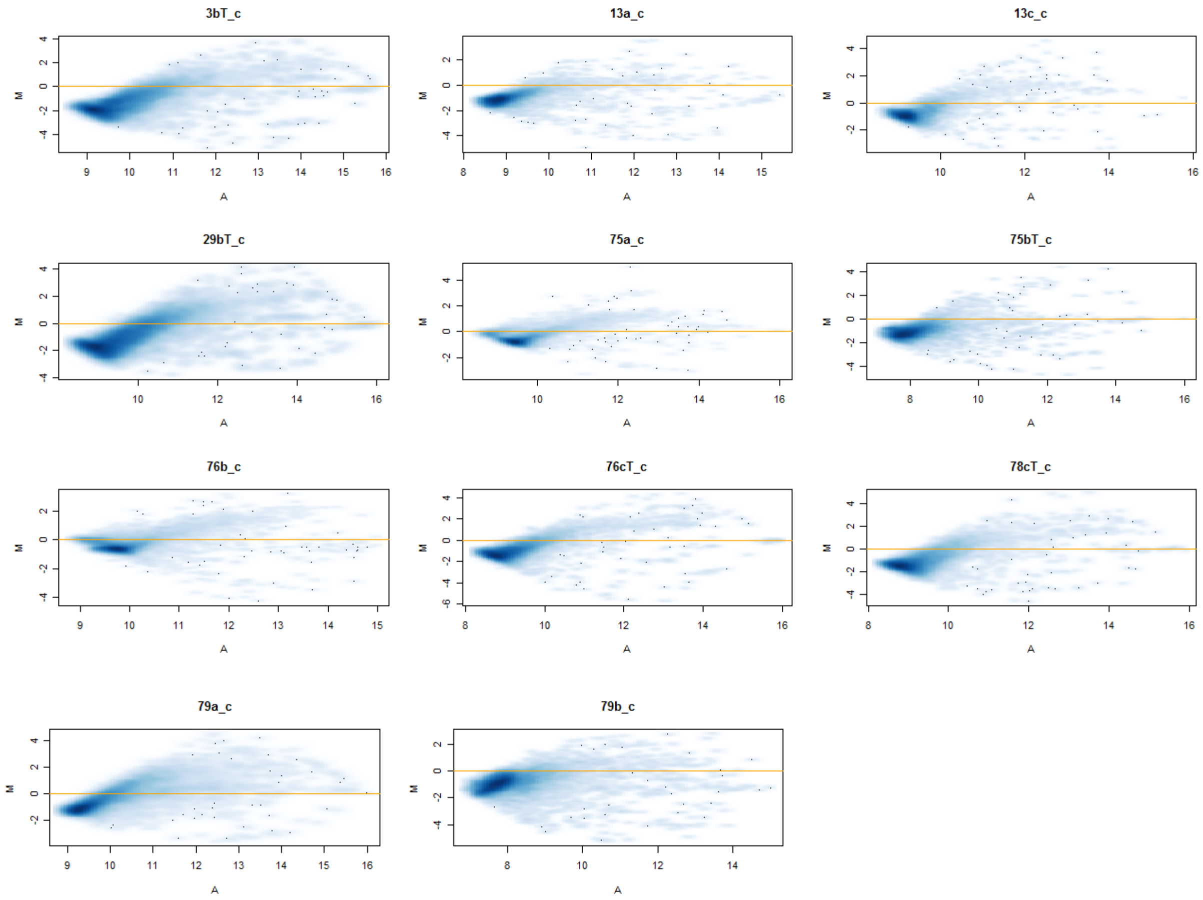
Healthy10 vs pseudo-median reference chip



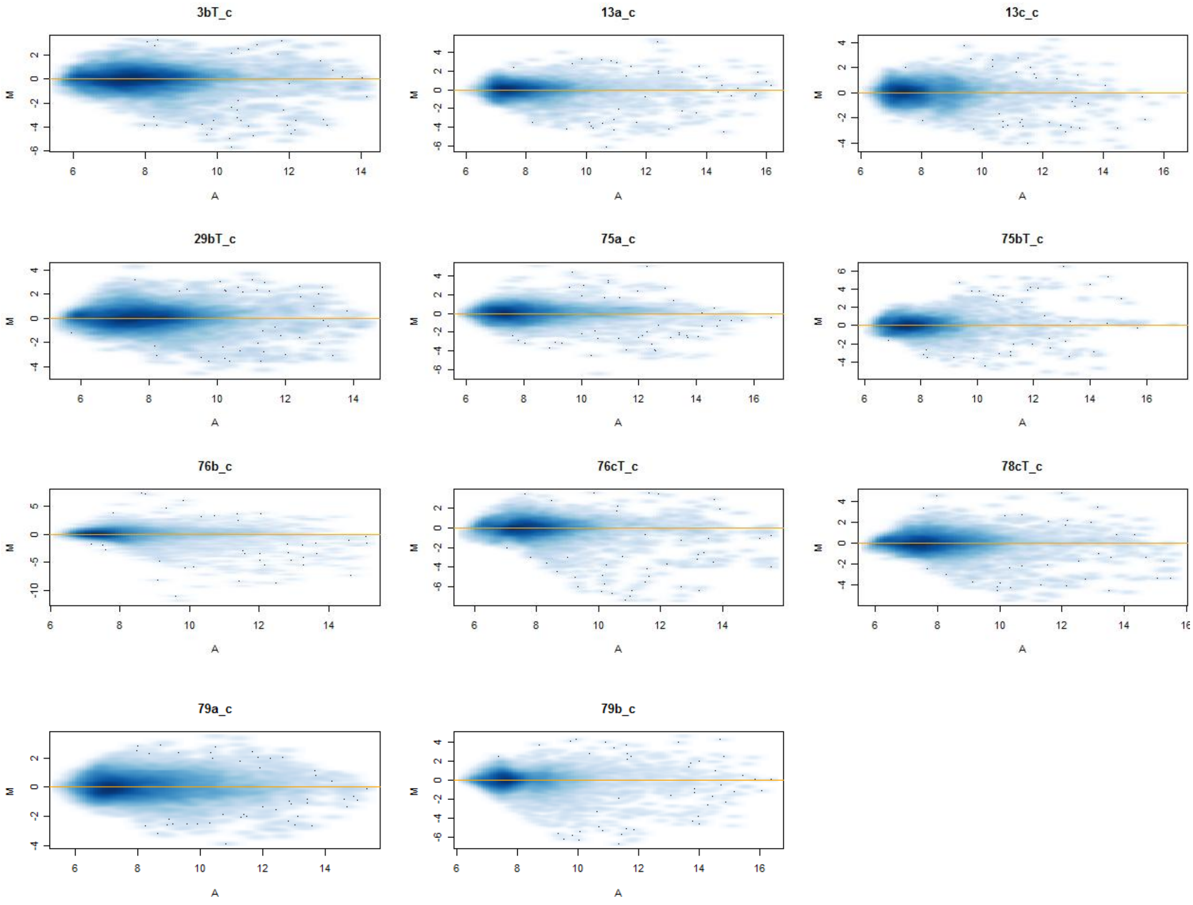
Appendix B

Assessment of cDNA Arrays

B.1 MA Plots of Untreated cDNA arrays



B.2 Post-normalization MA plots of cDNA arrays


3bT_c
M
A
13a_c
M
A
13c_c
M
A
29bT_c
M
A
75a_c
M
A
75bT_c
M
A
76b_c
M
A
76cT_c
M
A
78cT_c
M
A
79a_c
M
A
79b_c
M
A