Microarray-based expression profiling: Improving data mining and the link to biological knowledge pools

A thesis submitted in partial
satisfaction of the requirements for the degree of
Doctor of Philosophy by:

Peter James Giles

Department of Pathology
School of Medicine, Cardiff University

June 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Microarray based expression profiling provides a useful research tool to gain new insights into biological systems. Data analysis methods are in their infancy, with the answers to many basic questions shadowed by work on more complex computational and statistical methods. Understanding the fundamental concepts is essential in the application of statistical tests and to underpin work aiming to link microarray data to its biological annotation.

Centring on the detection of differential gene expression, the work presented in this thesis explores the effect of different statistical testing approaches and different expression metrics in their ability to correctly identify known changes in a dataset over a range of experimentally plausible sample sizes. Whilst certain combinations of methods are shown to have additional detection power in comparison to others, the results suggest that sample size, along with variability between samples, are probably the most important factors in analysis outcomes.

Having identified differentially regulated genes, the final and most labour intensive part of the analysis process is drawing biological conclusions and hypothesises about the data. A novel solution is presented which combines experimental data with a curated annotation sources along with analysis tools to assist the researcher in exploring the information contained within their dataset.

Acknowledgements

The End

Hopefully I've just saved you a few precious hours of your life, but if you still have an insane desire to delve into the depths of my investigations into making sense of microarray analysis then here are the important people that have made this voyage of discovery possible... for many of you this may be the only page of this thesis that will make any sense.

Above all I am indebted to Prof. David Kipling for everything he has done, taught and imparted onto me throughout the period of studentship. I think I have been very lucky in finding a supervisor with enthusiasm, inspiration, and time to fully engage with his student on a journey of joint exploration into microarray bioinformatics and feel I have learnt so much more than how to undertake good science. In return I hope that he has been enlightened to the marvels of computing beyond that provided by Macintosh!

The friendship, technical assistance and damned right stubbornness of Daniel Kirwilliam have been a fantastic asset to me over the last two years. It has been a pleasure to find somebody with similar views on the way to get the job done, how to endure basement existence, and a love of filling in *a capella* over some of music's finest works!

Thanks must go to Megan Musson for her tolerance and good humour and for enduring the undecipherable conversations in the office, along with providing valuable insight into the problems faced by average service users in analysing their data. The MADRAS project is indebted to the work from Matthew Peake and Kaye Smith in identifying our coding bugs and helping us mould a user friendly software tool.

I am thankful to Amanna, Jasper and Willow for their persistent interruptions in the pursuit of attention during my period of writing up. Many thanks you all for your love and support, and for enduring whatever negative parts of the process that I imparted in your direction, hopefully now I can now give you all the time and attention you deserve.

Finally I owe much thanks to my parents, who always believed I was capable of more than I achieved and somehow succeeded in steering me on a path through life. I find it amazing that my first work experience at the age of fourteen was in medical statistics and I'm now making a career using those skills... an inspired placement!

My studentship was generously funded by the Medical Research Council.

Contents

List of Abbreviations Used

AUC	Area Under the Curve
CEL	Affymetrix Cell Intensity File
dChip PMMM	dChip Perfect Match / Mismatch MBEI
dChip PM-Only	dChip Perfect Match Only MBEI
FDR	False Discovery Rate
MADRAS	Microarray Data Review and Annotation System
MAS	Microarray Suite
MBEI	Model Based Expression Indices
Q-Q	Quantile-Quantile
RMA	Robust Multi-Chip Average
VSN	Variance Stabilisation Normalisation

Chapter One
Introduction

In this Chapter, the concepts, technology and implementation of gene expression profiling are introduced and explored. Section 1.1 introduces functional genomics and the biological importance of gene expression. Section 1.2 introduces the techniques of microarray expression measurement and compares different technological approaches. Section 1.3 introduces the Affymetrix GeneChip system, the design and manufacture of a GeneChip, and the technological process in transferring a biological sample into experimental data. Section 1.4 discusses the analysis stages required to convert the scanned microarray image into a numerical result and the type of analysis that can subsequently be applied to the data to enable biological understanding.

1.1 Gene expression profiling

Significant advances in molecular biology have occurred as a result of the application of high-throughput techniques to experimental problems. Application of these technologies has accelerated the achievements in genome sequencing over the last few years and contributed to the completion of many sequencing projects including many microbial genomes, a few from higher organisms, as well as substantial parts of other eukaryotic genomes.

To biomedical researchers, the draft release of the human genome sequence (Lander, et al., 2001; Venter, et al., 2001) is purported to form the firm foundation for biomedical research in the decades ahead, allowing for further study into the mechanisms of human biology and the study of inherited disease.

However, the genome sequence only is only the start of the biological process and can be thought of as representing the *"parts list"* for an organism (Skolnick and Fetrow, 2000). To realize the full potential of the sequencing accomplishment, the information must be taken forward to assist in the understanding of how genomes function and to study the many interactions within an organism.

The ability to monitor the expression of many genes simultaneously at the transcript level has become possible due to the advent of DNA microarray technologies (Pease, et al., 1994; Schena, et al., 1995). Microarrays provide the possibility for examination of the expression patterns of many previously uncharacterised genes and may provide clues to their possible function by comparison analysis.

Combinations of this information allow for the formulation of metabolic schemas to understand how pathways are changed under varying conditions a cell is exposed to. Advancements in the application of these high-throughput technologies is set to allow data collection on a scale previously unparalleled and will allow for the comparative study of entire genomes and their resultant elements between species as well as characterise the mechanisms behind individual variations and the complex interplay that allows an organism to function.

1.2 Introduction to microarray technologies

Microarray technology makes use of the sequence resources created by the genome projects and other sequencing efforts to measure the cellular transcription of many genes simultaneously based on the principle of hybridisation. Hybridisation has been in use for many years in molecular biology and forms the basis of the established techniques of Northern and Southern Blotting.

1.2.1 Evolution of the microarray

In Southern blotting, a short labelled nucleic acid probe (either DNA or RNA) is used to hybridise to complementary fragments of DNA that have been separated according to size by gel electrophoresis. Radio-labelling of the oligonucleotide enables visualisation using photographic film sensitive to radiation (Southern, 1975). Northern blotting is similar but oligonucleotides bind to mRNA run through a gel and transferred to a membrane. The intensity of any resultant band on the film is a semi-quantitative measure of the amount of DNA or mRNA present, in comparison to a known standard.

Evolving from the insight that labelled nucleic acid molecules could be used to interrogate nucleic acid molecules attached to a solid support, arrays are based on the ideas of a mass parallel version of these blotting techniques (Lander, 1999). The principle difference between blotting techniques and arrays is the immobilisation of the probe to the substrate in a microarray. Evolution of the technique from blotting first lead to the development of *"macroarrays"*, which explored expression levels by hybridizing mRNA to cDNA libraries gridded on nylon filters.

Technical improvements lead to the development of the microarray, which utilises a non-porous solid support. In addition to the miniaturisation opportunities the rigid substrate offers, a technical advantage is obtained due to the fact that neither the target nucleic acids (normally cDNA) nor the post-hybridisation wash solutions are required to permeate into nitrocellulose pores, and therefore the rates of the hybridisation and washing steps are increased (Southern, et al., 1999).

1.2.2 Overview of microarray technology

Microarrays exploit the binding of complementary single-stranded nucleic acid sequences. Whereas a Southern blot utilises on only one probe, in a microarray, thousands of known probes are fabricated onto a solid substrate (e.g. a glass slide) in a specified order. There may be tens of thousands of spots on an array, each containing a huge number of identical DNA molecules or fragments of molecules, with lengths from twenty to hundreds of nucleotides.

Whilst there are differing implementations of the fundamental ideas behind microarray technology, the stages of production and experimentation are common between the different methodologies. Firstly the microarray must be designed and produced. DNA complementary to genes of interest is generated and laid out in microscopic quantities on solid surfaces at defined positions according to the array design. In the design stage, attention must be given to the sequences identifying genes of interest to overcome issues of families of similar genes sharing sequence and the effects of splice variants.

The next stage of experimentation converts the expression mRNA into a labelled complementary DNA, enabling it to be eluted over the surface and form complementary DNA binds. The presence of bound DNA is detected by fluorescence following laser excitation and then used to form an estimate of the relative level of gene expression.

Whilst the basic technology is the same, differences exist in the types of microarray available to the researcher based on the form of sequence used for detection, and the methods by which the sequence is applied to a solid substrate. The sequence can be applied to the slide using a robot to spot previously produced sequences to the array, or synthesized by photo-lithography or by ink-jet printing technologies.

The second difference is in the length of DNA sequences that are laid down on the array. Either full length complementary cDNA sequence or a particular segment in the form of a unique oligonucleotide sequence is placed on the slide to enable expression specific hybridisation.

Techniques involving oligonucleotide sequences and photo-lithographic production methods are generally the proviso of commercial microarray systems whereas the full length cDNA arrays are within the remit of the researcher to design and produce the array slides (e.g. Cheung V.G, et al. 1999) and as such are normally a cheaper solution.

1.2.2.1 cDNA microarrays

cDNA microarrays use robotic techniques to spot glass slides at precise points with complete gene/EST sequences in the form of a pre-synthesised cDNA to generate a probe. As a result of the circular probe formation left on the chip and the process of the applying liquid droplets of cDNA to the slide, these arrays are often termed spotted arrays. cDNA arrays offer a high degree of flexibility and can be designed and implemented using off the shelf hardware in a typical research environment, allowing the researcher complete control over the array design and features (Schulze and Downward, 2001). In addition the technology is sequence independent making it ideal for species with limited genome sequence availability (Chen, et al., 2004).

Experimentally cDNA arrays are normally subjected to differential expression by use of simultaneous, two-colour fluorescence hybridisation. Fluorescent probes are prepared from two mRNA sources to be compared one channel comprising of a green dye (Cy3) and the other a red dye (Cy5). The probes are mixed and then eluted over the microarray slide for an extended period to allow hybridisation to occur.

As a result of the simultaneous two-channel experimentation, complex issues of experimental design have evolved to overcome dye-biases and the general effects of simultaneous processing. It is generally accepted that cyclic design incorporating dye-swaps between channels is the optimal solution, but this has the limitation of closing the experimental loop and makes extension studies more difficult to plan (Churchill, 2002). Potential problems in the use of cDNA technologies include the maintenance of cDNA libraries needed for probe manufacture, and the occasional misidentification of probe sequences (Warner and Dieckgraefe, 2002).

1.2.2.2 High-density oligonucleotide arrays

In contrast to the high level of control the researcher has over the type, number, and identity of probes when using cDNA microarray systems, high-density oligonucleotide arrays are generally purchased as 'off the shelf' solutions with chips representing a standard range of probe sets (Lipshutz, et al., 1995). Custom designed arrays are available, and have been successfully developed by consortiums of researchers working on organisms for which catalogue arrays are not available, however for the average biological researcher the additional development and production costs for this form of array will most likely prove prohibitive to a customised design. However, the large number of probe sets (>47,000 on the HG-U133+ chip) incorporated on the chips compensates for this limitation.

Oligonucleotide arrays are manufactured using a combination of photolithographic and combinatory chemistry techniques (Lipshutz, et al., 1999), resulting in an extremely high feature density with complete control of the sequence laid down on the array. Typically, a set of unique oligonucleotide probes form a probe set providing information to represent a gene or expressed sequence tag (EST). However the array design requires sequence data to allow probe design, and there is a risk of uneven performance by individual array elements dependant on the rules used for oligo selection.

Despite potential issues of low specificity and sensitivity in short oligonucleotides and the converse issues of higher cost in purchase of large number of long oligonucleotides, high-density arrays offer a much more integrated system for experimental workflow by virtue of the commercial production of arrays and eliminate the potentially error-prone and time-consuming process of handling cDNA resources. Further details of the Affymetrix high-density oligonucleotide array system and the stages of experimental processing and analysis are described in Section 1.3

1.2.3 Application of microarray technology

Microarrays provide information on the relative expression levels of thousands of genes simultaneously. The information that such an experiment obtains can be used for a variety of purposes supporting research from the basic sciences, through pharmacocgenomic drug testing to applications in clinical diagnosis (Debouck and Goodfellow, 1999). Within the basic sciences microarrays provide the ability to explore the gene expression between differing cell types over time and in response to differing disease states. They also provide a potential mechanism for the development of interaction maps between genes and the downstream effects these produce (Clarke, et al., 2001).

Within pharmacogenomics, microarrays can assist in drug discovery by identifying the exact targets and actions of drugs and provide information of toxicological effects within the cell (Ivanov, et al., 2000; Nees and Woodworth, 2001). Clinically, microarrays have presented as a method to improve the pathological classification of disease and provide the potential for patient specific treatment derived from the results of identified bio-markers (Perez, et al., 2004). Microarrays have been shown to be effective in this regard with the identification of additional classes of breast cancer subgroups, which whilst appearing phenotypically identical, present with different transcription profiles (Hedenfalk, et al., 2003).

1.2.4 Limitations of microarray technology

Whilst microarrays are a powerful tool to enable a snapshot of the transcription events with a cell, they do present with limitations in their scope and of the information the technique can provide. The key proviso of the techniques is that the physiological state of an organism may not be reflected by gene expression or RNA levels. The amount of mRNA may not correlate with amount of translated protein and the expression of a protein may not always produce a detectable physiological activity or response.

In addition, mRNA is an unstable molecule, with the half lives of the messenger varying considerably. This makes the reproducible extraction of mRNA difficult. For mRNA that has been extracted without due care, all that may remain after extraction is the stable mRNA, which have not been subject to degradation and may not provide as much insight into the experimentally important transcription changes.

In Section 1.1.1 the concept of alternative splicing was introduced. Within the context of a microarray experiment, the researcher must consider to what extent changes in observed signal from a messenger are due to alternative splicing rather than a change in transcript abundance. Current knowledge of alternative splicing is limited, making array design to explore and account for this effect problematic. However, in theory the multiple probes present for each gene within high-density oligonucleotide systems should be able to reveal alternative splicing if probes span an alternative splice junction (Lee and Roy, 2004). However, looking for changes in the relative intensity of probes across a gene might reveal this but it doesn't exclude the possibility of changes in cross-hybridisation for some of the probes within a gene.

Technically quantification of transcription requires additional knowledge of how well each probe binds to its target before it can be used to deduce anything about the absolute mRNA concentration present in the cell. This can be achieved by using known concentrations of mRNA to calibrate each probe set. However, this is a labour intensive task if working with more than a few mRNAs. Whilst purists may argue that microarrays are therefore limited to measurements of whether a mRNA is present or not above the detection threshold, many researchers prefer to consider the microarray as a *"snapshot view"* of the current cellular transcription, one that can be used for hypothesis generation and is often supplemented with more accurate complimentary experiments (e.g. quantitative RT-PCR).

1.3 Affymetrix GeneChip technology

The Affymetrix GeneChip system is a commercial microarray system based on oligonucleotide array targets synthesised by a photolithographic process. The design and verification of the array is done commercially and is shipped as a plastic cassette containing the array, in which all hybridisation reactions occur (Lipshutz, et al., 1999). It is data from this microarray system that forms the basis for the work presented in this thesis.

1.3.1 Overview of technology features and production

Each gene is represented on the array by a series of different 25-mer oligonucleotide probes, which are directly synthesized onto the array. At the time of writing, each array contains up to 1.3 million different oligonucleotide probes (HG-U133+ array) with millions of oligonucleotide copies at a location on the array. The Affymetrix GeneChip design implements probes in pairs, consisting of a perfect match oligonucleotide and a mismatch oligonucleotide. The perfect match probe has a sequence exactly complimentary to the particular gene and thus measures the expression of the gene. The mismatch probe differs from the perfect match probe by a single base substitution at the centre base position.

The presence of the mismatch probe is argued to disturb the binding of the target gene transcript, help to determine the background signal and help control for any non-specific hybridisation that contributes to the signal measured for the perfect match oligonucleotide.

GeneChip arrays are commercially manufactured using a combination of photolithographic and combinatorial chemistry techniques building many arrays simultaneously on a 5-inch square quartz glass wafer.

In manufacture, linker molecules are applied to the glass wafer to form a covalently linked molecule-matrix which enables the synthesis of oligonucleotide strands onto the array. This process proceeds in a parallel process with the addition of A, C, T, or G nucleotides to multiple chains growing simultaneously by the programmed application of photolithographic masking techniques which expose or protect the linkers on each strand.

1.3.2 Chip Design

The design of each oligonucleotide arrays is reliant on the availability of accurate sequence information. Early chip designs (e.g. Hu6800) were based on sequence information obtained from GenBank exemplar sequences (a single sequence that represents a cluster of different sequences) from preliminary UniGene clusters. This was improved on in the design of the U95 chip series with the alignment of UniGene data using a cluster and alignment tool to form a consensus sequence where each base had agreement from 75% of the aligned sequences. The most recent design methodology for the latest U133 chips improved on previous practice and utilised sequence information from UniGene, dbEST, WUSTL, GenBank and RefSeq to form a the consensus sequence to which probes were then designed.

The probe spacing in early chip designs was approximately equal along an expressed sequence. However, the U133 design involved spacing to favour high quality independent probes in an attempt to ensure that multiple probes give measurements independent of the target (Affymetrix, 2001a). Probe uniqueness is required in order to minimise cross hybridisation to similar targets from unintended sequences. The Hu6800 and U95 chips were designed with probes that have 21 or more bases (out of 25) matching those in other probes excluded, as they were deemed too similar. In the design of the U133 chips the criteria were altered to exclude probes with two 8mer matches (including at least one with a 12mer match).

Figure 1.2

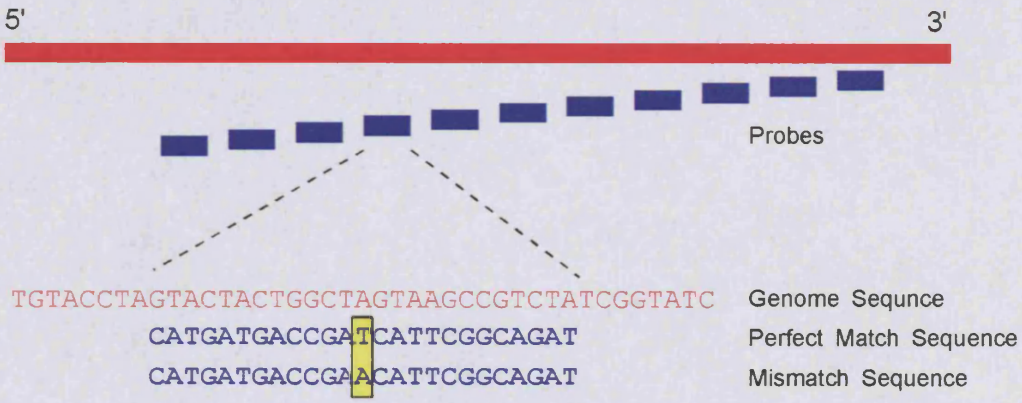


Figure 1.2 – Overview of probe location and design in relation to the genomic sequence

Mismatch probes are identical to the perfect match ones, but with an alteration of the middle (13th) base to be different to that expected from the expressed sequence, thus forming a mismatched hybridisation. A summary of the probe makeup of a probe set is shown in Figure 1.2. As part of the design process, probes for a given probe set are randomly assigned locations on the chip in an attempt to avoid location effects during the hybridisation process.

Improvements in the design algorithms and probe quality have allowed for a decrease in the number of probes in probe set from 16-20 down to 11 which has allowed for greater information density. The combination of fewer probes, along with a decrease in probe feature size on the array from 20 microns to 18 and most recently down to 11 microns, means that many more probe sets can be fitted onto a single array.

1.3.3 Sample processing

The experimental stages required to extract RNA from samples and hybridise these to GeneChips varies according to the experimental organism (Affymetrix, 2004). In eukaryotic samples, the process begins with total RNA isolated from cells (e.g. a tissue sample or cell line). The total RNA is reverse transcribed in two separate stages to produce double-stranded cDNA before a cleanup procedure is then carried out on the cDNA. The next stage involves the amplification of the cDNA into biotin labelled anti-sense cRNA which is also subjected to a cleanup procedure before fragmentation into segments typically 25-200 bases long.

At this stage, a number of controls are added which are used to locate the edges and corners of the array (oligo B2), along with control that can provide information on the hybridisation, washing and staining procedures (E-coli genes; BioB, bioC, bioD and cre). An additional five genes from B. subtilis, (dap, thr, trp, phe and lys) are also present on the chip, which are used by some researchers as additional controls or features used from normalisation between arrays (Hill, et al., 2001).

The fragmented biotin-labelled cRNA, along with the controls are mixed to form a hybridization cocktail which is inserted into the array cartridge before being placed in a hybridization oven for 16 hours, during which time the sample is eluted over the chip at optimal hybridisation temperatures. During hybridisation, the fragmented cRNA and controls bind to the oligonucleotides on the array utilising the complementary binding properties of DNA.

Following hybridisation, non-hybridised cRNA is removed from the cartridge and the array is placed in a *"fluidics station"*, where a series of washing and staining steps are applied to the array including the addition of a fluorescent staining agent streptavidin-phycoerythrin (SAPE) which binds with the biotin labelling on the cRNA.

After the washing and staining process the array is removed from the fluidics station and placed in a laser scanner. Laser light is applied to the array which excites the fluorescent staining agent. At locations where more cRNA is hybridised a brighter signal is observed. Each probe array is scanned twice, taking up to ten minutes, depending on the array format. The software calculates an average of the two images, defines the probe cells and computes an intensity for each cell. The double scan improves assay sensitivity and reduces background noise. The amount of signal emitted is recorded as a value in 16 bits, with many pixels comprising a single probe. The Affymetrix control software stores this image as a *DAT* file.

The final stage in the experimental processing is the application of a grid to the array according to the signal obtained from the corner and edge controls. These are used to superimpose and align a grid upon the image which is then used to produce a single expression value for each probe using the 75th percentile of the pixel intensities for each probe cell. The probe intensity values are then written into a *CEL* file. This *CEL* file forms the bases of the majority of data analysis options which are explored further in the course of this thesis.

1.4 Interpretation of microarray data

There have been several studies undertaken to validate the technology and explore the biological meanings of the resultant data. Chudin et al. compared signal level relative to spiked in transcript concentration as a sensitivity study and found linearity over the middle range of signal values with deviation at the extremes (Chudin, et al., 2002). Naef at al., undertook experiments to characterise the expression to noise ratio and found a relationship between the standard deviation of values and noise and an intensity dependence of the standard deviation of the resultant data (Naef, et al., 2002).

Li et al. examined the sensitivity and specificity between the Affymetrix oligonucleotide microarray systems and cDNA arrays and found large differences in the values produced for allegedly comparable data. They concluded that oligonucleotide microarrays is more reliable for interrogating changes in gene expression than data from long cDNA microarrays (Li, et al., 2002).

The central concept of any microarray data analysis is one of data reduction. If we look at a single Affymetrix U133A chip the values for millions of oligonucleotides are reduced to a single value for each of the 267,397 probes. The data obtained from each set of (on average) eleven probes is then used to calculate a score for each of 22283 probe sets representing a transcript. This is then repeated across a number of chips presenting a researcher with somewhere in the region of a quarter of a million data values for further analysis, from a 10 array experiment.

Dependent on the type of experiment being undertaken, the next stage in the analysis process may be the further reduction of this data to a list of significant changes within the dataset. Due to the volume of data being produced in the course of an experiment, the analysis stages undertaken during data reduction are an essential process to make data both comparable and usable.

1.4.1 Computing expression values from image data

The first important analysis step in any microarray experiment (both cDNA and high-density oligonucleotide) is the conversion of the image file representing the scanned array to a numerical representation of transcript expression. As part of this process various potential experimental artefacts in the data must be tackled in an attempt to overcome experimental variation and make data comparable between identical experiments. This is achieved with the application of normalisation and scaling methodologies.

Concentrating on the analysis of data from the Affymetrix GeneChip system, the first stage of analysis after the initial reduction of the full image to a representative CEL file image (see Section 1.3.4) are a number of stages resulting in the production of expression values for each probe set present on the chip. As each transcript is represented by a number of probes on the chip, these must be combined and reduced to form a single value for gene expression.

The production of an expression value requires the integration of data from numerous perfect match and mismatch probes representing both a measure of actual transcript expression and a component indicating the non-specific binding to the probes. In addition to the two methodologies for this process provided by Affymetrix within versions of their Microarray Suite software, numerous alternative methods of computing expression values have been proposed. Popular alternatives include the Model Based Expression Index (MBEI) (Li and Hung Wong, 2001a; Li and Wong, 2001b), the Robust Multi-chip Average (RMA) (Irizarry, et al., 2003) and an improved RMA version using sequence information in the analysis named gcRMA (Wu, et al., 2004).

Common to all these methods is a three stage analysis approach. Stage one applies a background correction to the array to make data comparable across the array and identification of the detection threshold. Stage two incorporates varying methodologies for the incorporation of the probe data into an expression value. Some of these methods follow a heuristic approach whilst others are strongly statistical, whilst others incorporate all the probe data and other disregard mismatch data. Stage three, overcoming data variability, is discussed in Section 1.4.2.

The varying methods for the low-level analysis of high-density oligonucleotide arrays all have a common goal; to produce biologically meaningful expression values by application of a specific blend of data manipulation and modelling of the probe intensity data. Rajagopalan compared the relative performance of different expression metrics for their accuracy in producing concentration curves and correct fold-change detection within the Affymetrix Latin square dataset. Rosetta Resolver and MAS 5.0 were found to out perform the dChip PM-Only model (Rajagopalan, 2003).

Optimally, the resultant expression values should be both precise (low variance between observations) and accurate. Chapters Three and Four concentrate on a comparison of these different expression metrics, and the ability of each to correctly identify changes within a dataset.

1.4.2 Issues of data variability and normalisation

Bakay et al. assessed the sources of variability in experimentation and experimental design using clinical biopsies and concluded that inter-patient differences were the most variable effect, with only a minor contribution from experimental variation (Bakay, et al., 2002).

However, scaling and normalisation steps are essential to overcome the effects of technical variation and systematic experimental error and aim to produce data that is comparable between arrays. Normalisation can summarised as a focused goal of getting numbers from one chip to mean the same as numbers from another chip.

Normalisation for microarrays generally makes the biological assumption that the vast majority of genes on the array are unchanged; however, some researchers have chosen to implement normalisation based on a selection of known invariant control genes or use a set of *"housekeeping genes"* whose expression are believed to remain constant and are used to scale the other expression values accordingly (Geller, et al., 2003).

Normalisation steps can be applied at many differing stages of the analysis process, and are key to the methodologies for many expression metrics; however they can also be applied post-expression analysis to overcome remaining issues of dissimilarity between biologically identical chips (Bolstad, et al., 2003; Durbin, et al., 2002). Investigations into the effect of post expression analysis normalisation are discussed and explored further in Chapter Five.

1.4.3 Experimental approaches using microarrays

Analysis of the expression values obtained from a series of experiments can be split into two main areas, pattern recognition (clustering and learning techniques) and detection of differential gene expression (Nadon and Shoemaker, 2002).

Pattern recognition involves the application of cluster analysis or class discovery, which results in grouping of chips and probe sets according to the experimental data. This can then be compared to clinical phenotype data for the purposes of identifying the genetic patterns (or fingerprint) of a given clinical condition. There are a variety of techniques available for the supervised and unsupervised clustering of data including hierarchical clustering, k-means clustering, self-organising maps and principal components analysis (Dudoit and Fridlyand, 2002; Knudsen, 2002).

For the biological researcher, however, many microarray experiments are of an exploratory nature, conducted to generate hypotheses to guide future research. Experiments of this type can include prior knowledge evaluation where a researcher is examining a wide series of effects within a known system of interest and inferring new information from the results, or more generic *"gene fishing"* experiments where a researcher is looking for new interesting results to potentially explain the phenotypic differences observed in the laboratory.

The common feature between these approaches is the desire to identify differentially regulated genes that may be of interest. The process of data analysis is typically concerned with the reduction of this data to a list of "significant" findings for follow up using complementary validation techniques (e.g. RT-PCR). This analysis stage forms yet another data reduction process, where the results for many probe sets across many chips are processed to a smaller list for further examination.

In many respects, data analysis methods are in their infancy, with the answers to many basic questions shadowed by work on more complex computational and statistical methods. Understanding the fundamental concepts is essential in the application of statistical tests and to underpin future work aiming to link microarray data to its biological annotation.

As Quackenbush comments, "Sophisticated computational tools are available but the methods that are used to analyse data can have a profound influence on the interpretation of the results. A basic understanding of these computational tools is therefore required" (Quackenbush, 2001).

1.4.4 Techniques to identify differential gene expression

The aim of all experiments and subsequent analysis is to make the strongest possible conclusion from limited amounts of data. Biological variability and differences in experimental accuracy can make it difficult to separate real differences from those occurring due to random variability.

The human brain has an exceptional ability for pattern recognition, even from random data. It is natural to conclude that these observed differences are real effects and to exclude the possibility for random variation from our judgement. Analysis techniques and statistics exists to prevent this occurring and add a degree of certainty to any conclusions reached as well as allowing for the efficient analysis of the vast volumes of data a microarray experiment produces.

1.4.4.1 Non-specific filtering techniques

After normalization and estimation of the expression levels some anomalies may exist. For example there are sometimes estimated expression levels that are negative or ones that are much too high. In addition, data obtained from the early versions of Affymetrix Microarray Suite software (prior to version five) and from certain dChip models (Li and Hung Wong, 2001a; Li and Wong, 2001b) contained some negative intensity scores in the output.

Negative numbers cause issues when working out ratios and applying transforms to data so it became customary to shift these values to positive numbers. Examples of approaches used to overcome this include the addition of a shifting factor to all signal values or the replacement of all numbers less than 20 with a signal value of 20. Whilst this overcomes the problem presenting, it also has the effect of changing the distribution of the data, and may change the outcome of analyses further downstream.

Other non-specific techniques are used to reduce the amount of data for downstream analysis, in many cases to remove potentially problematic data prior to the application of fold-change calculations (Kersten, et al., 2001; Teague, et al., 1999).

As a comprehensive example of the types of filtering that can be applied to a data set, Golub et al. (Golub, et al., 1999) pre-processed their data in three stages. First thresholds were applied to the data, removing the upper and lower most values. Secondly genes were removed from analysis where the maximum minus the minimum intensity of a gene across arrays was less than a pre-determined figure and those where the ratio of maximum intensity divided by minimum intensity was less than a defined threshold.

Finally Golub et al. transformed their data by taking base ten logarithms of the raw data. The application of a transform as part of analysis has been employed by many researchers, for a variety of heuristic reasons (Grant, et al., 2002; Katsuma, et al., 2001; Virtaneva, et al., 2001). Virtaneva comment the reasons for application of a log transform to their data was to *"reduce skew and the desired variability properties"*. An investigation of the application of these techniques and specifically transformation is addressed in Chapters Three and Four.

1.4.4.2 Identification of differentially regulated genes

Early work on microarray data involved the application of non-specific filtering techniques (the removal of probe sets that fail to adhere to a set of arbitrary rules (Golub, et al., 1999) and analysis of the ratio of gene expression between various grouped samples (Bowcock, et al., 2001; Cao, et al., 2001; Jiang, et al., 2001; Luthi-Carter, et al., 2000; Newton, et al., 2001; Porter, et al., 2001; Sandberg, et al., 2000; Unger, et al., 2001). Other researchers (Svaren, et al., 2000) simply chose the fifty highest expressed transcripts in their resultant dataset.

The limitation of this approach is a lack of a confidence measurement in any observation obtained from a microarray experiment. For example two genes which are both very low expressed in terms of their signal measurement may be identified as significant, whilst moderate changes in the measurement of a higher expressed gene may be deemed insignificant.

Scientists are often presented with data with small differences in the observations compared to experimental imprecision and biological variability and large variances in their observations. Statistical methods have been designed to deal with each of these problems and provide a robust measurement of any differences present. Application of non biased statistical tests give both access to information of which genes are likely to be differentially expressed, , but also provide a score for confidence in any result identified.

In the early days of statistics, it was assumed that most analyses would deal with large datasets and small samples were too small to be of any value. Opinion has gradually changed with the realisation that differences in large datasets are most often obvious. It is the smaller sample size that warrants the attention of statistical investigation in order to extract reliable conclusions from limited amounts of data.

Deterred by the cost of replication of comparatively expensive microarray experiments, some researchers have attempted to develop and apply techniques to minimise the requirement for experimental replication. Ron et al. reported good results when comparing the Affymetrix change p-values to the t-test and concluded this was a good method of reducing the requirement for replication (Ron, et al., 2003), whilst Kamb and Ramaswami used the technique of difference averaging to avoid multiple replicates (Kamb and Ramaswami, 2001).

Most statistical tests have been designed for situations where the number of observations exceeds the number of factors that may influence the result. Unfortunately microarray experiments consist of a large number of parallel observations. Wolkenhauer termed this issue in relation to microarrays as *"the curse of dimensionality"* (Wolkenhauer, et al., 2002).

Despite issues with the application of statistics to microarrays, they have proven a popular and ultimately useful tool for the identification of differentially expressed genes.

Many classic statistical techniques have been applied in the identification of differentially regulated genes in microarray data sets, including several variants of Student's t-test (Baldi and Long, 2001; Dow, 2003; Kooperberg, et al., 2002; Korn, et al., 2002; Lonnstedt and Speed, 2002; Tusher, et al., 2001) and ANOVA, an extension of the t-test for multiple groups of data. (Kerr, et al., 2000; Pavlidis and Noble, 2001). Non-parametric equivalents of these tests have also been utilised (McKay, et al., 2004).

Other researchers have chosen to undertaken analysis using multiple statistical techniques and then integrate the results to draw conclusions about the changes in their data. Welsh et al. employed a variety of methods including difference of mean, fold-change, t-test and clustering methods (Welsh, et al., 2001), Wang et al. combined results from the t-test and Westfall and Young permutation t-test for multiple correction (Wang, et al., 2002), whilst Tan et al. used the t-test and Mann-Whitney test to analyse their data for differential gene expression (Tan, et al., 2002).

The t-test and SAM technique (Tusher, et al., 2001) which implements a more robust version of the t-test have also proven popular analysis techniques (Mootha, et al., 2003; Tudor, et al., 2002).

In addition to the application of these more traditional testing approaches, research has developed novel techniques for the detection of differential gene expression. Jenssen et al. developed and applied a log rank test (Jenssen, et al., 2002), Jain et al., developed a local pooled error method and compared results versus the t-test and Westfall and Young permutation t-test (Jain, et al., 2003). Li et al. developed a rank method based on the sums of detection calls (absent / marginal / present) (Li, et al., 2002b). Other researchers (Hatfield, et al., 2003; Saidi, et al., 2004; Tong, et al., 2001) have implemented the Baldi and Long Bayesian t-test framework (Baldi and Long, 2001).

A variety of permutation techniques have also been developed for application to microarray data including Permax, which assigns a limited number of permutations to the dataset (Martin, et al., 2001; Mutter, et al., 2001). Pan utilised various permutations of simulated dataset utilising ROC curves to access performance of a variety of methods including non-parametric techniques, SAM and an empirical Bayes method (Pan, 2003). Xu and Li reported the application of a permutation method applied to dChip data and made comparisons of gene ranks in the output compared to non-parametric methods (Xu and Li, 2003) and Park et al. developed and applied a permutation based ANOVA test (Park, et al., 2003).

However, the use of statistical techniques to analyse microarrays have often been applied in a somewhat ad-hoc fashion, often due to a lack of options available to validate the application of a technique to a dataset, with many researchers choosing to implement multiple tests in order to obtain a consensus confidence in the results obtained. Chapters Three to Seven are concerned with the validation of statistical methodologies and their application to Affymetrix GeneChip arrays. The investigations have followed a *"back to basics"* approach in an attempt to overcome the hurdle of understanding and infer maximum significance from minimum resources.

1.4.5 Making sense of microarray experimental findings

Most important biological activities are not the result of a single molecular activity, and generally result from choreographed activities of multiple molecules often acting as functional pathways. To fully understand how organisms function we will have to understand the relevant pathways for that organism.

The parallelization of experimentation within a microarray experiment provides the information required to attempt conclusion about gene interaction by relation of the levels of many genes simultaneously.

The results of a statistical analysis may typically return a list of between 200 and 1000 genes tagged as significant. Practically a molecular biologist might consider following up around twenty genes and thus a list of 500 genes with no clue to relative priority may be of little help to guide future experiments. An important, but often overlooked final stage of a microarray experiment is the annotation of the data, allowing for functional links between findings to be made. Techniques and tools for the effective annotation and exploration of experimental findings are discussed in detail in Chapter Seven.

Chapter Two
Data distributions and their effect on analysis options

In this Chapter the issue of data distributions of microarray datasets is explored. Section 2.1 introduces the importance of data distributions and the effect these have n the choice of statistical test that can be applied to a dataset. Section 2.2 reviews the methods used to investigate the normality of data from a variety of expression metrics. Section 2.3 explores the data distributions of six expression metrics using a variety of techniques and investigates the effect of data transformations on data distribution. Section 2.4 discusses the results and observations regarding the observed data distributions and the effects these have on the use of statistical tests in an analysis.

2.1 Introduction

The importance of investigating differential gene expression is evident from the various cell types in an organism that all contain the same genetic information, but are phenotypically very different from each other in function and appearance. Many of these differences are fundamentally due to differences in gene expression between different cells.

The mass-parallelisation of techniques able to detect gene transcription provides potential insight into the full pattern of gene expression in a cell rather than the study of single genes and messengers. To the clinician the characterisation of patterns within a dataset can yield classification information for disease, such as the determination of transcriptional signatures that predict outcomes in breast cancers (Perou, et al., 1999).

2.1.1 Introduction to differential gene expression

However, to the laboratory-based researcher, the experimental goal may be centred more on determining the basic processes that control an organism's functioning or looking for malfunctions that may result in disease. At this level the focus may be identifying changes in gene expression between only two or three different states in order to yield potential useful information about groups of genes implicated in the phenotypic differences observed. In their simplest and most common application, microarrays experiments are focused on the detection of differences between two different types of sample, such as differences between diseased and normal states or between treated and untreated groups.

Analysis of the amount of data an experiment that just a simple experiment produces is not a trivial task due to both the amount of data a researcher is required to analyse in parallel, combined with the potential complexity of tools available (Quackenbush, 2001). It is the understanding of the tools available and the aim of determining best practice in analysis that forms the basis of the investigations undertaken and reported in this thesis.

2.1.2 Application of statistical techniques to microarray data

The aim of all experiments and subsequent analysis is to make the strongest possible conclusion from the available data. Biological variability and differences in experimental accuracy can sometimes make it difficult to separate real differences from those occurring due to random variability. The human brain has an exceptional ability for pattern recognition, even from random data. It is natural to conclude that these observed differences are real effects and to exclude the possibility for random variation from our judgement. Analytical methods exist to prevent this occurring and add a degree of certainty to any conclusions reached.

In the early days of the development of microarray technologies analytic methods were often based on pair-wise comparison between chips, often using the ratio of expression signal between two samples (the *"fold change"*) as the method for determining differential expression of a gene between samples. Fold change however, is limited by it failing to account for technical and biological variability and processing each data point as an exact value contributing to the overall result and a single extreme outlying value could results in a result being returned as significant. Statistical tests present as a solution to this problem and can provide results which identify which probe sets are significant along with a value indicating the confidence of a result being true.

Many different statistical approaches have been explored to identify differential gene expression (Kerr, et al., 2000; Miller, et al., 2001; Nadon and Shoemaker, 2002). However, few guidelines exist based on the comparative performance of each test in relation to each other, regarding which is best at extracting meaningful biological knowledge from a microarray experiment. An important avenue of research thus presents in undertaking a *"back to basics"* approach to the analysis of microarray data, building analysis methodologies with sound theoretical and empirical grounding in contrast to an ad hoc analysis decision making process.

The review of any basic statistics book on the subject of determining differences between groups of data yields a series of additional questions about the data in contrast to a unified solution to the most applicable statistical test. A summary of some of the questions typically encountered and the resultant suggestion for most applicable statistical test is shown as a flow chart in Figure 2.1.

Figure 2.1

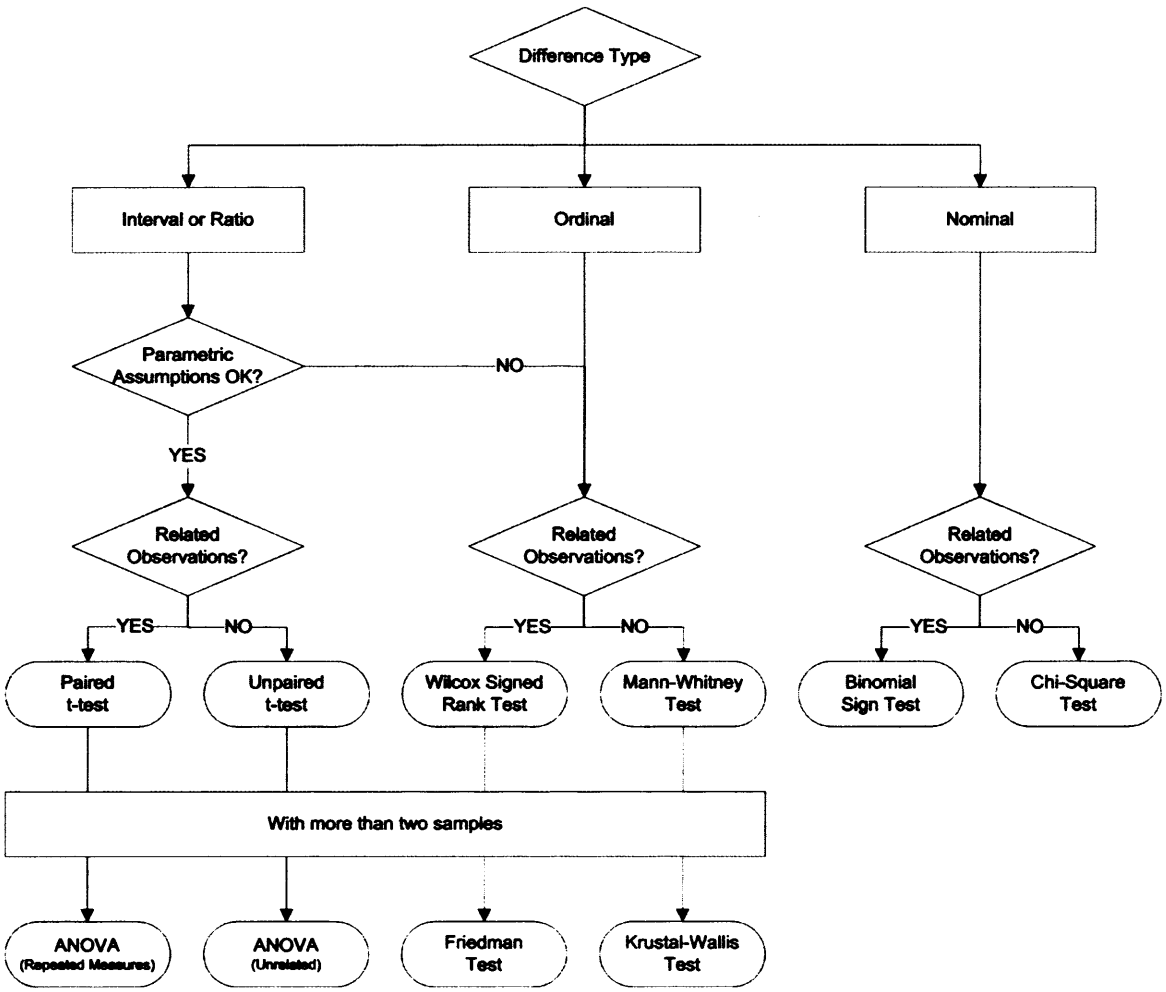


Figure 2.1 – Flow chart of typical questions encountered in the quest for a statistical test looking for differences between two groups of data. Interval and ratio data are continuous data where the differences are interpretable. Ratio data also contains a natural zero, allowing for interpretation of the ratios. Ordinal data are categorical and contain a logical ordering to the categories. Nominal data is categorical data where the order of the categories is arbitrary.

2.1.3 Reviewing key statistical assumptions and their relation to Affymetrix Gene Chip data

Each of the questions posed leads to the researcher making statements about the nature of their data; to a statistician, these answers are termed assumptions. There are very few large-scale validation datasets that have been made available by the companies that developed the technologies. As a result, researchers are often left attempting to solve issues of assumptions using real-world datasets which are limited in size and usually contain no "*known truth*" to test the outcomes of differing answers to these assumption questions.

Review of the questions posed in Figure 2.1 with the analysis of a microarray dataset in mind highlights the issue of whether the data meets parametric assumptions as a key question/assumption without an answer. This parametric assumption is concerned with the question of whether the distribution of the data is comparable to a distribution of which the equation is known. Applied to statistical techniques in which sampling is involved, this comparison is usually made to the normal (or Gaussian) distribution.

2.1.4 Application of parametric testing to microarray data

Parametric statistical tests are popular as a method of identifying differentially regulated genes in microarray datasets partially as result of small experiment size due to the associated cost considerations or the limitations of sample availability. A variety of parametric tests have been developed for the identification of differentially regulated genes in microarray datasets, including several variants of Student's t-test (Baldi and Long, 2001; Kooperberg, et al., 2002; Lonnstedt and Speed, 2002; Tusher, et al., 2001) and ANOVA (Kerr, et al., 2000).

These statistical tests have superseded earlier empirical methods that were often based on the magnitude of expression ratios, in combination with other heuristic filtering rules (Golub, et al., 1999; Knudsen, 2002). However, the majority of classical statistical tests have been designed for situations where the number of observations exceeds the number of factors that may influence the result.

Although increasingly sophisticated parametric techniques are being developed, many share the common assumption that the data (comprising the repeated measurement of the same gene across many chips) are drawn from a normal (Gaussian) distribution. As the researcher is unable to determine the answer to this question with the data from a typical microarray experiment, a logical approach is to consider the application of non-parametric testing.

2.1.5 Application of non-parametric testing to microarray data

Because of the lack of knowledge regarding data distributions resulting from the analysis of Affymetrix microarray data, statistical guidance would therefore suggest non-parametric tests as the analysis method of choice for the determination of differential gene expression. Non-parametric tests make fewer assumptions about the distribution of data (for example the requirement that the data comes from a Gaussian distribution). However, non-parametric tests are less powerful than their parametric equivalent, which assume the data distribution follows a known form.

Overall, it has been calculated that non-parametric tests require only 5% more observations in order to have the same power as parametric tests if the distribution is truly normal. If the sample is large then there is little difference between parametric and non-parametric tests, however at smaller sample sizes, to reach an identical statistical conclusion with the same confidence level, nonparametric tests require anywhere from 5% to 35% more data than parametric tests (Bethea and Rhinehart, 1991).

A typical exploratory microarray experiment may comprise of as little as three or four chips and unfortunately non-parametric testing provides insufficient power and resolution at this sample size and hence may not produce a useable result at this small sample size (Good, 2000; Grant, et al., 2002). Molutsky (Motulsky, 1999) states that with less than 7 samples it is impossible to get a p-value of less than 0.05, a commonly used confidence value.

Non-parametric tests have been used to provide such p-values (Troyanskaya, et al., 2002), but although they make no assumptions regarding the underlying data distribution, their use can be restricted by their relatively poor power when applied to small-scale, exploratory microarray datasets.

The major problem for a researcher designing a microarray experiment is one of cost. Current Affymetrix GeneChip experiments cost around £500 per sample. The researcher must be guided on dividing a limited resource between replicating experiments in order to calculate a statistical significance of the results, and undertaking another experiment. The view of an average researcher can be summarised in a quote by David Botstein *"If I had to replicate my experiments, I could only do half as much"* (Churchill, 2001).

2.1.6 Data distributions and Affymetrix microarray data

For researchers using Affymetrix GeneChip technology, the assumption of normality has been difficult to address, with the majority of extant datasets that are sufficiently large to critically investigate the issue of normality being confounded by underlying biological variability.

Inspection of existing Affymetrix datasets suggests that significant deviations from normality can occur. For example, when analysing the well-studied AML/ALL dataset (Golub, et al., 1999) one study found both normal and non-normal data, including bi-modal and tri-modal distributions (Grant, et al., 2002). A separate study profiling prostate cancers reported highly skewed distributions, with both positive and negative skew, leading the researchers to take the square root of each expression value in an attempt to control for this (Stamey, et al., 2001).

The limitation of these studies is the difficulty of separating the data distribution resulting from the analytical technology from that associated with biological variability. Indeed, for many of the large-scale cancer studies (Alon, et al., 1999; Golub, et al., 1999), the nature of the underlying biology would make it surprising if significant deviations from normality were not found. Fortunately, such datasets are large in scale (often in excess of fifty chips), and in these cases non-parametric testing would have the power to circumvent the problems raised by data that do not follow a normal distribution.

In contrast, well-designed small-scale exploratory experiments, involving as few as 3 or 4 chips in each sample group, usually do not suffer the same problems of genetic, patient and disease heterogeneity. Due to this small sample size and the limitations of non-parametric testing, we must look towards the application of parametric techniques, which offer increased power with smaller datasets. However, to achieve confidence in any result obtained, there is a need to review the nature of the data distributions obtained from microarray technology, and any address necessity to transform data prior to the application of parametric tests.

2.2 Technical Methodology

To address issues of data distributions and the assessment as to whether Affymetrix microarray data is suitable for parametric testing requires investigations applied to a dataset with sufficient replicates to explore these distribution issues in the absence of biological variance.

2.2.1 Experimental approach

As part of their software development cycle, Affymetrix made publicly available, a 59-chip "spike-in" dataset based on hybridisation of the same human pancreatic cRNA, together with various transcripts spiked-in at known concentrations, onto Affymetrix U95A GeneChips (Affymetrix, 2001b).

The original motivation for the production of this dataset was to address the experimental response to known transcript concentrations. However, removal of the spiked genes produces a dataset comprising 59 replicates from the same sample, large enough to enable exploration of the sample population, and to address whether the data follow a normal distribution and thus allows the valid application of a parametric test. Further details on the dataset can be found in Section 9.2.1

The dataset comprised the results from an experiment with 14 spiked-in probe-sets set against a complex background of human pancreatic mRNA hybridised onto U95A GeneChips containing 12559 probe-sets. Experimental details from Affymetrix are limited but it is believed that each of the 59 chips was hybridised with a technical replicate of the same mRNA. Removal of these 14 spikes thus leaves a dataset of 59 replicates which contain only experimental and technical variation.

2.2.2 Analysis with a series of expression metrics

Six different algorithms were applied to extract expression values for each probe set (generally comprising a series of sixteen probe pair intensities) from the image data. The data was analysed using Affymetrix Microarray Suite versions 4.0 (MAS 4.0) and 5.0 (MAS 5.0), the model based expression indices (MBEI) of Li and Wong (Li and Wong, 2001) calculated using two different models (PMMM and PM-Only) within the dChip software package, and the RMA (Irizarry, et al., 2003) and gcRMA (Wu, et al., 2004) algorithms released as part of the Bioconductor project (Gentleman, et al., 2004). Each of these analysis methods produced a single value for each probe set for each chip within the dataset. Further details of these expression metrics are contained in Section 9.2.2.

In addition to a quantitative expression value, four of the analysis methods (all MAS and dChip models) provide a qualitative measurement indicating if the transcript is detected (Present), not detected (Absent), or marginally detected (Marginal). These data were also extracted where present.

2.2.3 Post-analysis filtering

The expression values and calls for relative gene expression (Absent, Present or Marginal) were exported from each package into delimited text files. Data from the fourteen spiked-in genes (37777_at, 684_at, 1597_at, 38734_at, 39058_at, 36311_at, 36889_at, 1024_at, 36202_at, 36085_at, 40322_at, 407_at, 1091_at, 1708_at) and 67 control probe sets (with AFFX prefix) were removed, leaving 12545 probe sets for further analysis.

2.2.4 Technical Methodology

Normality testing using the Shapiro-Wilks (Shapiro and Wilks, 1965) test was undertaken using the R statistical language (Dudoit, et al., 2002; Ihaka and Gentleman, 1996). Futher insight into data normality was obtained by calculating the Pearson's correlation coefficient from a normal Quantile-Quantile (Q-Q) plot (Filliben, 1975) implemented using R. Measurements of the skew coefficient (Press, et al., 1993) were implemented within R using the e1071 library (Dimitriadou, et al., 2004). Box-Cox normality plots were implemented using R in combination with Microsoft Excel.

Datasets to be used for the investigation of data distributions were produced by extracting the expression values from the 59 CEL comprising the Affymetrix U95A Latin Square experiment. The CEL files were analysed using six different expression metrics, Affymetrix Microarray Suite versions 4.0 and 5.0, two dChip model based expression algorithms, and two variants of the RMA methodology. The resultant data was then processed to remove the 14 spiked-in genes and control probe sets (see Materials and Methods).

The expression matrix outputted from each analysis method comprised columns, relating to each GeneChip, with a row for each probe set. All the subsequent tests for normality refer to the data distribution for each row of data, comprising the 59 expression values for a single gene. This is distinct from analysing intra-chip distributions (Hoyle, et al., 2002), where a dataset would be a single column comprising the expression of many different genes

2.3 Results and Explorations

2.3.1 Statistical Normality Testing

The statistical assumption key to the application of parametric testing is the requirement for the data to follow a normal distribution. To test this assumption applied to Affymetrix microarray data, the normality of the data for each probe set was assed within the 59 chip Latin Square dataset (9.2.2)

Many statistical tests exist for the analysis of the distribution of a dataset, specifically the conformation to a normal distribution. The default test offered by many statistical packages (e.g. SPSS, SAS) is the Kolmogorov-Smirnov test. However, D'Agostino and Stephens have argued that "*the Kolmogorov-Smirnov test is only a historical curiosity - it should never be used*". In choosing a test to assess the normality of the exemplar dataset from Affymetrix, the issue of limited sample size restricted the choices of suitable tests; for example, the Kolmogorov-Smirnov test was designed for use against much larger sample sizes (D'Agostino and Stephens, 1986).

Formal testing of normality was undertaken by application of the Shapiro-Wilks test for normality to each probe set within the dataset from each of the six expression metrics under review. The Shapriro-Wilks test was chosen for its suitability for small and medium samples and ability to show good power across a range of non-normal distributions (Shapiro and Wilks, 1965).

The number of non-normal genes (scoring a p-value less than 0.05) for each of the four datasets is shown in Table 2.1. A large number of normal genes are observed within each of the expression metrics; however MAS 5.0 and gcRMA show a significantly increased level of non-normality when compared to the other four metrics.

Whilst this observation does not provide confidence in the application of parametric tests to these datasets, with 59 samples the Shapiro-Wilks test has the power to detect even small deviations from normality and score them as significant. In addition the nature of the analysis of the answers with a probe set either passing or failing the test for normality does not provide information regarding the magnitude of deviation from normality. It could be that many probe sets have only a modest deviation from normality, one that would not be a barrier to the application of a t-test (or similar).

Table 2.1

Analysis Method	Number of probe sets deviating from normality ($p<0.05$ from SW test)
MAS 4.0	3521 (28%)
dChip PM-MM	2213 (18%)
dChip PM-Only	2544 (20%)
MAS 5.0	5799 (46%)
RMA	3075 (25%)
gcRMA	6371 (51%)

Table 2.1 – Results from Shapiro-Wilks tests for normality.

Guidance about how much deviation can tolerated is hard to come by. However, Motulsky commented that "many parametric tests, such as the t-test, are resistant to moderate deviations from normality, although the degree to which such deviations are tolerated is dependent upon the shape of the distribution and the degree of deviation" (Motulsky, 1999). It is thus desirable to assess the degree of deviation from normality and investigate the nature of deviation observed.

2.3.2 Assessing the degree of deviation from normality

Investigations into the degree that each probe-set deviated from normality was addressed using normal Quantile-Quantile (Q-Q) plots. Q-Q plots are a graphical method for examining data distributions, showing the ordered data for each probe set plotted against the standard quantiles of the normal distribution. They provide a simple and effective visualisation of the data distribution, and any deviation from normality. For a perfect normal distribution a Q-Q plot would show data points in a straight line with a positive gradient (Hyndman and Fan, 1996). As an illustration, normal Q-Q plots for a typical reasonably-expressed probe set (32208_at) called Present in all six datasets is shown in Figure 2.1.

Since the visualisation of plots for over twelve thousand genes is impractical, the Pearson's correlation coefficient (R^2) for each Q-Q plot was calculated, to assess how close the data fit to a straight line, and thus normality. Any deviation from normality results in a correlation coefficient value significantly less than unity (Filliben, 1975). Frequency histograms of the correlation coefficient when compared to normality (R^2) for each of the datasets are shown in Figure 2.2.

Figure 2.1

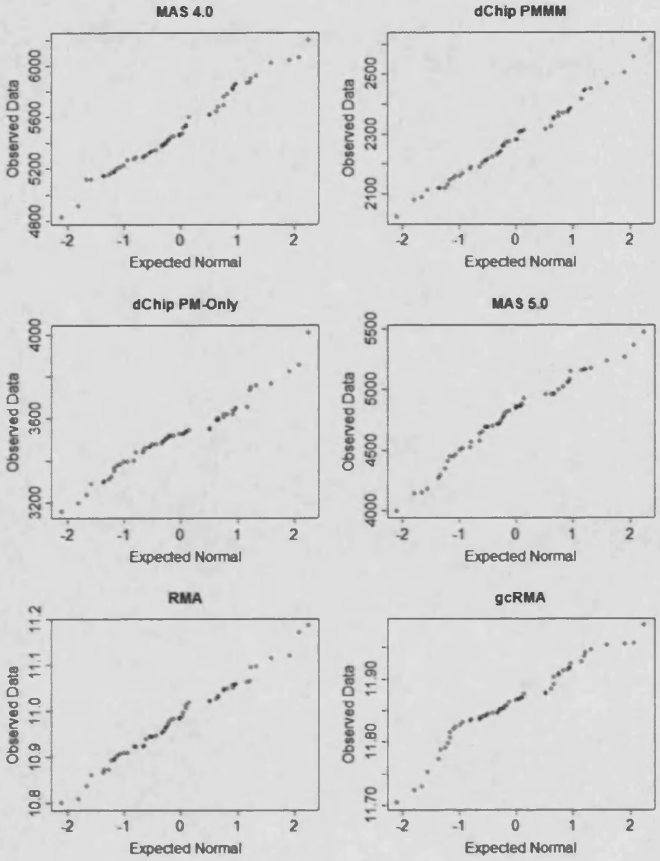


Figure 2.1 - Normal Q-Q plots for probe set 32208_at, which shows a good correlation to normality in all six datasets.

Figure 2.2

a) *b)*

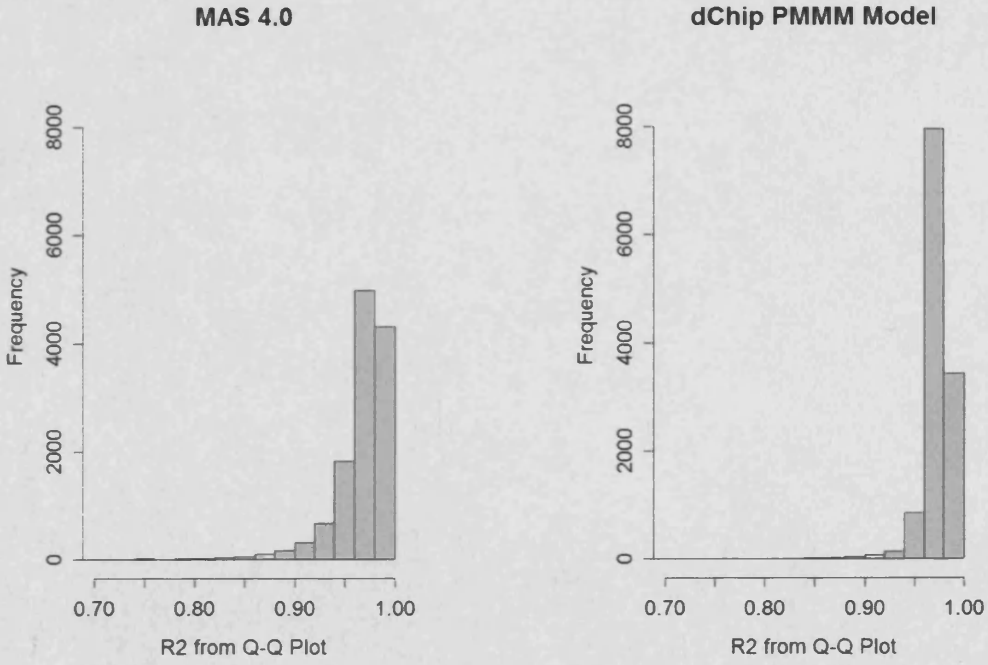


Figure 2.2

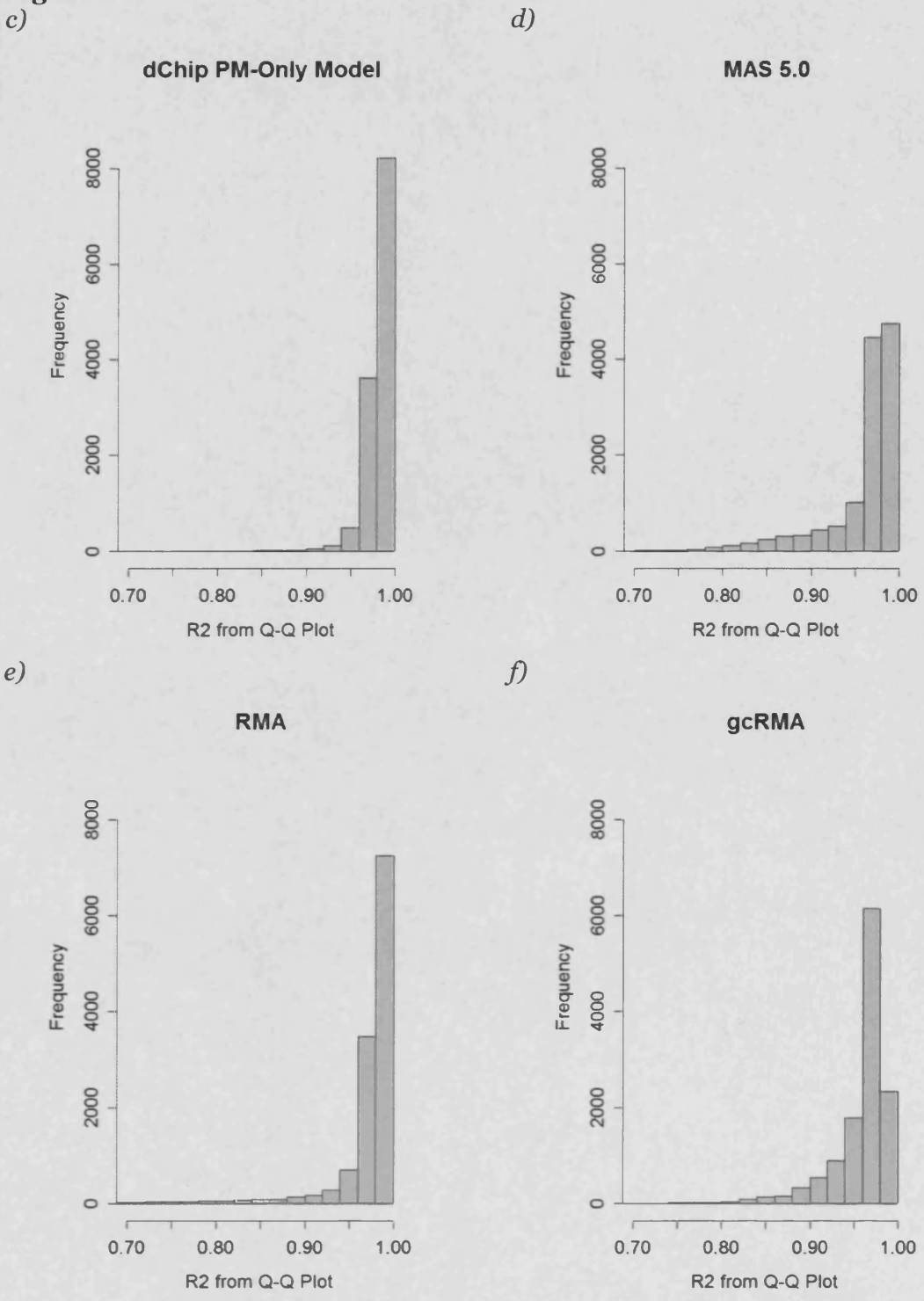


Figure 2.2 – Histograms showing the correlation to normality (Rp) extracted from normal Quantile-Quantile plots for individual probe sets.

The majority of probe sets from both dChip models, MAS 4.0 and RMA show a strong correlation with normality, with most probe sets scoring over 0.9. Interestingly, the MAS 5.0 dataset shows a pronounced tail to the histogram, with a significant number of probe sets with a low correlation value (3165 probe sets where R^2 <0.95). A similar effect is seen with data from gcRMA which displays an increase in tail density coupled with fewer probe sets scoring with a correlation coefficient value close to unity (R^2 >0.975).

Whilst the histogram plots are informative in producing information about the number of probe sets which present with differing degrees of correlation to normality, they do not provide information regarding the data distributions for the probe sets in each bin that result in non-normality. To investigate this issue further the correlation coefficient values from the Q-Q plots for each probe set were plotted against the mean expression signal for each probe set (Figure 2.3).

The majority of data from both dChip models and MAS 4.0 (Figures 2.3.a, 2.3.b and 2.3.c) group close to unity, with no apparent pattern to the expression levels of probe sets scoring as non-normal. The data distributions for a large proportion of RMA data correlates well to normality (Figure 2.3.e). However, a small group of low expressed probes show moderate deviation from normality and a few very low expressed probes show marked deviation from normality.

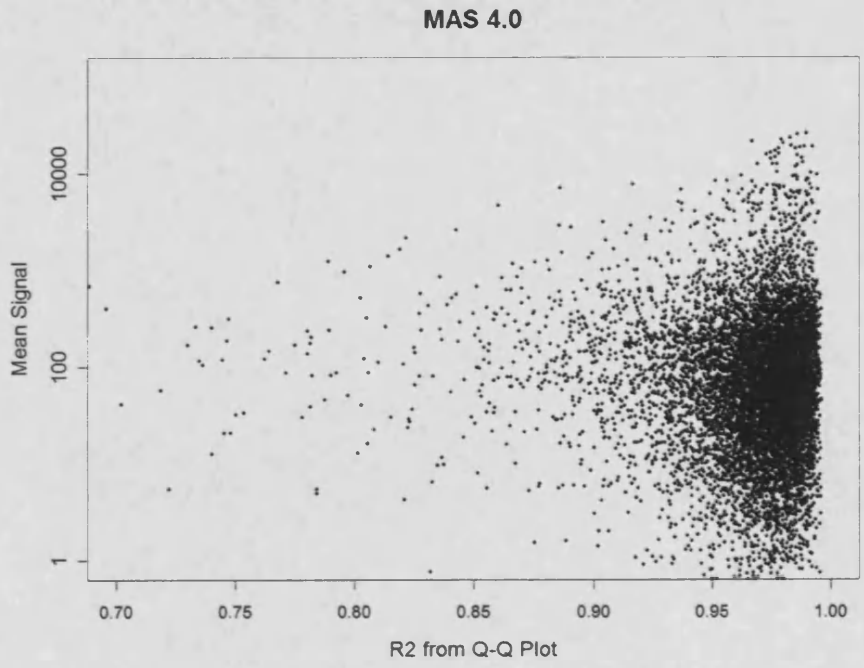
In contrast, MAS 5.0 and gcRMA (Figures 2.3.d and 2.3.f) show a large number of low expressed probe sets showing deviation from normality. Analysing the MAS 5.0 plot shows the majority of the poorly scoring probe sets have a relative expression level of less than 100 forming a pronounced tail to the plot. gcRMA shows a similar tail to the plot until reaching the very low expressed values, where their correlation to normality approaches unity again.

2.3.2.1 Comparison of correlation to normality and Shapiro-Wilks p-values

If the results of the Shapiro-Wilks test for normality are plotted against mean expression level a similar structure is seen to that visible in R^2 from Q-Q plots, versus expression plots (Figure 2.4). As both metrics address the degree of normality of the data, such a similarity could be expected, with any difference being attributed to differences in linearity and scaling of the normality data between methods. The use of the p-values to stratify data and look for more interpretation, whilst commonly done, is technically an incorrect use for the results of the test. In the strict statistical use of the Shapiro-Wilks test, an alpha level should be set prior to testing and the results interpreted as passing or failing at this level.

Figure 2.3

a)


MAS 4.0
Mean Signal
10000
100
1
0.70
0.75
0.80
0.85
0.90
0.95
1.00
R2 from Q-Q Plot

b)

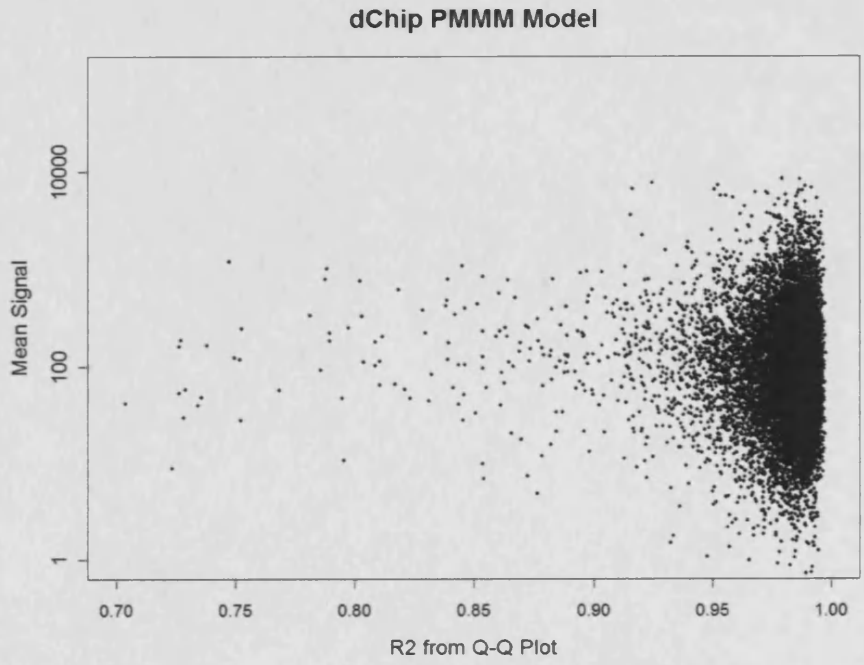
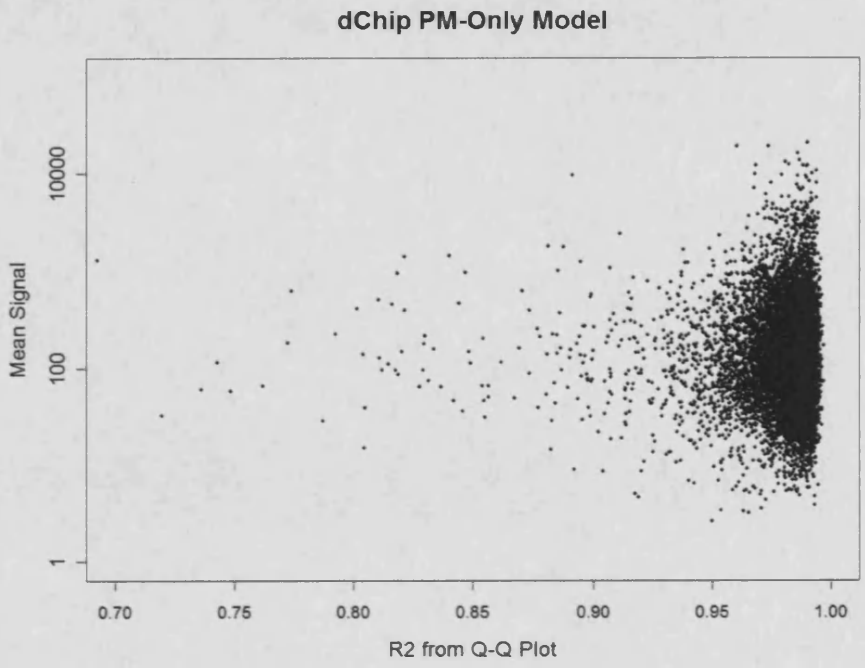

dChip PMMM Model
Mean Signal
10000
100
1
0.70
0.75
0.80
0.85
0.90
0.95
1.00
R2 from Q-Q Plot

Figure 2.3 - continued

c)



d)

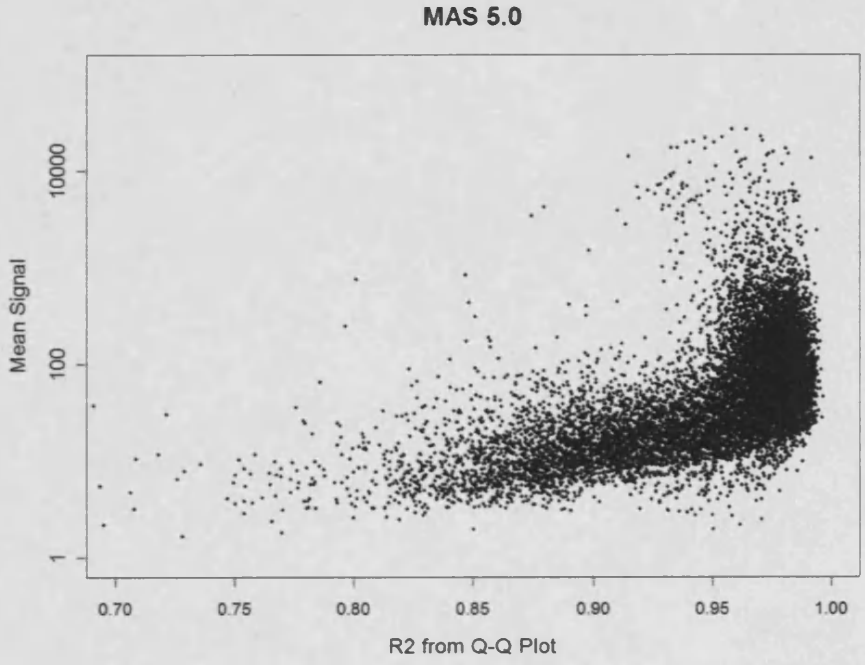
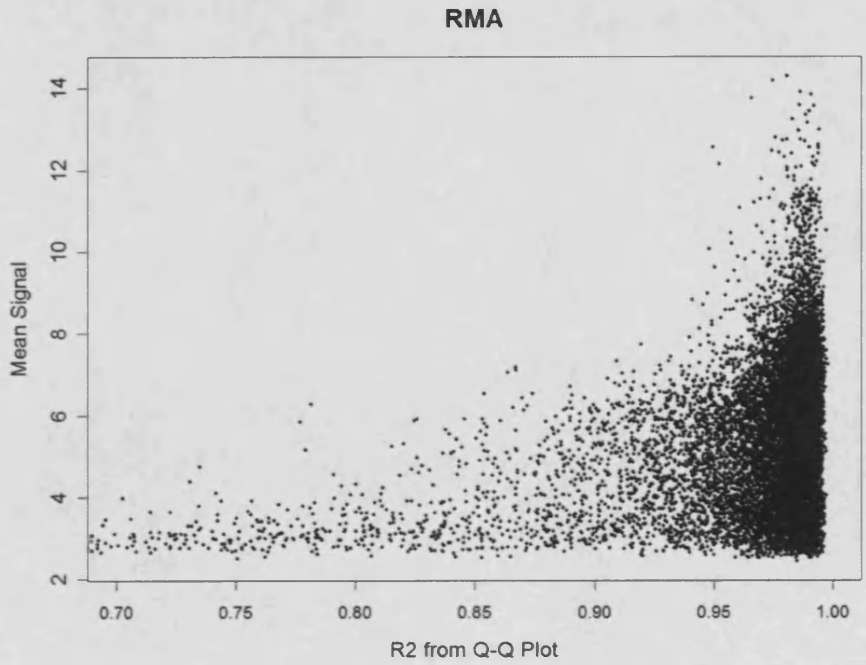


Figure 2.3 - continued

e)



f)

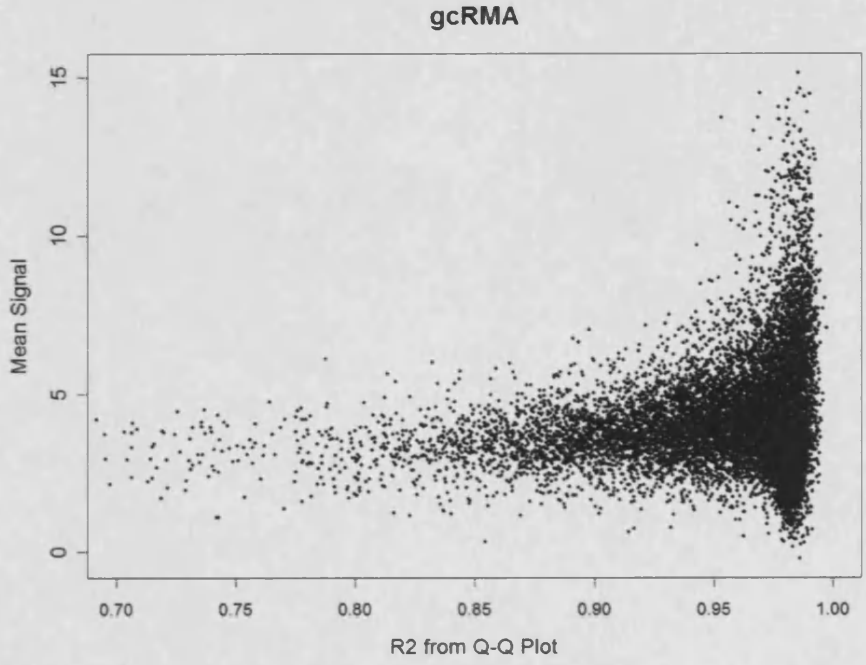


Figure 2.3 - Plots showing correlation to normality (R^2) versus mean expression signal for each probe set.

Figure 2.4

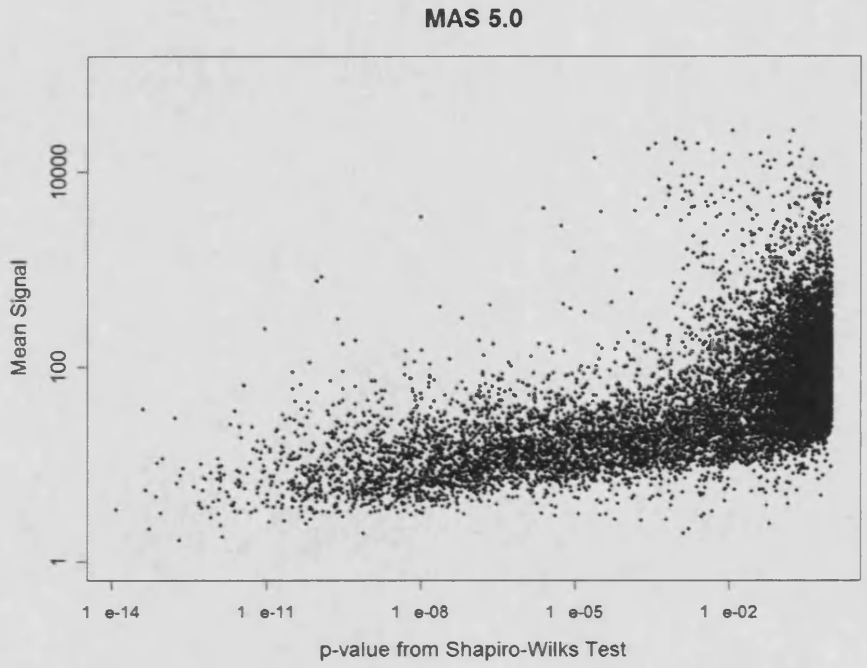


Figure 2.4 – Plot of mean expression signal from the MAS 5.0 dataset plotted against the p-value from Shapiro-Wilks test for normality for each probe set

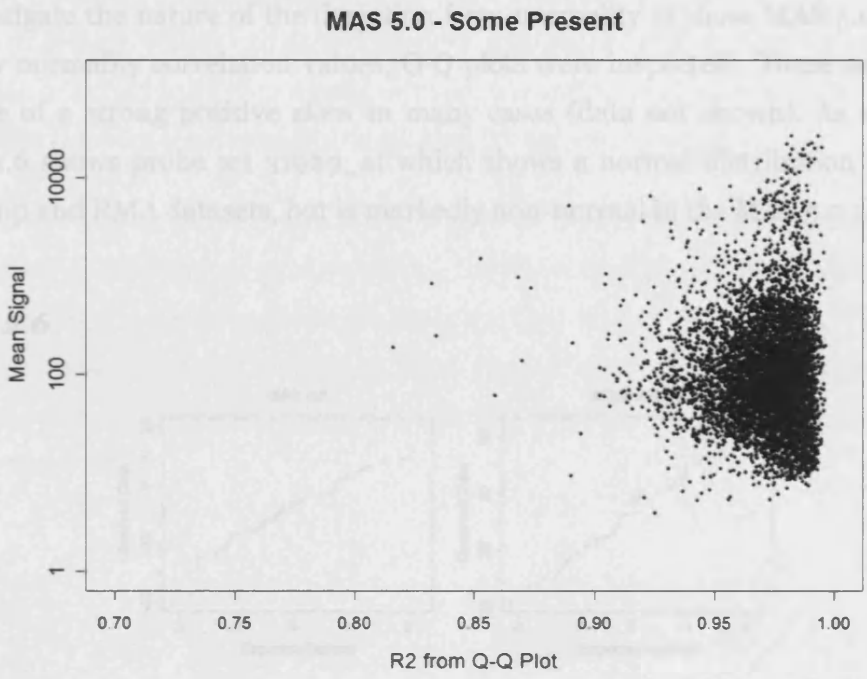
2.3.3 Investigating normality on a stratified MAS 5.0 dataset

Analysis using Affymetrix Microarray Suite provides an additional piece of information regarding each probe set, a detection call regarding the likely expression of a probe set. Each probe set is given a call of Absent, Marginal or Present derived from the resultant expression signal when compared to the background signal. Using this information, the R^2 data from the Q-Q analysis of MAS 5.0 data was stratified for re-analysis on the basis of the detection call.

Two groups were defined, one containing the data from probe sets called Absent in all samples, and the other where at least one sample was given a detection call of Present or Marginal. As shown in Figure 2.5, this somewhat arbitrary subdivision is nevertheless effective at separating the data into one group that is largely non-normal (the Absents) and another that is largely normal (the Present/Marginal probe sets). The difference in normality between the groups was scored highly significant using the Mann-Whitney test ($p < 2.2 \times 10^{-16}$).

Figure 2.5

a)



b)

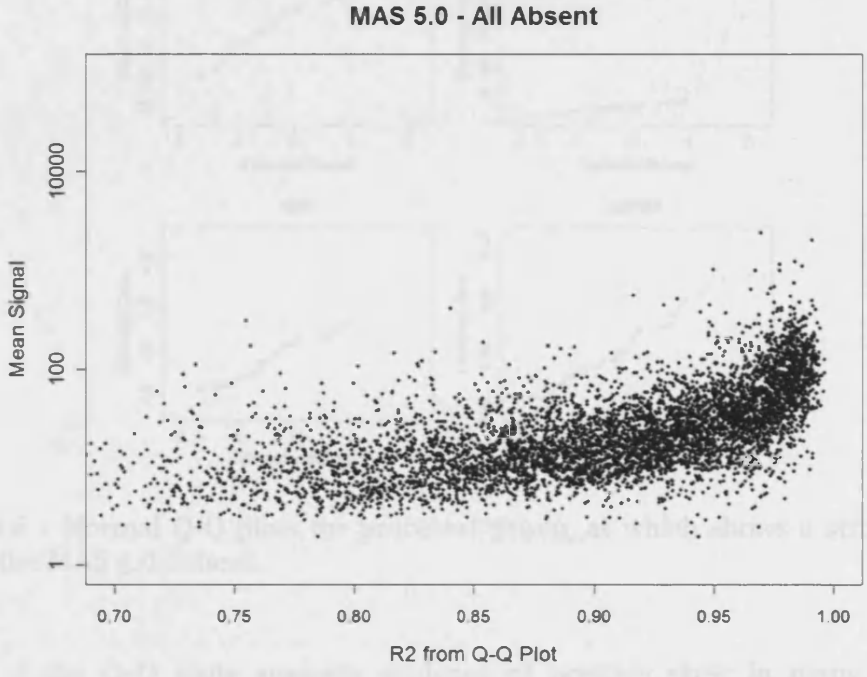


Figure 2.5 – Plots showing correlation to normality (R^2) versus mean expression signal for each probe set split according to the detection call from MAS 5.0 analyses. 2.5.a shows the data subset where at least one sample was given a detection call of Marginal or Present and 2.5.b shows the subset of data where all samples were called as Absent.

2.3.4 Characterising the nature of the data distributions leading to deviations from normality in the MAS 5.0 dataset

To investigate the nature of the deviation from normality in those MAS 5.0 probe sets with low normality correlation values, Q-Q plots were inspected. These suggested the presence of a strong positive skew in many cases (data not shown). As an example, Figure 2.6 shows probe set 31929_at which shows a normal distribution in the MAS 4.0, dChip and RMA datasets, but is markedly non-normal in the MAS 5.0 dataset

Figure 2.6

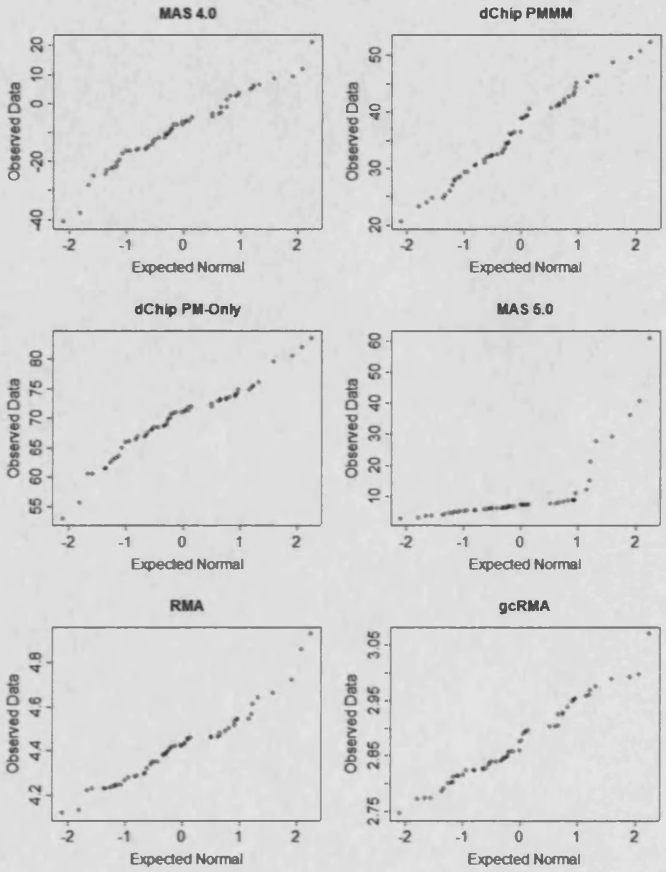


Figure 2.6 - Normal Q-Q plots for probe set 31929_at which shows a strong positive skew in the MAS 5.0 dataset.

Review of the Q-Q plots suggests evidence of positive skew in many probe sets presenting with a non-normal distribution. Skewness characterizes the degree of asymmetry of a distribution around its mean. Graphically a distribution is described as having either positive or negative skew describing the direction that the data is skewed in.

Due to the number of probe-sets under consideration a more numerical measure of skew is required in order to characterise different degrees of skewness. Numerically, skewness is defined in such a way as to make it non-dimensional. As such, skewness is defined in a way which characterises only the shape of the distribution, with the magnitude of a normally distributed dataset presenting with skew within the +2 to -2 range when the data are normally distributed (Press, et al., 1993).

Skew was calculated for each probe set in each of the six expression metrics using the e1071 library (Dimitriadou, et al., 2004). The skew coefficient for each probe set was plotted against the R^2 value obtained from the Q-Q plot of the data (Figure 2.8). For comparison the data was compared to a similar plot obtained by calculating the correlation to normality and skewness for 12000 normally distribution datasets with 59 values (Figure 2.7).

Figure 2.7

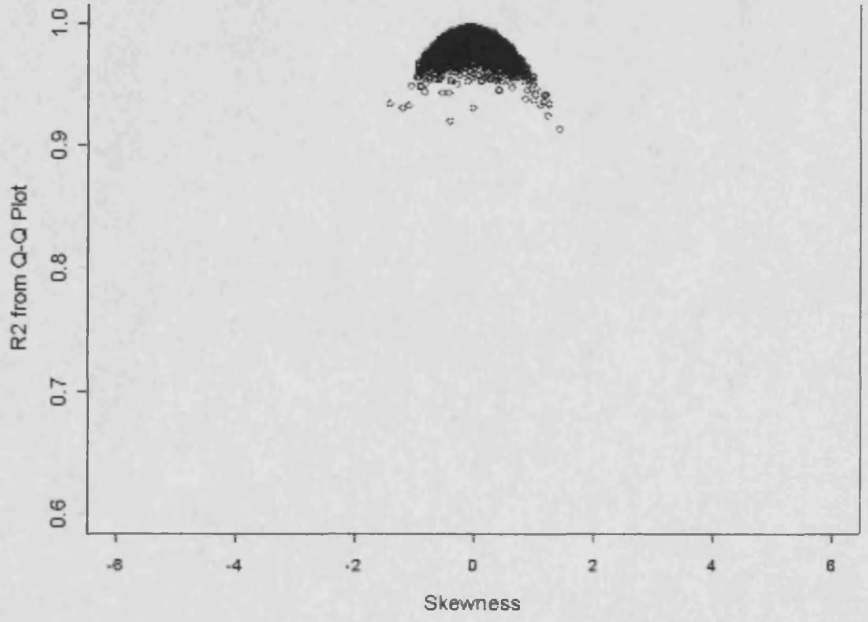
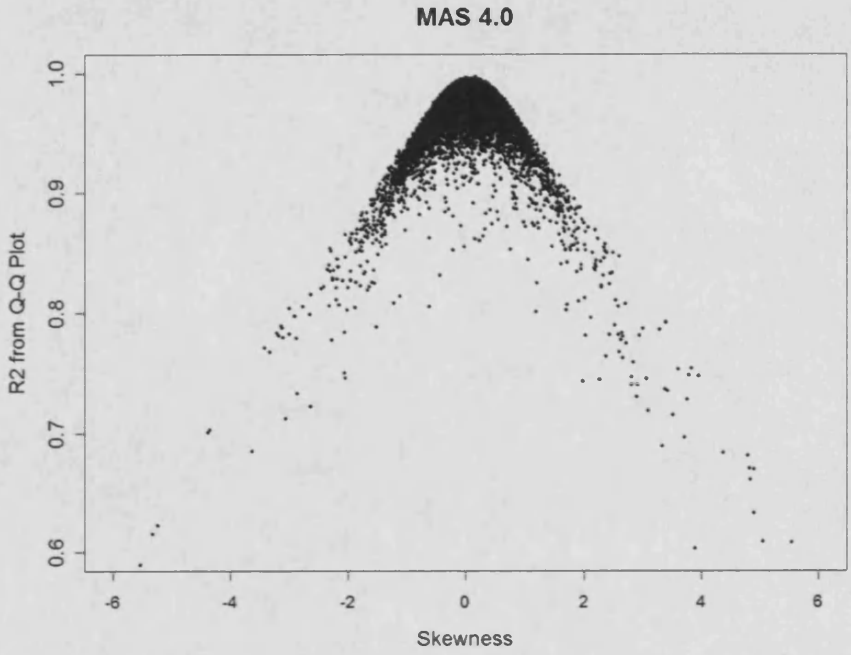


Figure 2.7 – Plot showing the correlation to normality and skewness for 12000 normally distributed datasets with 59 values.

The MAS 4.0 and both dChip datasets have few non-normal probe sets, with an even spread of skew scores (Figures 2.8.a, 2.8.b and 2.8.c). In contrast, the MAS 5.0, RMA and gcRMA datasets contains a large number of non-normal probe sets with a marked positive skew (Figures 2.8.d, 2.8.e and 2.8.f). In addition, there is strong correlation between the strong positive skew and a poor correlation to normality.

Figure 2.8

a)


MAS 4.0
R2 from Q-Q Plot
1.0
0.9
0.8
0.7
0.6
-6
-4
-2
0
2
4
6
Skewness

b)

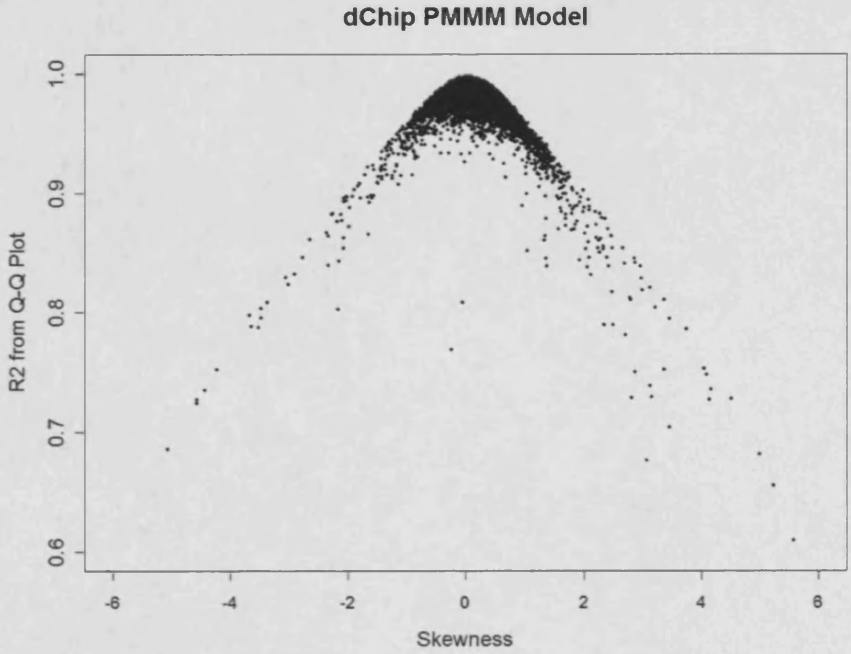
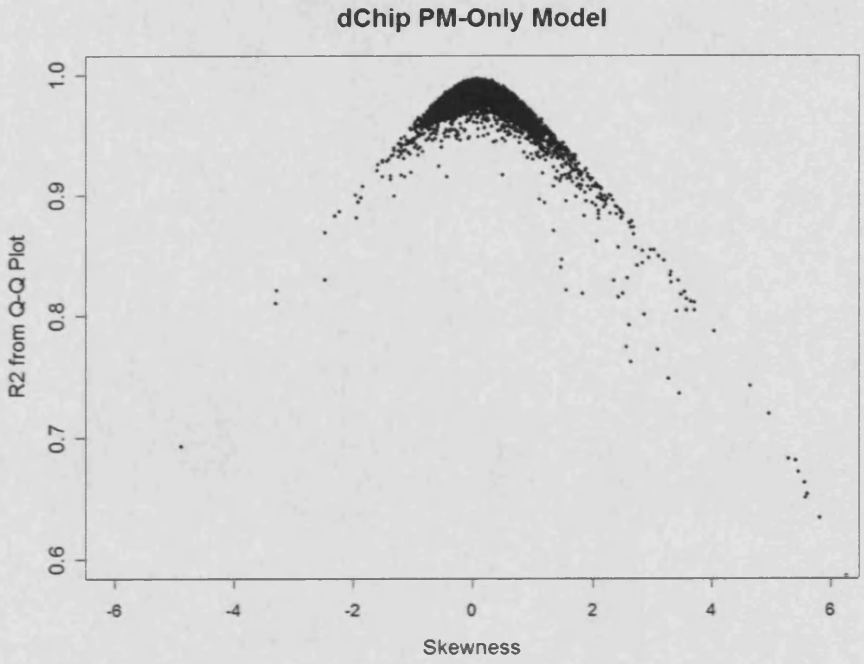

dChip PMMM Model
R2 from Q-Q Plot
1.0
0.9
0.8
0.7
0.6
-6
-4
-2
0
2
4
6
Skewness

Figure 2.8 - continued

c)



d)

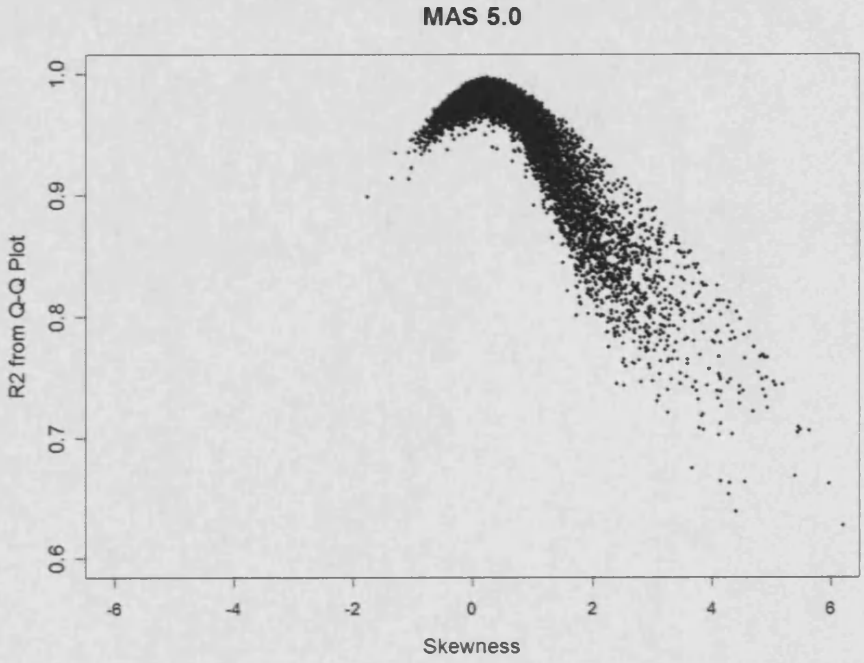
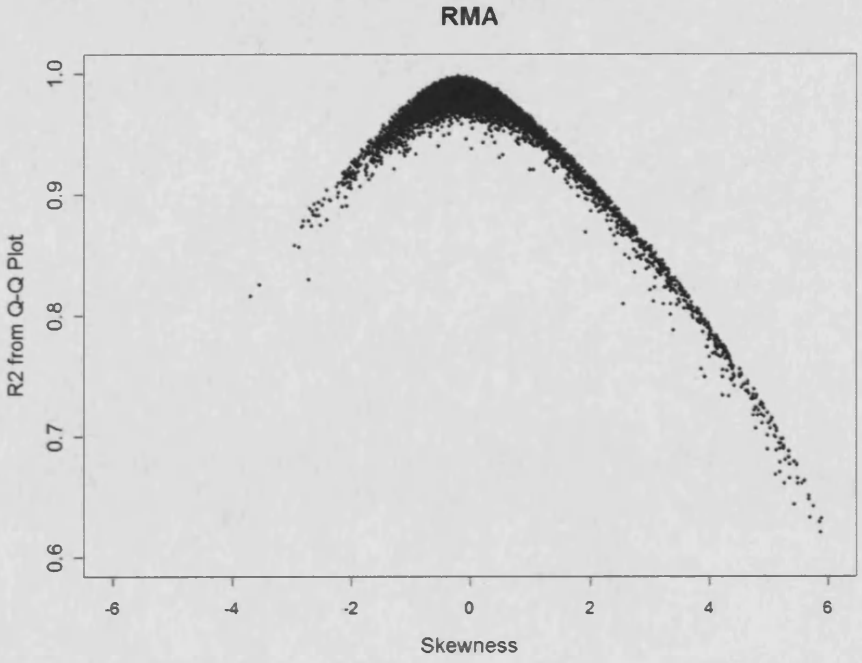


Figure 2.8 – continued

e)



f)

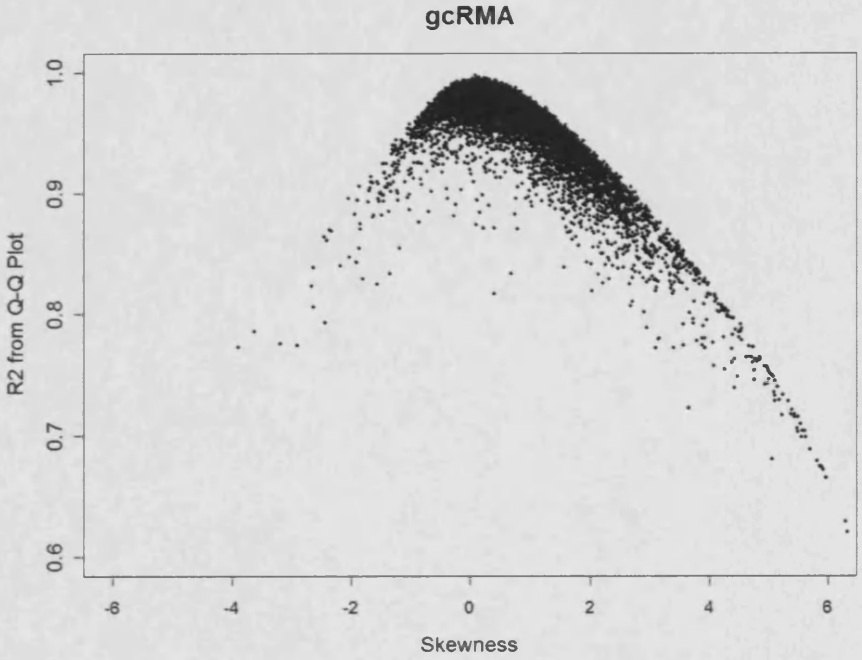


Figure 2.8 – Plots showing the correlation to normality (R^2 value) and skewness for each probe set.

2.3.5 The effect of data transformation filtering techniques

In light of the observations of non-normality in Absent called probe sets in MAS 5.0 data, gene filtering based on excluding all Absent calls would be extremely effective in removing the group of non-normal probe sets, and such an approach is commonly used by research prior to other analyses (e.g. clustering). However, there may be occasions when there is a desire to apply a statistical test to a gene called Absent in some samples, and present in others, such as an off-on regulation signal. Such a situation causes comparison to be made between data sets with very different distributions.

Functions can be applied to datasets that result in a transformation of the data distribution. An appropriate transformation of a dataset can often result in a dataset that approximately follows a normal distribution. Application of this technique can increase the number of tests available when their raw data does not meet the assumptions of standard tests.

Driven by the requirement of parametric tests to have data that follow a normal distribution, the question of whether skewed probe sets could be converted to normality by the application of a standard mathematical transformation was addressed. Examples of such transformations worthy of consideration include logarithmic, square root and inverse function since these can often correct the positively skewed data such as that presented by the MAS 5.0, RMA and gcRMA expression metrics.

2.3.5 Application of a logarithmic transform to correct non-normality in MAS 5.0, RMA and gcRMA datasets

Log_2 transformation is common with users of two-channel cDNA microarray systems in order to make fold changes symmetrical. In the case of Affymetrix data, many researchers have chosen to log transform their data to overcome distributional problems (Golub, et al., 1999; Katsuma, et al., 2001; Virtaneva, et al., 2001). In addition, it should be noted that the expression metrics RMA and gcRMA include a log transformation as the final step in their analysis algorithm.

However, the validity of such a transformation is unclear with many researchers justifying its use by commenting on the prevalence of log_2 data in biological data (Affymetrix, 2001b). The effect of log_2 transformation was investigated using Shapiro-Wilks testing and correlation to normality from Q-Q plots (Sections 2.3.1 and 2.3.2).

2.3.5.1 Logarithmic transformation of MAS 5.0 data

Shapiro-Wilks testing of log transformed MAS 5.0 data yielded 7047 probe sets deviating from normality ($p<0.05$), compared to 5799 in the un-transformed data. Comparison of the R_p data plotted versus mean (Figure 2.9) with that from the untransformed data (Figure 2.2d) showed a loss of the pronounced tail to the plot. However, the probe sets showing lack of correlation to normality has shifted from the low expressed probes to the mid-expressed probe sets.

Figure 2.9

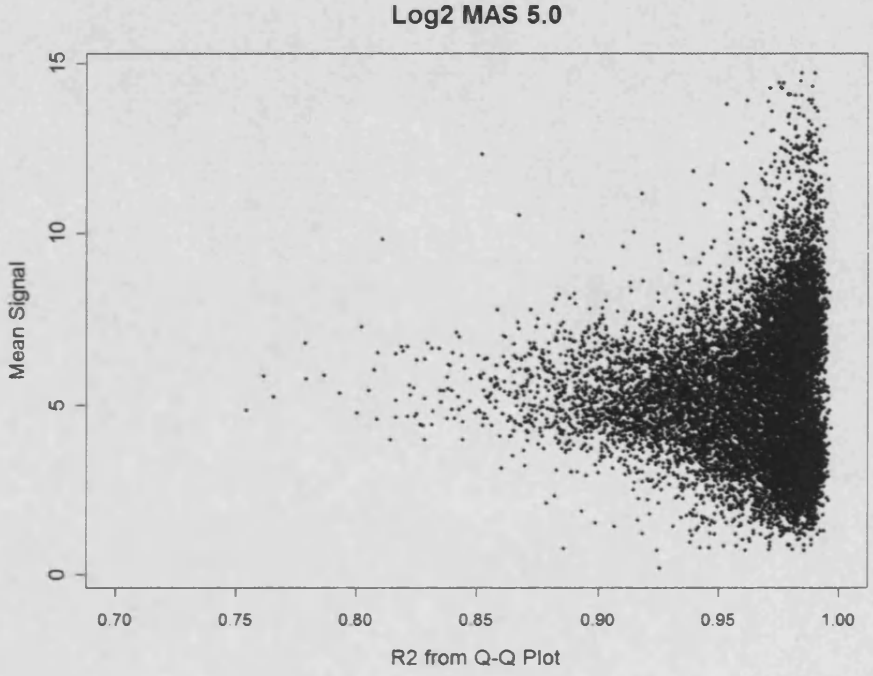


Figure 2.9 – R^2 from Q-Q plotted against mean expression value for log_2 transformed MAS 5.0 data

A plot comparing the correlation to normality values from the untransformed versus transformed data (Figure 2.10) shows that the majority of non-normal probe sets in the untransformed data have had their correlation to normality improved as a result of the transformation. In addition it shows that there are not a group of probe sets which are impossible to transform towards a normal distribution by application of a transformation. However, the plot also shows a number of probe set which correlated highly to normality in the untransformed have been moved away from unity in the transformed data. This suggests that the transformation may have corrected the normality for the non-normal low expressed probes, but altered the normality for a large number of mid to high expressed probe sets.

Figure 2.10

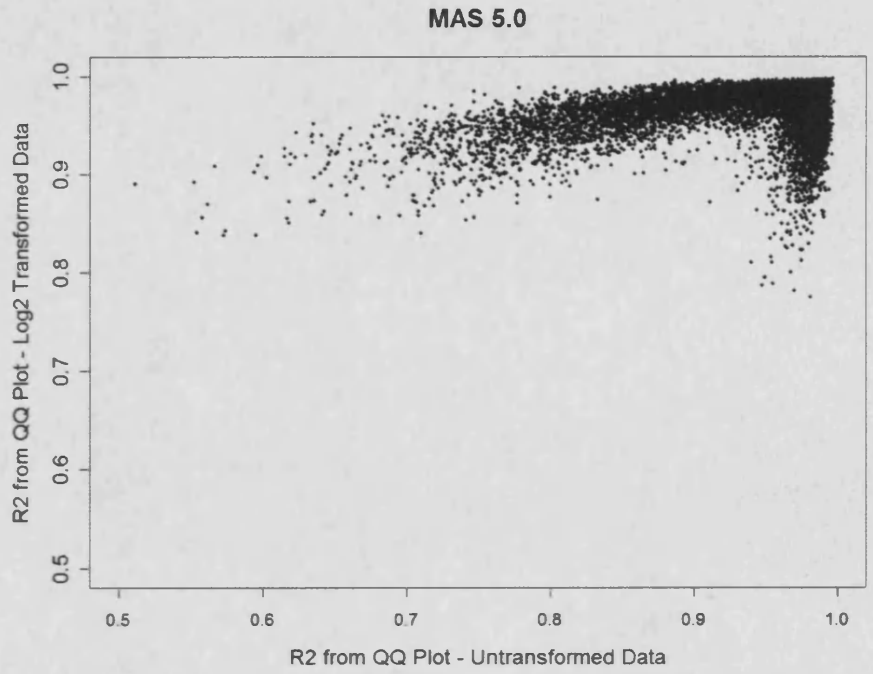


Figure 2.10 – Plot showing the correlation to normality scores for each probe set in untransformed and log_2 transformed MAS 5.0 datasets.

2.3.5.2 Logarithmic transformation of RMA and gcRMA data

RMA and gcRMA already incorporate a log_2 transformation as part of the default analysis algorithm. To facilitate meaningful comparison to other data, the transformation was removed using power transformation and the results compared to that from the default analysis.

$$x_{untransformed} = 2^{x_{transformed}}$$

Review of the R^2 versus Q-Q plots for each expression metric reveals very little change to the data distributions observed (for RMA compare Figure 2.11 to 2.2.e and for gcRMA compare Figure 2.12 to 2.2.f). Review of the Shapiro-Wilks data highlights some differences between the normality of the transformed to un-transformed data with RMA showing 3322 probe sets deviating from normality compared to 3075 in the default data. gcRMA yielded 7508 probe sets deviating from normality compared to 6371 in the untransformed data.

Figure 2.11

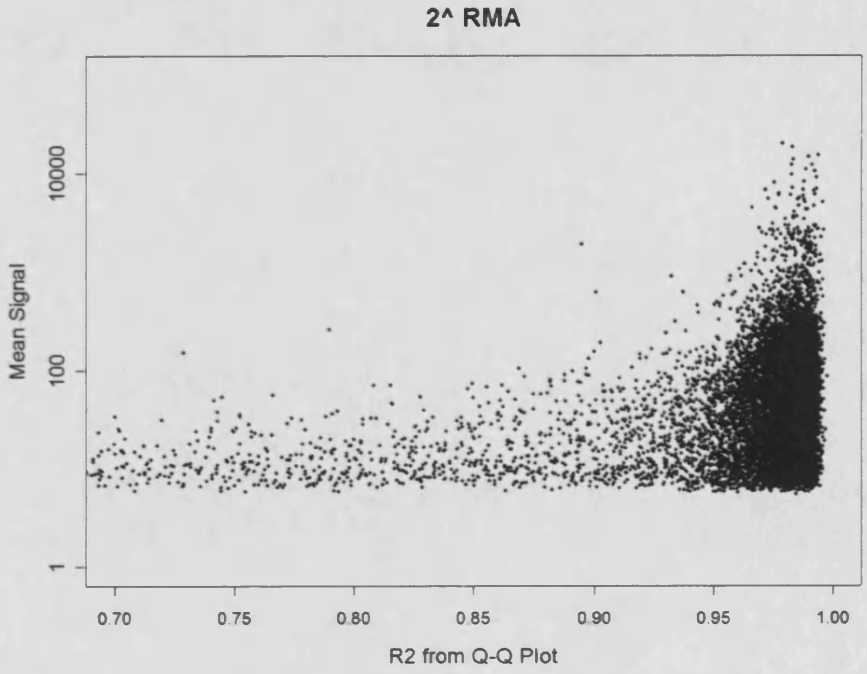


Figure 2.11 – R^2 from Q-Q plotted against mean expression value for power transformed (2^x) RMA data

Figure 2.12

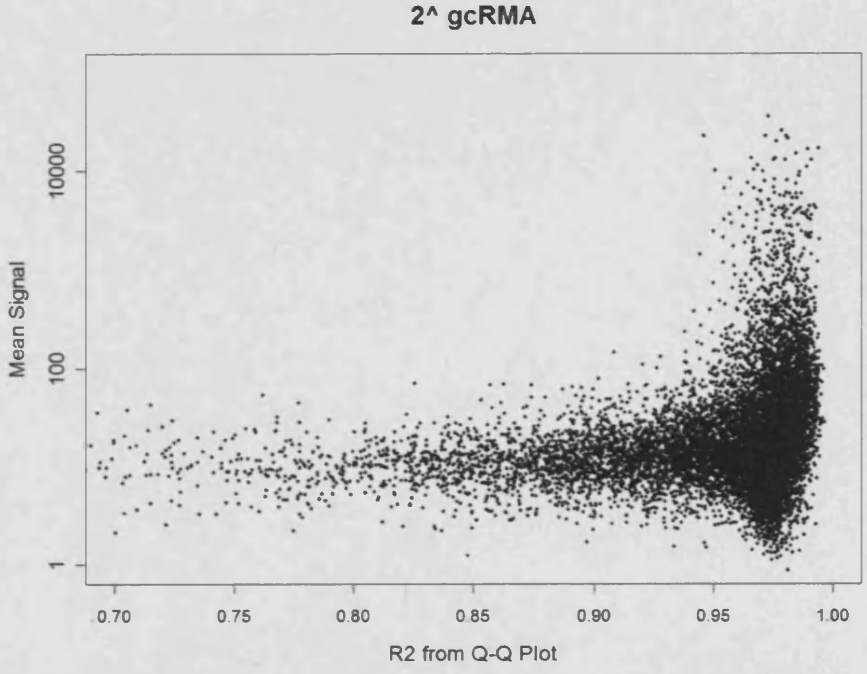


Figure 2.12 - R^2 from Q-Q plotted against mean expression value for power transformed (2^x) gcRMA data

2.3.6 Application of a Box-Cox transformation

Extending the possibilities of data transformation to improve the normality of resulting datasets the logical extension from the ad hoc application of a test and examination of the effects on normality is a systematic search for the optimal transformation. Box and Cox (Box and Cox, 1964) suggest a series of power transformation can be useful in determination of the ideal data transformation for correlation to normality using the transform:

$$Y' = Y^{\lambda}$$

The lambda value of a Box-Cox transformation correlates with other data transformations as follows; λ=1 ≡ no transform; λ=0.5 ≡ square root transform; λ=-1 ≡ inverse transform. To eliminate the problem of $Y^0 = 1$, the transformation is replaced with log(Y) when λ=0.

Given the results of a particular transformation it is helpful to define a measure of the normality of the resulting transformation. The R^2 correlation value from a standard Q-Q plot can be used to determine which transform to apply to a dataset in order to transform its distribution as close to normal as possible.

A wide range of different transformations were applied to the MAS 5.0, RMA and gcRMA datasets (lambda range -2 to 2 in steps of 0.1). For each probe set the value of lambda which produced data that was the most normal, as judged by the Rp/Q-Q plot method described above, was recorded.

Figure 2.13

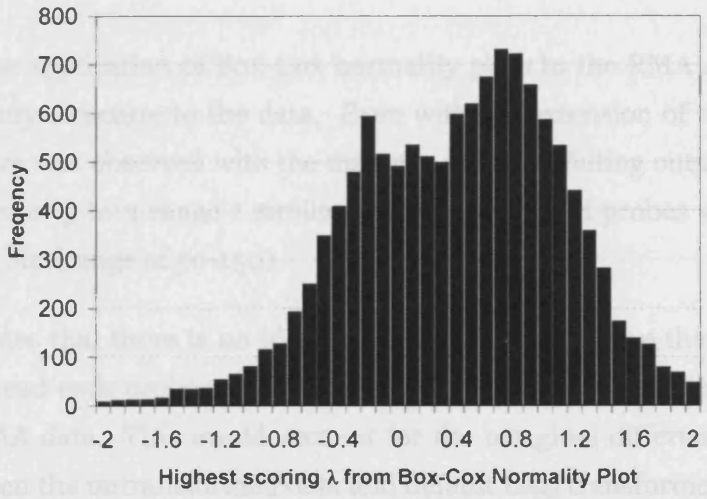


Figure 2.13 – Histogram of highest scoring transform value following Box-Cox normality plots of MAS 5.0 data.

Figure 2.13 shows a histogram of these optimal transformation values for each probe set for the MAS 5.0 dataset. The histogram suggests the presence of a bimodal distribution is observed with one distribution centred around λ = -0.2 (close to a log transformation) and another near λ = 1 (no transformation).

When the data were stratified into Present and Absent groups as before, this biphasic distribution was resolved into two separate monophasic distributions (Figure 2.14). Probe sets called Present require no transformation, which provides further support for the normality of these probe sets. In contrast, and as may be expected by the positive skewness observed in many of the probe sets, the optimal transformation for the Absent group is somewhere between a log and square root.

Figure 2.14

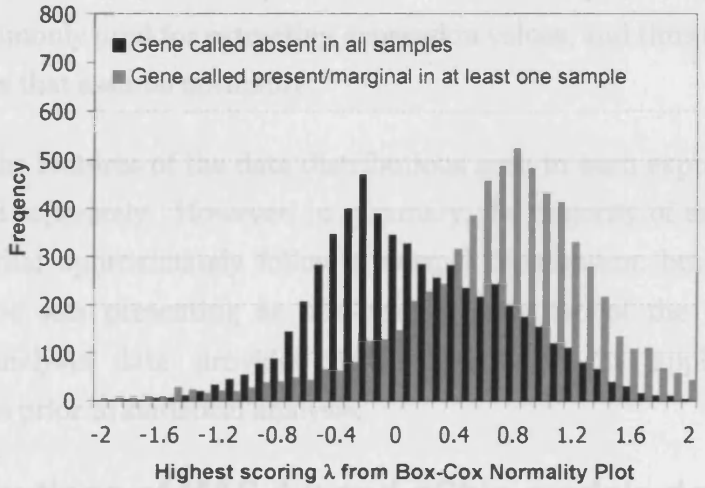


Figure 2.14 – Histogram of highest scoring transform value following Box-Cox normality plots of MAS 5.0 data split by MAS 5.0 detection call.

Results from the application of Box-Cox normality plots to the RMA and gcRMA data did not reveal any structure to the data. Even with the extension of the bin range, no obvious structure was observed with the majority of values falling outside of the -3 to 3 range. Within the -3 to 3 range a similarly small number of probes were observed in each histogram bin (range of 50-150).

The data indicates that there is no ideal transform that can move the dataset towards normality. Instead each probe set would require a differing transform when analysing RMA and gcRMA data. This would account for the marginal differences in normality observed between the untransformed data and default Log_2 transformed data discussed in section 2.3.5.

2.4 Summary and Discussion

Statistical testing is becoming more common in the analysis of microarray datasets, but many parametric tests (assuming normality of the data) have been applied without knowledge of the underlying data distribution. In some cases Affymetrix microarray expression data has been transformed with little empirical justification, simply because of a lack of data to address these issues of data distribution.

Analysis of the Latin Square 59-chip dataset provides an unprecedented opportunity to address relatively straightforward but key questions regarding the resultant data distribution obtained using Affymetrix microarray technology. The use of a combination of formalised statistical testing, graphical visualisation of the correlation to the normal distribution, and assessment of distribution features such as skew, provides support for the normality of much of the data produced by six different algorithms commonly used for extracting expression values, and thus the application of parametric tests that assume normality.

Discussion of the features of the data distributions seen in each expression metric are best considered separately. However, in summary, the majority of expression metrics produce data that approximately follow a normal distribution, but with some low-expressed probe sets presenting as non-normal with two of the analysis metrics. Overall the analysis data provides little support of the application of data transformations prior to statistical analysis.

4.2.1 Distributions of MAS 4.0 and dChip models data

Analysis of data from MAS 4.0 and both dChip models suggests that for most genes, no transformation of expression data is required. This is an important issue, as some previous studies using datasets produced using Affymetrix MAS 4.0 software chose to transform their data, having identified the presence of positive skew in some probe sets (Gieseg, et al., 2002; Stamey, et al., 2001).

It can be suggested that the apparently skewed data that was seen in these previous studies reflects not the underlying data distribution, but rather the presence of outlying data points because of biological heterogeneity. A few such extreme values can alter the distribution of a dataset away from normality, and logarithmic transformations can indeed provide an ad hoc route to reducing the influence of extreme outlier values (Iglewicz and Hoaglin, 1993).

The results from this Chapter suggest caution should be applied when applying transformations in such cases, because of the potential to distort the bulk underlying distribution. Instead, if outliers are an issue, consideration should be given to detection and accommodation of the outlying data points, using either more robust statistical tests (Lonnstedt and Speed, 2002) or by the use of more robust measures of location and spread such as the median or median absolute deviation (Iglewicz and Hoaglin, 1993). In contrast, there may be some special situations (e.g. variance stabilisation) when data transformation may be justified (Durbin and Rocke, 2003).

The analysis results suggest that datasets produced using dChip (either the PM or PM-MM model algorithms) and MAS 4.0 show a distribution close to normality that does not strongly violate the assumptions of classic parametric tests such as the t-test. It can, therefore, be argued that there is no a priori rationale for the application of a logarithmic (or any other) transformation to the expression values obtained from the analysis of Affymetrix image data prior to the use of parametric tests.

2.4.2 Distributions of MAS 5.0 data

Analysis of data generated using MAS 5.0 reveals an interesting situation. For probe sets identified as expressed (Present) by MAS 5.0, the outputted data appears normally distributed, making valid a t-test for a gene that is expressed in both groups.

However, a large number of low expressed genes, called as Absent by the package, instead follow a skewed distribution (e.g. Figure 2.6). This would compromise the use of the t-test without additional data transformation, although in practice it is unlikely that such a test would be applied to a gene that is not expressed in either sample. If necessary, something close to a log transformation may be suitable for such probe sets (see Figure 2.14). However, a generalised application of a log transformation to an entire MAS 5.0 dataset is not warranted, as for most genes this would distorts the data from the normal distribution (Figure 2.9)

It can only be speculated as to why MAS 5.0 is unique in distorting the data distribution of genes as they decrease in expression value. However, it is possible that it is the incorporation of the perfect match and mismatch probe information in combination with the background determination that result in the distortion observed; specifically how situations where the majority of mismatch probes score higher than the perfect match probes in a probe set.

A review of the algorithms behind each analysis method (Affymetrix, 2002a; Affymetrix, 2002b; Li and Hung Wong, 2001a; Li and Wong, 2001b) shows that dChip and MAS 4.0 employ a single algorithm to deal with all data, and if MM values are used this can produce negative expression values. In contrast, MAS 5.0 uses a different approach to using both PM and MM values in the expression calculation, one that avoids the production of negative values.

Affymetrix argue that the MM signal contains most of the background cross-hybridisation and stray signal affecting the PM probe and if this value is less than the PM value it is a physically possible estimate for background. However if the MM value is larger than the PM value, then the estimate is physically impossible and an idealised value must be substituted from the data in the rest of the probe set.

MAS 5.0 employs a scenario-based approach to expression calculations, and a decision process is used when this PM>MM assumption is broken. Generally MM values that are larger than their PM counter parts are replaced with a value determined as representative from the rest of the probe set. However, when the data is too close to the PM value, a value slightly less than the PM value is substituted for the experimental signal reading. It is probable that the result of this step to avoid negative values that produces the side-effect of skewed distributions.

Such an abrupt switch between models clearly has the potential to produce apparently distorted data if, in some experiments, a probe set is analysed using one algorithm and in other experiments a different process is applied. This effect can be modelled by producing a dataset with a normal distribution with a mean of zero and then shifting all values less than zero to be zero. Figure 2.15 shows the resultant Q-Q plot for such a model, which shows close correlation to that obtained from skewed MAS 5.0 probe sets.

However as the replacement values are typically sampled from real probe sets values, the predicted effect would be an overall distribution that would appear to originate from two separate normal distributions. Indeed, visual inspection of the normal Q-Q plots reveals not a smooth skewed distribution, but rather an abrupt change in gradient of the slope, producing two straight lines, each equating to a normal distributed portion of the dataset (Figure 2.6).

Figure 2.15

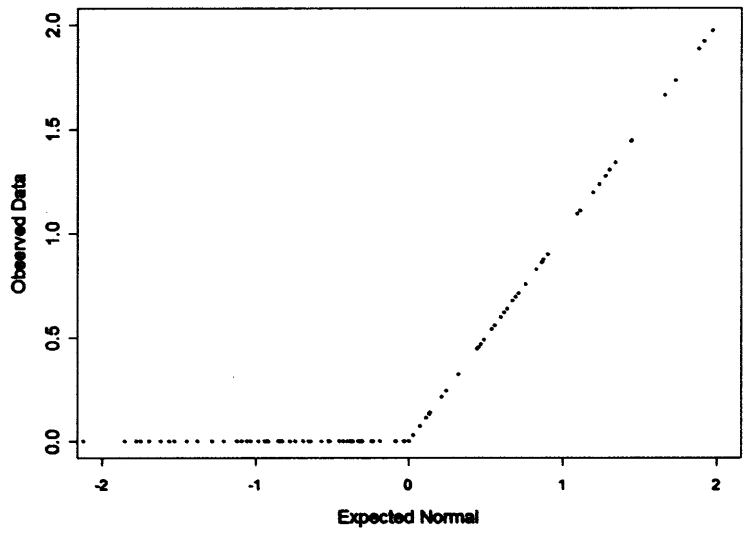


Figure 2.15 – Modelling of a normally distributed dataset with mean of zero, whose values less than zero are shifted to zero. Note the similarity to that shown for the MAS 5.0 dataset in Figure 2.6

2.4.3 Distributions of RMA data

Review of the data from RMA would suggest that application of parametric testing to these datasets would perform similarly to MAS 4.0 and dChip. There does not appear to be any pattern to the non-normality observed with these appearing in the very low expressed probe sets. One possible explanation for this observation is that it results as an artefact of determining real signal from background noise; RMA applies a single uniform background correction compared to a more complex zonal method employed in the Microarray Suite metrics.

Analysis of transformations applied to RMA data shows no benefit to the application of common transforms coerce the data into correlating better with normality. The data provides no support for the removal of the final stage log transformation from analysis using RMA.

2.4.4 Distributions of gcRMA data

Review of the gcRMA data indicates that the metric presents as the most problematic of all the expression metrics under consideration. Statistical testing for normality showed that 51% of probe sets present with a non-normal distribution and R_p from Q-Q plots versus mean expression show normality issues with many low to mid expression probe sets. Like the MAS 5.0 dataset the majority of these non-normal probes present with a positive skew.

However, unlike the MAS 5.0 data there seems little discernable pattern to the non-normality and no single transform could be identified to improve the normality of the whole dataset. The source of the problem may be as a result of the procedures the model uses to describe non-specific binding variation by integration of the MM probe data and sequence information into the background correction. Whilst the authors claim improvements in sensitivity and specificity of when compared to other methods, the evidence suggests that caution should be applied when considering the statistical analysis of image data analysed using gcRMA.

2.4.5 Conclusions

Overall the data provides strong support for the application of parametric statistical tests to data in four of the analysis algorithms (MAS 4.0, two dChip models and RMA).

Some care must be applied to certain probe sets called Absent in MAS 5.0 datasets when looking at differential gene expression situations presenting as on-off regulation using MAS 5.0 data due to apparent normality of one group and non-normality in the other. Control of the non-normality observed in the Absent group by application of a log transformation is one possible solution. However the transform degrades the normality seen in the Present group which is the data that a researcher may have the most confidence in experimentally. In this situation a researcher may wish to explore analysis methods that are more resilient to differences in the underlying data distribution and deviation from normality.

Further investigation is required to the suitability of data obtained with gcRMA. Whilst evidence for normality does not particularly support the use of parametric tests, gcRMA performs well when compared to other expression metrics across a range of performance indicators (Cope, et al., 2004).

With the exception of RMA and gcRMA data, where the application of transformation makes very little difference to the overall normality of the dataset, there is no evidence to suggest that application of a transformation improves a dataset, and in most cases actually makes the correlation with normality worse.

Although data distributions are an important issue when applying a statistical test, it should be considered that tests that violate assumptions perform perfectly adequately in the real world determining real biological differences. Further analyses of these expression metrics to address other related issues, such as sensitivity and bias, is therefore important (Lemon, et al., 2002).

Chapter Three
Statistical approaches to the detection of differentially regulated genes in Affymetrix datasets

In this Chapter the issue of applying statistical testing to microarray datasets is explored. Section 3.1 introduces key questions regarding basic data analysis, sample sizes, experimental design and the idea of developing best practice in analysis. Section 3.2 reviews the methods used to investigate the effects differing combinations of statistical tests and expression metrics over a range of experimentally plausible sample sizes undertaken in Section 3.3 using the U95A Latin Square dataset. Section 3.4 discusses the results and observations regarding the application of statistical tests within an analysis.

Conceptualisation and initial drafting of the FDR framework implemented in these explorations was undertaken by David Kipling.

3.1 Introduction

Over the last few years many funding bodies have invested heavily in projects using microarray technology and are in the process of producing very large quantities of experimental data. Despite this multi-million pound investment there remains a very real perceived weakness in our current ability to analyse these datasets: *"If the collection, analysis and interpretation of the data are flawed then it may not only be a waste of a valuable resource - we could draw faulty conclusions and potentially risk our health and environment."* Nick Fisher, President of the Statistical Society of Australia, quoted in (Tilstone, 2003).

In Chapter Two it was argued that, because of issues of small sample sizes, a researcher looking to apply statistical tests for differential gene expression should look towards parametric testing. However, these parametric tests typically come with the caveat that the data must follow a normal distribution, something that in general was met by each of the expression metrics reviewed, with the exception of some lower-expressed probe sets analysed using the MAS 5.0, RMA and gcRMA expression summaries.

However, these theoretical observations about which metric is "best", looking purely from a viewpoint of the data distributions, may not necessarily identify the expression metric and test that is most powerful when it comes to identifying differentially expressed genes in an experiment.

Similarly, even though the assumptions of a statistical test have been violated, or lower power can be inferred because of small sample size, the test may still be able to detect the differentially expressed genes of interest to the research.

This Chapter undertakes various investigations into the performance of each of the six expression metrics and investigates the *"real-world"* application of statistics to detect differentially regulated genes. Such data should provide guidance about the choice of methods for accurate microarray data analysis.

3.1.1 Defining *"best practice"* in data analysis

The idea of good analytical practice in microarray experiments is becoming well sounded in a variety of literature articles (Brazma, et al., 2001; Miller, et al., 2001; Tumor Analysis Best Practices Working Group, 2004). However, the majority of issues raised, and solutions proposed, are centred on complete annotation of experimental procedure and experimental design to allow comparative meta-analysis of data into the future.

In contrast, to date there has been relatively little guidance regarding *"best practice"* in the analysis of microarray data, with many researchers forced to choose a heuristic approach to the development of an analysis strategy due to a lack of firm evidence about the choices they should be making.

Miller makes comprehensive recommendations about the requirements for statistical analysis on microarray datasets, concluding that "Reports of differential gene expression should not be published unless they contain either significance tests or, at least, calculated estimates of the number of expected false positives". He goes on to suggest that all research proposals should include a formal power analysis relating to the number of samples being run, and makes the case for archiving the results to form a public resource (Miller, et al., 2001).

3.1.2 Sample size and statistical power

The issue of sample size has been previously discussed, specifically with relation to the potential application of parametric testing (2.1.3). More generally the issue of sample size is of interest to statisticians in the context of power. Statistical power is the probability that meaningful difference will be detected if present within a dataset (Motulsky, 1999).

In addition to sample size, other factors that influence the statistical power of a test include variability of the population, the desired detectable differences, the power of discrimination of a test and an acceptable error rate. In addition, experimental design, technical variability and data pre-processing play a role in the power of the statistical tests (Wei, et al., 2004). Interpretation of issues regarding sample size and power when applied to microarray data would appear to be even more difficult than simple estimation of each factor and their input to defined power calculations.

Tsai et al. raise the issue that current methods for power estimation do not consider the dependency of expression levels between genes (instead they are considered as un-related observations) (Tsai, et al., 2004) and Wei et al. (Wei, et al., 2004) reported that fewer individuals are needed for the same statistical power when using inbred animals rather than unrelated human subjects.

Practically, Dow reported that a meta-analysis of a larger experiment revealed that a sample size less than ten impaired the capacity to detect differentially expressed genes. As many researchers are interested in running experiments with far fewer than this reported threshold, it is of interest to know how additional replicates influence power, and ultimately the correct identification of differentially regulated genes within a dataset (Dow, 2003).

3.1.3 Application of statistics in the determination of differential gene expression

Supporting the idea of good analytical practice in microarray experiments are the development of many academic and commercial packages for microarray analysis that provide some excellent statistical capabilities. Academic examples include MeV (Saeed, et al., 2003) Genesis (Sturn, et al., 2002) and maxD (Hancock, et al., 2000), whilst popular commercial packages include GeneSpring (Silicon Genetics) and Rosetta Resolver (Rosetta Biosoftware).

However, where they often fall down is their all-encompassing, non-proscriptive nature; almost any test can be applied, irrespective of whether that is a sensible thing to do or not. Quackenbush describes this problem in detail, one where the sophistication of the computational tools often out-strips the understanding necessary to use them correctly (Quackenbush, 2001).

In Chapter Two, parametric tests such as the classic t-test were introduced as potential tests of interest for application to the small sample sizes faced in typical microarray research. A variety of parametric tests have been developed for the identification of differentially regulated genes in microarray data sets, including several variants of Student's t-test (Baldi and Long, 2001; Kooperberg, et al., 2002; Lonnstedt and Speed, 2002; Tusher, et al., 2001) and ANOVA (Kerr, et al., 2000; Pavlidis and Noble, 2001). Pan (Pan, 2002) highlighted the problem in applying statistics to microarrays, in that a researcher must make assumptions about their data in order to select a test, and in situations where multiple tests are available, there is a lack of information about how the results of a test compare to each other. Breitling et al. (Breitling, et al., 2004) add to this argument, commenting that the reasons for differing performance from current statistical techniques is poorly understood.

Focusing in on the t-test, many different variants on the basis variety have been proposed; Thomas et al (Thomas, et al., 2001) proposed a variant based on regression modelling, Pan (Pan, 2002) suggested a mixed model approach and Baldi and Long (Baldi and Long, 2001) developed a Bayesian framework approach. Although attempts have been made to compare these different methods (Pan, 2002), they have concentrated on results of analysis on real biological datasets, and thus conclusions could only be based on comparison of ranks between tests and examination of significant gene lists.

In section 2.1.4 the potential limitation of non-parametric testing in light of small sample size was highlighted. As an example, Thomas et al. demonstrated the price paid with loss in power when applying a non-parametric test, with no genes being returned as significant after the application of the Mann-Whitney test (Thomas, et al., 2001). However, there is a point at which the power between non-parametric and parametric testing should converge, and at this point the additional resilience to outlying data points and deviations in distribution are of use to the researcher.

3.1.4 Comparisons of expression metrics

In addition to the effect on output from the application of differing statistical tests, the other key area for differences in the analysis of microarray data is the expression metric used to convert image data from the scanner into numerical representations for analysis. Each development in metrics attempt to make an improved calculation of the *"truth"* within a GeneChip and minimise variability between chips to enable easier comparison and subsequent analysis.

Lemon et al. (Lemon, et al., 2002) made comparisons between MAS 4.0 and the dChip PMMM model using a variety of real-world and simulated datasets. Using a combination of analytical arguments, along with empirical data from data derived from spiked-in bacterial controls, they concluded that the model based expression metric (dChip) was the superior method. Cope et al. (Cope, et al., 2004) used the Affymetrix Latin Square dataset (9.1.2) along with a dilution dataset from GeneLogic to compare MAS 5.0 to dChip PMMM model and RMA and proposed a set of benchmark measures to judge expression metrics by. They concluded that RMA was the metric of choice for analysis of Affymetrix GeneChip data.

The common theme of both these studies was the application of a technique that examined sensitivity and specificity applied to a dataset with spiked-in data, providing known truth in the expected result. It is this methodology that will be built on, and extended, in the work presented within this Chapter.

3.1.5 Key questions regarding basic data analysis and experimental design.

Many of the issues that have been highlighted so far in the potential application of statistics and expression metrics can be reduced to a series of simple questions that a scientist new to microarrays would typically ask; *"How many chips do I need to run?"*, *"Should I log transform my data before analysis?"*, *"What statistical test should I use to identify differential expressed genes?" and "What method should I used to generate expression values?"*.

In order to address these issues, methods to accurately make comparisons between differing analysis outcomes are needed. The major limitation of much published work on microarray analysis is that it applies novel techniques to "real world" datasets addressing biological problems. This presents two major problems, firstly separating the biological variation from any technical variability, and secondly knowing what is being looked for in an analysis method.

Extending the work published by Cope et al. (Cope, et al., 2004), which used the known truth in a dataset to compare the responses from different expression summaries applied to Affymetrix GeneChip data, a testing framework was developed to compare the responses from combinations of expression metrics and statistical tests against a background of known truth, across a range of experimentally plausible sample sizes.

3.2 Technical Methodology

3.2.1 Overview

An integrated analysis script was written in R which took a subset of the Affymetrix U95A Latin Square dataset and produced a series of 20 replicates at a range of 3 to 12 chips in each of two groups using the MAS 5.0, RMA and gcRMA expression metrics. Next the script took each of these indexed matrices, extracted the relevant data and then runs a series of statistical tests with twenty replicates over a range of ten sample sizes. The statistical tests chosen for these initial investigations were fold-change, un-paired t-test, Welch's variant of the t-test and the non-parametric Mann-Whitney test.

The output for each test was fed into an FDR function which extracted the location of the spikes from the ordered p-values and calculated the area under the FDR curve as previously described. The output for each test was exported into a summary matrix containing the area under the FDR curve (AUC) for each sampling, grouped by sample size. Summary plots were produced charting the average area under the curve (with standard error, error bars), across the range of sample size under consideration.

Figure 3.1

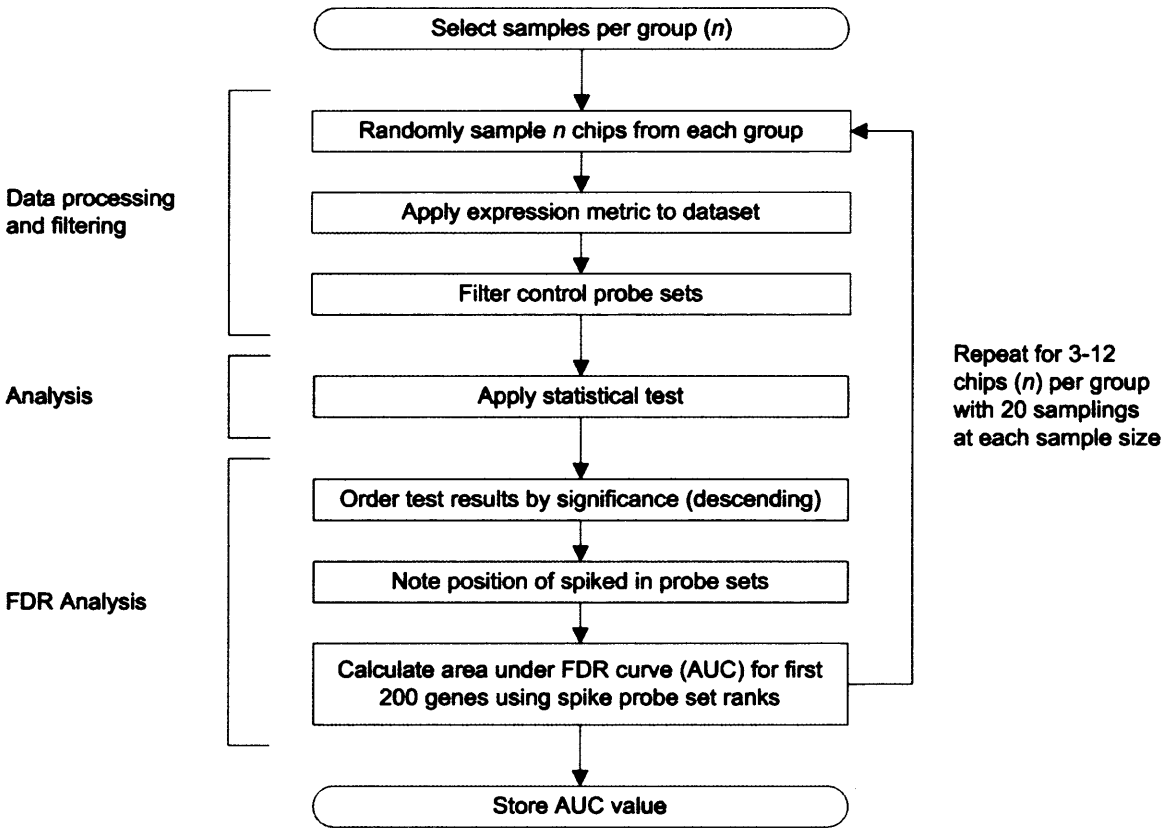


Figure 3.1 – Analysis flowchart

Datasets from MAS 4.0, MAS 5.0 and two dChip models were produced from the image CEL files using the stand-alone analysis package (as opposed to a Bioconductor implementation) and then split according to the samplings into indexed matrices using the R environment, totalling 200 datasets for analysis. The output from each run was formed into an indexed matrix for further analysis.

Full details of the technical methodology are given in Section 9.3.

3.3 Exploration and Results

3.3.1 Investigations into the detection sensitivity of fold change

Fold change describes the ratio of intensity values between two groups and are often specified as cut-off criteria for finding differentially expressed genes. The issue of calculating fold change using just a simple A/B calculation is the issue of the highly asymmetric scale created between up-regulated and down regulated genes by the same magnitude. For example an experiment with no change results in a fold change of 1, a 100 fold up regulation gives a fold change of 100, whilst a 100 fold down regulation gives a figure of 0.01.

To overcome this asymmetry, taking the log of the ratio gives a fold change value that is both symmetric and centred on zero. In addition, by taking the $\log_2$ of the data, each increment of the fold change figure represents a doubling of the fold change.

$$FoldChange = \log_2\left(\frac{\overline{X}}{\overline{Y}}\right)$$

Experimentally it was of interest to examine the power of simple fold change to correctly identify the spiked in probe sets with the Latin Square dataset and compare to those from a standard statistical test based on detection of the difference between means. The statistical test chosen was the Welch variant of the t-test. This test is an unpaired version of the t-test and does not assume heterogeneity of variance between the two sample groups.

Data from each of the 200 random sampling from the 24 chip Latin square experiment from each of the six expression metrics under consideration (MAS 4.0, MAS 5.0, dChip PMMM, dChip PM-Only, RMA and gcRMA) was analysed for the detection of the spiked in probe sets using fold change and the Welch t-test. The sensitivity and specificity of each test was examined using FDR curves.

The FDR curve implemented takes the p-values and orders them according to significance. A graph is then plotted working down the list of the first 200 probe sets in the list. For each probe set, the plot is advanced one mark horizontally with a plot of one mark vertically if the probe set is one of the spiked in transcripts (Figure 3.2). The larger the area under the curve (AUC) the more sensitive (the likelihood of detecting a true change in the dataset) and specific (the likelihood that observed results reflect true changes) the test is in the corrected identification of the spiked-in data. It is these AUC values that are returned to the framework for summation and interpretation.

Figure 3.2

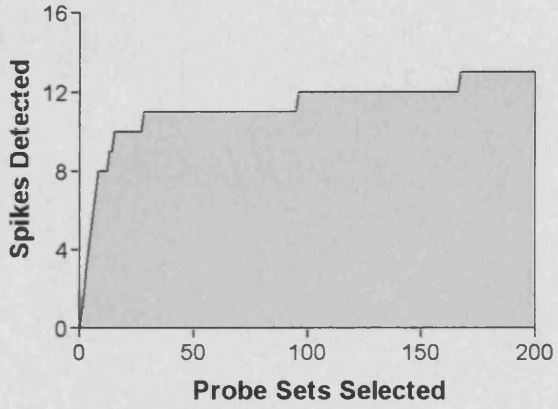


Figure 3.2 – Example FDR graph showing the sensitivity and specificity of an analysis method to detect the spikes within the Latin square dataset.

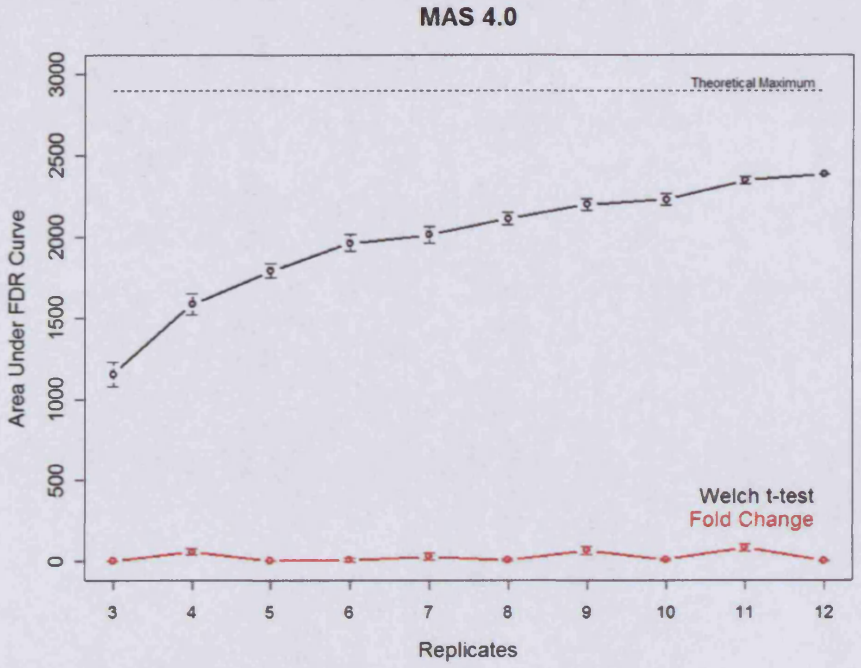
A summary graph was then drawn representing the 200 area under the curve values resulting from each analysis. The mean AUC for each sample size was plotted along with error bars representing the standard error. By reducing the data to this summary plot it is possible to view the results for many tests simultaneously and determine differences in their response at different sample sizes. The results for the Welch t-test and fold change data is shown in Figure 3.3.

Review of the summary graphs (Figure 3.3) shows that the statistical technique out performs basic fold change by a large amount across each of the expression metrics under review. Detection of spikes from data analysed using MAS 4.0, MAS 5.0 and dChip PMMM algorithms is found to be near impossible using fold change alone. dChip PM, RMA and gcRMA fair better with fold change being able to pick up some of the two-fold change in expression of spikes between groups.

The models that are able to detect spikes using fold change are all perfect match only probe models where the mismatch information is either ignored of down played in the analysis algorithm. To further characterise the data, MVA plots were examined. MVA plots have become a widely used tool in microarray analysis, comprising a vertical axis representing the difference between the logarithms of the signal (the fold change) and the horizontal axis being formed as the average logarithm of the signal intensity (Smyth et al., 2003). MVA plots are only able to examine the variability between two chips at a time, so for the purposes of investigation here, just two chips in the sample group were chosen. MVA plots for each of the six expression metrics are shown in figure 3.4.

Figure 3.3

a)


MAS 4.0
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
Welch t-test
Fold Change
3
4
5
6
7
8
9
10
11
12
Replicates

b)

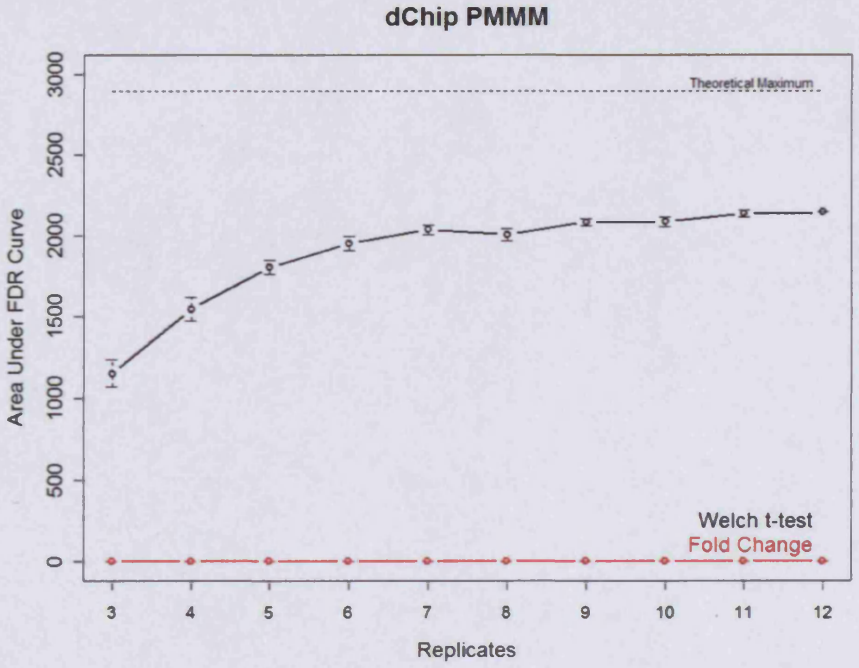
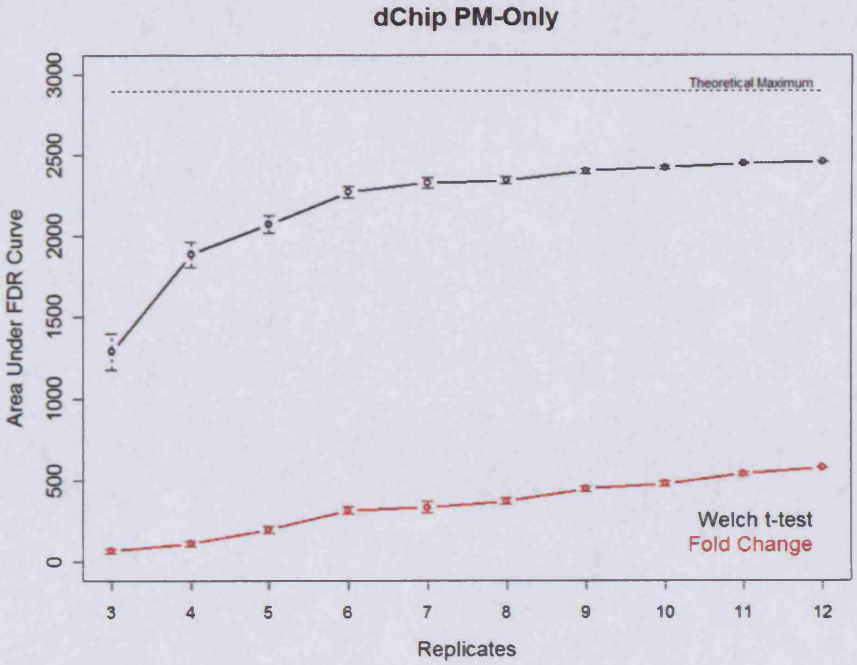

dChip PMMM
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
Welch t-test
Fold Change
3
4
5
6
7
8
9
10
11
12
Replicates

Figure 3.3 - continued

c)



d)

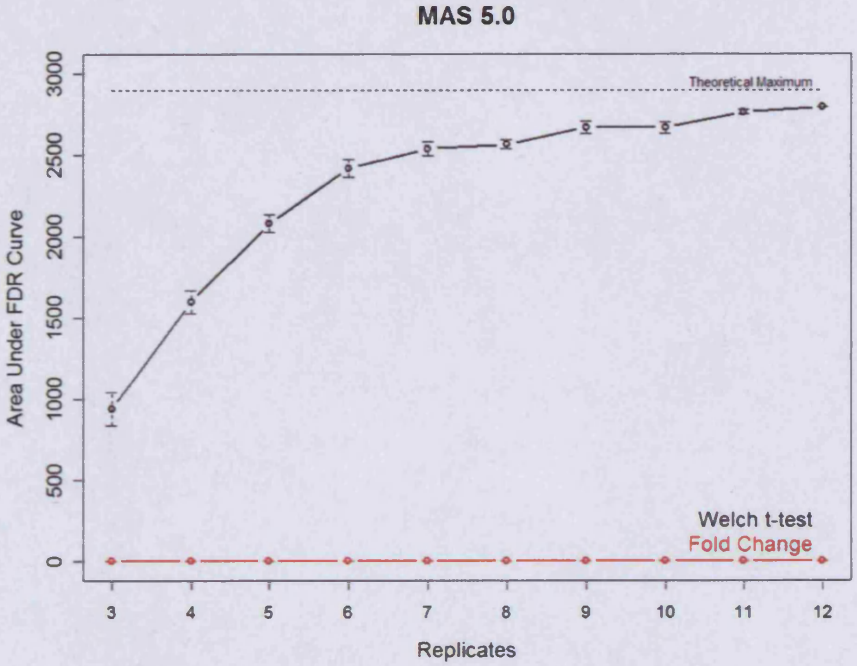
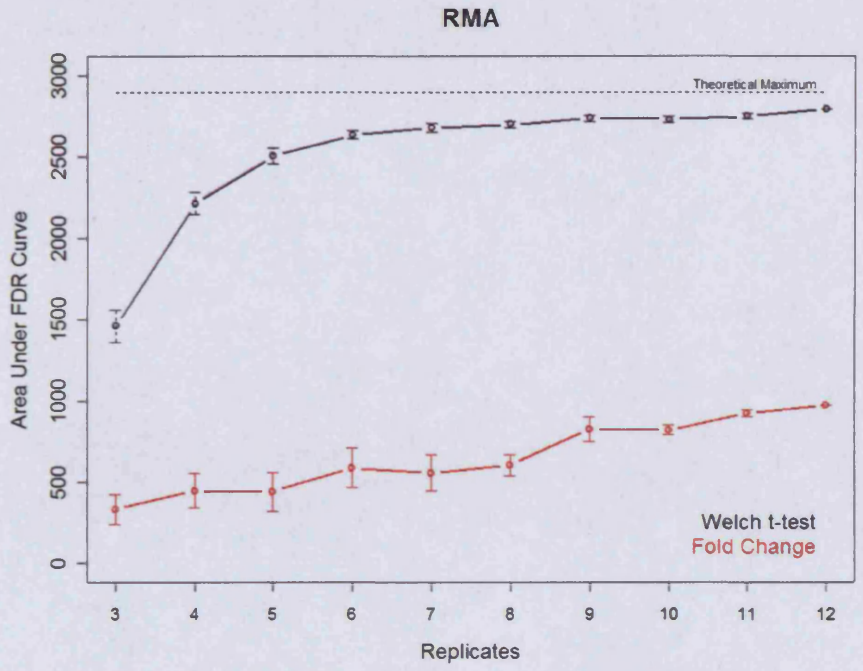


Figure 3.3 – continued

e)



f)

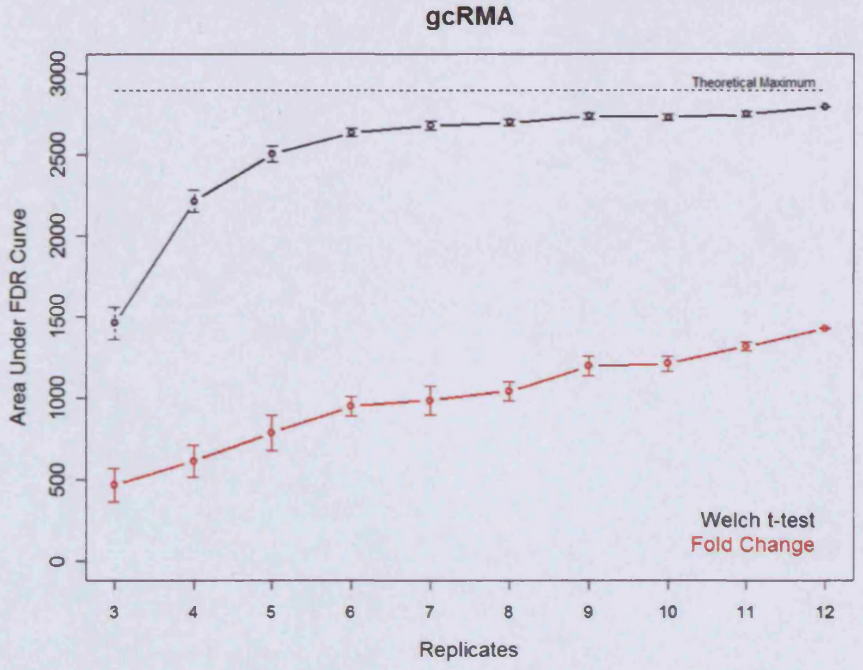


Figure 3.3 – Summary AUC plots for each of the six expression metrics, comparing the power of detection between fold change and the Welch t-test.

Figure 3.4

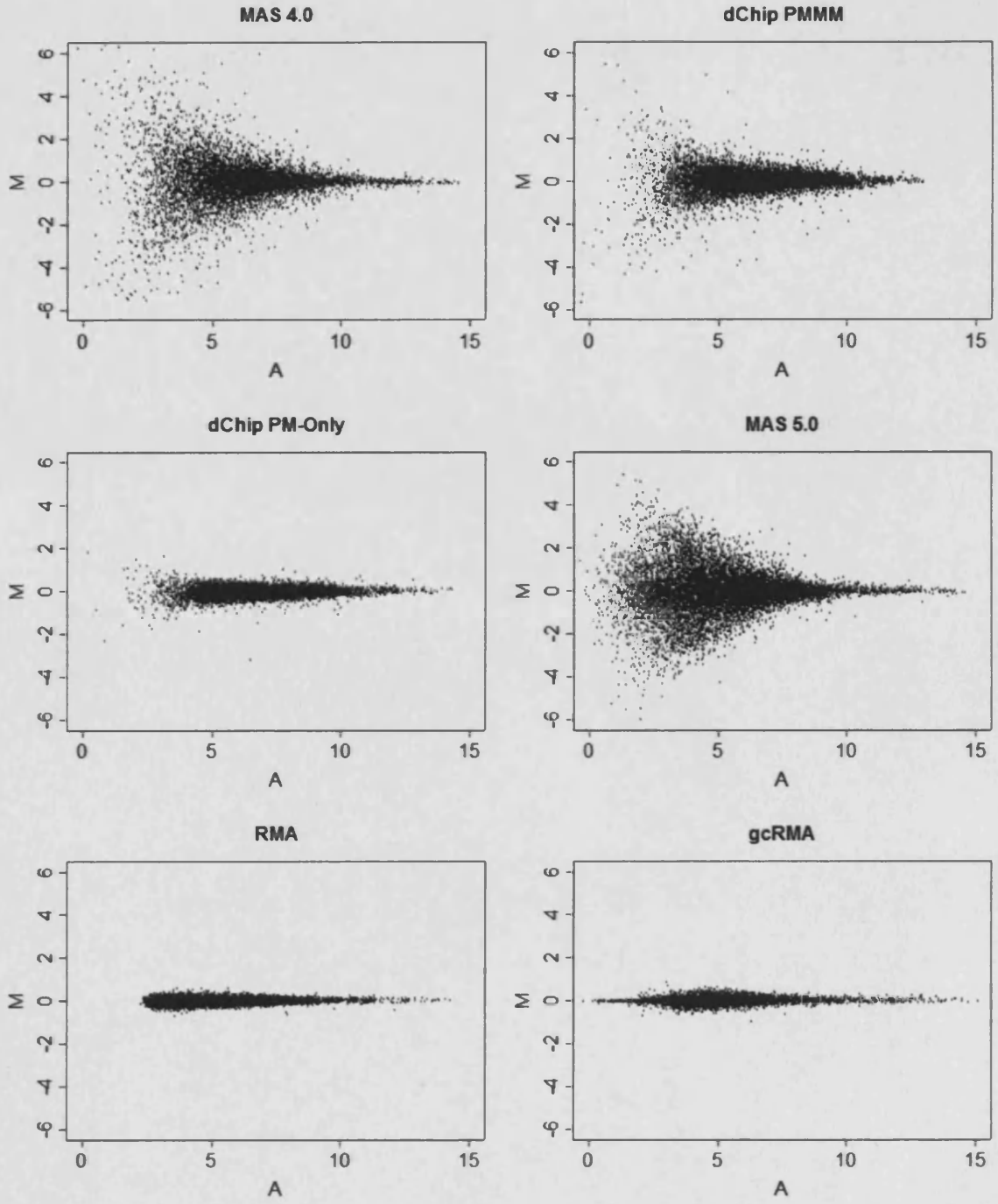


Figure 3.4 – MvA plots of each of the six expression metrics, showing the variability in expression levels between two technical variants of the same data.

It can be seen that each of the methods that give positive spike detection using fold change produce tight MVA plots with a tighter horizontal clustering to the graph. Fold change can be visualised onto these plots as the drawing of a horizontal line at a pre-determined level on the y-axis – in the case being considered here, each spike should be at +2. In both MAS models and dChip PMMM, the large cloud of highly variable probe sets at low expression values precludes the simple fold change threshold from detecting the spikes as many false positives are detected.

Whilst fold change can, to some extent, detect spikes in data analysed using dChip PM-Only, RMA and gcRMA expression metrics, the additional power achieved using a statistical technique suggest that whilst an intuitive technique, fold change is superseded in sensitivity by the statistical technique.

3.3.2 Does logarithmic transformation improve power of detection?

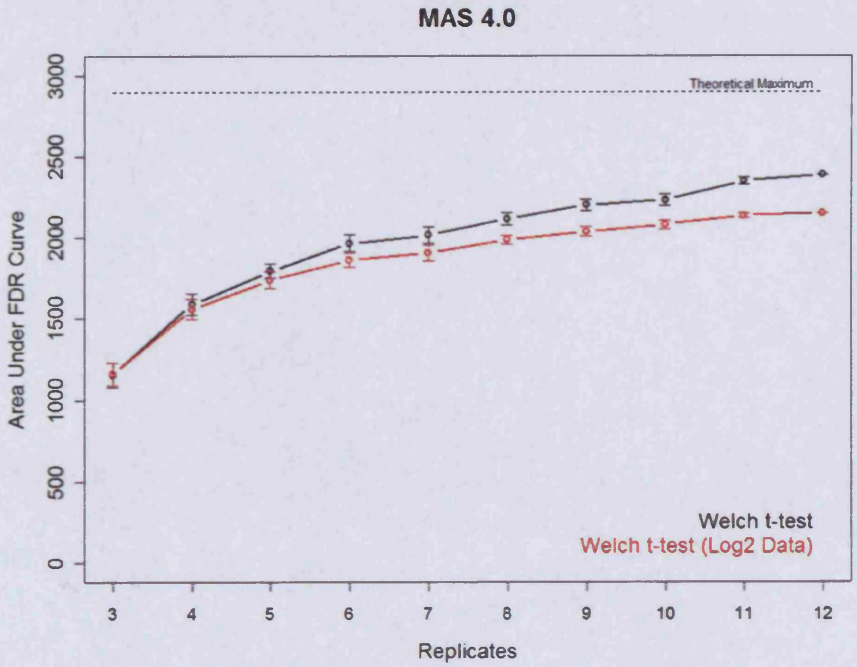
The application of logarithmic transformation was dealt with in the context of ensuring that the data complies with the assumption of normality for statistical testing in Chapter 2. Overall there seems little justification for the application of such a transformation prior to analysis, with the exception of RMA and gcRMA, where the application of a $\log_2$ transform did help conformation with normality to a very small extent.

However, the application of transforms is still a common experimental analysis proceeded and it was thus sought to investigate whether any advantage was achieved by the application of a $\log_2$ transform.

Data outputted from each expression metric was $\log_2$ transformed before analysis using Welch's t-test across a range of sample sizes before comparison using the previously described FDR methods. In the case of RMA and gcRMA data, which are provided ready-transformed, the data was untransformed with the application of a power transform resulting in an equivalent raw dataset for comparison. Welch's test was chosen as the safe variant of the t-test, making fewest assumptions about the data in exchange for a small reduction in power of detection. Summary FDR plots are shown in Figure 3.5.

Figure 3.5

a)


MAS 4.0
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
Welch t-test
Welch t-test (Log2 Data)
3
4
5
6
7
8
9
10
11
12
Replicates

b)

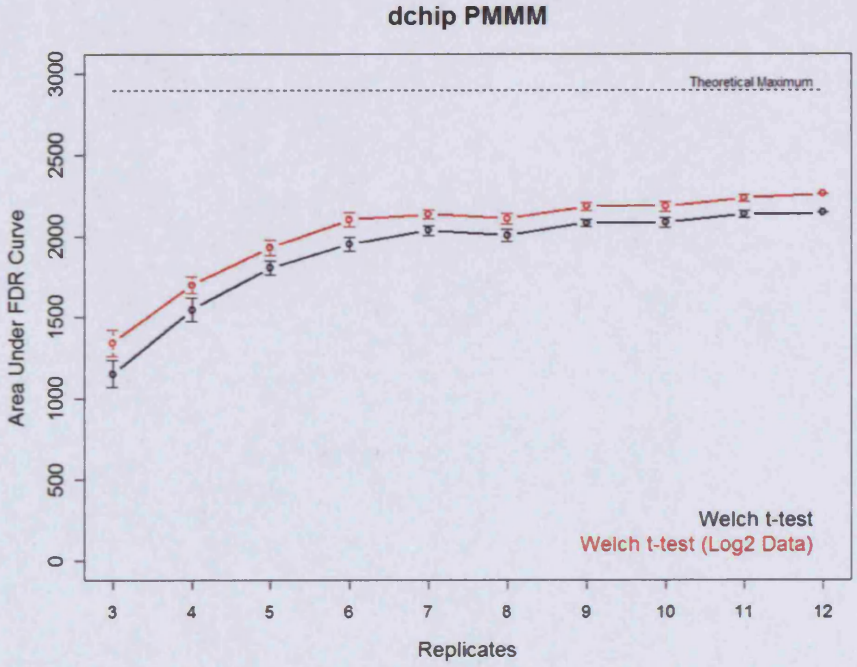
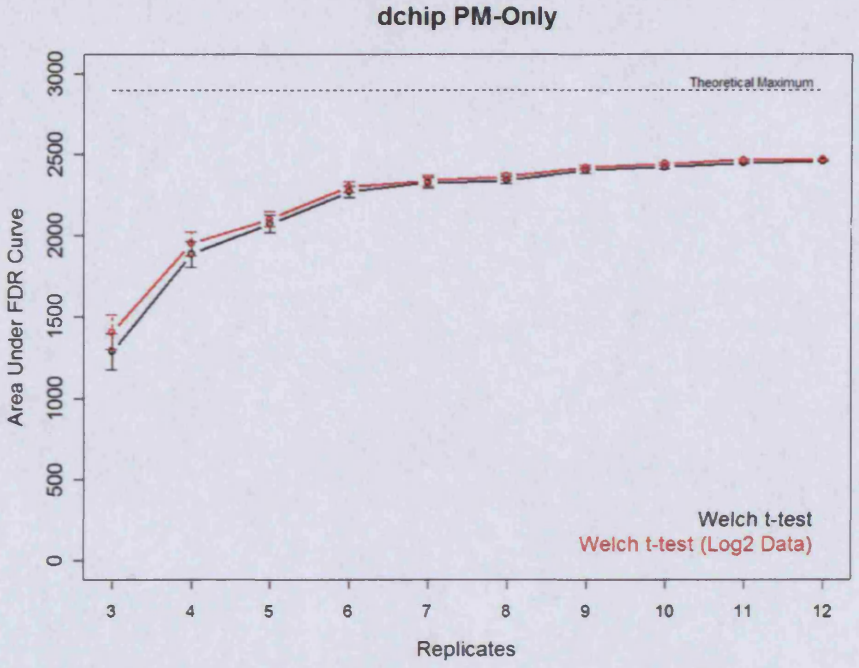

dchip PMMM
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
Welch t-test
Welch t-test (Log2 Data)
3
4
5
6
7
8
9
10
11
12
Replicates

Figure 3.5 – continued

c)



d)

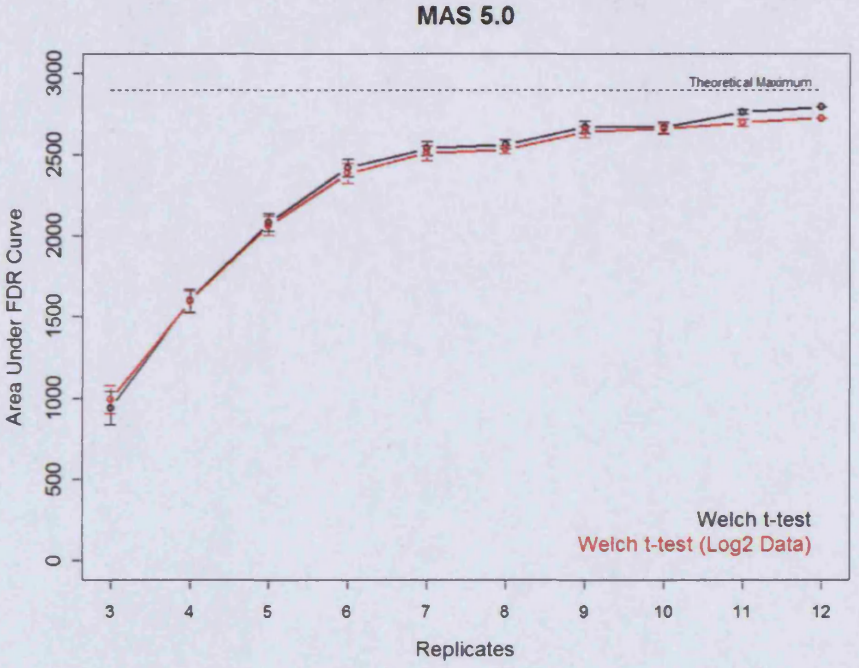
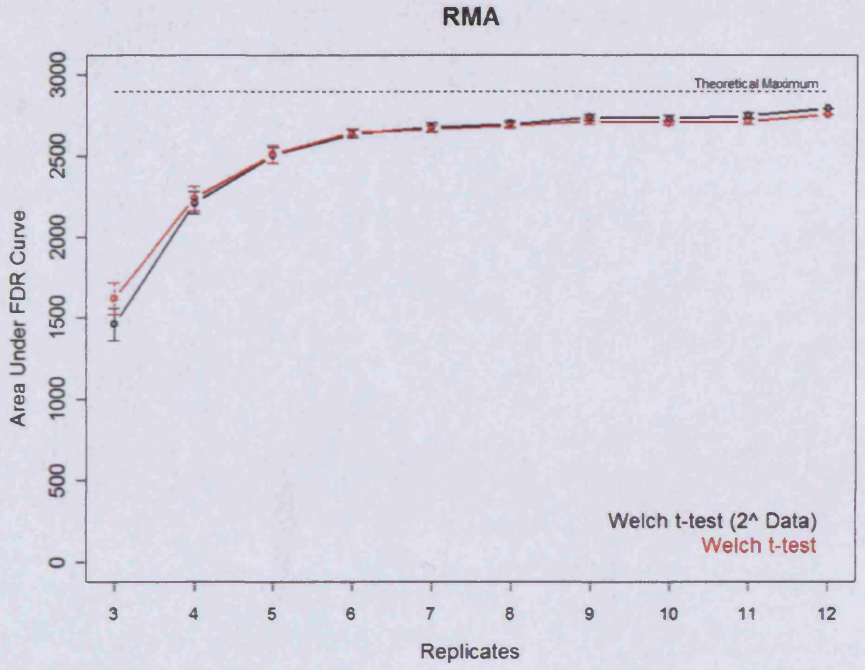


Figure 3.5 – continued

e)



f)

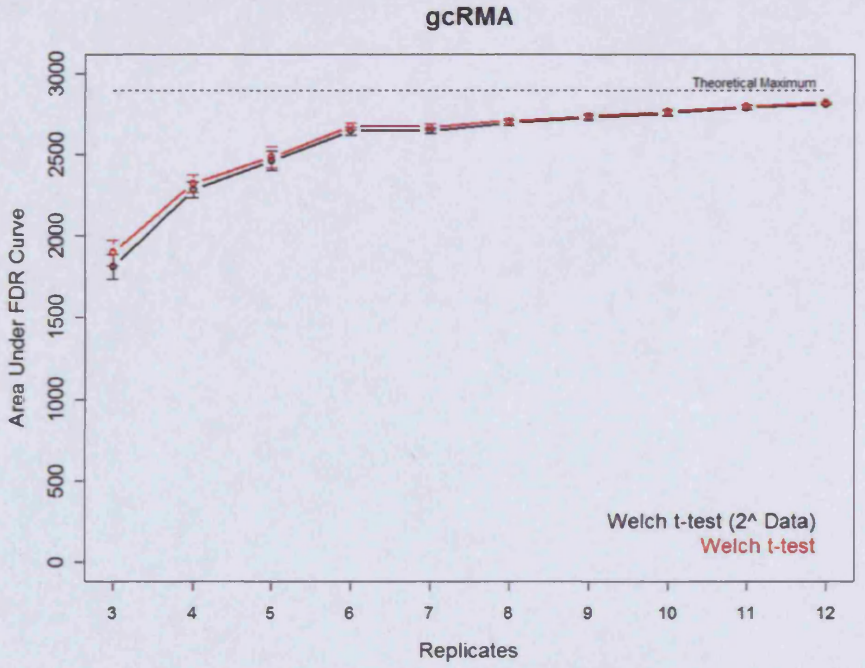


Figure 3.5 – Summary AUC plots for each of the six expression metrics, comparing the power to detect between untransformed and log_2 transformed data.

Summarised data from the area under FDR each of the FDR curves generated shows little difference in the performance between log and transformed and untransformed datasets, with the exception of MAS 4.0 and dChip PMMM model (Figure 3.5). The dChip PM-Only data was the tightest, with virtually identical results achieved with either transformed or untransformed data.

3.3.2.1 Effect of logarithmic transformation applied to MAS 4.0 and dChip PMMM data

In the MAS 4.0 data there was a slight decrease in detection ability with log transformed data (Figure 3.5.a). With the dChip PMMM model a slight improvement in spike detection was observed with the $\log_2$ transformed dataset (Figure 3.5.b). It should, however, be noted that MAS 4.0 and dChip PMMM expression metrics are the only two under consideration that produce negative numbers, and as negative numbers cannot be log transformed, these values had to be deal with before log transformation and subsequent application of the statistical test.

The approach used to deal with this problem was replacement of negative numbers with the smallest positive number within the probe set. The results of this analysis must therefore be treated with some caution as the analysis is thus undertaken on two different subsets of data, one reflecting the true data, and another modified set containing the lower expressed probe sets.

In MAS 4.0 50% of probe sets had data altered due to the presence of one or more negative values, whilst in dChip PMMM data 6% of probe sets had data adjusted. Any perceived advantage in the application of transformation or leaving data untransformed must therefore be treated with caution. As the margins of improvement are comparatively small with each expression metric, the prudent approach would be the choice of no data transformation.

3.3.2.2 Effect of logarithmic transformation applied to MAS 5.0, RMA and gcRMA data

The lack of difference in FDR performance between the transformed and untransformed data for MAS 5.0 (Figure 3.5.d) and both RMA (Figure 3.5.e and Figure 3.5.f) models correlates well with the observations in Chapter Two looking at the effect of transformation on the data distributions which showed that no marked improvement in the overall correlation to normality was achieved in each of the three datasets with the application of a $\log_2$ transform.

The data therefore provides no *a priori* reason to log transform MAS 5.0 data before analysis. Considering RMA, there is no consistent improvement in FDR results using either transformed or untransformed datasets and gcRMA performs marginally better using the log transformed data. There is thus no reason to diverge from the author's methodologies (Irizarry, et al., 2003; Wu, et al., 2004) which employ a logarithmic transform as the last stage in the analysis process.

3.3.3 How does pooling variance influence detection outcomes?

Referring to guidance about the application of the t-test to data we are faced with the option of two variants of the test, the standard unpaired t-test or the Welch variant of the t-test (Welch, 1947). The difference between the two is a question on whether to assume equality of variance, which describes the spread of the data.. The standard t-test offers additional power compared to the Welch t-test, gained by a better estimate of variance because the data can be pooled.

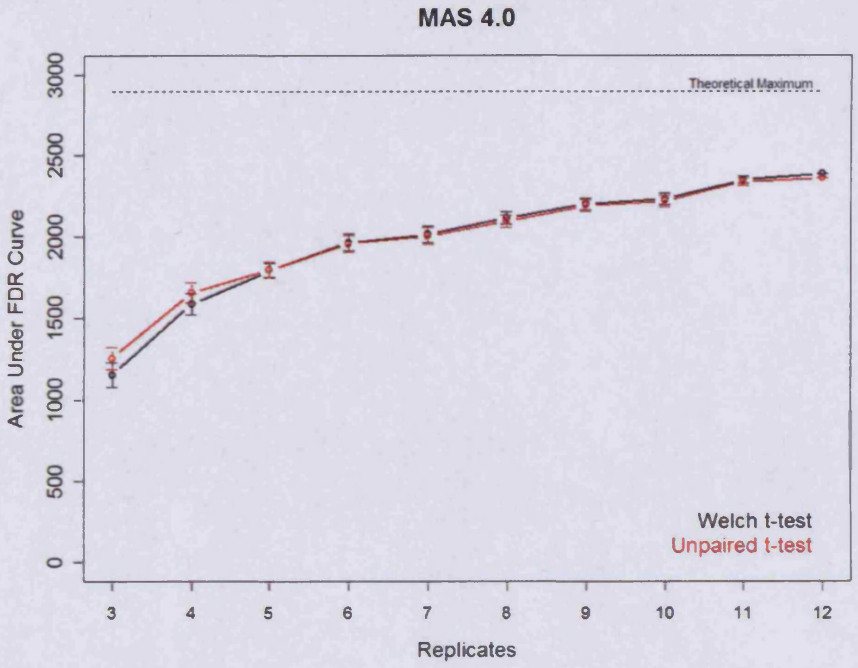
Published data (Baldi and Long, 2001) has indicated that in Affymetrix microarray data there is a relationship between variance and mean expression level; as the signal level increases for a probe set, so does the variance. In applying a test for differential gene expression a researcher will be interested in changes from a low to high (or vice versa) situation, and thus the test will be faced with groups with different measured levels of variance.

The natural conclusion to this observation is to apply the Welch variant to the t-test to overcome this assumption. However as it has already been shown that due to small sample sizes, there is a need to maximise all power available from the dataset when applying statistical testing, the question of whether pooling variance can improve power to detect differentially regulated genes (despite the apparent deviation in assumption) is of interest.

Using the already described framework (Figure 3.1) , both the unpaired t-test and Welch t-test was applied to the 24 chips with a two-fold change in the Affymetrix Latin square dataset and detection of the fifteen spikes assessed using FDR curves over a range of sample sizes with multiple samples. MAS and dChip was used in an untransformed form, RMA data was analysed as provided after analysis (i.e. in a $\log_2$ form). The resultant summary graphs showing the area under the FDR curve over a range of sample sizes is shown in Figure 3.6.

Figure 3.6

a)


MAS 4.0
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
Welch t-test
Unpaired t-test
3
4
5
6
7
8
9
10
11
12
Replicates

b)

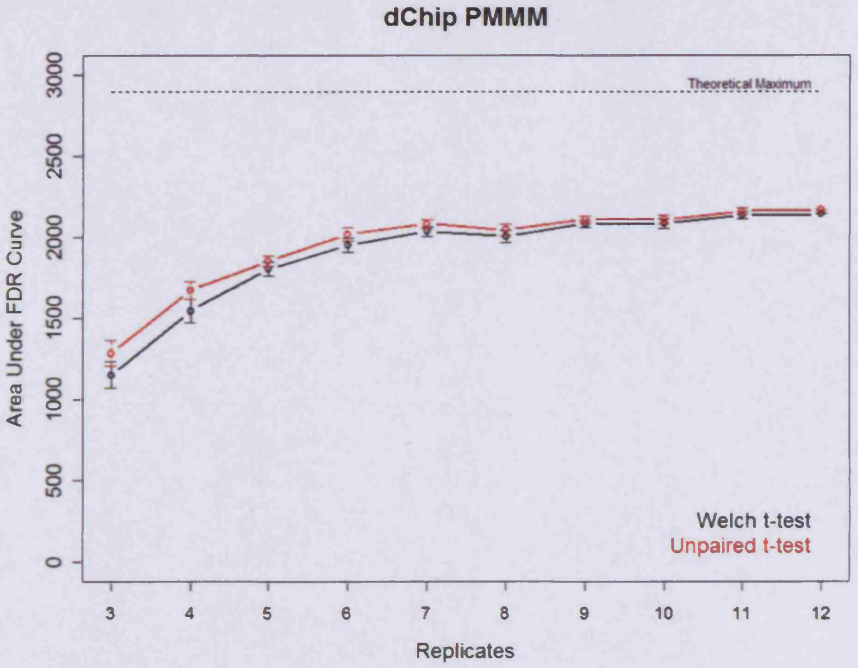
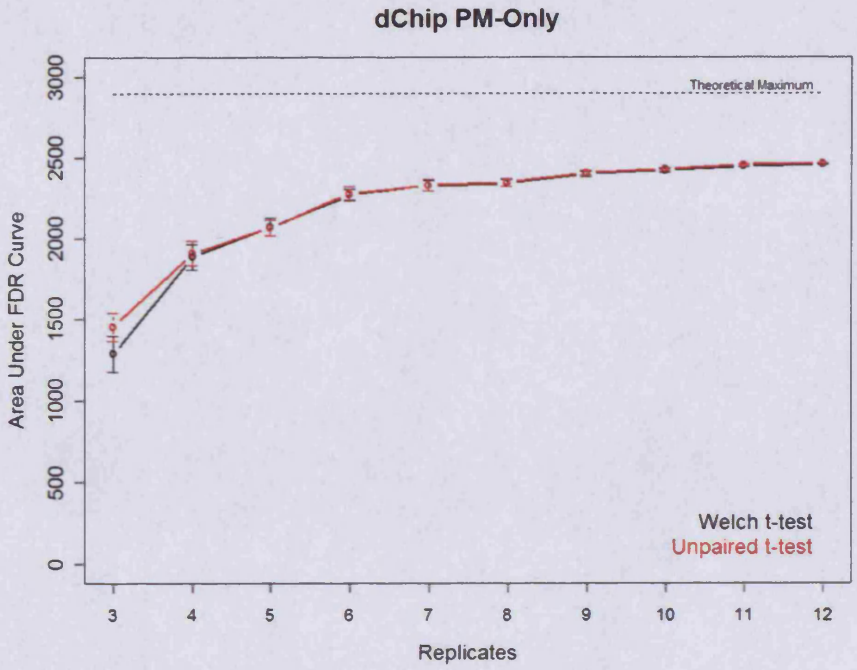

dChip PMMM
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
Welch t-test
Unpaired t-test
3
4
5
6
7
8
9
10
11
12
Replicates

Figure 3.6 – continued

c)



d)

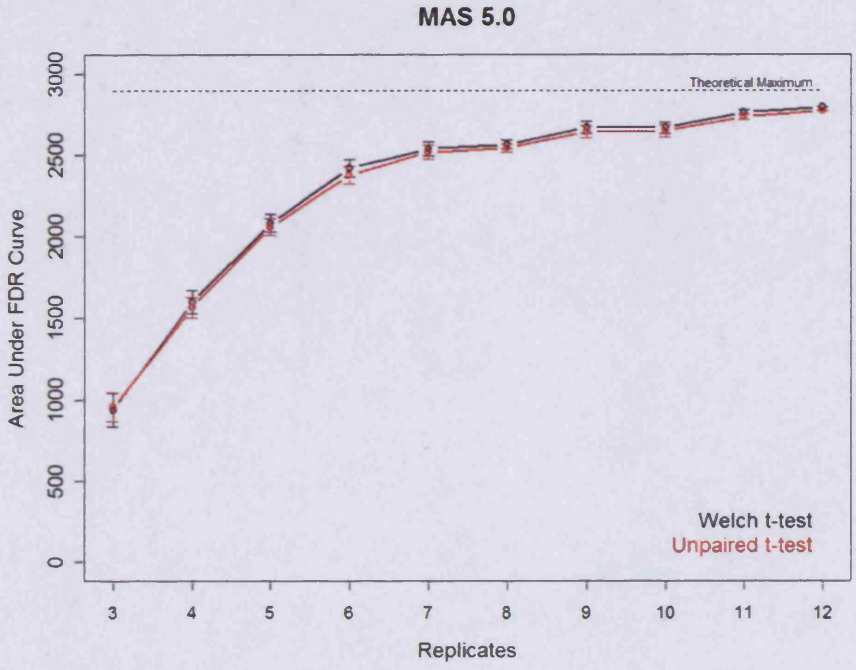
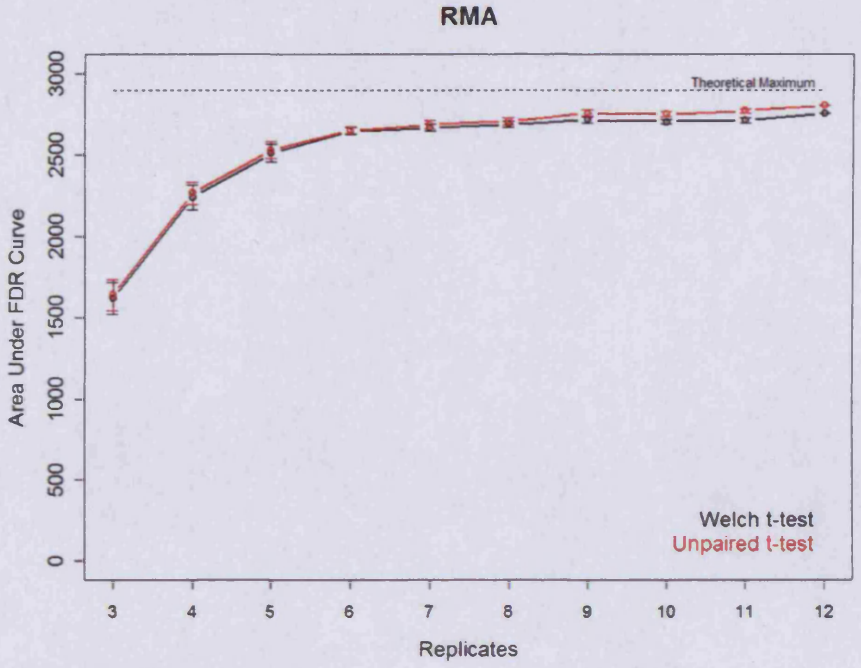


Figure 3.6 – continued

e)



f)

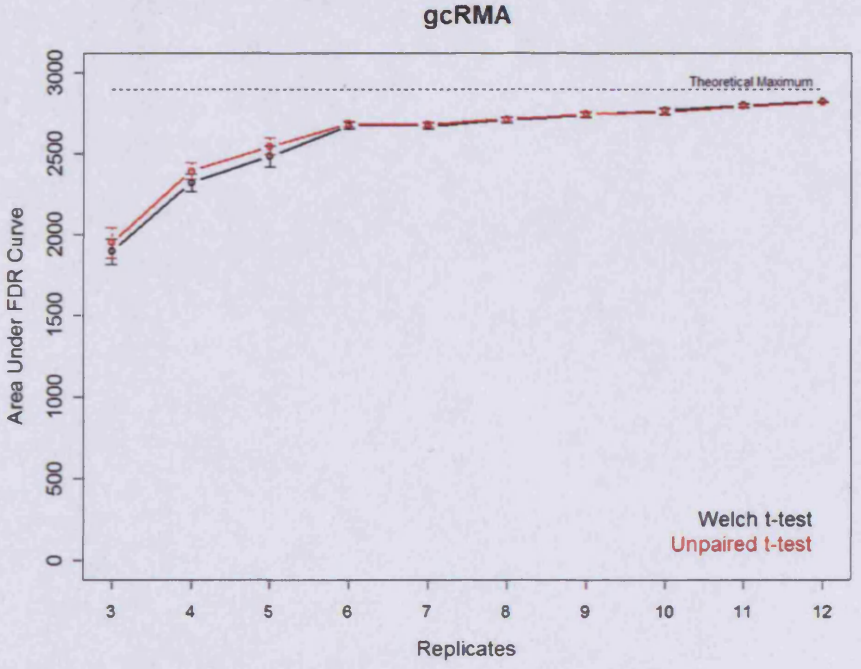


Figure 3.6 – Summary AUC plots for each of the six expression metrics, comparing the power to detect between the unpaired t-test (which pools variance) and the Welch variant of the t-test which calculates group specific variance.

Looking at the results from the two different t-tests, we can see that there is very little difference in performance (Figure 3.6). It could be argued that there is a small advantage in using the unpaired t-test on small sample sizes from dChip PMMM data and a similar advantage in large sample sizes on RMA data; however, the magnitude of improvement is small. The experimental data would appear to support the theoretical observations regarding variance and support the application of the Welch t-test to Affymetrix datasets.

So far data regarding how sample size affects detection outcome has been overlooked. Having settled on the application of the Welch t-test on the data resulting from each of the expression metrics (untransformed from MAS and dChip, $\log_2$ transformed from both RMA models) as best practice, observations about the power to detect across the range of sample sizes can be reviewed.

Marked improvements are seen across all expression metrics over sample sizes per group of three to seven; above this point comparatively small increases in sensitivity are obtained for each addition chip run. In the RMA models this critical size is reduced with a near consistent plateau of detection once at least six samples per group are analysed.

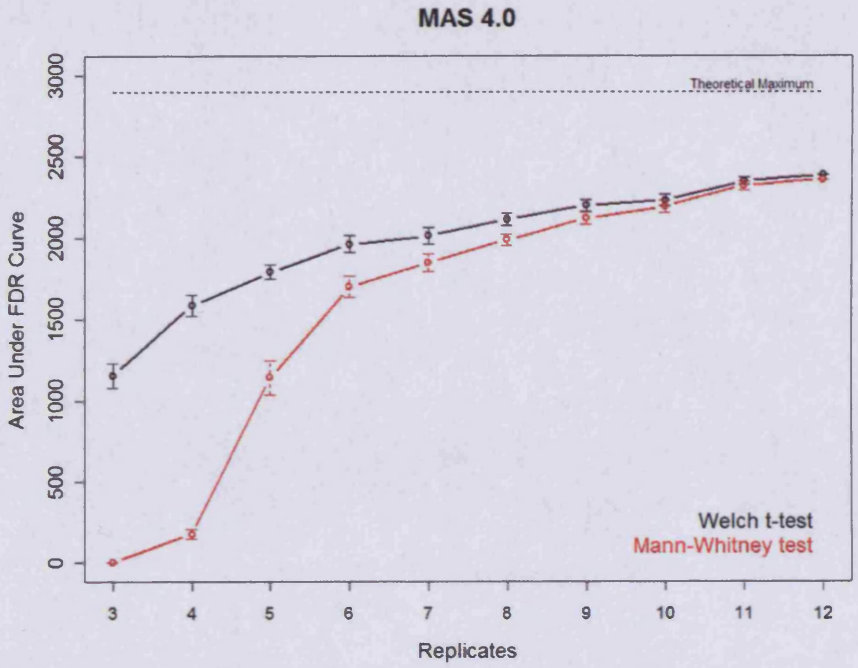
3.3.4 How applicable are non-parametric methods to microarray datasets?

By using the relative ranks of the data, non-parametric tests for differential expression are less sensitive to data distributions and outliers. However, the trade off for this improvement in resilience is a loss in power compared to the parametric test for the same size sample, including the observation that with less than seven samples it is impossible to obtain a p-value less than 0.05 (the typically applied significance level) with the Mann-Whitney test (Motulsky, 1999).

There are, however, situations where even with good experimental design the extra protection against outliers and other biological artefacts in the resultant data may lead a researcher to consider the application of a non-parametric test (e.g. the heterogeneous sample types found in many cancer datasets). Information about the comparative power of non-parametric versus parametric testing, along with data indicating at which sample sizes the power of the two groups of test converge, is thus important. The Mann-Whitney test was applied to the Latin Square dataset across each of the six expression metrics under consideration as a distribution free alternative to the t-test and the data analysed using the FDR method. For comparison, summary graphs of the Mann-Whitney data plotted against the Welch t-test data are shown in Figure 3.7.

Figure 3.7

a)


MAS 4.0
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
Welch t-test
Mann-Whitney test

b)

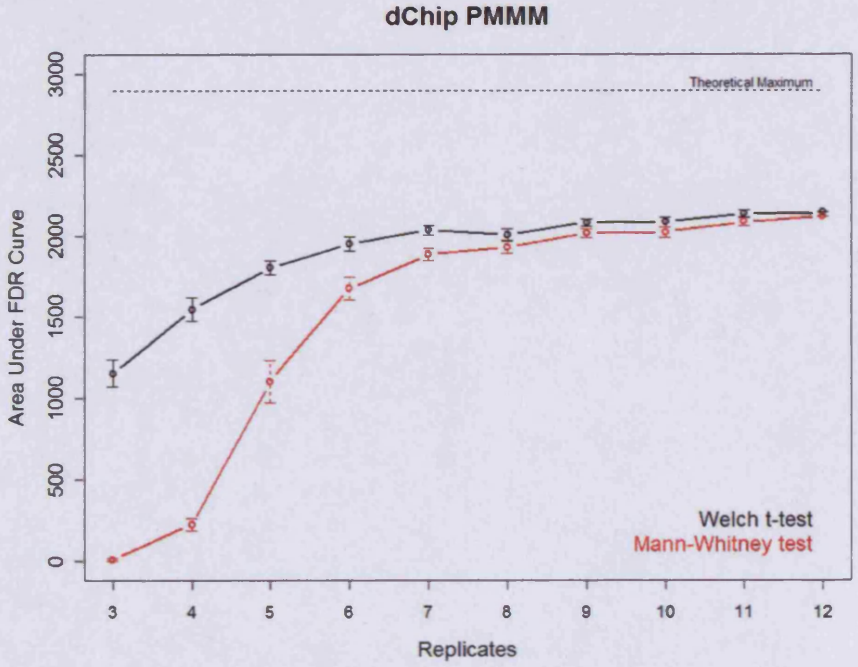
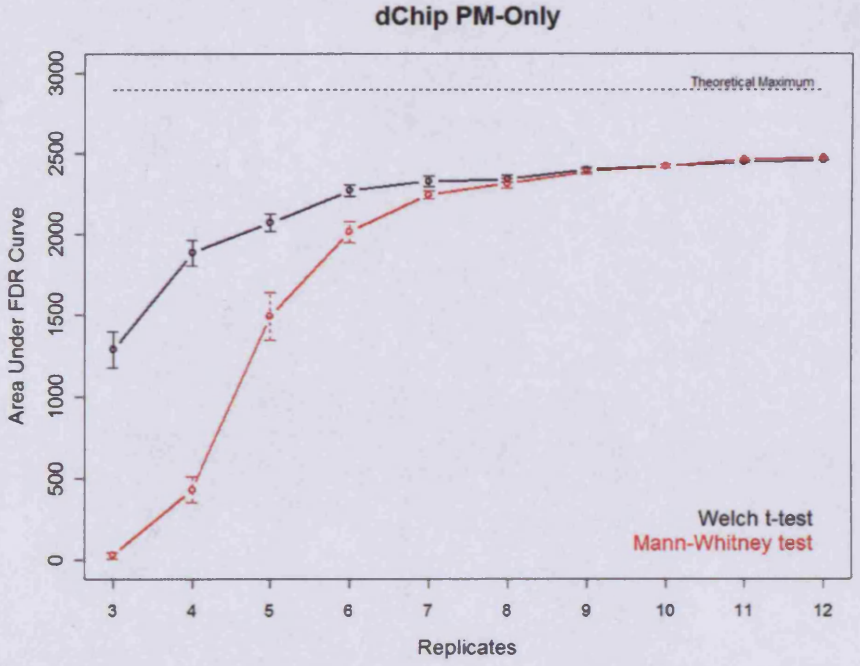

dChip PMMM
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
Welch t-test
Mann-Whitney test

Figure 3.7 - continued

c)



d)

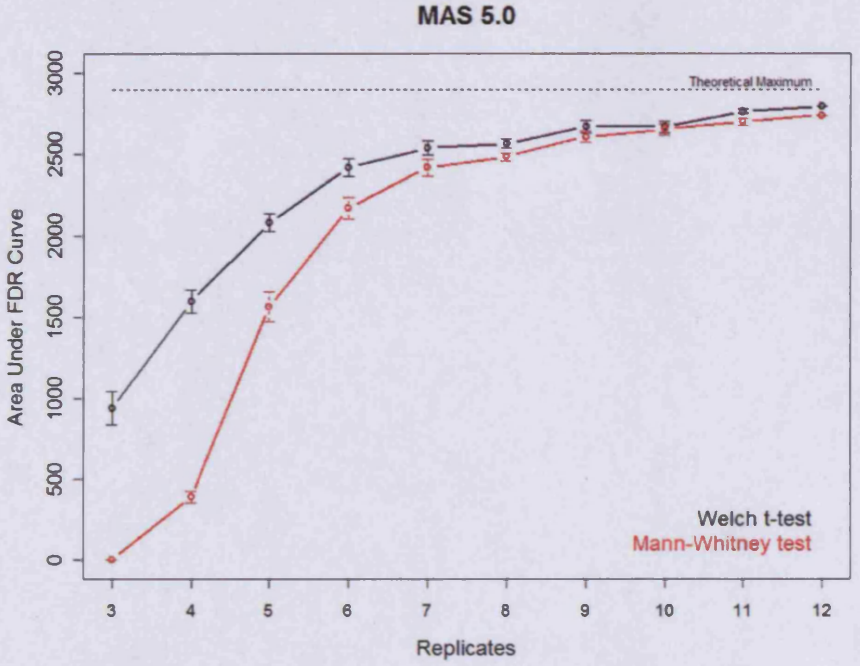
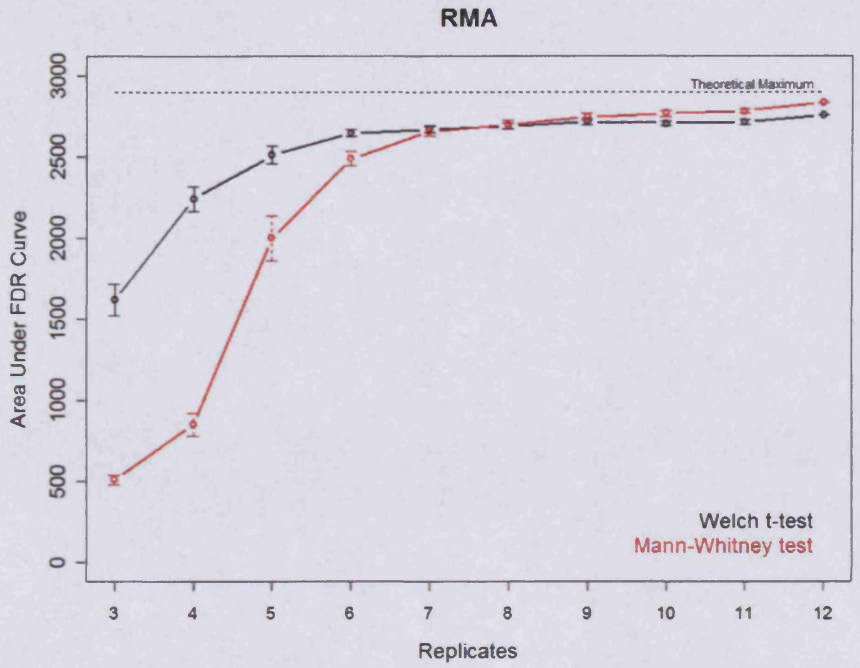


Figure 3.7 - continued

e)



f)

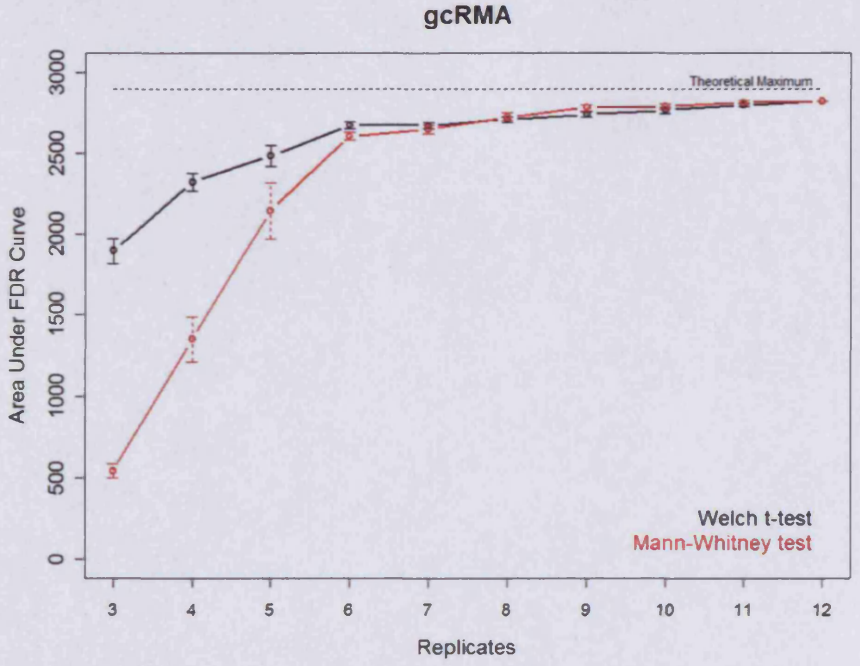


Figure 3.7 – Summary AUC plots for each of the six expression metrics, comparing the power to detect between the Welch variant of the t-test and its equivalent non-parametric alternative, the Mann-Whitney test.

Reviewing the summary FDR plots (Figure 3.7), it can be seen that the Mann-Whitney test performs particularly poor at smaller sample sizes, as would be expected. This effect is a combination of the lower power of the non-parametric test combined with the test being unable to stratify its results, with many probe sets being assigned an identical p-value.

Whilst at low sample size the test is practically useless, once six samples per group are run the test is closing the gap on the t-test with only a small loss in detection sensitivity and by the time the sample size is eight or more the results are nearly identical. Practically a researcher may wish to consider the non-parametric test what is believed to be a highly variable experimental design once six samples per group are being run.

For larger datasets the power on this very heterogeneous dataset are comparable to that from the parametric t-test. The additional resistance of the Mann-Whitney test to deviations from normality along with resilience to outliers beckon the suggestion that this be the standard test applied to datasets with more than 10 samples per group.

3.4 Discussion

3.4.1 Reviewing fold change

In all the expression metrics under comparison, fold-change was out-performed by statistical testing by a large margin across all sample sizes. In models incorporating mismatch probe information there was no power to detect the spiked-in probe sets in the dataset being examined, with only a small amount of power in those which utilise only the perfect match probes. This improvement in scoring from the PM-Only models is likely to be associated with an increase variability of low-expressed probe sets introduced by incorporation of MM-probe data in the expression metric algorithms.

Examination of the formulae behind fold-change and the t-test reveal that thought of in a simplistic manner, the t-test is an evolved implementation of fold change taking into account variance which produces the t-statistic. Integration of this variance information improves the power to detect changes in a dataset.

There is however, one situation where a research maybe forced to consider the use of fold-change - where insufficient chips have been run to apply statistical testing. Irizarry et al. (Irizarry, et al., 2003) have shown that the Affymetrix Signal Log ratio, which is computed using a Tukey-Biweight method on the probe pair intensities, is an efficient method of detecting changes when only two chips are to be compared – this method however is not expandable to experiments with more than two chips. Mariani et al. (Mariani, et al., 2003) suggest the application of a variable fold change threshold resulting from application of a lowess curve to the standard deviation information. These approaches whilst not examined for power to detect, do provide potential analysis methods for datasets with sample sizes too small for application of parametric testing.

3.4.2 Issues of sample size

Sample size is a major issue for a researcher designing a microarray experiment, since from a statistical point of view, the more replicates run the more confidence can be determined in the results. However, this comes at the trade off of cost; and these costs are currently not insignificant. The biologist designing an experiment is interested in the trade off between accuracy of results versus the cost to get those results.

The other factor to consider in this trade off is what the results actually mean. Whilst a statistician is interested in the exact meaning of a 95% confidence level and designing an analysis to reflect the exact nature of truth in a dataset, the biologist is typically using a microarray experiment as a hypothesis generation exercise, which will subsequently be followed by validatory techniques. A biologist must therefore be convinced that each additional replicate is worth running in terms of reinforcing confidence in the results.

Reviewing the data presented in this Chapter it is clear that the more samples run, the more power each test has. However summarising, it can be seen that over sample sizes of 3-6 per group large increases in power are seen for each additional chip run, between 7-10 samples per group there is continued improvement in sensitivity, and over 10 chips each additional chip adds little to the overall detection of spikes within the dataset.

Guidelines on sample size are difficult to deduce because of the underlying variability in biological data and experimental design which will affect the power of a test to detect differences. It should also be pointed out that samples do fail, and there are QC issues which may require the dropping of certain chips. These should be viewed as minimum estimates because of the clean and perfect nature of this dataset, and the likely influence of additional biological variation would require additional GeneChips in an experimental design.

3.4.3 Applying statistical testing to microarrays

In reviewing the data explored in this Chapter, there is the requirement to reduce answers to help answer the simple question *"Which test should I use?"*

As with all things surrounding the analysis of microarray data there is no simple answer. However, it is clear that the choice of test is dependent on the nature of the experiment and more critically the number of samples in each group. To review this three real-world experimental situations are presented, along with observations about the suitability of different tests.

Scenario 1: A small-scale experiment involving three to six replicates per group. In this experimental configuration, the t-test is the clear winner with much more power than fold change and non-parametric methods. However the test assumes data normality and is sensitive to outlying data points. Care would therefore be needed to ensure that noisy data (such as a comparison of tumour to normal tissue) are avoided.

Scenario 2: A middle-sized experiment comprising seven to ten arrays in each group. With this sample size the t-test continues to gain power for each additional GeneChip run per group. However, the non-parametric test is also beginning to perform well with only a marginal decrease in power over its parametric counterpart. Choice of test at this sample size must therefore be influenced by the type of sample being run. For homogeneous experimental designs (e.g. drug treatment of culture cells) the Welch variant of the t-test would be the choice of test. However, in a heterogeneous design (e.g. clinical samples) the additional confidence the non-parametric test provides would suggest adoption of the Mann-Whitney test.

Scenario 3: A large experiment with more than ten samples in each group. On the homogeneous dataset under consideration here, little additional power was achieved by addition of replicates. However, once this size of experiment has been reached, it must be concluded that the data is likely to be heterogeneous with an increased likelihood of deviations from normality and the presence of outlying data points. Here the Mann-Whitney test come into its own, with power is almost as good as parametric tests but with its resilience against deviations in data distribution and outliers.

3.4.4 Requirements and benefits of logarithmic transformation

In Chapter Two it was shown that from a distributional point of view there was little or no benefit to the logarithmic transformation of data from any of the expression metrics. It was noted that whilst the $\log_2$ transform included in RMA and gcRMA did not particularly benefit the correlation with normality, there was little to dispute the author's inclusion of this measure in their methods (Irizarry, et al., 2003; Wu, et al., 2004).

The data presented in this Chapter reveals no difference in the ability to correctly identify spikes in the dataset between logarithmically transformed and untransformed data with the exception of MAS 4.0 and dChip PMMM models. It was however noted that the negative numbers present in these datasets causes problems to the transformation and thus the results must be treated with caution.

These data provide no *a priori* reason for transforming data prior to statistical analysis. Detection in data from RMA and gcRMA are not altered by removal of the log transformation, so no reason is found to alter from the published methodology.

3.4.5 Comparative review of expression metrics

By grouping the data from each test together by expression metric the previously generated data can be reviewed in another way to provide information about the relative performance of different expression metrics. Figures 3.8, 3.9 and 3.10 show data for each expression metric grouped by fold change, the Welch t-test and the Mann-Whitney test.

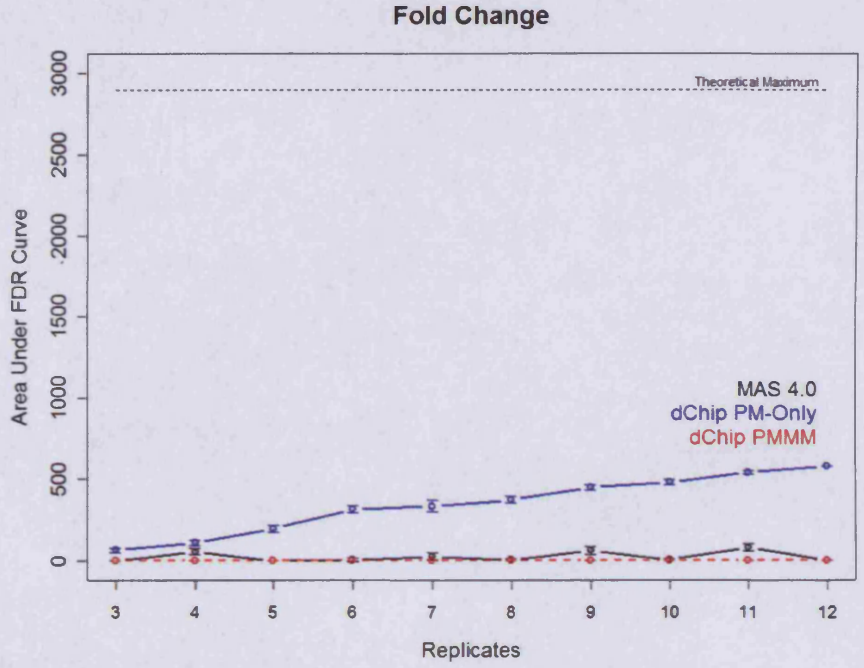
From these observations we can conclude that MAS 5.0, RMA and gcRMA would appear to be the best options for the conversion of image data into a workable expression summary. Both dChip models make attempts at improving the accuracy of the expression summary by using a model to assess standard errors for the expression metrics by pooling information across arrays. However, even in this highly homogeneous dataset, the technique does not perform as well as the other expression metrics and has a reduced ability to detect spiked transcripts using standard statistical methods.

It should be noted that the dataset used in these investigations formed part of the Affymetrix development process for the algorithms comprising MAS 5.0, and it could be the case that the default tuneable parameters in the methodology were influenced by the nature of this dataset. In addition the same dataset was key in the development of RMA and gcRMA.

However, in the absence of further datasets large enough and with sufficient replication to assess this, we must therefore conclude that there is no obvious reason to avoid these expression metrics for data analysis.

Figure 3.8

a)



b)

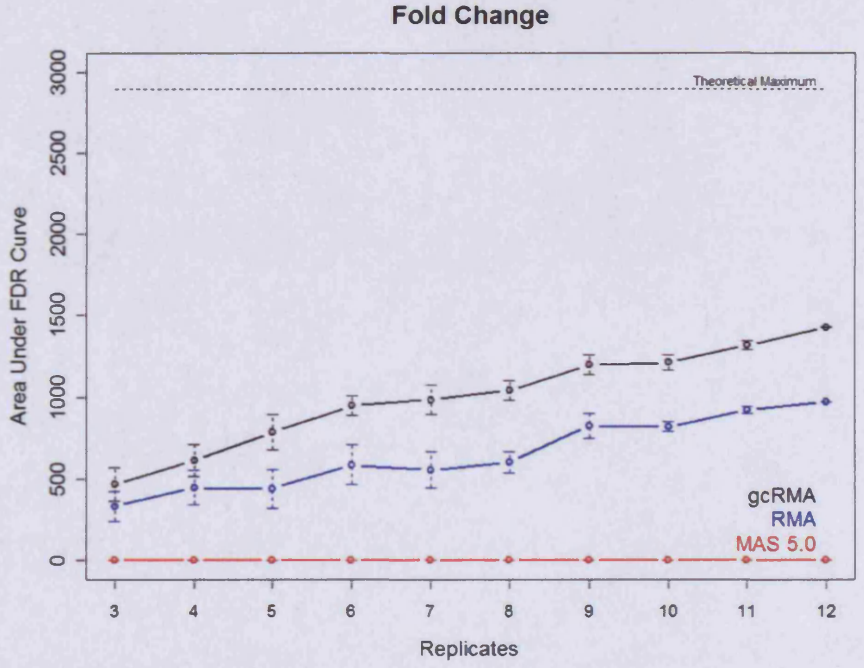
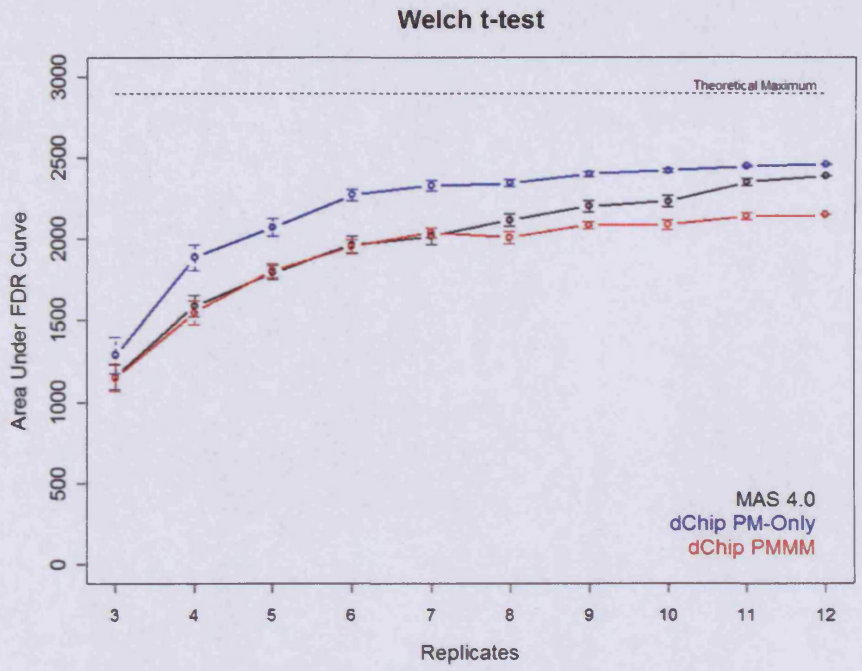


Figure 3.8 - Summary AUC plot comparing the power of each expression metric applied in combination with a fold-change test to detect the fifteen spiked-in data points of the Affymetrix Latin Square dataset.

Figure 3.9

a)



b)

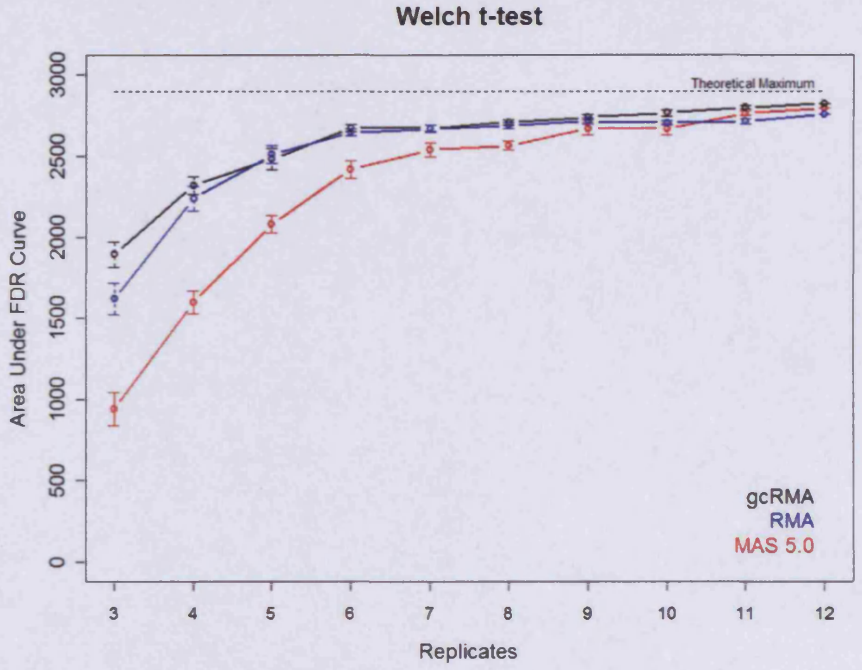
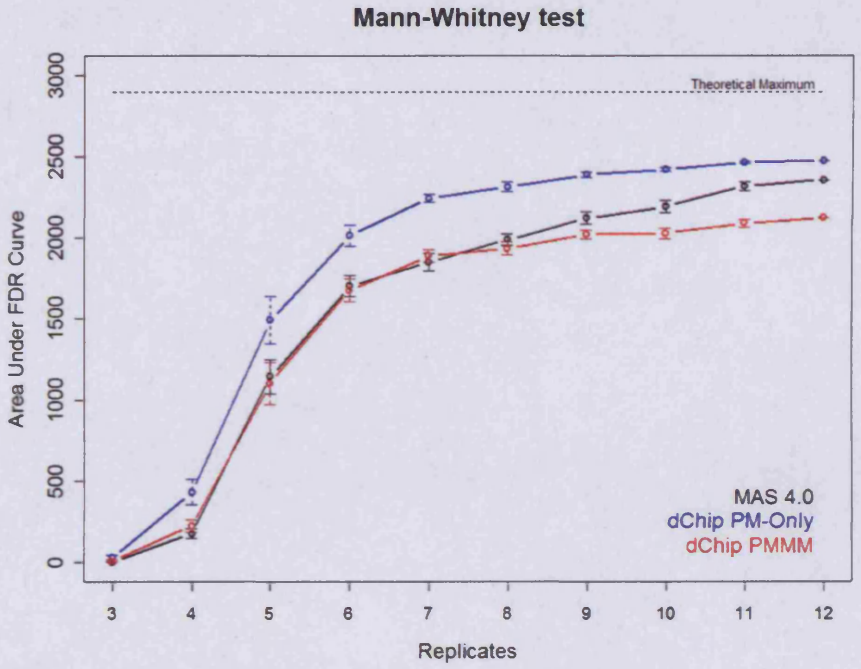


Figure 3.9 - Summary AUC plot comparing the power of each expression metric applied in combination with the Welch t-test to detect the fifteen spiked-in data points of the Affymetrix Latin Square dataset.

Figure 3.10

a)



b)

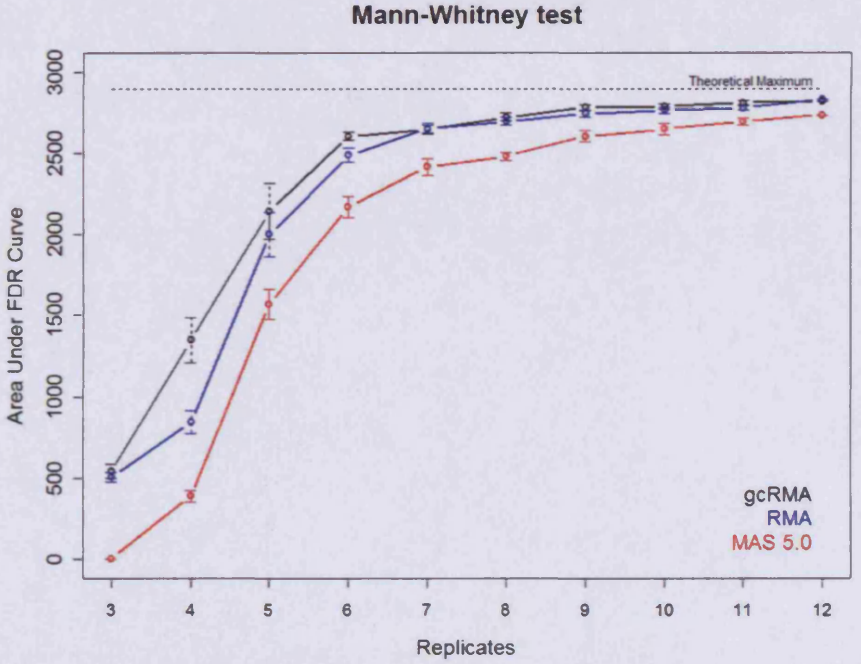


Figure 3.10 - Summary AUC plot comparing the power of each expression metric applied in combination with the Mann-Whitney test to detect the fifteen spiked-in data points of the Affymetrix Latin Square dataset.

Chapter Four
Approaches to data normalisation

In this Chapter the issue of data normalisation is explored. Section 4.1 introduces the concepts of normalisation and the application of normalisation within Affymetrix GeneChip data analysis. Section 4.2 reviews the methods used to investigate the effects of normalisation on experimental output. Section 4.3 explores the effects of QQ, VSN and a rank based normalisation method on the ability to detect the spikes in the U95A Latin Square dataset. Section 4.4 discusses the results and observations regarding the application of normalisation within an analysis.

4.1 Introduction

4.1.1 What is normalisation?

The Oxford dictionary defines *normalise* as *"verb: bring to a normal or standard state"*. Applied to microarray data, normalisation is an attempt to remove the impact of non-biological influences on biological data by correction of systematic biases. For example, this is conceptually similar to the adjustment of expression levels relative to a reference gene in Northern blot analysis. This variance in the data can occur for a variety of reasons including biological differences between samples, unequal quantities of starting RNA, differences in labelling and systematic biases in the measured expression levels (Quackenbush, 2002).

The common assumption of most normalisation methods is that the overall distribution of signal levels is similar between samples, and that the expression of the majority of genes changes little between the conditions under examination (Reimers, 2003). Whilst this would appear reasonable for the majority of laboratory treatments, clinical tumours are one example of a sample type that might present with very differing expression patterns between samples (Golub, et al., 1999).

4.1.2 Why is there a requirement to normalise data?

Variation in results arises naturally when dealing with results from multiple experiments. Many researchers have categorised this variation as being one of two types, interesting variation and obscuring variation (Bolstad, et al., 2003; Landis, et al., 2004). The biological difference that are being looked for in the course of microarray experiments are termed an interesting variation, whereas obscuring variation is used to describe the variation introduced as part of the experimental process, for example differences in sample preparation and labelling, differences resulting from array

production. and those introduced as part of the scanning process. Normalisation aims to overcome this obscuring variance and leave the researcher with a dataset representing a truer picture of the underlying differences between arrays.

As previously commented, most normalisation methods require the number of genes changing in expression between conditions to be small, or assume an equal number of increasing and decreasing genes. Zien et al. (Zien, et al., 2001) comment that when neither of these conditions are true then intra group normalisation should be applied instead of a global normalisation. However, in many cases poor experimental design and processing results in groups with distinct non-biological variation, and for this reason many researchers prefer to normalise all data together as a single group.

Applied to Affymetrix data, normalisation is applied in two distinct places, prior to the combination of probe signals into a probe set expression level, and/or after this incorporation.

4.1.3 Normalisation methodologies within expression metrics

The primary reason for the application of normalisation as part of expression meric analysis is to address variation in overall array performance due to issues of inconsistencies in array fabrication, subtle differences in experimental processing and differences occurring at the staining and scanning stage. As these sources of variation can contribute significantly to the data output from an array experiment, normalization is a critical first step in any analysis of gene expression data.

Normalisation can be applied to a single chip in order to shape the resultant data into an expected form, be applied to shift the data from many chips into that from a chosen baseline experiment, or use the data from all arrays in the experiment to determine a baseline to normalise to.

Historically, many normalisation procedures for microarrays arrays are global approaches, based on normalization of the overall mean or median array intensity to a common standard (Affymetrix, 2002b) and are linear in their transformation approach. Approaches using non-linear curves have been suggested (Li and Hung Wong, 2001a; Li and Wong, 2001b; Schadt, et al., 2001), however these approaches require the choice of a baseline value or array for normalisation to.

However this approach has its limitation when dealing with data with non-linear relationships, Bolstad et al. (Bolstad, et al., 2003) comment that it is common to see non-linear relations between arrays and the standard normalization provided by Affymetrix does not perform well in these situations.

In contrast a normalisation is an integral part of the three-step RMA and gcRMA methodologies, using a non-linear normalisation at the probe level. Typically the method employs a QQ normalisation method, although options are provided for the incorporation of different normalisation algorithms (Bolstad, et al., 2003).

Hartman et al. (Hartmann, et al., 2003) compared MAS 5.0 to RMA methods using VSN and QQ normalisation applied at the probe level, in their ability to detect a two-fold change of the spikes present in the Latin square dataset using two sample groups of four GeneChips. They conclude that RMA data with either QQ or VSN normalisation produce data where *"we are able to detect more of the spike in genes while getting less false positives"*. Wang et al. (2004) found the RMA normalization preferable to that within MAS 5.0 in analysis of inter-species conserved (ISC) probe sets

4.1.4 Post metric analysis normalisation of Affymetrix data

Post metric analysis normalisation can be applied in an additional attempt to overcome variation between arrays which are expected to produce near identical expression patterns. Whilst the RMA and gcRMA methods incorporate a cross-array normalisation as part of their methodology, the MAS algorithm treats each array as a separate entity during analysis. This has both advantages and disadvantages, since on one hand inter-array normalisation produces data which is less likely to need further manipulation before analysis; however it does produce data which is dependant on the other data in the experiment.

The approach taken by Affymetrix to normalisation within Microarray Suite is the scaling of data so that each array has the same average value (either a pre-defined value termed target intensity or a value determined from a baseline array). A shifting and scaling approach is applied after the statistical stages used to convert the image information to a numerical signal values for each probe set. As this is the final step in the analysis process and data can thus be re-scaled if required. All the data presented in this work analysed using MAS 5.0 have used a target scaling intensity of 100.

Post metric normalisation can vary from simple techniques which undertake additional shifting and scaling of the data to make it comparable to more complex ones which act in a non-linear fashion and attempt to address more complex measurement of data similarity such as the distributions of signal values.

An alternative approach to the global normalisation approaches which use all the available data to determine the normalisation baseline, normalisation can also be applied to shift and scale the data according to the values of a set of know spiked-in transcripts, added as part of the experimental process. Affymetrix have made provision for this form of normalisation with the addition probe sets which correspond to Bacillus genes that have been modified by the addition of poly(A)+ tails, then cloned into pBluescript. These probe sets can thus be used as controls for cDNA synthesis and subsequent sample preparation steps.

With the new generation of the U133 series of chips Affymetrix introduced support for a different type of normalisation, using a set of housekeeping genes, similar to that used in Northern blots. By using a set of 100 genes, the chips are re-scaled so that the averages of these housekeeping genes are identical across all chips.

In an investigation of median-interquartile range normalisation and quantile normalisation methods, Parrish and Spencer (Parrish and Spencer, 2004) found a substantial inflation of the number of genes identified by paired-t significance tests when compared to the non-normalised data. This observation was attributed to the power of the normalisation technique to overcome experimental variance. In their prostate cancer dataset the normalisation had a greater effect on RMA data that that from MAS 5.0.

4.1.5 Questions of normalisation

Although each of the expression metrics incorporates some form of normalisation as part of their methodology, it is of interest to examine the effect of common post-analysis normalisation steps on the data in an attempt to further overcome obscuring variance in the dataset and the effect of these steps on the identification of spiked-in probe sets in a dataset with known truth.

4.2 Technical Methodology

Data from MAS 5.0, RMA and dChip was examined for normality using the Shapiro-Wilks test and Q-Q correlation plots (as described in Chapter Two) following the application of QQ normalisation (Bolstad, et al., 2003), VSN normalisation (Huber, et al., 2002) and a rank based technique (Breitling, et al., 2004) using the full Affymetrix U95A Latin Square dataset with the spiked in samples and control probe sets removed (see Section 9.3.2)

The previously described integrated framework (Section 3.2), which took a subset of the Affymetrix U95A Latin Square dataset and produced a series of 20 replicates at a range of 3 to 12 chips in each of two groups using the MAS 5.0, RMA and gcRMA expression metrics, was modified to incorporate post-metric analysis normalisation (Figure 4.1).

Figure 4.1

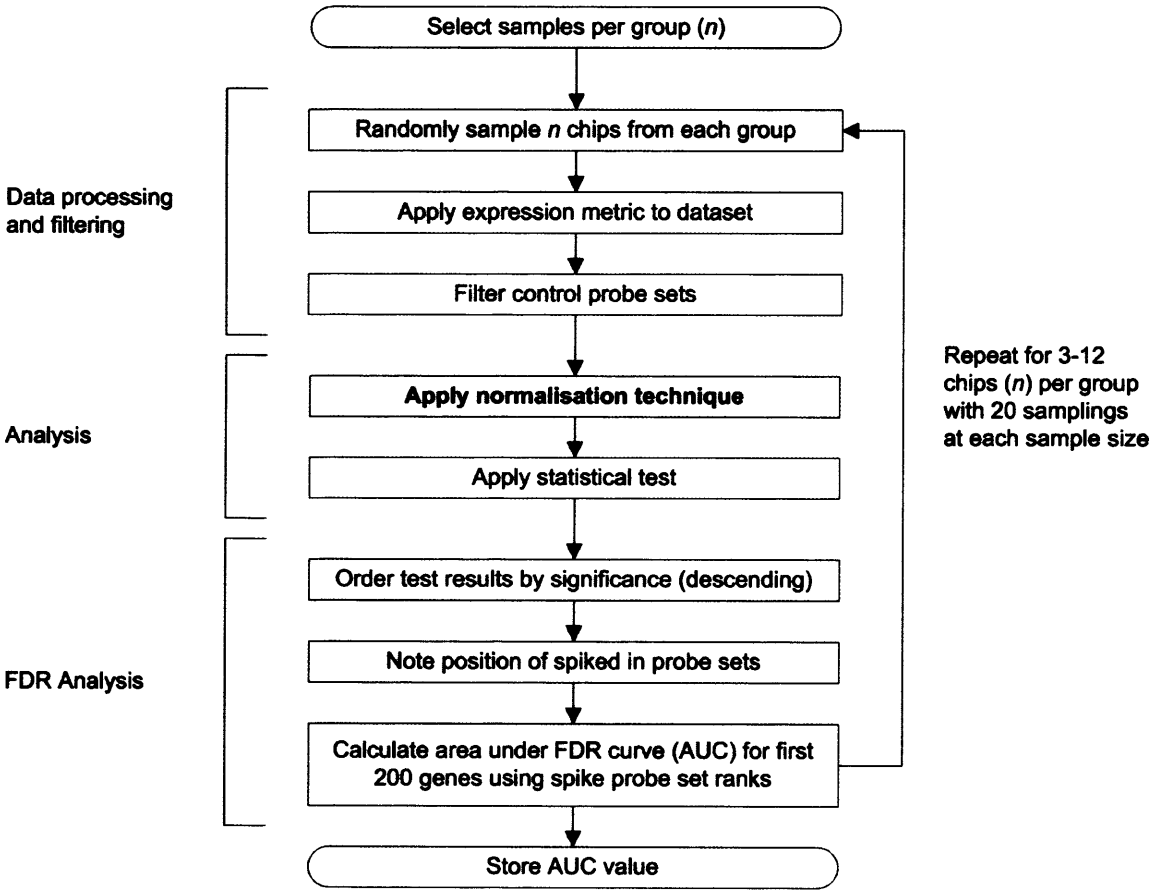


Figure 4.1 – Analysis flowchart

The resultant data from each expression metric was subject to normalisation using QQ normalisation, VSN and a rank based technique before the application of statistical tests for differential genes expression. The script undertook twenty samplings over a range of sample sizes of three to twelve arrays per group and subjected the normalised data to analysis using the Welch t-test. In the case of the rank adjusted data the data was analysed using the non-parametric Mann-Whitney test in addition to the Welch t-test.

The output for each test was fed into an FDR function which extracted the location of the spikes from the ordered p-values and calculated the area under the FDR curve as previously described. The output for each test was exported into a summary matrix containing each of the 200 AUC values calculated, grouped by sample size. Summary plots were produced charting the average area under the curve (with standard error, error bars), across the range of sample size under consideration.

Full details of the technical methodology are given in Section 9.4.1.

4.3 Exploration and Results

4.3.1 Application of Quantile-Quantile (QQ) normalisation

In Chapter Three the Quantile-Quantile (Q-Q) plot was introduced as a method of assessing correlation to normality. QQ normalisation extends this idea with the goal of making the frequencies and distribution of probe intensities for each array in a set of arrays the same. In the Q-Q plot, the distribution of two data vectors is the same if the plot is a straight diagonal line. In QQ normalisation, this concept is extended to n dimensions so that if all n data vectors have the same distribution, then plotting the quantiles in n dimensions gives a straight line along the line given by the unit vector (Bolstad, et al., 2003).

The method works by giving each array the same distribution, which is achieved by taking the mean of each quantile and substituting it as the value for each array value comprising the quantile data.

Practically, the data from each chip is sorted, independently from one another, with an index stored to keep their original order. From this new sorted matrix, the median value of each probe is calculated and used to replace the original value in each chip on that row. The matrix is then 'unsorted' using the original index, giving the normalised data.

The caveat for the use of a single standard for all chips is the assumption that there is no major change in distributions between chips. Whilst this appears a strong assumption about gene distributions, Reimers (Reimers, 2003) comments that in practice, the expression levels of genes move up and down roughly equally and it would need several hundred genes to be changed greatly and in one direction, to drive quantile normalization in error by more than 20%.

In the RMA and gcRMA expression metrics, QQ normalisation is applied at the probe level before summarisation of the probe values into a single signal level. However, it is also possible to apply QQ normalisation to the resultant expression metric data and assess the effect of the normalisation on the outcomes from statistical testing looking for the spiked in samples in the U95A Latin Square dataset.

As the purpose of the QQ normalisation is to synchronise the data distributions between chips there is the possibility that the method alters the data distributions across a single probe set. As has previously been discussed (Section 2.1.4) the data distribution of a single probe set is important if parametric testing is to be applied. It is therefore important to first analyse the effect of probe set distributions following QQ normalisation.

4.3.1.1 Normality of QQ normalised data

Using the strategies introduced in Section 3.3.1, the normality of data following the application of a QQ normalisation was first assessed using the Shapiro-Wilks test for normality. The number of non-normal genes (scoring a p-value less than 0.05) for each of the three datasets is shown in Table 4.1.

Table 4.1

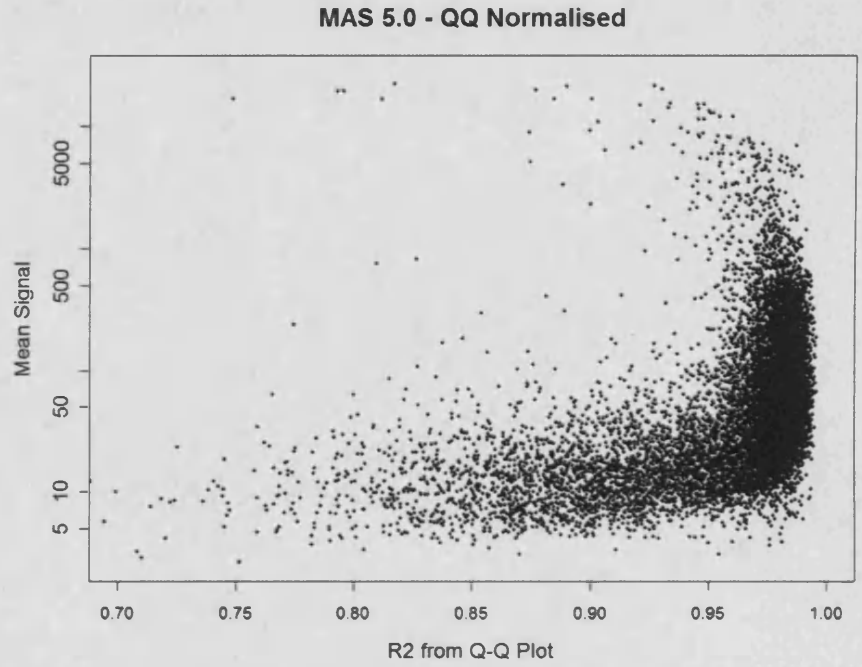
Analysis Method	Number of probe sets deviating from normality ($p<0.05$ from SW test)	
	Untransformed	QQ Normalised
MAS 5.0	5799 (46%)	5977 (48%)
RMA	3075 (25%)	3502 (28%)
gcRMA	6371 (51%)	9509 (76%)

Table 4.1 – Results from Shapiro-Wilks tests for normality.

The Shapiro-Wilks test data show that there is a slight increase in the non-normality of the data from each expression metric after the application of QQ normalisation. This observation highlighted the need for further exploration of the nature of this deviation from normality using plots showing mean signal plotted against the correlation to normality for a Q-Q plot (Section 2.3.2). The resultant plots are shown in Figure 4.2.

Figure 4.2

a)


MAS 5.0 - QQ Normalised
Mean Signal
5
10
50
500
5000
0.70
0.75
0.80
0.85
0.90
0.95
1.00
R2 from Q-Q Plot

b)

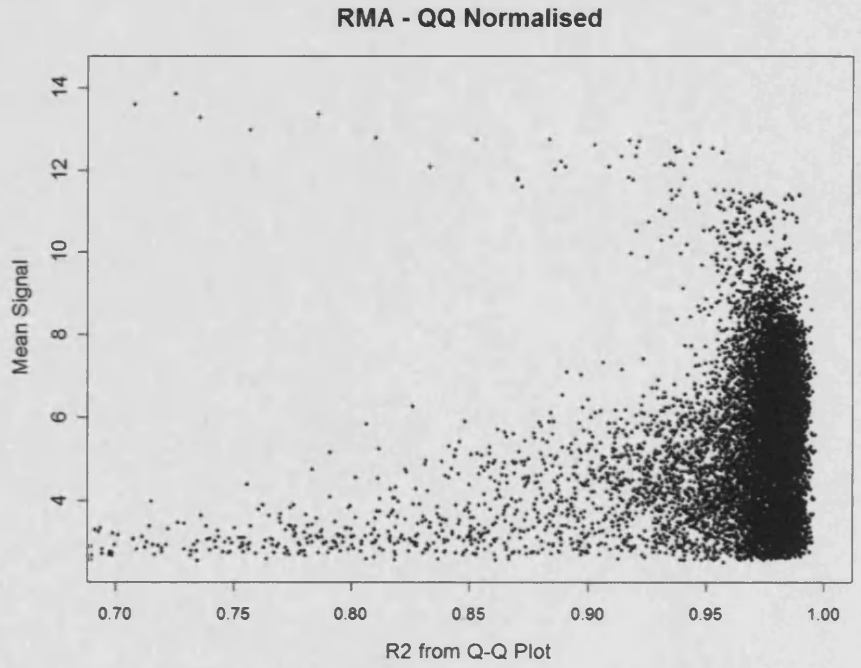

RMA - QQ Normalised
Mean Signal
4
6
8
10
12
14
0.70
0.75
0.80
0.85
0.90
0.95
1.00
R2 from Q-Q Plot

Figure 4.2 – continued

c)

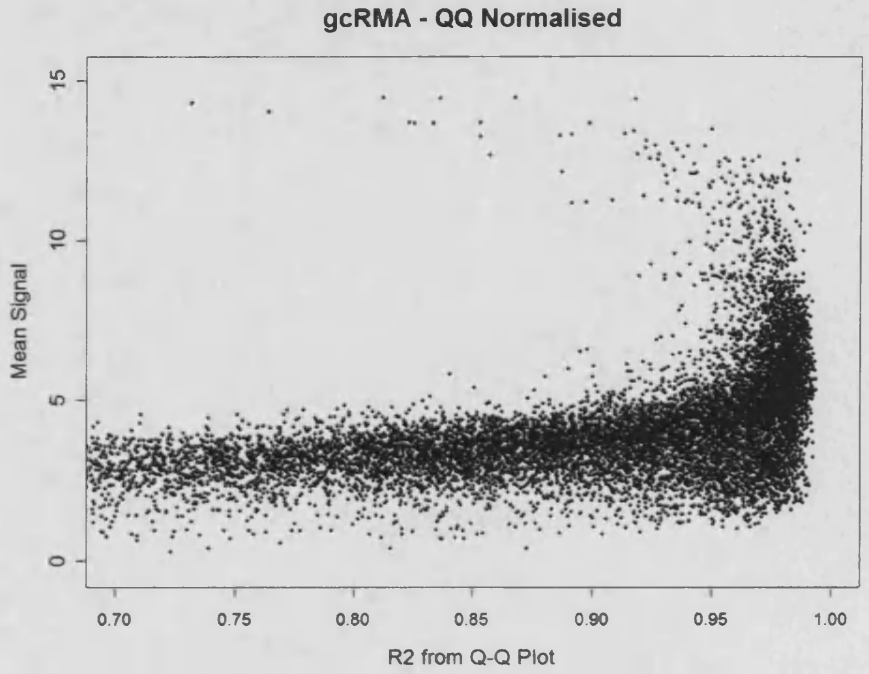


Figure 4.2 - Plots showing correlation to normality (R^2) versus mean expression signal for each probe set following the application of a QQ normalisation to each expression dataset.

Comparison of the plots to the equivalent plot in non-transformed data (Figure 3.2.d, 3.2.e and 3.2.f) shows little difference in the distributions. MAS 5.0 (Figure 4.2.a) and gcRMA (Figure 4.2.c) still show a large number of low expressed probe sets showing deviation from normality, with the MAS 5.0 plot showing the majority of the poorly scoring probe sets have a relative expression level of less than 100 forming a pronounced tail to the plot.

The pronounced return to normality in the lowest expressed gcRMA data is lost, forming a tail of non-normal probe sets at low expression levels. Little difference is seen in the RMA plot (Figure 4.2.b) except a possible widening of the cluster of genes correlating highly with normality. These data show that within this dataset QQ normalisation does not have a large effect on data distributions.

4.3.1.2 FDR performance of QQ normalised data

Performance of QQ normalised data was assessed using the already described framework (Section 4.2) using the Welch t-test to detect the fifteen spikes present in the 24 chips with a two-fold change in the Affymetrix Latin square dataset, assessed using FDR curves over a range of sample sizes with multiple samples. MAS 5.0 data was used in an untransformed form, RMA and gcRMA data was analysed as provided (i.e. incorporating a final $\log_2$ transformation). The resultant summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figure 4.3 plotted with the non-normalised data for comparison.

Figure 4.3.

a)

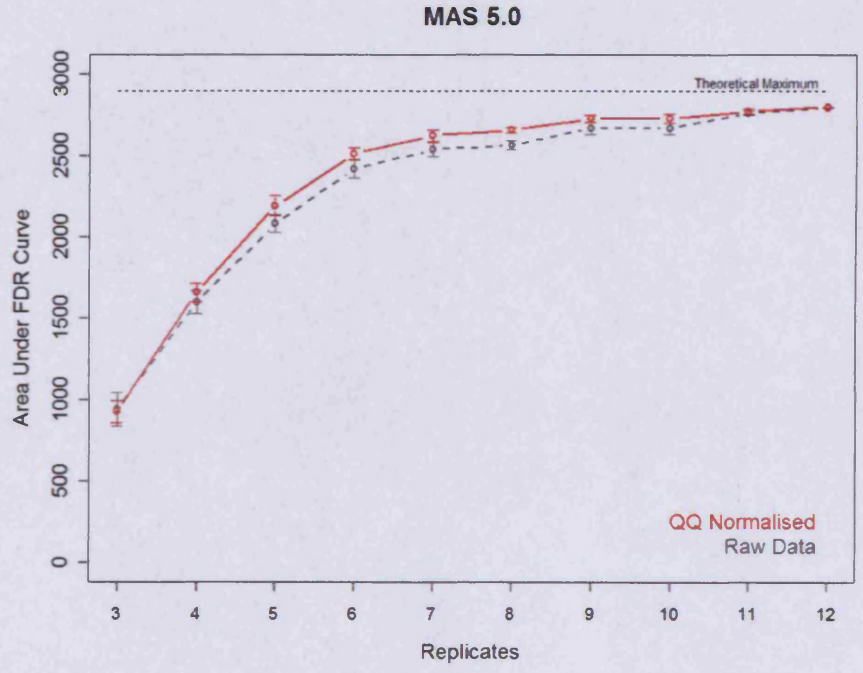
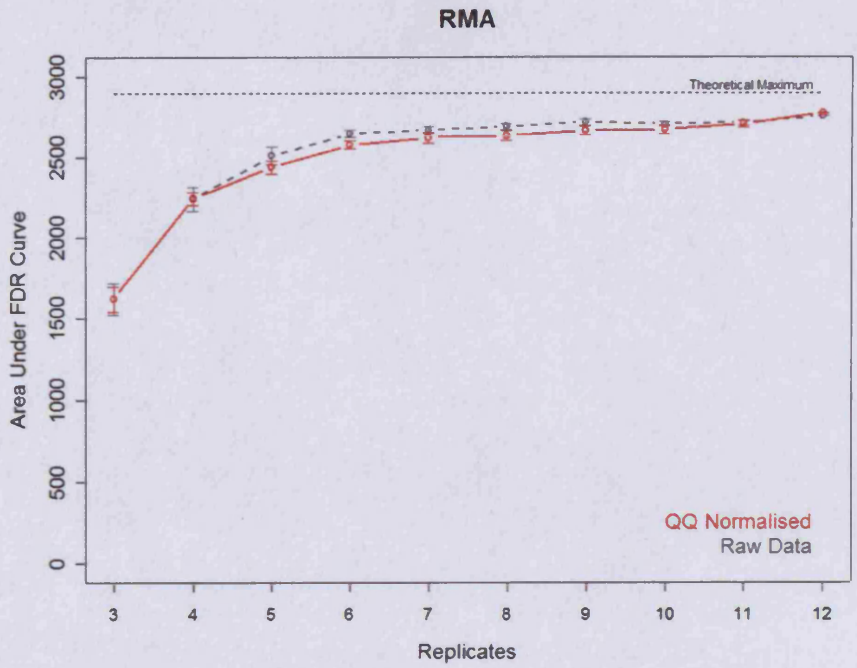


Figure 4.3 – continued

b)



c)

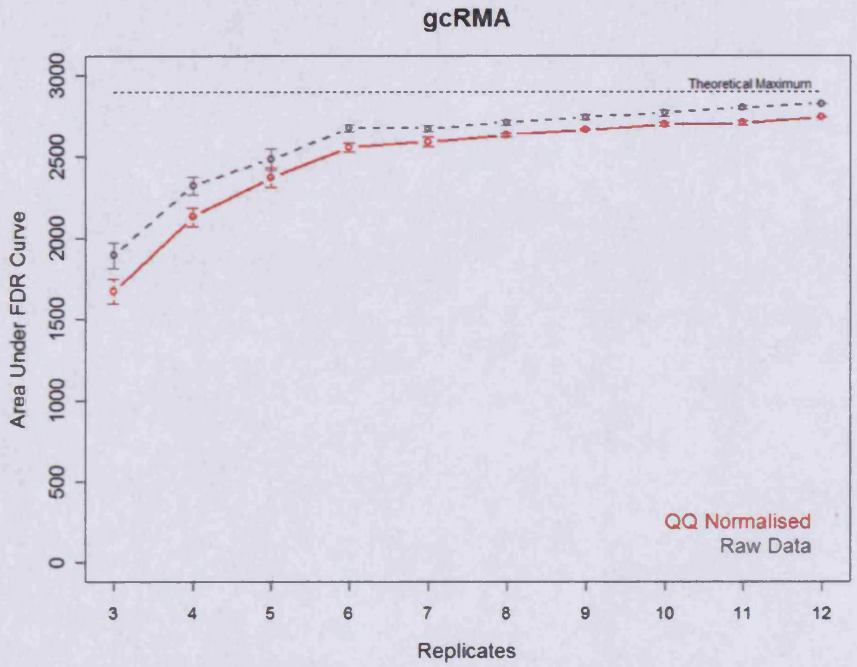


Figure 4.3 – Summary FDR plots showing the results of analysis using the Welch t-test on non-normalised data and QQ normalised data from MAS 5.0 (a), RMA (b) and gcRMA (c).

Review of the MAS 5.0 data (Figure 5.2.a) shows a marginal improvement in detection power of mid-ranged sample sizes (6-9 arrays per group), with convergence at small and large sample sizes. As MAS 5.0 does not incorporate any inter chip normalisation as part of the analysis the application of the QQ normalisation has the potential to draw any chips with a minor outlier into line with the other chips and improve detection.

Very little difference in detection power is seen with the RMA data (Figure 5.2.b) with a significant improvement in detection seen only for a few sample sizes using non-normalised data. Data from gcRMA shows a significant loss of power to detect across all sample sizes, suggesting that the application of a second round of normalisation to the data is detrimental to analysis of gcRMA data.

4.3.2 Application of Variance Stabilisation Normalisation (VSN)

In Chapter Four the Welch t-test was shown to outperform simple fold change by incorporation of information on the variance of the probe set expression values comprising a group of data. Various publications (Baldi and Long, 2001; Durbin, et al., 2002; Naef, et al., 2002) have indicated that in Affymetrix microarray data there is a relationship between variance and mean expression level, as the signal level increases for a probe set, so does the variance. VSN (Variance Stabilisation Normalisation) exploits this observation and provides a data transformation based on the mean-versus-variance dependency within the natural space data (Huber, et al., 2002).

A model is built of the variance-versus-mean dependence within the dataset before data transformation using a combination of data offset, gain and estimate of the true abundance for a probe set signal intensity, values that are derived from the modelling process. The data then undergoes an inverse hyperbolic sine transformation (asinh). The asinh transform is equivalent to the log transformation and gives near identical values for numbers greater than five; however the asinh transform does not produce negative numbers which cause problems for researchers wishing to implement a fold change comparison (Section 3.3.1).

VSN has been proposed as an alternative normalisation method within the RMA package as an alternative to QQ normalisation (Hartmann, et al., 2003) on the probe level data and was found to perform favourably when compared to MAS 5.0 and an RMA model incorporating QQ normalisation. It is therefore of interest to explore how the application of VSN to the resultant expression metric data can better inform an analysis, tested using spike detection of a dataset with known truth. To make results comparable to that from previous analysis, a sinh transform was used to reverse the effect of the asinh transform implemented as part of the VSN methodology.

4.3.2.1 Normality of VSN transformed data

Using the strategies introduced in Section 3.3.1, the normality of data following the application of a VSN transformation was first assessed using the Shapiro-Wilks test for normality, the number of non-normal genes (scoring a p-value less than 0.05) for each of the four datasets is shown in Table 4.2. Data from each expression metric was analysed in a natural form, requiring the removal of the final $\log_2$ transformation from RMA and gcRMA data.

Table 4.2

Analysis Method	Number of probe sets deviating from normality (p<0.05 from SW test)	
	Untransformed	VSN Transformed
MAS 5.0	5799 (46%)	5808 (46%)
RMA	3075 (25%)	3666 (29%)
gcRMA	6371 (51%)	9760 (78%)

Table 4.2 – Results from Shapiro-Wilks tests for normality.

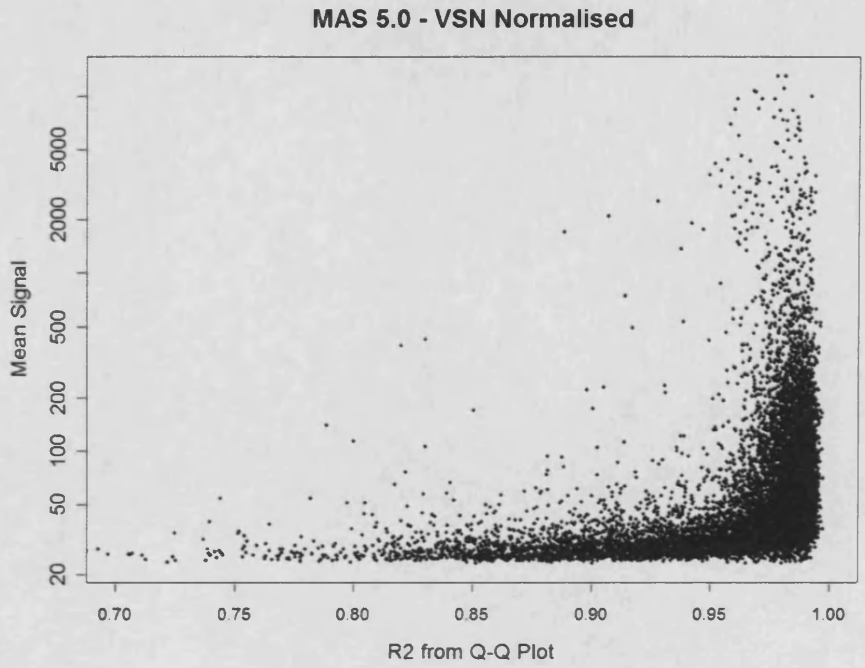
The Shapiro-Wilks tests show that the normality of data from MAS 5.0 is unchanged following VSN transformation. However, for RMA and gcRMA there is a marked increase in the number of probe sets presenting with a correlation differing from normality. This observation highlighted the need for further exploration of the nature of this deviation from normality using plots showing mean signal plotted against correlation to normality for a Q-Q plot (Section 2.3.2). The resultant plots are shown in Figure 4.4.

Comparison of the plots to the equivalent plot on non-transformed data (Figure 3.2.d, 3.2.e and 3.2.f) shows differences in the distributions. MAS 5.0 (Figure 4.4.a) has the closest match to the untransformed raw data, still presenting with a large tail of non-normal probe sets which have a lower expression value. Similarly to the results obtained from a QQ normalisation, the RMA plot (Figure 4.4.b) shows a widening of the cluster of genes correlating highly with normality (R^2 greater than 0.95).

Data from gcRMA (Figure 4.4.c) still shows a large number of low expressed probe sets showing deviation from normality. However there is a loss of the large cluster of data correlating highly with normality, and a general increase in the size of the tail along with an increase of density. As found in the results of the Shapiro-Wilks test, the majority of data after VSN transformation of gcRMA data is non-normal.

Figure 4.4

a)


MAS 5.0 - VSN Normalised
Mean Signal
20
50
100
200
500
2000
5000
0.70
0.75
0.80
0.85
0.90
0.95
1.00
R2 from Q-Q Plot

b)

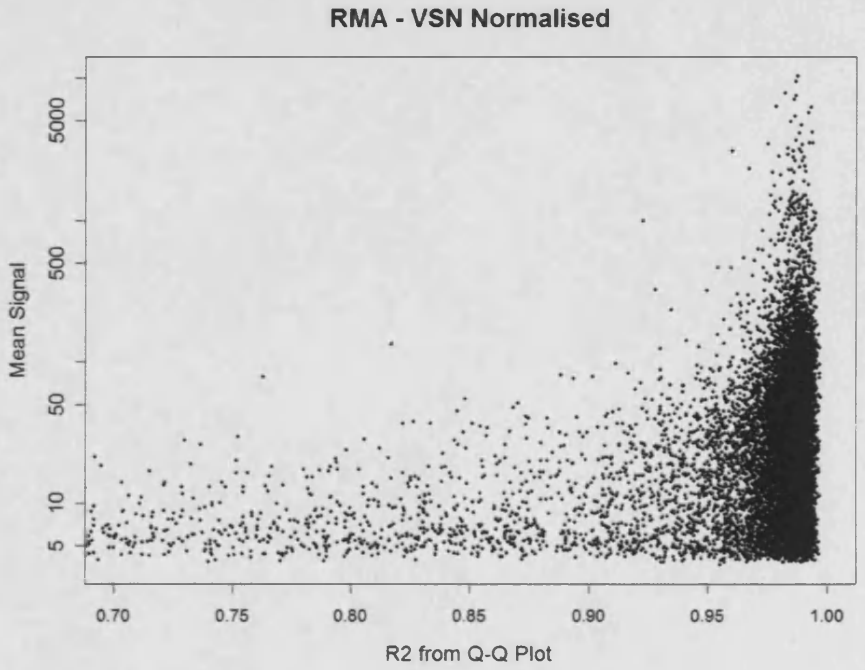

RMA - VSN Normalised
Mean Signal
5
10
50
500
5000
0.70
0.75
0.80
0.85
0.90
0.95
1.00
R2 from Q-Q Plot

Figure 4.4 - continued

c)

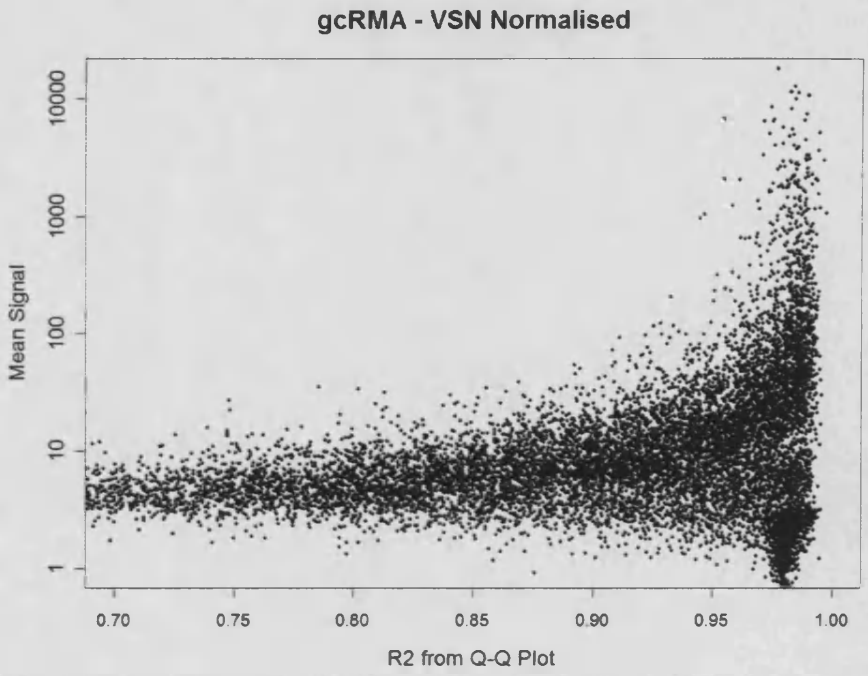


Figure 4.4 - Plots showing correlation to normality (R^2) versus mean expression signal for each probe set following the application of a VSN normalisation to each expression dataset.

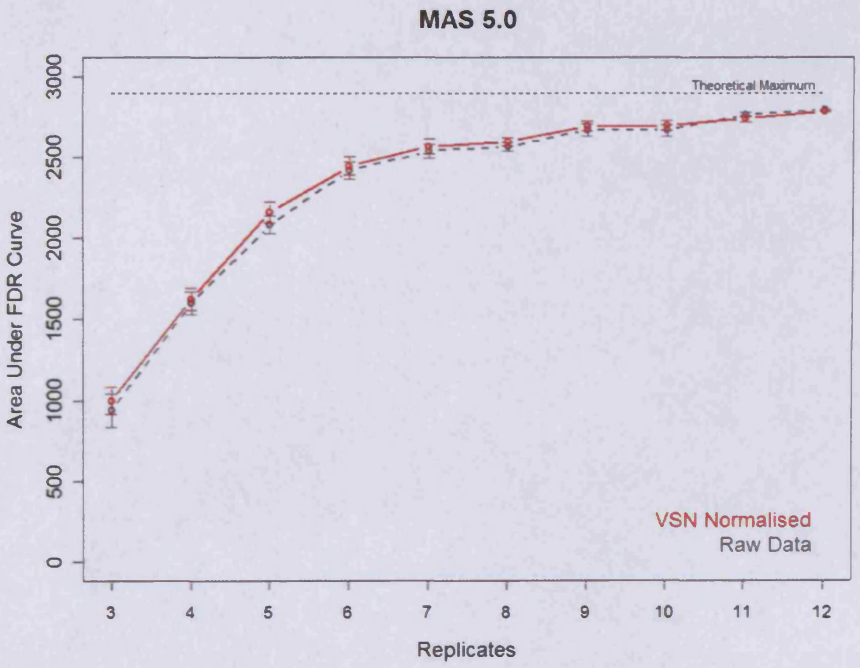
It can be seen that VSN transformation does not appear to have a large negative effect on the normality on data from MAS 5.0 and RMA, however the large changes in correlation to normality in data from gcRMA suggest the need for caution in the application of this method with parametric testing. It is therefore of interest to scrutinise the effect of this transformation in the detection of spiked in probe sets in the manner previously described.

4.3.2.2 FDR performance of VSN normalised data

Performance of VSN normalised data was assessed using the previously described framework (Section 4.2) using the Welch t-test to detect the fifteen spikes present in the 24 chips with a two-fold change in the Affymetrix Latin square dataset assessed using FDR curves over a range of sample sizes with multiple samples. Data from each expression metric was analysed in a natural form, requiring the removal of the final $\log_2$ transformation from RMA and gcRMA data. The resultant summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figure 4.5 plotted with FDR curves from the non-normalised data for comparison.

Figure 4.5

a)


MAS 5.0
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
VSN Normalised
Raw Data

b)

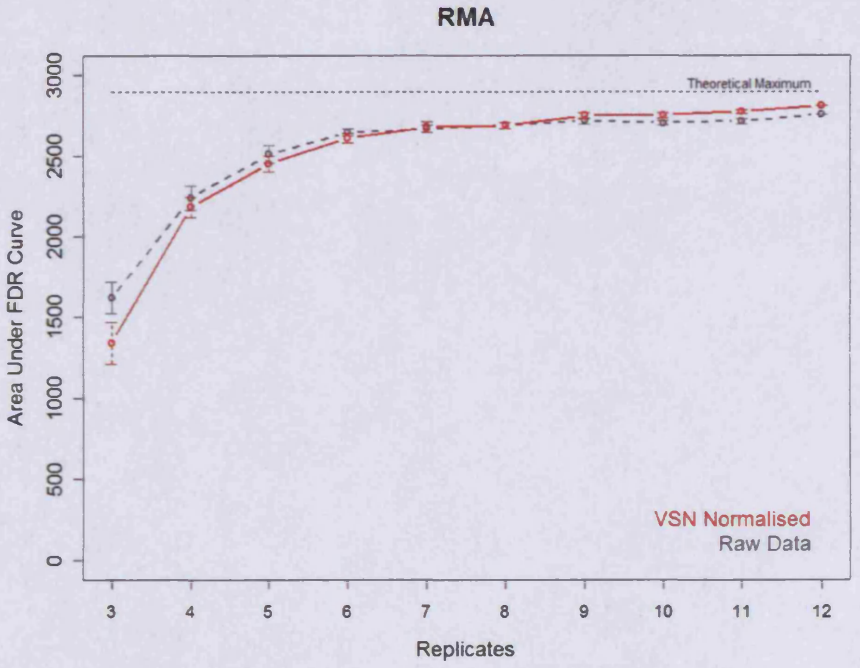

RMA
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
VSN Normalised
Raw Data

Figure 4.5 – continued

c)

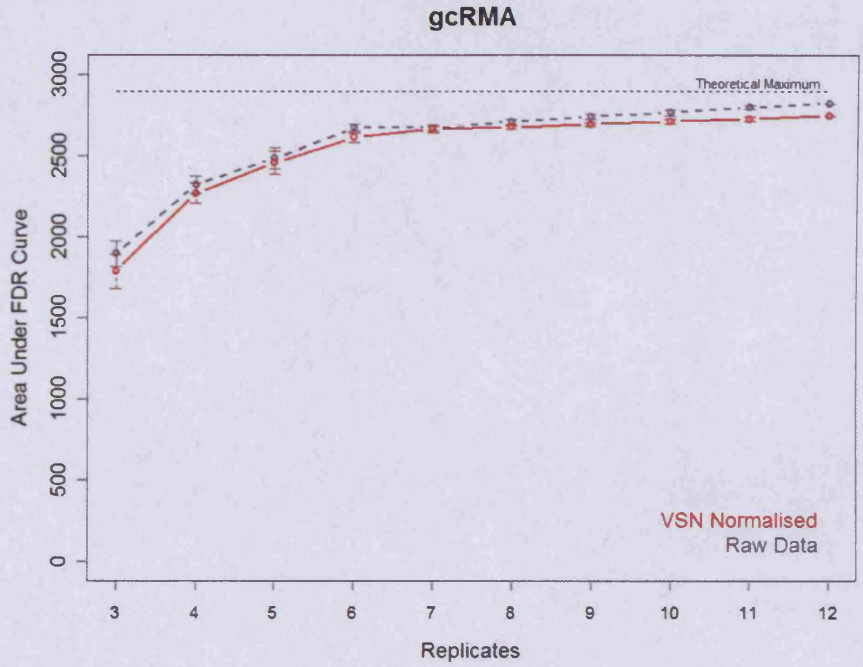


Figure 4.5 – Summary FDR plots showing the results of analysis using the Welch t-test on non-normalised natural space raw data and VSN normalised data from MAS 5.0 (a), RMA (b) and gcRMA (c).

It can be seen that the application of the VSN transform has very little effect on the detection outcomes in each of the expression metrics, with a slight loss of power observed in data from gcRMA. The heterogeneity of this sample would suggest that there is likely to be little variance between samples, and hence the observation of minimal changes in power between the VSN normalised and raw datasets.

5.3.3 The use of rank as an alternative signal measurement

The use of rank instead of actual data readings is a common technique in microarray analysis. In Chapter Four we introduced the Mann-Whitney test as a non-parametric alternative to the t-test, which uses the data's group ranking instead of the data value; Breitling et al. (Breitling, et al., 2004) introduced a technique based on the product of ranks following pair-wise comparison of individual arrays which performed favourably with fold-change calculation. Martin et al. (Martin, et al., 2004) proposed an analysis method based on Rank Difference Analysis of Microarrays (RDAM) where each signal value is replaced a ranking value between zero and one hundred, followed by pair-wise comparison of rank differences using the empirical null distribution. The authors comment that "*this simple transformation is a powerful normalizing procedure*".

Taking a similar approach to Martin, explorations were undertaken replacing data values with a rank value for each array independently (between 1 and 12545) and then use these values in the same manner as the signal values in the determination of differential gene expression. As this replacement takes the data from an interval data type to an ordinal form, examination of the effects of this step on the normality and distribution of the data is likely to be more important than that after QQ normalisation or VSN transformation.

4.3.3.1 Normality of rank replacement data

Using the strategies introduced in Section 2.3.1, the normality of data following the application of rank replacement was first assessed using the Shapiro-Wilks test for normality; the number of non-normal genes (scoring a p-value less than 0.05) for each of the four datasets is shown in Table 4.3.

The Shapiro-Wilks tests show that the normality of data from RMA is slightly increased following rank replacement. However, for MAS 5.0 and gcRMA there is a marked increase in the number of probe sets presenting with deviation from normality. This observation highlighted the need for further exploration of the nature of this deviation from normality using plots showing mean signal plotted against correlation to normality for a Q-Q plot (Section 2.3.2). The resultant plots are shown in figure 4.6.

Table 4.3

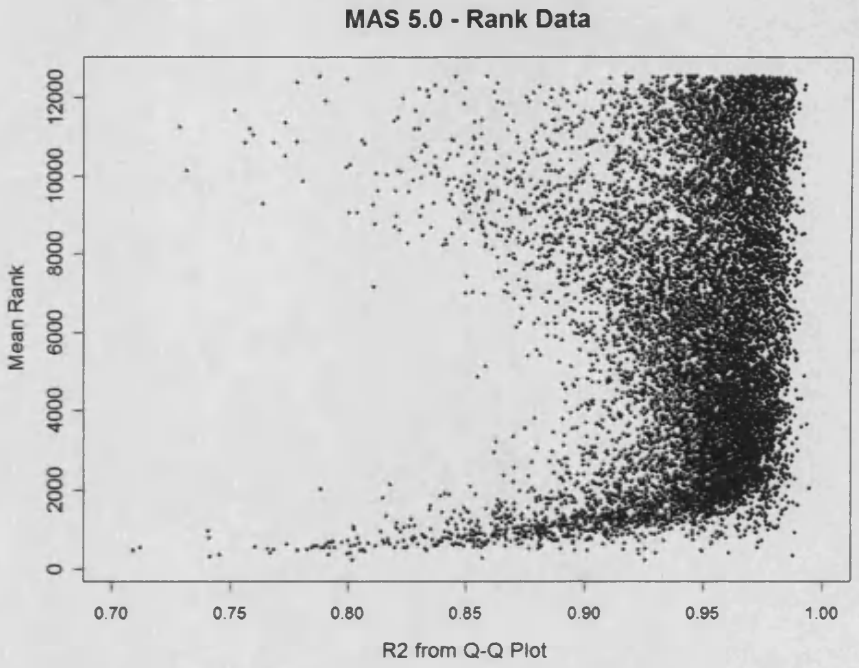
Analysis Method	Number of probe sets deviating from normality ($p<0.05$ from SW test)	
	Untransformed	Rank Transformed
MAS 5.0	5799 (46%)	8799
RMA	3075 (25%)	3810
gcRMA	6371 (51%)	9177

Table 4.3 – Results from Shapiro-Wilks tests for normality.

Data from MAS 5.0 (Figure 4.6.a) shows a large number of non-normal probe sets with tails forming in data with high and low mean ranks. Whilst many probe sets are non-normal many of these form a broad cluster in the higher scoring R^2 values, so whilst they present as *"failing"* normality testing, the robustness of statistical tests may overcome the issues this rank replacement generates.

Figure 4.6

a)


MAS 5.0 - Rank Data
Mean Rank
0
2000
4000
6000
8000
10000
12000
0.70
0.75
0.80
0.85
0.90
0.95
1.00
R2 from Q-Q Plot

b)

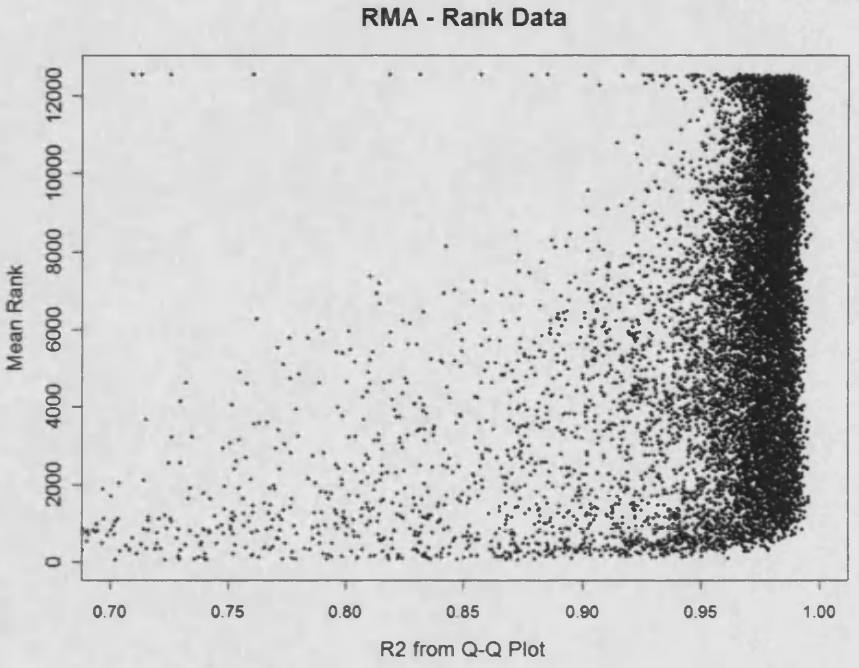

RMA - Rank Data
Mean Rank
0
2000
4000
6000
8000
10000
12000
0.70
0.75
0.80
0.85
0.90
0.95
1.00
R2 from Q-Q Plot

Figure 4.6 - continued

c)

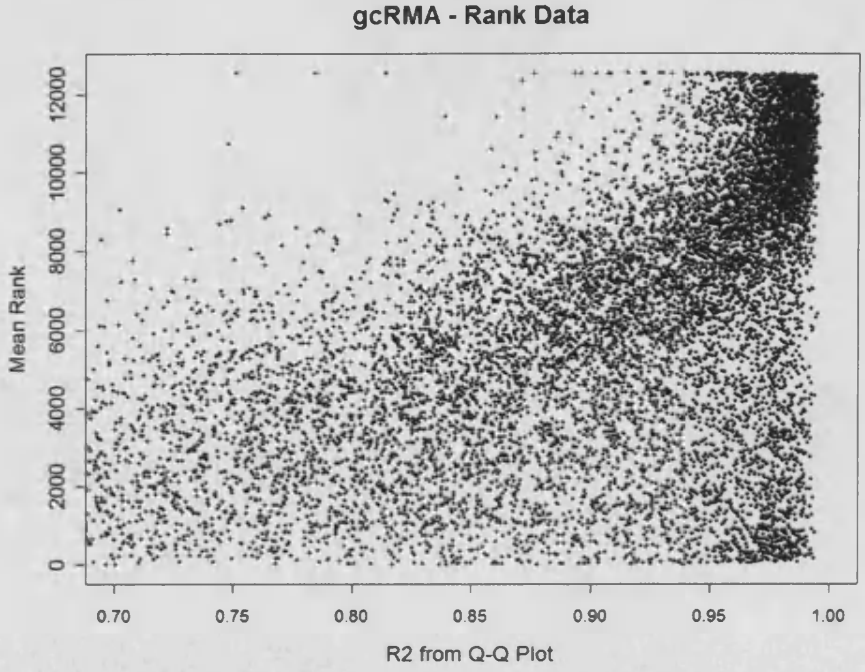


Figure 4.6 - Plots showing correlation to normality (R^2) versus mean expression signal for each probe set following the application of a rank based data substitution to each expression dataset.

Review of the data from RMA (Figure 4.6.b) shows a large amount of data correlating well with normality (R^2 greater than 0.95). A small tail is formed on the plot from expression values presenting with a low mean rank, and a few top scoring probe sets becoming non-normal.

gcRMA data (Figure 4.6.c) presents with the majority of probe sets being non-normal, with only a few high and low ranking probe sets correlating with normality. This observation once again suggests the need for caution in the application of gcRMA in combination with statistical based testing. It is therefore of interest to scrutinise the effect of this transformation in the detection of spiked in probe sets in the manner previously described.

5.3.2.3 FDR performance of rank transformed data

Performance of the rank transformed data was assessed using the previously described framework (Section 4.2) to detect the fifteen spikes present in the 24 chips with a two-fold change in the Affymetrix Latin square dataset assessed using FDR curves over a range of sample sizes with multiple samples. MAS 5.0 data was used in an untransformed form, RMA and gcRMA data was analysed as provided (incorporating a final $\log_2$ transformation).

Since transformation technically takes the data from an interval type to an ordinal form, reference back to the statistical flow chart in Chapter Three (Figure 2.1) shows that the correct choice of test is the Mann-Whitney test. Although not ideal, due to the lack of power at lower sample sizes, the rank data was subjected to the Mann-Whitney test within the FDR framework. The resultant summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figure 4.7 plotted with the Mann-Whitney data for the untransformed raw data for comparison.

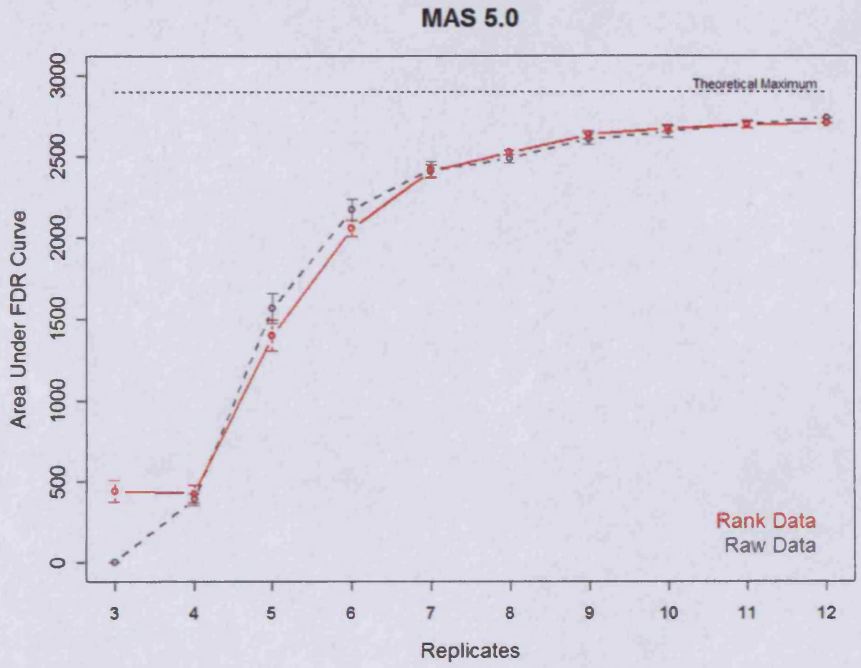
Interestingly at the smallest sample size of three replicates per group, both MAS 5.0 (Figure 4.7.a) and RMA (Figure 4.7.b) have an increased power to detect using the rank transformed data. Data analysed using MAS 5.0 has a similar power to detect using either method, with tight convergence at the higher sample numbers. With the exception of a few of the smaller sample sizes, the rank transformed had a lower power to detect when applied to both RMA and gcRMA data (Figure 4.7.b and 4.7.c), although there is indication of potential convergence as sample size increased.

Because of the power loss associated with using the Mann-Whitney test compared to a parametric test on an equivalent sample size it is of interest to examine the FDR performance of rank transformed data when subjected to the Welch t-test. Whilst technically incorrect because of the data type, it could be argued that data with a raage of zero to twelve thousand is a pseudo-interval data type. This is an approach that has been utilised (Breitling, et al., 2004) and whilst the data does present as mainly non-normal, examination of Figure 4.6 shows that for MAS 5.0 and RMA the majority of data does present with a correlation tending towards normality.

Rank data was subjected to the Welch t-test within the FDR framework. The resultant summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figure 4.8 plotted with the Welch t-test data for the untransformed raw data for comparison.

Figure 4.7

a)


MAS 5.0
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
Rank Data
Raw Data
3
4
5
6
7
8
9
10
11
12
Replicates

b)

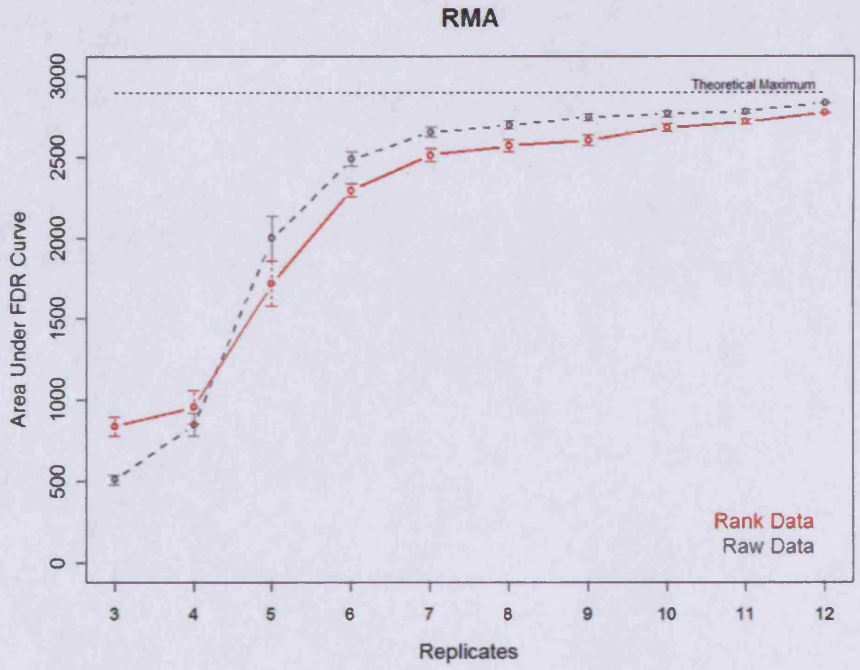

RMA
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
Rank Data
Raw Data
3
4
5
6
7
8
9
10
11
12
Replicates

Figure 4.7 – continued

c)

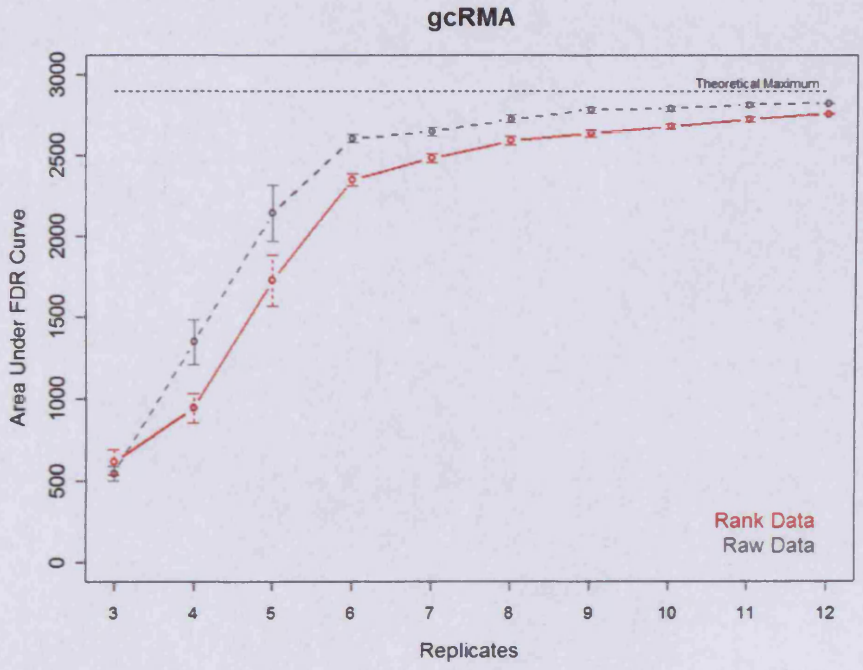


Figure 4.7 – Summary FDR plots showing the results of analysis using the Mann-Whitney test on the original and rank substituted data from MAS 5.0 (a), RMA (b) and gcRMA (c).

Figure 4.8

a)

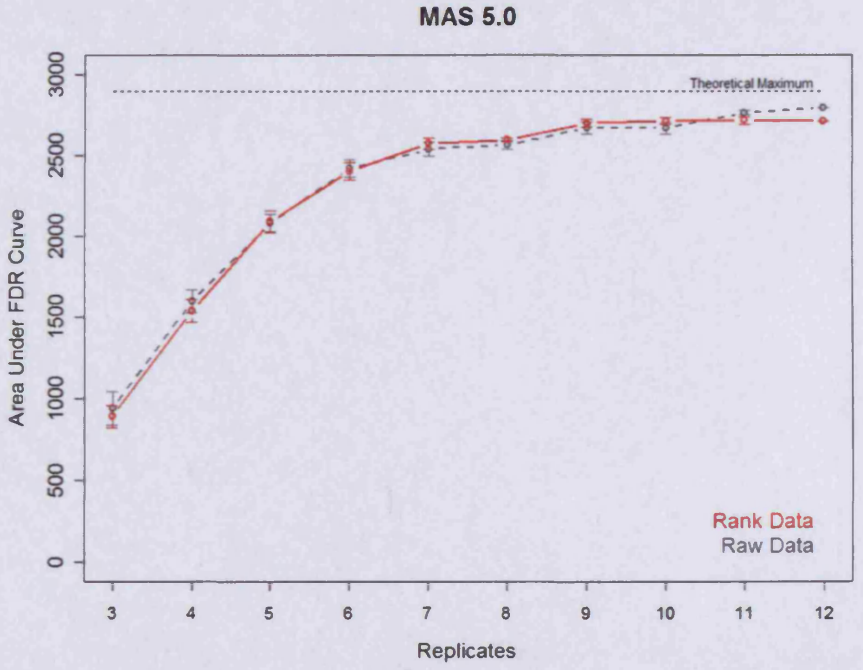
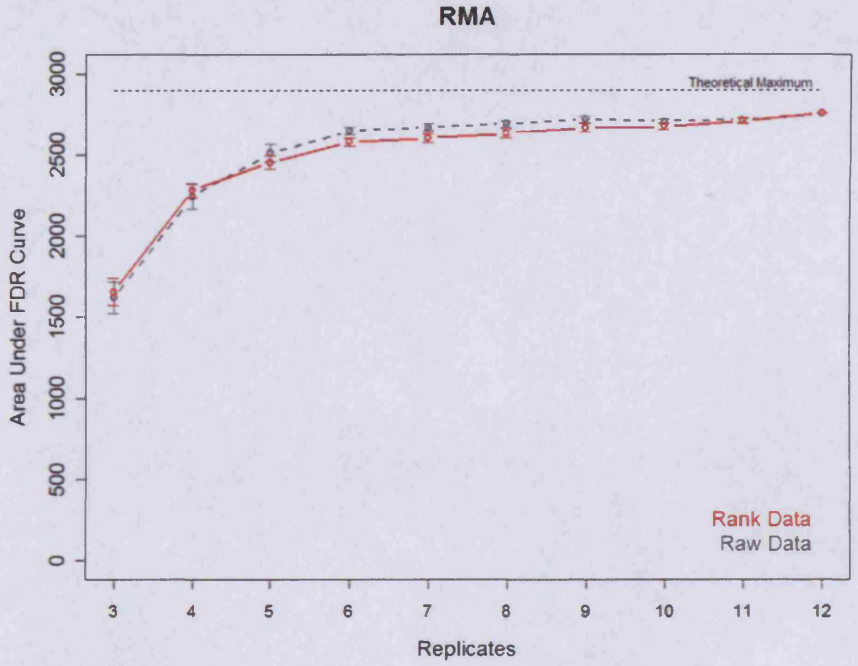


Figure 4.8 - continued

b)



c)

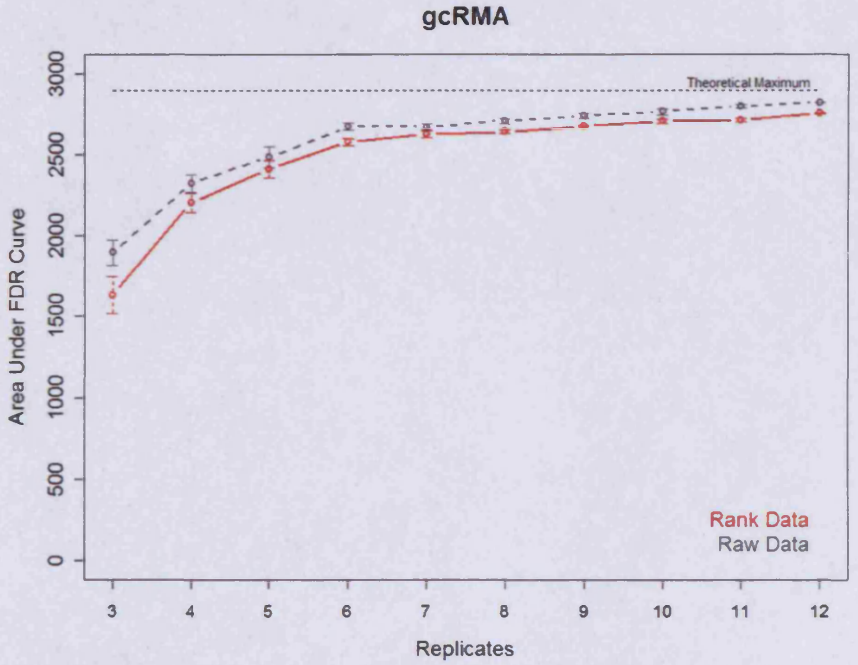


Figure 4.8 – Summary FDR plots showing the results of analysis using the Welch t-test on the original and rank substituted data from MAS 5.0 (a), RMA (b) and gcRMA (c).

Review of the data from MAS 5.0 (Figure 4.8.a) and RMA (Figure 4.8.b) shows very little difference in power to detect the spiked in probe sets using the Welch t-test between the rank transformed and untransformed data. Data from gcRMA (Figure 4.8.c) shows a small loss in power of detection when rank transformed, although this is a relatively small amount. Overall whilst application of this rank transformation does impact on the normality of the data it would appear to have very little impact on the outcome of spike detection in this Latin square dataset.

4.4 Discussion

In this Chapter the idea of normalisation was introduced along with experimental review of the effectiveness of the normalisation in improving power to detect spiked in probe sets within the Affymetrix Latin square dataset. Whilst the application of a normalisation step is logical in order to overcome obscuring variance in data, care must be taken in their application due to the risk of negative effects on the data. Wolkenhauer et al. (Wolkenhauer, et al., 2002) comment that whilst correcting for this non-biological variation normalisation can reduce the information content of the data.

4.4.1 Application of QQ normalisation

Following the application of a QQ normalisation the normality of datasets from all three expression metrics was slightly reduced, but the data broadly followed the same distributions as the untransformed data (compare Figure 5.1 to Figure 3.2 d, e and f). Power to detect the spikes in the truth dataset was marginally improved with QQ normalised MAS 5.0 data, had no effect on the outcome from RMA data, and slightly reduced power in the gcRMA dataset.

These results suggest little evidence against the application of a QQ normalisation, and suggest a positive benefit in the application of an inter chip normalisation method to MAS 5.0 data.

4.4.2 Application of VSN transformation

The application of a VSN transformation to data from MAS 5.0 had little effect on the number of probe sets presenting as being normally distributed, however increased the number of non-normal probes in data from RMA and gcRMA. However, VSN transformed data had a very similar power of detection when compared to the untransformed data.

Whilst there is little evidence to suggest support for the application of a VSN transformation as part of a default analysis, VSN may be of use in analyses which use analysis of covariance to detect differential gene expression by removing the relationship between mean expression signal and variance.

4.4.3 Application of rank transformation

Substitution of signal values with the data rank within an array resulted in large deviations from normality as would be expected as a result of a complete replacement of the data distributions. However, the transformation would appear to have very little impact on the outcome of spike detection in this Latin Square dataset with similar power seen with both the Mann-Whitney and the Welch t-test. A rank based data replacement therefore presents as an alternative normalisation method.

4.4.4 Overall Conclusions

Generally each of the normalisation techniques produced very little effect on analysis outcomes when applied to the Latin Square dataset. However, it should be noted that this is a very clean and homogenous dataset and more typical experimental data may benefit more from the application of these techniques.

Finkelstein et al. (Finkelstein, et al., 2002) comment that no single normalization or correction method currently available is able to address all the issues normalisation aims to overcome, but careful sequence selection, array design, experimental design and experimental annotation can substantially improve the quality and biological of microarray data. Another approach to deal with the variance is the application of more robust statistical testing; it is this that will be addressed in the following Chapters.

Chapter Five
Application of robust statistical testing

In this Chapter the application of statistical testing to Affymetrix microarray data is revisited looking at more robust testing methodologies designed to address the presence of outlier data within a dataset. Section 5.1 introduces the concepts of robust methods and examines the requirement to address outlying data within Affymetrix datasets. Section 5.2 reviews the methods used to investigate the practicalities and power of these tests before Section 5.3 explores the ability of a five robust methods to detect the spikes in the U95A Latin Square dataset. Section 5.4 discusses the results and observations regarding the application of statistically robust techniques within an analysis.

5.1 Introduction

In previous Chapters, the issue of applying statistical testing to microarray data highlighted problems of small sample size and the need to apply parametric testing to achieve maximal power from a dataset. The effect of systematic changes in the data distribution (such as the positive skew seen in some MAS 5.0 probe sets – Section 3.3.3) were identified as having the potential to have an effect on the test outcome, however, in practice, the non-normality seems did not appear to be extreme enough to have a marked effect on the classic t-test.

In an experimental environment there is often a source of data variation which is more difficult to control than shifts in data distribution. Problems with the biological processing of a sample, such as poor RNA quality or differences in reagents used for different chips within a study, can lead to one or more chips presenting with data very different to others representing identical data. A researcher may be reluctant to exclude such data from an analysis, due both the loss of power from discarding a sample and the cost of repetition.

In statistical terms these erroneous data points are termed outliers. Many statistical tests are sensitive to the presence of outliers. For example, a single grossly inaccurate data point may distort simple calculations of the mean and standard deviation (Iglewicz and Hoaglin, 1993). Statistical tests that are designed to overcome the challenges introduced by these potentially invalid data points are termed robust tests.

The Oxford English Dictionary describes robust as "*sturdy or able to withstand difficult conditions*". Applied to statistics, a test can be considered robust if it is not markedly affected by poorly structured data or outlying data points. In this Chapter a variety of methods are introduced to overcome outliers, including one method which aims overcome distributional issues without reverting to non-parametric methodologies.

5.1.1 Factors contributing to the presence of outliers in microarray data

A typical analysis looking for differential gene expression is concerned with taking a sample of a population, and comparing it to another set of data to draw conclusions about the differences. Pivotal to this approach is the belief that that data within a group is relativity consistent and is representative of its population.

Outliers in microarray datasets can originate from a number of causes including systematic technical outliers, sporadic technical outliers, biological outliers and chance outliers (occurring due to the finite sample sizes and sampling used). To detect and accommodate these outlying data points sufficient replication of data is required (Loguinov, et al., 2004).

Generalising, systematic outliers can be assigned two distinct causes; those occurring due to the biological information inputted into the microarray process, and technical outliers occurring as a result of the experimental process.

Whilst there is always the possibility that biological variation can be introduced to the experimental process, with good experimental design, and standardised sampling handling and extraction processes the chance of outliers occurring within an experiment can be reduced. There are many reasons for the presence of biological variation which ultimately presents as an outlier within an analysis, including differences in sample handling and treatment, differing degradation of RNA between samples, differences occurring as a result of temperature effects acting on a sample, differences in sampling or the issue that cells may be at differing stages of cyclic processes or development (e.g. the cell cycle) (Kadota, et al., 2003).

These chances of this type of biological outliers occurring can be reduced with good experimental procedure; however there is another source of biological outliers that is more difficult to control.

In some cases, heterogeneity of tissue types can result in outliers, for example biopsies may contain more than one distinct cell type however good the pathological extraction of the sample. This type of outlier can be more difficult to control and may require the ultimate elimination of a sample from analysis.

Because of the large number of steps involved in the experimental process from hybridization to image analysis, there are many stages at which systematic technical outliers can be introduced to an experiment. Example of reasons for these outliers include: dust, scratches on the array surface, imperfections in hybridisation, staining or scanning, errors in array production, or sample contamination or miss-labelling (Kadota, et al., 2003).

Another important stage at which outliers can also occur is in the initial analysis of the image file for conversion into numerical values for further analysis. Such an outlier can occur when the estimated difference between the background and the foreground intensities from the image analysis is small, or if an integral normalisation step is ineffective at producing the homogenous data required for further analysis (Gottardo, et al., 2003).

Reviewing the above exploring the origins of outliers, it can be summarised that an outliers can occur at any stage where insufficient care has been applied to reduce differences between samples, ultimately contributing to an inaccurate sampling and increased variation in the dataset.

5.1.1 Detecting outliers in a dataset

There is a considerable literature on the detection and accommodation of outliers (Barnett and Lewis, 1994; Iglewicz and Hoaglin, 1993). Iglewicz and Hoaglin comment that prior to considering the possible elimination of these points from the data; one should try to understand why they appeared and whether it is likely that similar values will continue to appear. They do, however, note that outliers are often just bad data points.

The ideal approach to examine outliers in a dataset is the examination of all outlying data points. This can be achieved by the application of manual or graphical techniques to explore the hypothesis that not all outliers are errors and that some outliers are genuine; indeed these genuine outliers may be the most important observations of the sample. However when applied to the thousands of probe sets within a typical microarray dataset this repetitive examination becomes an onerous task.

We must therefore look towards the less informative statistical techniques to assess the degree of outlying data in microarray datasets in order to examine how extreme an outlier is acceptable for the test chosen. Review of the statistical literature suggests a variety of tests including z-scores, Grubb's test, Dixon's test and Rosner's test to identify outliers. However, Hampel warns of the limitation of these techniques stating that many commonly applied tests cannot reject one distant outlier within a set of ten observations (Hampel, 1985).

5.2 Technical Methodology

Using the previously described methodology (Section 3.2), an integrated analysis script was written in R which took a subset of the Affymetrix U95A Latin Square dataset and produced a series of 20 replicates at a range of 3 to 12 chips in each of two groups using the MAS 5.0, RMA and gcRMA expression metrics. Next the script took each of these indexed matrices, extracted the relevant data and then ran a series of robust statistical tests with twenty replicates over a range of ten sample sizes. The statistical tests chosen for these robust investigations were three variants of the robust t-test, the trimmed t-test, the Winsorised t-test, Yuen's test, and a randomisation method.

Figure 5.1

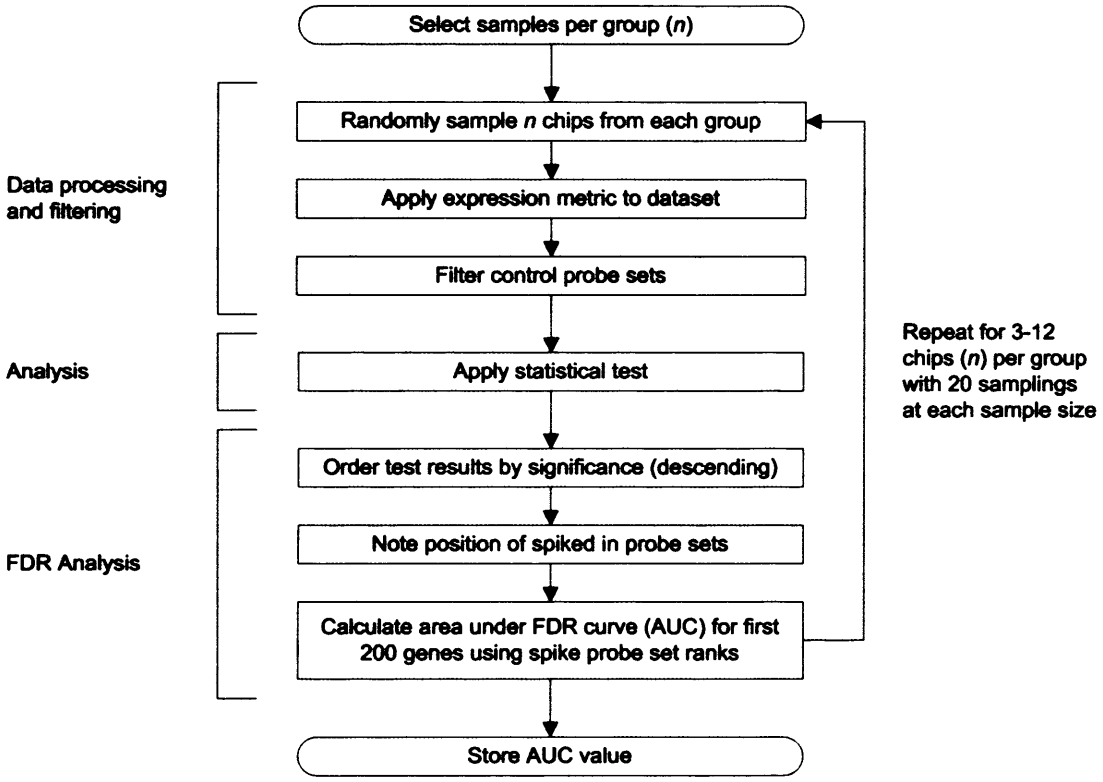


Figure 5.1 – Flow diagram of technical methodology for examination of power to detect performance of each robust statistical test.

The output for each test was fed into an FDR function which extracted the location of the spikes from the ordered p-values and calculated the area under the FDR curve as previously described. The output for each test was exported into a summary matrix containing each of the 200 AUC values calculated, grouped by sample size. Summary plots were produced charting the average area under the curve (with standard error, error bars), across the range of sample size under consideration.

Full details of the technical methodology are given in Section 9.5.1.

5.3 Exploration and Results

5.3.1 Improving the robustness of the Welch t-test

In previous Chapters it has been shown that the Welch t-test is and appropriate form of the t-test for analysis of Affymetrix microarray data because of its lack of assumption about equality of variance. In addition the test has been shown to perform well in its power to detect differential gene expression in the model Latin square dataset. However, in light of the observations regarding the number of outliers present in the dataset review of the test and explorations into improvement in robustness are of interest.

5.3.1.1 Fundamental estimators of the t-test

Review of the formulae behind the t-test yield two fundamental measures used in the calculation of the t-statistic (*t*): the sample mean and the sample variance. Review of the complex equation used to determine the degrees of freedom (*df*) of the Welch t-test show that standard deviation is the fundamental measure. The mean is an estimator of the location of the data and variance describes the spread.

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \qquad df = \frac{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right)^2}{\frac{(S_1^2/m)^2}{m-1} + \frac{(S_2^2/n)^2}{n-1}}$$

Where: $\bar{x}$, $\bar{y}$ *: Means of groups.* S_1^2, S_2^2 *: Standard deviations of groups.*
m, n : Number of observations in each group.

5.3.1.2 Robust estimators for the t-test

Expanding the ideas of Iglewicz and Hoaglin (1993) the values of mean and standard deviation can be substituted for more robust estimators of location and spread. The alternative estimators of median and median absolute deviation (MAD) are compatible with other variables within the t-test formulae and should in theory provide a more robust statistical test. The median absolute deviation (MAD) is a measure of the scale or dispersion of a distribution about the median, calculated as the median of the absolute-value distances of the points about the median.

Combinations of the estimators for location and spread into the t-test formula results in four different tests statistics.

	Standard Deviation (SD)	Median Absolute Deviation (MAD)
Mean	Welch t-test	Mean / MAD test
Median	Median / SD test	Median / MAD test

5.3.1.3 FDR performance of robust variants of the t-test

Performance of each robust variant of the Welch t-test was assessed using the previously described framework (Section 5.2). The ability of each test to detect the fifteen spikes present in the 24 chips with a two-fold change in the Affymetrix Latin square dataset was assessed using FDR curves over a range of sample sizes with multiple samples. MAS 5.0 data was used in an untransformed form, RMA and gcRMA data was analysed as provided incorporating the final $\log_2$ transformation. The resultant summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figure 5.1.

Review of the FDR plots (Figure 5.1) indicates that the technique with the most power to detect was the combination of mean and standard deviation (a standard Welch t-test). Substitution of the median for the mean still yielded good results; however, results indicated slight loss of power in comparison. Substitution of the standard deviation with the more robust median absolution deviation from the median significantly reduced the power of the test to determine the fifteen spikes in the dataset, and substitution of the median for the mean reduced this further.

Figure 5.1

a)

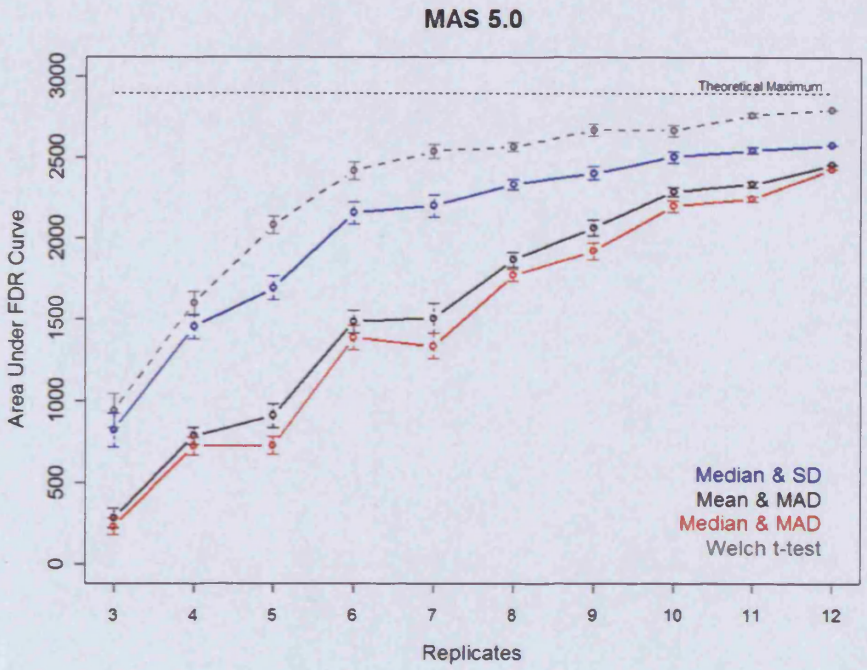
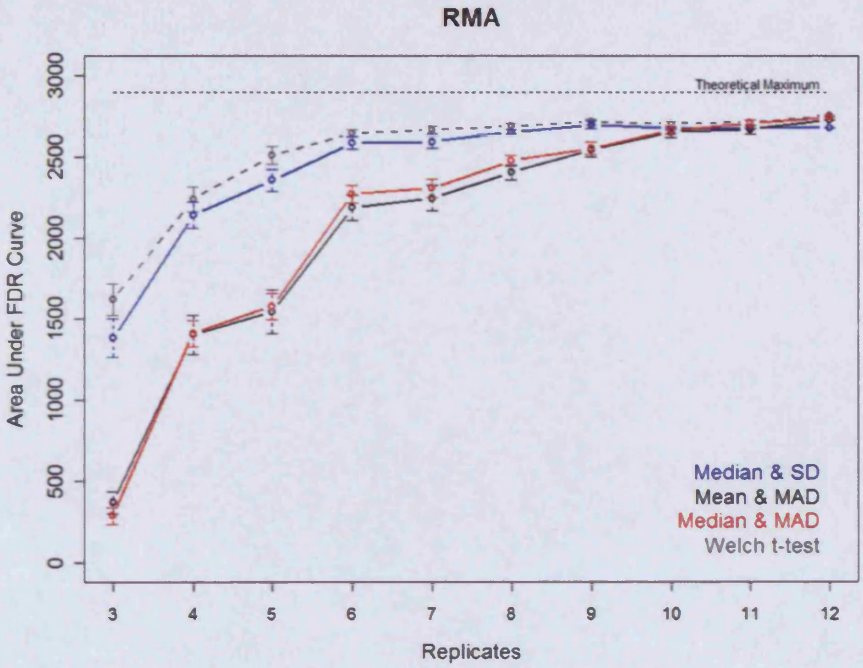

MAS 5.0
Theoretical Maximum
Median & SD
Mean & MAD
Median & MAD
Welch t-test
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates

Figure 5.1 - continued

b)



c)

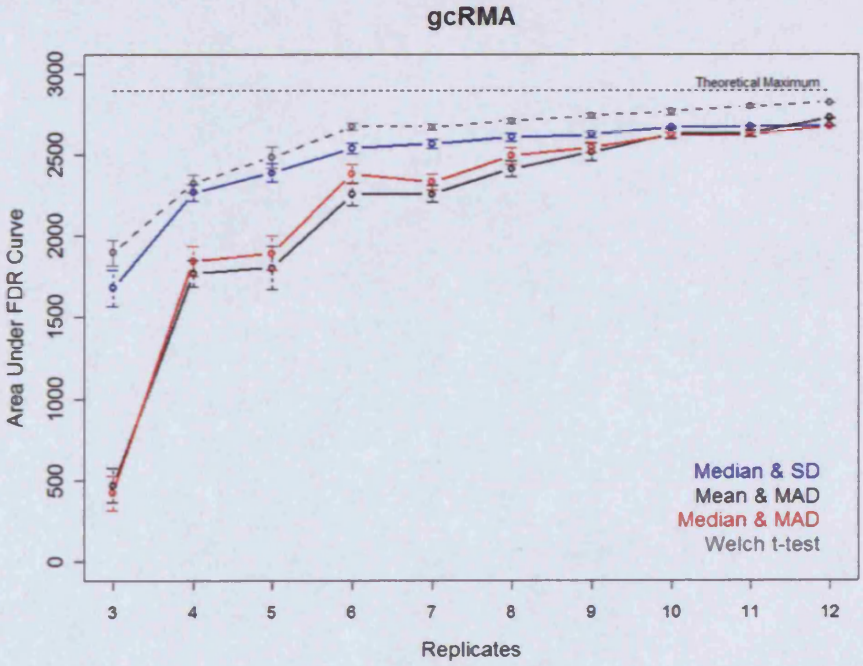


Figure 5.1 – Summary FDR plots showing the results of analysis using robust variants of the Welch t-test on data from MAS 5.0 (a), RMA (b) and gcRMA (c).

5.3.2 The trimmed t-test

The robust variants of the t-test introduced in Section 5.3.1 are based on the accommodation of outlying data points by use of more robust estimators within the analysis method. Another approach for dealing with outlying data points is the elimination of suspect data. Trimming is the simplest of methods available for dealing with the elimination of outliers from a dataset, whereby the largest and smallest observations are deleted from the sample.

This non-discriminatory approach to removing outliers results in a robust estimator of the mean that is relatively insensitive to the outlying values and an unbiased estimate of the population mean. However the trimmed data does not have a normal distribution even if the data are from a normal population and the trimming acts to reduce the sample variance (Dixon and Tukey, 1968; Tukey and McLaughlin, 1963).

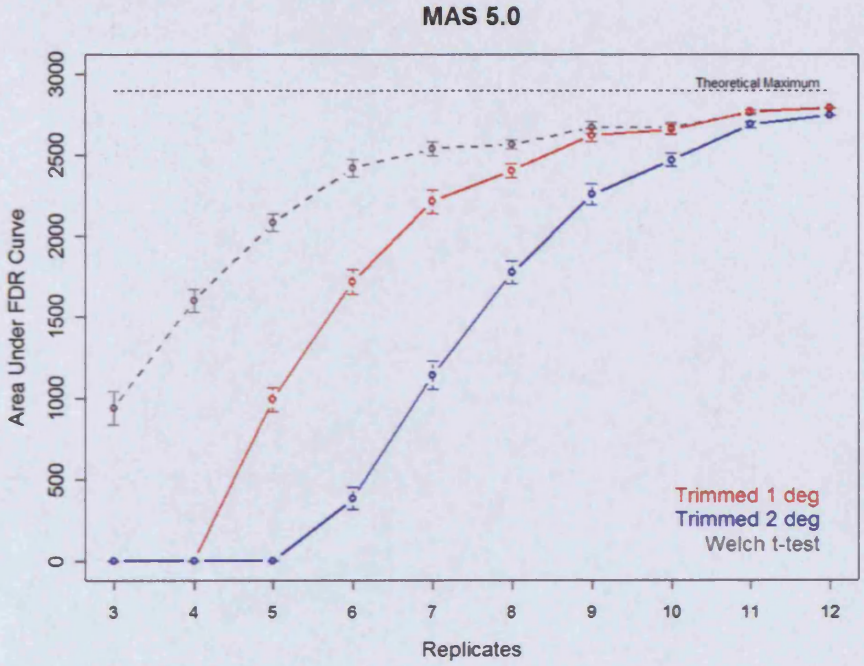
The amount of trimming applied to a dataset is described as degrees of trimming. Removal of one data point from each end of a dataset is termed 1 degree of trimming, two points from each end, 2 degrees of trimming and so on.

5.3.2.1 FDR performance of the trimmed t-test

Performance of the trimmed t-test was assessed using the previously described framework (Section 5.2) applying one and two degrees of trimming. The ability of each test to detect the fifteen spikes present in the 24 chips with a two-fold change in the Affymetrix Latin square dataset was assessed using FDR curves over a range of sample sizes with multiple samples. MAS 5.0 data was used in an untransformed form, RMA and gcRMA data was analysed as provided incorporating the final $\log_2$ transformation. The resultant summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figure 5.2.

Figure 5.2

a)


MAS 5.0
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
Trimmed 1 deg
Trimmed 2 deg
Welch t-test

b)

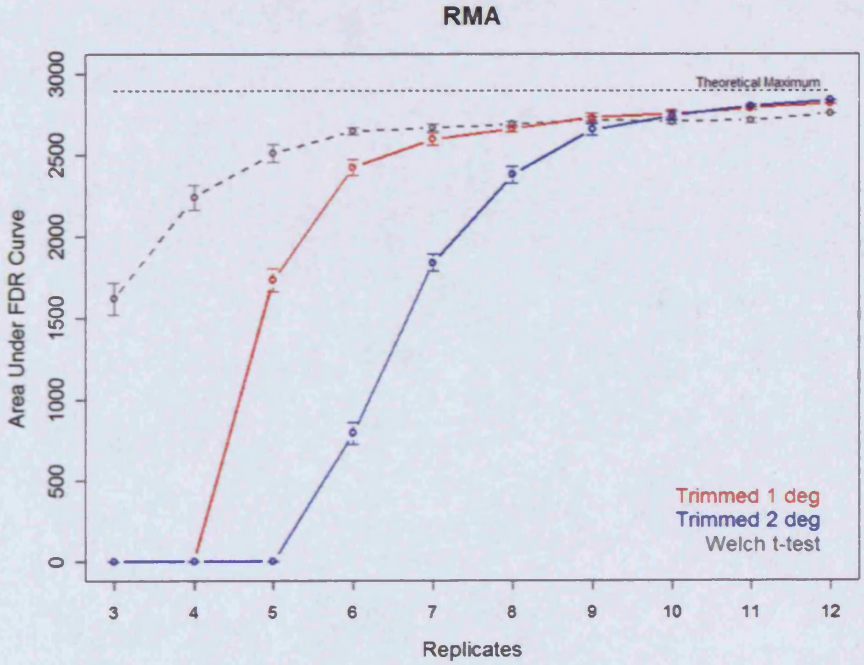

RMA
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
Trimmed 1 deg
Trimmed 2 deg
Welch t-test

Figure 5.2 - continued

c)

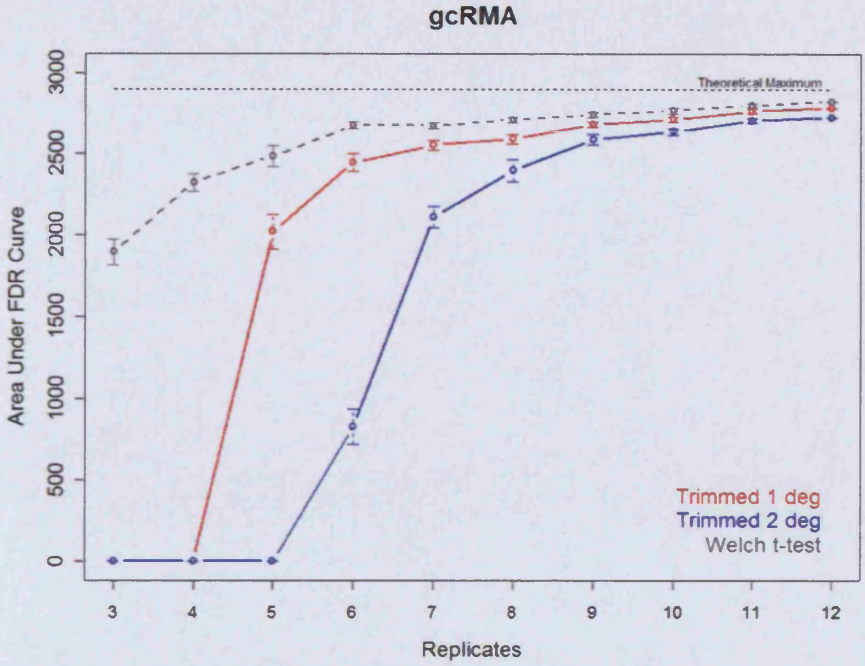
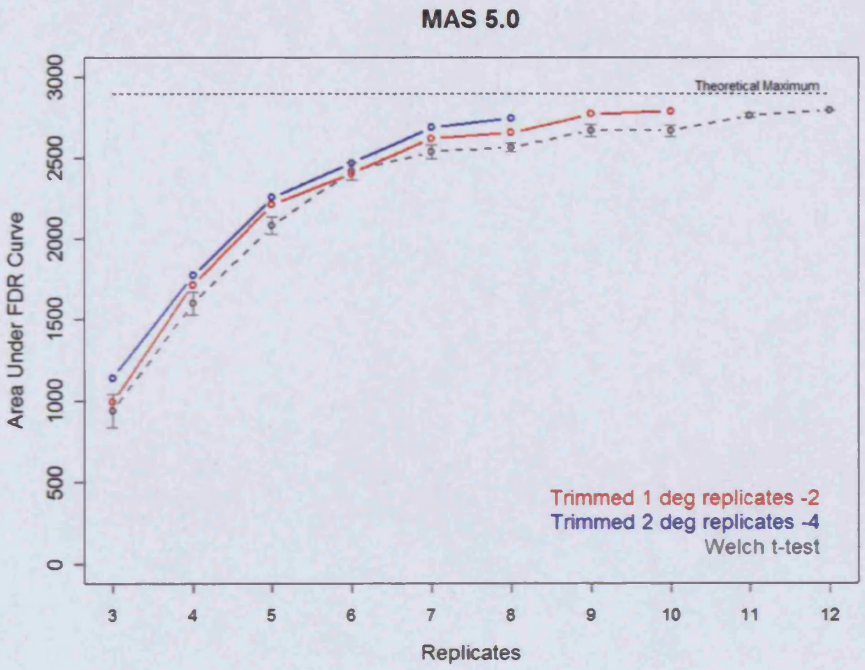


Figure 5.2 – Summary FDR plots showing the results of analysis using first and second degree trimmed t-tests on data from MAS 5.0 (a), RMA (b) and gcRMA (c).

It can be seen that across each of the three expression metrics the power to detect is severely reduced at each additional degree of trimming with some convergence of results towards the higher sample sizes. Comparison to the Welch t-test suggests that the power to detect may be similar to that obtained from a dataset equivalent in size to the resultant dataset after trimming. To explore this effect, the FDR graphs were re-drawn with the data from the trimmed samples shifted along the x-axis to their equivalent sizes after trimming. The resultant plots are shown in Figure 5.3.

Figure 5.3

a)


MAS 5.0
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
Trimmed 1 deg replicates -2
Trimmed 2 deg replicates -4
Welch t-test

b)

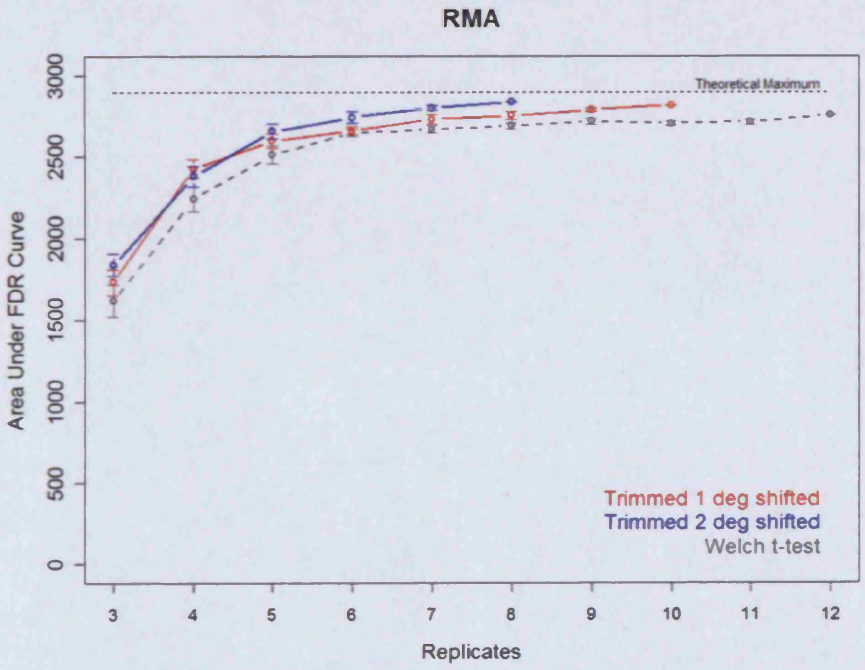

RMA
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
Trimmed 1 deg shifted
Trimmed 2 deg shifted
Welch t-test

Figure 5.3 - continued

c)

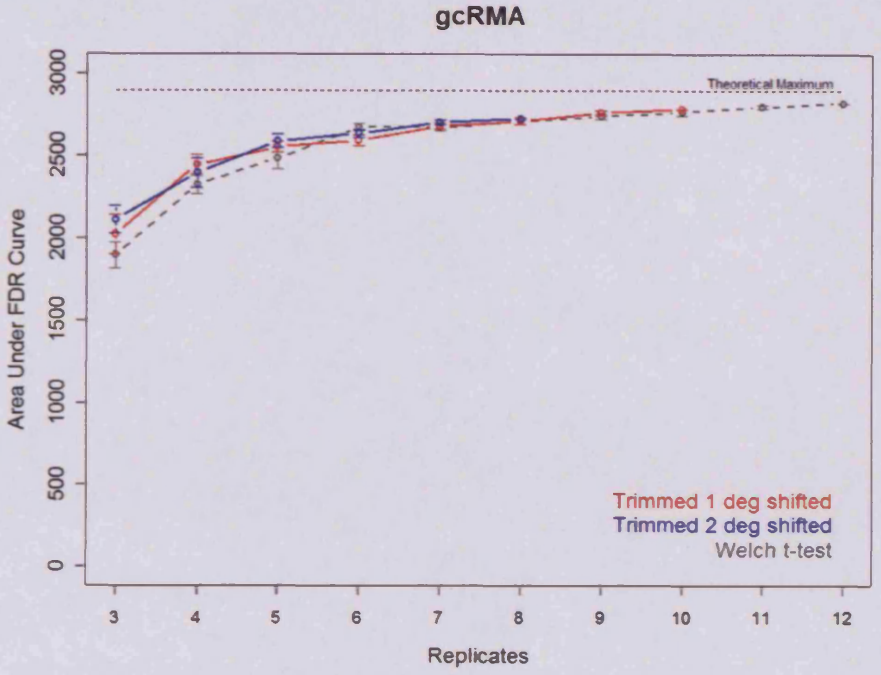


Figure 5.3 – Summary FDR plots showing the results of analysis using first and second degree trimmed t-tests on data from MAS 5.0 (a), RMA (b) and gcRMA (c). Data from the trimmed tests has been shifted to be equivalent to the Welch t-test data.

5.3.3 The Winsorised t-test

In the previous section the trimmed t-test was introduced in which both tails of the dataset are simply omitted. This had the effect of reducing the size of the dataset being analysed and hence the associated power to detect the spiked in samples of the Latin square dataset. Based on Winsor's principle that "All observed distributions are Gaussian in the middle", the Winsorised t-test works by eliminating the outliers and replacing them with data from the edge of the remaining distribution. In one degree Winsorisation, the smallest and largest values are given the values of their nearest neighbours. Thus:

	Resultant Dataset
Original	1 2 3 4 5 6 7 8 9
1 degree Winsorised	2 2 3 4 5 6 7 8 8
2 degree Winsorised	3 3 3 4 5 6 7 7 7

Similarly to the trimmed t-test, for a symmetric distribution, the symmetrically trimmed or Winsorised mean is an unbiased estimate of the population mean. However, the trimmed or Winsorised mean does not have a normal distribution even if the data are from a normal population (Iglewicz and Hoaglin, 1993).

The formula for calculating the t-statistic with Winsorised data is identical to that used for the Welch t-test. However the resultant t-statistic is revised to account for the changes in data, reflecting the amount of original data (h) left in the sample, sized (N).

$$t = \frac{h-1}{N-1} t_w$$

To obtain the p-value associated with a t-statistic from Winsorised data, standard significance table can be employed, with h-1 degrees of freedom (instead of N-1) (Tukey and McLaughlin, 1963).

5.3.3.3 FDR performance of the Winsorised t-test

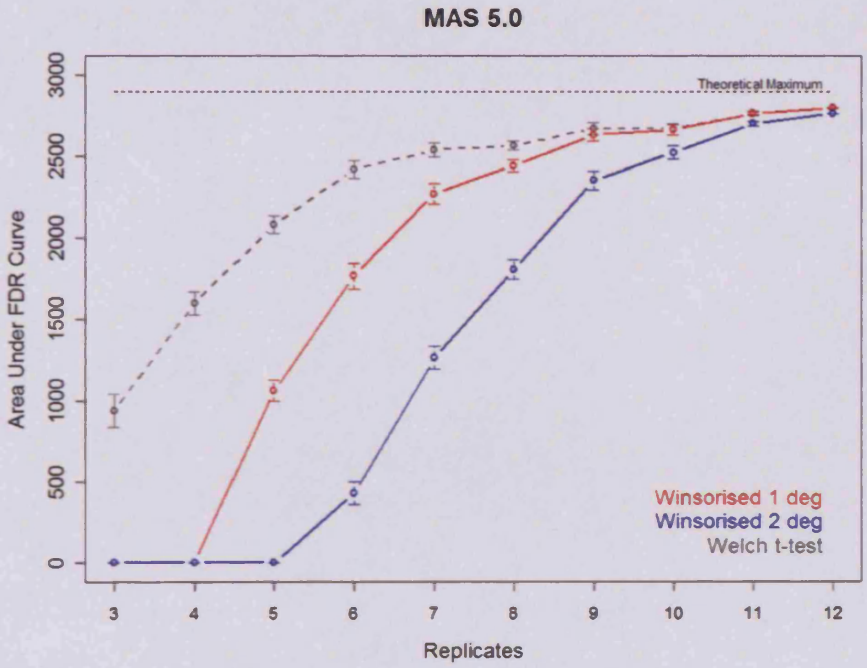
Performance of the Winsorised t-test was assessed using the previously described framework (Section 5.2). Wilcox (2001) asserted that the trimmed-mean approach is desirable if 20 percent of the data are trimmed under non-normal distributions, therefore one and two degrees of Winsorisation was applied to the data from each probe set.

The ability of each test to detect the fifteen spikes present in the 24 chips with a two-fold change in the Affymetrix Latin square dataset was assessed using FDR curves over a range of sample sizes with multiple samples. MAS 5.0 data was used in an untransformed form, RMA and gcRMA data was analysed as provided incorporating the final $\log_2$ transformation. The resultant summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figure 5.4.

As seen with the trimmed t-test, it can be seen that across each of the three expression metrics the power to detect is severely reduced at each addition degree of Winsorisation with some convergence of results towards the higher sample sizes. Comparison to the Welch t-test suggests that the power to detect may be similar to that obtained from a dataset equivalent in size to the resultant dataset after Winsorisation. To explore this effect, the FDR graphs were re-drawn with the data from the Winsorised samples shifted along the x-axis to their equivalent sizes reflecting the amount of original data remaining in the sample. The resultant plots are shown in Figure 5.5.

Figure 5.4

a)


MAS 5.0
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
Winsorised 1 deg
Winsorised 2 deg
Welch t-test

b)

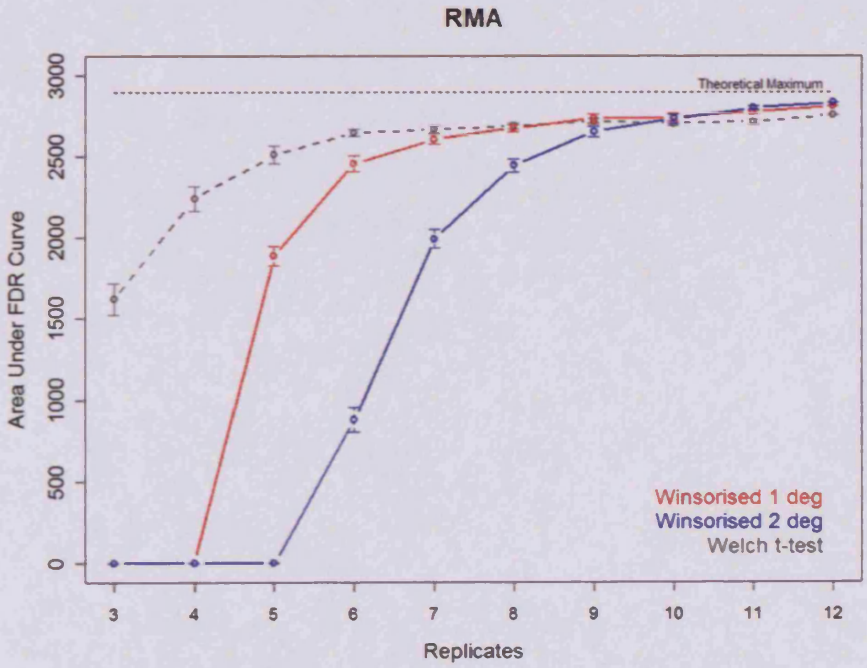

RMA
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
Winsorised 1 deg
Winsorised 2 deg
Welch t-test

Figure 5.4 - continued

c)

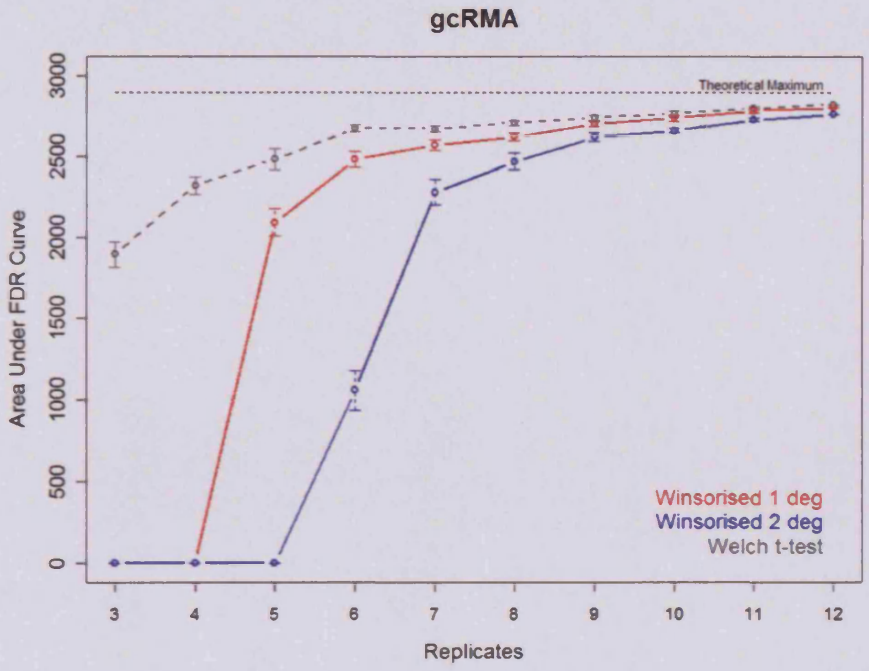


Figure 5.4 – Summary FDR plots showing the results of analysis using first and second degree Winsorised t-tests on data from MAS 5.0 (a), RMA (b) and gcRMA (c).

Figure 5.5

a)

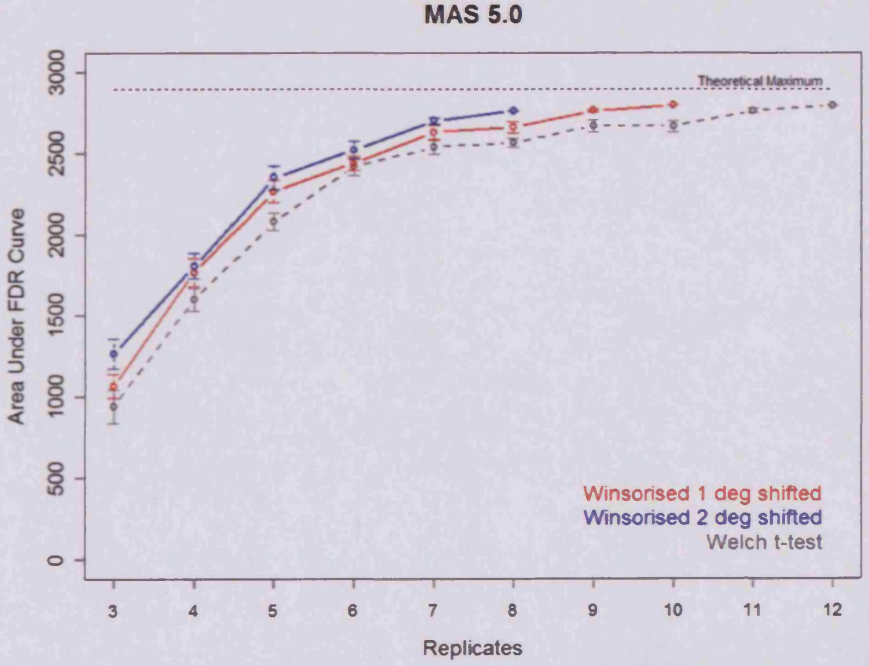
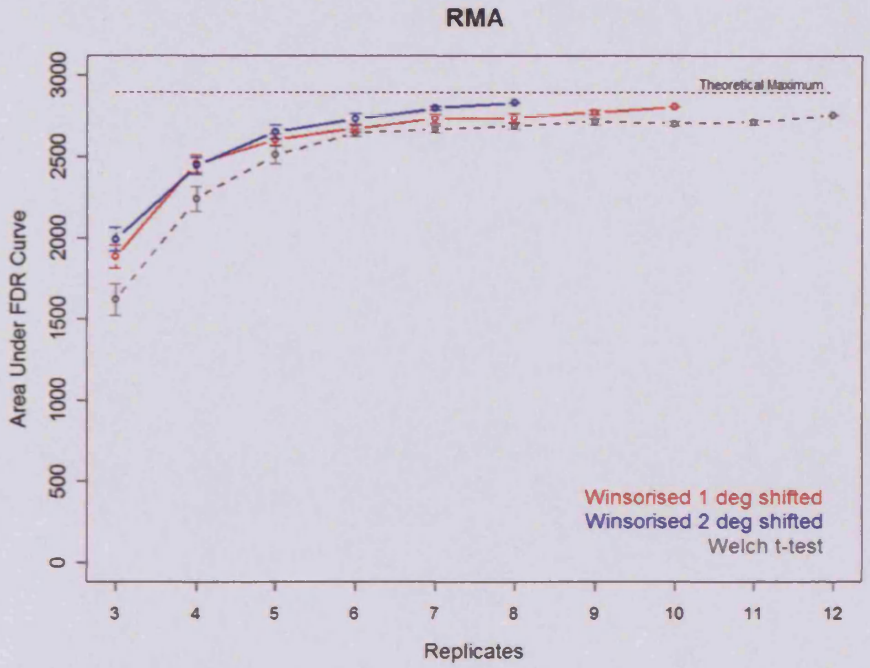


Figure 5.5 - continued

b)



c)

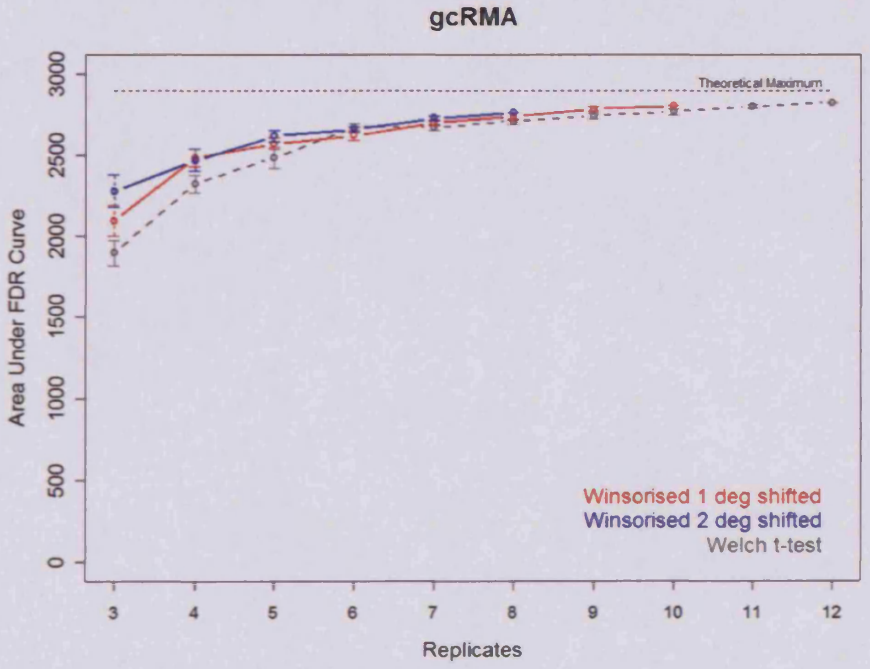


Figure 5.5 – Summary FDR plots showing the results of analysis using first and second degree Winsorised t-tests on data from MAS 5.0 (a), RMA (b) and gcRMA (c). Data from the Winsorised tests has been shifted to represent the amount of original data remaining.

5.3.4 Yuen's t-test

Yuen and Dixon (Yuen and Dixon, 1973) suggested an improved robust test could be achieved by combining the trimmed means and Winsorised variances of a dataset to be used in conjunction with Welch's t-test. The method was further developed by Yuen (Yuen 1974).

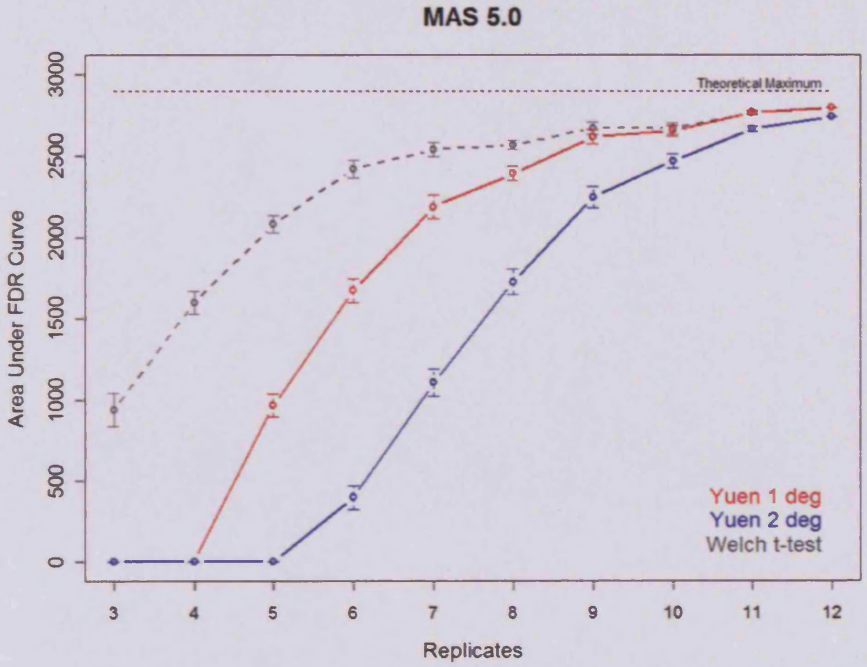
5.3.3.3 FDR performance of Yuen's t-test

Performance of the Yuen's t-test was assessed using the previously described framework (Section 5.2) applying one and two degrees of trimming/Winsorisation. The ability of each test to detect the fifteen spikes present in the 24 chips with a two-fold change in the Affymetrix Latin square dataset was \assessed using FDR curves over a range of sample sizes with multiple samples. MAS 5.0 data was used in an untransformed form, RMA and gcRMA data was analysed as provided incorporating the final $\log_2$ transformation. The resultant summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figure 5.5.

As seen with the trimmed and Winsorised t-tests, it can be seen that across each of the three expression metrics the power to detect is severely reduced at each addition degree of trimming/Winsorisation with some convergence of results towards the higher sample sizes. Comparison to the Welch t-test suggests that the power to detect may be similar to that obtained from a dataset equivalent in size to the resultant dataset after application of the Yuen techniques. To explore this effect, the FDR graphs were re-drawn with the data from the Yuen's t-test shifted along the x-axis to their equivalent sizes after trimming. The resultant plots are shown in Figure 5.7.

Figure 5.6

a)


MAS 5.0
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
Yuen 1 deg
Yuen 2 deg
Welch t-test

b)

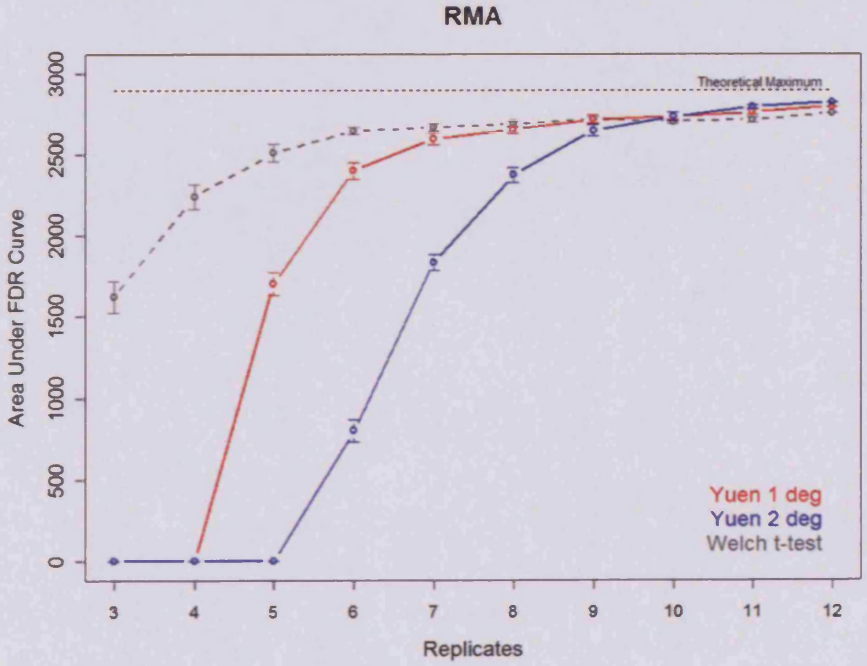

RMA
Theoretical Maximum
Area Under FDR Curve
0
500
1000
1500
2000
2500
3000
3
4
5
6
7
8
9
10
11
12
Replicates
Yuen 1 deg
Yuen 2 deg
Welch t-test

Figure 5.6 – continued

c)

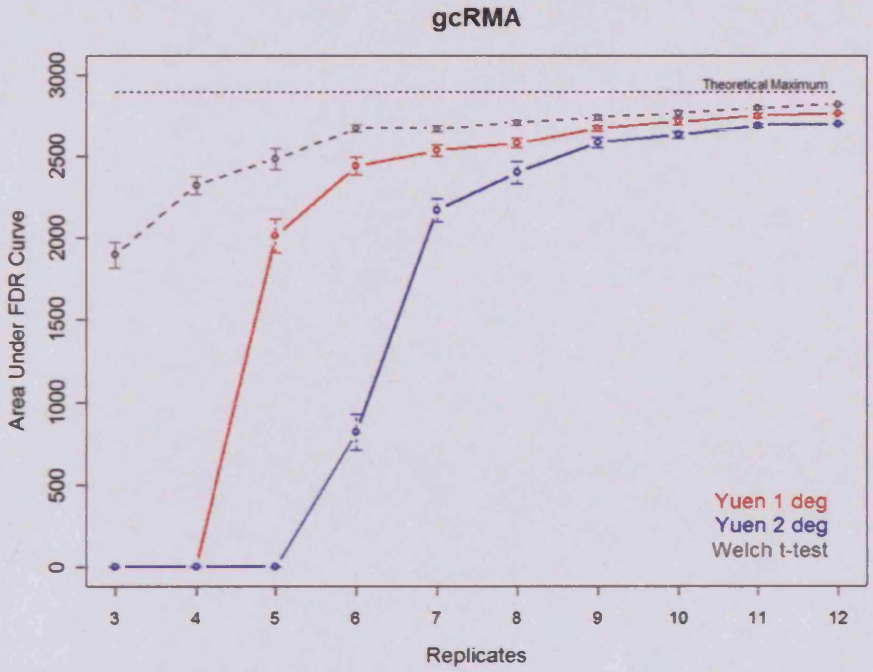


Figure 5.6 – Summary FDR plots showing the results of analysis using first and second degree Yuen's t-tests on data from MAS 5.0 (a), RMA (b) and gcRMA (c).

Figure 5.7

a)

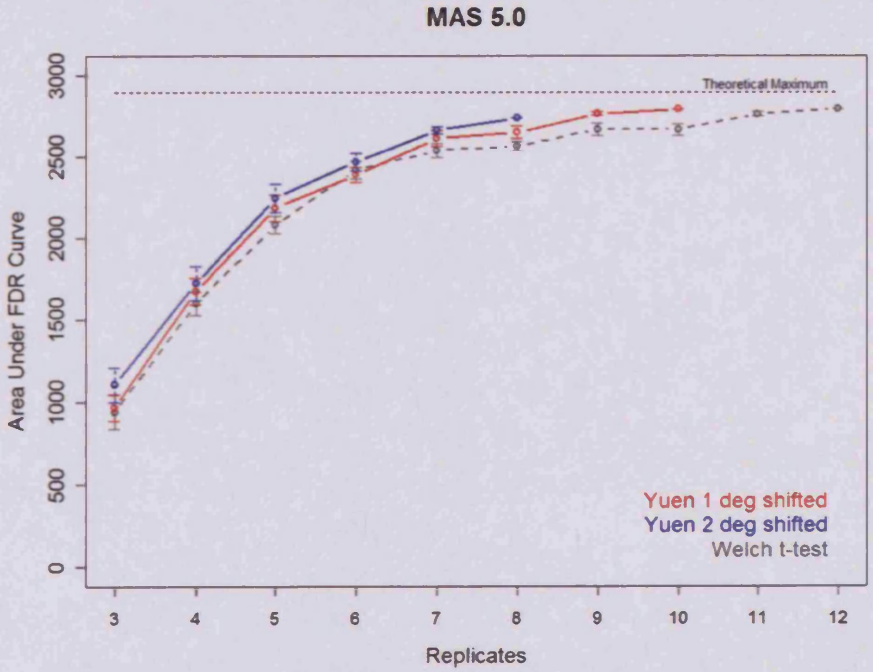
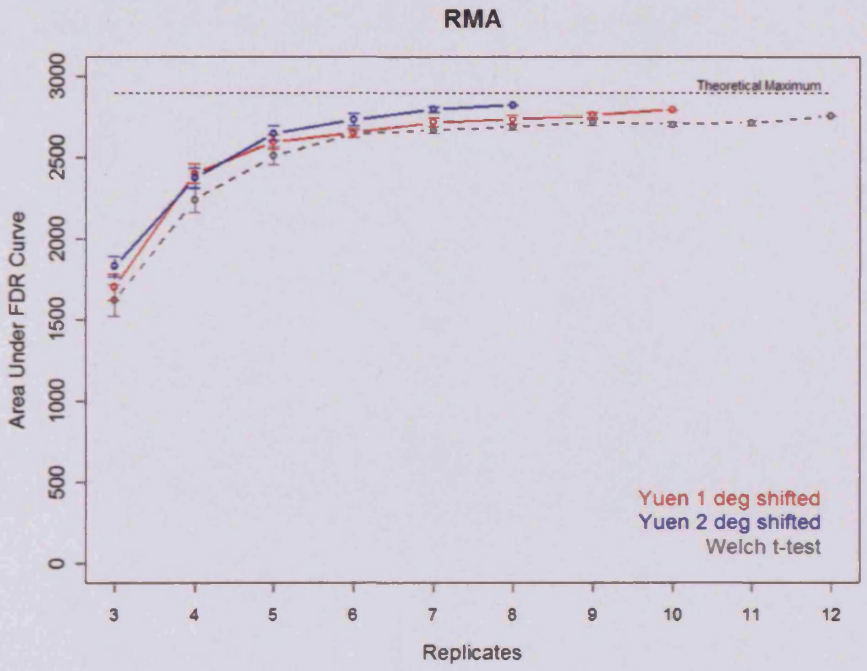


Figure 5.7–continued

b)



c)

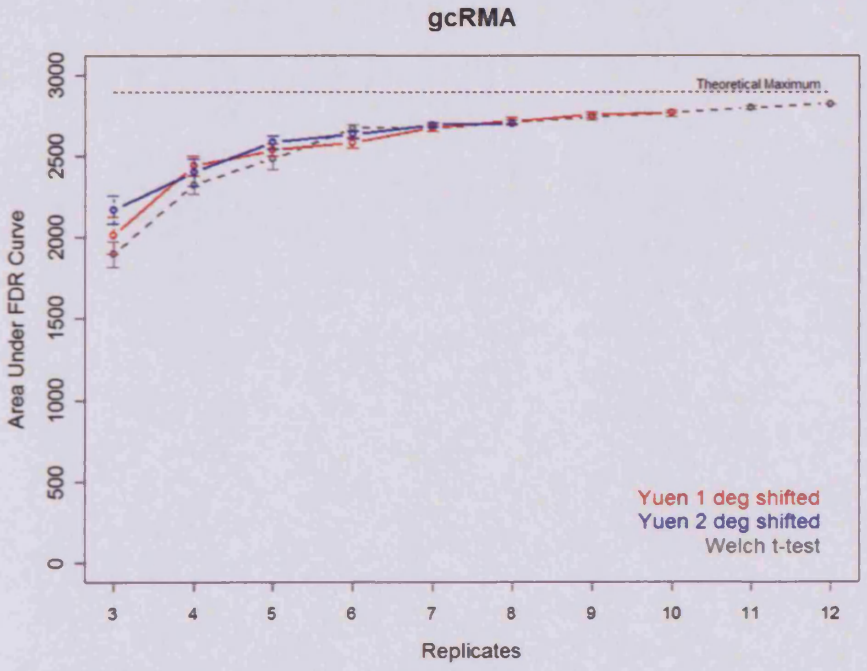


Figure 5.7 – Summary FDR plots showing the results of analysis using first and second degree Yuen's t-tests on data from MAS 5.0 (a), RMA (b) and gcRMA (c). Data from the Yuen tests has been shifted to be equivalent to the Welch t-test data.

5.3.5 Re-sampling based testing

With the exception of the Mann-Whitney test each of the tests examined so far has been a variant of the t-test which makes the assumption the data being drawn from a normal distribution. It was highlighted in Section 2.3 that problems exist with differences in data distributions between probe sets which cause technical problems with these conventional statistical testing methods (Motulsky, 1999).

The calculation of a p-value for each of these variant methods relies on comparison to results obtained from "perfect" datasets which conform to the t-distribution and match the assumptions of a test. Back as far as the 1930's Fisher and Pittman had suggested the development of tests that are not affected by the shape of the data distribution and yield an accurate confidence value. These tests were termed exact tests (Fisher, 1935; Pitman, 1937; Pitman, 1938).

However, it was not until the advent of computers however that these tests were developed and as computing power became more advanced and accessible they became popular research tools (Edginton, 1997). Re-sampling methods depart from theoretical distributions and instead derive inference based on repeated sampling with the same dataset.

Using these methods the tests derive an exact p-value for data instead of looking up from tables derived from asymptotic data. As a result the methods are robust against the effects of outliers within the data.

Within re-sampling there are two differing methodologies of randomisation and permutation. Permutation techniques exhaust all possible outcomes for data sampling re-arrangement and calculate a perfect exact p-value for the data, whilst randomisation tests simulate a large number of possible outcomes to derive a p-value for the dataset. (Edginton, 1997; Good, 2000).

Randomisation testing has been applied to microarray data (Dudoit, et al., 2000), however its use is not common and is generally available as an add-on module for complex statistical environments (e.g. SAS (Mehta and Patel, 1999)).

5.3.5.2 Pilot experiments using randomisation testing

Randomisation testing works to calculate a p-value based on the distribution of the data rather than that obtained from an asymptotic method. In previous Chapter it has been shown that the Welch t-test is the parametric test of choice for determination of differential gene expression, therefore, in developing a randomisation based test, it was the Welch t-test methodology that was chosen.

The methodology behind the method is a simple repetitive process. Initially the t-Welch t-statistic is calculated between the two groups of data (t_o). The data is then regrouped and then re-sampled into two groups of identical size to the initial dataset. The t-statistic is then calculated for the new groups and compared to the initial t-statistic (t_o). If the new value is greater than the original value then this is noted. This process is repeated a set number of times before a p-value is calculated by dividing the number of samplings whose t-statistic was greater than t_o by the number of samplings applied. A flow chart of this process is show in Figure 5.8.

An implementation of this method was written in as a R function and was applied to the 24 chip subset of the Latin Square dataset containing sixteen spiked in probe sets using 1,000,000 random re-sampling steps. Graphs of the p-value obtained from this analysis plotted against the p-value obtained from conventional Welch's t-test are shown in Figure 5.9. Review of this data indicates a good correlation between the two techniques. The p-values for each of the spiked-in transcripts are shown in Table 5.1.

Figure 5.8

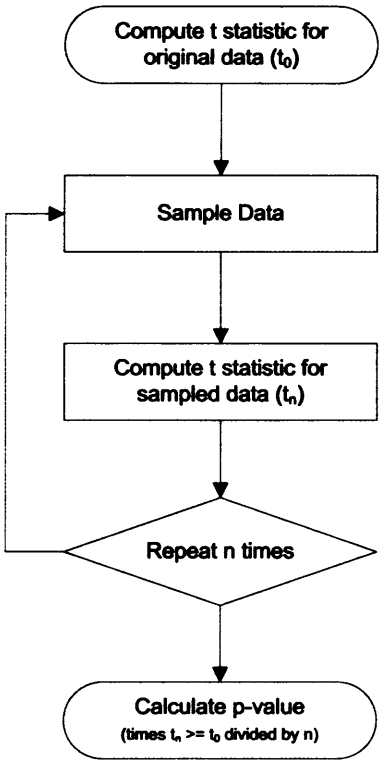
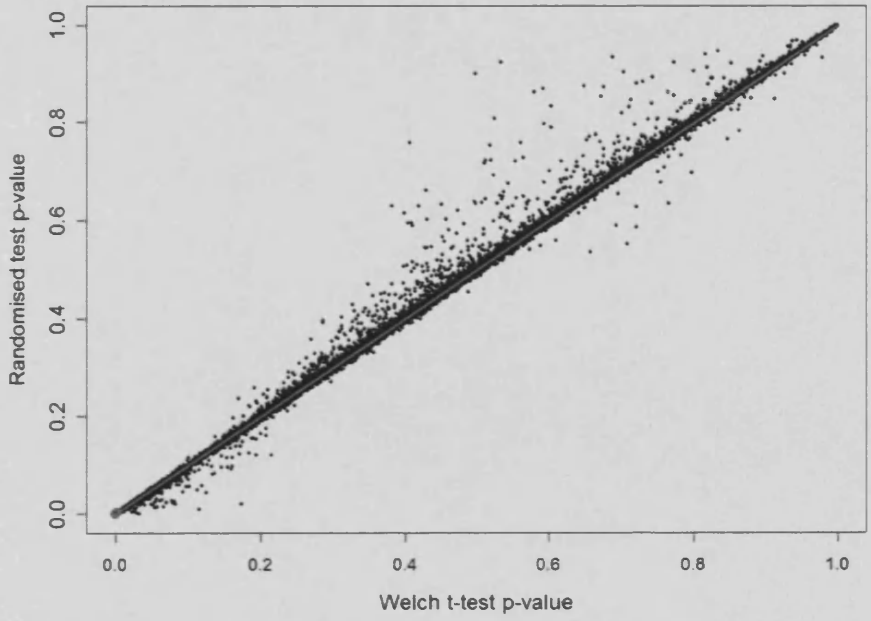


Figure 5.8 – Flow chart of the randomisation t-test methodology.

Figure 5.9

a)


Randomised test p-value
0.0
0.2
0.4
0.6
0.8
1.0
Welch t-test p-value

b)

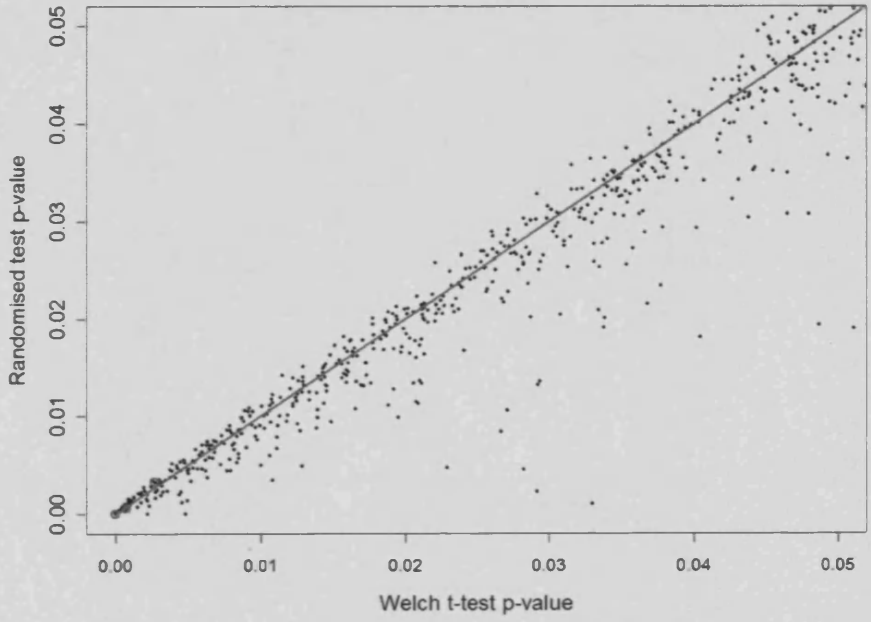
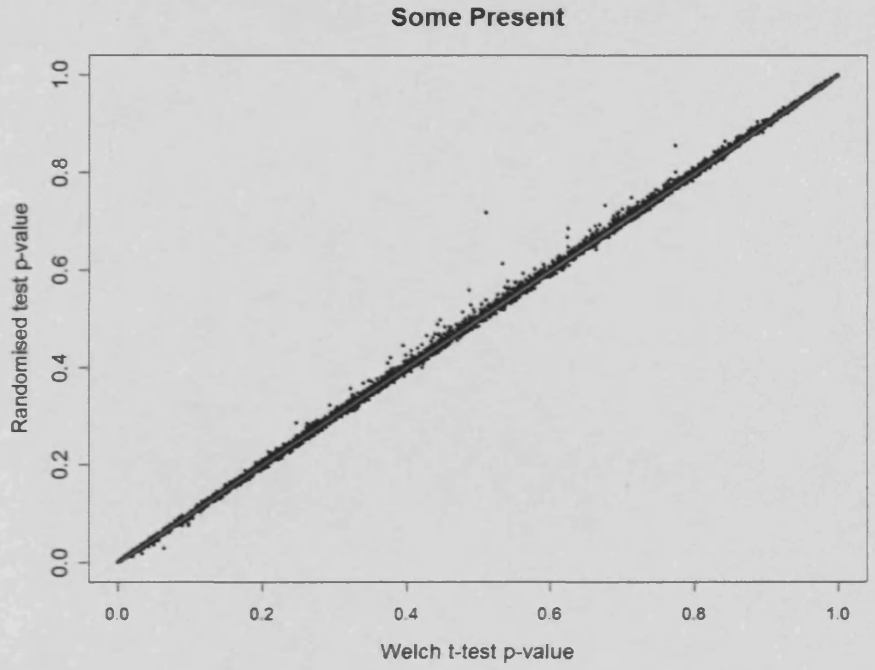

Randomised test p-value
0.00
0.01
0.02
0.03
0.04
0.05
Welch t-test p-value

Figure 5.9 - continued

c)



d)

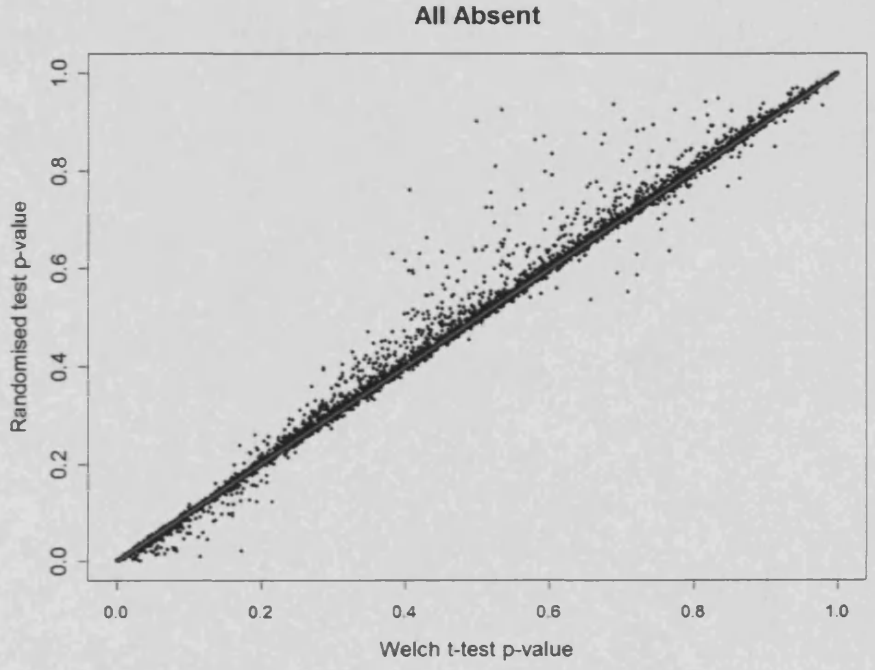


Figure 5.9 – Plots showing p-values obtained from randomised re-sampling based t-test plotted against those from the Welch t-test. Figure 5.9.b shows a expanded section of Figure 5.9.a with p-values from 0 – 0.05. Red circles around data points indicate data from the 16 spiked-in transcripts within the dataset. Figures 5.9.c and 5.9.d show the data from 5.9.a split according to MAS 5.0 detection calls.

Further review of the data identifies a group of genes with smaller p-values that score more significantly using the randomisation test (Figure 5.9.b). To investigate this effect further, the data according to the detection calls given as part of the MAS 5.0 analysis, with one group representing the probe sets where the value across all arrays are called as absent, and another containing probe sets where one or more arrays are called as marginal or present (an identical split to that used in Section 2.3.2 looking at data distributions).

Review of the relationship between these two groups of data and the Welch t-test data indicates that the probe sets with some probe sets called present, correlate very tightly with the results from the Welch t-test. Data with arrays called absent also shows good correlation, however it is this dataset which contains probe sets with large differences in p-values between the two methods.

Table 5.1

Spike	Randomised p-value	Welch t-test p-value
407_at	0.000001	1.99E-08
546_at	$p < 1^{-6}$	1.15E-07
1708_at	0.000558	7.66E-04
1024_at	$p < 1^{-6}$	5.20E-10
1091_at	0.000001	9.37E-07
1597_at	0.003308	2.68E-03
33818_at	$p < 1^{-6}$	9.64E-08
36085_at	0.000001	1.22E-09
36202_at	$p < 1^{-6}$	3.26E-14
36311_at	$p < 1^{-6}$	1.15E-05
36889_at	0.000006	1.40E-06
37777_at	0.00005	1.78E-05
38734_at	0.000009	3.99E-06
39058_at	0.000018	1.30E-05
40322_at	$p < 1^{-6}$	6.86E-09
684_at	0.000001	5.25E-11

Figure 5.1 – Comparison of p-values for spiked-in transcripts between randomised t-test (1,000,000 step randomisation) and the Welch t-test.

5.3.5.3 How many re-sampling steps are required?

In the initial testing of the method, an arbitrary figure of 1,000,000 was chosen for the number of random re-sampling steps applied to the data. As the technique is very computationally intensive it is essential to choose a figure for samplings that will produce statistically sensible and relevant results, but within a practical period of time. With the number of datasets that a microarray experiment produces, it is not feasible to run all the possible data permutations within a reasonable time (the 24 chip dataset has 5.2×10^{23} distinct permutations of the data)

To address this, a modified version of the program was coded which outputted the p-value after each random re-sampling step. Review of the data on a probe set by probe set basis indicated that there was convergence on a approximate p-value after a certain number of re-samplings, the number of steps before this convergence is seen being dependant on the final p-value. It is observed that probe sets resulting in a larger p-value converge on a single value much faster than those that have a smaller value. The smaller the resulting p-value the greater the number of permutations required to converge on this point. Example plots are shown in Figure 5.10.

It can be reasoned that for a gene to have a large p-value the observed statistic must exceed the actual statistic on many occasions. In contrast, probe sets that yield a smaller p-value will only exceed this threshold a fraction of the number of times, so as the p-value decreases the number of re-samplings between observing a significant result increases. This suggests that the run time for a full microarray dataset can be reduced by calculating the number of permutations required dependant on expected p-value.

5.3.5.4 Application of a binomial error model to reduce re-sampling steps

Research into potential methods for estimating the number of re-samplings required lead towards exploration of error modelling using standard binomial theory (Mehta, 1999). Binomial theory enables a confidence measurement to be made using the number of samplings (n) and the resultant p-value (p):

$$Error = \sqrt{\frac{p(1-p)}{n}} \qquad n = \frac{1-p}{\varepsilon^2 \cdot p}$$

Rearrangement of this formula gives a method for estimating the number of re-sampling steps required (n) if the researcher is willing to pre-define the confidence (ε) they wish to have in the resulting p-value (p).

Figure 5.10

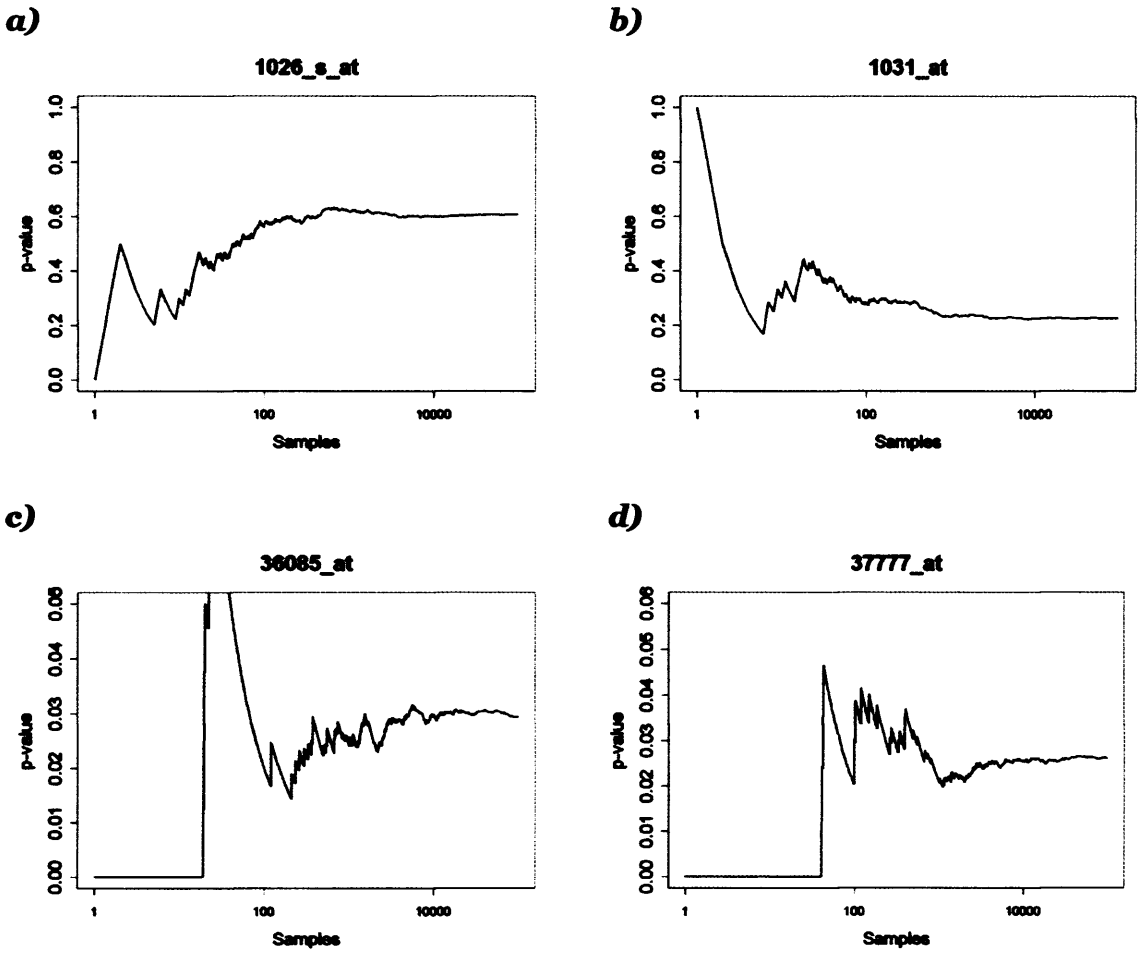


Figure 5.10 – Convergence of p-values from probe sets resulting in a large (5.10.a), mid (5.10.b) and small (5.10.c and 5.10.d) p-values over a range of samplings during the application of a 100,000 step randomisation t-test.

An example of the relationship formed between samplings and p-value is shown in figure 5.11 with error rates set at 0.1 and 0.05. Integrating this new model into the randomised re-sampling methodology gives a revised analysis which is represented in the flowchart in Figure 5.12.

Figure 5.11

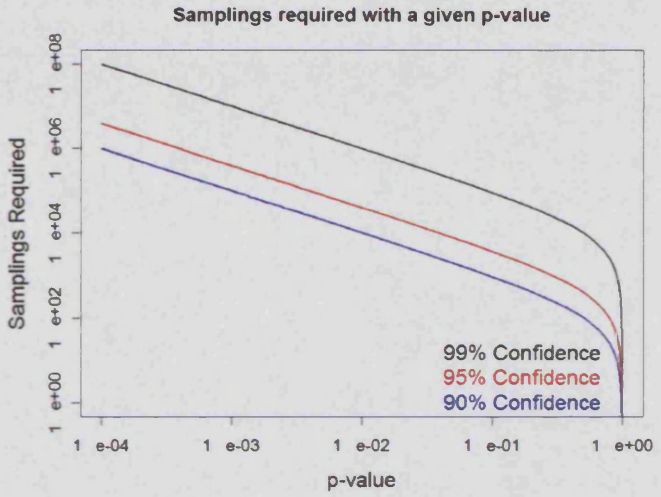


Figure 5.11 – The relationship between p-value and samplings required following the application of the binomial error model at 90% and 95% confidence in the resultant p-values

Figure 5.12

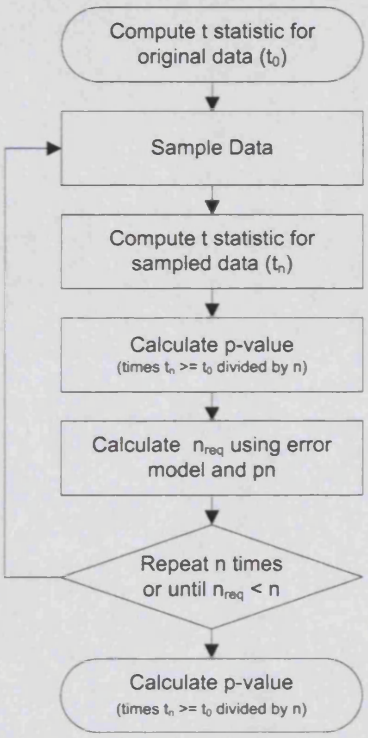


Figure 5.12 – Revised flowchart incorporating error model.

5.3.5.5 Testing the revised model

The efficiency of the error model was tested by implemented the revised functions into the code modified in Section 5.3.5.3 which outputted the p-value at each stage of the process. Figure 5.13 shows the results indicating that the error model yields near identical p-values (identical would no be expected due to the randomness of sampling) and terminated the run of the program at a point where it would appear the p-value has stabilised around a certain value.

5.3.5.6 Computationally defining the model

It has previously been commented that re-sampling based techniques are computationally intensive; indeed this is why after their conception is was many years before their usefulness and uptake into research was possible (Edgington, 1997; Good, 1999). In the case of the pilot testing undertaken in Section 5.3.5.2 the running of 1,000,000 permutations on 12545 probe sets took several days when run on a P4 3.0GHz machine within the R environment. To make re-sampling a useful research tool, avenues to reduce this time warrant exploration.

It is clear that the error model has the potential to reduce this time, by eliminating the need to run many samplings for probe sets which are likely to result in a high p-value. In addition explorations were made into the most computationally effective methods of implementing the methodology.

Investigations were undertaken in implementing the code in a variety of programming languages (Visual Basic, Perl and R), but these higher level interpreted languages were markedly superseded in speed by the lower-level compiled languages. The methodology was therefore re-coded into compiled C, taking at tab-delimited text files as it input. This solution was found to be the fastest implementation of the method.

In addition to the faster implementation of the basic algorithm, it was found that the revised algorithm incorporating the error model imposed an additional overhead to the execution of the method due to the requirement to calculate a p-value and the error model following each sampling. This issue was overcome by the application of a stepped model to the error model. Following each run with a set number of samplings, the data which could be removed according to the error model was removed and the p-value stored before re-analysis with a higher number of samplings. This was applied using a stepped approach up to a maximal number of samplings.

Figure 5.13

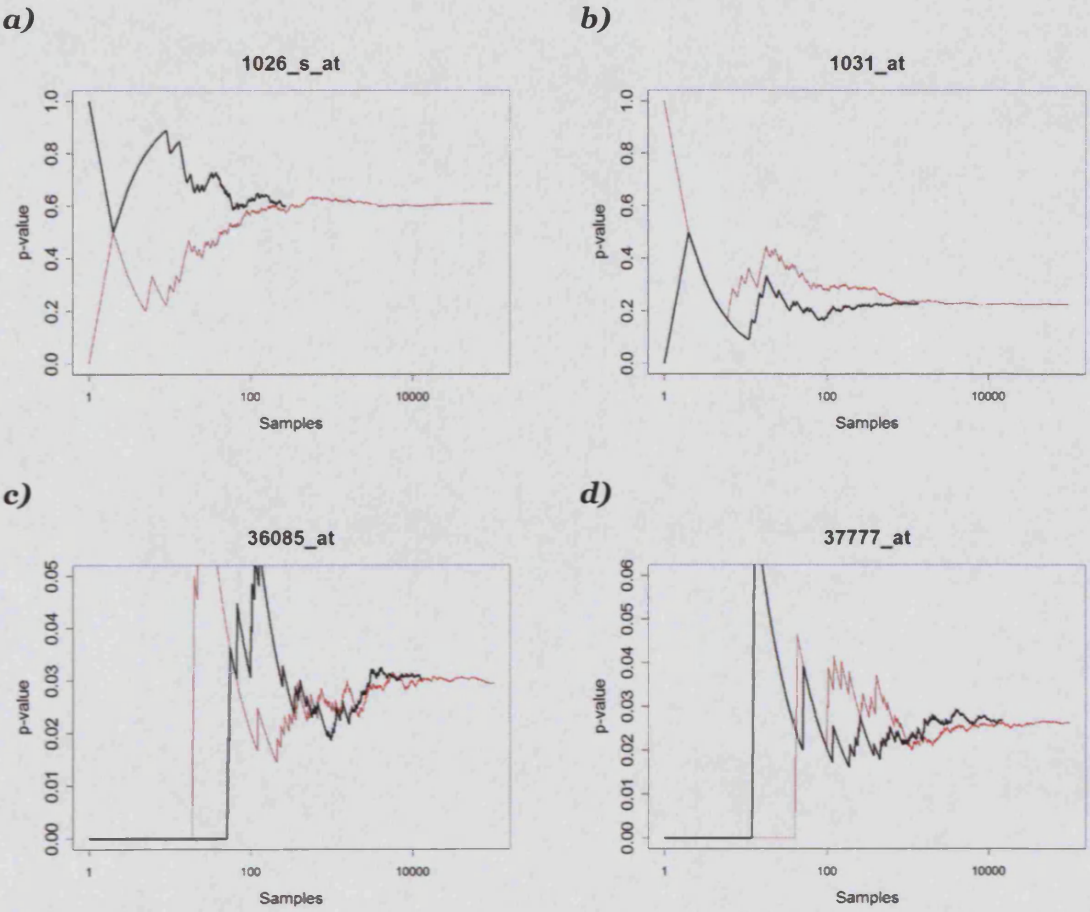


Figure 5.13 – Convergence of p-values from probe sets resulting in large (5.13.a), mid (5.13.b) and small l(5.13.c and 5.13.d) p-values over a range of samplings. The red lines show an example of data run without the error model for 100,000 randomisations (Figure 5.10) and the black lines show the same dataset run with the error model, terminating when confidence in the p-value was achieved.

5.3.5.7 FDR performance of the randomised re-sampling based t-test

Performance of the randomised t-test was assessed using the previously described framework (Section 9.3) using the stepped approach introduced in Section 5.3.5.6 starting at 10,000 samplings and increasing to 1,000,000 in steps of 10,000. The error level was set at 5%. Due to the computational requirements of this analysis only a single sample at each sample size was run for this test in contrast to the twenty run for the other tests.

The ability of each test to detect the fifteen spikes present in the 24 chips with a two-fold change in the Affymetrix Latin square dataset was assessed using FDR curves over a range of sample sizes with multiple samples. MAS 5.0 data was used in an untransformed form, RMA and gcRMA data was analysed as provided incorporating the final $\log_2$ transformation.

Initial review of the data suggested a similar FDR response to that obtained from the distribution independent Mann-Whitney test. Summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figure 5.14 incorporating the information from the Welch t-test and Mann-Whitney test for comparison.

In general the performance of the randomisation re-sampling based test was very similar in response to that from the Mann-Whitney test. At lower sample size (3-5 per group) there was very little power to detect, power which increased over the middling sample sizes until at those over 9 per group, a similar response was seen for all three tests.

Figure 5.14

a)

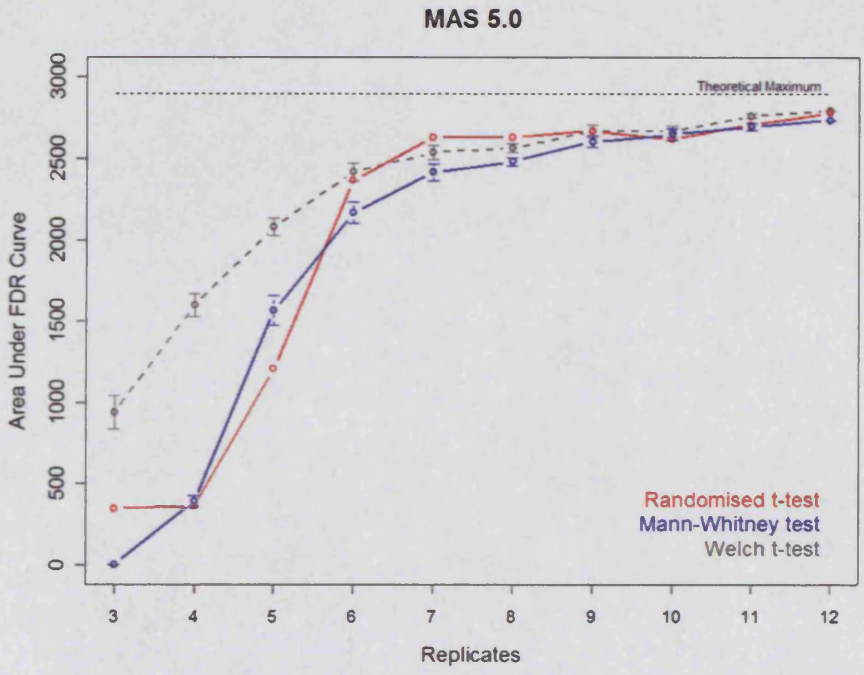
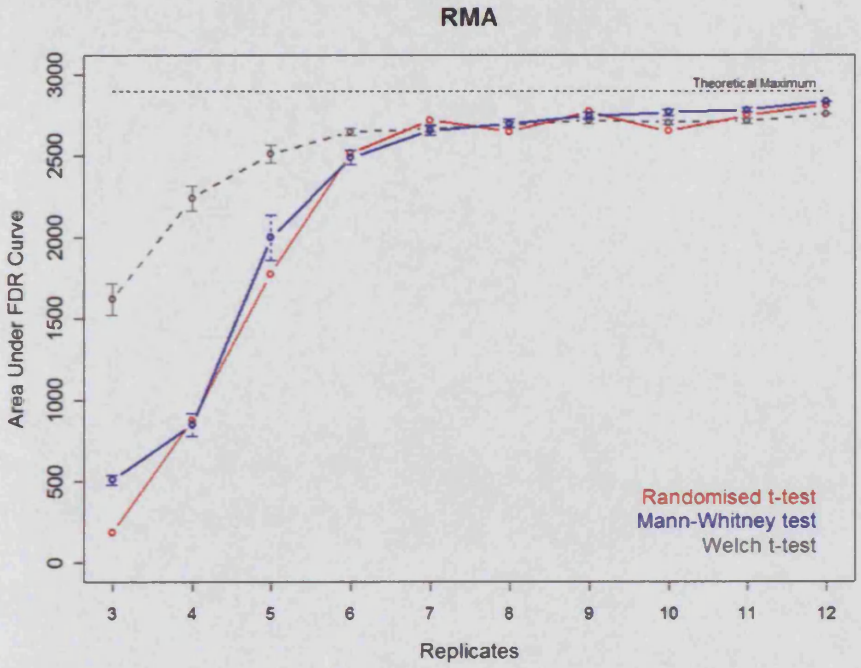


Figure 5.14 – Continued

b)



c)

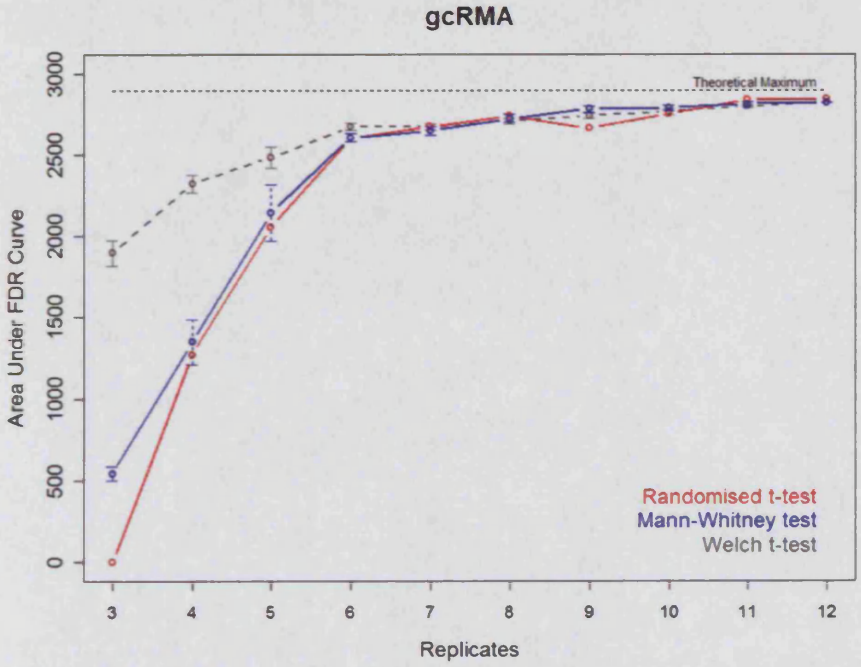


Figure 5.14 – Summary FDR plots showing the results of analysis using randomised re-sampling based t-tests on data from MAS 5.0 (a), RMA (b) and gcRMA (c). Additional lines from the Welch t-test and Mann-Whitney test are show for comparison.

Diaconis and Efron (Diaconis and Efron, 1983), recommend that when the sample size is small and does not conform to the parametric assumptions, re-sampling is recommended as a remedy. In contrast, the data presented here suggest that when applied to microarray data there is no power advantage to the randomisation test, with the Mann-Whitney test taking a fraction of the time to compute.

5.4 Discussion

In the introduction to this Chapter, the idea of outlying data points within Affymetrix microarray datasets was introduced.

5.4.1 Robust variants of the t-test

Building on the ideas suggested by Iglewicz and Hoaglin (Iglewicz and Hoaglin, 1993) for robust estimators, the t-test was reviewed and more robust measures of location and spread implemented within the t-test methodology. The resultant options for location (mean and median) and spread (standard deviation, median absolute deviation about the median) resulted in three variants of the Welch t-test, three of which can be considered more robust.

The technique with the most power to detect was the combination of mean and standard deviation (a standard Welch t-test). Substitution of the median for the mean still yielded good results; however, results indicated slight loss of power in comparison. Substitution of the standard deviation with the more robust median absolution deviation from the median significantly reduced the power of the test to determine the fifteen spikes in the dataset, and further substitution of the median for the mean reduced this further.

The measurement of the variance (spread) of the data would appear to be having a significant effect on the power of the test. This correlates with statements by Baldi and Long (Baldi and Long, 2001) who comment that estimating variance is the major limitation of the t-test at smaller sample sizes.

5.4.2 Trimmed, Winsorised and Yuen's t-tests

The robust variants of the t-test work by attempting to accommodated outliers into the statistical analysis. A differing approach is the elimination of these outlying data points from the dataset. Wilcox (Wilcox, 2001) and Lix & Keselman (Lix and Keselman, 1998) comment that when multiple factors such as non-normality and heterogeneity of variance occur, this has the effect of inflating the Type I error rate (false positives reported). They suggest the application of robust methods such as the trimmed means and Winsorised variances are applied.

The trimmed, Winsorised and Yuen's t-test all remove a predetermined amount of data from the tails of the data set working under the assumption that only one or two erroneous data points exist in the data. The Winsorised method replaces the lost values with others drawn from the new tails of the dataset, and Yuen's test combines trimming with Winsorisation.

Each of the above methods performed poorly in the detection of the spikes within the Latin square dataset. When the data was considered to represent the sample size of the data following the application of the trimmed techniques (equivalent to shift along the x-axis of the FDR graphs), power was slightly improved in comparison to the Welch t-test, however to achieve this much data has been discarded.

Removal of outliers would appear to have some success in the improvement of power to detect if data is pre-filtered and then compared to equivalent sample sizes. However the application of these techniques can be view as akin to throwing data away, with a large cost burden for each microarray experiment. The choice of a test which can overcome outliers in a dataset is a more preferable approach, combined with better experimental design to enable the inclusion of all data collected as part of an analysis. This observation supports research by Keselman and Zumo (Keselman and Zumbo, 1997) who found that the nonparametric approach has more power than the trimmed-mean approach.

5.4.4 Randomised Re-sampling based t-test

Randomised re-sampling presents an interesting compromise between the distribution issues encountered with the standard Welch t-test, without the need to choose a non-parametric approach and replace data with rank information. By re-sampling an pseudo exact p-value can be generated representing the distribution of the probe set under test.

Initial pilot testing of the technique showed high correlation between the results of the randomised method and the Welch t-test (Figure 5.9). This correlation can be used to support the application of the Welch t-test to microarray data; if the randomised test calculates a p-value according to the actual data distribution and the Welch t-test yields similar values then this suggests that the data follows a near normal distribution, which is a required assumption of the Welch test.

Application of the randomised method for the detection of spikes in the Latin square dataset yielded similar results as those obtained from the Mann-Whitney test. In comparison to the Welch t-test, the randomised test showed limited power at low sample sizes, slightly less power at middle sizes, and equivalent power at sample sizes over 9 per group.

Whilst this technique gives reasonable results, the computational and time overheads required for its application question the practicality of application in a research setting. If near identical results can be obtained using a test that is more robust against outliers and distribution issues in lesser time, then the Mann-Whitney test is the test of choice.

Chapter Six
Application of a Bayesian Framework

In this Chapter the problem of analysis on datasets with a small sample size is revisited with the application of a Bayesian framework applied to exploit trends in the data. Section 6.1 introduces the Bayesian methodologies and key work applying these techniques to microarray data. Section 6.2 details the modifications made to the previously described analysis framework (Section 3.2) to incorporate Bayesian methods for the identification of the spiked-in data in the Affymetrix Latin square dataset. Section 6.3 reviews the methods used to optimise the tests using spike detection in the Latin square data and compares the results to other statistical techniques. Section 6.4 discusses the results and observations about the application of Bayesian methods to Affymetrix microarray data.

6.1 Introduction

In previous Chapters the issue of small sample size has been a recurring issue influencing the options available for the application of statistical testing, specifically issues of estimating variance in the classic t-test for differences between the means. This lack of sample size results in poor estimates of within-treatment variance and hence, a corresponding poor performance of the t-test (Baldi and Long, 2001). Historically these tests have been developed for a single analysis between two groups of data, however in microarrays many thousands of these tests are run in parallel. Although there is no complete substitute for experimental replication to allow the sample results to represent those of the population, Bayesian methods aim to improve the confidence that can be achieved from an analysis by incorporating information regarding similarities between parallel observations and thus reduce the need for excessive replication.

6.1.1 Introduction to Bayesian methods

Named after Thomas Bayes (an English clergyman and mathematician), Bayesian data analysis combines Bayesian probability theory with statistical data analysis techniques to make predictions about future events based on our current information (Eddy, 2004). A single microarray comprises many parallel pseudo-replicated experiments, with some genes being represented by more than one observation, and similarities in the resultant data between other probe sets reporting distinctly different transcription profiles.

It has previously been noted that there is a relationship between the mean expression level and variance calculated for a dataset (Naef, et al., 2002). It is this observation that has been suggested as the basis for a Bayesian method using information provided by probe sets of similar expression levels to attain pseudo-replication of the experiment.

6.1.2 The Baldi and Long Bayesian Framework

In Section 4.1.2 the factors that affect the power of a statistical test to detect differences were highlighted and include variability of the population (Hatfield, et al., 2003), the desired detectable differences and an acceptable error rate (Wei, et al., 2004). In their 2003 paper, Baldi and Long introduced a framework for differential gene expression based on Bayesian estimates of a gene's variance incorporated into the Welch t-test to improve detection power.

The model is based on the observation that mean expression level and variance are related within many microarray datasets (Figure 6.8). Variance is calculated by combining the given variance for a probe set with a prior estimate of the expected variance for that gene - calculated from the average of the variance of other genes of similar expression level. These two figures for variance (the probe set variance and the local variance) are then combined according to a weighting factor. The weighting factor is controlled by the researcher and is a tuneable parameter dependent on how confident the experimenter is that the background variance of a closely related set of genes approximates the variance of the gene under consideration.

The analysis steps for the framework can be described in the following workflow description:

i) For each probe set calculate the mean, standard deviation (SD) for each group of data

ii) Separately for each group of data:

   a. order the data by mean expression value

   b. for each probe set, calculate a Bayesian SD measurement from a sliding window either side of the current probe set.

iii) Blend this Bayesian SD value with the probe set SD value

iv) Calculate the t statistic (assuming inequality of variance) using the probe set means and the Bayesian SD measures.

v) Calculate the Bayesian degrees of freedom

vi) Calculate a p-value using the Bayesian-t and degrees of freedom.

6.1.4 Application of the Bayesian framework to Affymetrix Data

Baldi and Long demonstrate the effectiveness of their algorithms using an E. coli dataset with four replicates per group from microarray data obtained from nylon membranes containing 4,290 probes (Arfin, et al., 2000; Long, et al., 2001). Review of the workflow in Section 6.1.3 indicates two main user tuneable parameters, the size of the window use for local variance estimation, and the confidence parameter used in the blending stage of the algorithm. In the initial publication (Baldi and Long, 2001), no guidelines for the tuneable parameters of the methodology are given, and in the parallel publication (Long, et al., 2001) little guidance is given as to the requirements for parameter selection.

In comparison to the dataset used to develop the framework, Affymetrix GeneChip data comprises a different biological technology and undertakes many more pseudo-parallel experiments (up to 47,000 probe sets in the current generations of GeneChips). Whilst the Bayesian framework has proven a popular tool within the cDNA microarray community, few publications exist regarding the application of this framework to Affymetrix data, and those that exist, do not details of the parameters selected (Hatfield, et al., 2003; Li, et al., 2005; Saidi, et al., 2004; Tong, et al., 2001).

It is therefore of interest to validate the application of the Bayesian framework to Affymetrix gene chip data and explore the optimal parameters for the local variance window size, differing options in the application of the blending parameter and alternative methods for the estimation of the localised variance.

6.2 Technical Methodology

Using the previously described methodology (Section 9.3.5), an integrated analysis script was written in R which took a subset of the Affymetrix U95A Latin Square dataset and produced a series of 20 replicates at a range of 3 to 12 chips in each of two groups using the MAS 5.0, RMA and gcRMA expression metrics. Next the script took each of these indexed matrixes, extracted the relevant data and then runs a variety of variants of the Bayesian framework (allowing tuning of the user alterable parameters) with twenty replicates over a range of ten sample sizes.

Figure 6.1

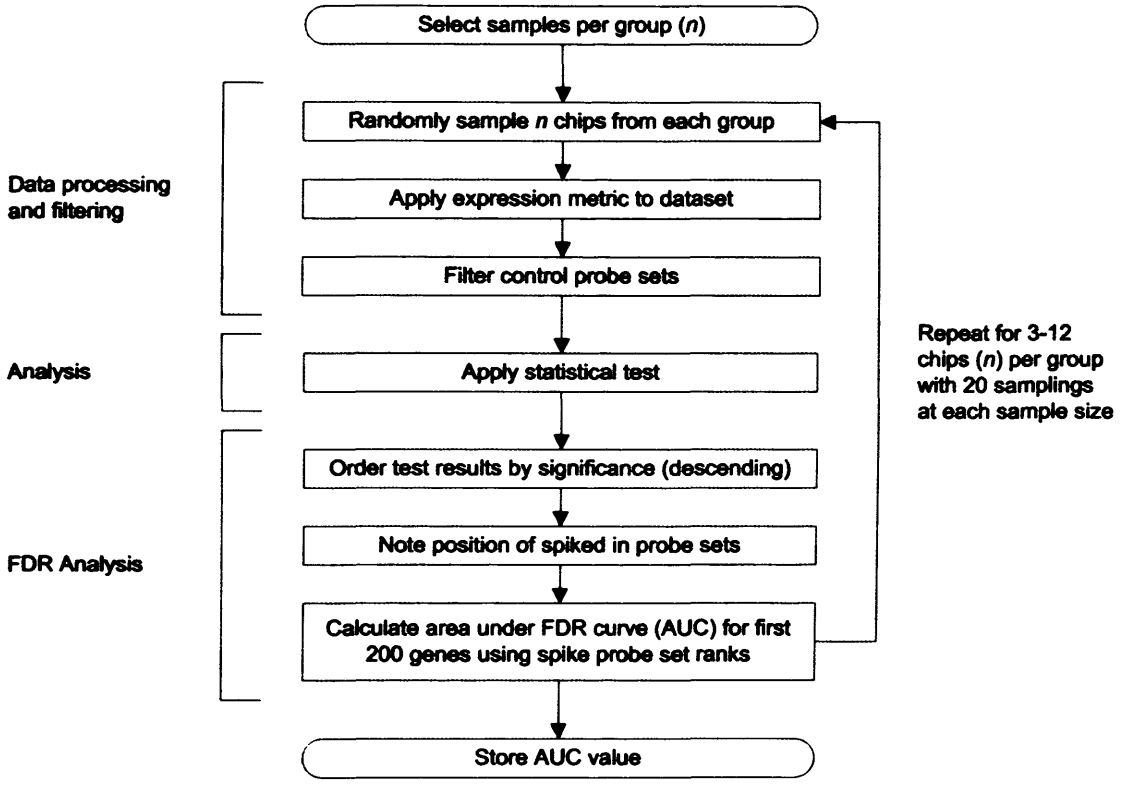


Figure 6.1 – Flow diagram of technical methodology for examination of power to detect performance of each variant of the Bayesian framework.

The output for each test was fed into an FDR function which extracted the location of the spikes from the ordered p-values and calculated the area under the FDR curve as previously described. The output for each test was exported into a summary matrix containing each of the 200 AUC values calculated, grouped by sample size. Summary plots were produced charting the average area under the curve (with error bars reporting the standard error), across the range of sample size under consideration.

Full details of the technical methodology are given in Section 9.6.

6.3 Results

In previous Chapters, the response to a test applied against a variety of different expression metrics has been considered. However, tuning the many parameters of the Bayesian framework with a variety of values against three expression metrics would yield more data than could sensibility be interpreted. Thus, in an attempt to minimise the variables under consideration, optimisation was undertaken using MAS 5.0 data. Following the identification of optimal values for each tuning parameter with this dataset, comparisons can made the RMA and gcRMA models using identical analysis settings.

6.3.1 Defining an optimal Bayesian window size

It has already been commented that the dataset used for the development of the Bayesian framework and used to illustrate its effectiveness at improved power of detection in differential gene expression between two groups was an E. coil dataset on nylon membranes containing 4,290 probes (Arfin, et al., 2000; Long, et al., 2001). Key to the analytical methodology is the application of a sliding window over probe sets sorted by mean expression level to determine an average local variance for probe sets of a similar expression level. As Affymetrix data contains many more probe sets than the development dataset it is thus of interest to determine the ideal window size for maximal power of detection.

Guidelines from the authors relate to the E. coli dataset (Baldi and Long, 2001), and suggest a window size of 101. In the follow-up paper (Long, et al., 2001) the issue of window size is further explored at values of 41, 101, and 161, with a value of 101 chosen for subsequent analysis. However, they also state *"Window sizes of larger than 100 genes often perform well."* In the supporting package Cyber-t *(http://visitor.ics.uci.edu/genex/cybert/)* the authors comment *"A sliding window of 101 genes has been shown to be quite accurate when analyzing 2000 or more genes, with only 1000 genes a window of 51 genes may work better".*

As Affymetrix data contains many more probe sets than the nylon membrane arrays used to develop the method it was hypothesised that a larger window may be more appropriate for estimation of the local variance. However is should also be noted that accurate determination of this window size is key to accurate results from the Bayesian method. If too small a size is chosen, insufficient information will be incorporated to determine an accurate variance estimate, however if too large a window is chosen the resultant local value will contain information from probe sets with have an expression level (and hence variance) vastly different from the probe set under consideration.

6.3.1.1 Using FDR performance to tune the window size

To assess the optimal window size, the output from the previously described FDR framework (Section 2.x) was used to compare results. The ability of the Bayesian test with different window sizes to detect the fifteen spikes present in the 24 chips with a two-fold change in the MAS 5.0 Affymetrix Latin square dataset was assessed using FDR curves over a range of sample sizes with multiple samples. Window sizes of 50, 100, 200, 300, 500, 750 and 1000 were chosen for consideration. The resultant summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figure 6.2. In addition data from the Welch t-test applied to MAS 5.0 data is shown for comparison to the standard test.

Figure 6.2

a)

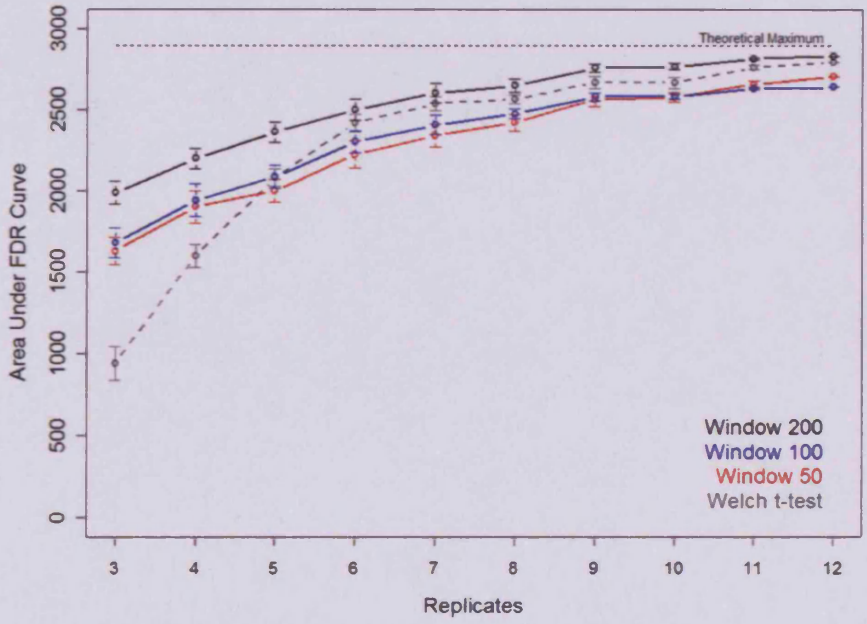
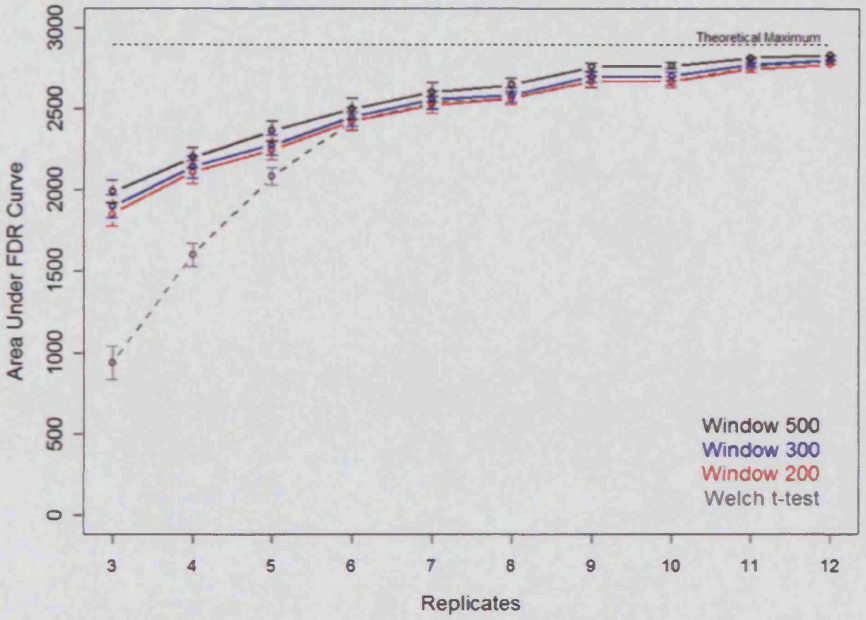


Figure 6.2 - continued

b)



c)

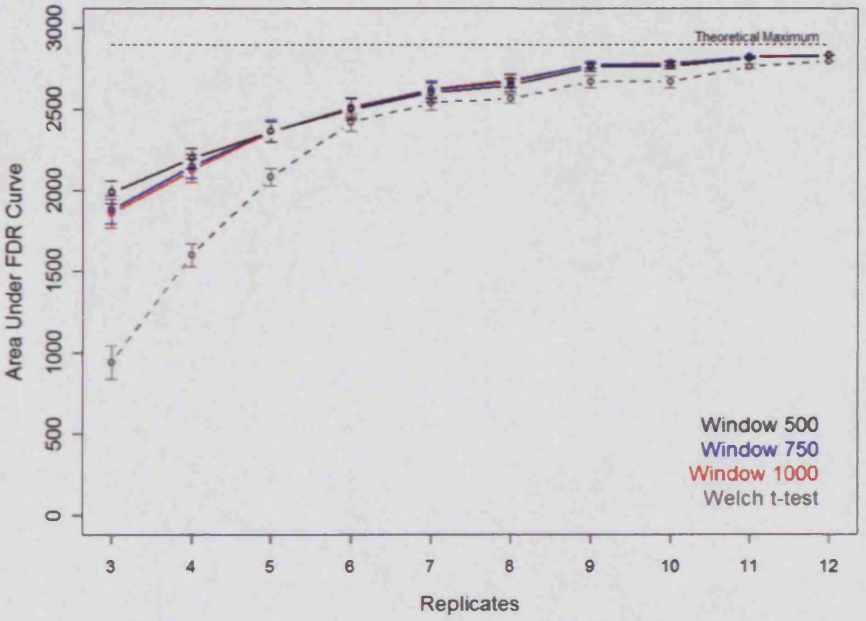


Figure 6.2 - Summary FDR plots showing the results of analysis using the Baldi and Long Bayesian framework on MAS 5.0 data with a variety of sliding window sizes.

Review of the data in Figure 6.2 shows that overall the power of the Bayesian method is greater than that of the standard Welch t-test, especially at the smaller sample sizes, where the conventional test has a marked loss of power. For comparison the power achieved from the Bayesian method with just three samples per group is equivalent to that obtained from the Welch test with five samples.

Analysis of the optimal window size indicates that an increase in power is observed as the window sizes rises from 50 to 500 (Figure 6.2.a and 6.2.b), beyond 500 the power begins to reduce again (Figure 6.2.c) indicating that a window size of 500 would appear the optimal size for application to this Affymetrix dataset. This value is approximately 5% of the probe sets to which the window is applied, which corresponds to guidelines for lowess smoothing, which also suggests a default window size at 5% of the size of the dataset (MathWorks, 2004).

6.3.2 Defining an optimal blending weighting

As previously described in Section (6.1.2) the Bayesian framework uses a combination of the actual probe set variance blended with the locally calculated variance from probe sets with a similar mean expression level. Within the methodology, the blending weighting is a number between zero and infinity which indicates the weight given to the Bayesian prior estimate of local variance; the larger the weighting the larger the confidence given to the local variance. The variances are then blended using the Dirichlet distribution according to the blending parameter (Baldi and Brunak, 2001). The Dirichlet distribution is the multivariate generalisation of the beta distribution and is assumed as the prior distribution of a multinomial distribution in Bayesian statistics (Wikipedia, 2005). The authors suggest the parameter is set to three times to number of experimental observations when the sample size is small, and that the value is reduced to a figure equal to the number of observations as the number of observations increases.

In choosing values for the weighting parameter, two separate application models were chosen. The first model altered the blending parameter as the sample size altered; thereby applying the desirable property of the Bayesian approach converging to the t-test as the experimenter carries out additional replications and thus becomes more confident in the observed estimate of within treatment variance. A second model was also applied in which the blending parameter was fixed and not dependant on the sample size.

6.3.2.1 Using FDR performance to tune the blending parameter

To assess the optimal blending parameter, the output from the previously described FDR framework (Section 9.6) was used to compare results. The ability of the Bayesian test with a window size of 500 to detect the fifteen spikes present in the 24 chips with a two-fold change in the MAS 5.0 Affymetrix Latin square dataset was assessed using FDR curves over a range of sample sizes with multiple samples. Blending parameters fixed at 5, 10, 15 and 20 were compared along with a variable blending parameter based on 1, 2, 3, 4 and 5 times the sample sizes in each group was applied. The resultant summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figures 6.3 and 6.4. In addition data from the Welch t-test applied to MAS 5.0 data is shown for comparison to the standard test.

Figure 6.4

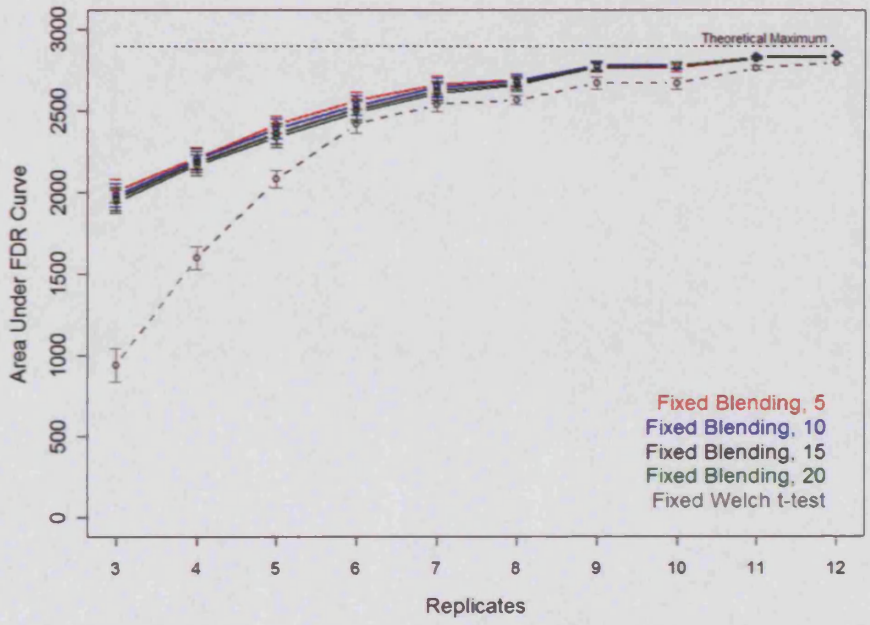


Figure 6.4 – Summary FDR plots showing the results of analysis using the Bayesian framework on MAS 5.0 data using different fixed blending weightings.

Review of the data for fixed blending shows very little difference between the different parameter values, however there is a general trend that the smaller the fixed blending parameter, the more powerful the resultant Bayesian test. However as this is not a mode the authors recommend the test is run in, it is the results from the variable blending that are of more interest.

Figure 6.5

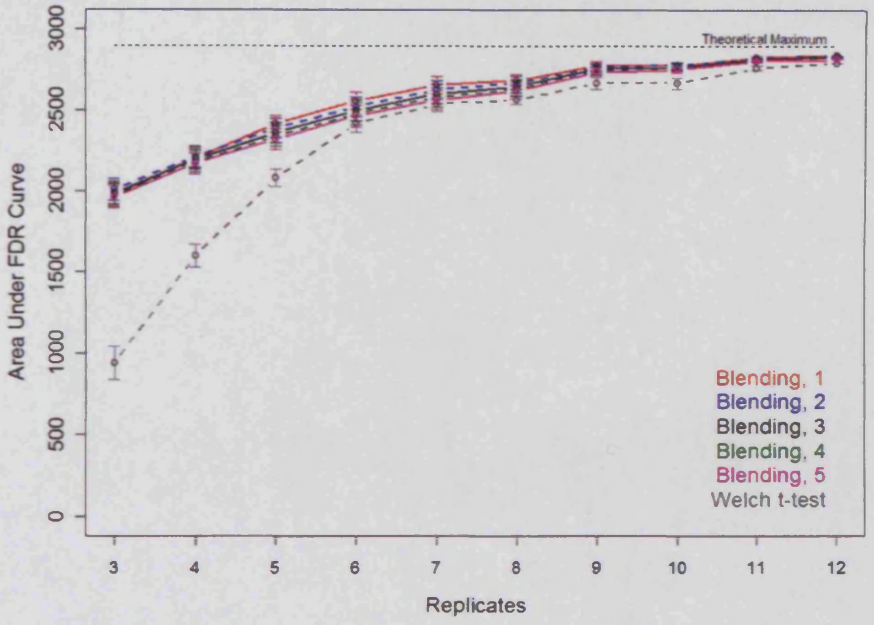


Figure 6.5 - Summary FDR plots showing the results of analysis using the Bayesian framework on MAS 5.0 data using different variable blending weightings calculated as a multiple of the sample size per group.

Review of the data assessing the power achieved from the Bayesian version of the t-test across of a range of variable blending values dependant on the sample size indicates similar results to those achieved from the fixed blending, namely that little difference is seen between the different values. In addition it should be noted that every variant of the Bayesian test performed better than the Welch t-test, especially at the smaller sample sizes.

As a final comparison, the FDR output from an analysis implementing a fully Bayesian estimate of the variance was compared to the results from fixed blending with a value of 5 and variable blending based on three times the sample size (Figure 6.6). A reasonably powerful test is observed, however it does lack power across all sample sizes when compared to the blended varieties of the test.

It can therefore be concluded that the blending stage of the Bayesian framework is key to the additional power obtained when compared to the standard methods, however the exact value of blending weighting seems less critical than the application of this stage itself. Thus there is no reason to deviate from the authors recommendation that the blending weighting be applied at a value three times the sample size in each group.

Figure 6.6

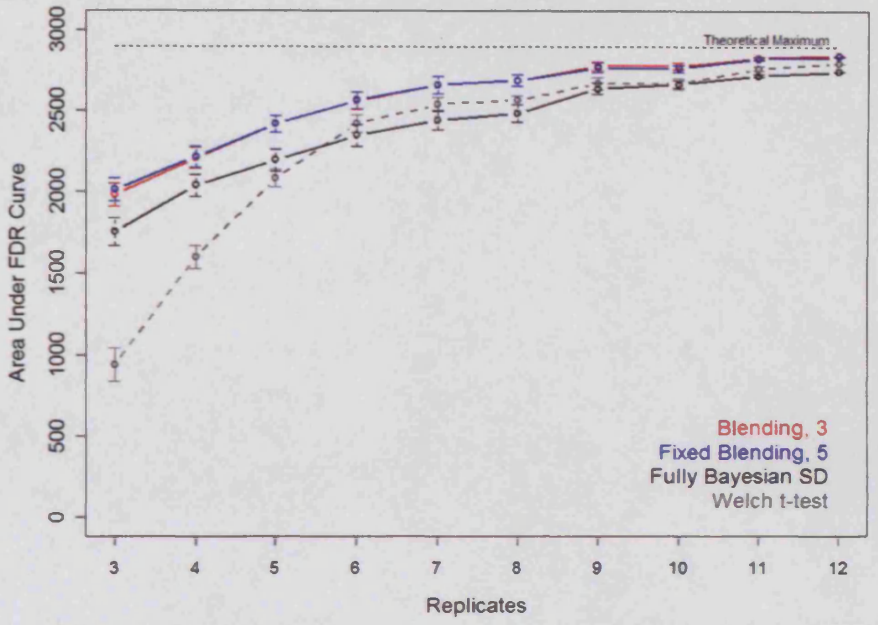


Figure 6.6 –Summary FDR plots showing the results of analysis using the Bayesian framework on MAS 5.0 data comparing the effect of variable and fixed blending weightings versus a fully Bayesian estimate of a probe set's variance.

6.3.2 Application of a robust local variance estimate

It had already been commented that accurate determination of the local variance is key to accurate application of the Bayesian model. In Section 6.3.1 the window size was tuned to achieve maximal power to detect. To achieve this power it can be assumed that at this optimal window size, an accurate measure of the localised variance has been obtained. As an alternative to this window size, another approach a more robust method for determining the variance value representing this window. In the framework as presented by the authors, the reported local variance is calculated as the mean of the variances for all probe sets within the current window. It is proposed that use of the median of variances, as an alternative to the mean, may yield improvements in the local variance estimation.

6.3.2.1 Using FDR performance to tune the blending parameter

Models using both the mean and median of the local variance were applied to untransformed data from MAS 5.0, and $\log_2$ transformed data from RMA and gcRMA with the results being assessed using the previously described FDR methodology (Section 9.6). The resultant summary graphs showing the area under the FDR curve over a range of sample sizes are shown in Figure 6.6. In addition data from the Welch t-test applied to MAS 5.0 data is shown for comparison to the standard test.

Review of the data comparing the use of mean versus median for estimation of the localised variance within the Bayesian framework yields markedly differing results dependant on the expression metric the methodology was applied to. When applied to MAS 5.0 data the mean estimate of the local variance has substantially more power than that obtained from the median.

Figure 6.7

a)

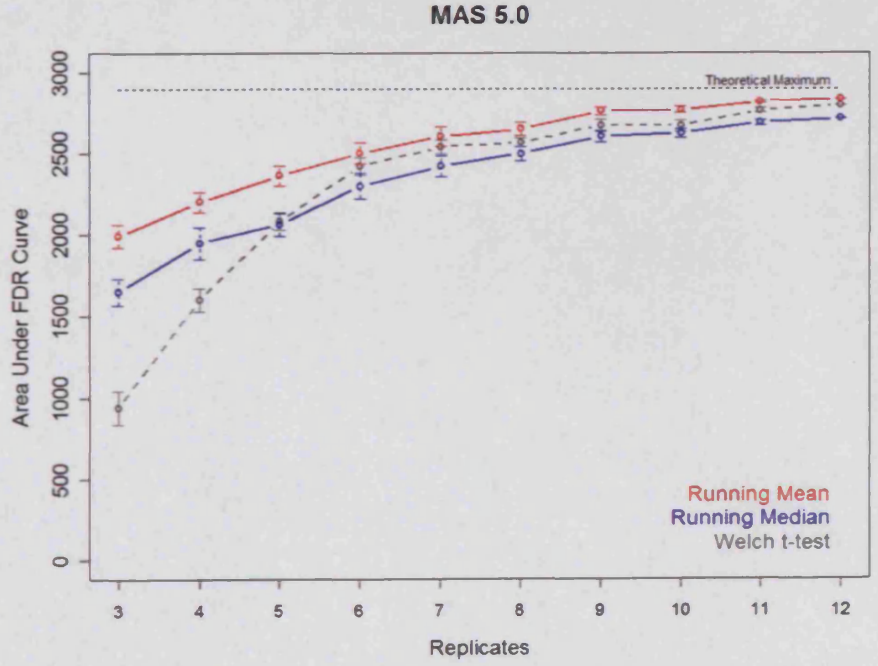
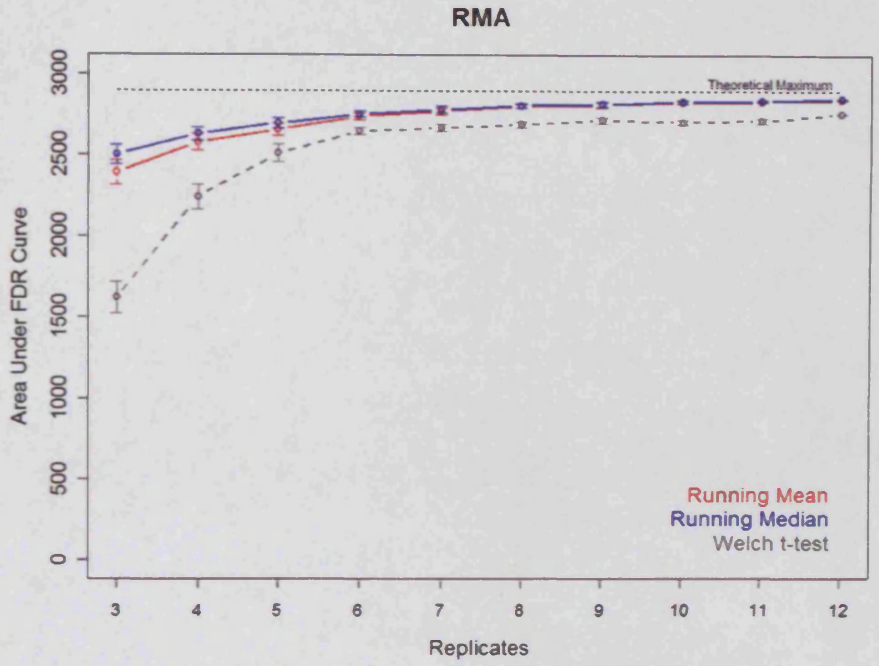


Figure 6.7 - continued

b)



c)

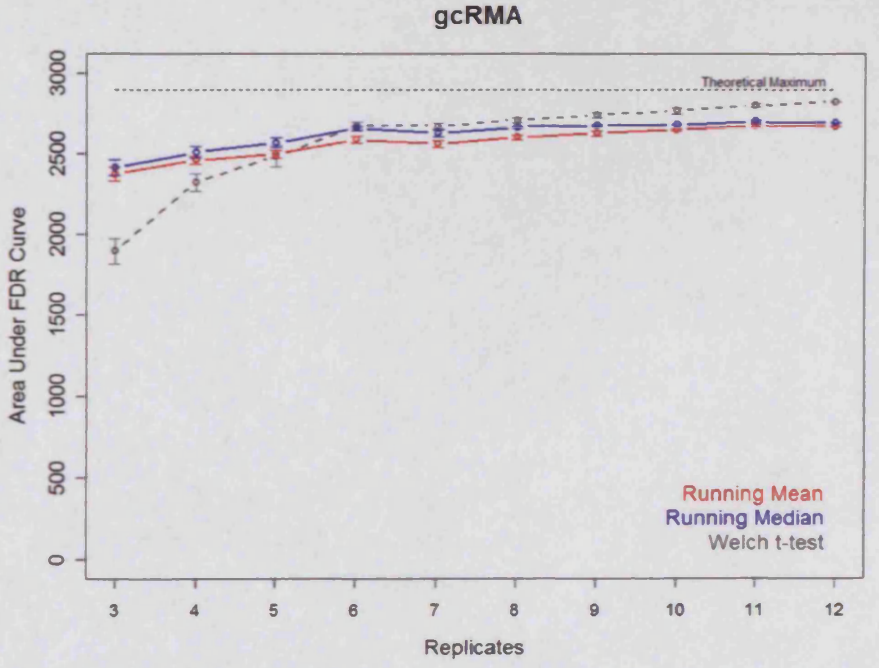


Figure 6.7 Summary FDR plots showing the results of analysis using the Bayesian framework using the running mean and median of the locally calculated variance applied to a) MAS 5.0 data, b) RMA data and c) gcRMA data.

With data obtained from RMA and gcRMA little difference was seen between the two variants of the model. There is therefore, little support for taking the median of the local variance window as an alternative to the mean. A possible reason for this observation was revealed when the relationship between mean and variance was examined for each of the three expression metrics (Figure 6.8.a, 6.8.d and 6.8.f). Whilst the reported relationship between mean and variance was observed with data from MAS 5.0, no clear relationship was observed with data from RMA or gcRMA, and data with variance being independent of the expression signal. Reversing the $\log_2$ transform introduced as the last stage of the RMA and gcRMA expression metrics re-established a relationship between the mean and variance (Figure 6.8.c and 6.8.e).

Whilst the Bayesian framework does not utilise the relationship between mean and variance, the blending stage with a typically constant variance across all expression levels would appear to act to control highly variable data and therefore improve the power of detection over the standard Welch t-test. This effect is similar to the procedures developed by Tusher et al. (Tusher, et al., 2001) who introduced a constant to the variance components of the t-test methodology to stabilise the denominator in the equation.

FDR analysis of the Bayesian framework was undertaken to explore the effect of the logarithmic transform of the data on the detection of the fifteen known spikes with data from MAS 5.0 and RMA (Figure 6.9). The results indicate that at smaller sample sizes there is an increase in power observed when the analysis is undertaken with natural space data.

Figure 6.8

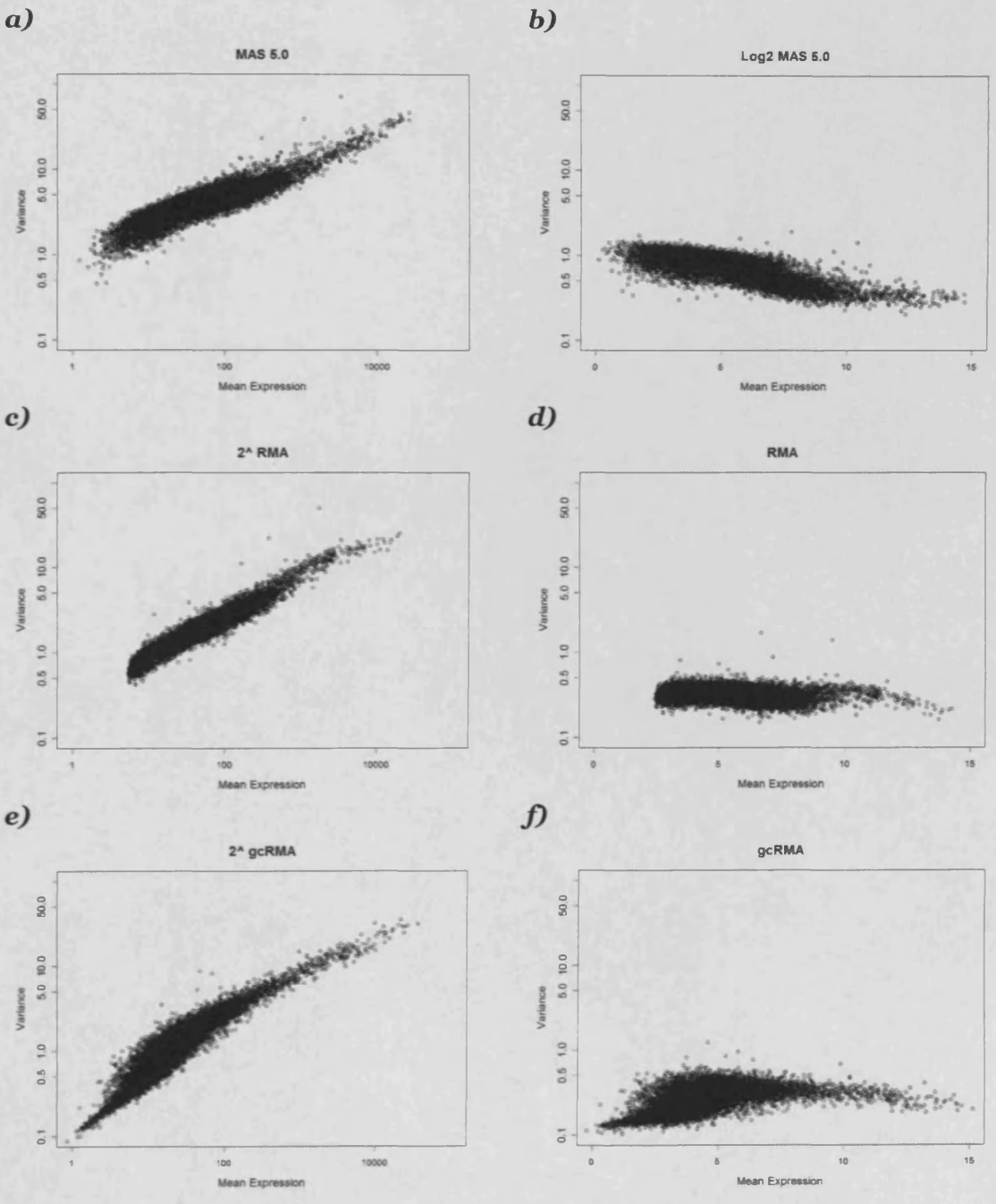
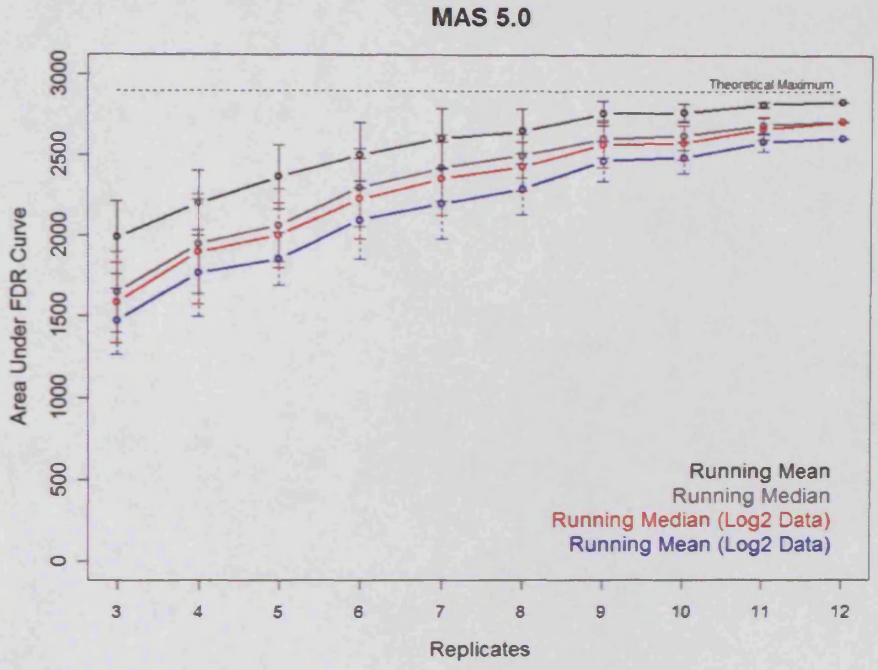


Figure 6.8 – Relationship between mean and variance for natural and log space data for each expression metric examined using a 12 sample dataset comprising one group of the Affymetrix Latin square dataset.

Figure 6.9

a)



b)

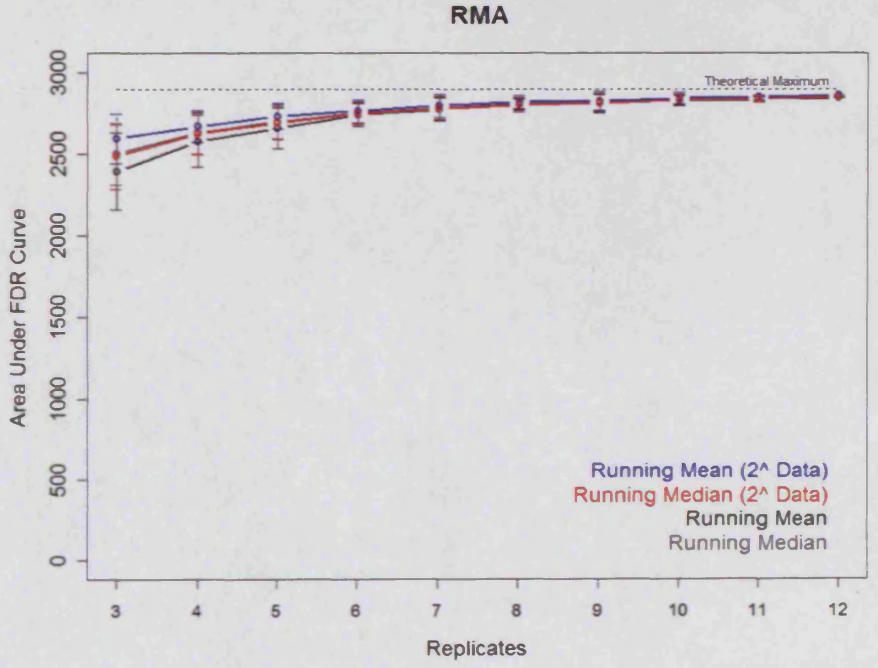


Figure 6.9 - Summary FDR plot showing the results of analysis using the Bayesian framework using the running mean and median of the locally calculated variance applied to data in logarithmic and natural space for MAS 5.0 and RMA data.

6.4 Discussion

Overall, the application of the Baldi and Long Bayesian framework for the detection of differential gene expression would appear to be a powerful technique that overcomes the limitations of some of the more standard statistical approaches at smaller sample sizes. Incorporation of information on the local variance determined from probe sets with a similar mean expression level along with the actual variance for the probe set under scrutiny yields power increases in the ability to detect differentially regulated genes when compared to the standard Welch t-test. An approximate comparison shows that the power achieved from the Bayesian method with just three samples per group is equivalent to that obtained from the Welch test with five samples.

The guidelines provided with the framework are somewhat limited in the effect that altering the various user tuneable parameters has on the experimental outcome, hence the requirement to investigate how changes in the local variance window and the blending weighting altered the outcome for the correct identification of spikes in the Affymetrix U95A Latin square dataset.

Investigations into the optimal sliding window size, used to estimate the local variance indicated suggested that with MAS 5.0 data, a window size of 500 gave the maximal power of detection of the sixteen spikes in the test dataset. This value is approximately 5% of the number of probe sets represented on the GeneChip under analysis and is comparable to guidelines given for the application of lowess smoothing, which uses a similar sliding window over a dataset. The author's recommendation of a window size of 101 represents a similar proportion of their developmental dataset and it is thus suggested that researchers apply a window size with a figure of 5% of the probe sets present on the array under analysis.

Explorations of the ideal blending parameters compared the application of fixed blending across a range of sample sizes versus a more dynamic model using a multiple of the replicates in each group as suggested by the authors of the methodology. Comparison of the two methods showed little difference in detection power, however when compared to a fully Bayesian method they proved superior. The data supports the application of a blending weighting as per the method author's suggestion, at a value three times the sample size per group.

Incorporating some of the ideas for statistical robustness introduced in Chapter Six, a final investigation was undertaken to explore the application of the median variance within the sliding window of localised variance. Data from MAS 5.0 yielded more power from the mean of the localised variance, RMA and gcRMA data gave similar results from either the mean or median. A final investigation of RMA data indicated that at small sample sizes, addition power was obtained by removing the logarithmic transform introduced as the final stage of the expression metric.

A summary plot comparing the best Bayesian results from MAS 5.0 and RMA compared to the Welch t-test show significant increases in power of detection at the lower levels (Figure 6.10). In summary, it is therefore suggested that researchers should employ the default mean estimate of the localised variance and consideration is given to the removal of the transformation when utilising the RMA expression metric. Overall the Bayesian framework presents as a very powerful analysis tool, allowing a research to extract the maximal amount of information from a cost-limited experiment and providing the greatest power of all of the statistical tests considered.

Figure 6.10

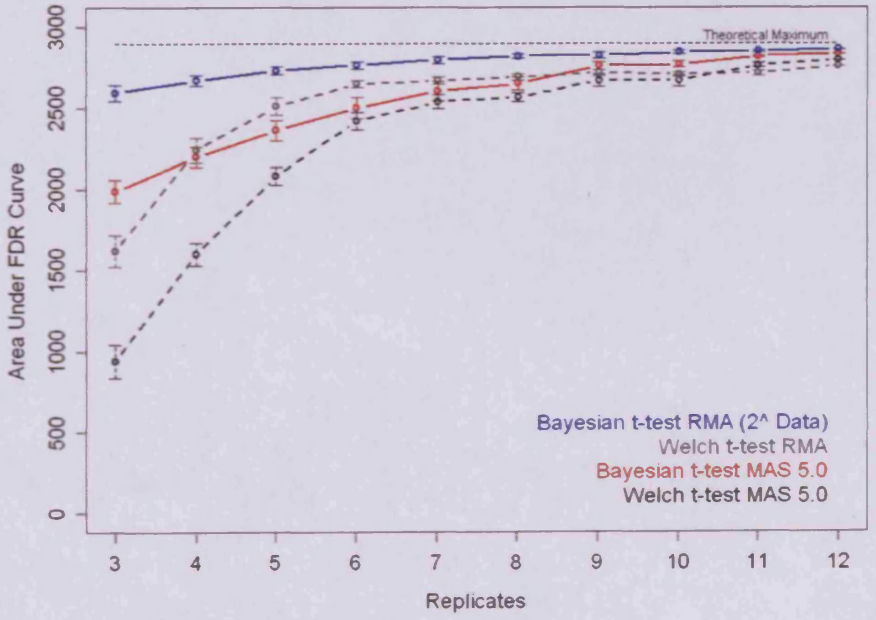


Figure 6.9 - Summary FDR plot showing the results of analysis using the Bayesian framework using the running mean and median of the locally calculated variance applied to data in logarithmic and natural space for MAS 5.0 and RMA data.

Chapter Seven
Approaches to annotation and exploration of Affymetrix microarray data.

In this Chapter the issue of annotation and exploration of the biological meaning of data is introduced. Section 7.1 introduces the need for annotation, the types of annotation available to assist in the interpretation of microarray data and highlights the limitations of many of the currently available tools and methodologies. Section 7.2 discusses the requirements and technical processes used in the development of a solution to data annotation and exploration. Section 7.3 introduces the key concepts identified and developed within the resultant software solution. Section 7.4 provides a functional overview of the MADRAS (MicroArray Data Review and Annoation System), highlighting the annotation and exploration tools available to the microarray researcher. Section 7.5 discusses how intuitive and comprehensive software tools are key to efficient experimentation using microarrays.

The work in this Chapter was undertaken in a joint collaboration with Daniel Kirwilliam, Bioinformatacist for the Wales Gene Park, Cardiff University.

7.1 Introduction

Work in previous Chapters has concentrated on investigating the optimal methods for the identification of differentially-regulated probe sets within an Affymetrix microarray dataset. However, identifying these probe sets is only the first start of an experimental process working towards the goal of determining and understanding the biological functions behind these experimental observations.

To draw conclusions about the functions operating within a dataset the key identifier linking observations between GeneChips, the probe set identifier (e.g. *"1708_at"*) must be annotated and expanded to provide a meaningful description of the biology the data is describes.

As part of the output from Affymetrix Microarray Suite (Affymetrix, 2001), the user is provided with a short textual description of the gene function the probe set is targeted at detecting, or an identifier for less well-characterised probe sets. However, RMA (Irizarry, et al., 2003) and gcRMA (Wu, et al., 2004) analysis metrics do not provide this information as part of the analysis process and the dChip (Li and Hung Wong, 2001a) analysis methodologies rely on external annotation sources to label probe sets.

In practice, the concise description offered by Affymetrix is often missing or poorly comprehensible due to the machine encoding and processing of the information, and some researchers have chosen to assemble their own Affymetrix annotation from the original probe sequences. As Zhong et al. comment, *"assembling comprehensive annotation information for all probe sets of any Affymetrix microarrays remains a time-consuming, error-prone and challenging task"* (Zhong, et al., 2003).

7.1.1 Making sense of experimental data

The volume of data that a microarray experiment produces is not insignificant, and the prospect of developing analysis strategies can be a daunting task for the novice user. For a researcher interesting in looking for differentially regulated genes within a dataset, this process can typically be reduced to a three step process:

i) Selection of genes of interest (extraction of genes with the largest fold change or the application of statistical testing).

ii) Collecting information on the probe set identifiers determined to be of interest, for example data describing a gene name and its functional information.

iii) Exploration and scrutiny of these findings in an attempt to identify links between significant observations and link these to meaningful biological hypotheses.

The first stage in the process, the statistical analysis of data and the production of a list of *"interesting genes"* is potentially the easiest stage in the process, and once decisions have been made about the type of analysis to be undertaken the results can be computed in minimal time and returned to the user. It is the subsequent annotation and exploration of these results that present as a potential bottleneck in analysis.

By the very nature of a microarray experiment, which provides a genome-wide picture of transcription within the cell, the results that are retuned are likely to take the researcher into areas they have no previous expertise in, highlighting new mechanisms and processes to explain the phenotypic differences between the cells of interest. If the list is small, this is easily achieved by reading database information and the available literature where appropriate. However, processing a list of hundreds of probe set identifiers typically returned from a statistical analysis is a much more onerous task.

Having established accurate annotation for each probe set, the potential then exists to assist in the summarisation of a list of significant results. Whilst many results may exist as stand-alone observations, many others may occur as a result of parallelisation

of results, either through replication of expression measurement on the microarray chip, or as a result of complementary biology acting on a series of linked transcripts.

Establishment of the resources and tools required to undertake such annotation and exploration of a dataset is thus key to the analysis of microarray data beyond the numerical stages of analysis. In addition the usability of these tools and establishment of effective workflows must be considered in order to undertake efficient exploration of a dataset's results.

7.1.2 Introduction to gene annotation

Annotation can be viewed as the commentary for each probe set and links the codes used to identify expression levels on microarrays to a variety of biological information about what the measurement represents. Annotation that may be of relevance to a researcher undertaking a microarray experiment can take many differing forms, from simple information such as the gene name and functional description, through more descriptive information about function such as gene ontology, through to information on the complex interactions of a gene within a pathway.

In humans, microarrays are attempting to measure the mRNA levels representing the 20,000-25,000 genes currently estimated to be present in the genome (International Human Genome Sequencing Consortium, 2004). Annotation of gene function, that is the linking a sequence within the genome to a biological process, can be achieved in a number of different ways; looking at the sequence information and identifying expressed sequence tags (ESTs), linking biological products (mRNAs or full length cDNAs) back to the original sequence or via homology to genes in other organisms. The primary consideration of annotation explored in this Chapter relates to Homo sapiens; however the annotation sources and challenges for other species are also briefly discussed.

7.1.2.1 Primary sequence annotation

Annotating an Affymetrix microarray experiment is the process of linking back to the sequence that the probe set represents, and then linking forwards to information about the function of the gene. GenBank (Benson, et al., 2004) is the primary public repository for human sequence information, with each accession number representing a unique submission to the database. GenBank identifiers are an important and common linker across all microarray technologies, but a GenBank identifier does not necessarily represent a gene, and a probe set may link to multiple accession numbers.

To overcome the limitations of the sequence based system of GenBank, UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters (Pontius, et al., 2003; Wheeler, et al., 2003). Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location. Currently UniGene contains 52,888 clusters of which 24,071 contain both mRNAs and ESTs. It is the UniGene database that is used as the basis for the design of each generation of human Affymetrix GeneChips with the latest clustering and genomic sequence information available utilised at the time of chip design (this information is reflected in the chip name, e.g. U95 series were built on UniGene build 95).

The default annotation produced by Microarray Suite for each probe set is a concatenated field containing information from the GenBank record (e.g. the record for probe set 41237_at is "*Homo sapiens /REF=D32129 /DEF=Cluster Incl. :Human mRNA for HLA class-I (HLA-A26) heavy chain, complete cds (clone cMIY-1) /cds=(5,1102) /gb= /gi=699597 /ug=Hs.181244 /len=1523 /LEN=1574*"). This field shows that the probe set is designed from UniGene cluster Hs.181244 representing GenBank sequence D32129.

However with each new build of the human genome, additional information is included which can yield new insight about the target of a particular probe set on a chip. The UniGene database changes regularly, since sequence data is re-clustered about once a month. As a result, the presence of new sequence information can change the clusters. As the sequences of an Affymetrix probe sets are fixed, these changes can result in a probe detecting the transcription of different genes or transcripts based on the current clustering. As such even though the binding of probes to transcripts in the experimental sample is fixed, the interpretation of what the resultant signal measurements represent is a much more dynamic entity.

7.1.2.2 Affymetrix probe set annotation

The key resource available to a researcher to link the probe set identifiers on GeneChip to information regarding biological function, is the NetAffx web portal (http://www.netaffx.com). Through this website, Affymetrix make current annotation data for each probe set available, including the information allowing the linking of a probe set between the various annotation databases available. Whilst some researchers have chosen to undertake their own annotation processing working form the probe sequences and building upwards (Gautier, et al., 2004), to the majority of researchers choose to use the information provided by Affymetrix.

It should be pointed out, however, that there are limitations to the annotation methodologies employed by Affymetrix. The approach has assumed a *"one probe set-one target"* relationship for the interpretation of the data and therefore the system does not deal well with situations where a subset of oligonucleotide probes in a probe set may be assigned to another gene (or more than one gene) based on the current UniGene clustering and genome annotation. Whilst this issue will hopefully be resolved over time as the stability of UniGene cluster increases and additional methods can be developed to incorporate all of the information that the experimental design provides, the issue of multi-gene families will still remain.

7.1.2.3 Secondary meta databases

Whilst the RefSeq and UniGene databases are concerned with collation and simple annotation of the primary sequence data, there are a variety of second level, meta databases which undertake to collate further relevant information and relate them to primary databases. Examples include LocusLink (Maglott, et al., 2000; Pruitt, et al., 2000; Pruitt and Maglott, 2001), centered on genomic location (e.g. curated sequence and descriptive information about genetic loci); PFAM (Bateman, et al., 2004; Sonnhammer, et al., 1997) for protein domain structure; OMIM (Online Mendelian Inheritance in Man) (Hamosh, et al., 2005; Machet, 1998) for disease-related gene information; and GeneCards (Rebhan, et al., 1997; Rebhan, et al., 1998; Safran, et al., 2003) for comprehensive information from other databases on human genes.

Table 7.1

LocusLink	Curated sequences and descriptions of genetic loci	*http://www.ncbi.nlm.nih.gov/LocusLink*
OMIM	Online Mendelian inheritance in man: a catalog of human genetic and genomic disorders	*http://www.ncbi.nlm.nih.gov/Omim*
Pfam	Protein families: multiple sequence alignments and profile hidden Markov models of protein domains	*http://www.sanger.ac.uk/Software/Pfam*
GeneCards	Integrated database of human genes, maps, proteins and diseases	*http://bioinfo.weizmann.ac.il/cards*

Table 7.1 – Summary of selected metabase information and URLs from Nucleic Acid Research Molecular Biology Database Collection 2004 (Galperin, 2004).

These secondary databases are designed to provide information from the perspective of genes, disease or proteins and provide a much more user-friendly gateway to a variety of information. However, the information is less complete than in the primary sequence databases.

The nature of the information in such databases is generally human curated, and more readable than the coded nature of the primary sequence databases and thus represents an essential resource in the determination of probe set function and links to other results within the biological system of study.

LocusLink is a descriptive database centered on the idea of a single genomic locus representing a single gene, with an emphasis on well-characterised loci (Maglott, et al., 2000; Pruitt, et al., 2000; Pruitt and Maglott, 2001). LocusLink provides a single query interface to curated sequence information and descriptive information about genetic loci. This includes official nomenclature (symbol, name), aliases, sequence accession numbers, phenotypes, MIM numbers, UniGene clusters, homology and map locations. At the time of writing LocusLink contains 33325 loci, which all have well established links with the UniGene database.

The LocusLink database is in the process of being discontinued and superseded by the newer Entrez Gene system (Maglott, et al., 2005). However, at the current time, a large proportion of Affymetrix microarray annotation resources use LocusLink identifiers as a key linker of information. LocusLink information is likely to remain key to the interpretation of microarray data during the transition period between NCBI systems; indeed the LocusLink identifiers are identical to those within Entrez Gene. However, some of the supplemental information provided for each identifier is different.

OMIM is a curated database of human genes and genetic disorders which currently contains around 15,000 records describing a single gene and the disorders relating to it (Hamosh, et al., 2005; Machet, 1998). OMIM focuses primarily on inherited, or heritable, genetic diseases and is considered to be a phenotypic companion to the human genome project. The OMIM record typically includes information about which diseases appear to be linked to specific genes, along with primary references that explain how the gene was sequenced and mapped to specific chromosomal regions. The record is hand-curated by various expert contributors and often presents as a lengthy textual description of the science surrounding a gene.

7.1.3 Linking gene functions

Each of the annotation sources that have been introduced have been very gene-centric and concentrate on the data relating to the single gene under scrutiny. One of the powers of microarrays is the parallel observation of many different genes simultaneously looking for patterns in expression values. To exploit this potential annotation must be examined between probe sets, looking for links in the probe sets flagged as significant. Intuitively, if several methods deliver the same functional annotation then one might have a higher confidence in the results.

Manual pattern searching is one approach that can be used, looking for correlation in the annotation between probe sets. However, this is a somewhat *"hit and miss"* approach dependent on the sources used and extent of annotation available for a gene. In addition, different database curators may use different words for the same function, or may mean different things by the same word. The context in which a gene was found (e.g. "TGFβ- induced gene") may not be particularly associated with its function.

The key resource needed to assist in the linking of probe set observations is a linking network between different genes with linking in expression or common function. Two potential gateways to this type of information are the various gene pathways and the application of gene ontology to functionally annotate a gene of interest.

7.1.3.1 Gene Ontology

The Gene Ontology consortium *(http://www.geneontology.org)* has undertaken a project to produce a controlled vocabulary that can be applied to all organisms and be robust against change whilst the knowledge of gene and protein roles in cells is accumulating and changing (Ashburner, et al., 2000; Harris, et al., 2004). It is a collaborative effort to address the need for consistent descriptions of gene products in different databases.

The developed ontologies allow the description of the attributes of a gene product in three non-overlapping domains of molecular biology. The Molecular Function domain contains ontologies describing the tasks performed by individual gene products (e.g. *"carbohydrate binding"* or *"ATPase activity"*), whereas, the Biological Process domain describes broad biological goals, such as mitosis or purine metabolism, which are subsequently linked to ordered assemblies of molecular functions. The final domain, Cellular Component describes the sub-cellular structures, locations, and macromolecular complexes (e.g. *"nucleus"* or *"telomere"*) which a gene is believed to be functional in.

In addition to the controlled vocabulary comprising the gene ontology project, collaborating databases provide linking database objects and gene ontology terms, which are hand curated and documented along with supporting documentation. These ontologies represent a unified, consistent system, terms occur only once, and there is a dictionary of allowed words. The terms have free text definitions and stable unique identifiers. Furthermore, terms are related to each other: the hierarchy goes from very general terms to very detailed ones.

A variety of methods have been developed to assist in the annotation and analysis of microarray data using gene ontology to provide information and links between observations. On of the simplest tools FATIGO, (Al-Shahrour, et al., 2004), undertakes simple counts of ontology terms for each of the genes in a list against each the terms in a single layer of the gene ontology hierarchy. Tuneable parameters allow for selection of the ontology domain and hierarchy level. However, each option requires a re-run of the program, which is a repetitive error prone task to fully explore a microarray dataset.

EASE, the Expression Analysis Systematic Explorer (Hosack, et al., 2003) makes attempts to improve on the limitations of FATIGO by undertaking a statistical analysis on the significance of the number of counts in a particular category and presenting results as a p-value. To overcome the issue of multiple probe sets linking to the same gene (and hence gene ontology annotation) over-representing and raising significance, EASE converts all accessions to LocusLink accession numbers before reporting counts. GO Miner (Zeeberg, et al., 2003) is another tool which undertakes a similar analysis approach, but integrates diagrammatical representation of the ontology hierarchy to supplement results. Use of gene ontology would appear to be a powerful method of supplementing basic gene annotation with a more system biased annotation helping to identify links between significant results (Li, et al., 2004).

7.1.3.2 Pathways

Pathways contain information, most commonly presented in diagrammatical form of how genes interact. A pathway can represent a series of interacting proteins allowing a biological function to occur, or can chart the links between the genes involved in initiating a response. Historically pathways were often determined for diagrams in textbooks, but these have the limitation of not being in a computer-readable format. A variety of pathway databases have evolved to collate information in a machine readable form including the KEGG (Kanehisa, et al., 2004; Ogata, et al., 1999), GenMAPP (Dahlquist, et al., 2002) and BioCarta (BioCarta, 2005) projects.

Each of these projects relies on contribution from experts within the global research community to submit data and thus each project contains only a few thousand well annotated pathways (more in section 7.4.4). GenMAPP is provided by the authors as a software tool for the production of pathway diagrams, and has the facility to overlay the results of a microarray experiment in the form of shaded gene squares, providing a visual representation of gene expression within a pathway.

7.1.4 Limitations of current tools

So far in this introduction the issue of exploration and annotation of data has been covered from the technical provision of information, rather than the interaction with it in order to undertake an analysis. In contrast to the *"black box"* approach in which statistics chooses the significant changes in a dataset and an automated system makes sense of the results and returns a series of hypothesis about the observations in a dataset, biologists often approach their data with a set of pre-determined ideas about what they might be observing, and may wish to approach their data from a variety of angles.

Another issue to the typical researcher with limited computing skills is the management and processing of data between experiments and formats required for differing software packages. An example would be the combination of selected data from two stages of an experiment. As analysis tools are often developed by computationally-expert users, issues of user interaction and complexity are often overlooked. Researchers are often frustrated with the tools provided for them, and demand simpler, more user friendly programs.

Current tools available for the exploration of data follow a gene centric approach, where each line of interest must be followed from the initial probe set identified through a series of linked resources to gain the desired insight regarding that observation. In addition, current approaches to the annotation and exploration of microarray data keep the data and annotation separate, with researchers often relying of spreadsheet tools such as MS Excel to bring together information.

Experience shows that exploration for a single probe set can result in the creation of multiple web windows to access the required information. Whilst this manual process may be acceptable for a small list of genes, it becomes an onerous task when typical lists of several hundred results are encountered. This problem can be summarised as data is available rather than accessible.

7.2 Development of an analysis tool

The issue of exploring and annotating data is an area that has received surprisingly little attention from the academic development community, with most work concentrating on development of algorithmic based analysis problems (e.g. RMA, dChip, SAM). Effective use of computing and bioinformatics resources applied to the problems faced by biologists are however important and can reduce the time required to perform routine repetitive tasks.

Many annotation sources are publicly available and there is increasing linking between resources, however the requirement to firstly obtain and then review the numerous pages of text and screens containing differing segments of information required to understand the biological functions represented by the data from a single probe set cannot be view as an efficient analysis workflow. A solution was thus envisaged allowing a biologist to drive the simultaneous visualisation of data and annotation of their data.

7.2.1 Developmental drivers

As a result of interactions with researchers accessing the Cardiff University GeneChip service it became apparent that in contrast to the black box approaches to analysis (e.g. statistical analysis and clustering) where data is inputted and a list of definitive answers is output, biologists typically have ideas about the results of their data and wish to explore their data from a variety of angles and test pre-existing hypotheses.

One of the key drivers for the development of a new analysis solution was frustration in answering simple questions a novice user may have about their data. Examples include *"Can I see my data for all cyclin genes?"*, *"Can I see the data for this list of candidate genes?"*, *"Is there any link in the underlying biology from these significant results?"*, *and "Are there any other probes for this gene in my dataset?"*

Whilst the questions appear simple, providing answers requires the combination of a variety of differing tools and often repetitive submission of queries to databases in order to attempt to construct answers. Efficient tools to process these queries therefore present as a major bottleneck in analysis. The biologists who generate the datasets must ultimately undertake the analysis and interpretation of the data, rather than rely on the non-expert guidance given by a bioinformatician in order to achieve the maximum amount of interpretation of the biology contained within a dataset.

7.2.2 Requirements for an analysis environment

Many functionally excellent analysis tools have found limited uptake in their use due to issues of complexity for the research with average computation skills. This is exemplified by R environment (R Development Core Team, 2004) in which much microarray development has been undertaken. However, as the environment is almost entirely command line driven this has limited its usage by those undertaking microarray research, and is often limited to the development community rather than the user base.

In developing a solution to the challenges of combining data and annotation, the decision was made to develop a solution for intensive usage by the average biological researcher. As such the need for an easy to use, intuitive interface with clear functionality was identified along with challenges of the multiple operating systems in use within an establishment. An environment accessed through a webpage was viewed to be an ideal solution to overcome these issues, and provides the added benefit of the portability of access to data from differing workstations.

The technical requirements of a web-delivered system result in the need for a central data repository and delivery system which provides additional scope for sharing data between users. The system was therefore envisaged to allow the sharing of data within a research group, enabling interesting and key results to be shared between users without the need to leave the analysis environment for transfer.

7.2.3 Developmental aims

In an attempt to overcome the limitations of current systems and to assist the average microarray user with the issues and challenges highlighted regarding the annotation, a solution is desired, built on the fundamental idea of drawing together user data with relevant and meaningful annotation. Such a solution would enable the exploration of results and annotation in order to draw biologically meaningful conclusions from GeneChip data in an rapid exploration environment.

Forming an integral part of the typical analysis workflow, the system should build on the idea that an analysis is typically concerned with the linked observations between about collection of results (not just a single gene value), and thus allow a move from typical gene-by-gene analysis to looking and analysing groups of data, looking for links in the underlying biology. Ideally, this functionality would be supported by relevant analysis tools to assist in the identifications of links between results highlighted as significant.

7.3 MADRAS
Microarray Data Review and Annotation System

7.3.1 Key concepts

Review of the tools available to the microarray researcher revealed that typically the functions of analysis and annotation were kept very separate by the majority of common packages. Having undertaken a statistical analysis, one tool is required to view the experimental data and a series of others are required to annotation the findings of the statistical analysis. In addition the majority of tools work from flat text files, which makes the combination of data a tricky task for the average computationally shy researcher. Key to the development of MADRAS was the identification of three separate data streams which when combined can provide an environment suitable for the rapid exploration of data.

7.3.1.1 Experiments

Within MADRAS, the data from each GeneChip array is indexed and stored as a separate entity within the database. However, the text file output from each of the common expression metrics consists of a data matrix containing data from many differing arrays forming the columns of the matrix, with the probes and their associated signal intensities forming the rows of the matrix. The basic upload file for MADRAS is based on the MAS 5.0 Pivot Table export (Affymetrix, 2001), where each chip is represented with up to three columns of data with column name suffixes identifying the data type. More information on data formats is contained in Section 9.3.1.

As part of the upload users are given the chance to give an appropriate short name for each GeneChip, and add a description for each chip. At this point the data is then split for database storage with entries being added to a "chips" table containing the summary information for each chip, and the raw data values (including A/M/P calls) are uploaded into a single "data" table, all uniquely indexed for retrieval later.

The uploaded data can then be combined in the system to form "experiments" which form collections of chips. The key flexibility by adopting this approach is the ability to edit and combine collections of chips for analysis *"on the fly"*, without the need to combine the original experimental output files using tools such as Microsoft Excel. In addition data can be grouped within an experiment as the user desires using functions for re-ordering the chips within an experiment.

The MIAME guidelines have attempted to develop a common set in information that should be stored for each array to describe the Minimum Information About a Microarray Experiment (MIAME) (Brazma, et al., 2001). The attempt of this initiative is to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. In developing the MADRAS database the decision was made not to burden the user with inputting the substantial amounts of information regarding the experimental processes, experimental design and sample descriptions required for a MIAME compliant database. The MADRAS database should be seen as a constituent of an experimental tool, rather than a repository for the storage and retrieval of publishable data, such a function is better served by initatives such as the ArrayExpress (Brazma, et al., 2003; Parkinson, et al., 2005), GEO database (Barrett, et al., 2005; Edgar, et al., 2002) or Stanford Microarray Database (Gollub, et al., 2003; Sherlock, et al., 2001).

7.3.1.2 Probelists

The concept behind the probelist is to provide functionality in allowing the creation of subsets of data of interest to the user. A probelist is a list of Affymetrix probe set identifiers (present on a specific chip type) which are grouped together with a collective name and description (e.g. cell cycle genes). The source of identifiers that form a probelist can come from a variety of sources and include the results of statistical analyses, collections of prior knowledge relating to an analysis (e.g. published lists of genes involved in DNA repair), or another researcher's published results.

In addition MADRAS has additional information which can be of interest to the user and is delivered in the form of a probelist. Examples of functions returning a probelist include; searches of the annotation database across a variety of fields, finding probe sets on a chip relating to a list of gene names, and pathway data from BioCarta, GenMAPP and KEGG (further details in section 7.4.4)

7.3.1.3 Annotation

There are many annotation sources available to a researcher, some linking directly from Affymetrix identifiers (e.g. NetAFFX) others relying on linker information (e.g. LocusLink, Entrez Gene), and others referenced by gene name (e.g. OMIM). The problem when attempting to annotate just a single probe set using these tools is the number of windows to switch between on a desktop, and the amount of information that must be sifted before the key information about a probe set can be comprehended.

The approach taken in MADRAS was to review all of the common annotation sources and to extract only the fields deemed to be of most use after consultation with a variety of microarray users. The outcome was a combination of information that could take a researcher from the basic information about a probe set all the way to a short description of function, all contained within the same page. The way the annotation database was built also allows for easy searching for other probe sets representing the same gene on a chip and presence within pathways from BioCarta, GenMAPP and KEGG. In addition links to the original sources were made available, along with information of homologues of the probe set in other species.

7.3.1.4 Combining the three streams

Combining these three data streams is the central feature of MADRAS and forms the basis of the exploration functions (see Figure 7.1 for a graphical overview of the design). Selection of a probe set from within a probelist draws the user data (formatted into an experiment) in the graphical form of a bar chart and heatmap, along with a variety of annotation sources into a single overview for that probe set. Navigation tools then allow easy shifting between probe sets in a probelist allowing for rapid exploration of data.

The heatmap is a method of visualising the intensity of data using colour intensity, in the MADRAS implementation data is first log transformed and then median centred. Data below the median is given a colour value between 0 and 255 on the green spectrum, scaled according to the data value where the data point furthest from the median is given maximal colour saturation, with values near the median representing with a shade of green nearing black. Data above the median is similarly coloured using the red spectrum. The results are represented with a series of coloured cells below each bar on the bar chart (Figure 7.2).

7.3.2 Technical details

Having highlighted the desire to develop a web accessible system to overcome the design challenges of platform dependence, multi-location accessibility, and allowing for the potential of data sharing, issues still remained on the optimal set of programming and delivery tools to be used for the system. In designing the MADRAS, the current state of development environments suitable for delivery of a database drive web program was reviewed and a number of solutions examined.

Figure 7.1

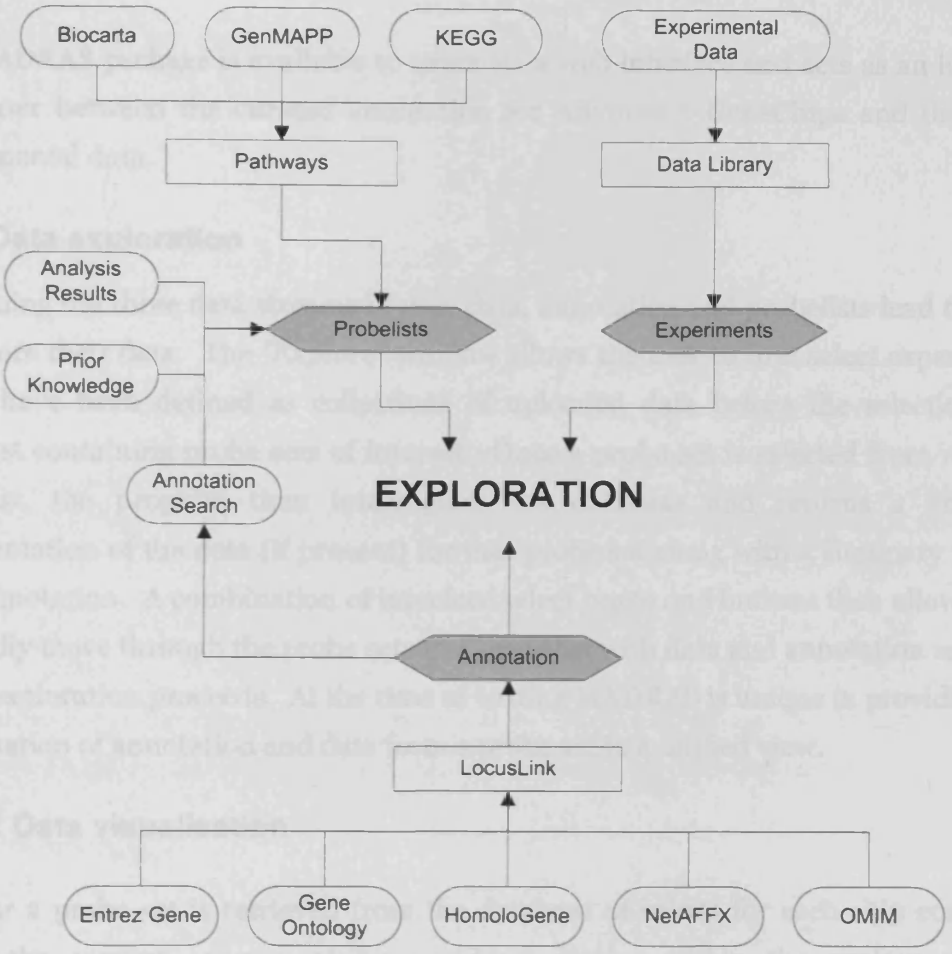


Figure 7.1 – Diagrammatical representation of the data relationships in MADRAS.

The result was the adoption of the open source PHP/mySQL platform. PHP (PHP: Hypertext Preprocessor) *(http://www.php.net)* is a CGI scripting language, interfacing with an Apache *(http://www.apache.org)* web server. Linking with PHP and Apache, the system interfaces with an open source mySQL *(http://www.mysql.com)* database engine for storage and retrieval of user and annotation data.

Further details of the technical concepts and data structures comprising the MADRAS system are contained in Section 9.7.3.

7.4 Functional Overview of MADRAS

The MADRAS package is available to users via a web interface and acts as an interface for a user between the curated annotation for Affymetrix GeneChips and their own experimental data.

7.4.1 Data exploration

Combining the three data streams of user data, annotation and probelists lead the user to explore their data. The *"Explore"* window allows the user to first select experiments which have been defined as collections of uploaded data before the selection of a probelist containing probe sets of interest. Once a probe set is selected from within a probelist, the program then interrogates the database and returns a graphical representation of the data (if present) for that probe set along with a summary page of gene annotation. A combination of interface select boxes and buttons then allow a user to rapidly move through the probe sets in a probelist with data and annotation returned as the exploration proceeds. At the time of writing MADRAS is unique in providing this combination of annotation and data from a probe set in a unified view.

7.4.1.1 Data visualisation

Data for a probe set is retrieved from the database of values for each chip contained within the current experiment for graphical display within the explore view of MADRAS. Once retrieved from the database the data is primarily displayed in the form of a bar chart (Figure 7.2), with additional colouring of the bars to represent the presence call (absent / marginal / present) from Microarray Suite when this additional information is uploaded with the expression values.

Additional options are provided to enhance the display of the data through the application of a logarithmic scale and/or the elimination of the values of expression level from each bar. As an alternative to the bar chart representation of the user data, a heatmap style diagram for the data is also shown. The heatmap is generated by taking all of the data available for the probe set under scrutiny with the current experiment and assigning a colour (further technical details in Section 7.3.1.4).

A chip with a low expression value compared to others in the experiment will be coloured green, whilst high expressers are coloured red, with those in the middle having no/little colour assigned. In the example visual representation (Figure 7.2), the high expressing heart data is clearly seen as red, whilst the low expressing testis data is green, with the middle expressing lung data having little colour representation.

Figure 7.2

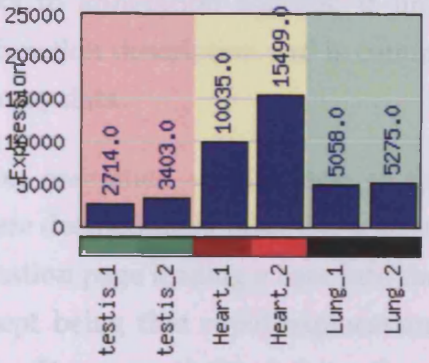


Figure 7.2 – Graphical representation of user data comprising a bar chart and heatmap diagram (see Section 7.3.1.4). The background of the chart can be shaded by the user to highlight different groups of data contained within the experiment.

In practice this alternative display methods would appear to be a good alternative to the bar chart when rapidly exploring data within a probe list as the eye is not required to move to reassess the information presented and intuitively shows which chips are low, middle and high expressers relative to the probe set's experimental data.

7.4.1.2 Annotation

In the introduction to this Chapter the various sources available to annotate a probe set were discussed along with the pro and cons of each methods and direction. The major limitation encountered was the separate nature of the sources, and the multiple windows required to fully assimilate information about a probe set. MADRAS attempts to overcome these issues by providing a custom assembled series of sources of annotation which can be displayed on a single page in combination with the user's experimental data.

To gauge the sources that would provide maximal information and insight for a probe set, a series of local biologists and system users were quizzed regarding their views on what they view as good annotation sources, along with a feel for the relative importance and use of the suggestions made. The outcome of these discussions and the following design stages is an annotation page leading from simple gene annotation (e.g. name and symbol), through more comprehensive fields of information to homologues to genes in other species.

The primary sources of annotation information for the MADRAS system are NetAffx (Liu, et al., 2003), LocusLink (Maglott, et al., 2000; Pruitt and Maglott, 2001), HomoloGene (Wheeler, et al., 2005), OMIM (Hamosh, et al., 2005; Machet, 1998) and

Biocarta (BioCarta, 2005). Practical examination of theses sources and the links between them resulted in the Locuslink identifier being deemed most appropriate for linking Affymetrix probesets to annotation sources; it links directly to the Refseq (Pruitt and Maglott, 2001) function description and is commonly used to link to other annotation sources and pathway data.

The plethora of information contained within each of the sources identified was reduced to the fields that were deemed most descriptive about the function of the gene and ordered down the annotation page leading a user into the specifics of the probe set / gene's function, the concept being that rapid exploration required only a few key details about the probe set. However, if this information proved to be of interest, further details were also available without the need to link out to another page. Whilst the system was designed to reduce the need to link to other databases, links were provided back to the original annotation information for users who would prefer to consult the original data sources. The sources chosen are able to give a user a quick overview of gene function, disease related information, gene ontology and pathway information.

Gene names and function were obtained from the RefSeq and Locuslink function information, while disease and phenotypic information is provided in the form of the OMIM text entry. Whilst the OMIM record does provide a very comprehensive summary of the gene which has been hand curated, the record is often lengthy and has the potential to obscure important information. The decision was thus made to make the OMIM information available by revealing information already presented in the page. This can be viewed as the computational equivalent of opening a flap in a book containing extra information, whilst retaining the original structure of the book.

Information regarding the potential links between the probe set of interest and others involved in a system are provided by provision of the bottom level gene ontology field, along with pathway information obtained from GenMAPP, Biocarta and KEGG. In addition the chromosome mapping location information is displayed, obtained from the LocusLink record.

Recognising the fact that there often multiple probe sets supposedly representing the same gene, the nature of the links in the LocusLink database allows for information to be presented regarding the presence of other probe sets for the locus of interest, and probe sets present on other generations of gene chips. This information allows for the corroboration of findings between datasets even if these have been run at differing stages of the Affymetrix GeneChip release cycle.

The inclusion of HomoloGene information allows the expansion of insight and provision of linked information between species. This information enables sparsely annotated genes to achieve an enhanced description by using information from other species. Whilst this functionality is of limited use for researchers using human samples, those undertaking studies using other species with more limited annotation (e.g. rats) may find this information of use to overcome the limitations of the species specific annotation.

An example explore page is shown in Figure 7.3 showing the combination of data and annotation, all on a single page.

7.4.2 Annotation searching

The establishment of a database linking a variety of annotation sources provides the possibility of rapidly answering the difficult questions that were proposed in the developmental drivers for the MADRAS system. The idea of the probelist is fundamental to the exploration of data within MADRAS, and the search facility provides all its results in the form of probe set identifiers which can then be explored in combination with the user's data.

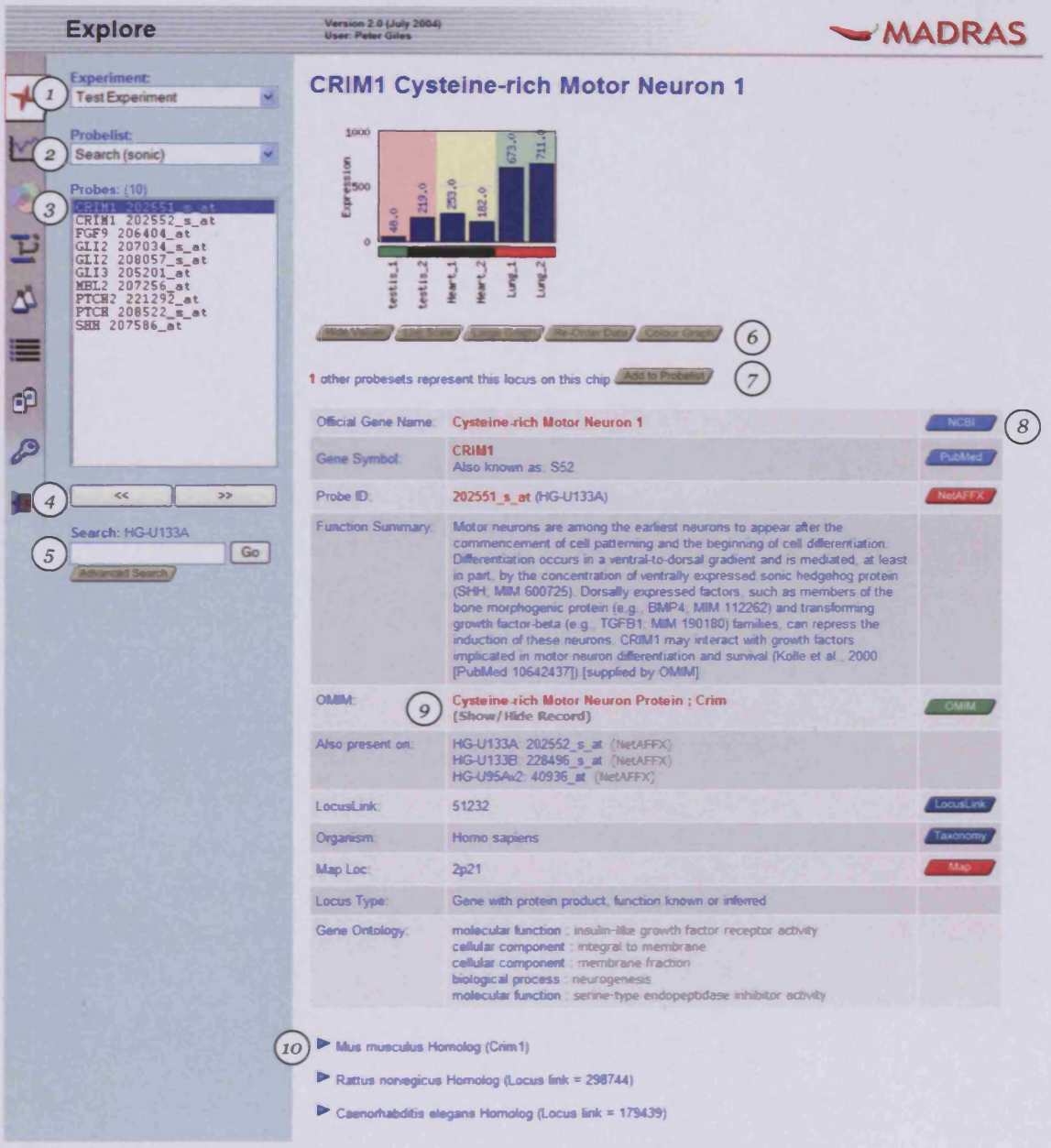
The explore page contains a simplified universal search box which seeks the search term across a range of annotation fields including gene name, gene symbol, gene product and the functional summary. To re-use the example of searching for cyclin genes, the power of this approach was highlighted when seeking for these genes on the HG-U133A chip. MADRAS was able to find additional genes as a result of this simple search compared to those hand curated over a period of several months by the user.

To expand the functionality of the generic simple search, an advanced search page is provided which seeks the term using a wildcard search against specific annotation fields including the official gene name, gene symbol, gene product, functional summary, LocusLink identifier, alias gene symbols, chromosome location and gene ontology. In addition the advanced search is able to search the annotation for homologues of the genes and relate these back to the GeneChip type the search was undertaken on (Figure 7.4).

7.4.3 Gene pattern finder

The parallel nature of microarray experiments can often yield result for system which are seen to respond in parallel to the experimental situation, and a researcher will often be interested in finding genes with a similar expression pattern to a gene already highlighted as being of interest.

Figure 7.3



Data Selection
Data exploration requires the selection of an experiment ① (a selection of uploaded chip data) and a probelist ② (a list of Affymetrix probe identifiers.

Data Navigation
Data can be navigated by direction selection of probes ③ or by browsing back and forward through the probelist ④. The user is presented with an annotation summary and graphical representation of the experimental data for the selected probe.

Search Options
The search box ⑤ on the explore page undertakes a generic search of gene names, functional summary and identifier fields and returns results as a new probelist. Options are provided for more specific search options in the advanced search page.

Graph Formatting
A variety of display options are provided for the data along with function to re-order data and colour the background of the graph ⑥.

Alternative representation
The presence of other probe sets representing the same transcript is flagged to the user ⑦, along with a link to create a probelist containing these additional probe sets.

Annotation Links
Direct links to the relevant pages from each original annotation source are provided along side probe set annotation ⑧.

OMIM Annoation
The large OMIM text field is hidden by default, but the page is expanded to show the record as part of the annotation with a simple click ⑨.

Homologues
A simplified annotation table is provided using homologue data where available, this data is initially hidden from view, but easily expanded into view ⑩.

Figure 7.2 – Example of the "Explore" page with overview of key functions.

In contrast to other packages which require the user to select a gene of interest and then do a *"find similar"* search MADRAS provides the functionality to undertake a search for a theoretic gene pattern as well as experimentally existing ones. This is achieved by use of a unique 'slider interface' (similar to a graphic equaliser) (Figure 7.5), making it possible to design a pattern of gene expression of expected observation and then search for all the genes that match that profile (e.g. *"find me genes with high expression in the heart samples"*).

Figure 7.4

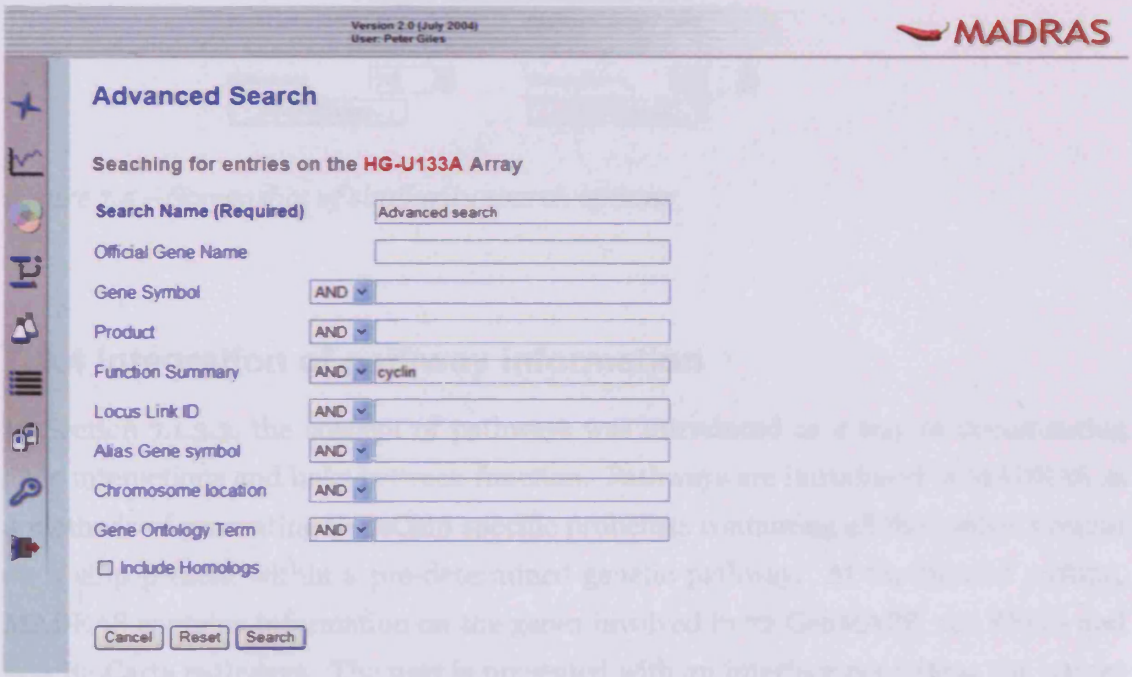


Figure 7.4 – Screenshot of advance search options

There are mathematically numerous methods to undertake the pattern matching (Sturn, et al., 2002) each of which has advantages and disadvantages and sensitivity issues when presented with data of a particular makeup. To overcome these issues and to reduce the burden on the user to make these statistical decisions MADRAS provides a single robust pattern matching algorithm in the form of a Pearson's correlation. Each probe set within the experimental dataset is compared to the master expression pattern defined by the user and a R^2 value calculated. Options are then provided regarding the filtering of these results before ultimate return of the information in the form of a probe list of similar genes.

Figure 7.5

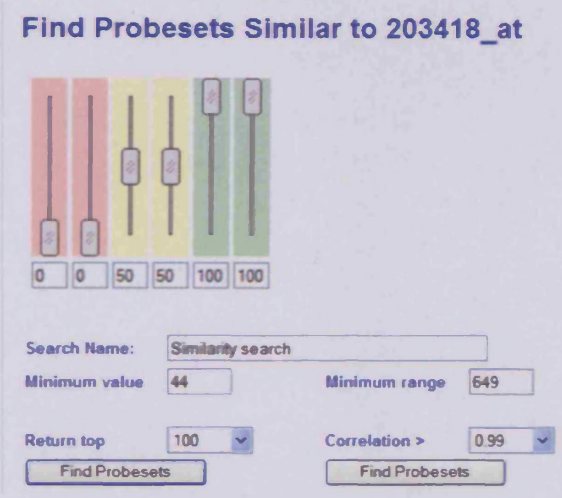


Figure 7.5 – Screenshot of similarity search options

7.4.4 Integration of pathway information

In Section 7.1.3.2, the concept of pathways was introduced as a way of documenting gene interactions and links between function. Pathways are introduced in MADRAS as a methods of generating GeneChip specific probelists containing all the probes present on a chip present within a pre-determined genetic pathway. At the time of writing, MADRAS contains information on the genes involved in 72 GenMAPP, 121 KEGG and 502 BioCarta pathways. The user is presented with an interface containing the names of each pathway, and upon selection the data is returned as a chip-specific probelist of probe sets present within the pathway of interest, which can then be explored and interrogated (Figure 7.6).

7.4.5 Probelist analysis

The annotation methods above are very much gene centric and are concerned with the information regarding a probe set in isolation. The formation of probelists within MADRAS provides an ideal and unique opportunity for further analysis of the data they contain looking for potential links and patterns with them. This is achieved using a combination of visualisation and data mining techniques.

7.4.5.1 Probelist clustering techniques

Patterns within a probelist can be visualised using gene clustering techniques (Quakenbush, 2001; Sturn 2001). In a similar approach to that used in the similarity search, single robust methods are implemented in MADRAS for the visualisation of clustered data from the data linked to a probelist. Experiments can be clustered and the resultant dendrogram viewed and explored. Probe sets clustering is visualised in the form of a log transformed median centred red-green heatmap (Figure 7.7)

Figure 7.6

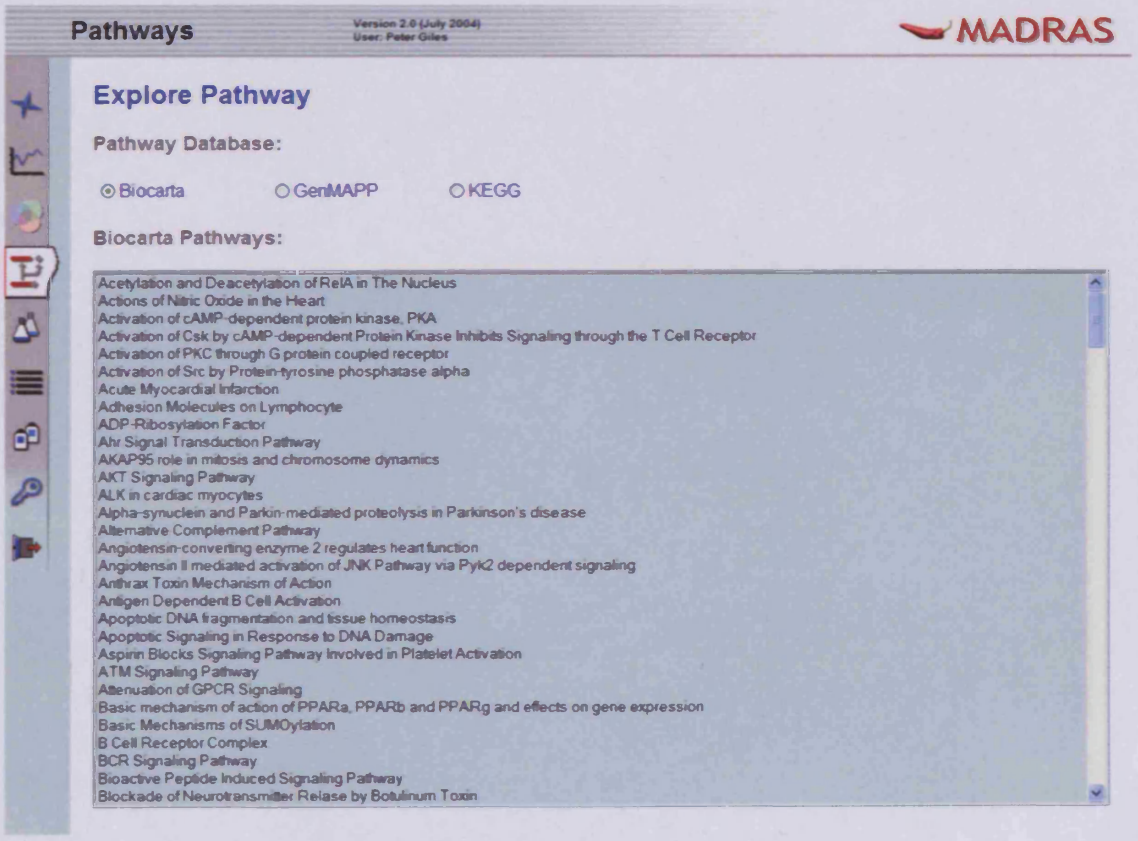


Figure 7.6 – Screenshot of pathways selection screen. Once a pathway is selected the probe sets on the currently selected GeneChip present within the pathways are returned as a probelist.

The tree view and heatmap graphics are produced using the free R statistical language (R Development Core Team, 2004). Following data export from within the MADRAS system. The output from R is then read back into the system for display. The graphics are also produced in Adobe Acrobat format and the user is provided with links to download copies for reference.

7.4.5.1 Methods to access probelist similarity

A user will often have undertaken a series of analyses on a dataset, or have prior knowledge of implicated genes and wish to compare these looking for similarities between the results. Unfortunately the methods available to undertake such a comparison are limited, and based on simple counts of matches between lists. MADRAS contains two tools for the examination of similarity between lists; a Venn diagram tool, and a probelist similarity tool for multiple lists.

Figure 7.7

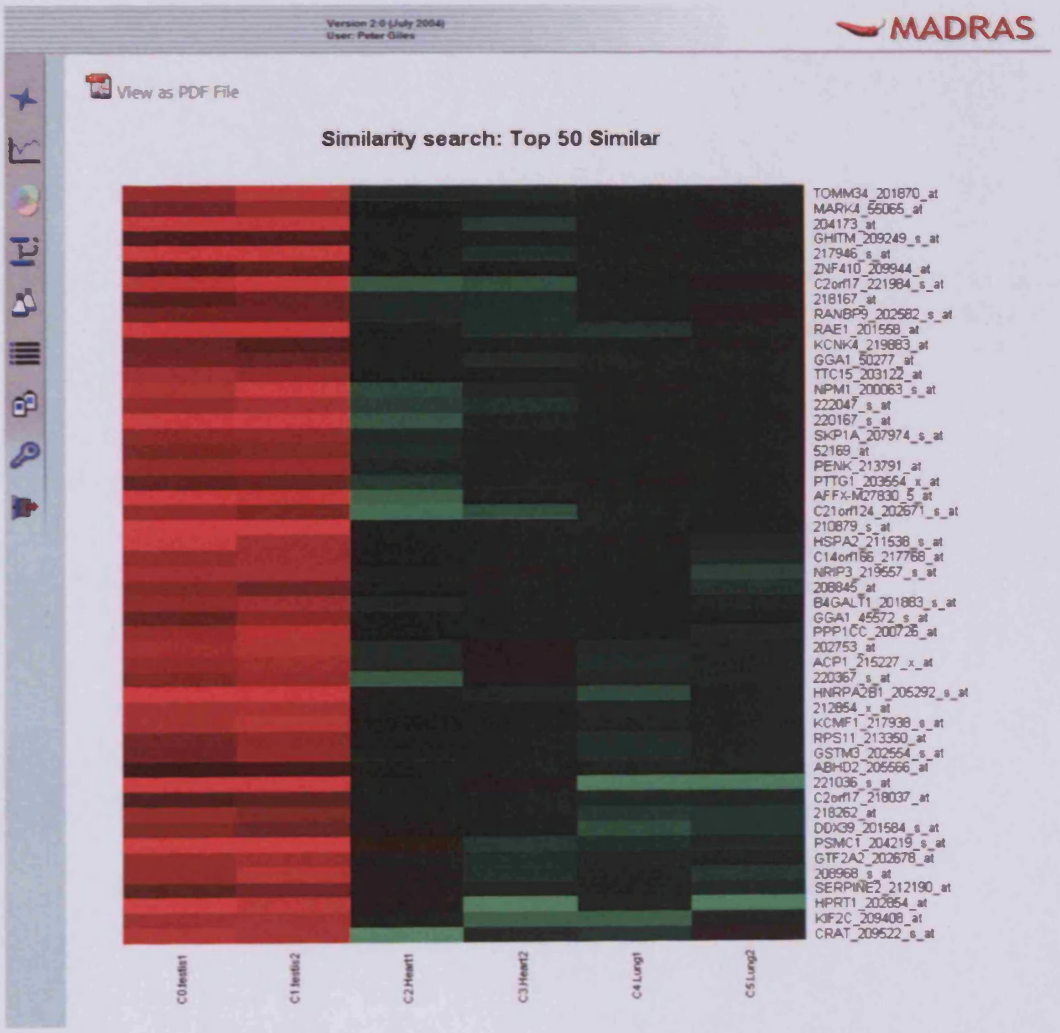


Figure 7.7– Example of probe set clustering visualised as a heatmap.

The Venn diagram tool is able to simultaneously assess the similarities between three separate probelists, and undertakes counts for similarity between each of the groups. The output shows these counts overlaid on a Venn diagram (Figure 7.8) indicating similarity between the contents of each probelist. By selecting a group, the user is presented with a probelist of the common probe sets which can easily be saved as a probelist for further analysis and exploration.

To address situations where a user may wish to examine similarities between more than three probelists simultaneously, a probelist similarity tool is provided with an output similar to the distance calculator commonly found in a road atlas. The method returns a table showing the number of probes common between each pairing of the probelists selected (Figure 7.9)

Figure 7.8

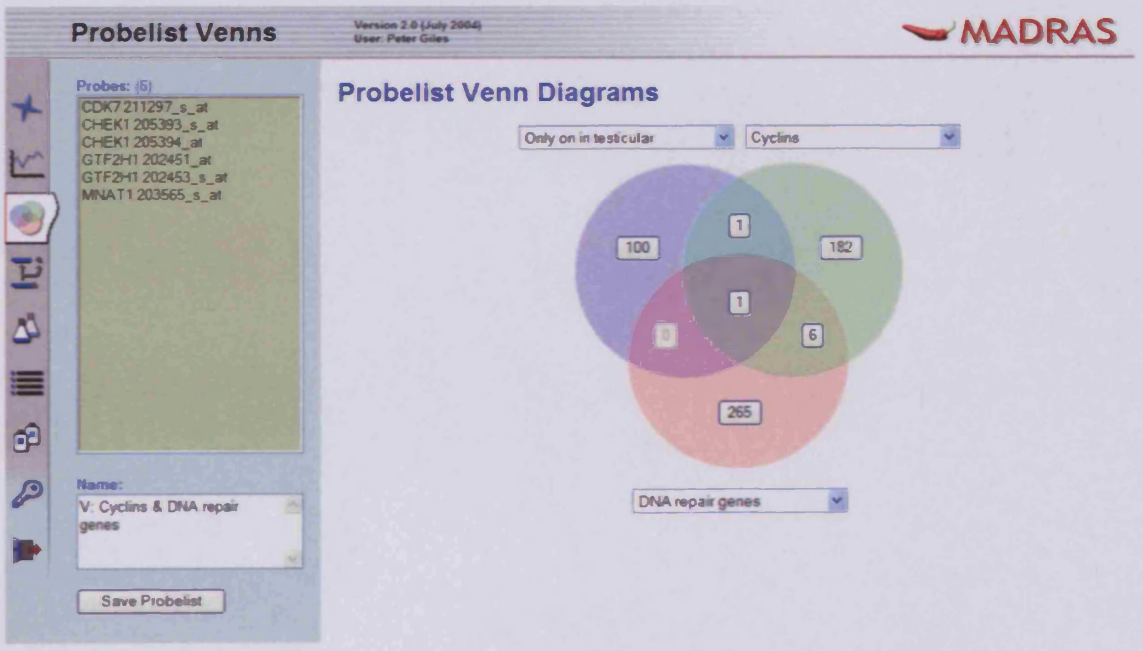


Figure 7.8 – Venn diagram showing similarities in content between probelists

7.4.5.2 Over-representation analysis

Perhaps the most time consuming aspect to microarray analysis is converting gene lists into meaningful biological answers. To achieve this, the user must access the probe sets within their probelist of interest looking for links biological links and patterns between the observations. Links between probe sets can be formed by common wording in their title, presence in a pathway, or shared gene ontology.

Figure 7.9

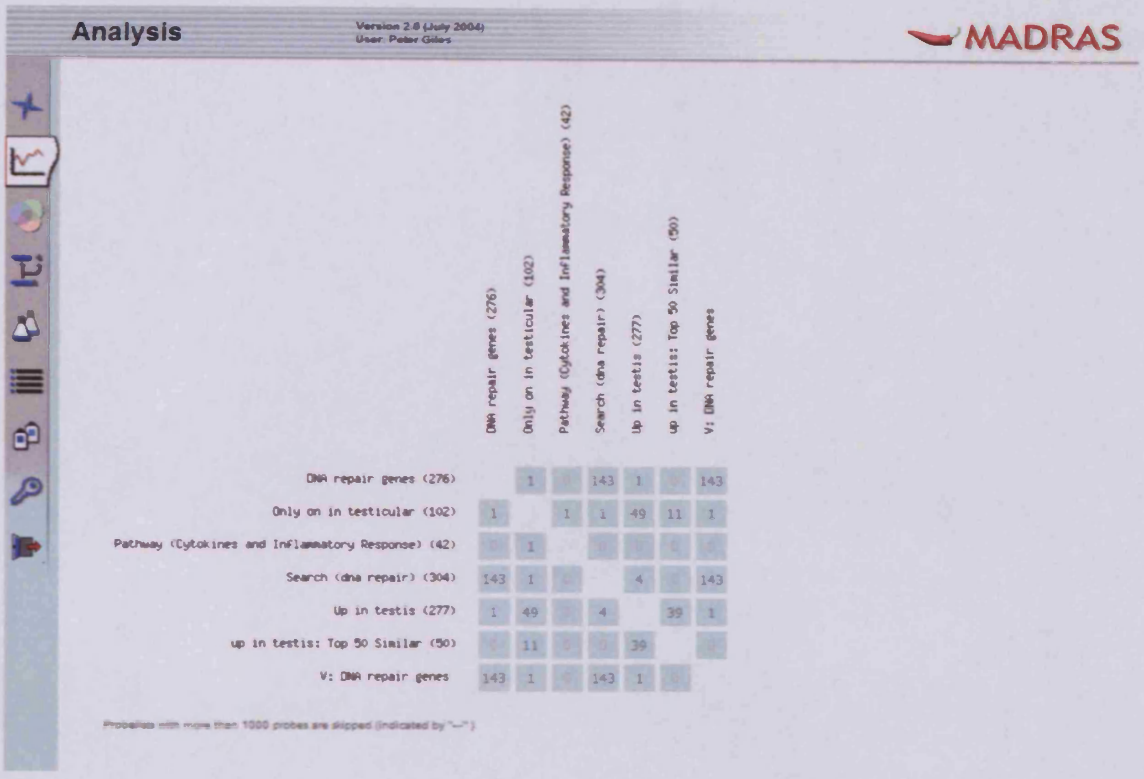


Figure 7.9 – Similarity counts between probelists

Building on the ideas presented in the EASE package (Hosack D.A., et al., 2003) MADRAS applies the idea of over-representation of terms across a wide range of annotation fields and returns a statistical significance value for the observations made.

Over-representation analysis is concerned with counting features within a sample and then comparing it to observations within the population and judging for significant over-representation of an observation in the sample. This can be conceptualised by using a bag of white and black balls. If the bag contains 50 white and 50 black balls and a user selects 10 balls and 7 of these are white and 3 are black, is the number of white balls selected significant over-represented? The start point for applying this method to microarray data is undertaking counts of the terms of interest for each probe set in a probelist and then the application of statistical techniques to examine significance.

If a user is interest in the over-representation of the term *"cardiac"* in their probe list of interest, the first stage is to undertake counts of this term in the population. This is achieved by retrieving all the available annotation for the GeneChip of interest and searching for matches. However as there are multiple probe sets for the same gene there is the potential for over representation due to chip design, and not necessarily biological significance. To overcome this, the population annotation must be reduced

to those representing loci only, achieved by the reduction of the population probe set list to unique LocusLink identifiers. The number of hits for *"cardiac"* is noted along with the number of unique loci represented on the GeneChip under scrutiny. A similar process is then undertaken on the probelist with the reduction of the data to unique loci and counts of matches to the search term.

As in most situations, there are many statistical techniques which can be applied to the resultant table of four values to access the significance between the sample and population. EASE (Hosack, et al., 2003) uses a modified version of the Fisher's exact test for this comparison, with a contingency sample set up to compare the reference set (GeneChip population), to the sample set (probelist information). However, it has been shown that Fisher's exact test is not necessarily the optimal test for this type of comparison and should be used only when very few genes are involved (Man, et al., 2000). As an alternative, MADRAS implements the statistical assessment of significance by using the hypergeometric distribution, a method primarily concerned with sampling without replacement from a binomial population (Doniger, et al., 2003).

MADRAS uses its large multi-source annotation to facilitate this over-representation analysis. In addition to comparison for presence within well defined pathway (BioCarta, GenMAPP and KEGG) fields and bottom level gene ontology information, MADRAS also undertakes text analysis on the gene name and functional summary fields. These text fields are broken down into individual words and counts undertaken on each word within the description. In addition word-pairs are accessed to overcome language issues which may results in under-representation of significant terms (i.e. *"cell"* will be well represented in the population; however *"cell cycle"* may be a significant word pairing in a probe list).

The question of correction for multiple testing on this data is one not fully resolved. Due to the number of different assessments made on the data for the same probelist this form of correction should be applied. However arguments persist regarding the independence of categories (especially for the hierarchical gene ontology information), so it is not clear how to correct for multiple testing (Hosack, et al., 2003; Zeeberg, et al., 2003). MADRAS incorporates a Bonferonni correction as part of the over-representation analysis, but leaves the choice of application to the user.

7.5 Discussion

The issue of combining data in an intuitive, accessible format has becoming a major challenge for the bioinformatics community (Orphanoudakis, et al., 2005). Work on MADRAS has demonstrates the benefits of combining experimental data, theoretical data and annotation, along with analytical tools aimed at assisting the interpretation of a dataset, in a single rapid exploration environment. Such an environment forms the final part of a microarray analysis workflow, taking in the results from the experiment, along with those from other analyses and statistical tests and aims to accelerate the time spent deducing conclusions for further experimentation or publication from the data.

Although bioinformatics software is best written by trained bioinformaticists, many have argued that to best capitalise on this investment it is ultimately the researcher who created those datasets (and are thus the custodians of it) who must be empowered to undertake the data analyses (Tilstone, 2003). Although early-stage data analysis focuses on statistical and mathematical tools, often in collaboration with experts in these disciplines, there soon becomes a point where the biomedical researcher themselves must take the lead in data analysis and understanding. This point is typically when biological and clinical data must be integrated with the expression data, and when the results must be interpreted in the wider context of other prior knowledge (typically from a range of public databases).

MADRAS has proven to be a highly usable platform for microarray research. Its browser interface has proven popular and it allows easy cross-platform compatibility, along with easy access from multiple locations. Linking together the required resources for analysis into a single "snapshot" enables rapid data exploration stemming from meaningful biological questioning combined with analytic tools to add confidence in any patterns emerging.

Whilst issues of usability and accessibility are not at the forefront of many developers working in the field, the extra work required to achieve these assets can only be view as a positive step, enabling the tool be applied to the drivers of the biomedical research, the biological researcher.

Chapter Eight
Summary and Discussion

8.1 Microarray data analysis

There has been an explosion in the development and provision of microarray technology in the last decade, with the number of probe sets represented on each generation of Affymetrix GeneChips following Moore's Law and doubling every two years (Affymetrix, 2000).

There has been great investment by the research community into establishing microarray services within many institutions, and provide the expertise to assist researchers undertake experiments on a chosen platform. The protocols to undertake the biological process in the experiment have been well characterised and much expertise exists as to *"best practice"*.

However, there has been a perceived lack of investment and exploration into the effects of the varying analysis approaches on the quality of results obtained, and guidance on *"best practice"* for data analysis. A lack of firm guidance, along with a lack evidence to demonstrate the comparative effectiveness of one methodology versus another has created a knowledge and guidance gap.

Many microarray services do not have the expertise to analyze the data nor the long term interest to learn it and as a result many researchers are left in a position where they can generate vast amounts of data, to which they have little idea how to analyse and understand the results it contains. Consequently, the researcher may be left feeling let down by the technology and those who deploy it.

Following a back to basics approach the work undertaken in this thesis has attempted to introduce the reader to the many steps involved in the analysis of a microarray experiment centred on the detection of differential gene expression once it has left the experimental system, and interrogate the effect of each stage on the ultimate experimental outcomes.

8.1.2 Analysis stages to determine differential gene expression

There are many numerical analysis stages involved in the process of converting the data obtained from the scanned array image into the desired results from an experiment. The analysis required for a typical analysis to identify differentially regulated genes between two groups of data can be broken down into a five step process:

1. Application of an expression metric to convert the information contained within the scanned array image into a summary expression signal level for each probe set.

2. Mathematical transformation of the resultant dataset to overcome distributional issues.

3. The utility of a normalisation stage to reduce comparative variability between experimentally similar data and address differences in data distributions between arrays.

4. The application of techniques to identify differentially regulated genes, which are typically returned as a list of *"interesting genes"*

5. Exploration and annotation of these "interesting genes" in an attempt to a draw meaningful biological conclusion to the observations and substantiate the hypothesis of the experiment.

In this thesis a variety of combinatory techniques were applied to stages one to four in order to determine the effects of different options on the detection of genes with known differential expression, with the goal of providing guidance on the issues involved with an analysis, and suggestions on techniques to consider dependant on experimental design (Chapters One to Six).

Completing the experimental process with the annotation and exploration of the results from previous stages requires the incorporation of information from a wide range of other sources in time efficient manner, along with tools to assist in the interpretation of sizable amounts of information (Chapter Seven).

8.3 Exploring best practice using the Affymetrix Latin square dataset

Datasets suitable for the exploration of the many analysis options presenting to the researcher, along with information regarding the expected results are few. In this work, a single dataset (the Affymetrix HG-U95A Latin square dataset) was identified and exploited to explore the effect of many different analysis issues and combinations. At each stage the results were reviewed and analysis undertaken to assist in the determination of the optimal methodologies that could be applied for the analysis of an experiment designed to determine the differential expression of genes between groups of data.

8.2.1 Factors influencing the choices of expression metric

Initial investigations into data distributions were applied to data from six expression metrics; MAS 4.0, MAS 5.0, dChip PMMM and PM-Only models, RMA and gcRMA used to analyse the complex background component of the 59-chip Latin square dataset. These results indicated that results from MAS 4.0, both dChip models and RMA on the whole correlated well with the assumption of normality; however data from MAS 5.0 and gcRMA presented with a proportion of the data with marked non-normality.

Investigations into the application of commonly applied statistical tests (fold-change, t-test and Mann-Whitney test) to determine the power to detect fifteen known two-fold changes with a 24-chip subsection of the data indicated that analysis using MAS 5.0, RMA and gcRMA consistently outperformed the earlier empirical MAS 4.0 and model based expression metrics of dChip. It should be noted that the dChip PM-Only model did perform markedly better than the PMMM model, suggesting that inclusion of the MM probes reduces the clarity of detection within the dataset. Use of MM match probes would appear to add to the technical variation observed, and explains the improved power of detection with expression metric utilising only the perfect match probe information.

A potential caveat in the identification of MAS 5.0, RMA and gcRMA being most successful in the detection of the spiked-in data in the dataset, is the fact that the dataset used in these investigations formed a key part of the development process for each algorithm, and the values of tuneable parameters were influenced according to the features of this dataset.

However, the findings correlate well with work by Chloe et al. () who explored many different combinations of the many stages involved in expression metric analysis using a fully synthetic spiked-in experiment (200 spiked in RNA with known fold-change, and a background of 2551 transcripts with known concentration). They concluded that the optimal set of algorithms for maximal spike detection included the background correction and perfect match probe adjustment from MAS 5.0, followed by the application of the median polish steps from RMA (although the MAS 5.0 expression summary was also a top performer).

Although gcRMA consistently outperformed the power of the other metrics to detect the spikes in this dataset, it is a relatively new methodology, which has presented with data distributions which may cause issues with the application of subsequent analysis stages, and it is suggested that further exploratory analysis using real-world datasets is required before full confidence can be achieved in this metric. In contrast, MAS 5.0 and RMA are established expression metrics which present with good power of detection and validation using other datasets. The ultimate choice in which expression metric to be applied must be determined having considered whether the researcher wishes to accept the small loss in detection power incurred by the inclusion of the data obtained from the miss-match probes.

8.2.2 Application of post-metric analysis normalisation

Applied to the homogenous Latin square dataset, the observation that application of post-metric analysis normalisation using QQ-normalisation, VSN and a rank based method produced very little effect on analysis outcomes and power of detection is unsurprising.

However, the application of a post-metric normalisation step in a typical biological experiment may form an important analysis stage which will enable data to be comparable and be a powerful ally to facilitate the extraction of meaningful and accurate analysis results reflecting the likely biological changes with the systems under exploration.

Normalisation can be applied in a variety of ways, dependant on the type of nature of the variability that requires control and reduction. Normalisation can be applied to the full dataset which can be useful if there is the presence of a few chips with differing distributions to other chips in the experiment, or to the arrays within a single group of data where the nature of the biological samples with that group suggests that high variability may occur with a subsection of supposedly identical observations.

The choice as to whether to apply normalisation stages as part of a data analysis should be influenced by issues of data variability, and a series of exploratory tests to examine the nature of the resultant data is suggested (e.g. MVA plots, distribution plots) to determine the sources of variability within a dataset undertaken. The results of these observations will influence the choice of normalisation method. The methodology behind QQ normalisation suggest its application to data with skewed or differing distributions, VSN is designed to overcome issues of differing variability between samples. Chloe et al. (Choe, et al., 2005) reported that the loess normalisation was their preferred method for post-metric analysis normalisation.

8.2.2 Data transformation

Data transformation has been a popular analysis stage, applied to microarray data in an to control distributional differences and to allow shift data distributions towards a more normal form, allowing for the application of parametric statistical tests. In Chapter Two it was shown that from a distributional point of view there was little or no benefit to the logarithmic transformation of data from any of the expression metrics. It was noted that whilst the $\log_2$ transform included in RMA and gcRMA did not particularly benefit the correlation with normality, there was little to dispute the author's inclusion of this measure in their methods (Irizarry, et al., 2003; Wu, et al., 2004).

Data presented in Chapter Three, looking at the effect of transformation on the detection of spikes within the Latin square dataset reveals no difference in the ability to correctly identify spikes in the dataset between logarithmically transformed and untransformed data. It was thus concluded that there is no a priori reason for transforming data prior to statistical analysis. Detection power in data from RMA and gcRMA were not altered by removal of the log transformation, so no reason was found to alter from the published methodology.

The effect of the logarithmic step within the RMA metric was re-visited in Chapter Six investigating the application of a Bayesian framework exploiting the relationship between mean and variance to better inform the parameters within the t-test and therefore improve power of detection. It was found that logarithmic step of the metric eliminated the mean/variance relationship and thus slightly diminished the power to detect the spiked-in transcripts of the dataset.

Overall little support was found for the application of a transformation step as part of each analysis. Whilst the removal of the transformation step from RMA data would appear to slightly improve detection power for data from this metric, the researcher may wish to retain the transformation step and apply the metric in the form the authors believe best represents the biological meaning within a dataset.

8.2.3 The utility of statistical tests to identify differential gene expression

A variety of statistical tests were applied to the 24-chip subsection of the Affymetrix Latin square dataset to compare the ability of each test to significantly identify the two-fold change in expression for each of 15 spiked-in transcripts. The relative power of each test compared to others was unaffected by which expression metric was used for the first stage analysis of the CEL files. Compared to fold-change alone, statistical tests incorporating the spread of the data (in addition to just the location used in fold-change) significantly increase the power to detect the spiked-in transcripts.

8.3.3.1 Comparison of classical parametric tests versus the comparable non-parametric alternative

Issues of sample size have made the parametric t-test a popular choice for the analysis of microarray data. Having validated the use of the test by examination of the data distributions (with caveats for the MAS 5.0 data, and marked non-normality with gcRMA data), two variants of the t-test (dependent on the assumption of equality of variance between groups) were compared to the non-parametric equivalent Mann-Whitney test across a range of sample sizes.

At small sample sizes (3-6 per group), the non-parametric test had significantly reduced power to detect the spikes, with the gap closing over the mid-ranged sample sizes (7-10 arrays per group). At the comparatively large sample sizes (11-12 samples per group), the power of detection was similar between the parametric and non-parametric techniques.

There was no difference in detection power between the versions of the t-test which do and do not assume heterogeneity of variance and as a link has been established between variance and signal magnitude it is suggested that researcher accept the potential for a very small loss in power, and utilise the Welch variant of the t-test which does not assume equality of variance between groups.

8.3.3.2 The utility of more robust variants of the t-test

A variety of variants of the t-test designed to accommodate for outliers in the sample dataset were compared using the same framework for assessing detection of spikes in the Latin square dataset. The results indicated that the utility of such tests is akin to throwing data away and shows comparative power to standard tests with a reduced sample size.

The results suggest that primary analysis of the array results and pre-filtering to remove or normalise outlying arrays is preferable over an approach that attempts to accommodate the erroneous data and produce meaningful results. The magnitude of sample size in a typical exploratory microarray experiment is not sufficient to allow for the application of these more robust statistical tests.

8.3.3.3 Implementation of a robust randomisation based testing method

Application of the randomised method, which uses randomisation to define the actual data distribution for the sample rather than comparison to a known distribution, to detect the spiked-in data of the Latin square dataset yielded similar results as those obtained from the Mann-Whitney test. When compared to the Welch t-test, the randomised test showed limited power at low sample sizes, slightly less power at middle sizes, and equivalent power at sample sizes over 9 per group.

Whilst the technique gives reasonable results, the overheads in time and computational power required for its application question the practicality of application within an typical research environment. The observation, that near identical results can be obtained using Mann-Whitney test which is more robust against outliers and distribution issues in lesser time further reduce the merit of randomisation based testing.

8.3.3.4 The utility of a Bayesian framework method

The Balid and Long (Baldi and Long, 2001) Bayesian framework utilised the relationship between the signal mean and variance within a group of data to better information the variance parameters of the t-test and improve power of detection. Comparing results from the Bayesian analysis on just three samples per group, to those from the standard Welch t-test indicate that similar power is seen to five arrays per group in the Welch test.

The difference in power between the tests reduces as sample size increase, but the Bayesian test continually outperforms the standard t-test methodology. Overall, the Bayesian framework presents as a very powerful analysis tool, allowing a research to extract the maximal amount of information from a cost-limited experiment and providing the greatest power of all of the statistical tests considered.

8.3 Drawing conclusions on "best practice"

Review of the work undertaken in this thesis (Section 8.2) identified a series of conclusions regarding the relative performance of differing combinations of analysis methodologies to detect truth in a dataset, and which can assist in the definition of *"best practice"* in the analysis of Affymetrix microarray data. Whilst the investigations provide an important corpus of information and background to the analysis of GeneChip data, there are limitations to the explorations undertaken and more general issues of experimental design which significantly contribute to the accuracy of any analysis.

8.3.1 Reviewing the inference obtained from the Latin square dataset

The Latin square data represents a very hard analysis situation, with the focus of investigations being on methods to detect a known number of un-related changes represented with just a two-fold change in expression signal between the two groups. More typically and experiment will contain many more probe sets with transcriptional differences between groups presenting with a variety in the magnitude of difference between the groups.

In addition, the nature of unchanged background data across all chips is uncharacteristically homogenous. In contrast to the single mRNA sample which was hybridised as a "complex background" for the Latin square experiment, the background in a typical experiment will accompany the differentially regulated data, and whilst much data will be comparable differences between cells and patient samples will produce a much more variable background overlaid with the *"real changes"*, which must be identified.

Biological variation is one of many compounding factors on the number of replicates, and is likely to be substantially more in many experiments than that from technical variance.

As a result, conclusion which can be drawn about the relative power of different techniques to detect the spikes, and to relative power dependant on sample size can only guide as to the best methods to overcome the technical variation within the system. Accordingly, when undertaking an experiment it is prudent to use the best tools to extract information from the data, but however sharp the tool, it will be unable to extract meaningful results if the variability between arrays is the overwhelming factor of the experimental observation.

8.3.2 Experimental design and sample size

Guidelines from the Affymetrix Best Practices Expression Analysis work group reported that to obtain high quality data which can be readily compared between studies, site and over time are dependant on three critical aspects of good experimental design, each designed to reduce the variability between samples, and therefore facilitate more accurate analysis. The three recommendations were the standardisation of tissue sampling, storage and processing procedures, experimental design which makes comparisons between equivalent tissue types and undertaking sufficient biological replicates. Each one of these stages is concerned with the reduction of variability between resultant array data.

Sample size is probably the single most important factor in microarray experimental design for a variety of reasons. The number of replicates in each sample group is key to achieving confidence in the obtained results and influences the choices of test which can be applied to a dataset. Whilst more replicates does result in a *"better experiment"* this improvement has a significant cost implication with each addition replicate addition over £500 to the experimental budget. There is therefore a requirement to identify sensible compromises to these two divergent requirements to the number of replicates, the researcher will often wish to undertake the minimum number possible, whilst the statistician would argue, the more data the better.

Although statistical methods exist for power-calculation the key piece of information these require is the variance expected within the system, as it requires experimental results to determine the factors, the researcher is left in a circle of cyclic logic. As an example, the `power.t.test` function in R requires four pieces of information; the number of observations per group, the Standard deviation for the dataset, the required significance level and the true difference in means. The relationship between variance and expression observed further obscures the matter, a single variance level cannot be determined for an experiment; as the information on the magnitude of change expected and prediction of signal levels would be required to fully derive a power calculation.

Guidelines on sample size are difficult to deduce because of the underlying variability in biological data and experimental design which will affect the power of a test to detect differences. It should also be pointed out that samples do fail, and there are QC issues which may require the dropping of certain chips.

Drawing conclusions regarding the number of replicates needed to achieve confidence in results is a complex matter, and it can only be advised that more replicates are needed than the attempts to define *"best practice"* indicate. The findings should thus be viewed as minimum estimates because of the clean and perfect nature of this dataset; the likely influence of additional biological variation would require additional GeneChips in an experimental design.

8.3.4 The requirements to better define "best practice"

Ultimately it is the quality of input to the microarray experimentation process that will have the greatest effect on the output. A well designed experiment built on a good understanding of the biology most likely to affect the results will yield the best results. However, the application of the *"best"* analysis techniques a well designed experiment will maximise the number of true results returned, and can assist in overcoming some of the inherent issues of undertaking many thousand experiments in parallel.

Key to the definition of *"best practice"* are reference datasets to with known truth to which differing analysis methodologies can be applied and the power of detection quantitavely described. The work in this thesis used the Affymetrix HG-U95A Latin square dataset as its truth dataset, which at the time of writing was the only publicly available dataset with known truth which can be sectioned to provide biologically meaningful sample sizes. However, when compared to other reference datasets (e.g. the GeneLogic spiked-in set) it is limited in the number of spiked-in samples.

Ultimately, the analysis community requires a variety of datasets to benchmark different analysis combinations against from a range of sample types, over a range of sample sizes, with a significant number of known changes. Such datasets should represent a variety of sample types, over a full range of relevant sample sizes, with differing fold-changes of spiked-in samples, over multiple groups of data to enable full exploration of the range of statistical techniques available.

Such an approach should yield datasets which can be used to both train and validate new models for expression analysis. The provision of a number of diverse datasets should also eliminate the current risk of over-training and optimisation of methodology based on the truth in a single dataset. It is ultimately possible that the optimal analysis methodology may vary according to experimental design, variability and the desired outputs from an experiment.

8.3.4 Future horizons in expression analysis

The work undertaken in this thesis has followed the traditional analysis approach of converting the image data into an expression value for each probe set, before the transformation, normalisation and analysis of the data to identify the differentially regulated genes. The processes involved in the conversion of the image data into signal values is a complex one, which as it has been shown can yield a variety of differing results which should be directly comparable.

The gcRMA metric is one example of how integration of information regarding the array design can improve results by incorporation of other well defined thermodynamic binding information. One key area for development is the integration of the information that each GeneChip array contained, and the techniques to produce the most biologically meaningful data.

Barrera et al. (Barrera, et al., 2004) have shown that the application of two-way ANOVA techniques to the probe level data can increase the power of detection by incorporating more data into the analysis and providing additional links between related experimental observations comprising the data submitted for analysis.

Reduction of the number of analysis stages required for an analysis can only assist in the reduction of variability of experimental outcome, and the probe level approaches which eliminate further stages of data reduction are likely to improve experimental efficiency.

However, control of variability would appear to be the biggest hurdle to each stage of analysis and experimentation, with variability occurring from samples, through processing, through arraying and through analysis. A badly designed and executed experimental cannot be accurately analysed, no matter how good the analysis approach is.

8.4 Making biological sense of analysis results

Initial stages of data analysis are concerned in the manipulation and processing of the numerical information that an experiment produces, and is involved in a substantial data reduction exercise, typically producing a list of *"interesting"* results in the form of a list of probe set identifiers. At this point the expertise in analysis shifts from that of the computational biologist back to the biomedical researcher who must take the lead in making sense of the data analysis and either form or confirm hypotheses from the results.

Whilst previous stages have been concerned with data reduction, the process of exploring the results will typically require the integration of additional data from a variety of sources to assist understanding. The nature of the task, taking a researcher into areas of biology that they may not be familiar, coupled with the limitations of available tools and exploration environments make this a significant task in terms of both time and difficulty.

8.4.1 Integration of resources to improve exploration efficiency

Work on the MADRAS (Microarray Data Review and Annotation System) environment centred on the development of a rapid exploration environment which would enable researchers to question their data, review their results along with selected relevant annotation fields, and provide analysis tools to assist in the complex task of drawing conclusions regarding the biological processes contributing to the observed results.

Such an environment forms the final part of a microarray analysis workflow, taking in the results from the experiment, along with those from other analyses and statistical tests and aims to accelerate the time spent deducing conclusions for further experimentation or publication from the data. Feedback from the local user community suggests that the concept and implementation bridges a gap in current workflow systems and has proven a very useful research tool.

The annotation in MADRAS was chosen after consultation with a variety of users as to the information they would find most useful to rapidly understand the transcript being represented by an Affymetrix probe set. It is important that this process is ongoing as advancements in genome annotation may alter the importance of databases and the fields they contain.

As an example the NCBI LocusLink database is in the process of transition to EntrezGene which aims to form and provides a unified query environment for genes defined by sequence, however as a result certain data made available in the LocusLink database have been removed and must therefore be sourced from alternative locations.

The formation of a comprehensive annotation may prove a useful resource which can be applied and utilised to a wide range of emerging high-throughput technologies as many of the issues of interpreting and exploring data are likely to be shared issues with common solutions.

8.4.2 Improving overrepresentation analysis

Whilst the annotation and exploration of experimental data is key to completing analysis of a dataset, there is also a need to place such data in the context of the wider annotation and metadata resources available to the researcher. Although it is possible to undertake manual literature review using the information contained within the MADRAS annotation summary, this is a labour-intensive process which is subject to user bias. To assist in overcoming these barriers, MADRAS incorporates over-representation analysis applied to an extended range of data fields, gene annotation and pathway data.

It should be pointed out however, that there are limitations to the techniques used to assess any significant in the data within a probelist. The basis of analysis within MADRAS is the presence of linked significant results with defined linked between the annotation genes, either via association into a pathway, or via the textual description. As a result, a set of well annotated yet less signiciant genes are likely to be identified over a single probe set as significant, due to issues of data overload and systematic bias in analysis approaches.

A potential future avenue for this type of analysis is the scoring of annotation quality, so that well explored parts of the genome do not over-shadow less represented, but potentially more important parts, with the possibility down weighting probe set with a quality and prevalence of annotation.

Whilst, automated analysis does prove a useful and convenient tool in many analysis scenarios, it does underpin the requirement for the researcher to fully understand their experiment and the analysis applied, and to fully explore the results from a variety of angles in contrast to a *"black box"*, data in answers out system which may report significant, yet ultimately uninteresting results.

8.4.3 Building and extending the MADRAS system

MADRAS allows the facile exploration of expression data from one specific microarray platform in a gene-centric fashion in combination with key metadata from several public databases (LocusLink and OMIM in particular). Work on MADRAS has demonstrated the benefits of combining experimental data, theoretical data and annotation, along with analytical tools aimed at assisting the interpretation of a dataset, in a single rapid exploration environment. A logical progression of this work would be the application of the obtained knowledge and experience to expand the system to accept data from other microarray platforms and technologies.

Overcoming the limitations of the over-representation analyses could potentially be achieved by the incorporation of additional information to supplement the currated data contained in the annotation fields. One potentially interesting development idea would be the extension of mass-parallel annotation mining tools to larger corpuses of text (e.g. the PubMed abstracts text corpus) (Chaussabel and Sher, 2002). Such an approach would require a combination of data mining, statistical and natural language processing (NLP) techniques with the aim to identify the more subtle and detailed links between genes which are lost in reduced data that is contained within many annotation summaries (Jenssen, et al., 2001).

High throughput technologies in biomedical science are yielding huge volumes of experimental and clinical data including gene, protein, chromosome and tissue information. As they investigate the same underlying biology each of these techniques are ultimately complimentary, thus improving the insight that can be achieved. The evolving challenge of biomedical informatics is to begin to integrate different sorts of experimental data with public annotation data, along with analysis tools to make the best possible conclusions from the assembled information.

By drawing information together and then analysing it in a mass-parallel approach, patterns in the observations can be assembled and then explored for similarity yielding new insight into the underlying mechanisms for disease. Microarray data, and the MADRAS system, form only one part of the overall biological picture that can be obtained using current experimental techniques. It is therefore hoped that the successes of MADRAS and its fundamental ideas can be developed to incorporate data from other technologies allowing parallel exploration of results from multiple systems.

An example would be the incorporation of mass-parallel proteomics data in a situation, where a researcher may have mRNA and protein level data from the same sample and wishes, for example, to ask how the genes that are changed at the mRNA level behave at the protein level, and vice versa.

Fundamentally, the issue of providing tools to the biomedical researcher which can accelerate the experimental process is one worthy of much attention by the bioinformatics community in parallel to the large amount of algorithmic and statistical research being applied to the high-throughput technologies. Ultimately it is the interpretation of the biological data that should be providing the key research driver.

8.5 Conclusions

Microarray based expression profiling provides a useful research tool to gain new insights into biological systems. During the period that the work presented here was undertaken, data analysis methods have greatly advanced, with more on developing the accuracy of analysis, and the techniques which can best extract meaningful information from the data. However, with many more analysis experts entering the field and wishing to make a mark, there is a risk that answers to basic questions are over shadowed by work on more complex computational and statistical methods and many simple yet powerful methods are overlooked.

Understanding the fundamental concepts is essential in the application of statistical tests and to underpin future work aiming to link microarray data to its biological annotation. Although it can be argued that bioinformatics software is best written by trained bioinformaticists, computer scientists and statisticians, it is ultimately the researcher who created the experiment and commissioned dataset who must be empowered to undertake the data analyses as they are the ones who wish to best capitalise on their investment.

To the researcher, there is the requirement of a large investment of time and energy required to understand microarray analysis and its implementation. To this end, many groups have made decisions to *"out source"* their analysis and rely on external expertise. The risk of such an approach is that a lack of understanding of the limitations and factors affecting analysis will lead to sub-optimal experimental design and implementation, which will results in frustration from the user when their data is returned without significant meaningful results. A similar outcome is likely from black box analysis approaches.

Consequently, a middle ground must be found where the required user knowledge is modular, so they can understand the system at whatever level they require for comfortable acceptance of any additional leaps of faith. To back up this approach the analysis community needs to be able to provide firm answers to simple questions raised by the non-expert user and practical substantiated guidance as to best practice.

Adoption of the elements of good experimental design, good experimental processing, best practice analysis and tools to fully exploit the information returned from an experiment should enable the full potential of microarray technology to be realised, allowing the researcher to move from a system able to detect the obvious transcriptional changes in a system, to one which can detect even the smallest of nuances between sample groups and assist in the further annotation and understanding of genome functions.

Chapter Nine
Materials and Methods

9.1 Introduction

The following sections provide supplemental information to support the exploration and results in previous chapters. These sections aim to inform the reader of the technical methodology behind each stage of analysis and identify the key functions and ideas utilised. In addition, further information is provided on the datasets and expression metrics chosen for analysis and comparison.

9.2 Data distributions and their effect on analysis options

The following information supports the work undertaken in Chapter Two which is concerned with the assessment of data distributions on Affymetrix GeneChips data.

9.2.1 Introduction to the Affymetrix Latin Square dataset

The Latin Square data was release by Affymetrix and formed part of the development cycle and validation for the MAS 5.0 software suite (Affymetrix, 2001b; Affymetrix, 2002a; Affymetrix, 2002b). The dataset consist of a series of transcripts spiked-in at known concentrations and arrayed in a Latin Square format along with a complex background of human pancreatic mRNA.

The supporting documents from Affymetrix state that the Latin Square design for the human data set consists of 14 spiked-in gene groups in 14 experimental groups. The concentration of the 14 gene groups in the first experiment is 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1024pM. Each subsequent experiment rotates the spike-in concentrations by one group; i.e. experiment 2 begins with 0.25pM and ends at 0pM, on up to experiment 14, which begins with 1024pM and ends with 512pM. Each experiment contains at least 3 replicates, with two concentrations being represented with twelve replicates.

The 59-chip data set was downloaded from the Affymetrix website (*http://www.affymetrix.com/support/technical/sample_data/datasets.affx*) as a series of pre-processed cell intensity (CEL) files, each containing a list of intensity values for each oligonucleotide probe on a GeneChip.

9.2.2 Overview of Expression Metrics

Six different algorithms were applied to extract a single expression values for each transcript represented by a series of probes on the image data. In addition to a quantitative expression value, four of the analysis methods also provide a qualitative measurement indicating if the transcript is detected (Present), not detected (Absent), or marginally detected (Marginal). The expression values and calls for relative gene expression (Absent, Present or Marginal) were exported from each package into delimited text files. Further details of the algorithms and methodology for each expression metric follow in Sections 9.2.1.1 – 9.2.1.6.

9.2.2.1 Microarray Suite 4.0

Microarray Suite (MAS) 4.0 (Affymetrix, Santa Clara CA) calculates expression values using an empirical method summing the PM-MM value and correcting for background noise. The expression level outputted by MAS 4.0 is termed Average Difference.

The Average Difference for each probe set is an average of the differences between each PM probe cell and its control MM probe. Thus, in those probe sets where the mismatch probe has a higher intensity value than the perfect match the Average Difference will be negative. In general, MM is greater than PM for about 1/3 of the probes on any given array (Irizarry et al., 2003a)

MAS 4.0 also generates an Absolute Call which indicates a confidence for the Average Difference being sufficiently different from background for the probe set to be considered expressed. A decision matrix is employed to determine the presence or absence of each transcript on each chip.

In this study, data was analysed using global normalisation and target intensity (TGT) of 100 and exported to tab delimited text file within Microarray Suite 4.0. TGT is an arbitrary target intensity value which each chip is scaled to in order to facilitate comparison between chips.

9.2.2.2 dChip PMMM Model

Li and Wong (Li and Hung Wong, 2001a; Li and Wong, 2001b) were the first to propose model-based expression measures. They observed that PM values are often less than MM values and identified the need for non-linear normalization, and proposed the use of multi-array summaries for the detection and removal of outliers.

dChip attempts to make improvements over the MAS 4.0 algorithms by application of an invariant set normalization method and model-based expression values (MBEI). The MBEI value is the weighted average of PM/MM differences. dChip also applies an outlier detection algorithm to eliminate values from potentially cross-hybridizing probes and subtracts away the cross-hybridization signals equally from a probes PM and MM value, that results in sensitivity to expression changes at the low concentration level. However, in many cases the resultant expression values are similar to those obtained from MAS 4.0 as Average Difference values.

CEL file data was analysed in dChip version 1.3, using the PMMM model with default settings, incorporating default median chip normalisation and application of the MBEI model.

9.2.2.3 dChip PM-Only Model

The dChip package contains algorithms which can calculate an expression value using either both the PM and MM probes, or the PM probe values only, dependent on analysis settings. The PM-Only algorithm uses only PM probes and results in the production of all-positive expression values, which is important for researchers wishing to apply fold-change analysis. However, the model has a slight caveat, with the absorption of some background signal into the expression values due to the MBEI calculating the weighted average of background-adjusted PM values. This effect reduces sensitivity at lower concentration levels (Li and Wong, 2001b; Naef, et al., 2002).

CEL file data was analysed in dChip version 1.3, using the PM-Only model with default settings, incorporating default median chip normalisation and application of the MBEI model.

9.2.2.4 Microarray Suite 5.0

In response to consumer demand to eliminate negative values and produce a more robust expression metric, Affymetrix integrated new statistical algorithms into MAS 5.0. MAS 5.0 utilises a series of statistical techniques in the conversion of image intensity data across a series of probes into an expression value, and steps are taken to avoid the production of negative numbers. Average Difference was replaced with a new expression level termed Signal (Affymetrix, 2001b; Affymetrix, 2002a; Affymetrix, 2002b).

The analysis process first undertakes a zoned based background correction using the lowest 2% of probe intensity values. These values are then smoothed to ensure an even

transition across the chip. For each probe, an ideal mismatch value is calculated (to eliminate the possibility of negative values) and subtracted to adjust the PM intensity for non-specific hybridisation. The adjusted PM intensities are log-transformed to stabilize the variance before application of a One-Step Tukey's biweight estimator, to yield a robust mean of the values. Signal is outputted as the trimmed mean of the antilog of the biweight value.

The Absolute Call from MAS 4.0 is replaced with a detection call and associated detection p-value. Detection p-value is calculated using a Wilcoxon rank test of the probe values, looking for a statistical difference between the MM values and the PM values. The detection call is determined by cut-offs applied to the p-value and are user tuneable.

In this study, data was analysed using default parameters within Microarray Suite 5.0, with the application of a TGT of 100.

9.2.2.5 Robust Multi-chip Average (RMA)

The Robust Multi-chip Average (RMA) method (Irizarry, et al., 2003) consists of three steps: a background adjustment, quantile normalization and finally summarization. RMA can be broken down into a three step process, firstly probe-specific background correction to compensate for non-specific binding using the PM probe distribution (it should be noted that the background adjustment step in RMA ignores the MM values present on each array).

Secondly, the application of a probe-level multi-chip quantile-quantile (Q-Q) normalization to unify PM distributions across all chips and finally the generation of a robust probe-set summary by reducing information for individual probes to a single value for the probe set using the log-normalized probe-level data by median polishing and a robust multi-chip linear model fit on the log scale.

RMA analysis was implemented using the *"affy"* module (Gautier, et al., 2004) of the BioConductor project (Gentleman, et al., 2004) implemented within the R environment (Ihaka and Gentleman, 1996). Whilst many different analysis models have been developed by the BioConductor community, default settings were employed for background correction, normalisation (Q-Q) and polishing.

9.2.2.6 gcRMA

gcRMA is a development of the RMA model which aims to improve the background correction by incorporating information about predicted hybridisation (Wu, et al., 2004). The method builds on the work of Naef and Magnasco (Naef and Magnasco, 2003) who propose a solution useful for predicting specific hybridization effects using the base composition of the probes. Probe affinity is modelled as a sum of position-dependent base effects

The model takes advantage of sequence information to appropriately describe non-specific binding variation. The authors claim that this practical adjustment results in improvement in overall sensitivity and specificity when compared to other methods (MAS 5.0, RMA).

In this study, data was computing using the gcRMA module of BioConductor (Gentleman, et al., 2004) implemented in R (Ihaka and Gentleman, 1996) using default settings, and options for prior computation of the binding affinities from the sequence information provided by Affymetrix.

9.2.3 Data filtering and preparation

Data from the 14 spiked-in genes (37777_at, 684_at, 1597_at, 38734_at, 39058_at, 36311_at, 36889_at, 1024_at, 36202_at, 36085_at, 40322_at, 407_at, 1091_at, 1708_at) and 67 control probe sets (with AFFX prefix) were removed from the resultant files from each of the six expression metrics, leaving 12545 probe sets for further analysis.

9.2.4 Technical Methodologies

9.2.4.1 Formal testing of normality

Formal testing of normality was undertaken by application of the Shapiro-Wilks test to the 59 values for each probe set using the `shapiro.test` function in the R base system (Ihaka and Gentleman, 1996). This analysis was repeated for data from each of the six expression metrics under review.

9.2.4.2 Assessing correlation to normality

Correlation to normality was calculated as the Pearson's R^2 value from a Quantile-Quantile (QQ) plot of the 59 data values for each probe set against 59 values generated to be representative of a normal distribution. Analysis was undertaken using the R function `rnorm` to generate a normally distributed dataset, which was correlated each probe sets data using the `cor` function.

9.2.4.3 Calculation of skew

The skew for the data from each probe set was calculated using the `skewness` function from the e1071 library (Dimitriadou, et al., 2004).

9.2.4.4 Application of the Box-Cox transformation and assessment to normality

The Box-Cox function was applied to data from each probe set using values of lamba from -2 to 2 in steps of 0.1 (Section 2.3.6). Code to undertake this process was written using the R language and the results of the transformation assessed for correlation to normality using QQ plots (Section 9.2.4.2). The value of maximal correlation to normality was returned as the result of the Box-Cox function.

9.3 Statistical approaches to the detection of differentially regulated genes in Affymetrix datasets

The following information supports the work undertaken in Chapter Three which is concerned with the determining the power to detect "truth" in a dataset with known spiked-in values. The initial image data was subject to analysis using a variety of expression metrics and testing to determine differential gene expression across a range of sample sizes.

9.3.1 Defining the exploratory dataset

Reviewing the associated documentation of the Latin Square experimental design *(http://www.affymetrix.com/support/technical/sample_data/datasets.affx)* identified replicate chips where most spikes vary by only 2-fold. Further review of the information indicated that each of these, four replicates in the Latin square design were replicated over three separate chip wafers, resulting in 12 replicates in each group for the 14 spiked in transcripts with a 2-fold change in concentration.

Table 9.1

	Spiked transcript concentration (pM)	
Probe set	Chips M, N, O, P	Chips Q, R, S, T
1597_at	0	0.25
38734_at	0.25	0.5
39058_at	0.5	1
36311_at	1	2
36889_at	2	4
1024_at	4	8
36202_at	8	16
36085_at	16	32
40322_at	32	64
407_at	512	1024
1091_at	128	256
1708_at	256	512
37777_at	512	1024
684_at	1024	0

Table 9.1 – Spiked transcript concentration for the spiked in data of the Affymetrix Latin Square dataset.

9.3.2 Review of the spiked-in transcripts

In Section 9.2.1 considering data distributions it was stated that data for fourteen probes spiked in by Affymetrix was excluded prior to investigations. However, review of the literature (Cope, et al., 2004; Wolfinger and Chu, 2002) suggests inconsistencies in the information provided by Affymetrix.

The data suggests that 407_at and 37777_at follow the same spike-in pattern, and probe set 33818_at fills the missing gap in the Latin Square grid. Probe set 546_at is designed against the same target as 36202_at (Unigene ID Hs. 75209) and the data support this observation. The Latin Square dataset can thus be considered to contain sixteen probe sets which contain known truth in terms of the spiked-in transcript according to the Latin Square design (407_at, 546_at, 1708_at, 1024_at, 1091_at, 1597_at, 33818_at, 36085_at, 36202_at, 36311_at, 36889_at, 37777_at, 38734_at, 39058_at, 40322_at, 684_at).

Extracting the information for the 24 chips containing replicates across two groups of 12 chips shows that the majority of probe sets present with a two-fold change in the spiked-in transcripts, with the exception of probe set 684_at, which is represented with spikes of 0 and 1024 pM in the two groups. This probe set was therefore removed from further analysis, leaving fifteen probe sets for detection.

9.3.3 Overview of analysis

The aim of the analysis undertaken in this section of work was the real world exploration and validation of the ability of combinations of expression metrics and statistical testing to identify known changes in a dataset. Such an analysis also allows for investigations into the issue of data transformation.

In addition the number of replicates present in the Latin Square dataset allow for the consideration of the relative performance of differing analysis approaches using experimentally plausible sample sizes (ranging from 3 to 12 replicates in each group representing a single state).

Key to the approach is the identification of methods to identify the detection of "truth" from the analysis results, and make meaningful comparison between expression metrics and approaches to detection differential gene expression.

9.3.4 Performance assessment using FDR curves

In order to determine the relative performance of differing methods to correctly identify the known truth in the form of our 2-fold change in expression of the 16 spiked in probes FDR curves were generated visualising the performance in differing tests from each resultant dataset and at a range of sample sizes.

FDR curves are based on the idea of ROC (Receiver Operator Characteristics) curves plotting specificity against sensitivity. ROC curves are generated using the whole dataset and were found to produce small differences because of the small amount of known truth in the dataset. Instead the FDR curve focuses on a smaller amount of the dataset, in our case the first 200 hundred genes ranked on the basis of statistical significance from a test statistic - behaviour after that is not recorded. A graph is then plotted scoring one mark vertically for detection of known truth and one mark horizontally for a false positive.

This provides a more intuitive visualisation of the performance of each method and is in the order of magnitude that are likely to be considered for follow up study using complementary methodologies, i.e. it is quite acceptable to review a list of 200 genes for further annotation and follow up. Using this method we can get an idea of the false positive rate in such a strategy. In order to quantify the results of the many graphs we can use the area under the curve to quantify performance. A perfect result gives the largest area where the 16 spikes are the first probe sets identified in the ordered list, less perfect tests yield smaller areas (Figure 9.2)

9.3.5 Development of an analysis framework

The analysis framework can be broken down into three main processing stages, generation of random samplings from the data set, application of expression metric analysis and application of tests for differential gene expression and the assessment of power using FDR curve analysis. The results of each stage for the inputs to the next, with various variations applied before submission to the next stage.

9.3.5.1 Random sampling of data at a variety of sample sizes

To obtain an idea of confidence that can be inferred from the results of downstream testing, the 24 chip dataset was randomly sampled to produce 20 replicate datasets over a sample size range of 3-12 (Figure 9.1). The result of this sampling was an index of filenames at each of the 200 samplings.

Figure 9.1

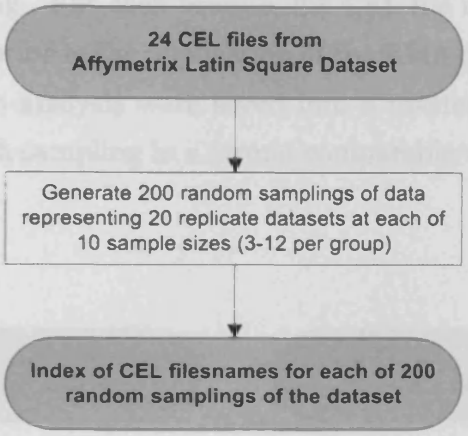


Figure 9.1 – Overview of sampling from Latin Square dataset.

9.3.5.2 Application of expression metric analysis to each dataset

The index produced in 9.3.5.1 was combined with the CEL file data to apply differing expression metrics to the data. The result in each case was a 200 element array containing matrices of expression values for each of the random samples previously generated. Analysis using MAS 4.0, MAS 5.0, and the dChip models was undertaken in the respective software package (as opposed to Bioconductor implementation) as a single analysis run of all 24 chips. The resultant text files were then imported into R along with the index data and matrices produced by extracting the relevant columns from the uploaded file (Figure 9.2). For each expression metric a master array was formed containing the expression data for each sampling.

Figure 9.2

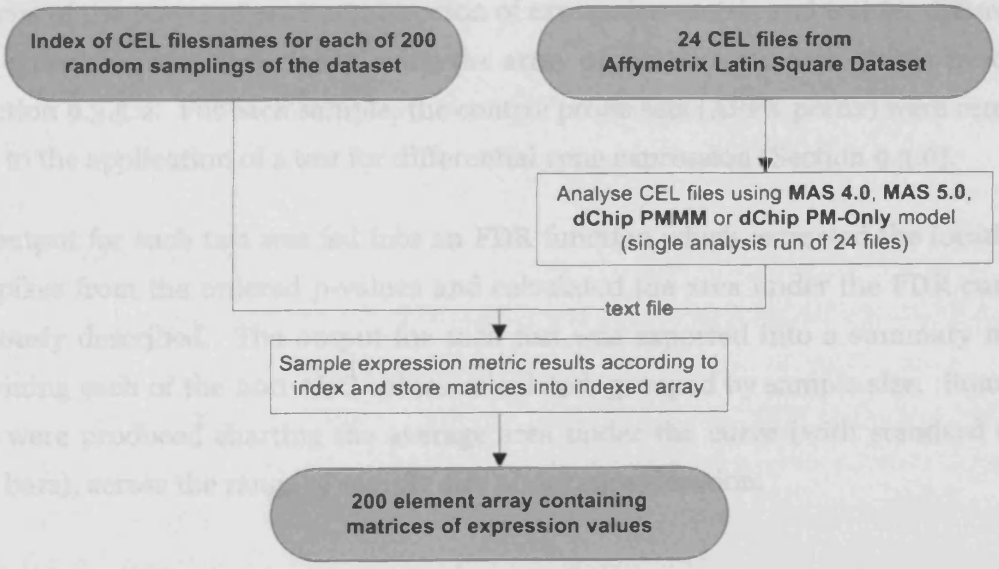


Figure 9.2 – Overview data analysis using Microarray Suite and dChip expression metrics.

The multi-chip normalisation undertaken in RMA and gcRMA required the individual analysis for each sampling. For each sample, the CEL file data was imported into R using the `affy` module, prior to the application of the RMA or gcRMA analysis (Figure 9.3). The results of each analysis were saved into a master array containing all the expression results for each sampling in a format comparable with that from Microarray Suite and dChip analysis.

Figure 9.3

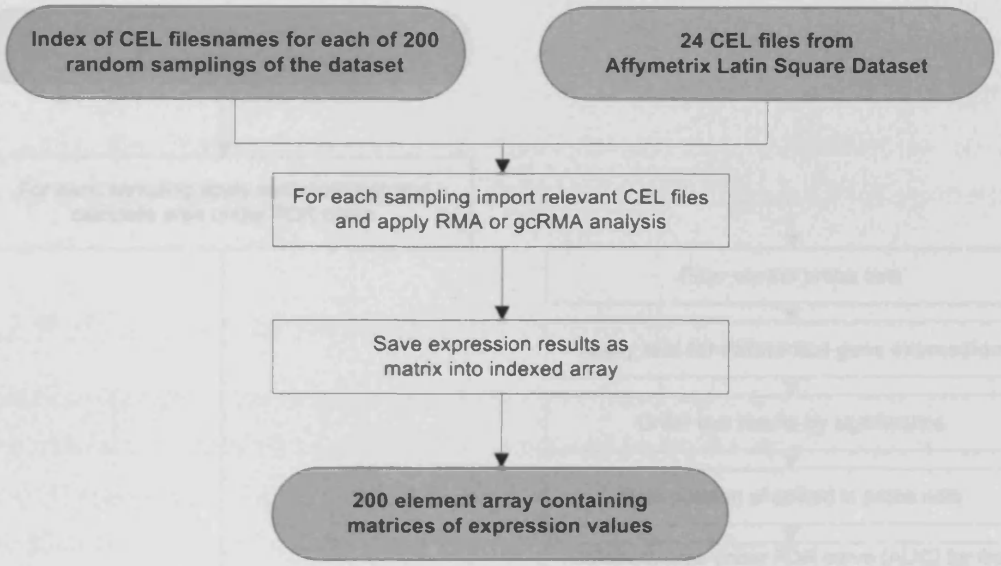


Figure 9.3 – Overview data analysis using RMA and gcRMA expression metrics.

9.3.5.3 Application of tests to determine differential gene expression and calculate power

Analysis of the power of each combination of expression metric and test for differential gene expression was undertaken using the array of matrices for each metric produced in section 9.3.5.2. For each sample, the control probe sets (AFFX prefix) were removed prior to the application of a test for differential gene expression (Section 9.3.6).

The output for each test was fed into an FDR function which extracted the location of the spikes from the ordered p-values and calculated the area under the FDR curve as previously described. The output for each test was exported into a summary matrix containing each of the 200 AUC values calculated, grouped by sample size. Summary plots were produced charting the average area under the curve (with standard error, error bars), across the range of sample size under consideration.

By combining the results from different combinations of expression metric and test for differential gene expression difference in power can be observed by the relative positioning of the lines on the plot. By implementation of error bars using the standard error of the mean, the presence of a gap between error bars at a chosen sample size is equivalent to a significance value of $p < 0.05$ for a standard paired t-test on the same data (Motulsky, 1999).

Figure 9.4

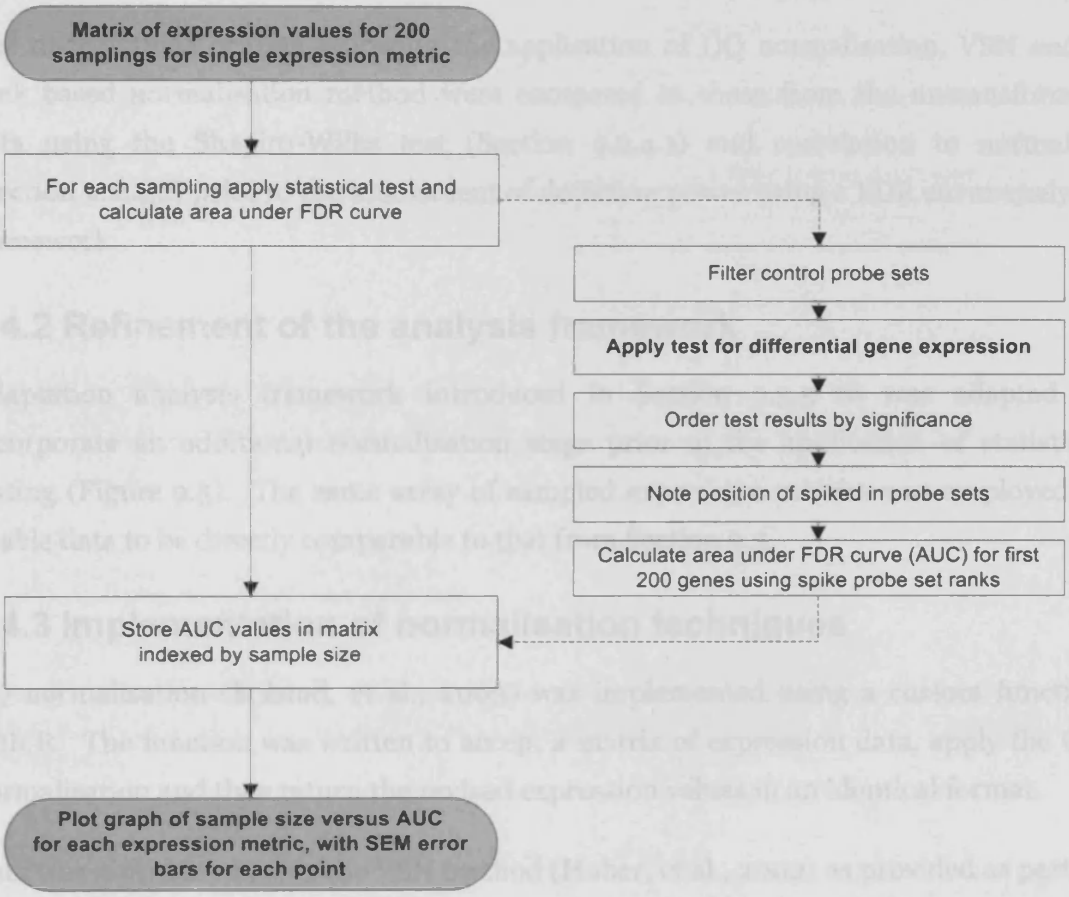


Figure 9.3 – Overview of FDR analysis for each expression metric and test for differential gene expression.

9.3.6 Statistical testing

Fold change analysis was undertaken using a custom function which calculated the absolute value of the ratio between the logs of the means of each group of data. Analysis of data using the Welch t-test and unpaired t-test were initially undertaken using the R `t.test` function, however testing identified a performance advantage when using customised code for the calculation values for t and degrees of freedom and using the `pf` function to calculate p-values. The non-parametric Mann-Whitney test was implemented using options of the `wilcox.test` function.

9.4 Approaches to data normalisation

The following information supports the work undertaken in Chapter Four investigating the effect of differing data normalisation techniques to improve the power to detect "truth" in detection of the spiked-in transcripts within the Affymetrix Latin Square dataset.

9.4.1 Examination of data distributions following normalisation

The distributions of data following the application of QQ normalisation, VSN and a rank based normalisation method were compared to those from the untransformed data using the Shapiro-Wilks test (Section 9.2.4.1) and correlation to normality (Section 9.2.4.2) prior to the assessment of detection power using a FDR curve analysis framework.

9.4.2 Refinement of the analysis framework

Adaptation analysis framework introduced in Section 9.3.5 to was adapted to incorporate an additional normalisation stage prior to the application of statistical testing (Figure 9.5). The same array of sampled expression metrics was employed to enable data to be directly comparable to that from Section 9.3.

9.4.3 Implementation of normalisation techniques

QQ normalisation (Bolstad, et al., 2003) was implemented using a custom function with R. The function was written to accept a matrix of expression data, apply the QQ normalisation and then return the revised expression values in an identical format.

Data was normalised using the VSN method (Huber, et al., 2002) as provided as part of the bioconductor VSN module. Similarly to the QQ normalisation, the function returns a matrix of identical format to the input matrix with revised values following the application of variance stabilisation normalisation.

The relative rank of data within each array was calculated by the application of the R function `rank` to each column within the sample under analysis. The resultant matrix being of identical format to the input linked to the next stage of analysis, the application of statistical testing.

Figure 9.5

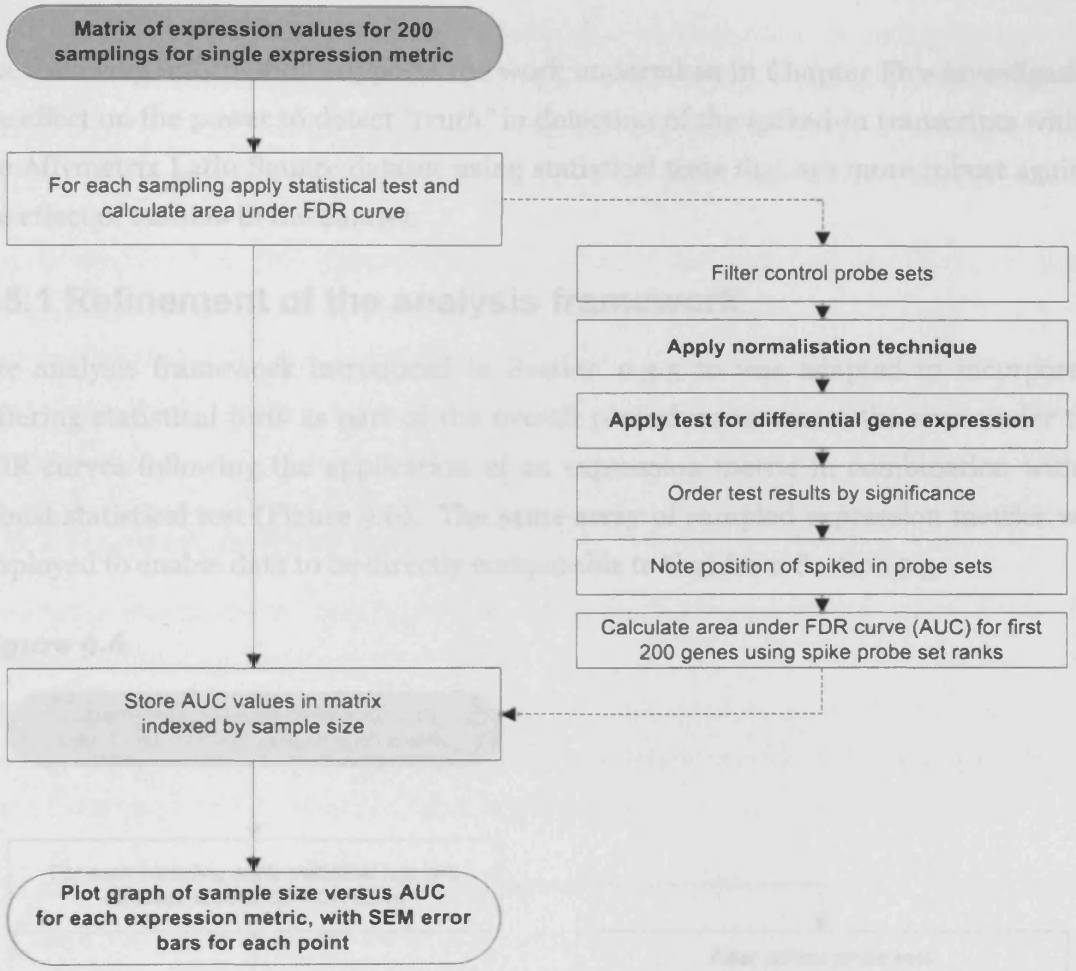


Figure 9.5 – Overview of FDR analysis for each expression metric and test for differential gene expression following the additional application of a normalisation technique.

9.4.4 Implementation of statistical testing

FDR analysis of data following the application of QQ normalisation and VSN were assessed using the Welch t-test implemented in the customised code identified in Section 9.3.6. Rank transformed data was assess using both the Welch-t test and the non-parametric Mann-Whitney test implemented using options of the `wilcox.test` function.

9.5 Application of robust statistical testing

The following information supports the work undertaken in Chapter Five investigating the effect on the power to detect *"truth"* in detection of the spiked-in transcripts within the Affymetrix Latin Square dataset using statistical tests that are more robust against the effect of outliers in the dataset.

9.5.1 Refinement of the analysis framework

The analysis framework introduced in Section 9.3.5 to was adapted to incorporate differing statistical tests as part of the overall procedure to assess the area under the FDR curves following the application of an expression metric in combination with a robust statistical test (Figure 9.6). The same array of sampled expression metrics was employed to enable data to be directly comparable to that from Section 9.3.

Figure 9.6

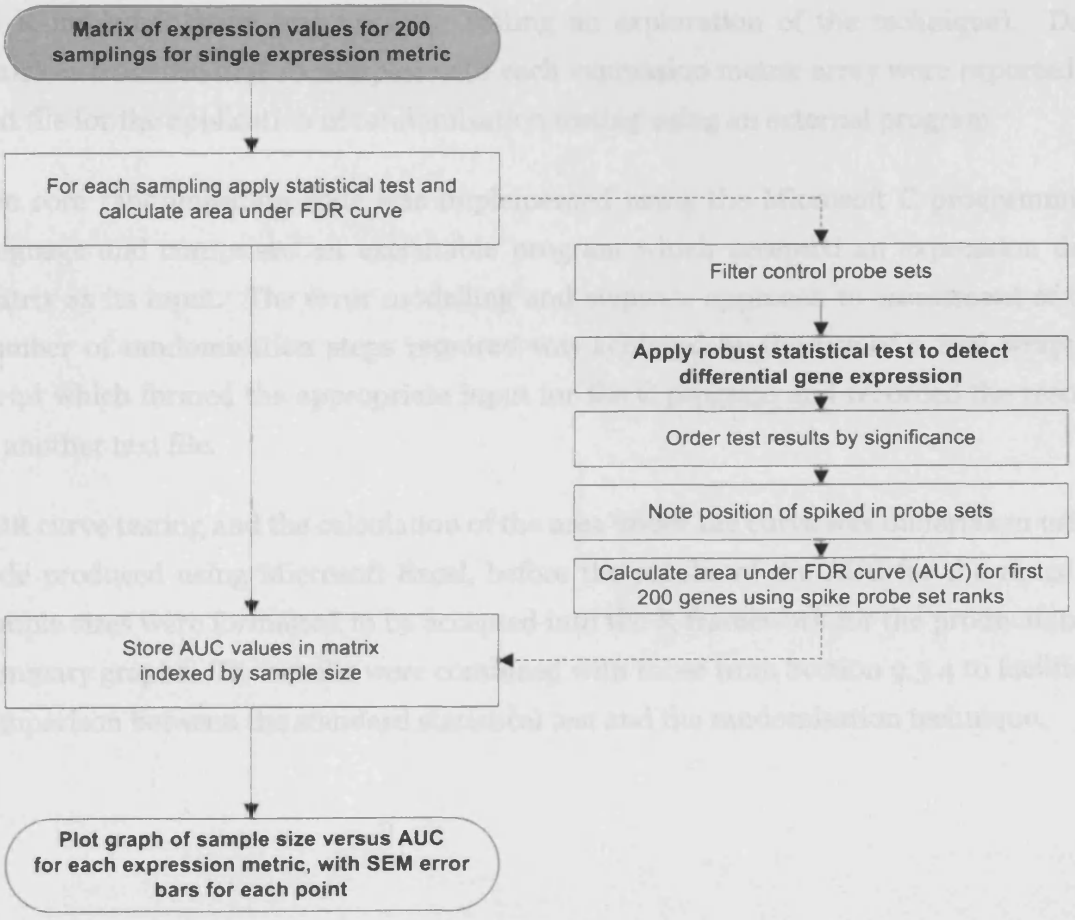


Figure 9.6 – Overview of FDR analysis for each expression metric incorporating more robust statistical tests for differential gene expression.

9.5.2 Implementation of robust statistical tests

Four alternatives to the t-test incorporating a combination of mean/median and standard deviation/median absolute deviation were implemented by alterations to the Welch t-test function identified in Section 9.2.4. The R functions of `mean` and `sd` were replaced by those of `median` and `mad` at appropriate points of the t-test formulae.

The trimmed t-test, Winsorised t-test and Yuen t-test were all implemented using variations of the Welch t-test function with alterations made to the data inputted to the function before application of the standard Welch t formulae. Modifications were made accordingly to the degrees of freedom before the calculation of a p-value using the `pf` function.

9.5.3 Randomisation testing

Because of the issues of computational intensity and speed of execution, it was not possible to fully implement randomisation testing within the R environment (although an R implementation was used for testing an exploration of the technique). Data matrices from the first 10 samples with each expression metric array were exported to text file for the application of randomisation testing using an external program.

The core randomisation code was implemented using the Microsoft C programming language and comprised an executable program which accepted an expression data matrix as its input. The error modelling and stepwise approach to assessment of the number of randomisation steps required was achieved by the use of a Perl wrapper script which formed the appropriate input for the C program and recorded the results to another text file.

FDR curve testing and the calculation of the area under the curve was undertaken using code produced using Microsoft Excel, before the results of the AUC for the range of sample sizes were formatted to be accepted into the R framework for the production of summary graphs. The results were combined with those from Section 9.3.4 to facilitate comparison between the standard statistical test and the randomisation technique.

9.6 Application of a Bayesian Framework

The following information supports the work undertaken in Chapter Six investigating the effect on the power to detect "truth" in detection of the spiked-in transcripts within the Affymetrix Latin Square dataset using the Baldi and Long (Baldi and Long, 2001) framework for Bayesian testing.

9.6.1 Refinement of the analysis framework

The analysis framework introduced in Section 9.3.5 to was adapted to incorporate the Bayesian framework for detection of differential gene expression as part of the overall procedure to assess the area under the FDR curves following the application of an expression metric in combination with a statistical test (Figure 9.7. The same array of sampled expression metrics was employed to enable data to be directly comparable to that from Section 9.3.

Figure 9.7

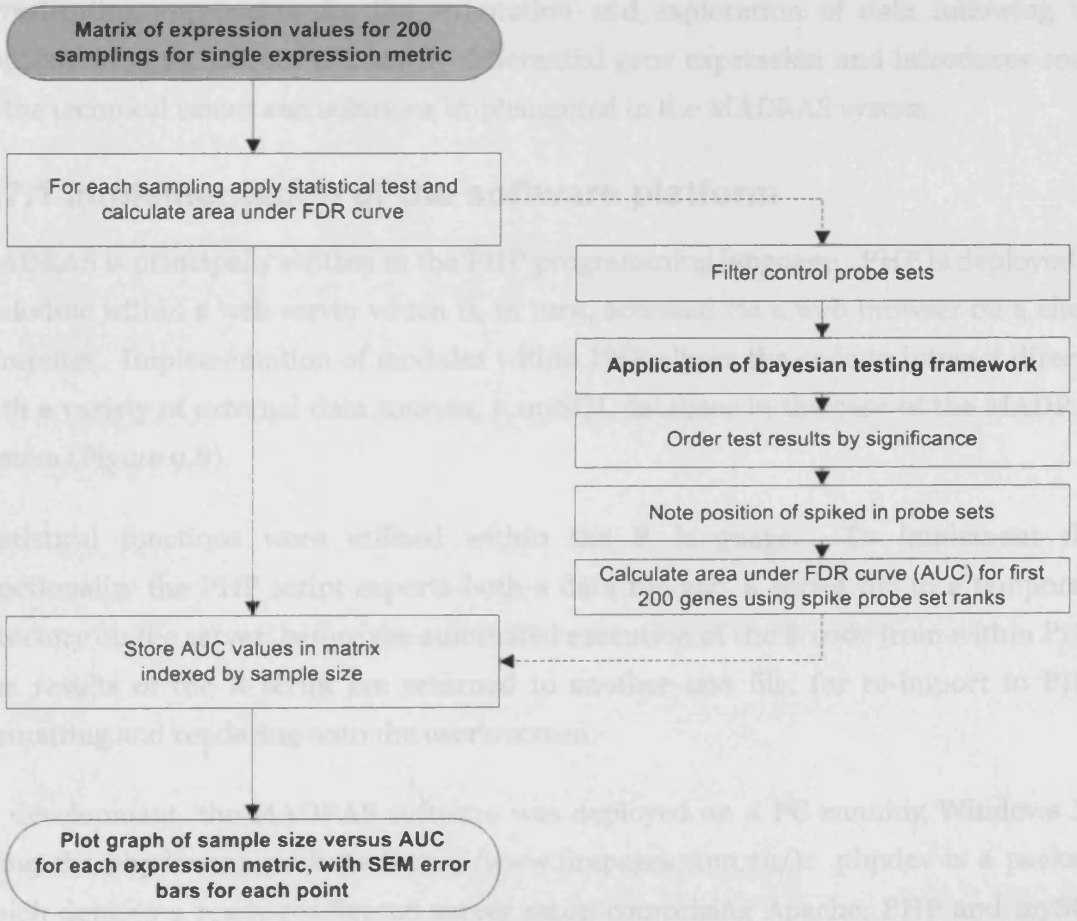


Figure 9.7 – Overview of FDR analysis for each expression metric incorporating more variations of the Baldi and Long Bayesian framework for differential gene expression.

9.6.2 Implementation of a Bayesian framework

The base code for the execution of Bayesian testing in the R environment was downloaded from the author's website (http://visitor.ics.uci.edu/genex/cybert/). Review of the code identified many sections that could be parsed from the script and a refined and simplified version was produced which would enable simple integration within the FDR framework.

Additional options were added to the function to allow for variations in the window size and blending options as well as alterations in the algorithm utilised from the FDR script (see Section 9.5.2). The results of the Bayesian analysis were returned into the script for summarisation and analysis of detection power.

9.7 Approaches to annotation and exploration of Affymetrix microarray data

The following technical information underpins the work undertaken in Chapter Seven investigating approaches for the annotation and exploration of data following the application of techniques to identify differential gene expression and introduces some of the technical issues and solutions implemented in the MADRAS system.

9.7.1 Implementation of the software platform

MADRAS is principally written in the PHP programming language. PHP is deployed as a module within a web server which is, in turn, accessed via a web browser on a client computer. Implementation of modules within PHP allows the code to interact directly with a variety of external data sources, a mySQL database in the case of the MADRAS system (Figure 9.8).

Statistical functions were utilised within the R language. To implement this functionality the PHP script exports both a data file and R script file to a temporary directory on the server, before the automated execution of the R code from within PHP. The results of the R script are returned to another text file, for re-import to PHP, formatting and rendering onto the user's screen.

In development, the MADRAS software was deployed on a PC running Windows XP using the phpdev423 package (http://www.firepages.com.au/). phpdev is a package which deploys a ready-configured server setup comprising Apache, PHP and mySQL and allows for simplified system installation. The software platform for MADRAS can be implemented on a variety of operating systems using differing platform releases of the open source tools.

Figure 9.8

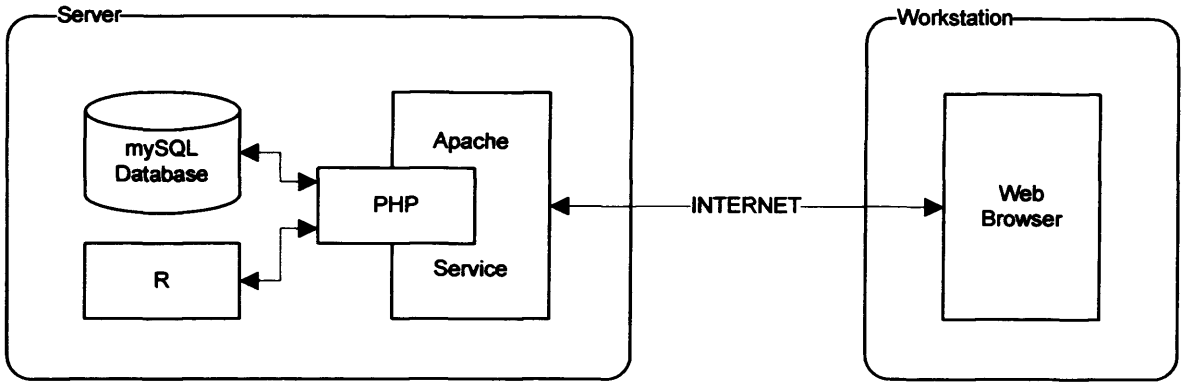


Figure 9.8 – Overview of software implementation for the MADRAS system

9.7.2 MADRAS database design and implementation

The MADRAS database is principally designed around two modules, an annotation database, and user data and exploratory information database. Addition information is required to separate the users information, and allow sharing of data between research groups. This information is stored in additional database tables for user information.

The user data tables store the uploaded GeneChip array information, along with information to format individual chips into groups termed *"experiments"*. Key to the exploration of data in manageable segments is the provision of *"probelists"* which store lists of linked probe set identifiers, which can be combined along with annotation information and actual user data from an *"experiment"* to form the rapid exploration environment that MADRAS forms.

The annotation database is formed from a variety of text files downloaded from the following sources:

LocusLink from NCBI (ftp://ftp.ncbi.nlm.nih.gov/refseq/LocusLink/ARCHIVE/)

Array specific annotation from Affymetrix (http://www.netaffx.com)

Homologene from NCBI (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/)

Taxonomy from NCBI (ftp://ftp.ncbi.nih.gov/pub/taxonomy/)

OMIM from NCBI (ftp://ftp.ncbi.nih.gov/repository/OMIM/)

GenMAPP pathways from GENMAPP
(http://www.genmapp.org/download-MAPPs.asp)

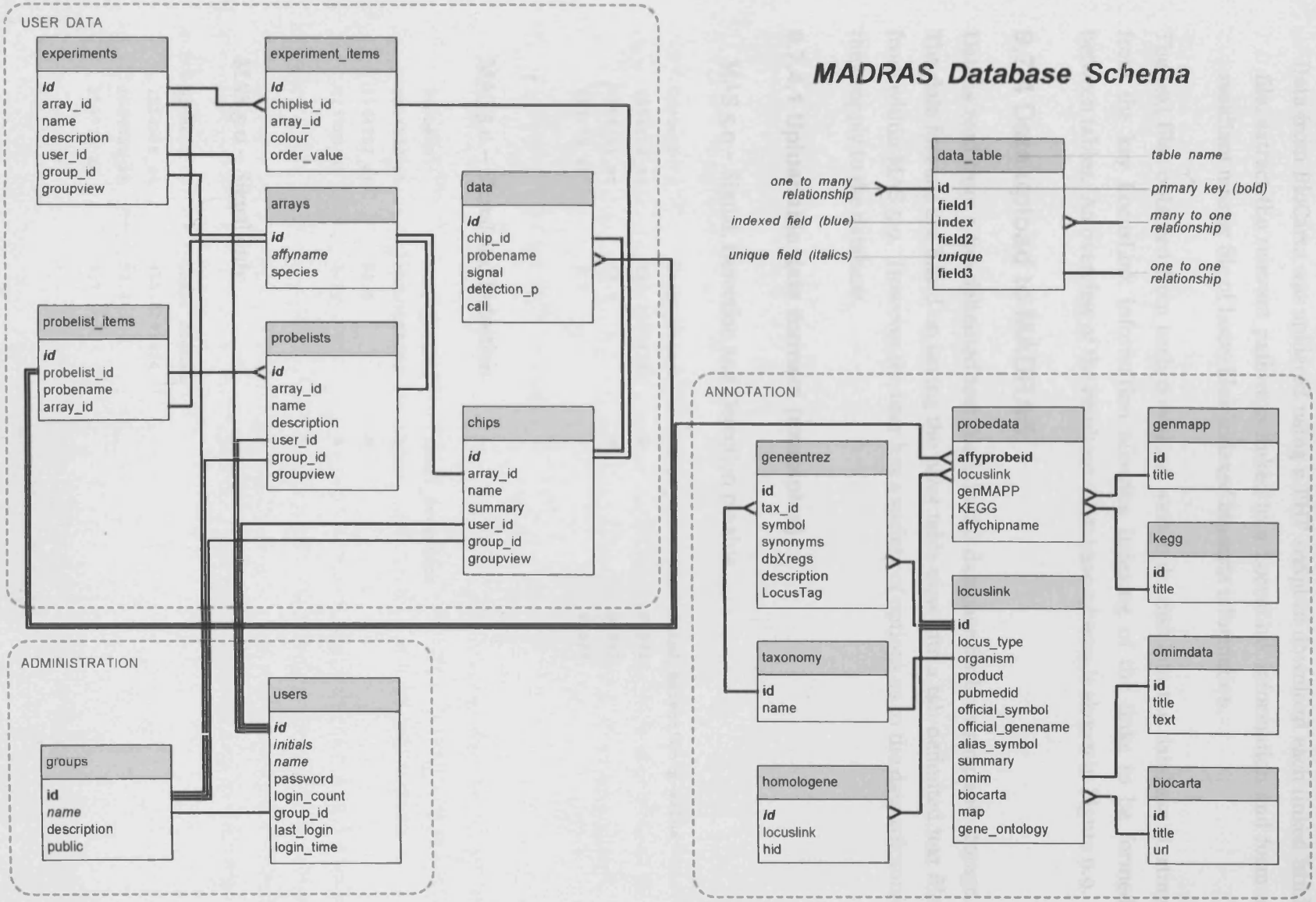

MADRAS Database Schema
USER DATA
experiments
id
array_id
name
description
user_id
group_id
groupview
experiment_items
id
chiplist_id
array_id
colour
order_value
arrays
id
affyname
species
data
id
chip_id
probename
signal
detection_p
call
probelist_items
id
probelist_id
probename
array_id
probelists
id
array_id
name
description
user_id
group_id
groupview
chips
id
array_id
name
summary
user_id
group_id
groupview
ADMINISTRATION
users
id
initials
name
password
login_count
group_id
last_login
login_time
groups
id
name
description
public
data_table
id
field1
index
field2
unique
field3
table name
primary key (bold)
many to one relationship
one to one relationship
one to many relationship
indexed field (blue)
unique field (italics)
ANNOTATION
probedata
affyprobeid
locuslink
genMAPP
KEGG
affychipname
genmapp
id
title
kegg
id
title
geneentrez
id
tax_id
symbol
synonyms
dbXregs
description
LocusTag
locuslink
id
locus_type
organism
product
pubmedid
official_symbol
official_genename
alias_symbol
summary
omim
biocarta
map
gene_ontology
omimdata
id
title
text
taxonomy
id
name
homologene
id
locuslink
hid
biocarta
id
title
url

Figure 9.9

KEGG data from GenomeNet (ftp://ftp.genome.ad.jp/pub/kegg/)

Data from BioCarta was spidered using a PHP script to download each linked html file, extract the relevant pathways linked to a LocusLink information and form a resultant master file of LocusLink indexed biocarta information.

The text files obtained from each download were uploaded into the database starting from the key LocusLink information allowing indexing of the links to be formed between tables. An overview of the resultant ddatabase schema is shown in Figure 9.9.

9.7.4 Data upload to MADRAS

Data is read from a tab delimited text file into the database for browsing and storage. The data formats are based on saving the pivot table view into a tab-delimited text file from within MAS 5.0. However the user has a variety of options as to the data columns they supply to the database.

9.7.4.1 Uploadable data formats (examples)

MAS 5.0 – Signal, Detection and Detection p-value

ProbeID	Chip1_Signal	Chip1_Detection	Chip1_Detection p-value
1554526_at	453.7297434	P	0.942
1564251_at	53.4	M	0.042
213779_at	3.7	A	0.028

MAS 5.0 – Signal and Detection

ProbeID	Chip1_Signal	Chip1_Detection
1554526_at	453.7297434	P
1564251_at	53.4	M
213779_at	3.7	A

MAS 5.0 – Signal only

ProbeID	Chip1_Signal
1554526_at	453.7297434
1564251_at	53.4
213779_at	3.7

9.7.4.2 Upload of other data formats

At present the system only accepts MAS 5.0 data formats (other data may be uploaded if forced into this format). To force other data into the required format the key things to note are the fact that the ProbeID column is optional and column labels must end with MAS 5.0 suffixes (_Signal, _Detection, _Detection p-value) and must be present in the combinations indicated above (Section 9.7.4.1).

9.7.3 Technical implementation of analysis features

9.7.3.1 Bar chart and heatmap graphics

The bar chart and heatmap graphics integrated into the explore view (Section 7.4.1) are formed following the extraction of relevant data from the database and then converted into graphics using the jpgraph PHP library (*http://www.aditus.nu/jpgraph/*). jpGraph is a fully OO (Object-Oriented) graph creating class library which can produce dynamic graphics for inclusion into the web-page output, eliminating the requirement generate fixed graphic files on the server.

A variety of options are passed to the jpgraph code describing the graphing options required and include options for a linear or logarithmic scale, display or non-display of data values for graphs, the size of the graph and information for the colouration of the bars according to detection call and background according to user defined parameters.

9.7.3.2 Experimental data heatmap

The experimental heatmap function plots an image of the clustered data for all probe sets in the current probelist. The data is subject to average linkage hierarchical clustering after the application of a log transformation and median centring step.

The clustering is undertaken using customised R code which accepts a text file in a similar to the *"MAS 5.0 signal only"* data upload format (Section 9.7.4.1). The script outputs two files, a graphical file for inclusion in the returned web-page (png file) and an Adobe Acrobat file available for download. The clustering is applied to the probe sets only and not the experiments.

9.7.3.3 Dendrogram

Experimental clustering for data for a selected probelist is undertaken using a modified version of the heatmap function (Section 9.7.3.2). However, clustering is applied to the experiment columns only, and data returned as a dendrogram graphic.

9.7.3.4 Over-representation analysis

The key function used in the execution of over representation analysis is the `p.hyper` function within the R environment. However, the computationally intensive process of an over-representation analysis is the generation of the four figures the hypergeometric p-value function requires. The analysis can review the terms contained within the Gene Ontology, BioCarta, KEGG and GenMAPP fields as well as the text based gene name and summary fields. For the text based fields, these are split into individual and word-pairs for individual over-representation analysis for each word/word pair.

First the probe lists is reduced to a list of unique probes by reference to the LocsuLink database and array design information tables. Next, data for the field under investigation (e.g. BioCarta pathways) is extracted and counts made within the list of unique loci for each different term encountered within the field of interest (e.g. list and counts of all matches to BioCarta pathways).

For each resultant term, a search is then undertaken against the full corpus of data for that field in relevant LocusLink database entries for the GeneChip under examination, resulting in the four pieces of information required from application of the hypergeometric test; number of hits against the unique loci in the probelists under examination, the number of unique loci in the probelist, the number unique LocusLink loci represented on the current array, and the number of these loci which contain the term under analysis.

As an example, a list of DNA repair genes containing 276 probe sets on a HG-U133A array can be reduced to 134 loci in the LocusLink database. Examination of the Gene Ontology field indicates that 13 of these 134 loci match the annotation "Molecular Function : Damaged DNA Binding". Review of the relevant LocusLink data for the array shows that 22 out of 12726 loci match the annotation. These four figures are those forwarded to the `p.hyper` function in R.

This process is repeated against all identified terms for the field of interest (e.g. all Gene Ontology terms represented in the list), with each one returning a separate p-value from the analysis. The user is provided with the option of applying a Bonferonni correction for multiple testing on each of these analysis results. The results are returned as a table to the user with columns indicating the search term, probe list loci counts, global loci counts and the resultant p-value from hypergeometric probability analysis.

Appendix One
Overview of R functions created for distribution and FDR investigations

File: boxcox.r

Function: box.cox

Returns a dataset following the application of a Box-Cox power transform according to the supplied lambda value.

Usage

```
box.cox(x, lam)
```

Details

`x` vector of data values

`lam` lambda value

Function: box.cox.norm

Returns a lambda value reflecting the Box-Cox transformation correlating most closely to normality. The Box-Cox transformation is applied between `lmin` and `lmax` with a step value of 0.1.

Usage

```
box.cox.norm(x, norm="NA", lmin=-3, lmax=3)
```

Details

`x` vector of data values

`norm` array of normal distribution values (length must equal length of `x`)
if none supplied, the array is filled using the `rnorm` function

`lmin` minimum value of lambda value

`lmax` maximum value of lambda value

File: basic_tests.r

Function: foldchange

Returns the fold-change between two groups of data, calculated as the absolute value of the log (base two) transformed ratio of the means of the two groups of data.

Usage

```
foldchange(x, ng1="NA", ng2="NA")
```

Details

`x` vector of data values - group one data should proceed group two data

`ng1` number of values in group one

`ng2` number of values in group two

default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.

Function: ttest.p

Returns the p-value resulting from an unpaired t-test assuming homogeneity of variance between the two groups of data.

Usage

```
ttest.p(x, ng1="NA", ng2="NA")
```

Details

`x` vector of data values - group one data should proceed group two data

`ng1` number of values in group one

`ng2` number of values in group two

default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.

Function: welch.p

Returns the p-value resulting from the unpaired Welch t-test assuming heterogeneity of variance between the two groups of data.

Usage

```
welch.p(x, ng1="NA", ng2="NA")
```

Details

`x` vector of data values - group one data should proceed group two data

`ng1` number of values in group one

`ng2` number of values in group two

default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.

Function: mwu.p

Returns the p-value resulting from the Mann-Whitney test (non-parametric equivalent to the unpaired t-test)

Usage

```
mwu.p(x, ng1="NA", ng2="NA")
```

Details

`x` vector of data values - group one data should proceed group two data

`ng1` number of values in group one

`ng2` number of values in group two

default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.

File: robust_tests.r

Function: med_mad.p

Returns the p-value resulting from a robust implementation of the Welch t-test, with substitution of the median and median absolute deviation of each group of data for the mean and standard deviation respectively.

Usage

```
med_mad.p(x, ng1="NA", ng2="NA")
```

Details

`x` vector of data values - group one data should proceed group two data

`ng1` number of values in group one

`ng2` number of values in group two

default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.

Function: med_sd.p

Returns the p-value resulting from a robust implementation of the Welch t-test, with substitution of the median of each group of data for the mean.

Usage

```
med_sd.p(x, ng1="NA", ng2="NA")
```

Details

`x` vector of data values - group one data should proceed group two data

`ng1` number of values in group one

`ng2` number of values in group two

default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.

Function: mean_mad.p

Returns the p-value resulting from a robust implementation of the Welch t-test, with substitution of the median absolute deviation of each group of data for the standard deviation.

Usage

```
mean_mad.p(x, ng1="NA", ng2="NA")
```

Details

`x` vector of data values - group one data should proceed group two data

`ng1` number of values in group one

`ng2` number of values in group two

default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.

Function: trimmed.p

Returns the p-value resulting from a trimmed version of the Welch t-test, with the tails of the data trimmed.

Usage

```
trimmed.p(x, ng1="NA", ng2="NA", deg=1)
```

Details

`x` vector of data values - group one data should proceed group two data

`ng1` number of values in group one

`ng2` number of values in group two

default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.

`deg` number of data elements to remove from each end of the ordered dataset for each group. Note: if more elements are to be removed than exist in the dataset the function returns a p-value of 1.

Function: windsor.p

Returns the p-value resulting from a Windsorised version of the Welch t-test, with the tails of the data substituted according to Windsor's principle.

Usage

```
windsor.p(x, ng1="NA", ng2="NA", deg=1)
```

Details

`x` vector of data values - group one data should proceed group two data

`ng1` number of values in group one

`ng2` number of values in group two

default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.

`deg` number of data elements to Windsorise at each end of the ordered dataset for each group. Note: if more elements are to be removed than exist in the dataset the function returns a p-value of 1.

Function: yuen.p

Returns the p-value resulting from a Yuen's version of the Welch t-test, implementing a trimmed mean and Windsorised variance for each group of data.

Usage

```
windsor.p(x, ng1="NA", ng2="NA", deg=1)
```

Details

`x` vector of data values - group one data should proceed group two data

`ng1` number of values in group one

`ng2` number of values in group two

default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.

`deg` number of data elements to trimmer of Windsorise at each end of the ordered dataset for each group. Note: if more elements are to be removed than exist in the dataset the function returns a p-value of 1.

File: bayesian_tests.r

Function: bayesian.t

Returns the p-value resulting from Bayesian implementation of the Welch t-test utilising a localised estimate of the dataset's variance to improve detection. The function returns a vector contain the p-values for all probe sets in the dataset with Affymetrix identifiers as the element names.

Usage

```
bayesian.t(h, winsize, running = "mean",
variance="unequal", conf=3, mad=FALSE, fixed=FALSE)
```

Details

`h`	matrix of data values (rows containing probe set identifiers, and columns GeneChip array names)
`winsize`	window size in which to calculate the local variance
`running`	options of `"mean"` or `"median"` reflect the calculation of the mean or median of values in the current window of variances
`variance`	window size in which to calculate the local variance
`conf`	blending constant, if `fixed = FALSE` then blending is implemented using `conf` times the sample size in each group
`mad`	implementation of median/MAD version of t-test
`fixed`	fixed blending constant independent of sample size

File: latin2testing.r

Function: makeRTindex

Generates an index of file names representing samplings of a subset of the Affymetrix HG-U95A Latin square dataset over a range of sample sizes between three and twelve. The returned file `RTindex.RData` contains an array with 10 times `runs` elements, each containing the CEL file name for that sampling.

Usage

```
makeRTindex(runs=5)
```

Details

`runs`	number of samplings to take over the full range of sample sizes

Function: makeRTdata

Applies expression metrics to the samplings extracted from the `RTindex.RData` file created using `makeRTindex`. The function has a number of dependences according to the expression metric applied and outputs an R data object containing an array of expression matrices according to each of the samplings.

Usage

```
makeRTdata(expsum="mas5", celfile.path = getwd())
```

Additional Files Required

```
latin24_mas4_exaffy.csv
latin24_mas5_exaffy.csv
latin24_MBEI_PMMM_exdchip13.csv
latin24_MBEI_PM_exdchip13.csv
1521m99hpp_av06.CEL
1521n99hpp_av06.CEL
1521o99hpp_av06.CEL
1521p99hpp_av06.CEL
1521q99hpp_av06.CEL
1521r99hpp_av06.CEL
1521s99hpp_av06.CEL
1521t99hpp_av06.CEL
1532m99hpp_av04.CEL
1532n99hpp_av04.CEL
1532o99hpp_av04.CEL
1532p99hpp_av04.CEL
1532q99hpp_av04.CEL
1532r99hpp_av04.CEL
1532s99hpp_av04.CEL
1532t99hpp_av04r.CEL
2353m99hpp_av08.CEL
2353n99hpp_av08.CEL
2353o99hpp_av08.CEL
2353p99hpp_av08.CEL
2353q99hpp_av08.CEL
2353r99hpp_av08.CEL
2353s99hpp_av08.CEL
2353t99hpp_av08.CEL
```

Details

`expsum` choice of expression metric to apply, options of `rmaall`, `rmasep`, `gcrmaall`, `gcrmasep`, `dchippm`, `dchippmmm`, `mas5` and `mas4`

`celfile.path` Path information to access additional required files

Function: fdr.test

Undertakes comparative analysis of the power of differing combinations of statistical testing to identify the spiked-in transcripts within the Affymetrix Latin square dataset. Combinations of tests can be applied to the series of data samplings derived using `makeRTindex` and `makeRTdata`. Following analysis the positions of the spikes is determined and used to output a matrix of area under the FDR curves according to sample size.

Usage

```
fdr.test(expsum="mas5", test="welch", do.qq=FALSE,
do.rank=FALSE, log.it=FALSE, power.it=FALSE,
log10.it=FALSE, do.vsn=FALSE, graph.out=TRUE)
```

Additional Files Required

```
dchipPMMMRT.RData
dchipPMRT.RData
gcrmaAllRT.RData
gcrmaSepRT.RData
mas4RT.RData
mas5RT.RData
rmaAllRT.RData
rmaSepRT.RData
exclude.txt (list of Affymetrix identifiers to be excluded from analysis)
```

Details

`expsum` choice of expression metric to apply, options of `rmaall`, `rmasep`, `gcrmaall`, `gcrmasep`, `dchippm`, `dchippmmm`, `mas4` and `mas5`

`test` statistical test to apply to each dataset:

`foldchange` Absolute value of the log ratios of the mean expression level

`ttest` Unpaired t-test

`welch` Welch t-test

`mwu` Non-parametric Mann-Whitney test

`medsd` Robust Welch t-test using median of the dataset in formulae.

`meanmad` Robust Welch t-test using mean and median absolute deviation (MAD) in formulae.

`medmad` Robust Welch t-test using median and median absolute deviation (MAD) in formulae.

`trimmed1` Trimmed t-test (1 degree trimming)

`trimmed2` Trimmed t-test (2 degree trimming)

`winsor1` Winsorised t-test (1 degree trimming)

`winsor2` Winsorised t-test (2 degree trimming)

`yuen1` Yuen's t-test (1 degree trimming)

`yuen2` Yuen's t-test (2 degree trimming)

`bayes_welch_mean` Apply Bayesian framework utilising the Welch t-test, running mean of local variance with a window size of 500

`bayes_welch_median` Apply Bayesian framework utilising the Welch t-test, running median of local variance with a window size of 500

`bayes_ttest_mean` Apply Bayesian framework utilising the unpaired t-test, running mean of local variance with a window size of 500

`bayes_ttest_median` Apply Bayesian framework utilising the unpaired t-test, running median of local variance with a window size of 500

`bayes_mad_mean` Apply Bayesian framework utilising the Welch t-test, running mean of local median absolute deviation, with a window size of 500

`bayes_mad_median` Apply Bayesian framework utilising the Welch t-test, running median of local median absolute deviation, with a window size of 500

`do.qq` Apply Quantile-Quantile normalisation to datasets

`do.rank` Transform data to ranks within each array

`log.it` Apply $\log_2$ transform data to each dataset

`power.it` Apply 2^ transform data to each dataset

`log10.it` Apply $\log_{10}$ transform data to each dataset

`do.vsn` Apply VSN normalisation to datasets

`graph.out` Display FDR graph upon completion of analysis (uses `fdr.plot`)

Function: fdr.plot

Draws a line graph of the output from `fdr.test`, plotting the sample size on the x-axis, with area under the FDR curve. Where multiple values exist in the input matrix at each sample size, the mean of values is plotted, together with error bars representing the standard error of the mean.

Usage

```
fdr.plot(x, addit=FALSE, col="black", test="NA",
metric="NA", transform="NA", add.text=TRUE, title="",
lty=1)
```

Details

`addit` add line to existing `fdr.plot`

`col` line colour

`test`	text to add to plot (line 1)
`metric`	text to add to plot (line 2)
`transform`	text to add to plot (line 3)
`add.text`	display text fields
`title`	plot title
`lty`	line type (standard R `par` values)

File: randomised.r

Function: randomised.p

Returns a p-value resulting from the application of a randomised t-test to a dataset following a number of random samplings of the data.

Usage

```
randomised.p(x, ng1="NA", ng2="NA", its=10000)
```

Details

`x`	vector of data values - group one data should proceed group two data
`ng1`	number of values in group one
`ng2`	number of values in group two
	default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.
`its`	number of random samplings to take to derive a p-value

Function: randomised.p.error

Returns a p-value resulting from the application of a randomised t-test to a dataset following a number of random samplings of the data derived using an error model based on binomial theory.

Usage

```
randomised.p.error(x, ng1="NA", ng2="NA", imax=10000,
k=0.05)
```

Details

`x` vector of data values - group one data should proceed group two data

`ng1` number of values in group one

`ng2` number of values in group two

default values for `ng1` and `ng2` are `"NA"`. If no parameters are passed, values are calculated by dividing the dataset into two equal groups.

`imax` number of random samplings to take to derive a p-value

`k` confidence required in the resultant p-value

Appendix Two
Overview of script and code to implement the randomisation t-test

File: wrapper.pl

Implements the randomised t-test utilising a bionomical error model using a `perl` script and the `randc2.exe` executable. For efficiency data is subjected to randomisation testing for a set number of cycles, the data is then reviewed according to the model, before data being deemed to have reached an acceptable p-value is excluded. The remaining data is then subject to repeated stepped analysis until the maximum number of permutations is achieved.

Usage

```
wrapper.pl input output rows cols1 cols2 step max conf
```

Details

`input`	input filename (tab delimited text file, with ID's as column one and no header information
`output`	output filename
`rows`	number of data rows
`cols1`	number of columns for group 1 data
`cols2`	number of columns for group 2 data
`step`	permutation step between error modelling
`max`	maximum number of permutations to perform
`conf`	confidence required (0.1 = 10%)

File: randc2.exe

Implements the randomised t-test in a C code executable.

Usage

```
randc2 input output rows cols1 cols2 perms
```

Details

`input`	input filename (tab delimited text file, with ID's as column one and no header information
`output`	output filename
`rows`	number of data rows
`cols1`	number of columns for group 1 data
`cols2`	number of columns for group 2 data
`perms`	number of randomisation steps to apply

Bibliography

Affymetrix (2000) News Release: NASDAQ Stock Exchange Report. Santa Clara, CA.

Affymetrix (2001) Affymetrix Microarray Suite Users Guide - Version 5.0. Affymetrix, Santa Clara, CA.

Affymetrix (2001a) Array Design for the GeneChip® Human Genome U133 Set. Affymetrix Inc,, Santa Clara, CA.

Affymetrix (2001b) New Statistical Algorithms for Monitoring Gene Expression on GeneChip Probe Arrays. Affmetrix, Inc., Santa Clara, CA.

Affymetrix (2002a) Statistical Algorithms Reference Guide. Affmetrix, Inc, Santa Clara, CA.

Affymetrix (2002b) Statistical Algorithms Description Document. Affmetrix, Inc, Santa Clara, CA.

Affymetrix (2004) GeneChip Expression Analysis Technical Manual. Affymetrix Inc., Santa Clara, CA.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc Natl Acad Sci U S A, 96, 6745-6750.

Al-Shahrour, F., Diaz-Uriarte, R. and Dopazo, J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes, Bioinformatics, 20, 578-580.

Arfin, S.M., Long, A.D., Ito, E.T., Tolleri, L., Riehle, M.M., Paegle, E.S. and Hatfield, G.W. (2000) Global gene expression profiling in Escherichia coli K12. The effects of integration host factor, J Biol Chem, 275, 29672-29684.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat Genet, 25, 25-29.

Bakay, M., Chen, Y.W., Borup, R., Zhao, P., Nagaraju, K. and Hoffman, E.P. (2002) Sources of variability and effect of experimental approach on expression profiling data interpretation, BMC Bioinformatics, 3, 4.

Baldi, P. and Brunak, S. (2001) Bioinformatics: The Machine Learning

Approach. MIT Press, Cambridge, MA.

Baldi, P. and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes, Bioinformatics, 17, 509-519.

Barnett, V. and Lewis, T. (1994) Outliers in Statistical Data. Wiley,, New York, NY.

Barrera, L., Benner, C., Tao, Y.C., Winzeler, E. and Zhou, Y. (2004) Leveraging two-way probe-level block design for identifying differential gene expression with high-density oligonucleotide arrays, BMC Bioinformatics, 5, 42.

Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W. and Edgar, R. (2005) NCBI GEO: mining millions of expression profiles--database and tools, Nucleic Acids Res, 33 Database Issue, D562-566.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004) The Pfam protein families database, Nucleic Acids Res, 32, D138-141.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2004) GenBank: update, Nucleic Acids Res, 32, D23-26.

Bethea, R.M. and Rhinehart, R.R. (1991) Applied Engineering Statistics. Marcel Dekker, Inc, New York, NY.

BioCarta (2005) Biocarta Pathways.

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

Bowcock, A.M., Shannon, W., Du, F., Duncan, J., Cao, K., Aftergut, K., Catier, J., Fernandez-Vina, M.A. and Menter, A. (2001) Insights into psoriasis and other inflammatory diseases from large-scale gene expression studies, Hum Mol Genet, 10, 1793-1805.

Box, G.E.P. and Cox, D.R. (1964) An analysis of transformation., J Royal Statistial Society, 26, 211-243.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data, Nat Genet, 29, 365-371.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., Oezcimen, A., Rocca-Serra, P. and Sansone, S.A. (2003) ArrayExpress--a public repository for microarray gene expression data at the EBI, Nucleic Acids Res, 31, 68-71.

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

Cao, S.X., Dhahbi, J.M., Mote, P.L. and Spindler, S.R. (2001) Genomic profiling of short- and long-term caloric restriction effects in the liver of aging mice, Proc Natl Acad Sci U S A, 98, 10630-10635.

Chaussabel, D. and Sher, A. (2002) Mining microarray expression data by literature profiling, Genome Biol, 3, RESEARCH0055.

Chen, Y.A., McKillen, D.J., Wu, S., Jenny, M.J., Chapman, R., Gross, P.S., Warr, G.W. and Almeida, J.S. (2004) Optimal cDNA microarray design using expressed sequence tags for organisms with limited genomic information, BMC Bioinformatics, 5, 191.

Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M. and Halfon, M.S. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset, Genome Biol, 6, R16.

Chudin, E., Walker, R., Kosaka, A., Wu, S.X., Rabert, D., Chang, T.K. and Kreder, D.E. (2002) Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays, Genome Biol, 3, RESEARCH0005.

Churchill, G.A. (2001) Statistical Design and the Analysis of Gene Expression Microarray Experiments.

Churchill, G.A. (2002) Fundamentals of experimental design for cDNA microarrays, Nature Genetics, 32, 490-495.

Clarke, P.A., te Poele, R., Wooster, R. and Workman, P. (2001) Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential, Biochem Pharmacol, 62, 1311-1336.

Cope, L.M., Irizarry, R.A., Jaffee, H.A., Wu, Z. and Speed, T.P. (2004) A benchmark for Affymetrix GeneChip expression measures, Bioinformatics, 20, 323-331.

D'Agostino, R.B. and Stephens, M.A. (1986) Goodness-of-fit Techniques. Dekker, New York.

Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C. and Conklin, B.R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways, Nat Genet, 31, 19-20.

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

Diaconis, P. and Efron, B. (1983) Computer-intensive methods in statistics, Scientific American, 5, 116-130.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel, A. (2004) Misc Functions of the Department of Statistics (e1071). CRAN, TU Wien.

Dixon, W.J. and Tukey, J.W. (1968) Approximate Behavior of the Distribution of Winsorized t (Trimming/Winsorization 2), Technometrics, 10, 83 -98.

Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C. and Conklin, B.R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data, Genome Biol, 4, R7.

Dow, G.S. (2003) Effect of sample size and P-value filtering techniques on the detection of transcriptional changes induced in rat neuroblastoma (NG108) cells by mefloquine, Malar J, 2, 4.

Dudoit, S. and Fridlyand, J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset, Genome Biol, 3, RESEARCH0036.

Dudoit, S., Hwa Yang, Y., Callow, M.J. and Speed, T.P. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report #578. Stanford University School of Medicine.

Dudoit, S., Yang, Y.H. and Bolstat, B. (2002) Using R for the analysis of DNA microarray data., R News, 2, 24-32.

Durbin, B. and Rocke, D.M. (2003) Estimation of transformation parameters for microarray data, Bioinformatics, 19, 1360-1367.

Durbin, B.P., Hardin, J.S., Hawkins, D.M. and Rocke, D.M. (2002) A variance-stabilizing transformation for gene-expression microarray data, Bioinformatics, 18 Suppl 1, S105-110.

Eddy, S.R. (2004) What is Bayesian statistics?, Nat Biotechnol, 22, 1177-1178.

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

Edginton, E.S. (1997) Randomization Tests. Basel, New York, NY.

Filliben, J.J. (1975) The Probability Plot Correlation Coefficient Test for Normality, Technometrics, 17, 111-117.

Finkelstein, D., Ewing, R., Gollub, J., Sterky, F., Cherry, J.M. and Somerville, S. (2002) Microarray data quality analysis: lessons from the AFGC project. Arabidopsis Functional Genomics Consortium, Plant Mol Biol, 48, 119-131.

Fisher, R.A. (1935) The Design of Experiments. Hafner, New York.

Galperin, M.Y. (2004) The Molecular Biology Database Collection: 2004 update, Nucleic Acids Res, 32, D3-22.

Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004) affy--analysis of Affymetrix GeneChip data at the probe level, Bioinformatics, 20, 307-315.

Gautier, L., Moller, M., Friis-Hansen, L. and Knudsen, S. (2004) Alternative mapping of probes to genes for Affymetrix chips, BMC Bioinformatics, 5, 111.

Geller, S.C., Gregg, J.P., Hagerman, P. and Rocke, D.M. (2003) Transformation and normalization of oligonucleotide microarray data, Bioinformatics, 19, 1817-1823.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G.,

Tierney, L., Yang, J.Y. and Zhang, J. (2004) Bioconductor: open software development for computational biology and bioinformatics, Genome Biol, 5, R80.

Gieseg, M.A., Cody, T., Man, M.Z., Madore, S.J., Rubin, M.A. and Kaldjian, E.P. (2002) Expression profiling of human renal carcinomas with functional taxonomic analysis., BMC Bioinformatics, 3.

Gollub, J., Ball, C.A., Binkley, G., Demeter, J., Finkelstein, D.B., Hebert, J.M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J.C., Schroeder, M., Brown, P.O., Botstein, D. and Sherlock, G. (2003) The Stanford Microarray Database: data access and quality assessment tools, Nucleic Acids Res, 31, 94-96.

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

Good, P.I. (2000) Resampling Methods: A Practical Guide to Data Analysis. Birkhauser, Boston, MA.

Gottardo, R., Raftery, A., Yeung, K. and Bumgarner, R. (2003) Robust Estimation of cDNA Microarray Intensities with

Replicates. Technical Report no. 438. Department of Statistics, University of Washington.

Grant, G.R., Manduchi, E. and Stoeckert, C.J. (2002) Using non-parametric methods in the context of multiple testing to determine differentially expressed genes. Kluwer Academic Publishers, Norwell MA.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic Acids Res, 33 Database Issue, D514-517.

Hampel, F. (1985) The breakdown points of the mean combined with some rejection rules, Technometrics, 27, 95-107.

Hancock, D.J., Morrisson, N., Rattray, M., Brass, A., Cornell, M. and Manchester, U.o. (2000) maxd - A Data Warehouse, Analysis, and Visualisation Environment for Expression Data.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C.,

Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. and White, R. (2004) The Gene Ontology (GO) database and informatics resource, Nucleic Acids Res, 32, D258-261.

Hartmann, O., Samans, S. and Schäfer, H. (2003) Microarray data normalization and transformation: Comparing MAS, VSN and RMA for Affymetrix GeneChips. Philipps-University Technical Report.

Hatfield, G.W., Hung, S.P. and Baldi, P. (2003) Differential analysis of DNA microarray gene expression data, Mol Microbiol, 47, 871-877.

Hedenfalk, I., Ringner, M., Ben-Dor, A., Yakhini, Z., Chen, Y., Chebil, G., Ach, R., Loman, N., Olsson, H., Meltzer, P., Borg, A. and Trent, J. (2003) Molecular classification of familial non-BRCA1/BRCA2 breast cancer, Proc Natl Acad Sci U S A, 100, 2532-2537.

Hill, A.A., Brown, E.L., Whitley, M.Z., Tucker-Kellogg, G., Hunter, C.P. and Slonim, D.K. (2001) Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls, Genome Biol, 2, RESEARCH0055.

Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C. and Lempicki, R.A. (2003) Identifying biological themes within lists of genes with EASE, Genome Biol, 4, R70.

Hoyle, D.C., Rattray, M., Jupp, R. and Brass, A. (2002) Making sense of microarray data distributions., Bioinformatics, 18, 576-584.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression, Bioinformatics, 18 Suppl 1, S96-104.

Hyndman, R.J. and Fan, Y. (1996) Sample Quantiles in Statistical Packages., American Statistician, 50, 361-365.

Iglewicz, B. and Hoaglin, B.C. (1993) How to Detect and Handle Outliers. In. American Society for Quality, Milwaukee, WI.

Ihaka, R. and Gentleman, R.R. (1996) A language for data analysis and graphics., J Comp. Graph. Statist.,, 5, 299-314.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome, Nature, 431, 931-945.

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

Ivanov, I., Schaab, C., Planitzer, S., Teichmann, U., Machl, A., Theml, S., Meier-Ewert, S., Seizinger, B. and Loferer, H. (2000) DNA microarray technology and antimicrobial drug discovery, Pharmacogenomics, 1, 169-178.

Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connell, M. and Lee, J.K. (2003) Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays, Bioinformatics, 19, 1945-1951.

Jenssen, T.K., Kuo, W.P., Stokke, T. and Hovig, E. (2002) Associations between gene expressions in breast cancer and patient survival, Hum Genet, 111, 411-420.

Jenssen, T.K., Laegreid, A., Komorowski, J. and Hovig, E. (2001) A literature network of human genes for high-throughput analysis of gene expression, Nat Genet, 28, 21-28.

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

Kadota, K., Tominaga, D., Akiyama, Y. and Takahashi, K. (2003) Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification, Chem-Bio Informatics Journal, 3, pp.30-45.

Kamb, A. and Ramaswami, M. (2001) A simple method for statistical analysis of intensity differences in microarray-derived gene expression data, BMC Biotechnol, 1, 8.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome, Nucleic Acids Res, 32, D277-280.

Katsuma, S., Shiojima, S., Hirasawa, A., Suzuki, Y., Takagaki, K., Murai, M., Kaminishi, Y., Hada, Y., Koba, M., Muso, E., Miyawaki, S., Ohgi, T., Yano, J. and Tsujimoto, G. (2001) Genomic analysis of a mouse model of immunoglobulin A nephropathy reveals an enhanced PDGF-EDG5 cascade, Pharmacogenomics J, 1, 211-217.

Kerr, M.K., Martin, M. and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data, J Comput Biol, 7, 819-837.

Kersten, S., Mandard, S., Escher, P., Gonzalez, F.J., Tafuri, S., Desvergne, B. and Wahli, W. (2001) The peroxisome proliferator-activated receptor alpha regulates amino acid metabolism, Faseb J, 15, 1971-1978.

Keselman, R.C. and Zumbo, B. (1997) Specialized tests for detecting treatment effects in the two-sample problem., Journal of Experimental Education, 65, 355-366.

Knudsen, S. (2002) A biologists guide to analysis of DNA microarray data. John Wiley and Sons,. Hoboken, NJ.

Kooperberg, C., Sipione, S., LeBlanc, M., Strand, A.D., Cattaneo, E. and Olson, J.M. (2002) Evaluating test statistics to select interesting genes in microarray experiments, Hum Mol Genet, 11, 2223-2232.

Korn, E.L., McShane, L.M., Troendle, J.F., Rosenwald, A. and Simon, R. (2002) Identifying pre-post chemotherapy differences in gene expression in breast tumours: a statistical method appropriate for this aim, Br J Cancer, 86, 1093-1096.

Lander, E.S. (1999) Array of hope, Nat Genet, 21, 3-4.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M.,

Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S. and Chen, Y.J. (2001) Initial sequencing and analysis of the human genome, Nature, 409, 860-921.

Landis, G.N., Abdueva, D., Skvortsov, D., Yang, J., Rabin, B.E., Carrick, J., Tavare, S. and Tower, J. (2004) Similar gene expression patterns characterize aging and oxidative stress in Drosophila melanogaster, Proc Natl Acad Sci U S A, 101, 7663-7668.

Lee, C. and Roy, M. (2004) Analysis of alternative splicing with microarrays: successes and challenges, Genome Biol, 5, 231.

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

Li, C. and Hung Wong, W. (2001a) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, Genome Biol, 2, RESEARCH0032.

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

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

Li, J., Lee, J.M. and Johnson, J.A. (2002b) Microarray analysis reveals an antioxidant responsive element-driven gene set involved in conferring protection from an oxidative stress-induced apoptosis in IMR-32 cells, J Biol Chem, 277, 388-394.

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

Li, S., Becich, M.J. and Gilbertson, J. (2004) Microarray data mining using gene ontology, Medinfo, 2004, 778-782.

Li, Y., McClintick, J., Zhong, L., Edenberg, H.J., Yoder, M.C. and Chan, R.J. (2005) Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf4, Blood, 105, 635-637.

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

Lipshutz, R.J., Morris, D., Chee, M., Hubbell, E., Kozal, M.J., Shah, N., Shen, N., Yang, R. and Fodor, S.P. (1995) Using oligonucleotide probe arrays to access genetic diversity, Biotechniques, 19, 442-447.

Liu, G., Loraine, A.E., Shigeta, R., Cline, M., Cheng, J., Valmeekam, V., Sun, S., Kulp, D. and Siani-Rose, M.A. (2003) NetAffx: Affymetrix probesets and annotations, Nucleic Acids Res, 31, 82-86.

Lix, L.M. and Keselman, H.J. (1998) To trim or not to trim: Tests of location equality under heteroscedasicity and nonnormality., Educational and Psychological Measurement, 58, 409-429.

Loguinov, A.V., Mian, I.S. and Vulpe, C.D. (2004) Exploratory differential gene expression analysis in microarray experiments with no or limited replication, Genome Biol, 5, R18.

Long, A.D., Mangalam, H.J., Chan, B.Y., Tolleri, L., Hatfield, G.W. and Baldi, P. (2001) Improved statistical inference from DNA microarray data using analysis of variance

and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12, J Biol Chem, 276, 19937-19944.

Lonnstedt, I. and Speed, T.P. (2002) Replicated microarray data, Statistical Sinica, 12, 31-46.

Luthi-Carter, R., Strand, A., Peters, N.L., Solano, S.M., Hollingsworth, Z.R., Menon, A.S., Frey, A.S., Spektor, B.S., Penney, E.B., Schilling, G., Ross, C.A., Borchelt, D.R., Tapscott, S.J., Young, A.B., Cha, J.H. and Olson, J.M. (2000) Decreased expression of striatal signaling genes in a mouse model of Huntington's disease, Hum Mol Genet, 9, 1259-1271.

Machet, L. (1998) [Genetic diseases on the Internet: OMIM], Ann Dermatol Venereol, 125, 645.

Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2005) Entrez Gene: gene-centered information at NCBI, Nucleic Acids Res, 33 Database Issue, D54-58.

Maglott, D.R., Katz, K.S., Sicotte, H. and Pruitt, K.D. (2000) NCBI's LocusLink and RefSeq, Nucleic Acids Res, 28, 126-128.

Man, M.Z., Wang, X. and Wang, Y. (2000) POWER_SAGE: comparing statistical tests for SAGE experiments, Bioinformatics, 16, 953-959.

Mariani, T.J., Budhraja, V., Mecham, B.H., Gu, C.C., Watson, M.A. and Sadovsky, Y. (2003) A variable fold change threshold determines significance for expression microarrays, Faseb J, 17, 321-323.

Martin, D.E., Demougin, P., Hall, M.N. and Bellis, M. (2004) Rank Difference Analysis of Microarrays (RDAM), a novel approach to statistical analysis of microarray expression profiling data, BMC Bioinformatics, 5, 148.

Martin, K.J., Graner, E., Li, Y., Price, L.M., Kritzman, B.M., Fournier, M.V., Rhei, E. and Pardee, A.B. (2001) High-sensitivity array analysis of gene expression for the early detection of disseminated breast tumor cells in peripheral blood, Proc Natl Acad Sci U S A, 98, 2646-2651.

MathWorks (2004) Bioinformatics Toolbox For Use with MATLAB. The MathWorks, Inc., Natick, MA.

McKay, B.C., Stubbert, L.J., Fowler, C.C., Smith, J.M., Cardamore, R.A. and Spronck, J.C. (2004) Regulation of ultraviolet light-induced gene expression by gene size, Proc Natl Acad Sci U S A, 101, 6582-6586.

Mehta, C. and Patel, N. (1999) Proc-StatXact 4 for SAS Users. Cytel Software Corperation.

Miller, R.A., Galecki, A. and Shmookler-Reis, R.J. (2001) Interpretation, design, and analysis of gene array expression experiments, J Gerontol A Biol Sci Med Sci, 56, B52-57.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L.C. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, Nat Genet, 34, 267-273.

Motulsky, H. (1999) Analysing data using GraphPad Prism. GraphPad Prism, Inc.

Mutter, G.L., Baak, J.P., Fitzgerald, J.T., Gray, R., Neuberg, D., Kust, G.A., Gentleman, R., Gullans, S.R., Wei, L.J. and Wilcox, M. (2001) Global expression changes of constitutive and hormonally regulated genes during endometrial neoplastic transformation, Gynecol Oncol, 83, 177-185.

Nadon, R. and Shoemaker, J. (2002) Statistical issues with microarrays: processing and analysis, Trends Genet, 18, 265-271.

Naef, F., Hacker, C.R., Patil, N. and Magnasco, M. (2002) Characterization of the expression ratio noise structure in high-density oligonucleotide arrays, Genome Biol, 3, PREPRINT0001.

Naef, F. and Magnasco, M.O. (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays, Phys Rev E Stat Nonlin Soft Matter Phys, 68, 011906.

Nees, M. and Woodworth, C.D. (2001) Microarrays: spotlight on gene function and pharmacogenomics, Curr Cancer Drug Targets, 1, 155-175.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data, J Comput Biol, 8, 37-52.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res, 27, 29-34.

Orphanoudakis, S., Kafetzopoulos, D. and Tsiknakis, M. (2005) Biomedical Informatics in Support of Individualized Medicine, ERCIM News.

Pan, W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments, Bioinformatics, 18, 546-554.

Pan, W. (2003) On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression, Bioinformatics, 19, 1333-1340.

Park, T., Yi, S.G., Lee, S., Lee, S.Y., Yoo, D.H., Ahn, J.I. and Lee, Y.S. (2003) Statistical tests for identifying differentially expressed genes in time-course microarray experiments, Bioinformatics, 19, 694-703.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G.G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S. and Brazma, A. (2005) ArrayExpress--a public repository for microarray gene expression data at the EBI, Nucleic Acids Res, 33 Database Issue, D553-555.

Parrish, R.S. and Spencer, H.J., 3rd (2004) Effect of normalization on significance testing for oligonucleotide microarrays, J Biopharm Stat, 14, 575-589.

Pavlidis, P. and Noble, W.S. (2001) Analysis of strain and regional variation in gene expression in mouse brain, Genome Biol, 2, RESEARCH0042.

Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P. and Fodor, S.P. (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis, Proc Natl Acad Sci U S A, 91, 5022-5026.

Perez, E.A., Pusztai, L. and Van de Vijver, M. (2004) Improving patient care through molecular diagnostics, Semin Oncol, 31, 14-20.

Perou, C.M., Jeffrey, S.S., van de Rijn, M., Rees, C.A., Eisen, M.B., Ross, D.T., Pergamenschikov, A., Williams, C.F., Zhu, S.X., Lee, J.C., Lashkari, D., Shalon, D., Brown, P.O. and Botstein, D. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, Proc Natl Acad Sci U S A, 96, 9212-9217.

Pitman, E.J.G. (1937) Significance tests which may be applied to samples from any population, Royal Staistical Society Supplement, 4, 119-130 and 225-132.

Pitman, E.J.G. (1938) Significance tests which may be applied to samples from any population. Part III. The analysis of variance test., Biometrika, 29, 322-335.

Pontius, J.U., Wagner, L. and Schuler, G.D. (2003) UniGene: a unified view of the transcriptome. In, The NCBI Handbook. National Center for Biotechnology Information, Bethesda, MD.

Porter, J.D., Khanna, S., Kaminski, H.J., Rao, J.S., Merriam, A.P., Richmonds, C.R., Leahy, P., Li, J. and Andrade, F.H. (2001) Extraocular muscle is defined by a fundamentally distinct gene expression profile, Proc Natl Acad Sci U S A, 98, 12062-12067.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1993) Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge.

Pruitt, K.D., Katz, K.S., Sicotte, H. and Maglott, D.R. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI, Trends Genet, 16, 44-47.

Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources, Nucleic Acids Res, 29, 137-140.

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

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

R Development Core Team (2004) R: A language and environment for

statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1997) GeneCards: integrating information about genes, proteins and diseases, Trends Genet, 13, 163.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support, Bioinformatics, 14, 656-664.

Reimers, M. (2003) An (Opinionated) Guide to Microarray Data Analysis. Karolinska Instutitute, Dept. of Biosciences.

Ron, M., Rea, M., Weller, J., Band, M., Lewin, H., Medrano, J. and Gregg, J. (2003) Analysis of Microarray Suite 5.0 change p-values in Affymetrix replicated microarray experiments. Agricultural Research Organisation.

Rosetta Biosoftware Rosetta Resolver. Seattle, WA.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V. and Quackenbush, J. (2003) TM4: a free, open-source system for microarray data management and analysis, Biotechniques, 34, 374-378.

Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., Rosen, N., Shmoish, M., Peter, Y., Glusman, G., Feldmesser, E., Adato, A., Peter, I., Khen, M., Atarot, T., Groner, Y. and Lancet, D. (2003) Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE, Nucleic Acids Res, 31, 142-146.

Saidi, S.A., Holland, C.M., Kreil, D.P., MacKay, D.J., Charnock-Jones, D.S., Print, C.G. and Smith, S.K. (2004) Independent component analysis of microarray data in the study of endometrial cancer, Oncogene, 23, 6677-6683.

Sandberg, R., Yasuda, R., Pankratz, D.G., Carter, T.A., Del Rio, J.A., Wodicka, L., Mayford, M., Lockhart, D.J. and Barlow, C. (2000) Regional and strain-specific gene expression mapping in the adult mouse brain, Proc Natl Acad Sci U S A, 97, 11038-11043.

Schadt, E.E., Li, C., Ellis, B. and 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.

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

Schulze, A. and Downward, J. (2001) Navigating gene expression using microarrays--a technology review, Nat Cell Biol, 3, E190-195.

Shapiro, S.S. and Wilks, M.B. (1965) An analysis of variance test for normality., Biometrika, 52, 591-611.

Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J.C., Dwight, S.S., Kaloper, M., Weng, S., Jin, H., Ball, C.A., Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. and Cherry, J.M. (2001) The Stanford Microarray Database, Nucleic Acids Res, 29, 152-155.

Silicon Genetics GeneSpring. Redwood City, CA.

Skolnick, J. and Fetrow, J.S. (2000) From genes to protein structure and function: novel applications of computational approaches in the genomic era, Trends Biotechnol, 18, 34-39.

Sonnhammer, E.L., Eddy, S.R. and Durbin, R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments, Proteins, 28, 405-420.

Southern, E., Mir, K. and Shchepinov, M. (1999) Molecular interactions on microarrays, Nat Genet, 21, 5-9.

Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis, Journal Molecular Biology, 98, 503-517.

Stamey, T.A., Warrington, J.A., Caldwell, M.C., Chen, Z., Fan, Z., Mahadevappa, M., McNeal, J.E., Nolley, R. and Zhang, Z. (2001) Molecular genetic profiling of Gleason grade 4/5 prostate cancers compared to benign prostatic hyperplasia, J Urol, 166, 2171-2177.

Sturn, A., Quackenbush, J. and Trajanoski, Z. (2002) Genesis: cluster analysis of microarray data, Bioinformatics, 18, 207-208.

Svaren, J., Ehrig, T., Abdulkadir, S.A., Ehrengruber, M.U., Watson, M.A. and Milbrandt, J. (2000) EGR1 target genes in prostate carcinoma cells identified by microarray analysis, J Biol Chem, 275, 38524-38531.

Tan, F.L., Moravec, C.S., Li, J., Apperson-Hansen, C., McCarthy, P.M., Young, J.B. and Bond, M. (2002) The gene expression fingerprint of human heart failure, Proc Natl Acad Sci U S A, 99, 11387-11392.

Teague, T.K., Hildeman, D., Kedl, R.M., Mitchell, T., Rees, W., Schaefer, B.C., Bender, J., Kappler, J. and Marrack, P. (1999) Activation changes the spectrum but not the diversity of genes expressed by T cells, Proc Natl Acad Sci U S A, 96, 12691-12696.

Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, L.P. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles, Genome Res, 11, 1227-1236.

Tilstone, C. (2003) DNA microarrays: vital statistics, Nature, 424, 610-612.

Tong, L., Shen, H., Perreau, V.M., Balazs, R. and Cotman, C.W. (2001) Effects of exercise on gene-expression profile in the rat hippocampus, Neurobiol Dis, 8, 1046-1056.

Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D. and Altman, R.B. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data, Bioinformatics, 18, 1454-1461.

Tsai, C.A., Wang, S.J., Chen, D.T. and Chen, J.J. (2004) Sample size for gene expression microarray experiments*, Bioinformatics.

Tudor, M., Akbarian, S., Chen, R.Z. and Jaenisch, R. (2002) Transcriptional profiling of a mouse model for Rett syndrome reveals subtle transcriptional changes in the brain, Proc Natl Acad Sci U S A, 99, 15536-15541.

Tukey, J.W. and McLaughlin, D.H. (1963) Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization 1, Sankhya A, 25,, 331 -352.

Tumor Analysis Best Practices Working Group (2004) Expression profiling--best practices for data generation and interpretation in clinical trials., Nature Reviews Genetics, 5, 229-237.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, Proc Natl Acad Sci U S A, 98, 5116-5121.

Unger, M.A., Rishi, M., Clemmer, V.B., Hartman, J.L., Keiper, E.A., Greshock, J.D., Chodosh, L.A., Liebman, M.N. and Weber, B.L. (2001) Characterization of adjacent breast tumors using oligonucleotide microarrays, Breast Cancer Res, 3, 336-341.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge,

W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. (2001) The sequence of the human genome, Science, 291, 1304-1351.

Virtaneva, K., Wright, F.A., Tanner, S.M., Yuan, B., Lemon, W.J., Caligiuri, M.A., Bloomfield, C.D., de La Chapelle, A. and Krahe, R. (2001) Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics, Proc Natl Acad Sci U S A, 98, 1124-1129.

Wang, E.Q., Lee, W.I., Brazeau, D. and Fung, H.L. (2002) cDNA microarray analysis of vascular gene expression after nitric oxide donor infusions in rats: implications for nitrate tolerance mechanisms, AAPS PharmSci, 4, E10.

Warner, E.E. and Dieckgraefe, B.K. (2002) Application of genome-wide gene expression profiling by high-density DNA arrays to the treatment and study of inflammatory bowel disease, Inflamm Bowel Dis, 8, 140-157.

Wei, C., Li, J. and Bumgarner, R.E. (2004) Sample size for detecting differentially expressed genes in microarray experiments, BMC Genomics, 5, 87.

Welch, B.L. (1947) The generalization of 'students' problem when several different population variances are involved, Biometrika, 34, 28-35.

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

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pontius, J.U., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L. and Yaschenko, E. (2005) Database resources of the National Center for Biotechnology Information, Nucleic Acids Res, 33 Database Issue, D39-45.

Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. and Wagner, L. (2003) Database resources of the National Center for Biotechnology, Nucleic Acids Res, 31, 28-33.

Wikipedia (2005) Dirichlet distribution.

Wilcox, R. (2001) Fundamentals of modern statistical methods. Substantially improving power and accuracy. Springer Verlag, New York, NY.

Wolfinger, R. and Chu, T. (2002) Who are those strangers in the latin square. CAMDA 2002.

Wolkenhauer, O., Carla Moller-Levet, C. and Sanchez-Cabo, F. (2002) The curse of normalization, Comparative and Functional Genomics, 3, 375–379.

Wu, Z., R.A., I., Gentleman, R., Martinez Murillo, F. and Spencer, F. (2004) A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Johns Hopkins University.

Xu, R. and Li, X. (2003) A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data, Bioinformatics, 19, 1284-1289.

Yuen , K.K. (1974) The two-sample trimmed t for unequal population variances, Biometrika, 61, 165-170.

Yuen, K.K. and Dixon, W.J. (1973) The approximate behavior and performance of the two-sample trimmed t, Biometrika, 60, 369-374.

Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S., Bussey, K.J., Riss, J., Barrett, J.C. and Weinstein, J.N. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data, Genome Biol, 4, R28.

Zhong, S., Li, C. and Wong, W.H. (2003) ChipInfo: Software for extracting gene annotation and gene ontology information for microarray analysis, Nucleic Acids Res, 31, 3483-3486.

Zien, A., Aigner, T., Zimmer, R. and Lengauer, T. (2001) Centralization: a new method for the normalization of gene expression data, Bioinformatics, 17 Suppl 1, S323-331.