COMPARISON BETWEEN AFFYMETRIX AND ILLUMINA GENE EXPRESSION MICROARRAY PLATFORMS

COMPARISON BETWEEN AFFYMETRIX AND ILLUMINA GENE EXPRESSION MICROARRAY PLATFORMS

By

YIQIANG LUO, B.SC.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2007) McMaster University

(Statistics) Hamilton, Ontario

TITLE: Comparison Between Affymetrix And Illumina Gene Expression Microarray Platforms

AUTHOR: Yiqiang Luo, B.Sc.
(McMaster University)

SUPERVISOR: Dr. Angelo Canty
(McMaster University)

NUMBER OF PAGES: xiii, 88

Abstract

DNA microarray technology is an exploratory tool that can be used to measure thousands of gene expression values simultaneously. Many different microarray platforms have been developed. Various studies have shown that the same experiment performed in different laboratories using the same or different microarray platforms sometimes produce very different results. The lack of reproducibility is becoming a major challenge when comparing the microarray experiments across different labs and different platforms.

Our study is focused on the cross-platform reproducibility between the Affymetrix GeneChip and Illumina BeadChip using two rodent experimental models for Type-1 diabetes. Within-platform reproducibility is also checked for the Illumina platform. Comparisons are carried between the two platforms in terms of the physical array features, the data quality and the chip effects, and finally the SAM analysis. Our study suggested that the overall comparability between these two platforms is fairly poor, however some findings are promising and the results show some agreement with the conclusions found in Barnes *et al.* (2005).

Acknowledgements

I would like to thank my supervisor, Dr. Angelo Canty, for his constant support, giving me suggestions and guidance throughout the whole process of this interesting project.

I am also grateful to all the people from Dr. Jayne Danska's group at The Hospital for Sick Children in Toronto for providing the data. My special thanks go to Michael Sung, for his hard work in producing the first and second Illumina mouse data sets. I learned a lot from meeting all these biologists, which is essential in helping me to understand the research project better. It's a great pleasure to work with all these great people.

I would also like to dedicate this report to my wife, Yanhui Li and my parents, Jianhua Luo and Aizheng Fu. Thank you for all your support, encouragements and understanding you showed throughout all these years.

Finally, I wish to thank Professor Peter McDonald for his teaching and instruction during the years I spend at McMaster University, and Professor Roman Viveros-Aguilera and Noori Akhtar-Danesh, for taking time to be my defence examiners.

List of Tables

List of Figures

Contents

Chapter 1

Background

DNA microarray technology has been widely applied as a tool in scientific research, especially in biotechnology and pharmaceutical applications. It gives us the ability to look at the gene expression profile – the expression pattern of tens of thousands of genes of a given tissue at a given time. This large-scale, high-throughput approach dramatically boosted the pace of biological research.

Microarray technology is still at an early stage of development. Being an interdisciplinary science, it requires collaboration from other related disciplines such as bioinformatics and statistics to borrow the computational power and innovated methodology for microarray data analysis. Compared to the hardware development of the array platform itself, the software support from these co-disciplines is becoming more and more important. Biostatisticians need more sophisticated and robust data analysis methodology to deal with the challenges that arise from various applications and array platforms. One big challenge in microarray experiment is the lack of comparability or reproducibility between studies from different laboratories and different microarray platforms, see Irizarry *et al.* (2005). The first discordance was reported

in Kuo *et al.* (2002). It has been shown that, even within the same platform but different versions of the same array products, it could generate very different data and results in terms of detecting the differentially expressed genes, see Shoemaker and Lin (2005). This reproducibility issue must be verified and better understood when we try to interpret and compare the results from different researchers and different microarray platforms.

This chapter gives the background review to prepare the reader with basic genetic knowledge and the statistical methods that are required in this thesis. The basic biology concepts will be introduced in Section 1.1. Section 1.2 will introduce the physical and annotational features and the comparisons of some major gene expression microarray platforms. Section 1.3 describes the experiments that were performed. Section 1.4 will give the statistical methodologies that have been applied in the data analyses.

1.1 Protein, DNA, Genes and Gene Expression

Watson and Crick first proposed the double-helix structure of the Deoxyribonucleic Acid (DNA) in 1953. DNA is a long polymer composed of simple units called nucleotides (See Figure 1.1). The nucleotides on one strand are linked to each other by phosphodiester bonds, while between the two strands of the helix, the nucleotides are linked together by the hydrogen bonds between the bases. There are only four bases found in DNA, namely, adenine (abbreviated as A), cytosine (C), guanine (G) and thymine (T). Each type of base on one strand forms a bond with just one type of base on the other strand, that is, A with G and C with T. This is called the Watson-Crick complementary base-pairing rule. This one-to-one complementary relationship

controls the DNA duplication that makes it possible for the genetic information to be passed onto the next generation. The sequence information on DNA is vital for the gene expression process which also inspired the DNA microarray technology.

Gene expression is a process whereby DNA sequences are converted into functional proteins. It is a multiple-step process that begins with "transcription", when DNA is transcribed to RNAs (including mRNA, tRNA and other RNAs). The next step is called "translation", in which mRNAs are used as a template to form strings of amino acid as the material to assemble the actual proteins. The assembling is precisely performed according to certain protein coding rules that are "installed" in the DNA sequences. Genes are the particular DNA segments coded with above rule information that are responsible for the particular protein synthesis. Knowing this, we say gene expression actually studies the abundance of the transcribed mRNAs in a biological system at a certain condition.

1.2 Gene Expression Profiling DNA Microarray Platforms

Compared to the early stage of molecular biology experiment, for example the Southern blotting and Northern blotting that work in a "one gene at a time" manner, gene expression profiling microarray platforms put tens of thousands of genes in one single array/chip. Using a library of probe sequence created by some genome sequencing projects, it allows us to look at the overall expression profile of all those genes and their interactions at a given time and under given conditions. For instance, one can compare gene expression between the normal and diseased cells to find the

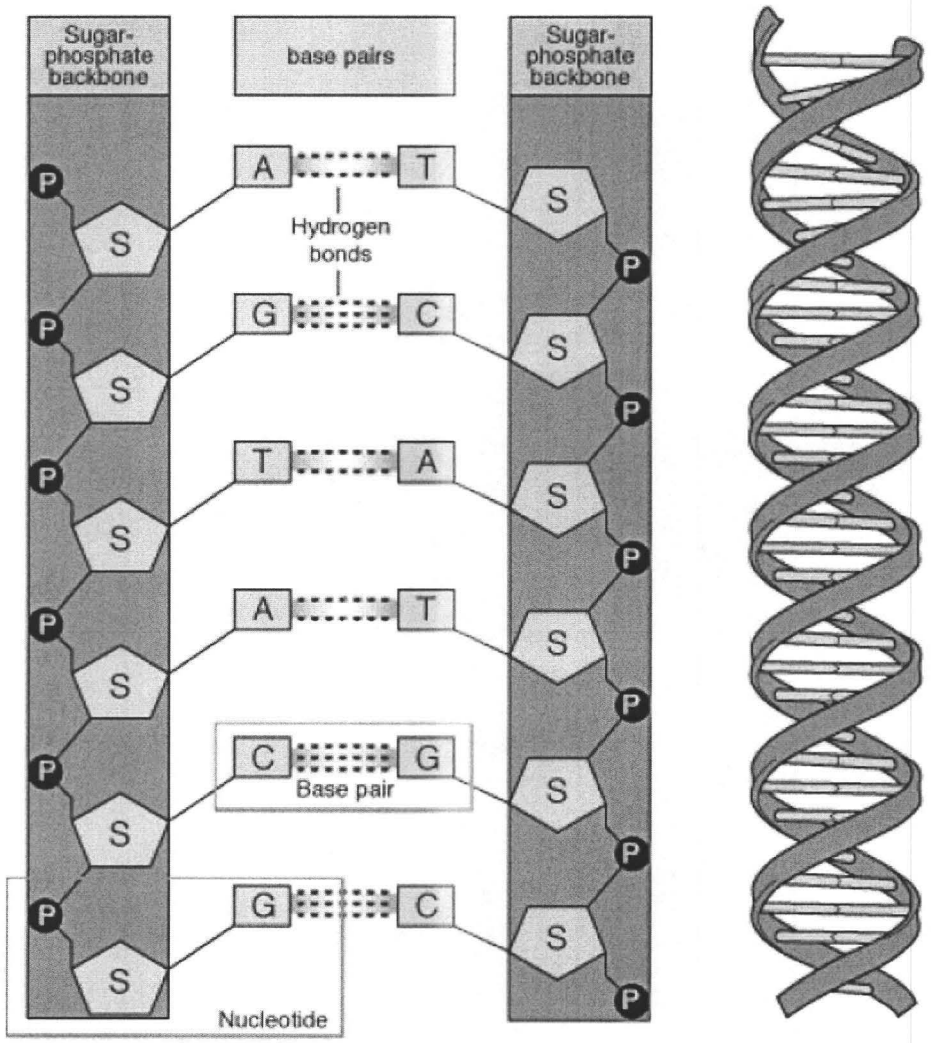


Figure 1.1: *The double helix structure of the DNA. The nucleotides on one strand are linked to each other by the phosphodiester (P) bonds. Between the backbone of strands are the base pairs of adenine (A) with thymine (T) and cytosine (C) with guanine (G). The figure comes from the slides of "Introduction to Genome Biology" presented by Sandrine Dudoit and Robert Gentleman at Bioconductor Short Course 2003.*

differentially expressed genes between two conditions.

DNA microarray technology has been evolving from Southern blotting methodology for almost two decades. Many companies are producing their own commercial array platforms using different designs and materials. But despite these manufacturer differences, it has two major categories: One is the "spotted" microarray, also called two-channel or two-colour microarray. Here colour is referred to the different fluorescent signal of Cy3 and Cy5. The other one is oligonucleotide or single-channel microarray. The major difference between these two platforms is that, the first array hybridizes two biological samples onto the same array such that two fluorescent colors are required to distinguish them during the image processing, while the second array only hybridizes one biological sample on each array.

Oligonucleotide DNA microarrays can be further divided into two subgroups: The long oligonucleotide arrays, whose probes are composed of 60-mer or 50-mer DNA sequences (e.g., Illumina Beadarray), and short oligonucleotide arrays that use 25-mer (e.g., Affymetrix GeneChip) or 30-mer of probe sequence design.

Among these gene expression profiling microarray platforms, Affymetrix GeneChip and Illumina BeadChip are two major ones. Especially for Affymetrix GeneChip, it has been trusted by many researchers and laboratories and various commercial software packages have been developed to deal with the GeneChip data. Compared to Affymetrix GeneChip, Illumina BeadChip is a relatively new and promising array technology which is becoming more and more popular because of its many special features. The following sections discuss the physical and technical features and the differences between these two microarray platforms.

1.2.1 Affymetrix GeneChip^{TM} Platform

Affymetrix GeneChip uses technology similar to that used in computer silicon chip manufacturing. But instead of masking the silicon material on the array surface, Affymetrix uses masks and photolithographic process to control the oligonucleotide synthesis on the glass/plastic array surface. For the probe design, "in situ" 25-mer oligonucleotide gene-specific probes are used, more specifically, probe sets that constructed by 11 to 20 different probe pairs are used to match different genes. The probe pair design including one mismatch (MM) probe and one perfect match (PM) probe. The MM probe is used to control non-specific bindings during hybridization. One special feature of GeneChip array is that, each probe pair is attached to a predefined location on the array surface. Figure 1.2 illustrates the structure of the GeneChip and the hybridization mechanism and the laser scanned image. In our experiments, Affymetrix "Mouse430 v2" and rat "RAE230 v2" GeneChips are used for the mouse and rat experiments respectively.

1.2.2 Illumina Sentrix^{TM} BeadChip Platform

Compared to Affymetrix GeneChip platform, Illumina's BeadChip is a relatively new technology using different assays and designs. We used "Mouse_Ref-8_v1" and rat "RatRef-12_v1" BeadChips in our mouse and rat experiments respectively.

Figure 1.3 illustrates the probe design of the Illumina BeadChip. Instead of using 25-mer "in situ" RNA sequences that attached to predefined locations, Beadarray immobilizes standard 50-mer long oligonucleotide probes along with a 29-mer address sequence onto the $3\mu m$ silica microbeads. For each gene-specific long oligonucleotide probe (probe/bead type), the number of bead is a random variable with a mean

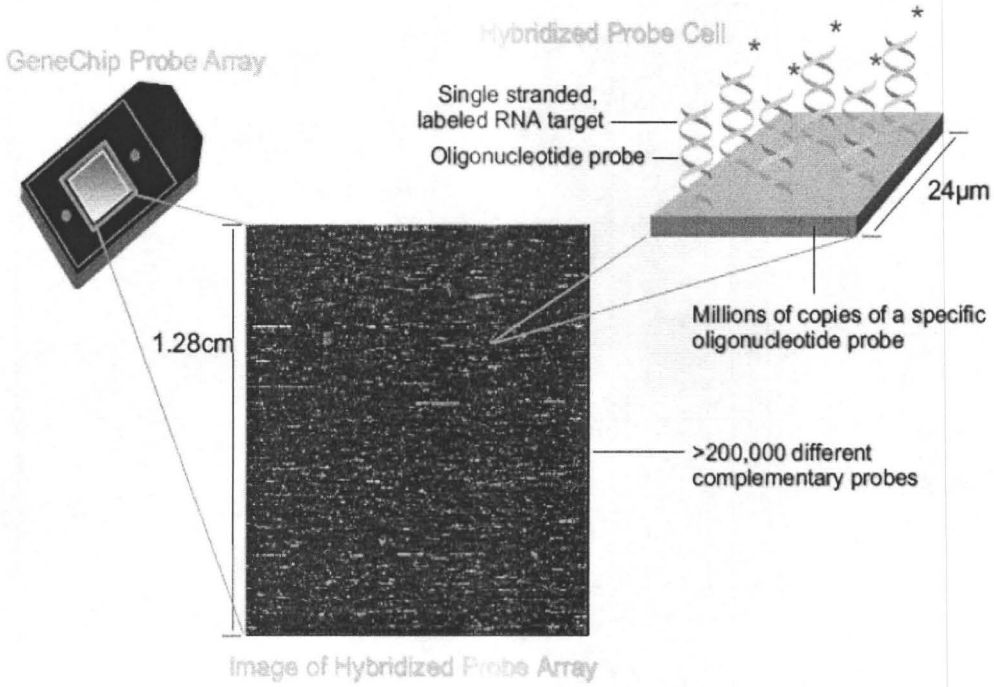


Figure 1.2: *An illustration of structure and design of the Affymetrix GeneChip (oligonucleotide) array. The figure comes from the slides of "DNA Microarray Data Oligonucleotide Arrays" presented by Sandrine Dudoit, Robert Gentleman, Rafael Irizarry and Yee HwaYang at Bioconductor Short Course 2003.*

equals to 30. With these "internal replicate", trimmed mean values are calculated for the gene-specific probes as the gene expression value, see Kuhn *et al.* (2004) for more details. This redundancy ensures the statistical accuracy and robustness in gene expression measurements. According to Illumina documentation, it is found that the 50-mer long gene-specific oligonucleotide showed superior performance than the GeneChip 25-mer probe sequence sets. Another feature of the Illumina BeadChip is that all arrays are randomly assembled. All different bead types come from a master beads pool and then randomly placed onto the wells on the array substrate (see Figure 1.4). This randomness minimizes the effects of spatially localized artifacts, besides, the beads pool can be customer designed and so adjustable from application to application. Also because of the randomness, all probes need to be identified in order to get the gene expression data. This is done by a procedure called decoding, in which the address sequence information of each bead type is decoded and translated into (X, Y) co-ordinates by some algorithm in Gunderson *et al.* (2004). Instead of using MM and PM probe set design, Illumina Beadarray uses "negative control" and other quality control probes. In addition to the above, another important and innovative feature of BeadChip platform is its packaging design, it packs multiple (6-12) Beadarrays on each single BeadChip, which indicates that hybridization and other processes are performed in a parallel manner for all Beadarrays on the same chip, while in the Affymetrix experiment, all GeneChips are processed separately and individually. In all, as mentioned in Steemers and Gunderson (2005), the Illumina BeadChip platform "enables parallel interrogation of multiple whole genomes or focused sets of target genes across large sample populations with low cost per sample". For more details about Illumina Sentrix^{TM} technology please refer to Kuhn *et al.* (2004); Illumina (2004, 2005a,b).

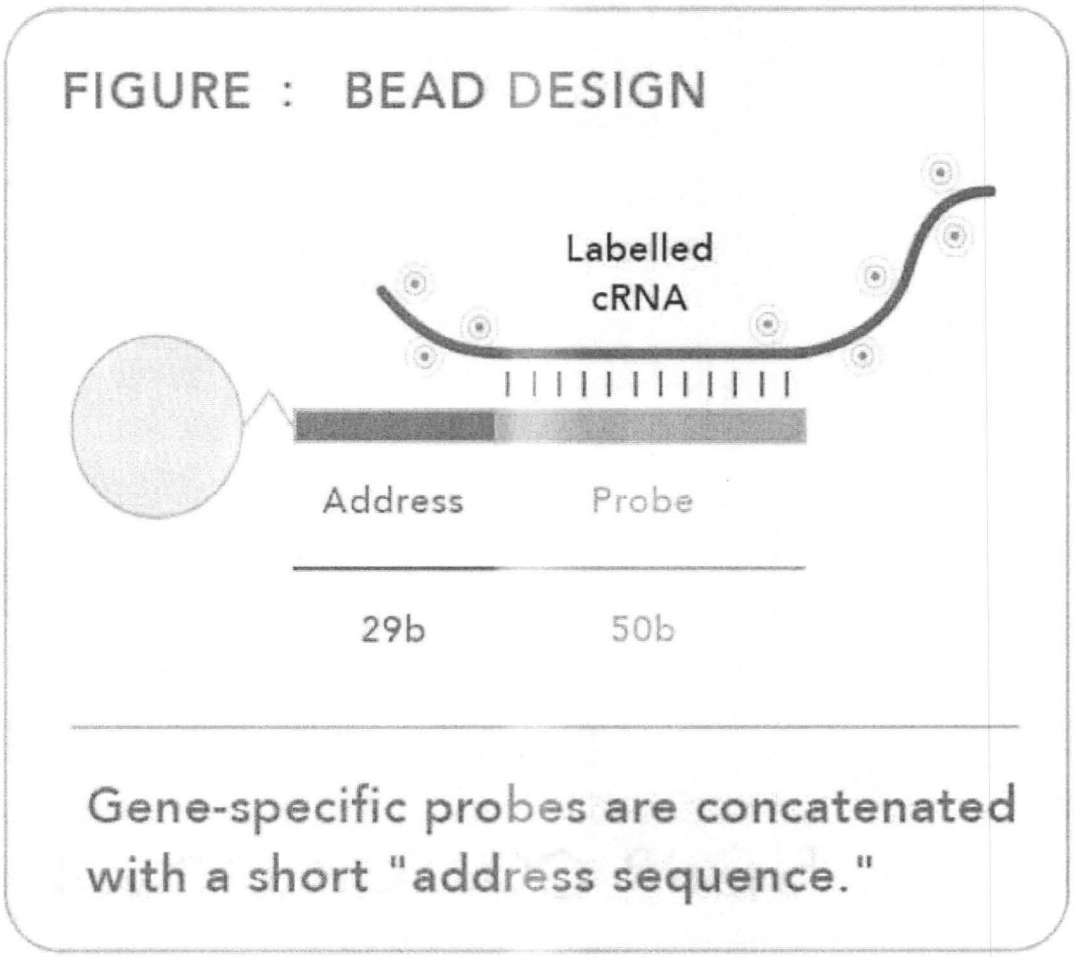


Figure 1.3: *Illustration of the probe design of Illumina direct hybridization assay technology. The figure comes from Illumina website (www.illumina.com).*

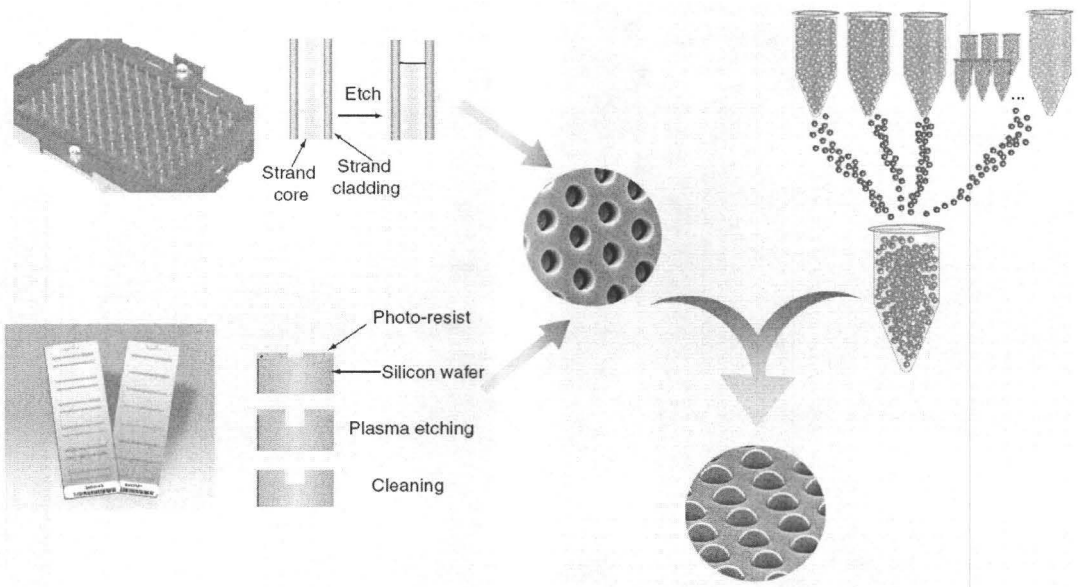


Figure 1.4: *The Illumina BeadChip design of two patterned substrate formats: the Sentrix Array Matrix consisting of 96 optical fiber bundles with 50,000 wells created by acid etching, and the Sentrix BeadChip created using a MEMS-patterned slide substrate. Oligonucleotide are individually immobilized on each bead type, which are subsequently pooled. Bead pools are self-assembled into the patterned substrate, and decoding is performed to determine the identity and location of each bead type. The figure comes from Illumina company profile by Steemers and Gunderson (2005).*

1.2.3 Physical and Annotational Differences Between Affymetrix GeneChip and Illumina BeadChip Platforms

In Table 1.1, we summarize the major physical and annotational differences between the Illumina BeadChip and the Affymetrix GeneChip arrays used in our mouse and rat experiments.

The physical differences in probe design and array packaging have been discussed in previous sections. The rest of this section shows the annotational differences between the two platforms.

We applied the Illumina annotation file provided by the company. Table 1.1 illustrates that the "Mouse_Ref-8_v1" BeadChip employs a design of 24049 different type of probes that map to 18241 unique known mouse genes, also according to the manufacturer's annotation file, each array on the "RatRef-12_v1" BeadChip has a total of 22523 type of probes which map to 21909 unique known rat genes.

For the GeneChip platform, the annotation is based on the package from Bioconductor, where the mappings are based on latest data (2007 April, by the time of analysis) provided by Entrez Gene (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/). The library of "Mouse430 v2" GeneChip includes 45102 probe sets in total, which map to 20958 unique known genes. The rat "RAE230 v2" library consists of 31099 probe sets in total and 14295 unique known genes mapped.

Physical Features

	Affymetrix GeneChip	**Illumina BeadChip**
Probe (set) Design	uses 11-20 probe pairs (MM and PM) of 25-mer "in situ" short-oligonucleotide for each gene	uses 30 copies of same probes of 50-mer standard long-oligonucleotide for each gene
Physical Attachment	predefined location for each probe set	probes are randomly allocated (decoding process required)
Array Packaging	one array per GeneChip	multiple arrays per BeadChip
Per Array Cost	relatively more expensive	relatively cheaper

Annotations

	"Mouse430v2.0" GeneChip	**"MouseRef-8 v1" BeadChip**
Number of probes (sets)	45102	24049
Number of annotated probes (sets)	41382 (91.75%)	23821 (99.05%)
Number of unique genes	20958	18241
	"RAE230 v2" GeneChip	**"RatRef-12 v1" BeadChip**
Number of probes (sets)	31099	22523
Number of annotated probes (sets)	22657 (72.85%)	22326 (99.13%)
Number of unique genes	14295	21909

Table 1.1: *The physical and annotational differences between the Affymetrix GeneChip and Illumina BeadChip microarray platform.*

1.3 Description of Microarray Experiments

1.3.1 Mouse Experiment

In our Type-1 diabetes (T1D) mouse model, two parental strains of mice are used, which are the Non-Obese Resistant (NOR) and Non-Obese Diabetic (NOD) mice. The biological interest focuses on the Idd5 locus on the mouse genome, and the question is: how does this genetic locus contribute to T1D? Microarray experiments are designed using the NOD.NOR-Idd5 congenic and the NOD mice to show the difference in their gene expression profiling. The congenic mice are bred from the two original parental strains through multiple back-crosses to the background parental strains. After a number of generations, the NOD.NOR-Idd5 congenic mice have such a purified genome like the NOD parental mice that the only difference in their genomes is at the Idd5 locus, which includes only dozens of annotated genes. The mouse microarray experiments use four biological replicates either from the NOD.NOR-Idd5 single congenic strain or the NOD mice strain. Two technical replicates BeadChips are prepared, namely, each replicate BeadChip has eight arrays, four of which are from the congenic strain and the other four are from the NOD strain.

1.3.2 Rat Experiment

In our T1D rat model, two strains of rat called BB and BB7B are used in our rat model. BB7B rat is a double congenic strain on a BB background with the different loci on chromosomes 8 and 13 coming from a WF (Wistar-Firth) rat. Again, we are trying to answer the same question: how different is the gene expression profile (especially for genes on those loci) across the two rat strains? The rat microarray

experiment design uses three biological replicates from either the BB7B or the BB rat strain, and two technical replicate BeadChips are prepared. Notice that, compared to the mouse model, the BB7B double congenic rat strain has a relatively wider genome region of interest than the single Idd5 locus of the mouse model. More specifically, the rat model contains more annotated genes that are expected to be differentially expressed than the mouse model.

The objective of the experiments is to find a list of most differentially expressed genes between these two mouse or rat strains. The technical replicates are used to investigate the reproducibility within the Illumina platform. For the cross-platform comparison, we will examine the lists of differentially expressed genes produced using the Illumina and Affymetrix microarray platforms.

1.4 Statistical Analysis

In the early stage of microarray data analysis, the statistical analyses stayed in a "one-gene-at-a-time" paradigm. All you needed to know was the two-sample t-test or some other "t-like" tests. However, when it comes to microarray data analysis, instead of testing one gene, we are testing tens of thousands of genes, which means we are dealing with a family of tens of thousands of t-tests simultaneously. This is exactly the setup of a multiple hypothesis testing problem that will be discussed in Section 1.4.3.

Some other statistical analyses are used for microarray data quality assessment, for example cluster analysis, which will be discussed in Section 1.4.2.

Sample size in microarray experiments is another problem in statistical analysis. Microarray experiments are expensive, so doing a large number of biological repli-

cates is not feasible due to resource limitations. To achieve the maximum statistical reliability and robustness, permutation and/or Bayesian techniques (not discussed in this thesis) are required.

1.4.1 Gene Expression Data and Data Preprocessing

All raw microarray data have to be preprocessed before further statistical analysis. Preprocessing is a three-stage procedure including background correction, normalization, and summarization. In the following sections we will briefly discuss the robust multi-array average (RMA) and the quantile normalization methods used in our analysis, both methods are described in Irizarry *et al.* (2003).

Normalization is an important step in microarray data analysis. In general, all normalization algorithms are designed to remove variation of non-biological noise and systematic artifacts within or between arrays so that their values can be made comparable. All forms of normalization achieve this goal by making assumptions about the experimental samples and adjusting their values in a way that would factor out intensity changes arising from experimental variation without affecting the true biological differences. Knowing this, one has to be careful when applying normalization methods and make sure to understand the underlying assumptions of each method and to decide if they apply in the case of your experiment.

Under the assumption of our rodent models, that most probes are not differentially expressed, the distributions of the expression data are expected to be similar across the arrays. The simplest method works just by multiplying each array by a constant to make the mean (median) intensity same for each individual array. Other methods like quantile normalization proposed by Bolstad (2001) can deal with non-linearity

quite well. The algorithm assumes that there is an underlying common distribution of intensities across the arrays. The algorithm first ranks all intensities inside each array/data set, then average intensity per rank is calculated across the arrays and the intensities for each data set are then recalculated according to the original ranking. This method recalculates the intensities in each array according to the original ranking and also makes sure that all arrays have the same quantile distribution across different arrays. More discussion of normalization methods and other preprocessing methods can be found in Yang (2006), Bolstad *et al.* (2003) and Irizarry *et al.* (2003).

First proposed by Irizarry *et al.* (2003), robust multi-array average (RMA) is a preprocessing method particularly designed for Affymetrix GeneChip data. It summarizes the expression data using the background-corrected, normalized, and log-transformed PM values on GeneChips. The normalization method we used in RMA is the default quantile normalization that proposed by Bolstad *et al.* (2003). More details about the RMA method can be found in Irizarry *et al.* (2003).

In our study we only deal with the summary data produced from the platform software such as the RMA for Affymetrix and Beadstudio software for Illumina data. More specifically, in our analyses, Affymetrix GeneChip "CEL" raw data are preprocessed to produce the summary data using the RMA methods. The BeadChip summary data are produced from the Beadstudio software, and the quantile normalization is then applied to both the Affymetrix and BeadChip summary data.

After the preprocessing, all DNA microarray data has a matrix format of $J \times n$ gene expression values, where J represents the total number of probes (or probe sets) that were represented in one single array and n represents the number of samples or replicates used in the experiment.

1.4.2 Array and Data Quality Assessment Using Graphical Diagnostic Plots

Data quality is an important issue in the microarray experiment because there are many sources of variation that can introduce bias into microarray data. The data quality assessment can be performed before or after the preprocessing step.

The graphical diagnostic plots, such as the "XY" plot, "MA" plot, histograms and box plot are often used in microarray data analysis to visually check the overall data quality before and after the normalization. Like the XY scatter plot that plots the Y-values (expression values from array 2) on the Y-axes against the X-value (expression values from array 1) on the X-axes, MA plot replaces X-values with A-values on the horizontal axes and M-values on the vertical axes instead of using the Y-values, where the A-values and M-values are the average and difference of the expression values of the two arrays which can be defined in the following formulae. Besides, the axes of MA plot can be transformed by some scalar function like $\log_2$.

In oligonucleotide DNA array, the values of M_i and A_i for each probe i could be calculated by the formulae:

$$M_i = \log_2(E_{i1}) - \log_2(E_{i2})$$
$$A_i = \tfrac{1}{2}(\log_2(E_{i1}) + \log_2(E_{i2}))$$

Where E_{i1} and E_{i2} are the expression intensity of probe or probe set i on array 1 and array 2 respectively.

By the above definition, the MA plot shows the difference of the log-intensity against their average for each probe (set) on two individual arrays. If two arrays are very similar to each other, we would expect to see a MA plot such that all points are centered around the zero horizontal line (because M values are near zero). Meanwhile,

all points on the XY plots should be all centered around the $Y = X$ line. These scatter plots can be used to diagnose problems such as bias between arrays, etc.

Cluster analysis is a common statistical technique to classify or partition the whole data set into several subgroups according to some predefined distance metric. In gene pattern discovery, hierarchical clustering is the most commonly used technique for extracting the underlying gene cluster structure. The traditional representation of this hierarchical relationship is a tree diagram called the cluster dendrogram as illustrated in Figure 2.4. It groups the most similar individual elements (leaves) at one end and a single cluster (root) containing every element at the other using certain distance metric. The algorithm builds (agglomerative) or breaks up (divisive) a hierarchy of clusters, more specifically, divisive algorithm begins at the top of the tree, whereas agglomerative algorithm starts at the bottom to construct the clusters. The most commonly used distance metrics in gene expression hierarchical clustering are the Euclidean distance and $1 - \rho$, where ρ is the Pearson's correlation coefficient measurement. Other distance metrics are also available. We used the Euclidean distance as the distance metric in our analysis for its familiarity to the biologists. More discussion of hierarchical clustering can be found in Arabie *et al.* (1996).

1.4.3 Evaluating Differential Expression (DE) - Multiple Hypothesis Testing

The problem of Traditional Statistical Hypothesis Testing - "t-test" As mentioned previously, the common objective of the microarray experiment is to identify the differentially expressed genes across two biological conditions.

The traditional procedure of selecting differentially expressed genes uses the t-test

to test the underlying null hypothesis:

$H_{0,i}$: *gene i is not differentially expressed across the two biological conditions*

Conventional t-statistic and other modified t-like test statistics are applied to log-expression values for each gene and a p-value can be found for each gene. The next step is to choose a significance level for the p-values (e.g., $p < 0.05$) according to the study objective. But this approach could be problematic sometimes as we will discuss in the following sections.

In microarray data analysis, it is essentially a multiple hypothesis testing problem that can be described as in Table 1.2. For a total of m individual hypothesis tests, which is equal to the total number of genes on each array, suppose m_0 is the total number of actual true H_0 hypothesis for which we make U correct decisions and R wrong decisions (type I error/false positive); $m_1 = m - m_0$ is the total number of actual true H_1 hypothesis comprising S correct decision and T wrong decision (type II error/false negative). Notice that all these numbers are assumed known, but we actually will never know the exact numbers of m_0, S, T, U and V, they are all unobservable random variables. Only R is an observable random variable.

To control the multiple test error rate, various ways of error measurements have been proposed. These including the traditional Family Wise Error Rate (FWER) and Bonferroni correction as well as more recent methods using the False Discovery Rate (FDR) or Positive False Discovery Rate (pFDR). We will discuss these concepts in the following paragraphs.

Family Wise Error Rate (FWER) In the literature, the Family Wise Error Rate (FWER) is proposed and defined as the probability of making more than one false

	Not Reject	Reject	Total
True H_0	U	V	m_0
True H_1	T	S	$m - m_0$
Total	$m - R$	R	m

Table 1.2: *The problem of Multiple Tests.*

discoveries, which can be formulated by the following:

$$FWER = \Pr(V > 1)$$

Accordingly, Bonferroni correction suggested to use a much smaller p-value $= \alpha/m$ for each individual testing, where the α is the usual overall error rate control for the tests. FWER and Bonferroni correction are too conservative usually yielding no significant genes at all. Compared to the FDR approach, these methods are especially of low power when large number of true H_1 are expected for large m, see Benjamini and Hochberg (1995).

False Discovery Rate (FDR), pFDR and Q-value The False Discovery Rate (FDR) was first proposed by Benjamini and Hochberg (1995) as the expected proportion of the false positive among those rejected hypothesis using a sequential p-value controlling approach, and FDR is defined as zero if $\mathbf{R} = 0$. It uses the observed p-values to estimate the rejection region such that on average the $FDR < \alpha$ for some pre-chosen α value. But Benjamini and Hochberg's FDR is controlled for all values of m_0, in other words, it did not use the information in the data when estimating the $\hat{m_0}$.

$$\mathbf{FDR} = E\left(\frac{\mathbf{V}}{\mathbf{R}}\right)$$

The Positive False Discovery Rate (pFDR) and the concept of q-value was proposed in Storey (2003). It is defined as the same expected proportion as FDR but conditioned on that there is at least one rejected hypothesis ($\mathbf{R} > 0$):

$$\mathbf{pFDR} = E\left(\frac{\mathbf{V}}{\mathbf{R}}\middle|\mathbf{R} > 0\right)$$

In Storey (2002), Storey proposed a different methodology and applied it to his definition of pFDR and the calculation of q-values. He uses the opposite approach to Benjamini and Hochberg's sequential p-value method by fixing the rejection region and then estimating the corresponding pFDR, and simulation study shows that the new method offers increased applicability, accuracy and power compared to Benjamini and Hochberg's method.

The pFDR analogue to p-values is the q-value which has a special statistical relationship with the p-value. It measures the significance in terms of the FDR rather than the type I error. The relationship between the p-values and the q-values can be illustrated in the following formula:

In a multiple hypothesis testing problem, for each individual test, traditional p-value for an observed t-statistic when $T = t$ is:

$$p\text{-values} = \min_{\Gamma: t \in \Gamma}\{Pr(T \in \Gamma \mid H_0)\}$$

where $\{\Gamma\}$ is the rejection region(s). While q-value of this test can be defined as:

$$q\text{-value} = \inf_{\Gamma: t \in \Gamma}\{pFDR(\Gamma)\}$$

From the formulae we discovered that q-value measures the significance of the test with respect to the pFDR – it is the minimum pFDR that can be obtained given that the test was significant (rejected for the rejection region $\{\Gamma\}$). For more details see Storey (2002).

Significance Analysis of Microarray(SAM) The SAM method proposed by Tusher *et al.* (2001) was designed to evaluate the significance of changes in gene expression pattern in comparing two different biological conditions. SAM assigns a score, called d-statistic which is a modified t-statistic to each gene (actually probes or probe sets) as a measure of the relative difference between two conditions. A gene is then labeled as significant if its score exceeds a threshold. The ideas regarding the error rate control discussed in previous paragraphs, namely the pFDR and q-values are implemented in the SAM package.

Let's suppose we have a multiple test problem of m genes and n arrays from two experimental conditions (say, n_1 arrays from condition 1 and $n-n_1$ from condition 2). Let $\bar{x}_{i1}$ and $\bar{x}_{i2}$ be the average gene expression values for gene i in the two conditions. Let $s_i = SE(\bar{x}_{i1} - \bar{x}_{i2})$. The detail of SAM procedure can be summarized in the following four steps:

Step 1: Forming the Test Statistic (d-statistic) For each individual gene or test, the modified t-statistic for gene i is:

$$d_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{s_i + s_0}$$

where s_0 is an adjustment in order to avoid extreme large values of d_i caused by very low variability of a gene across the samples.

Step 2: Calculating the Null Distribution The null distribution of each gene is calculated by permuting the condition labels. For example, if we have $n_1 = n_2 = 3$ samples from each condition and thus 6 samples in total. The permutation is done by permuting the label pool (1 1 1 2 2 2), i.e., randomly assign 3 samples for label 1, then the rest three are set to label 2. Then recalculate the test statistic $d_i^{\star b}$

for the b^{th} possible permutation for gene i, the ordered statistics for this permutation is $d^{\star b}_{(1)} < d^{\star b}_{(2)} < \cdots < d^{\star b}_{(J)}$ for all J genes. Notice that we would have $B = \binom{6}{3} = 20$ permutations in total in this example.

To achieve more accurate and powerful test statistics, Storey and Tibshirani (2001) pooled the null statistics across all genes by taking the average of all ordered permuted d-statistics for gene $j = 1, 2, 3, \cdots, J$:

$$\bar{d}_j = (1/B) \sum_{b=1}^{B} d^{\star b}_{(j)}$$

See Storey and Tibshirani (2001) for more details regarding the pooled null distribution.

Step 3: SAM Plotting and Threshholding (Choosing the Rejection Region) Instead of using the traditional symmetric rejection region ($|d_i| > t$), SAM uses two data-driven cutoff points t_1 and t_2 to reject the hypothesis and calls gene-i significant if $|d_i| < t_1$ or $|d_i| > t_2$. More specifically, SAM plots a scatter plot of the observed and expected ordered d-statistics found from the permutations, then a band of two parallel lines at a distance Δ (threshold) from the 45° line is drawn. Starting from the origin, counts up and to the right until we find the first point that falls outside this band, then all genes to the right of that point are called significant (over-expressed genes), even if they fall inside the band again. Do the same thing to the bottom left corner we find all the significantly under-expressed genes. The upper and lower Y-values form the cutoff points t_1 and t_2. See Figure 2.11 for an illustration.

Step 4: Control the FDR in Some reasonable Fashion Once we have chosen the threshold Δ, the FDR and pFDR can be estimated using the following

estimators and q-value can be calculated for each gene according to the definition. The FDR and pFDR are controlled using the methods as described in Storey (2002).

$$\widehat{FDR}_{\Delta'}(\Delta) = \hat{\pi}_0(\Delta') \cdot \frac{R^0(\Delta)}{R(\Delta) \vee 1}$$

$$\widehat{pFDR}_{\Delta'}(\Delta) = \hat{\pi}_0(\Delta') \cdot \frac{R^0(\Delta)}{Pr(R^0(\Delta) > 0) \cdot [R(\Delta) \vee 1]}$$

where $R(\Delta)$ is defined as the number of $\{d_i : d_i \leq t_1(\Delta) or d_i \geq t_2(\Delta)\}$, and

$$R^0(\Delta) = \frac{\sum_{b=1}^{B} \#\{d_i^{b\star} : d_i^{b\star} \leq t_1(\Delta) \text{ or } d_i^{b\star} \geq t_2(\Delta)\}}{B}$$

which is derived from the permuted d-statistics. $\hat{\pi}_0(\Delta')$ is an estimator of the overall proportion of true null hypothesis by defining another Δ' value, which can be adaptively chosen to minimize the bias and the variance of the estimates, see (Storey and Tibshirani, 2001). Δ' has to be carefully chosen since $\hat{\pi}_0(\Delta')$ is very important to the estimator of FDR and pFDR.

A more practical way to control the FDR is to pre-specify a tolerable level of α that is biologically meaningful. Then take the smallest Δ such that $\widehat{FDR}_{\Delta'}(\Delta) \leq \alpha$.

SAM package allows us to interactively change Δ, as a result, an FDR-by-Delta table can be produced. One can choose the proper threshold by choosing a meaningful FDR. For example, a larger FDR could be appropriate in some application because the two samples are too similar that result too few significant genes with a small FDR choice. In this way, we can use the FDR as the cutoff point that works in a reverse way to control the multiple test error rate.

In our analyses, we used the SAM function written in R by Dr. Angelo Canty. This SAM function was originally designed to investigate the strain effect and give a list of the differentially expressed genes between two different strains. Strain variable

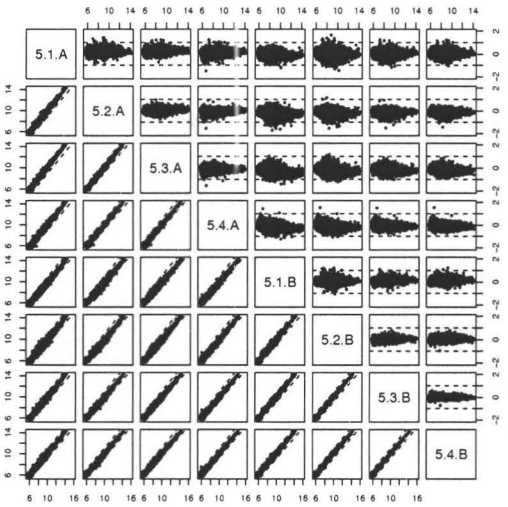
could refer to different biological conditions in a case-control experiment (i.e., the control and treatment groups). In our experiment, the strain variable refers to the same tissue of two biologically different rodent species, for example, the NOD and NOD.NOR-Idd5 mice. In addition to the strain variable, block variables are allowed and thus the permutation technique must be restricted within the blocks to retain the block effect. Block variables could be used to investigate some confounding variables such as the day effect, etc.

Chip reproducibility can be verified by showing two similar SAM plots and similar lists of the significant genes from two SAM analyses, usually of two technical replicate chips. The comparison of the SAM analyses will be discussed in the next few chapters.

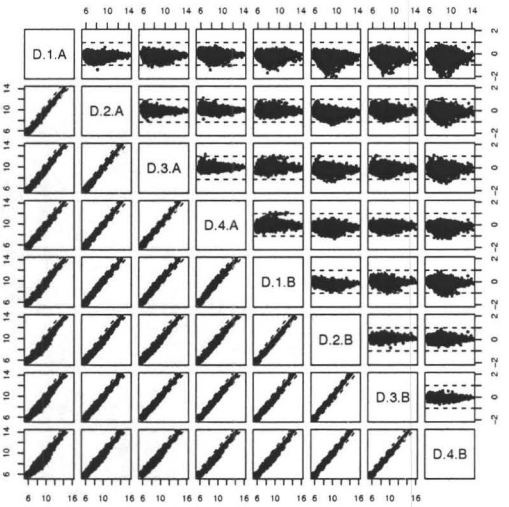
Chapter 2

Within Platform Reproducibility (Illumina)

In this chapter we investigate the chip reproducibility or the chip effect within the Illumina BeadChip platform using our rodent models for Type-1 diabetes. In the mouse model (see mouse data description in Section 1.3.1), two microarray experiments were done using the Illumina Sentrix^{TM} "MouseRef-8 v1" BeadChip, while for the rat model (see rat data description in Section 1.3.2), the "RatRef-12" BeadChips were used. The BeadChip summary data were produced from the Beadstudio software. The within-platform reproducibility is investigated through the data quality assurance and then the SAM analysis. The Affymetrix mouse data was taken from an earlier experiment using the same RNA samples but on "Mouse430v2.0" GeneChips, and the Affymetrix rat data is based on six rat "RAE230v2" GeneChips. Nonetheless, due to the lack of the technical replicates and the platform nature, the reproducibility or chip effect within the Affymetrix GeneChip platform cannot be verified and compared with the Illumina BeadChip platform.



(a) NOD.NOR-Idd5 arrays

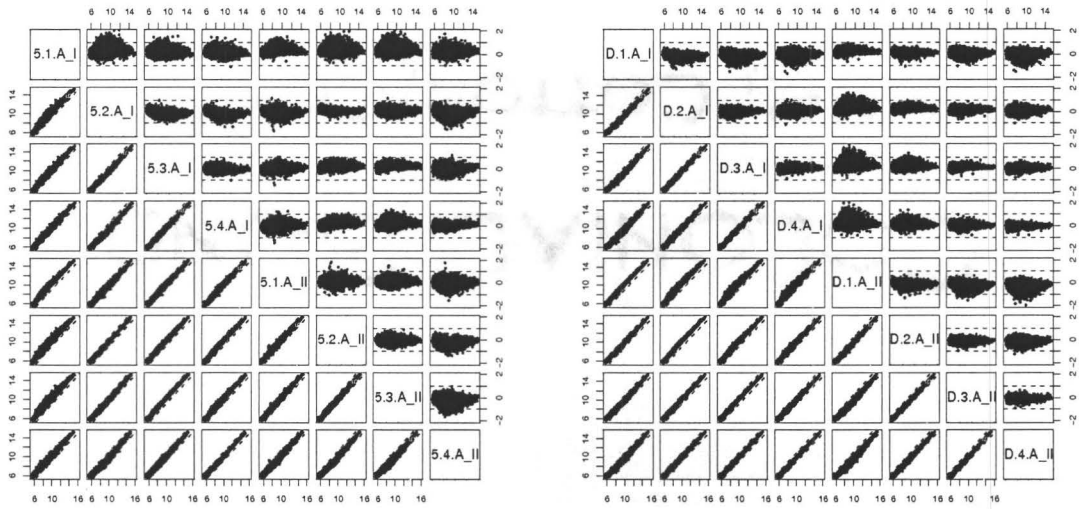


(b) NOD arrays

Figure 2.1: *"MA" and "XY" plots for mouse Beadarrays from NOD.NOR-Idd5 strain and NOD strain of the first experiment. The expression data are* $\log_2$*-scaled but unnormalized.*

2.1 Visual verification of the Data Quality and the Chip Effect

As mentioned in the previous chapter, data quality is a very important issue in microarray experiments and it is necessary to look into it using some quality control methods. Here, we visually check the data quality by applying some graphical diagnostic methods such as the scatter plots, boxplots, histograms and the hierarchical clustering dendrogram of arrays for both Illumina and Affymetrix platforms.



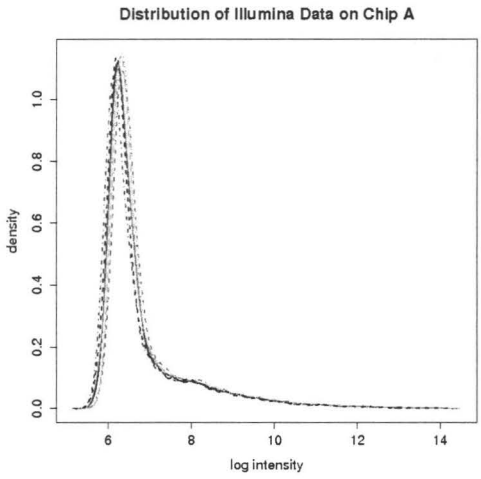
(a) NOD.NOR-Idd5 arrays

(b) NOD arrays

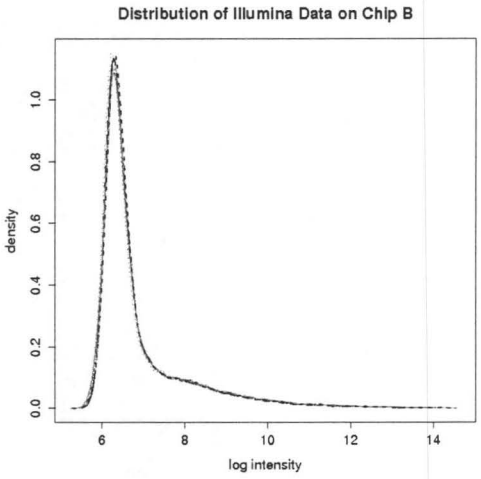
Figure 2.2: *"MA" and "XY" plots for mouse Beadarrays from NOD.NOR-Idd5 strain and NOD strain of the second experiment. The expression data are* $\log_2$*-scaled but un-normalized.*

2.1.1 Illumina "MouseRef-8 v1" BeadChip Data

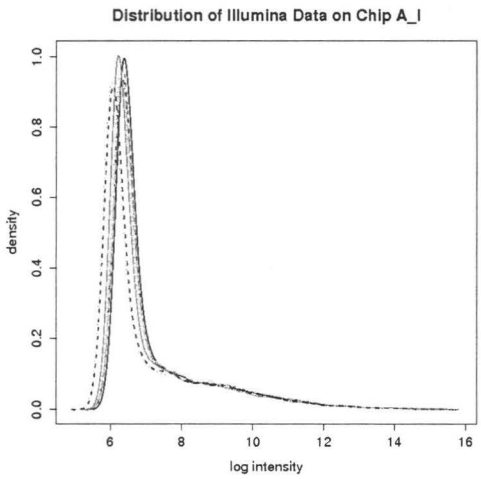
In each of the mouse model experiments, we have two Illumina "MouseRef-8 v1" BeadChips that form the technical replicates, namely, the chip A that replicates chip B in the first experiment. Notice, however, that due to the way the RNA sample was prepared in the first experiment, these technical replicates are actually not pure technical replicates. In the the second mouse experiment, chip A_I and chip A_II are pure technical replicates. On each single "MouseRef-8 v1" BeadChip, there are eight arrays with four biological replicates from conditions of either NOD.NOR-Idd5 strain or NOD strain. To check the BeadChip data quality, we will discuss the following diagnostic plots:



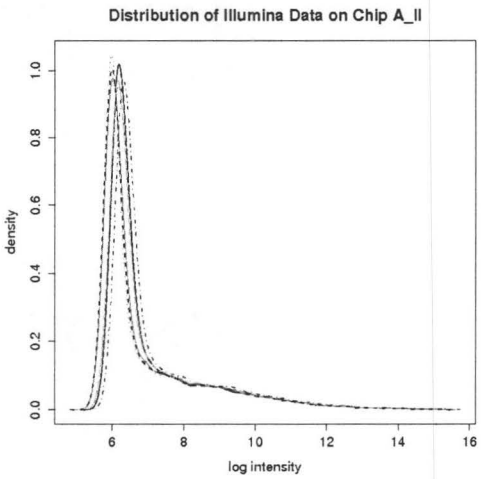
(a) Chip A



(b) Chip B

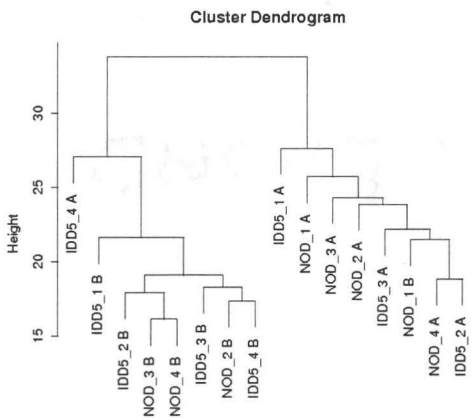


(c) Chip A_I

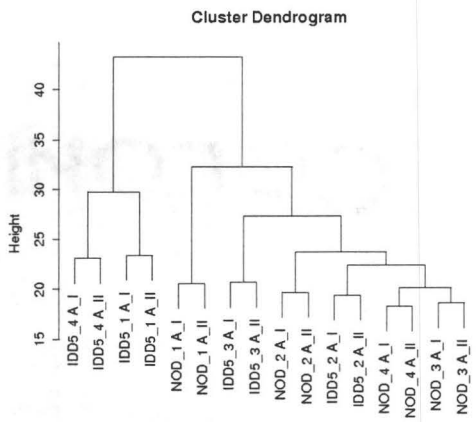


(d) Chip A_II

Figure 2.3: *Smoothed histograms of the un-normalized Beadarray mouse data on chip A, B, A_I and A_II.*

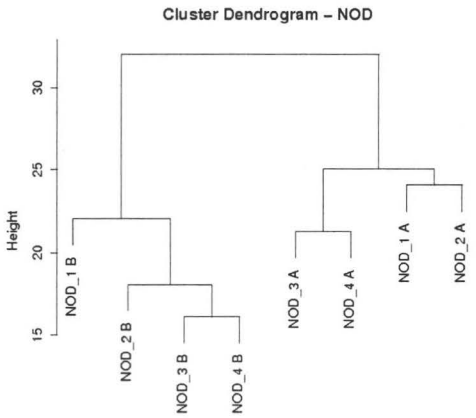


(a) Arrays in first experiment

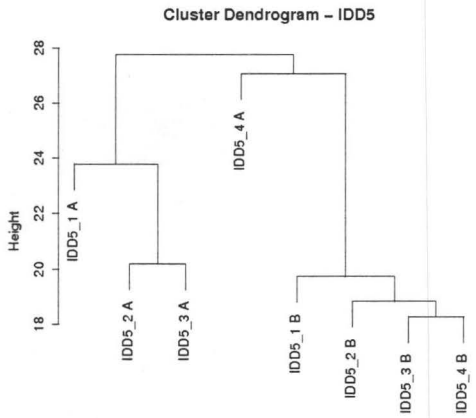


(b) Arrays in second experiment

Figure 2.4: *Clustering all 16 arrays on two BeadChips from both experiments. For first mouse experiment, the array names are distinguished by chip A and B at the end of the names. For the second mouse experiment, the array names are distinguished by chip A_I and A_II.*

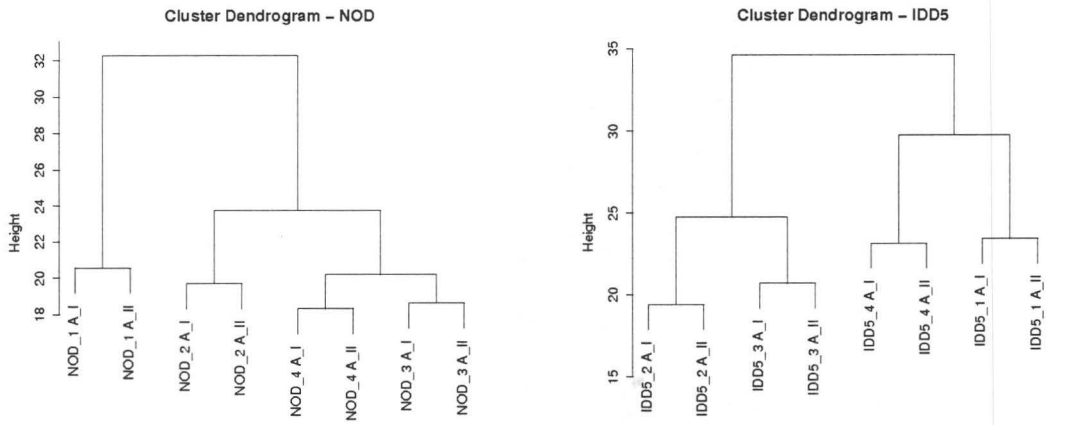


(a) Dendrogram of NOD strain



(b) Dendrogram of NOD.NOR-Idd5 strain

Figure 2.5: *Clustering within the same strain for first mouse experiment.*



(a) Dendrogram of NOD strain

(b) Dendrogram of NOD.NOR-Idd5 strain

Figure 2.6: *Clustering within the same strain for second mouse experiment.*

"MA" and "XY" plot and the distribution plot of the Beadarrays We combine the "MA" and the "XY" pairwise plots in one big plot as in Figures 2.1 and 2.2. Located in the upper right corner of the plot are the "MA" scatter plots and the lower left corner are the "XY" scatter plots. To investigate the chip effect, we group the arrays by different strains. Within the same group (including the technical replicates), all arrays are biologically identical to each other. If the chip effect is small, as discussed in section 1.4.2, we can expect to see an "MA" plot with all points scattered around the zero horizontal line because M's are near zero, and a "XY" plot with all points scattered around the $Y = X$ line. Figures 2.1 and Figure 2.2 give the scatter plots within different strains for both experiments. Both figures showed that the data quality in both experiments is reasonably good and there is no single Beadarray having extreme variation. All arrays are highly correlated since pairwise Pearson's correlation coefficients are greater than 98% for both experiments. This

suggests that the chip effect is small in terms of the correlation coefficients. However, especially for the technical replicate arrays, the second plot does show slightly better similarity between the Beadarrays than the first one in terms of the overall image and the correlation coefficient matrix. Figure 2.3 shows the smoothed histograms of the Beadarray data on chip A and B in first experiment, A_I and A_II in second experiment.

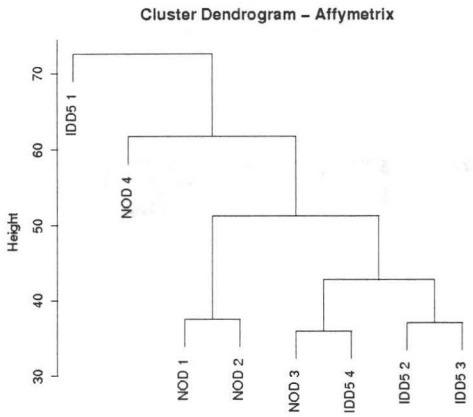
Hierarchical clustering of Beadarrays After the quantile normalization and $\log_2$-scaling, hierarchical cluster analyses were performed using Euclidean distance metric for different groups of Beadarrays to show the similarity/dissimilarity within the same group. To see the big picture, we put all 16 arrays from both the first and the second experiment together. Figure 2.4 shows the overall clustering dendrogram for both experiments, from which we see that, except for two individual arrays (array IDD5.4.A and NOD.1.B), dendrogram of Figure 2.4(a) shows a big cluster between chip A and B in the first experiment. This suggests that the chip effect is much more significant compared to the biological effect. Figure 2.4(b) is the overall dendrogram arising from the second Illumina experiment, in which we find a much smaller chip effect between the BeadChip A_I and A_II compared to Figure 2.4(a). In addition to the grouping within the same experiment, the hierarchical clustering within the same strain group (either NOD or NOD.NOR-Idd5) across two experiments were investigated, Figure 2.5 once again confirms the big chip effect in the first experiment compared to the second experiment as showed in Figure 2.6.

The RNA Preparation The clear chip effect we have seen above is investigated and it reveals that the effect is partially caused by the RNA sample preparation in

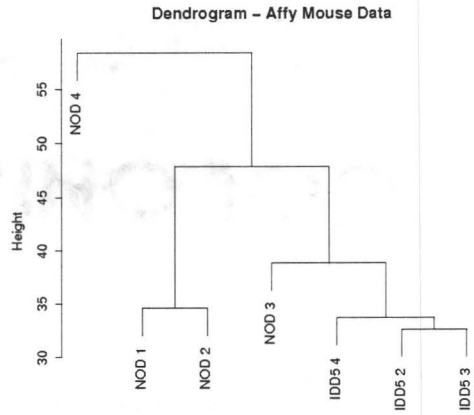
the first experiment. As we pointed out at the beginning that chip A and chip B from the first experiment are actually not pure technical replicates. Compared to the second experiment, some effect due to RNA preparation was introduced in the RNA sample in the first experiment, where the samples were prepared in a parallel way for both chip A and chip B. While in the second experiment, the RNA samples were prepared together before being divided into two parts for the hybridization onto BeadChip A_I and A_II. In this way, it avoided many possible sources of variation. As a result, we point out that in order to take advantage of the Illumina BeadChip platform, the RNA samples for technical replicates on multiple BeadChips should be processed together. This can significantly reduce the experimental bias and variation.

2.1.2 Affymetrix "Mouse430v2.0" GeneChip Data

The array quality of the Affymetrix GeneChips mouse data is investigated through the Pearson's correlation coefficient matrix, the cluster analysis and the smoothed histograms of the density distribution of array data. There are eight GeneChips from one experiment with no technical replicates involved, four of which form the biological replicates that come from either the NOD or NOD.NOR-Idd5 strain. As in the Beadarray analyses, the pairwise Pearson's correlation coefficients (greater than 97%) between the eight GeneChip arrays suggest that all arrays are highly correlated. Figure 2.8 gives the histograms and the boxplots of all eight GeneChips, from the boxplots in Figure 2.8(a) we notice that, the chip "IDD5 1" has a very different mean and quantile values compared to the others. This extremeness is also reflected in the histogram plot of Figure 2.8(b), where this "problem" chip has a shifted distribution from the others.

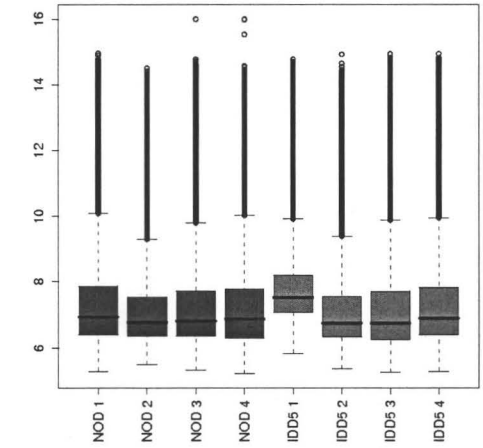


(a) Dendrogram of 8 Affymetrix chips

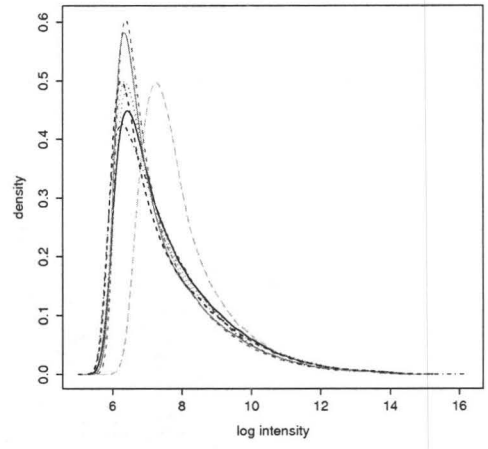


(b) Dendrogram of 7 Affymetrix chips after excluding the "problem" chip

Figure 2.7: *Dendrogram of Affymetrix mouse GeneChips before and after excluding the "problem" chip.*

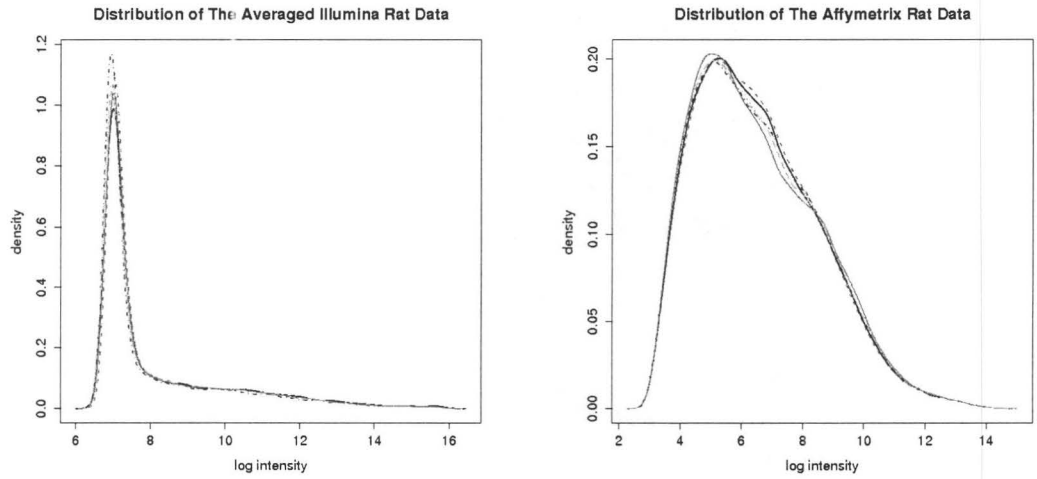


(a) boxplot of 8 Affymetrix chips



(b) Histogram of 8 Affymetrix chips

Figure 2.8: *The boxplot and histogram for the original GeneChip Mouse data, which is un-normalized and* $\log_2$*-scaled.*



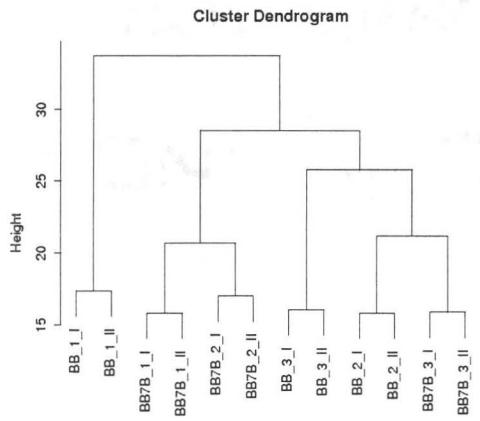
(a) Illumina (b) Affymetrix

Figure 2.9: *Smoothed histograms of the un-normalized rat data.*

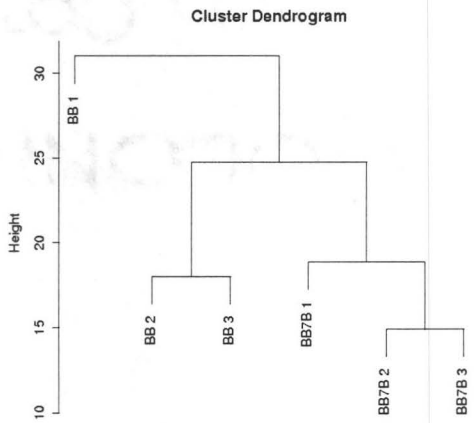
After we normalized the GeneChip data by default RMA method, the clustering is performed between the GeneChip arrays. Figure 2.7 shows the cluster dendrogram before and after excluding the problem chip of "IDD5 1", Figure 2.7(b) suggests that when excluding the chip "IDD5 1", there seems to be a good clustering between the two biological strains, although chip "NOD 4" seems quite different from the other NOD chips.

2.1.3 Illumina Rat "RatRef-12 v1" BeadChip Data

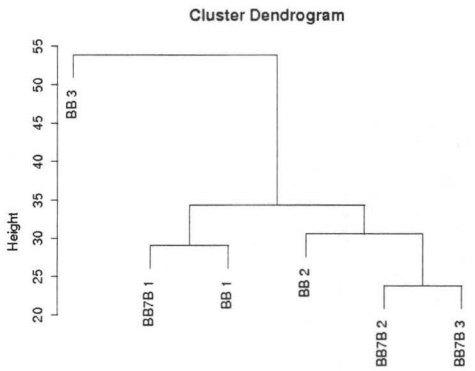
The Illumina rat data come from 12 Beadarrays on two Illumina "RatRef-12 v1" BeadChips that form the technical replicates. On each BeadChip, three biological replicates come from two conditions of either BB or BB7B strain (as described in Section 1.3.2). The rat data quality is assessed in terms of the pairwise Pearson's correlation coefficients and the smoothed histograms of the un-normalized Beadarray



(a) All 12 Illumina arrays



(b) Averaged Illumina data



(c) Affymetrix rat data

Figure 2.10: *Clustering dendrogram of the Illumina and Affymetrix rat data.*

data. The coefficients suggest that data quality is quite good as illustrated in Figure 2.9(a). The clustering dendrogram in Figure 2.10 suggests that the chip effect is relatively small compared to the biological strain effect.

2.1.4 Affymetrix Rat "RAE230v2" GeneChip Data

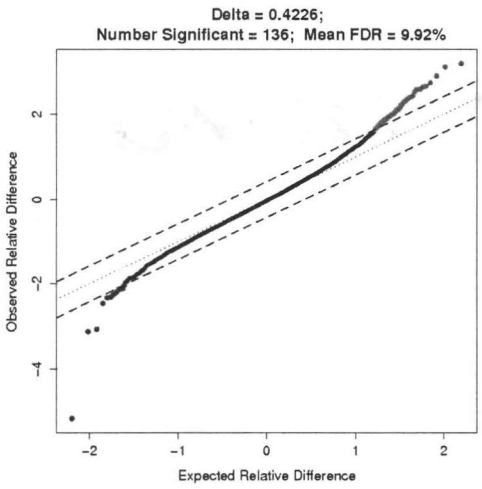
In the Affymetrix rat experiment, we used six rat "RAE230v2" GeneChips, three of which are biological replicates from the conditions of either BB or BB7B strains. Analyses show that, although it shows a relatively larger variance than the rat BeadChip data, the rat GeneChip data also have pretty good quality in terms of the Pearson's correlation coefficient (the smoothed histograms of density distributions) and the clustering dendrogram, see Figure 2.9(b) and Figure 2.10(c).

2.2 Illumina Within-platform Reproducibility through SAM Analysis

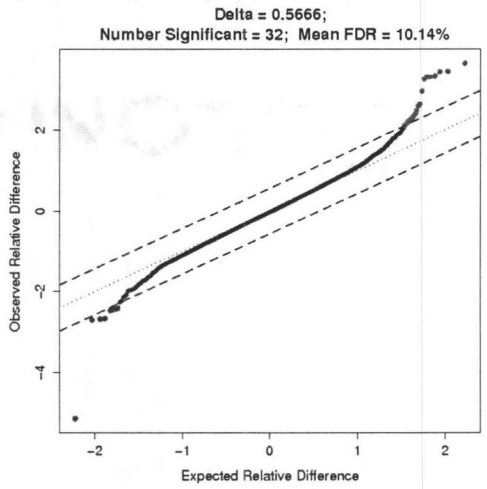
Within-platform reproducibility can be verified through comparing the SAM analyses of two technical replicate chips as described in Section 1.4.3. In this section we investigate the reproducibility within the Illumina BeadChip platform using both mouse and rat data.

2.2.1 Comparing SAM Analyses Using the Mouse Data

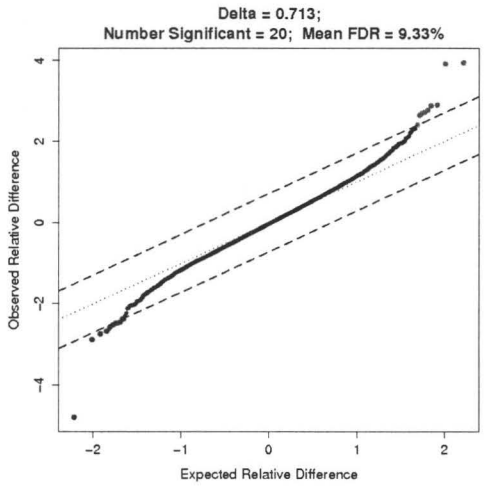
Due to the data problem in the first Illumina experiment, we decided to use the second Illumina data set, namely, the data on chip A_I and A_II. Besides, in order to use the information on both chips, we generate the "averaged" data by taking the average



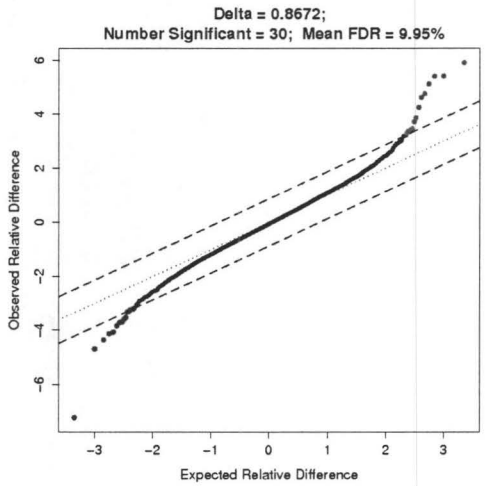
(a) Chip A



(b) Chip A_I

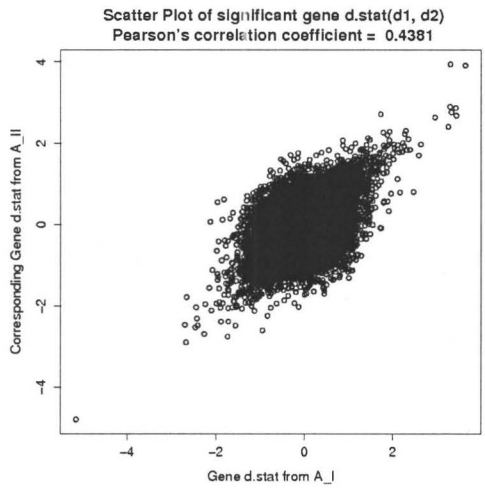


(c) Chip A_II

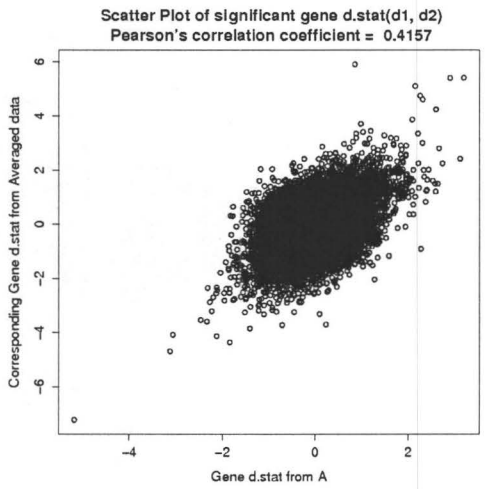


(d) "Averaged Data"

Figure 2.11: *SAM plots of chip A, A_I, A_II and the "Average" mouse data.*



(a) chip A_I v.s A_II



(b) chip A v.s Averaged data

Figure 2.12: *Pairwise scatter plots of the d-statistics from different SAM analyses of chip A_I against A_II and chip A against the averaged mouse data.*

BeadChip		**A_I**	**Comm**	**A_II**	**A**	**Comm**	**Average**
FDR	5%	11	7	8	35	9	12
	10%	32	13	20	136	15	30
	20%	64	33	63	227	37	109
	30%	118	61	142	549	65	227

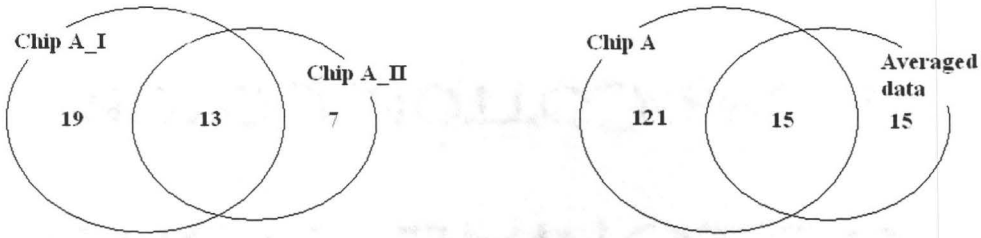
Table 2.1: *Number of significant probes from different SAM analysis of mouse BeadChip A, A_I, A_II and the averaged data at different FDR cutoff points $\simeq$ 5, 10, 20 and 30%, along with the number of those common probes (column "Comm") in both lists of chip A_I and A_II; chip A and the averaged data.*

	Gene Symbol	Probe ID	d-stat	q-value	Rank in Chip A_II
1	Il1b	scl18674.7.1_35-S	-5.1504	0.0135	1
2	Asgr2	scl41344.9.1_32-S	3.6447	0.0186	3
3	Rtn1	scl00104001.2_121-S	3.4425	0.0186	8
4	Ccl24	scl25923.3.1_0-S	3.4290	0.0186	5
5	Asb2	scl42115.9.1_121-S	3.3299	0.0186	6
6	Rnase6	scl078416.2_288-S	3.3079	0.0186	2
7	Pdk3	scl54041.11.98_19-S	3.3002	0.0186	4
8	Ifitm1	scl068713.2_9-S	3.2592	0.0186	20
9	Ifitm6	scl30503.2.1_167-S	2.9591	0.0225	9
10	Igfbp5	scl0016011.2_295-S	2.6401	0.0460	47
11	Itgb7	scl46686.16.1_15-S	2.5948	0.0504	109
12	Adam23	scl023792.26_1-S	2.4741	0.0552	4966
13	Fos	scl42959.4_58-S	-2.6969	0.0667	17
14	Cbln1	scl34520.5_191-S	-2.6738	0.0667	10
⋮	⋮	⋮	⋮	⋮	⋮
25	Dpp4	scl19224.27_390-S	2.2580	0.0784	55
26	Dcn	scl0013179.1_275-S	2.2305	0.0806	45
27	Itgae	scl0001586.1_298-S	2.2149	0.0846	86
28	AB124611	scl0382062.2_10-S	2.1693	0.0942	5681
29	Klk11	scl32730.3.1_40-S	2.1475	0.0960	87
30	Il8ra	scl16618.1.1_330-S	2.1154	0.1014	51
31	Col2a1	scl0012824.2_219-S	2.1116	0.1014	1045
32	Ifitm2	scl080876.2_10-S	2.1089	0.1014	73

Table 2.2: *Significant probes from mouse BeadChip A_I and their corresponding rankings in chip A_II when FDR $\simeq$ 10%.*

	Gene Symbol	Probe ID	d-stat	q-value	Rank in Chip A_I
1	Il1b	scl18674.7.1_35-S	-4.7825	0.0130	1
2	Rnase6	scl078416.2_288-S	3.9430	0.0173	6
3	Asgr2	scl41344.9.1_32-S	3.9134	0.0173	2
4	Pdk3	scl54041.11.98_19-S	2.9031	0.0389	7
5	Ccl24	scl25923.3.1_0-S	2.8767	0.0389	4
6	Asb2	scl42115.9.1_121-S	2.7691	0.0432	5
7	Klk8	scl32735.4_7-S	2.7217	0.0444	82
8	Rtn1	scl00104001.2_121-S	2.6850	0.0486	3
9	Ifitm6	scl30503.2.1_167-S	2.6426	0.0533	9
10	Cbln1	scl34520.5_191-S	-2.8886	0.0635	14
11	2310046K01Rik	scl20130.6_7-S	-2.7473	0.0712	107
12	Axud1	scl35215.8_496-S	-2.6846	0.0712	33
13	9030416H16Rik	scl071521.2_63-S	-2.5908	0.0762	1237
14	Dusp1	scl50147.4_283-S	-2.5207	0.0762	20
15	Plk3	scl23942.14.1_32-S	-2.4759	0.0762	192
16	Hist1h2ah	scl0319168.1_1-S	-2.4724	0.0762	23
17	Fos	scl42959.4_58-S	-2.4553	0.0762	13
18	Fpr-rs2	scl014289.2_177-S	-2.3792	0.0764	78
19	Hist2h2ac	scl00319176.1_318-S	-2.3784	0.0764	114
20	Ifitm1	scl068713.2_9-S	2.4138	0.0933	8

Table 2.3: *Significant probes from mouse BeadChip A_II and their corresponding rankings in chip A_I when FDR* $\simeq$ *10%.*



(a) Chip A_I v.s A_II

(b) Chip A v.s The averaged data

Figure 2.13: *The Venn diagram of the number of significant probes and those in common from the SAM analyses between chip A_I, A_II, A and the averaged mouse data when FDR* $\simeq$ *10%.*

of the replicates on chip A_I and A_II. SAM analyses are then performed using these data sets. This is because the SAM code only considers biological replicates. As mentioned in previous chapter, the RNA samples in second experiment are from the same preparation as those used on chip A in the first experiment. Thus we can treat these three Illumina BeadChips A, A_I and A_II as technical replicates prepared on two different days. The SAM results are compared between chip A and the averaged data in order to see the day effect. We show all the SAM plots in Figure 2.11, and Table 2.1 gives the number of significant probes at different levels of FDR. The pairwise SAM analysis comparisons are performed in the following cases:

First, we compare the SAM analysis of chip A_I against that of chip A_II. SAM plots of chip A_I and A_II are illustrated in Figure 2.11(b) and 2.11(c), the picture of the SAM plots suggest that two SAM analyses are quite similar to each other in terms of the shape, the tail behavior and the number of significant genes. For example at a 10% FDR, the SAM analysis of chip A_I gives 32 significant probes and the number is

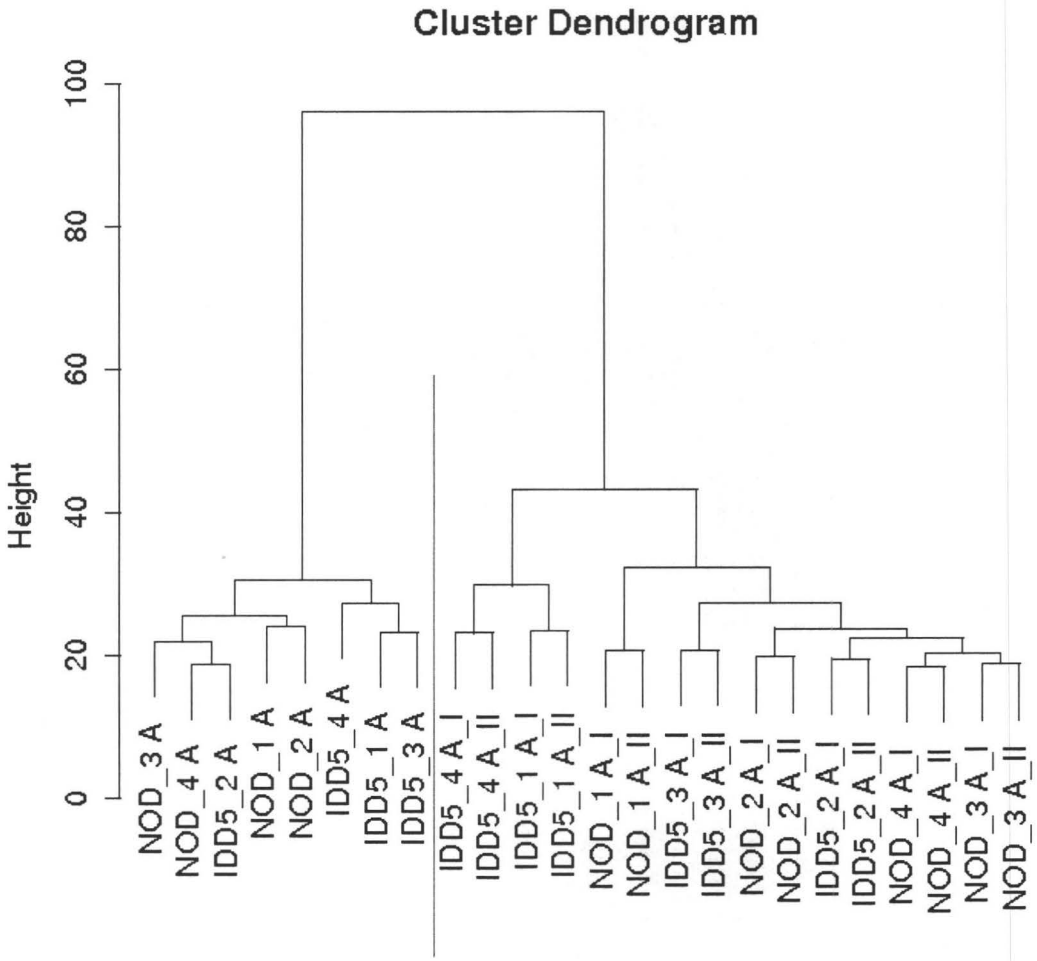


Figure 2.14: *Cluster dendrogram of all 24 Beadarrays on three technical replicate BeadChips (chip A from first experiment, A_I and A_II from second experiment). The line indicates and distinguishes the biggest cluster of the arrays from the first and second mouse experiment.*

64 if we set the FDR equal to 20%. And the corresponding numbers are 20 and 63 from the SAM analysis of chip A_II (see Table 2.1 for the comparisons at different FDR cutoff points). Using a 10% FDR cutoff, we listed the significant probes from chip A_I and their corresponding rankings in chip A_II in Table 2.2. From the table, we find that excepting a few probes with some extreme disagreements in their rankings between the two analyses (Itgb7, Adam23, AB124611 and Col2a1 ranked at 11th, 12th, 28th and 31st in chip A_I), most probes are in reasonably good agreement with their rankings in chip A_II. The comparison shows that the rankings in the top 10 of both analysis reach a particularly good agreement with only two exceptions (Ifitm1 and Igfbp5). However, the overall Pearson's correlation coefficient of the d-statistics of the two SAM analyses is only 0.4381 as showed in Figure 2.12(a). The reason for this low correlation coefficient could be the large amount of noise in the data. Under the assumption that most probes are not differentially expressed, these probes form the points in the middle of the plot that make the plot looks "fat".

Secondly, at the same FDR $\simeq$ 10%, we compare the SAM analyses of chip A_II against that of chip A_I. In addition to the SAM plots in Figure 2.11, Table 2.3 lists the 20 significant probes from the SAM analysis of chip A_II and their corresponding rankings in chip A_I. From this table, we find that excepting a few probes (9030416H16Rik, Plk3, Hist2h2ac, Klk8, 2310046K01Rik and maybe Fpr-rs2), a similar conclusion can be reached that the overall agreement of the two significant probe ranking lists is pretty good and this is especially true for the top 10 significant probes.

As a summary, Table 2.1 lists the numbers of significant probes from different SAM analyses of chip A, A_I, A_II and the averaged data using different FDR cutoff points $\simeq$ 5, 10, 20 and 30%. In middle column of "Comm", where it gives the number of those probes that are common in both lists when comparing chip A_I and A_II as

well as chip A and the averaged data. From the table we see that, for example, at FDR $\simeq$ 10%, 32 and 20 probes are called significant from the SAM analysis of chip A_I and chip A_II accordingly. Among these two significant probe lists, 13 probes are found in common (appeared in both lists) which gives the percentage of 41% and 65% for corresponding list if ignoring their rankings. Furthermore, from the Table 2.2 and 2.3 we find that 8 out of the 13 common significant probes belong to the top 10 significant probes in both lists.

In addition to the numbers outlined in Table 2.1, the Venn diagram in Figure 2.13 shows the relationship of the numbers of significant probes and of those common genes in both lists at 10% FDR.

Illumina Day Effect Although in this report we are not interested in evaluating the day effect and we didn't formally test the significance of the day effect. We do see some day effect from the cluster analysis and the comparison of SAM analyses.

As illustrated in Table 2.1, the day effect is revealed from the difference in the "number of significant probes" from the SAM analyses (136 of chip A v.s 32, 20, 30 of others at the 10% FDR). Furthermore, day effect is verified after comparing the significant gene lists from the SAM analyses of the chip A and the averaged data (not included in this report), we find that the agreement between the two SAM analyses is quite poor in terms of the significant probe rankings, and the overall Pearson's correlation coefficient of the d-statistics from the two SAM analyses is 0.4157 as shown in Figure 2.12(b). Lastly, when we put together and compare the cluster analysis of all three data set of chip A, A_I and A_II, the cluster dendrogram in Figure 2.14 clearly shows the day difference between the first and second experiments.

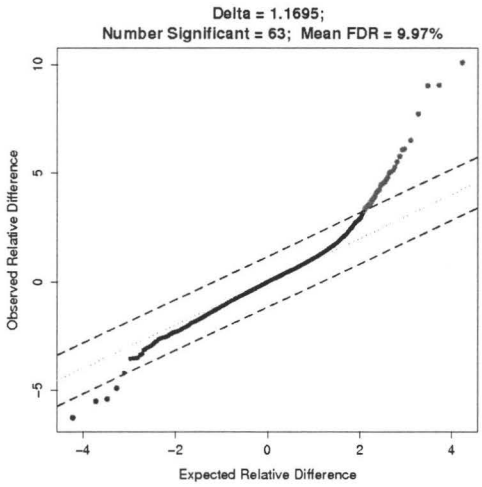
BeadChip		I	Comm	II	Average
FDR	5%	8	1	1	8
	10%	61	48	66	57
	20%	152	100	141	127
	30%	224	141	209	267

Table 2.4: *Number of significant probes from different SAM analysis of rat BeadChip I, II and the averaged data at different FDR cutoff points $\simeq$ 5, 10, 20 and 30%, along with the number of those common probes (column "Comm") in both lists of chip I and II.*

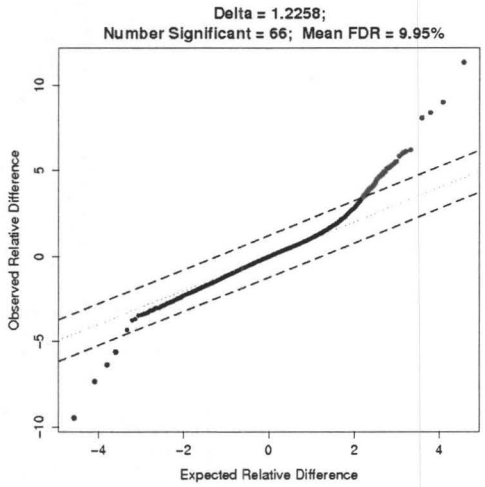
2.2.2 Comparing SAM Analyses Using the Rat Data

To check the reproducibility within the Illumina rat BeadChips, we compare the SAM analyses of rat chip I and II that form the technical replicates.

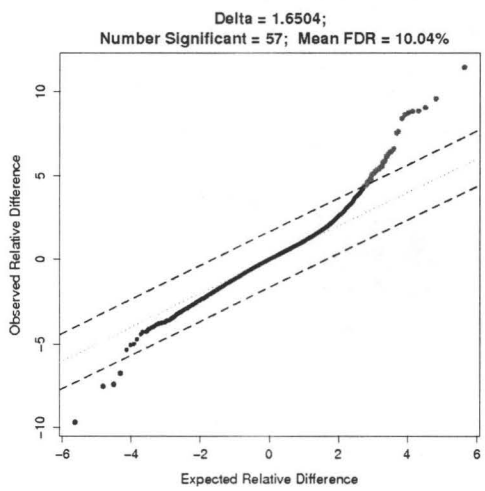
First, we compare the SAM analysis of rat chip I against that of chip II. SAM plots of chip I and II are illustrated in Figure 2.15 using a 10% FDR, the picture of the SAM plots suggest that two SAM analyses are quite similar to each other in terms of the shape, the tail behavior and the number of significant genes. For example, the SAM analysis of chip I gives out 63 significant probes and the number is 152 if we set the FDR equal to 20%. And the corresponding numbers are 66 and 141 from the SAM analysis of chip II (see Table 2.4 for the comparisons at different FDR cutoff points). We examine the top 20 significant probes in both significant lists from chip I and chip II along with their corresponding rankings in the other list in Table 2.5 and Table 2.6. Table 2.5 suggests that the top 20 significant probes reach a pretty good agreement between the two chips. We also find that, as with the mouse data, the rankings in the



(a) Chip I



(b) Chip II



(c) Averaged data

Figure 2.15: *SAM plots of chip I, II and the "Average" rat data.*

Rank in Chip I	Gene Symbol	d-stat	q-value	Rank in Chip II
1	Defa	10.0901	0.0465	1
2	Np4	9.0537	0.0465	5
3	LOC498659	9.0133	0.0465	2
4	Ms4a3_predicted	7.7194	0.0465	4
5	Nradd	6.501	0.0465	7
6	LOC310395	6.1	0.0524	17
7	MGC93766	6.0689	0.0524	10
8	RGD1307811	5.7718	0.0524	19
9	Tmepai_predicted	5.5066	0.0621	18
10	Ldhc	5.27	0.0632	23
11	Ugt8	5.1098	0.0632	16
12	LOC502819	5.052	0.0632	9
13	LOC502903	5.0224	0.0632	25
14	Ctsg_predicted	5.0063	0.0632	24
15	Cklfsf8	4.7828	0.0652	40
16	Nkg7	4.6947	0.0657	30
17	Galnt3_predicted	4.6327	0.0657	21
18	Cdkn3_predicted	4.5745	0.0686	15
19	Orm1	4.5647	0.0686	28
20	Ccr1	-6.2546	0.0756	29

Table 2.5: *The top 20 significant probes from rat BeadChip I and their corresponding rankings in chip II.*

Rank in Chip II	Gene Symbol	d-stat	q-value	Rank in Chip I
1	Defa	11.3300	0.0469	1
2	LOC498659	8.9988	0.0657	3
3	LOC363158	-9.4466	0.0657	26
4	Ms4a3_predicted	8.3797	0.0657	4
5	Np4	8.0582	0.0657	2
6	Snx14_predicted	-7.3378	0.0670	25
7	Nradd	6.2011	0.0670	5
8	RGD1311259_predicted	6.1032	0.0670	301
9	LOC502819	6.0157	0.0670	12
10	MGC93766	5.8558	0.0670	7
11	Cspg5	-6.3740	0.0670	64
12	Coro2a_predicted	5.5246	0.0670	39
13	Gca_predicted	5.4249	0.0670	22
14	LOC299354	5.3161	0.0670	23
15	Cdkn3_predicted	5.2707	0.0670	18
16	Ugt8	5.2058	0.0670	11
17	LOC310395	5.1545	0.0670	6
18	Tmepai_predicted	5.1339	0.0670	9
19	RGD1307811	5.1058	0.0670	8
20	Gfi1	4.9502	0.0670	24

Table 2.6: *The top 20 significant probes from rat BeadChip II and their corresponding rankings in chip I.*

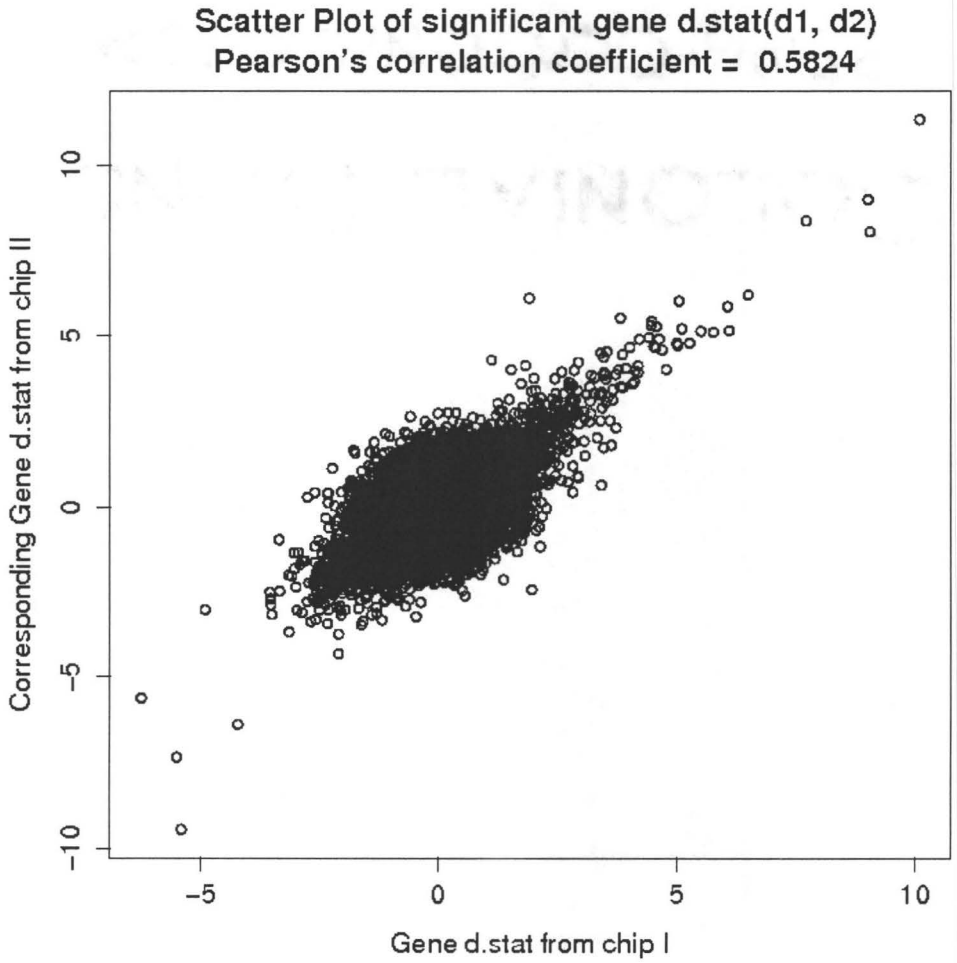
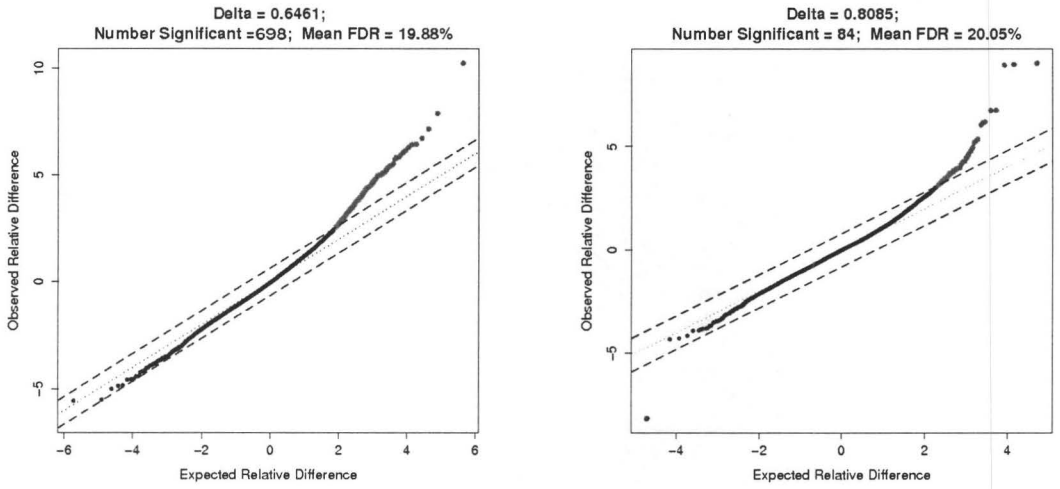


Figure 2.16: *Pairwise scatter plots of the d-statistics from different SAM analyses of chip I against II of the rat data.*



(a) Mouse platform

(b) Rat platform

Figure 2.17: *SAM plots using both mouse and rat Affymetrix GeneChip data.*

top 10 of both analyses reach a particularly good agreement. However, the overall Pearson's correlation coefficient of the d-statistics of the two SAM analyses is only 0.5824 as shown in Figure 2.16. Although it is not highly correlated because of the large amount of noise in the data, compare to the mouse data of chip A_I and A_II (0.4381), this correlation is much higher.

Comparing the SAM analyses of Chip II against Chip I, excepting one probe with some extreme disagreement in its ranking between the two analyses (RGD1311259 predicted, ranked at 8th in Chip II but 301st in Chip I, see Table 2.6), a similar conclusion can be made that the overall agreement of the two significant probe ranking lists is pretty good.

	Gene Symbol	d-stat	q-value	Gene Name
1	Speg	10.2462	0.0242	SPEG complex locus
2	Zfp64	7.8966	0.0363	zinc finger protein 64
3	Psmd11	7.1489	0.0565	proteasome (prosome, macropain) 26S subunit, non-ATPase, 11
4	Hmox1	6.7240	0.0588	heme oxygenase (decycling) 1
5	C5ar1	6.4464	0.0588	complement component 5a receptor 1
6	Ifitm1	6.4224	0.0588	interferon induced transmembrane protein 1
7	Commd3	6.2963	0.0588	COMM domain containing 3
8	Ywhaz	6.1414	0.0645	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide
9	Ltb	6.0617	0.0645	lymphotoxin B
10	Gna11	5.9440	0.0670	guanine nucleotide binding protein, alpha 11
11	Marcks	5.8497	0.0670	myristoylated alanine rich protein kinase C substrate
12	Tsc22d3	5.8306	0.0670	TSC22 domain family 3
13	Pfn1	5.8224	0.0670	profilin 1
14	Apba3	5.7410	0.0674	amyloid beta (A4) precursor protein-binding, family A, member 3
15	Ctnnd2	5.5282	0.0719	catenin (cadherin associated protein), delta 2
16	Ddost	5.4648	0.0719	dolichyl-di-phosphooligosaccharide-protein glycotransferase
17	Omt2a	5.4403	0.0719	oocyte maturation, alpha
18	2610510E02Rik	5.3930	0.0719	RIKEN cDNA 2610510E02 gene
19	Mgst3	5.3833	0.0719	microsomal glutathione S-transferase 3
20	Smpd4	5.3385	0.0719	sphingomyelin phosphodiesterase 4

Table 2.7: *The top 20 Significant probe sets (genes) from the mouse Affymetrix GeneChip data.*

	Gene Symbol	d-stat	q-value	Gene Name
1	Defa	9.0358	0.0480	defensin, alpha 5, Paneth cell-specific
2	Np4	8.9842	0.0480	defensin NP-4 precursor
3	RatNP-3b	8.9473	0.0480	NA
4	Fam31b_predicted	-8.1344	0.0680	family with sequence similarity 31, member B (predicted)
5	RGD1560913_predicted	6.7699	0.0680	similar to expressed sequence AW413625 (predicted)
6	Mthfs	6.7569	0.0680	5,10-methenyltetrahydrofolate synthetase (5-formyltetrahydrofolate cyclo-ligase)
7	IgG-2a	6.1957	0.0680	gamma-2a immunoglobulin heavy chain
8	Anxa1	6.1589	0.0680	annexin A1
9	Nck1_predicted	6.0578	0.0680	non-catalytic region of tyrosine kinase adaptor protein 1 (predicted)
10	Tln1	5.3661	0.0680	talin 1
11	Nt5e	5.2715	0.0680	5' nucleotidase, ecto
12	Rnpep	5.2130	0.0680	arginyl aminopeptidase (aminopeptidase B)
13	Nt5e	4.9527	0.0812	5' nucleotidase, ecto
14	NA	4.7707	0.0822	NA
15	Pacsin1	4.6739	0.0900	protein kinase C and casein kinase substrate in neurons 1
16	Fut2	4.5625	0.0900	fucosyltransferase 2 (secretor status included)
17	Tf	4.4532	0.0933	transferrin
18	Defa7	4.4369	0.0933	defensin alpha 7
19	NA	4.2847	0.1028	NA
20	NA	4.2800	0.1028	NA

Table 2.8: *The top 20 Significant probe sets (genes) from the rat Affymetrix GeneChip data.*

2.3 SAM Analysis of Affymetrix Platforms

Due to the lack of technical replicates in the data set, the chip effect within the Affymetrix GeneChip platform cannot be investigated through comparing two replicate SAM analyses. In the rest of this section, we perform the SAM analysis using the GeneChip data of both mouse and rat data and give out the SAM results as following:

2.3.1 Mouse platform

Based on the results we find in section 2.1.2 and comparing the sensitivity analysis by running the SAM analysis with and without that "problem" chip ("Idd5 1"), we decided to abandon the "problem" chip that caused the inconsistency in the histogram and the boxplots, But the reason for this inconsistency is not clear yet.

The SAM analysis is then performed using the seven mouse GeneChip data (four from NOD and three from NOD.NOR-Idd5). The RMA preprocessing method with the quantile normalization is applied. Figure 2.17(a) shows the SAM plot at FDR cutoff points around 20%, which results in 99 probe sets called significant. Table 2.7 shows only the top 20 significant probe sets in the list.

2.3.2 Rat platform

The rat SAM analysis is performed using the six GeneChips of BB and BB7B rat data. The RMA preprocessing method with the quantile normalization is applied. Figure 2.17(b) shows the SAM plot at 20% FDR cutoff point which results in 84 probe sets called significant. Table 2.8 shows only the top 20 significant probe sets along with their d-statistics and gene descriptions.

2.4 Summary

In this chapter, we analyzed both the mouse and the rat data from both Illumina BeadChip and Affymetrix GeneChip platforms. Except for one problem GeneChip in the Affymetrix mouse platform, the data quality is quite good for both platforms in terms of the correlation coefficients and other diagnostic plots.

The chip reproducibility is pretty good within the Illumina BeadChip platform in terms of the correlation coefficients, the cluster analysis and the SAM analyses. The agreement in rankings of the significant probes is particularly good for those top differentially expressed genes. However, the within platform reproducibility of Affymetrix platform cannot be observed through SAM analysis due to the lack of technical replicates.

The day effect does inevitably exist in Illumina BeadChip platform as it does in the Affymetrix GeneChip platform. In addition to the day effect, RNA preparation artifact was found in our mouse model experiment. In order to take advantage of the Illumina BeadChip platform and other microarray platforms as well, caution must be advised in the procedure of the RNA sample preparation.

In next chapter, we will compare the results from the two platforms.

Chapter 3

Across Platform Reproducibility Using the Mouse Data

In this chapter we check the reproducibility of the Affymetrix GeneChip and Illumina BeadChip platforms through the SAM analyses using the mouse data from both platforms. At the end of the chapter another paper comparing these two platforms is discussed and compared to our study, and some conclusions are made.

3.1 Difficulties of the Comparison

The primary questions of interest in comparing different microarray platforms are: How do they perform in detecting the differentially expressed genes in the same experimental setting? Do they detect a similar pattern of genes, especially among the top significant genes on both platforms?

Under the assumption that most genes are not differentially expressed between our two mouse RNA samples, these genes are producing only noise in the gene ranking lists

and so we are not interested in comparing the concordance among these genes. We only compare those differentially expressed genes, in other words, the top significant genes from the SAM analysis.

A number of problems arise in comparing different microarray platforms. One big challenge is the probe annotation. Unlike the within-platform situation where we are comparing two ranking lists within the same annotation space, now we are comparing products of two different manufacturers that use different probe designs and different probe annotations. Namely, we are comparing two lists of (significant) genes of different length and partially different names. For example for the mouse platforms, as shown in Table 1.1, Affymetrix "Mouse430v2.0" GeneChip has 45101 probe sets in total, which annotated to 20958 unique genes. While in the Illumina "MouseRef-8 v1" BeadChip platform, the annotation package gives 24049 probes that mapped to 18241 unique genes. Fortunately, among those annotated genes, the two platforms have 13234 genes in total that are mutually comparable (have the same gene symbol and functional description) to each other. The coverage of these common genes are reasonably good at 63.15% for Affymetrix and 72.55% for Illumina. In addition, there are 7724 Affymetrix genes not presented in the Illumina platform, which means these genes only make sense for the Affymetrix platform but not for the Illumina platform. Similarly, there are 5007 annotated genes on the Illumina platform that are not available on the Affymetrix platform. Figure 3.1 shows the Venn diagram of these comparable and non-comparable genes in the two platforms. Obviously, we cannot compare these genes that are unique to only one platform. All we can do is to make comparisons within the subset of the 13234 common genes (Subset B in the Venn diagram).

In all, the cross-platform comparison becomes much more complicated because of

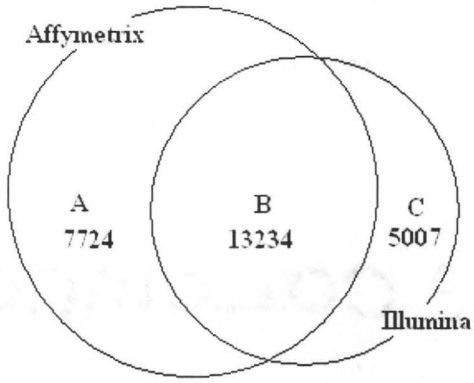
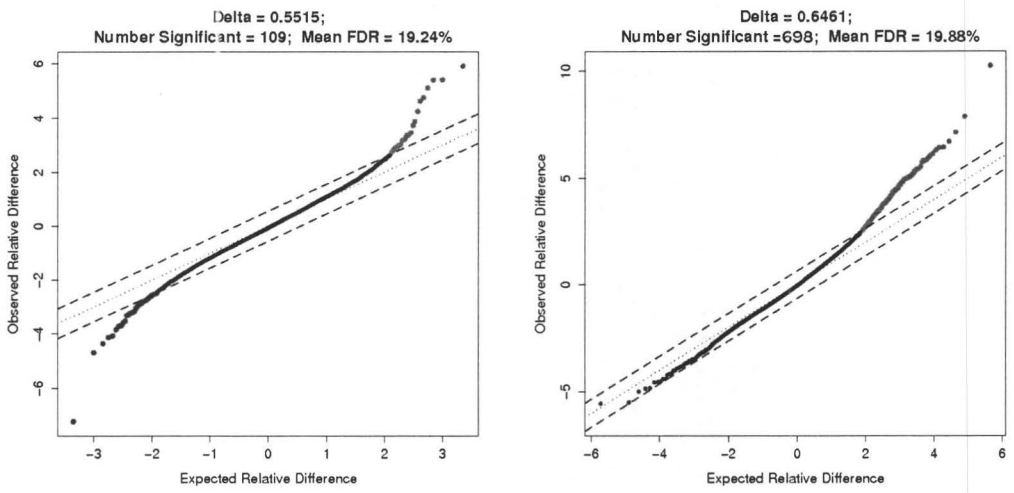


Figure 3.1: *Venn diagram of the number of annotated genes in "Mouse430v2.0" GeneChip and Illumina "MouseRef-8 v1" BeadChip. B denotes the number of common genes for both platforms.*

the heterogeneity in the annotation spaces caused by different manufacturers. One possible solution is to find such a third-party linkage between these two annotation spaces that we can make an unique projection for all probes from one space to the other. But unfortunately, there is not such a linkage available at least for now. Another compromised approach is to compare the concordance only within the subset of the 13234 common genes with a hope that most significant genes will fall in this subset. That is, we ignore those genes that are only available in one platform even though they are significant. We used gene symbol as the linkage between two annotations, which seems quite feasible as there are quite a large number of common genes with the same gene symbols for both platforms. Besides the gene symbol, another annotation mapping file provide by Illumina company is also used. It maps the Illumina probeIDs to the Affymetrix probeIDs for all possible annotated probes between these two platforms. The analyses using these two linkage IDs produced very similar



(a) Illumina

(b) Affymetrix

Figure 3.2: *SAM plot comparison between the Illumina and Affymetrix analysis using the mouse data.*

results. The results using the mapping file are showed in following sections.

3.2 Comparing the SAM Analyses

We chose the SAM analysis using the averaged mouse data (as described in last chapter) for the Illumina platform, and compare its result with the SAM analysis using the Affymetrix GeneChip mouse data. Both analyses are compared at different levels of FDR cutoff points roughly equal to 5%, 10% and 20%. Figure 3.2 only gives the SAM plots at the FDR $\simeq$ 20%.

3.2.1 Comparing the SAM Plots and Boxplots

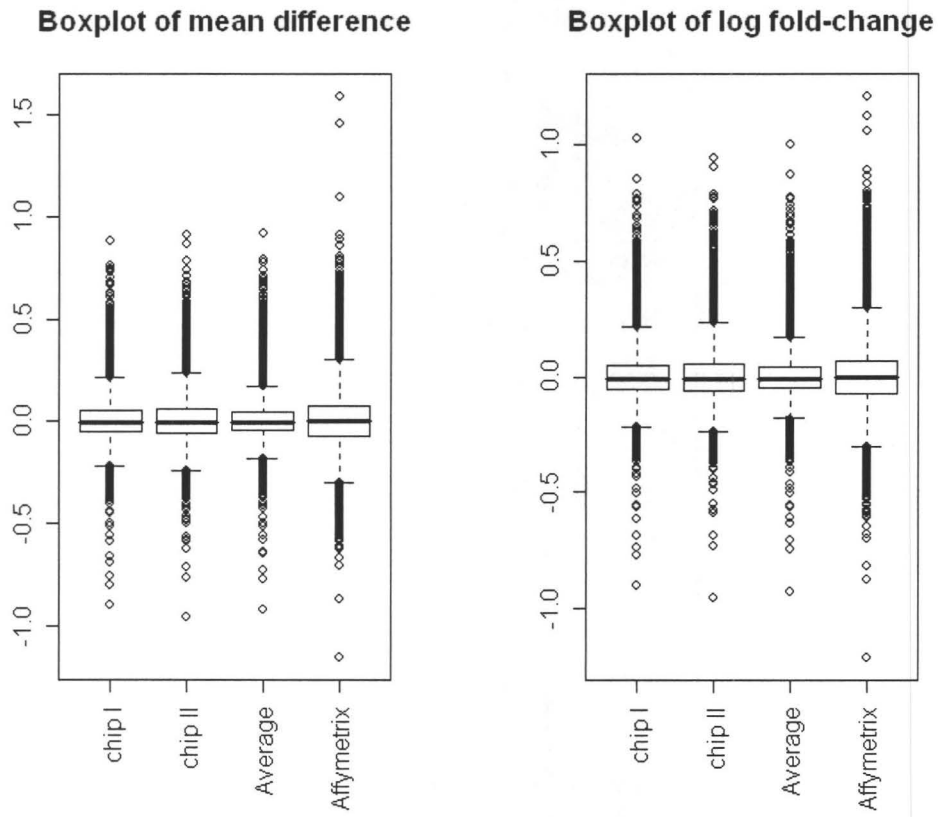
Comparing the two SAM plots in Figure 3.2, we find that they differ in terms of the scale of the axes, the number of significant probes or probe sets and the tail behaviors as well. From the Illumina SAM plot, we surprisingly find that Affymetrix analysis gives no "under-expressed" genes (in the lower-left tail) as all the d-statistics are positive (refer to Table 2.7), while there are both "under-expressed" and "over-expressed" genes (in the upper-right tail) from the Illumina analysis. Notice that the terms of "under-expressed" and "over-expressed" are referred to the gene expression behavior comparing the congenic (Idd5) mouse against the parental (NOD) mouse strain.

The boxplots in Figure 3.3 show that the mean difference of the gene expression value on the Illumina platform has a smaller variance than the data on the Affymetrix platform, and it is also true for the log fold-change variances.

3.2.2 Comparing the (Comparable) Significant Genes

As stated before, we only investigate the concordance of those comparable genes across the two platforms. The comparable genes are subsetted based on the Illumina's annotation mapping file that maps the Illumina probeIDs on the "MouseRef-6 v1" BeadChip to the Affymetrix probeIDs on "Mouse430v2.0" GeneChip. Since the probes (24049 in total) used on "MouseRef-8 v1" BeadChip form a subset of those probes used on "MouseRef-6 v1" BeadChip platform, the mapping file (with a total of 20175 linkages) between the annotation of "MouseRef-6 v1" and "Mouse430v2.0" will cover all the linkages between "MouseRef-8 v1" and "Mouse430v2.0" platforms.

Instead of giving the d-statistic plot for the whole genome of either platform,



(a) Mean difference

(b) Log fold-change

Figure 3.3: *Boxplots for the mean difference expression and the log fold-change comparison between the Illumina and Affymetrix analysis using the mouse data. Note: Average data is the averaged Illumina data as described in Section 2.2.1.*

Figure 3.4 gives the d-statistic plot and the Pearson's correlation coefficient of these comparable genes in two platforms. From which we find that the Pearson's correlation coefficient is 0.1464, which is very poor. The shape of this plot does not support an assumption of good concordance between these two platforms.

The SAM analyses give two significant gene lists from either platform. For example in the Illumina list, we find that there are 109 significant probes in total at FDR $\simeq$ 20%. According to the mapping file, 70 probes (63.3%) among the total whose probeIDs are linked to certain probeIDs in the Affymetrix platform, which means these 70 probes are comparable and the other 39 (36.7%) probes are non-comparable across the two platforms. On the other hand in the Affymetrix list, at the same FDR cutoff point, the SAM analysis gives 698 significant probe sets in the significant list, of which 277 (39.7%) are comparable and 421 (60.3%) are non-comparable across the platform.

From the significant probe list tables (refer to Table 3.2 and Table 3.3) we find that, for example, the first comparable significant gene in the Illumina list is "Asgr2" (3rd is the original overall ranking in the Illumina platform if including those non-comparable genes), the second is "Pdk3" with an overall ranking as the 5th, and the third is "Asb2" with an overall ranking as the 6th in the list. The corresponding ranking for these top 3 comparable significant genes in the Affymetrix list are the 78th, 28th and 31st (the original overall rankings). Furthermore, we find 19 of such comparable genes that are in both significant lists at the FDR level of 20%. This forms the percentage of 27.14% for the Illumina list and 6.86% for the Affymetrix list for these common significant genes correspondingly. We summarize the percentage statistics of these comparable and common significant genes in Table 3.1, along with the Venn diagrams to show the agreement between these two significant lists, at three

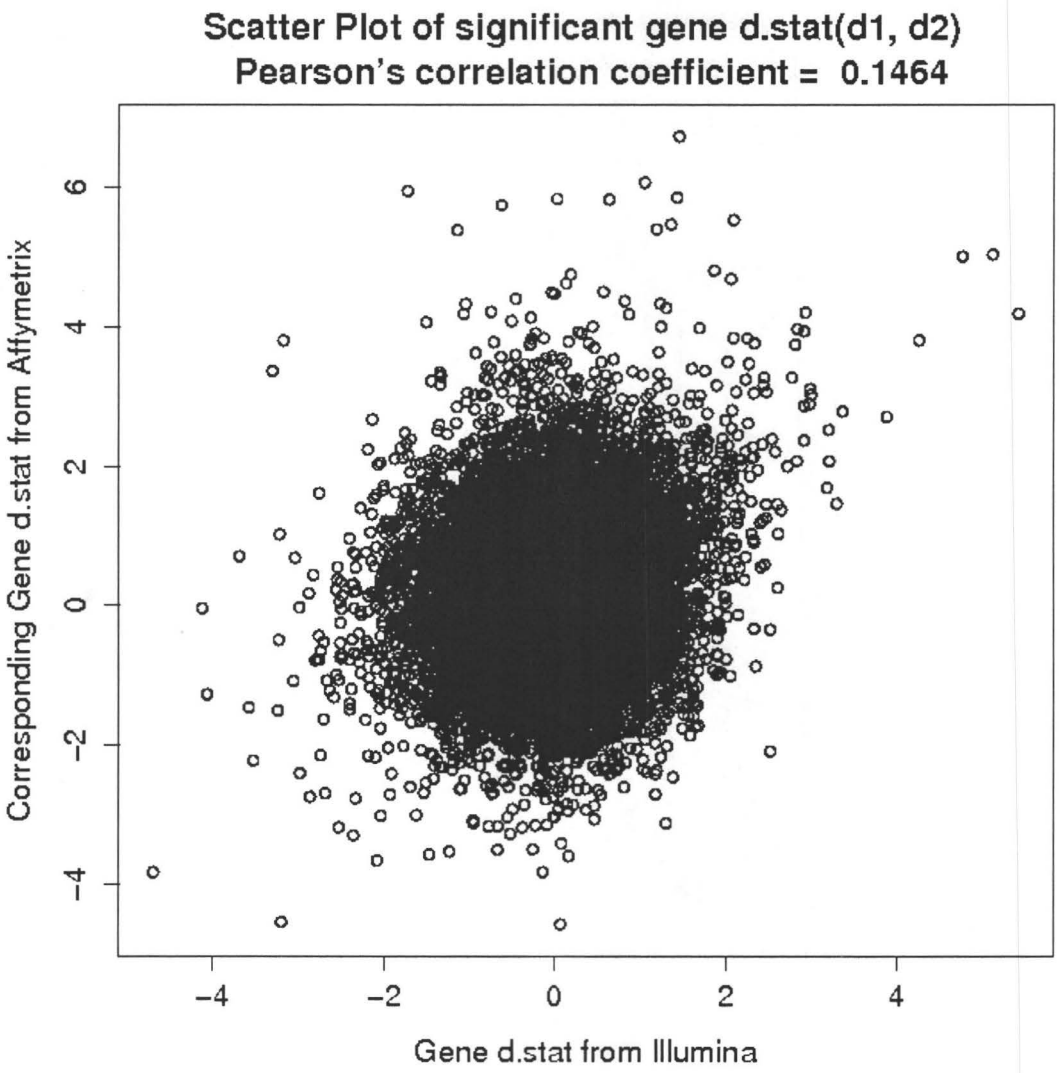


Figure 3.4: *The pairwise scatter plots of the d-statistics of those comparable genes from two mouse microarray platforms and the Pearson's correlation coefficient.*

different FDR cutoff points of 20%, 10% and 5%.

Table 3.2 shows the comparable top 40 significant genes from the Illumina platform and their corresponding ranking and d-statistics in the Affymetrix platform. From which we find that the overall concordance between the two platform is very poor. Furthermore, we surprisingly find that genes Zfp148 and Supt16h, ranked at 26th and 35th on Illumina, showed completely opposite expression behavior between the two platforms, both are significantly under-expressed (at 20% FDR) on the Illumina platform but become over-expressed on the Affymetrix platform.

In addition to the above finding, however, only looking at the top 10 significant genes from the Illumina list (Table 3.2) suggests a quite good agreement in rankings, four out of five (80%) comparable genes are in both lists, which is pretty good compared to the overall percentage of 27.14% at a 20% FDR and 16.7% at a 10% FDR (Table 3.1). This tells us that, when mapping from the Illumina analysis to that for the Affymetrix data, the concordance of the significant genes across the two platforms is significantly higher for those extremely differentially expressed genes (top significant genes) than the average. However, this finding is not true if we are mapping from the Affymetrix analysis to that for the Illumina data. We find that none of the comparable top 10 genes in the Affymetrix list shows up in the Illumina list (Table 3.3). And the overall percentages of those common significant genes are 6.86% at a 20% FDR and 9.38% at a 10% FDR, which are all much lower than those from the Illumina platform (Table 3.1).

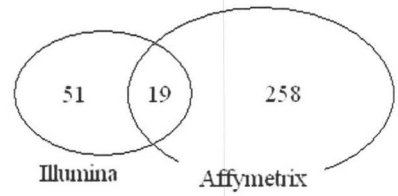
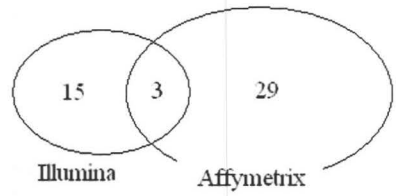
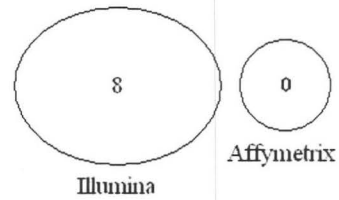
$FDR \simeq 20\%$			
	Illumina	Affymetrix	
Total number of significant genes	109	698	
Comparable genes	70 (63.3%)	277 (39.7%)	
Non-comparable genes	39 (36.7%)	421 (60.3%)	
19 genes in both lists	27.14%	6.86%	
$FDR \simeq 10\%$			
	Illumina	Affymetrix	
Total number of significant genes	30	91	
Comparable genes	18 (60%)	32 (35.2%)	
Non-comparable genes	12 (40%)	59 (64.8%)	
3 genes in both lists	16.7%	9.38%	
$FDR \simeq 5\%$			
	Illumina	Affymetrix	
Total number of significant genes	13	2	
Comparable genes	8 (61.5%)	0 (0%)	
Non-comparable genes	5 (38.5%)	2 (100%)	
No genes in both lists	0		

Table 3.1: *Summary of concordance of the significant genes from two mouse platforms at different* $FDRs \simeq 20\%, 10\%$ *and* 5%.

ILMN Rank	ILMN Symbol	d.stat1	Affy Symbol	d.stat2	Affy Rank
3	Asgr2	5.4133	Asgr2	4.1814	78
5	Pdk3	5.1086	Pdk3	5.0290	28
6	Asb2	4.7553	Asb2	5.0027	31
8	Axud1	-4.6899	Axud1	-3.8237	1136
9	Ccl24	4.2490	Ccl24	3.7986	122
11	Nfil3	-4.1250	Nfil3	-0.0407	43763
12	Cbln1	-4.0688	Cbln1	-1.2688	13249
13	Ifitm6	3.8747	Ifitm6	2.7051	525
17	Mrpl55	-3.6922	Mrpl55	0.7110	26530
18	Dusp1	-3.5801	Dusp1	-1.4542	11323
19	Fpr-rs2	-3.5255	Fpr-rs2	-2.2196	5835
22	Itgb7	3.3568	Itgb7	2.7823	475
24	4632408A20Rik	3.2909	4632408A20Rik	1.4602	4234
26	Supt16h	-3.3031	Supt16h	3.3647	216
27	2310046K01Rik	-3.2404	2310046K01Rik	-1.4984	10991
28	Trps1	-3.2296	Trps1	-0.4951	26666
29	Card4	-3.2178	Card4	1.0254	10032
30	Fos	-3.1981	Fos	-4.5395	801
31	Dscam	3.2028	Dscam	2.0667	1499
32	Il18r1	3.1967	Il18r1	2.5179	682
33	Npy	3.1775	Npy	1.6912	2866
35	Zfp148	-3.1764	Zfp148	3.7954	124
37	D16Ertd472e	-3.0588	D16Ertd472e	-1.0797	15563
39	Plk3	-3.0393	Plk3	0.6902	27663

Table 3.2: *The table of comparable significant genes from the Illumina platform and the corresponding ranking in the Affymetrix mouse platform.*

Affy Rank	Affy Symbol	d.stat1	ILMN Symbol	d.stat2	ILMN Rank
4	Hmox1	6.7240	Hmox1	1.4485	3417
9	Ltb	6.0617	Ltb	1.0483	10316
10	G-11	5.9440	G-11	-1.7239	532
11	Marcks	5.8497	Marcks	1.4241	4279
12	Tsc22d3	5.8306	Dsip1	0.0245	23375
13	Pfn1	5.8224	Pfn1	0.6335	16014
14	Apba3	5.7410	Apba3	-0.6253	5966
15	Ctnnd2	5.5282	Catnd2	2.0893	446
16	Ddost	5.4648	Ddost	1.3549	5082
18	2610510E02Rik	5.3930	1700025G04Rik	1.1861	7805
19	Mgst3	5.3833	Mgst3	-1.1420	2046
28	Pdk3	5.0290	Pdk3	5.1086	5
31	Asb2	5.0027	Asb2	4.7553	6
39	Por	4.7989	Por	1.8608	804

Table 3.3: *The table of comparable significant genes from the Affymetrix platform and the corresponding ranking in the Illumina mouse platform.*

3.3 Compare to Another Paper

In another paper by Barnes *et al.* (2005), they compared reproducibility of the Affymetrix and the Illumina platforms based on a series of dilution studies using the human microarray products. It was found that the agreement between these two platforms is very high, especially for those genes that are predicted to be differentially expressed. It was shown that, firstly, the level of gene expression is an important factor in making a good cross-platform comparison and within-platform comparison as well, secondly, the precise location of the probe on the genome also plays an important role when comparing the two platforms.

Compare the findings from our study to the above paper, although the two studies used different experimental designs and different microarray products and approaches of statistical analysis, our studies more focused on the SAM perspective. We arrived at similar findings to a certain degree; both studies suggest a good within-platform reproducibility on the Illumina platform. Besides, our study partially verified the finding that the level of gene expression plays an important role during the within- and cross-platform comparison. We say partially because what we find is only true for the Illumina platform but not for the Affymetrix according to the mapping file. Plus, as pointed in Barnes' paper, we believe that the precision of the annotation mapping file has a huge impact on the cross-platform comparison. We hope that the mapping file does map the two different probeIDs to the same location on the genome, so we know that they are measuring the same thing on both platforms.

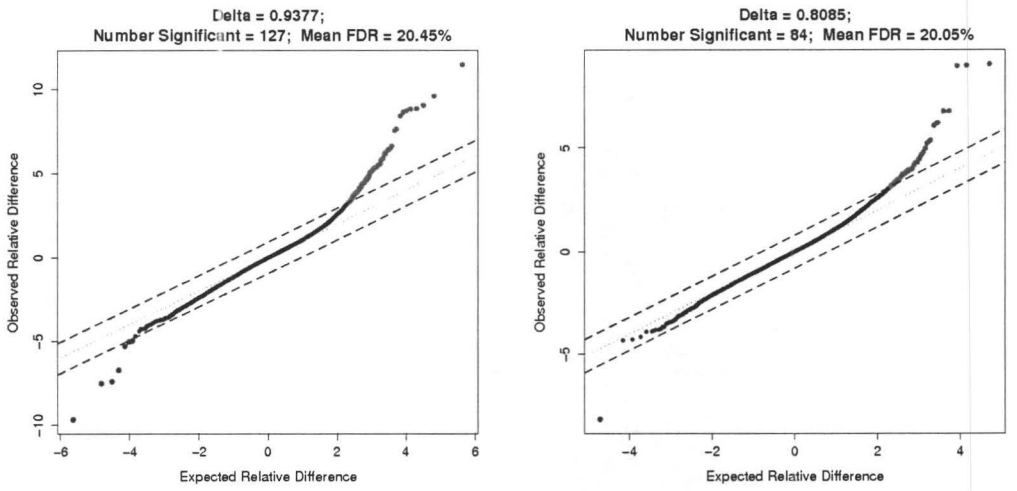
Chapter 4

Across Platform Reproducibility Using the Rat Data

In this chapter we check the reproducibility across the Affymetrix GeneChip ("RAE230v2") and Illumina BeadChip ("RatRef-12 v1") platforms through the SAM analyses using another experiment in the rat model. The rat data is described in Section 1.3.2. We investigate the concordance of those comparable significant rat genes arising from the SAM analyses of both platforms. The result is also compared to that of the mouse data from the previous chapter.

4.1 Comparing the SAM Analyses

As we did in the mouse chapter, we produced the "averaged Illumina rat data" by averaging the expression values from the technical replicates. We performed the SAM analysis and compare its result with the SAM analysis using the Affymetrix GeneChip rat data. Quantile normalization methods for both platforms are used in the same



(a) Illumina

(b) Affymetrix

Figure 4.1: *SAM plot comparison between Illumina and Affymetrix analysis using the rat data.*

way as we did for the mouse data. Figure 4.1 illustrates both SAM plots at the FDR $\simeq$ 20%. We will compare the SAM analyses of both platforms at different FDR cutoff points at 5%, 10% and 20% in the following sections.

4.1.1 Comparing the SAM Plots and Boxplots

Comparing the two SAM plots in Figure 4.1 (FDR $\simeq$ 20%), we find that they differ slightly in terms of the number of significant probes or probe sets, but there is quite a big difference in the tail behavior. Compared to the Illumina SAM plot, we find that Affymetrix analysis gives fewer "under-expressed" genes in the lower-left tail than the Illumina analysis, e.g., only one significant probe set with negative d-statistic was detected to be significantly under-expressed. Notice that the terms of "under-expressed" and "over-expressed" are referred to the gene expression behavior

comparing the parental (BB) rat against the congenic (BB7B) rat strain.

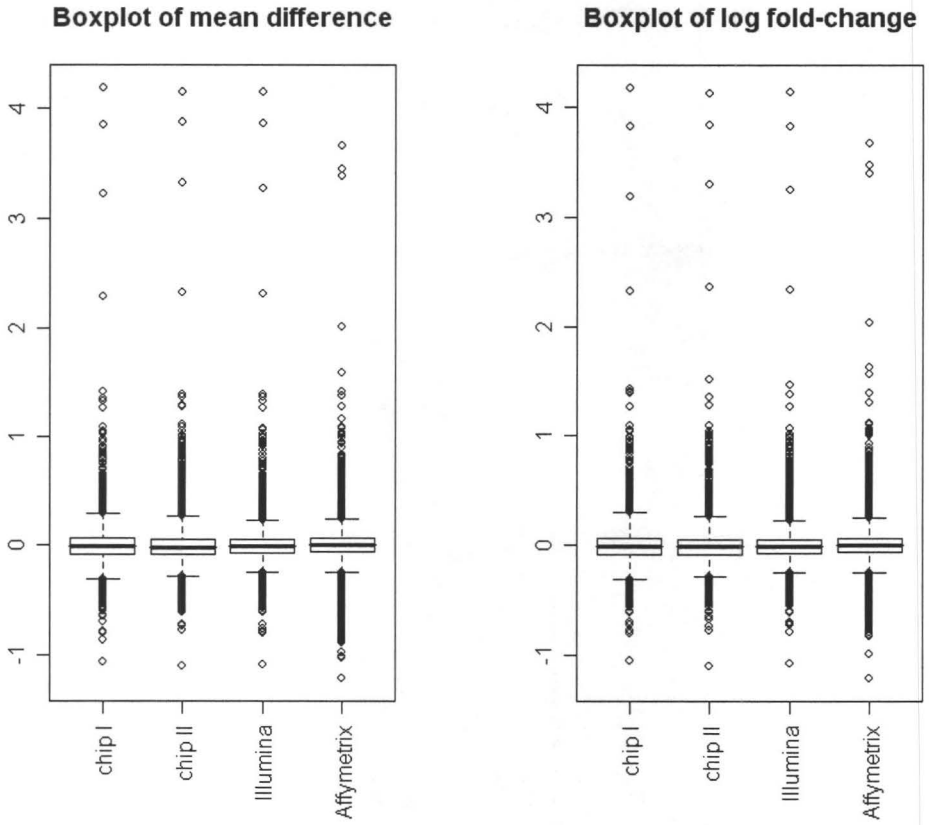
The boxplots in Figure 4.2 show the variances of the mean difference of the gene expression value and log fold-change on the Illumina and the Affymetrix platforms. Compared to the mouse platform (Figure 3.3), we find the rat data suggests a very similar variation of the mean difference and the log fold-change across the two microarray platforms. That is, the variation in Affymetrix platform from rat data is much smaller than that from the mouse data.

4.1.2 Comparing the (Comparable) Significant Genes

Again, we only investigate the concordance of those comparable rat genes across the two platforms. The comparable genes are subsetted based on the Illumina's annotation mapping file that maps the Illumina probeIDs on the "RatRef-12 v1" BeadChip to the Affymetrix probeIDs on rat "RAE230v2" GeneChip.

Figure 4.3 gives the scatter plot of the d-statistics and the Pearson's correlation coefficient among those comparable genes in two platforms, from which we find that the Pearson's correlation coefficient is 0.4223 which is not very high but much better than the coefficient we found in the mouse data (0.1464).

From the significant probe list tables (refer to Table 4.2 and Table 4.3) we find that, at 20% FDR, within the list of 127 significant probes from the Illumina analysis, there are 71 probes which are comparable to the Affymetrix platform according to the mapping file. On the other hand for the Affymetrix list, only 29 out of the 84 significant probe sets are comparable to the Illumina platform. For example, the first comparable significant gene in the Illumina list is "Defa", and the second is "Np4" (3rd in the original overall ranking in the Illumina platform when including those



(a) Mean difference

(b) Log fold-change

Figure 4.2: *Boxplots for the mean difference expression and the log fold-change comparison between the Illumina and Affymetrix analysis using the rat data.*

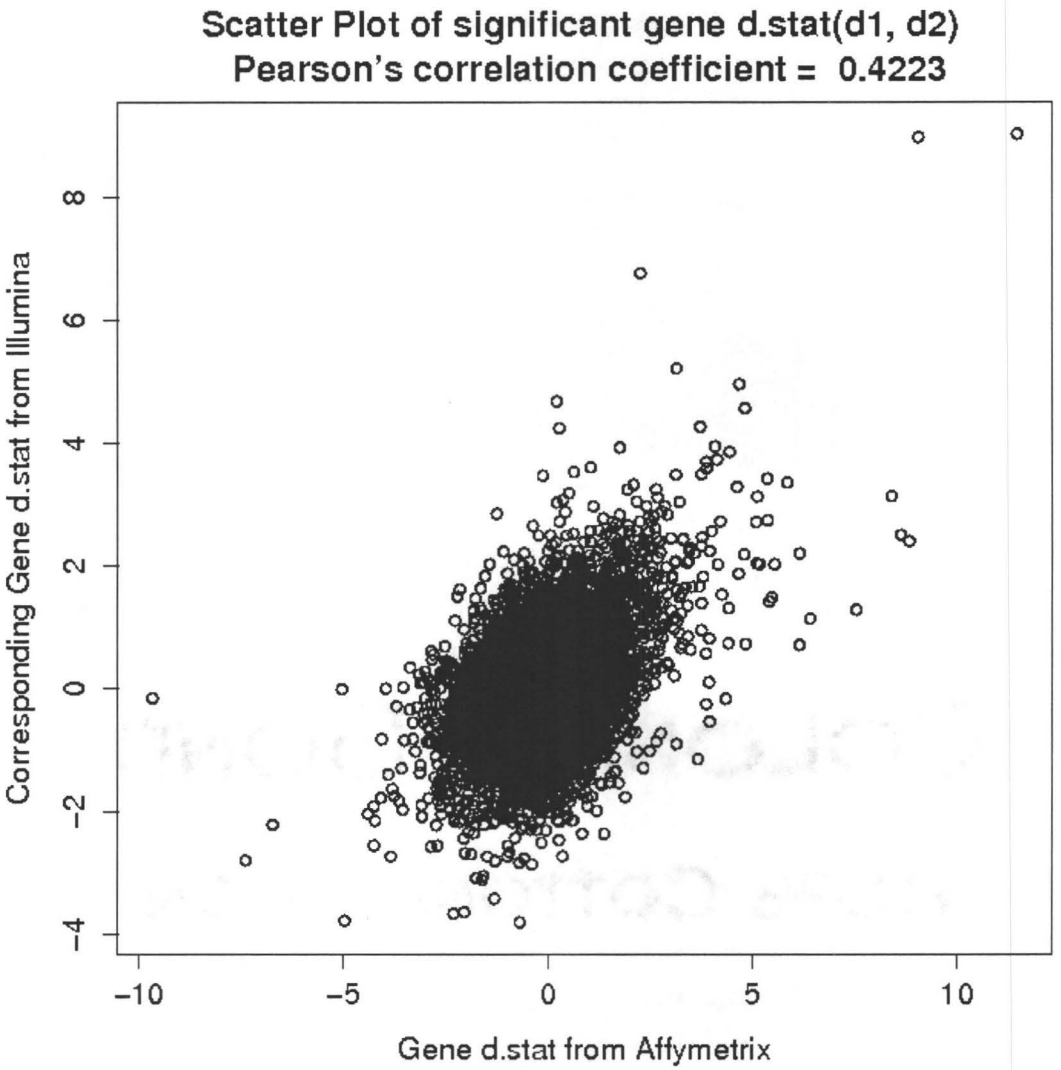


Figure 4.3: *Pairwise scatter plot of the d-statistics of those common genes from two rat microarray platforms and the Pearson's correlation coefficient.*

non-comparable genes). The corresponding ranking for these two comparable significant genes in the Affymetrix list are the 1st and the 2nd (original overall ranking). Furthermore, we find 16 such comparable genes that are in both significant lists at this FDR level of 20%. This forms the percentage of 22.5% for the Illumina list and 55.2% for the Affymetrix list for these comparable significant genes. As we did in the mouse data analysis, we summarized these percentage statistics of these comparable and common significant genes in Table 4.1, along with the Venn diagrams to show the agreement between these two significant lists, at three different FDR cutoff points of 20%, 10% and 5%.

Table 4.2 and Table 4.3 shows the comparable top 40 significant genes from both the Illumina and Affymetrix platforms and the corresponding rankings and d-statistics in the other platform. From these tables we find that the overall concordance between the two platforms is still fairly poor, but seems better than the ranking tables from the mouse analysis. This is also reflected in the higher correlation coefficient. Furthermore, we didn't find any opposite signs of the d-statistics which indicate the totally different expression behaviors.

Comparing these percentage statistics in Table 4.1, we find that, a lower (5%) FDR cutoff point improves the comparability of the significant genes between the two platforms. The percentage of those common significant genes increases to 40% and 100% from 12.9%, 50% when FDR = 10% and from 22.5%, 55.2% when FDR = 20%. The increase in the percentages indicates a better concordance for those extremely significant genes in the two platforms. Again, this finding is similar to what we have found from the mouse data in the previous chapter.

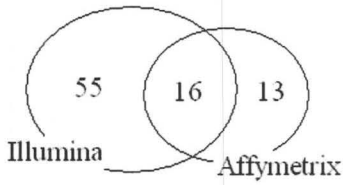
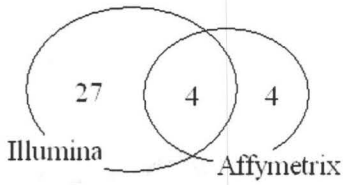
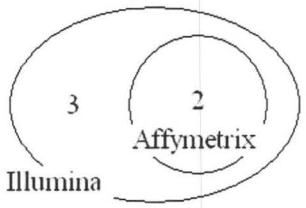
$FDR \simeq 20\%$			
	Illumina	Affymetrix	
Total number of significant genes	127	84	
Comparable genes	71 (55.9%)	29 (34.5%)	
Non-comparable genes	56 (44.1%)	55 (65.5%)	
16 genes in both lists	22.5%	55.2%	
$FDR \simeq 10\%$			
	Illumina	Affymetrix	
Total number of significant genes	56	21	
Comparable genes	31 (55.4%)	8 (38.1%)	
Non-comparable genes	25 (44.6%)	13 (61.9%)	
4 genes in both lists	12.9%	50%	
$FDR \simeq 5\%$			
	Illumina	Affymetrix	
Total number of significant genes	8	3	
Comparable genes	5 (62.5%)	2 (66.7%)	
Non-comparable genes	3 (37.5%)	1 (33.3%)	
2 genes in both lists	40%	100%	

Table 4.1: *Summary of concordance of the significant genes from the two rat platforms at different FDRs $\simeq 20\%, 10\%$ and 5%.*

ILMN Rank	ILMN Symbol	d.stat1	Affy Symbol	d.stat2	Affy Rank
1	Defa	11.4662	Defa	9.0358	1
3	Np4	9.0542	Np4	8.9842	2
5	MGC93766	8.8321	Abhd14b	2.3946	280
7	RGD1307811	8.6268	Spbc25	2.4979	228
8	Nradd	8.3999	Nradd	3.1356	83
11	Tmepai_predicted	7.5463	Tmepai_predicted	1.2771	2120
14	Cklfsf8	6.4120	Cmtm8	1.1348	2819
17	Ccr1	-7.4030	Ccr1	-2.8037	856
19	Cdkn3_predicted	6.1521	Cdkn3_predicted	2.1955	401
20	Ugt8	6.1513	Ugt8	0.6999	7254
22	Ldhc	5.8520	Ldhc	3.3581	63
24	Cspg5	-6.7348	Cspg5	-2.2186	2358
26	LOC362626	5.5358	RGD1359529	2.0191	546
28	Stk16	5.4746	Stk16	1.4787	1452
29	S100a9	5.4091	S100a9	1.4124	1635
30	Orm1	5.3726	Orm1	2.7379	146
31	Gfi1	5.3614	Gfi1	3.4216	59
35	LOC363028	5.1781	Spbc24_predicted	2.0162	550
37	Nkg7	5.1216	Nkg7	3.1254	84
38	LOC360847	5.1104	Ube2t_predicted	2.0310	532
39	LOC308607	5.0923	E2f8	2.7090	157

Table 4.2: *The table of comparable significant genes from the Illumina platform and the corresponding ranking in Affymetrix rat platform.*

Affy Rank	Affy Symbol	d.stat1	ILMN Symbol	d.stat2	ILMN Rank
1	Defa	9.0358	Defa	11.4662	1
2	Np4	8.9842	Np4	9.0542	3
5	RGD1560913_predicted	6.7699	LOC499322	2.2627	542
12	Rnpep	5.2130	Rnpep	3.1457	143
13	Nt5e	4.9527	Nt5	4.6760	45
15	Pacsin1	4.6739	Pacsin1	0.2144	17218
16	Fut2	4.5625	Fut2	4.8189	42
21	Zbp1	4.2633	Zbp1	3.7265	91
22	Ceacam1	4.2420	Ceacam1	0.2824	16150
26	Lcn2	3.9471	Lcn2	4.0892	66
28	Rrm2_mapped	3.9280	Rrm2	1.7558	1764
31	Ms4a2	3.8555	Ms4a2	4.4456	52
39	LOC24906	3.7298	LOC24906	4.1379	64

Table 4.3: *The table of comparable significant genes from the Affymetrix platform and the corresponding ranking in the Illumina rat platform.*

4.2 Comparison with the Mouse Data Analysis

In addition to the above findings, after comparing the rat analysis shown in Table 4.1 with the mouse analysis shown in Table 3.1, we find that the rat data suggest a better overall agreement across the two platforms for those comparable significant genes than the mouse data. This is true both in terms of the Pearson's correlation coefficient (0.4223 v 0.1464) and the percentage of common significant genes. Especially, with a lower FDR at 5% (for the top significant genes), the rat data suggests a much better agreement across the two platforms compared to the mouse data analysis. The two most significant comparable genes are the same in both platforms (Defa and Np4).

Besides, as shown previously in the boxplots of the mean difference and Log fold-change, the variations for the rat data are greatly improved compared to the mouse data (Figure 4.2 and Figure 3.3).

We will discuss and summarize the possible reasons for this improvement in the overall comparability between these two rodent models in the conclusion chapter.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this report, we discussed the within and across platform reproducibility between two major expression profiling microarray platforms, the Affymetrix GeneChip and Illumina BeadChip. From the study we find that, the reproducibility within the Illumina platform is pretty good in terms of the data quality and SAM analysis, especially for the most differentially expressed genes. However, as expected and seen on the Affymetrix platform, day effect does exist for the Illumina BeadChip platform as well. We also find that, due to the way the Beadarrays are packed on the Illumina Sentrix^{TM} platform, proper procedure of RNA preparation and hybridization can efficiently reduce the Illumina chip effect.

For the across platform reproducibility we find that, using the mapping file (both mouse and rat) provided by Illumina gives a fairly poor overall reproducibility of those comparable significant genes across the platforms. However, the rat data do show better overall concordance and smaller variance in the mean difference than the mouse

data do. We suspect that the improvement is due to the relatively larger genomic difference between the two rat strains. This shows the evidence that the genomic difference in biological samples could be a confounding variable that will affect the across platform reproducibility in microarray experiment. In addition, we find there is obviously better reproducibility for those extremely differentially expressed genes (top significant genes) across the two platforms. These findings partially agree with Barnes *et al.* (2005).

In addition to the agreements between the two platforms, we also find a lot of disagreements. Especially for the mouse data, it reveals that some genes are significantly under-expressed in the Illumina platform are actually significantly over-expressed in the Affymetrix platform. We conclude several reasons that might cause these disagreements as below:

First, as suggested in Barnes *et al.* (2005), the accuracy of the annotation plays a very important role when comparing the two platforms. The annotation that we used in our analysis (the mapping file from the Illumina) could be a major source of bias that caused the discrepancy in the resulting significant gene lists.

Secondly, under our biological assumption that most genes on the microarray are not differentially expressed, the experiments produce a lot of noise in the data. This low signal/noise ratio initially makes the across platform reproducibility lower.

Lastly, we used the same normalization method for both platforms, but this could also be a source to introduce bias to the comparison. We should use the optimal preprocess method to achieve the best SAM results for each individual platform and thus obtain the best comparability across the two platforms. However, the optimal normalization/preprocessing methods for both platforms, especially the BeadChip data, are still being developed and more study needs to be done in the future work.

5.2 Future Work

A lot of work could be done to look further into the questions of across platform reproducibility.

As mentioned above, more research should be done in the preprocessing methods in order to get the optimal normalized expression data for both Affymetrix and Illumina platforms. For this reason, for example, the GCRMA and other algorithms could be applied or need to be developed to use the probe level information for both GeneChip and BeadChip platforms.

Secondly, rather a more bioinformatical than a biostatistical research task though, another avenue for the future work is to find a more accurate and comparable annotation files for different microarray platforms. We believe a more accurate annotation linkage file between the Affymetrix and Illumina platforms would greatly improve the across platform reproducibility.

Finally, although we didn't test the significance of the suspicious confounding variable as mentioned previously, in order to see a more clear picture of the comparability (for different biological applications) across the Affymetrix and Illumina platforms, more experimental design could be performed to investigate how these confounding variables (e.g., the biological difference on genome of the comparison) can contribute to the reproducibility across the platforms.

Bibliography

Arabie, P., Hubert, L. J. and De Soete, G. (eds) (1996) *Clustering and Classification.* Singapore: World Scientific.

Barnes, M., Freudenberg, J., Thompson, S., Aronow, B. and Pavlidis, P. (2005) Experimental comparison and cross-validation of the affymetrix and illumina gene expression analysis platforms. *Nucleic Acids Research* **33**(18), 5914–5923. doi:10.1093/nar/gki890.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**(1), 289–300.

Bolstad, B. (2001) Probe level quantile normalization of high density oligonucleotide array data. http://www.stat.berkeley.edu/bolstad.

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**(2), 185–193.

Gunderson, K. L., Kruglyak, S., Graige, M. S., Garcia, F., Kermani, B. G., Zhao, C., Che, D., Dickinson, T., Wickham, E., Bierle, J., Doucet, D., Milewski, M., Yang, R., Siegmund, C., Haas, J., Zhou, L., Oliphant, A., Fan, J.-B., Barnard, S.

and Mark (2004) Decoding randomly ordered dna arrays. *Genome Research* **14**, 901–907.

Illumina (2004) *BeadStation 500X Gene Expression System.* Doc# 11176829 Rev. C.

Illumina (2005a) *Illumina BeadArray Reader User Guide.* Doc# 11179510 Rev. B.

Illumina (2005b) *BeadStudio User Guide - Data Analysis Software for Use with Illumina Gene Expression Products.* Doc# 11179632 Rev. B.

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**(2), 249–264.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G. N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martínez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q. and Yu, W. (2005) Multiple-laboratory comparison of microarray platforms. *Nature Methods* **2**(5), 345–350.

Kuhn, K., Baker, S. C., Chudin, E., Lieu, M.-H., Oeser, S., Bennett, H., Rigault, P., Barker, D., McDaniel, T. K. and Chee, M. S. (2004) A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Research* **14**, 2347–2356.

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

Shoemaker, J. S. and Lin, S. M. (eds) (2005) *Methods of Microarray Data Analysis.* Iv edition. New York: Springer.

Steemers, F. J. and Gunderson, K. L. (2005) Illumina, inc. *Pharmacogenomics* **6**(7), 777–782. DOI: 10.2217/14622416.6.7.777.

Storey, J. D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B* **64**(3), 479–498.

Storey, J. D. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics* **31**(6), 2013–2035.

Storey, J. D. and Tibshirani, R. (2001) Estimating false discovery rates under dependence, with applications to dna microarrays. *Technical of Report* Department of Statistics, Stanford University.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**(9), 5116–5121.

Yang, R. (2006) *Comparison of Normalization Methods in Microarray Analysis.* Master's thesis, McMaster University.

Appendix A

Glossary

Adenine (A) One of the four bases in DNA. Adenine (A) is always paired with thymine (T) according to the complementary base-pairing rule.

Affymetrix GeneChip One of the gene expression microarray platforms produced by Affymetrix, which consists of hundreds of thousands of probe cells, each containing several million copies of a specific oligonucleotide probe.

Annotation A sequence representation of genetic material with information relating position to gene names, regulatory sequences, repeats, and protein products, etc. This annotation is usually stored in predefined fields in biological databases, especially sequence databases.

Bioinformatics The field of science in which mathematics, computer science and information technology are merge into a single discipline to solve biology problems. This includes recording, annotation, storage, analysis, and searching/retrieval of nucleic acid sequence (genes and RNAs), protein sequence and structural information, etc.

Chromosome A large, threadlike macromolecule in the cell nucleus that car-

ries the genes in a linear order. e.g., the human genome consists of 23 pairs of chromosomes; in each of these pairs, one chromosome comes from the mother and the other from the father.

Complementary Base-Pairing The two complementary strands of DNA are connected via hydrogen bonds between base pairs: Adenine (A) is always paired with thymine (T), and cytosine (C) is always paired with guanine (G).

Cytosine (C) One of the four bases in DNA. Cytosine (C) always is paired with guanine (G) according to the complementary base-pairing rule.

DNA Acronym for deoxyribonucleic acid. DNA carries the genetic instructions, in very long sequences of nucleotides, for making living organisms. Two long strands of DNA in the form of a double-helix structure make up each chromosome.

DNA Microarray A small glass slide that comprises thousands – or even hundreds of thousands – of spots, or probe cells. Each of these spots contains specific genetic material for measuring the expression of a single gene. The most prominent examples are the Affymetrix GeneChip, cDNA microarrays and Illumina BeadChips.

Gene A segment of DNA involved in producing a polypeptide chain. It can include regions preceding and following the coding DNA as well as introns between the exons. Considered a unit of heredity.

Gene Expression A multiple-step process of converting a DNA sequence into a protein. It consists of "transcription" and "translation" steps.

Genome The full DNA sequence of a organism, constituting a blueprint for all cellular structures and activities in that organism. Virtually all cells of an organism contain a copy of the complete genome.

Guanine (G) One of the four bases in DNA. Guanine (G) is always paired with cytosine (C) according to the complementary base-pairing rule.

Hybridization The process of joining two complementary strands of DNA.

Illumina BeadChip Another gene expression microarray platform produced by Illumina.

Locus A location in the DNA sequence.

Mismatch (MM) An oligonucleotide probe used on an Affymetrix microarray to adjust the corresponding perfect match (PM) for background noise and nonspecific binding. Each MM oligo is a sequence consisting of 25 bases, which is almost identical to the sequence of corresponding perfect match. Only the thirteenth base of the MM is complementary to the thirteenth base of the corresponding PM.

Nonspecific Binding Hybridization of a mRNA or cDNA sequence to a spot that actually corresponds to a different mRNA or cDNA sequence.

Oligonucleotide Abbreviated "oligo". A sequence of single-strand RNA or DNA often used on microarrays as probe to measure gene expression. The oligo sequences can consist of different length, usually 25 to 60 bases.

Over-expression A gene is sometimes expressed in increased quantity comparing from one biological condition against the other. The increase can be measured in abundance of transcribed mRNAs, or the fluorescent intensity values between two biological conditions.

Perfect Match (PM) An oligonucleotide consisting of 25 bases used on Affymetrix GeneChips to measure gene expression. Each PM is used with its corresponding mismatch (MM).

Probe A piece of labeled RNA or DNA used to measure the expression of a gene.

Probe Set A set of probes pairs that represents a particular gene. On Affymetrix GeneChips, a probe set typically consists of 11-20 pairs of perfect match and mismatch probe pairs.

Protein A large, complex molecule. Proteins are responsible for virtually everything that happens in an organism.

RNA Acronym for ribonucleic acid. A chemical similar to a single strand of DNA but carrying a different sugar molecule (ribose instead of deoxyribose) and a different base (uracil [u] instead of thymine).

Southern Blot A method named after its inventor, Edwin Southern. It is used in molecular biology to check for the presence of a DNA sequence in a DNA sample.

Thymine (T) One of the four bases in DNA. Thymine (T) is always paired with adenine (A) according to the complementary base-pairing rule.

Under-expression Contrary to over-expression, a gene is sometimes expressed in decreased quantity comparing from one biological condition against the other. See over expression.