

國立交通大學

統計學研究所

碩士論文

使用效度與信度來比較艾菲爾微陣列基因晶片的
預處理方法與表現量差異方法的組合

Validity and Reliability of Combinations of

Preprocessing and Differential Expression Methods for

Affymetrix GeneChip Microarrays

研究生：王雅莉

指導教授：黃冠華 博士

中華民國九十六年七月

使用效度與信度來比較艾菲爾微陣列基因晶片的
預處理方法與表現量差異方法的組合
Validity and Reliability of Combinations of
Preprocessing and Differential Expression Methods for
Affymetrix GeneChip Microarrays

研究生：王雅莉 Student: Ya-Li Wang

指導教授：黃冠華 Advisor: Dr. Guan-Hua Huang



碩士論文

A Thesis
Submitted to institute of Statistics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics
July 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年七月

使用效度與信度來比較艾菲爾微陣列基因晶片的 預處理方法與表現量差異方法的組合

研究生：王雅莉

指導教授：黃冠華 博士

國立交通大學統計學研究所

摘要

微陣列晶片的技術已被廣泛地應用了好幾年，許多計算分析的工具也已被發展出來，而我們著重的平臺是已被廣泛應用的艾菲爾(Affymetrix)公司所製造的基因晶片。為了評估各種預處理方法與表現量差異方法組合的表現，我們考慮了四種常用的預處理方法：MAS 5.0、RMA、dChip及PDNN，與五種常用的表現量差異方法：fold-change、two sample t-test、SAM、EBarrays及limma。為了評估各種方法組合的效度，我們使用了三組嵌釘(spike-in)資料以及接收器運作指標曲線來做評估；而為了評估信度，我們採用另一組來自「微陣列晶片品質管制計畫」的資料組，此資料是將樣本分送至兩個同樣使用艾菲爾晶片平台的不同檢測站所生成的資料，用此兩檢測站的資料所選出的表現量差異基因的重複率作為比較信度的準則。若同時注重信度與效度，我們推薦幾種方法組合：當表現量差異基因個數少時，推薦 RMA+fold-change、RMA+SAM、RMA+limma、PDNN+fold-change、PDNN+SAM 與 PDNN+limma 此六種組合；而當表現量差異基因個數多時，則推薦 dChip(PM-only)+fold-change、dChip(PM-only)+SAM 與 dChip(PM-only)+limma 此三種組合。

關鍵字：微陣列晶片、艾菲爾基因晶片、接收器運作指標曲線

Validity and Reliability of Combinations of Preprocessing and Differential Expression Methods for Affymetrix GeneChip Microarrays

Student: Ya-Li Wang Advisor: Dr. Guan-Hua Huang
Institute of Statistics
National Chiao Tung University

ABSTRACT

Microarray technology has been widely used for several years and a large number of computational analysis tools have been developed. We focus on the most popular platform, Affymetrix GeneChip arrays. To evaluate which combinations of preprocessing and differential expression method perform well, we consider 4 popular preprocessing methods (MAS 5.0, RMA, dChip and PDNN) and 5 popular differential expression methods (fold-change, two sample t-test, SAM, EBarrays and limma). We use three spike-in datasets to assess the validity, and ROC curves are used for the evaluation. To evaluate the reliability, we use another dataset from MAQC project, which was generated using samples hybridized to Affymetrix platform at two different test sites. Overlap rates between two test sites are compared. To give consideration to both validity and reliability, six combinations are recommended when differentially expressed genes are less, RMA+fold-change, RMA+ SAM, RMA+ limma, PDNN+ fold-change, PDNN+SAM, and PDNN+limma. Three combinations are recommended when differentially expressed genes are more, dChip(PM-only)+ fold-change, dChip(PM-only)+SAM, and dChip(PM-only)+limma.

Key words: Microarray, Affymetrix GeneChip, ROC curve

誌 謝

這兩年，真的是酸甜苦辣樣樣來，精采的兩年、快樂的兩年、辛苦的兩年、充實的兩年！特別感謝黃冠華老師在這兩年的指導，我老是無預警地跑去找您 Meeting，甚至有時只是一件我不能確定的小事，但忙碌的您總是先放下手邊的工作來替我解決問題與疑惑，謝謝老師的包容與耐心指導，我才能夠順利地完成這篇論文；也謝謝老師所給的訓練，讓我寫論文的這一年來收穫良多，也過得很充實。另外，還要感謝所上其他老師們的教導，讓我對統計有更深入的瞭解。

此外特別感謝泰賓學長以及雪芳對於我論文及程式上的指導，謝謝你們的教導與幫忙我才能順利完成我的程式；也謝謝同門的素梅與阿淳一路來的鼓勵與支持，忍受我壓力大時的鬼叫鬼叫，因為有你們的陪伴，我才能撐過這辛苦的一年，很高興有你們一起奮鬥到最後，我最好的夥伴們！還有一起待在嚴寒的電腦室奮鬥的美惠及柯柯，有你們的陪伴、打氣與取暖，我才不至於凍僵在電腦室；另外還有熱心的建威，謝謝你總是認真地幫我一起想辦法解決問題；很有大哥風範的永在，雖然最後我們不需要改成 Latex，但還是很謝謝你從以前一路以來的幫忙。

很開心在交大遇見了這群可愛的同學們，也讓我見識到何謂真正的高手，聰明又擅長寫程式的雪芳、雖然說話很冷但一起做報告後驚覺很厲害的小米、博學多聞電腦又很強的大哥永在、又帥又聰明英文又好的吳益銘…等等！我喜歡研究所的學習環境，擁有自己的研究室，可以恣意地在自己的地盤撒野、打鬧，偶爾想認真一下也能找到同學互相請教、討論，研究室像是我在新竹的另一個家，感謝所有 408、409 同學的一路陪伴，我才能快樂地渡過這寫論文辛苦的一年！

最後，最感謝從小到大帶給我許多溫暖與關懷的父母及家人，謝謝你們的支持與鼓勵，讓我無憂無慮快樂地過著生活，永遠有個溫暖的家當後盾，在我失意、不如意的時候給我安慰與鼓勵，謝謝你們！

在此，謹以此篇論文，獻給我最愛的父母及家人，還有許多陪伴我的好友們！

王雅莉 2007.07.08

Contents

Abstract (in Chinese)	i
Abstract (in English)	ii
Acknowledgements (in Chinese)	iii
Contents	iv
List of Tables	vi
List of Figures	vii
1 Introduction	1
2 Literature Review	3
2.1 Background of microarray	3
2.2 Affymetrix GeneChip array	3
2.3 Overview of preprocessing method	5
2.3.1 Background adjustment	5
2.3.2 Normalization	5
2.3.3 Summarization	6
2.4 Four preprocessing methods used	6
2.5 Five differential expression methods used	14
2.6 Datasets	22
3 Materials and Methods	27
3.1 Implementation of methods selected	27
3.1.1 Four preprocessing methods used	27
3.1.2 Five differential expression methods used	30
3.2 Assessment of validity	32
3.3 Assessment of reliability	35

4	Results	38
4.1	Assessment of validity by ROC curve	38
4.2	Assessment of reliability by overlap rate	40
5	Conclusions and Discussion	43
5.1	Conclusions	43
5.2	Discussion	44
	Reference	46



List of Tables

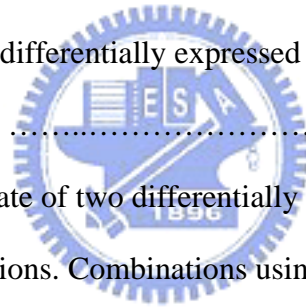
Table 1. Summary of the four preprocessing methods used	49
Table 2. Summary of the three spike-in datasets used	49
Table 3. Affymetrix human genome U95 dataset contains 14 spike-in gene groups in each of 14 experimental groups. This table shows the spiked-in concentrations (pM)	50
Table 4. Affymetrix human genome U133 dataset contains 14 spike-in gene groups in each of 14 experimental groups. This table shows the spiked-in concentrations (pM)	51
Table 5. Area under ROC curve (FP<100) for HGU95 dataset	52
Table 6. Area under ROC curve (FP<100) for HGU133 dataset	53
Table 7. Area under ROC curve (FPR<0.1) for Golden Spike dataset	54



List of Figures

Figure 1-1. ROC curves for all combinations using HGU95 dataset. Combinations using the same preprocessing method are assigned to the same color.	55
Figure 1-2. ROC curves for all combinations using HGU95 dataset but FP<100.	55
Figure 1-3. ROC curves for all combinations using HGU133 dataset.	56
Figure 1-4. ROC curves for all combinations using HGU133 dataset but FP<100.	56
Figure 1-5. ROC curves for all combinations using Golden Spike dataset.	57
Figure 1-6. ROC curves for all combinations using Golden Spike dataset but false positive rate<0.1.	57
Figure 2-1. For HGU95 dataset, ROC curves of all combinations are divided by preprocessing method.	58
Figure 2-2. For HGU133 dataset, ROC curves of all combinations are divided by preprocessing method.	58
Figure 2-3. For Golden Spike dataset, ROC curves of all combinations are divided by preprocessing method.	59
Figure 3-1. ROC curves for all combinations using HGU95 dataset. Combinations using the same differential expression method are assigned to the same color.	59
Figure 3-2. ROC curves for all combinations using HGU133 dataset.	60
Figure 3-3. ROC curves for all combinations using Golden Spike dataset.	60
Figure 4. Overlap rate of two differentially expressed gene lists generated using different combinations.	61
Figure 5-1. Overlap rate of two differentially expressed gene lists generated using different combinations for K_AA treatment/control.	62

Figure 5-2. Overlap rate of two differentially expressed gene lists generated using different combinations for L_AA treatment/control.	62
Figure 5-3. Overlap rate of two differentially expressed gene lists generated using different combinations for L_CFY treatment/control.	63
Figure 5-4. Overlap rate of two differentially expressed gene lists generated using different combinations for L_RDL treatment/control.	63
Figure 6. Overlap rate of two differentially expressed gene lists generated using different combinations with EBarrays as differential expression method. ...	64
Figure 7. Overlap rate of two differentially expressed gene lists generated using different combinations. Combinations using the same differential expression method are assigned to the same color.	65
Figure 8. Overlap rate of two differentially expressed gene lists generated using nine different combinations.	66
Figure 9-1. Average overlap rate of two differentially expressed gene lists generated using different combinations. Combinations using the same preprocessing method are assigned to the same color.	67
Figure 9-2. Average overlap rate of two differentially expressed gene lists generated using different combinations. Combinations using the same differential expression method are assigned to the same color.	68



1 Introduction

Microarray is a device designed to simultaneously measure the expression levels of many thousands of genes in a particular tissue or cell type. It is widely used in many areas of biomedical research especially Affymetrix GeneChip platform. Millions of probes with length of 25 nucleotides are designed on an Affymetrix array. Two categories of probes are designed: “perfect match (PM)” probe perfectly matches its target sequence and “mismatch (MM)” probe is created by changing the middle (13th) base of its paired perfect match probe sequence. The purpose of designing MM probe is to detect the nonspecific binding because its perfect match partner may be hybridize to nonspecific sequences. A paired PM and MM is called a “probe pair” and each gene will be represented by 11-20 probe pairs typically. Owing to this distinctive design, preprocessing Affymetrix expression arrays usually involves three main steps. That are background adjustment, normalization, and summarization. Nowadays, a large number of preprocessing methods have been developed to estimate expression levels of genes.

Another fundamental goal of a microarray experiment is to identify those genes that are differentially expressed within different samples. For example, a disease may be caused by large expression of particular genes resulting in variation between diseased and normal tissues. The method used to detect the genes expressed differentially between different samples is called differential expression method. Various preprocessing and differential expression methods have been proposed, and their developers using different datasets and criteria claimed there are some features superior to other methods. In this thesis, we use the common datasets to evaluate combinations of the most popular preprocessing and differential expression methods in terms of validity and reliability. We try to help users of Affymetrix to select the best

method for their own microarray data.

Here we consider four commonly used preprocessing methods, Microarray Suite software Version 5.0 (MAS 5.0: Affymetrix, 2002), DNA-Chip Analyzer (dChip: Li and Wong, 2001a and 2001b), Robust Multi-array Analysis (RMA: Irizarry *et al.*, 2003a) and Position-Dependent Nearest-Neighbor (PDNN: Zhang *et al.*, 2003), and five popular differential expression methods, Fold change(FC), two sample t-test, Significance Analysis of Microarrays (SAM: Tusher, Tibshirani and Chu, 2001), Parametric Empirical Bayes methods (EBarrays: Newton *et al.*, 2001, Kendziorski *et al.*, 2003, and Newton and Kendziorski, 2003), and Linear Models and Empirical Bayes methods (limma: Smyth, 2004). Four datasets in total are used. Three are spike-in datasets used to assess the validity: two from Affymetrix Latin square datasets and one from the Golden Spike Project. ROC curves are used for the evaluation. To evaluate the reliability, we use another dataset from MicroArray Quality Control (MAQC) project, which was generated using samples hybridized to Affymetrix platform at two different test sites. Overlap rates between two test sites are compared.

2 Literature Review

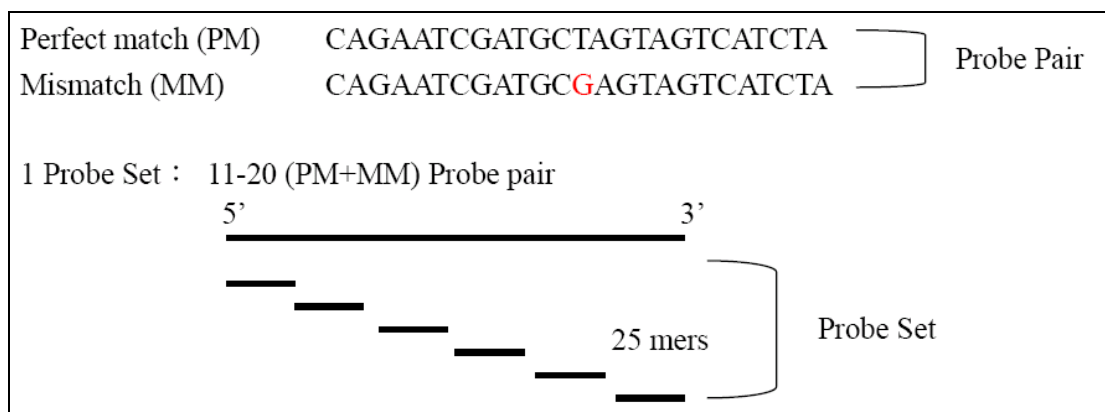
2.1 Background of microarray

Microarray is a device designed to simultaneously measure the expression levels of many thousands of genes in a particular tissue or cell type. It is widely used in many areas of biomedical research. Microarray technology makes use of the sequence resources created by the genome projects and other sequencing efforts to detect which genes are expressed in a particular cell type of an organism. Many different microarray technologies have been developed, and can be classified into three main categories: cDNA array (highly variable in length), short oligonucleotide array (25-30 base) and long oligonucleotide array (50-80 base). The high-density oligonucleotide array produced by Affymetrix is one kind of the short oligonucleotide array. Affymetrix GeneChip arrays have become a widely used microarray platform and numerous of methods have been proposed for analyzing this type of microarray data. This thesis focuses on the analysis of data from Affymetrix GeneChip expression arrays.

2.2 Affymetrix GeneChip array

Affymetrix GeneChip array are high throughput assays for measuring the expression levels of many thousands of gene transcripts simultaneously in a particular tissue or cell type. The technology takes advantage of hybridization properties of nucleic acid. To measure how much quantity of specific nucleic acid transcripts of interest present in the sample, complementary molecules are used to attach to a solid surface. The specific nucleic acid transcripts of interest presented in the sample are referred as “target”, and the complementary molecules attached to a solid surface are referred as “probe”. Millions of probes with a usually length of 25 nucleotides are

produced on an Affymetrix array. Two categories of probes are designed, “perfect match (PM)” probe perfectly matches its target sequence, and “mismatch (MM)” probe is created by changing the middle (13th) base of its paired perfect match probe sequence. The purpose of designing MM probe is to detect the nonspecific binding because its perfect match partner may be hybridize to nonspecific sequences. A paired PM and MM are called a “probe pair”, and a gene represented by multiple probe pairs is called a “probeset”. Typically, each gene will be represented by 11-20 probe pairs. For more comprehensible, we show these in following graph.



After RNA samples were prepared, labeled and hybridized to an array with millions of probes, the array is scanned and pixel intensity values are calculated using peculiar instruments by Affymetrix. According to these values, intensity values for each probe, called probe-level intensities, are computed and stored in a CEL file. The next step is to find a way to combine the 11-20 probe pair intensities together to a summary value for a given gene. The summary value for a given gene is defined as a measure of expression that represents the amount of the corresponding mRNA species. However, due to many systematical biases from different sources in miroarray experiments, data preprocessing becomes more necessary and important. The goal of data preprocessing is to obtain a corrected intensity value that represents the abundance of mRNA, instead of an uncertain brightness biased by other sources. Preprocessing Affymetrix expression arrays usually involves three main steps:

background adjustment, normalization, and summarization, that is, low-level analysis of Affymetrix microarray. Details for preprocessing methods are described later.

Another fundamental goal of a microarray experiment is to identify those genes that are differentially expressed within different samples. For example, a disease may be caused by large expression of a particular gene or genes resulting in variation between diseased and normal tissues. Numerous of differential expression methods are proposed to detect those differentially expressed genes between diseased and normal tissues. Throughout this thesis, we attempt to compare those combinations of the most commonly used preprocessing methods and differential expressed methods.

2.3 Overview of preprocessing method

Here we interpret the three main steps of data preprocessing briefly before mentioning these preprocessing methods used to compare.

2.3.1 Background adjustment

Because partial measured probe intensities maybe caused by non-specific hybridization or the noise in the optical detection system, background adjustment is essential to rid of these intensities not exactly expressed from genes. Observed probe intensities need to be adjusted to give the accurate expression levels of specific hybridization (Huber et al., 2005). Some methods make use of MM probes to adjust, but some are not.

2.3.2 Normalization

During the process of carrying out the microarray experiment involving multiple arrays, there are many obscuring sources of variation involved, such as physical problems with the arrays, laboratory conditions, hybridization reactions, labeling, and scanner difference. In order to compare measurements from different arrays, implying different tissue, some proper normalization is necessary.

2.3.3 Summarization

Due to Affymetrix platform designing multiple probes to represent a gene, summarization is needed to combine these probe intensities to a single value. For each gene, the background adjusted and normalized intensities are used to be summarized into one measurement that estimates the expression level.

2.4 Four preprocessing methods used

Notations:

$i = 1, \dots, I$: representing the different array (sample)

$j = 1, \dots, J$: representing the probe pair in the gene

$g = 1, \dots, G$: representing the probe set (gene)

MAS 5.0

MAS 5.0 (Microarray Suite software, Version 5.0) is offered by Affymetrix (Affymetrix, 2002). The gene expression level is calculated from the combined, background-adjusted, PM and MM values of the probe set. At the beginning, both PM and MM probe intensities must be preprocessed for background adjustment.

To do the background adjustment, the array is divided into K rectangular zones (default $K = 16$). The probes are ranked and the lowest 2% is chosen as the background b_{z_k} for that zone. Then each probe intensity is adjusted based on a weighted average of each of the background values, $b(x, y)$.

$$b(x, y) = \frac{1}{\sum_{k=1}^K w_k(x, y)} \sum_{k=1}^K w_k(x, y) b_{z_k}.$$

The weights for zone k , $w_k(x, y)$, are dependent on distance from the probe location (x, y) to each of the zone centers. In particular, the weight is defined as:

$$w_k(x, y) = \frac{1}{d_k^2(x, y) + smooth},$$

where $d_k^2(x, y)$ is the Euclidean distance from the probe location (x, y) to the center

of zone k. The default value of *smooth* is 100, which is added to $d_k^2(x, y)$ to ensure that the value will never be zero. The calculated background, $b(x, y) - 1$, establishes a “floor” to be subtracted from each raw probe intensity. There are some rules for avoiding leading to the negative intensity.

After each probe intensity is preprocessed for background adjustment, an ideal mismatch value is calculated and subtracted to adjust the PM intensity. Originally, the suggested purpose of the MM probes was that they could be used to adjust the PM probes for non-specific binding. The naïve approach is subtracting the intensity of MM probe from the intensity of the corresponding PM probe. However, this becomes problematic because the MM value is sometimes larger than the PM value. To avoid taking the negative expression value, Affymetrix introduced the concept of an *Ideal Mismatch (IM)*, a quantity derived from the *MM* value that is never bigger than its corresponding *PM* value. IM is defined as a quantity equal to MM when $MM < PM$, but adjusted to be less than PM when $MM \geq PM$. This is done by computing the *specific background*, SB_g , for each probe set g . If the $g=1, \dots, G$ is the probe set and $j=1, \dots, J$ is the probe pair, then the SB_g is defined as:

$$SB_g = Tukey\ Biweight\left\{\log_2(PM_{jg}) - \log_2(MM_{jg}) : j = 1, \dots, J\right\},$$

and the IM_{jg} for probe pair j in probe set g is defined as:

$$IM_{jg} = \begin{cases} MM_{jg} & , MM_{jg} < PM_{jg} \\ \frac{PM_{jg}}{2^{SB_g}} & , MM_{jg} \geq PM_{jg} \text{ and } SB_g > \text{contrast } \tau \\ \frac{PM_{jg}}{2^{\left(\frac{\text{contrast } \tau}{1 + \left(\frac{\text{contrast } \tau - SB_g}{\text{scale } \tau}\right)}\right)}} & , MM_{jg} \geq PM_{jg} \text{ and } SB_g \leq \text{contrast } \tau \end{cases}$$

where *contrast* τ (with a default value of 0.03) and *scale* τ (with a default value of 10) are tuning constants. The adjusted PM intensity is obtained by subtracting the corresponding IM from the observed PM intensity. Then, MAS 5.0 use a one-step Tukey Biweight to combine the probe intensities in log scale.

$$signal_g = \text{the anti log of Tukey Biweight} \left\{ \log_2 (PM_{jg} - IM_{jg}) \right\}.$$

Finally, signal is scaled using a trimmed mean. They defined a scaling factor *sf* and a normalization factor *nf* in their algorithm.

$$sf = \frac{Sc}{TrimMean(signal, 0.02, 0.98)},$$

where *Sc* is the target signal (default *Sc*=500). MAS 5.0 offers two analysis for user to choose, that are absolute analysis and comparison analysis. According to which analysis you want to perform, *nf* has different definition.

$$nf = \begin{cases} 1 & , \text{ for absolute analysis} \\ \frac{TrimMean(SPVe, 0.02, 0.98)}{TrimMean(SPVb, 0.02, 0.98)} & , \text{ for comparison analysis} \end{cases}$$

where *SPVb* is the baseline array signal, and *SPVe* is the experiment array signal. More details are described in the Statistical Algorithms Description Document (Affymetrix, 2002). The reported value of MAS5.0 of probe set *g* is:

$$ReportedValue(i) = nf \times sf \times signal_g.$$

dChip

dChip (DNA-Chip Analyzer) is also a popular software for Affymetrix platform probe-level and high-level analysis of gene expression microarrays (Li and Wong, 2001a) and SNP microarrays. This software can be downloaded from the website <http://biosun1.harvard.edu/complab/dchip/>. dChip can be used to fit the Model Based Expression Index (MBEI) (Li and Wong, 2001a), and obtain what we refer to as the dChip expression measure. Li and Wong reported that variation of a specific probe across multiple arrays (the between-array variance) is in general smaller than the

variance across probes within a probe set (the within-probe set variance) (Li and Wong, 2001a). To account for this strong probe affinity effect, they proposed a multiplicative model, for any given gene:

$$\begin{aligned} MM_{ij} &= \nu_j + \theta_i \alpha_j + \varepsilon \\ PM_{ij} &= \nu_j + \theta_i \alpha_j + \theta_i \phi_j + \varepsilon = \nu_j + \theta_i (\alpha_j + \phi_j) + \varepsilon \end{aligned} \quad (1)$$

Here PM_{ij} and MM_{ij} denote the PM and MM intensity values from the i -th array and the j -th probe pair for this gene. θ_i denotes the expression index for this gene in the i -th array. Here multiple arrays are available for analysis. Assume that the intensity value of a probe will increase linearly as θ_i increases, but different increasing rate for different probes. And within the same probe pair, the PM_{ij} will increase at a higher rate than the MM_{ij} . α_j and ϕ_j represent the increasing rate of the MM_{ij} probe and the additional increasing rate in the corresponding PM_{ij} probe respectively. The increasing rates are assumed to be nonnegative. ν_j is the baseline response of the j -th probe pair due to nonspecific hybridization, and ε are assumed to be independent normally distributed errors.

The model for individual probe responses implies an even simpler model for the PM–MM differences:

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}, \quad i = 1, \dots, I, \quad j = 1, \dots, J. \quad (2)$$

The model above is called PM-MM difference model (Li and Wong, 2001a).

Li and Wong discovered that because of doubting the efficiency of using MM probes, some investigators design custom arrays using PM probes exclusively. Thus, they proposed another model later to estimating gene expression levels, called PM-only model (Li and Wong, 2001b). The PM-only model focus only on PM probes, using the description of PM in model (1). The PM-only model is as follows:

$$PM_{ij} = v_j + \theta_i \alpha_j + \theta_i \phi_j + \varepsilon = v_j + \theta_i \phi_j' + \varepsilon. \quad (3)$$

Notations in the PM-only model represent the same meaning as well as PM-MM difference model, except that ϕ_j' merges the two increasing rates α_j and ϕ_j .

No matter what model above is referred, Li and Wong's measure is defined as the maximum likelihood estimates of the expression index θ_i and outlier probe intensities are removed as part of the estimation procedure. Before computing model-based expression levels, dChip use the "Invariant Set" normalization method to normalize arrays at PM and MM probe levels for PM-MM difference model or PM probe levels for PM-only model. Using a baseline array, arrays are normalized by selecting invariant sets of probes then using them to fit a non-linear relationship between the "treatment" and "baseline" arrays. A set of probe is said to be invariant if ordering of probe in one chip is the same in other set. By default, an array with median overall intensity is chosen and all other arrays are normalized to it.

In order to summarize the probe intensities, dChip performs the "Invariant Set" normalization method, then fit the normalized probe intensities to the alternative model for any given gene. Maximum likelihood estimates of the expression index θ_i is the expression measure for this gene in array i .

RMA

RMA (Irizarry *et al.*, 2003a), Robust Multi-array Analysis, is an expression measure consisting of three particular preprocessing steps: convolution background correction, quantile normalization, and a summarization based on a multi-array model fit robustly using the median polish algorithm. Many preprocessing methods, such as MAS 5.0 and dChip, calculating their measures rely on the difference PM-MM with the intention of correcting for non-specific binding. However, the exploratory analysis presented in Irizarry *et al.* (2003a) suggests that the MM probe may be a mixture

probe for which detects not only non-specific binding and background noise but also the transcript signal just like the PM probe. Thus, subtracting the MM intensity from the PM intensity as a way of correcting for non-specific binding and background noise is not always appropriate. These RMA authors proposed a procedure ignoring the MM intensities and using only the PM intensities.

The RMA convolution background correction method is motivated by looking at the distribution of probe intensities. The model observed PM as the sum of a background intensity bg_{ijg} caused by optical and nonspecific binding, and a signal intensity s_{ijg} .

$$PM_{ijg} = bg_{ijg} + s_{ijg}, i = 1, \dots, I, j = 1, \dots, J, g = 1, \dots, G$$

with i representing the different array, j representing the probe pair, and g representing the different probe set. Under the model above, the background corrected probe intensities will be given by $B(PM_{ijg})$, where $B(PM_{ijg}) \equiv E(s_{ijg} | PM_{ijg})$. To obtain a computationally feasible $B(\cdot)$ we consider the closed-form transformation obtained when assuming that s_{ijg} is distributed exponential and bg_{ijg} is distributed normal, and the results obtained using $B(\cdot)$ work well in practice (Irizarry *et al.*, 2003a).

Next, perform the quantile normalization, which is to make the distribution of probe intensities for each array the same (Bolstad *et al.*, 2003). In order to summarize the probe intensities, RMA introduced a log scale linear additive model. The model is:

$$T(PM_{ij}) = e_i + a_j + \varepsilon_{ij},$$

where PM_{ijg} represents the PM intensity of array $i=1, \dots, I$ and probe pair $j=1, \dots, J$, for any given probe set g . $T(\cdot)$ represents the transformation that background corrects, normalizes, and logs the PM intensities, e_i represents the log2 scale

expression value found on arrays i , a_j represents the log scale affinity effects for probes j , and ε_{ij} represents error (Irizarry *et al.*, 2003b). To protect against outlier probes, they use a robust procedure, such as median polish, to estimate model parameters (Irizarry *et al.*, 2003a). The estimate of e_i as the log scale measure of expression refers to as robust multi-array average (RMA).

PDNN

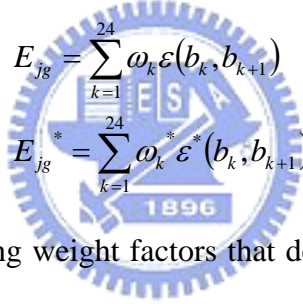
Zhang *et al.* (2003) propose a simply free energy model over the probe signals that enables to estimate the gene expression levels, called “position-dependent nearest-neighbor (PDNN) model”, for the formation of RNA-DNA duplexes on Affymetrix microarray. Different from most methods focused on statistical models, it is a physical model taking into account the sequence of nearest-neighbors (adjacent two bases) and the position of these nucleotide pairs. It has been suggest that the effect of nearest-neighbor nucleotide pairs is the most important factor in determining RNA/DNA duplex stability. Their model also describes binding interactions complicated by many factors such as steric hindrance on the chip surface, probe-probe interaction and RNA secondary structure formation.

The model is based on the nearest-neighbor model (Sugimoto *et al.*, 1995) with two modifications: (1) a positional weight factor is added to reflect the different contributions from different part of the probe; (2) two different types of binding on the probes are considered. The two types of binding are gene-specific binding (GSB), representing the formation of DNA-RNA duplexes with exact complementary sequences, and non-specific binding (NSB), representing the formation with many mismatches between the probe and the attached RNA molecule. Notice that PDNN assumes that the majority of probes are designed specifically for their target, and only PM probes are used for GSB and NSB estimation. PDNN model divides signal of a

probe into three components, GSB, NSB and uniform background B, as follows:

$$\hat{I}_{jg} = \frac{N_g}{1 + e^{E_{jg}}} + \frac{N^*}{1 + e^{E_{jg}^*}} + B,$$

where \hat{I}_{jg} is denoted as the expected intensity of the j -th probe in a probe set targeted to detect gene g , N_g as the true expression level for gene g , and N^* as the population of RNA molecules that contributes to NSB. E_{jg} is defined as the free energy for formation of the specific RNA-DNA duplex with the targeted gene, and E_{jg}^* is the average free energy for NSB, that is, formation of duplexes with many different genes. E_{jg} and E_{jg}^* are computed as weighted sums of stacking energies with the sequence of a probe is given as $(b_1, b_2, \dots, b_{25})$.



$$E_{jg} = \sum_{k=1}^{24} \omega_k \varepsilon(b_k, b_{k+1})$$

$$E_{jg}^* = \sum_{k=1}^{24} \omega_k^* \varepsilon(b_k, b_{k+1})$$

with ω_k and ω_k^* representing weight factors that depend on the position along the probe from the 5' end to the 3' end, and $\varepsilon(b_k, b_{k+1})$ is defined the same as the stacking energy used in the nearest-neighbor model (Sugimoto *et al.*, 1995). Both of GSB and NSB are involving 16 stacking energy parameters and 24 weight factors.

The unknown parameters are obtained by minimizing the fitness function F to optimize the match between the expected probe intensity \hat{I}_{jg} and the observed probe intensity I_{jg} .

$$F = \sum \frac{(\ln \hat{I}_{jg} - \ln I_{jg})^2}{M},$$

where M is the total number of probes on an array. A Monte Carlo simulation procedure is used to minimize the fitness function F. When the parameters are given,

the gene expression level N_g can be calculated and are scaled to an average of 500 on an array.

For more comprehensible, we give a summary table for the four preprocessing methods above in Table 1.

2.5 Five differential expression methods used

Fold-change

Fold-change is the most commonly used method of detecting differentially expressed gene between two compared condition samples. It is often the first method used in microarray analysis. For any given gene, fold-change is calculated by the probeset intensity ratio of two compared condition samples. If there are replicates, we usually average across the samples for each condition in advance. Then the ratio of these averaged values is referred as fold-change. Larger fold-change leads the gene to be more likely differentially expressed gene. Biologist favors fold-change equal to 2 as the threshold of differential expression.

Two sample t-test

The simplest statistic method for comparing means between two groups is two sample t-test. It is widely applied in microarray analysis when detecting the differentially expressed genes between two compared condition samples. For any given gene, assume that the measurements of the first condition sample arise independently and identically from normal distribution with mean μ_1 and variance σ_1^2 , and the measurements of the second condition sample arise independently and identically from normal distribution with mean μ_2 and variance σ_2^2 . When carrying out a two sample t-test, the variances of the two samples may be assumed to

be equal or unequal. The approach of unequal variance assumption is also called Welch's t-test. For any given gene g , suppose that the number of samples in condition1 and in condition2 are M and N respectively. Here we describe the two tests briefly.

Two sample t-test for equal variance:

$$\begin{aligned}
 & \text{condition 1: } X_{g1}, \dots, X_{gM} \sim N(\mu_1, \sigma^2) \\
 & \text{condition 2: } Y_{g1}, \dots, Y_{gN} \sim N(\mu_2, \sigma^2) \\
 & H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2 \\
 & \text{test statistic: } \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{M} + \frac{1}{N}} S_p} \sim T_{M+N-2}, \\
 & \text{where } S_p^2 = \frac{\sum_{i=1}^M (X_i - \bar{X})^2 + \sum_{i=1}^N (Y_i - \bar{Y})^2}{M + N - 2}.
 \end{aligned}$$

Two sample t-test for unequal variance (Welch's t-test):

$$\begin{aligned}
 & \text{condition 1: } X_{g1}, \dots, X_{gM} \sim N(\mu_1, \sigma_1^2) \\
 & \text{condition 2: } Y_{g1}, \dots, Y_{gN} \sim N(\mu_2, \sigma_2^2) \\
 & H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2 \\
 & \text{test statistic: } \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{S_X^2}{M} + \frac{S_Y^2}{N}\right)}} \sim T_v \text{ (approximately)}, \\
 & \text{where } S_X^2 = \frac{1}{M-1} \sum_{i=1}^M (X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 \text{ and} \\
 & v = \frac{\left(\frac{S_X^2}{M} + \frac{S_Y^2}{N}\right)^2}{\frac{S_X^4}{M^2(M-1)} + \frac{S_Y^4}{N^2(N-1)}}.
 \end{aligned}$$

After performing the test and the conclusion leads to reject H_0 , we consider that this gene is a differentially expressed gene.

SAM (Significance Analysis of Microarrays)

SAM is a method for identifying genes on a microarray with statistically significant changes in expression. It was proposed by Tusher, Tibshirani and Chu

(2001). The method is based on a modified version of the standard t-statistic. Standard t-statistic method is popular but having the problem of multiple testing. That is, when thousands of hypotheses are tested simultaneously in microarray experiment, it would increase chance of false positives. For example, if we have 10000 genes in our microarray experiment and all of them are non-differentially expression. Choosing significance level $\alpha = 0.01$, we would expect that there are $10000 \times 0.01 = 100$ genes called significant (having p -value < 0.01). Even if we choose a small $\alpha = 0.01$ to evaluate small numbers of genes, we still get 100 genes called significant because of multiple testing. This problem led them to develop a statistical method adapted specifically for microarrays, Significance Analysis of Microarrays (SAM).

For each gene g , the “relative difference” d_g in gene expression is:

$$d_g = \frac{r_g}{s_g + s_0}.$$

Here r_g is a score, s_g is a standard deviation, and s_0 is an exchangeability factor (Chu *et al.*). SAM software can be adapted for many types of experimental data, such as a simple unpaired two-group data, multiple-group data, paired data, censored survival data, \dots , etc. For each type of experimental data, SAM defines different r_g and s_g in a different way (Chu *et al.*). We now focus only on the experiment of two groups. In this case, r_g and s_g have the following definition:

$$r_g = \bar{x}_{g2} - \bar{x}_{g1}$$

$$s_g = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times \frac{\sum_{i \in \text{group1}} (x_{gi} - \bar{x}_{g1})^2 + \sum_{i \in \text{group2}} (x_{gi} - \bar{x}_{g2})^2}{n_1 + n_2 - 2}},$$

where \bar{x}_{g1} and \bar{x}_{g2} are defined as the average levels of expression for gene g in group 1 and group 2, and x_{gj} is defined as the expression level for gene g and

sample i . Group 1 and 2 have n_1 and n_2 genes, respectively. Comparing with standard two sample t-statistic for equal variance, the test statistic is the same as

$$\frac{r_g}{s_g} .$$

It is a problem with low expression levels genes. That is, variance in $\frac{r_g}{s_g}$ can

be high because of small values of s_g . But in order to compare values of d_g across

all genes, the distribution of d_g should be independent of the gene expression level.

Thus, SAM adds s_0 in the denominator to ensure that the variance of d_g is

independent of gene expression level. The value for s_0 is chosen to minimize the

coefficient of variation. Rank all genes from small to large by d_g and denote new

arrangements as $d_{(g)}$. In other words, $d_{(g)}$ is the g -th smallest relative difference.

To identify differentially expressed genes, a scatter plot of the observed relative difference $d_{(g)}$ vs. the expected relative difference $\bar{d}_{(g)}^*$ is used. The definition of

the expected relative difference $\bar{d}_{(g)}^*$ is as follows. Take B sets of permutations of

the samples, and re-calculate a new “relative difference” d_g^{*b} for each permutation b .

Obtain the corresponding order statistics $d_{(g)}^{*b}$ by ranking d_g^{*b} from small to large

for each permutation b . For each permutation b , estimate the expected order statistics

$$\text{by } \bar{d}_{(g)}^* = \frac{1}{B} \sum_{b=1}^B d_{(g)}^{*b} .$$

In the scatter plot mentioned above, each points represents a specify gene.

Choosing an adequate value as threshold Δ , the genes apart from the $d_{(g)} = \bar{d}_{(g)}^*$

line by a distance greater than the threshold Δ are regarded as differentially

expressed genes. Using the samr package in R, the differentially expressed genes can

be identified by giving a threshold Δ .

EBarrays

EBarrays package is implemented in the Bioconductor which is an open source and open development software project for the analysis and comprehension of genomic data (<http://www.bioconductor.org/>). EBarrays is an empirical Bayes analysis for identifying differentially expressed genes between two or among more than two conditions. The models attempt to describe the probability distribution of a set of expression measurements taken on a gene g , and select differentially expressed genes by posterior probability of expression pattern, which is computed for each gene and for each pattern. For more details on the methodology, see Newton *et al.* (2001), Kendziorski *et al.* (2003) and Newton and Kendziorski (2003).

Measurements are considered as arising from an observation distribution $f_{obs}(\cdot | \mu_g)$, where μ_g is a gene-specific mean value. The number of mean expression patterns possible depends on the number of conditions in a microarray experiment. For example, with a typical two conditions experiment, there are two possible patterns of expression - equivalent expression and differential expression between the two conditions. With three conditions, there are five possible patterns among the means. One pattern is equivalent expression across all conditions, and one pattern is distinct expression in each condition. Notice that different conditions may be sharing a common mean expression level, thus there are three patterns for altered expression in just one condition.

Suppose in the general case of I arrays including N conditions, there are $m+1$ possible distinct patterns. For gene g , $\mathbf{d}_{\sim g} = (d_{\sim g1}, \dots, d_{\sim gN})$ denotes the data vector where the measurements among the same condition cluster together. For any pattern k , the expression measurements sharing the common mean expression level group into the same subset. Thus, the N conditions are partitioned into $r(k)$ mutually exclusive

and exhaustive subsets $\{S_{t,k}; t = 1, 2, \dots, r(k)\}$. Assume that measurements sharing a common mean expression level μ_g arise independently and identically from an observation component $f_{obs}(\cdot | \mu_g)$, and μ_g arise from some genomewide distribution $\pi(\mu_g)$. Two parametric forms, Gamma-Gamma and Lognormal-Normal models, are considered later. Denote $f(d_{\sim g, S_{t,k}})$ as the pdf for the data indexed by subset $S_{t,k}$.

$$f(d_{\sim g, S_{t,k}}) = \int \left(\prod_{s \in S_{t,k}} f_{obs}(d_{g,s} | \mu_g) \right) \pi(\mu_g) d\mu_g$$

The pattern specific predictive density for pattern k is given by

$$f_k(d_{\sim g}) = \prod_{t=1}^{r(k)} f(d_{\sim g, S_{t,k}})$$

where $k=0$ denotes the null hypothesis which is equivalent expression among all conditions. For each gene, discrete mixing parameters p_k , $k=1, \dots, m+1$ are introduced to denote the unknown probabilities of expression pattern k, and describe the marginal distribution of the data by a mixture of the form

$$\sum_{k=0}^m p_k f_k(d_{\sim g})$$

The posterior probability of expression pattern k is then

$$P(k | d_{\sim g}) \propto p_k f_k(d_{\sim g})$$

and the posterior odds in favor of pattern k is

$$odds_{g,k} = \frac{p_k}{1-p_k} \frac{f_k(d_{\sim g})}{1-f_k(d_{\sim g})}$$

The authors consider two particular distributional forms of the general mixture model described above. The way to specify the model is determined by the choice of

observation component $f_{obs}(\cdot | \mu_g)$ and mean component $\pi(\mu_g)$.

Gamma-Gamma model (GG model):

Assume that the observation component $f_{obs}(\cdot | \mu_g)$ is a gamma distribution having shape parameter $\alpha > 0$ and scale parameter $\lambda = \frac{\alpha}{\mu_g}$ for measurements greater than zero, and assume a constant coefficient of variation $\frac{1}{\sqrt{\alpha}}$ in this distribution. They take the mean component $\pi(\mu_g)$ to be an inverse gamma, i.e. the quantity $\lambda_g = \frac{\alpha}{\mu_g}$ has a gamma distribution with shape parameter α_0 and scale parameter ν . Thus three parameters are involved in GG model, $\theta = (\alpha, \alpha_0, \nu)$.

Lognormal-Normal model (LNN model):

Assume that the observation component $f_{obs}(\cdot | \mu_g)$ is a log-normal distribution with mean μ_g and variance σ^2 , and assume a constant coefficient of variation on the raw scale in this distribution. A conjugate prior for the μ_g is normal with mean μ_0 and variance τ_0^2 . Thus three parameters are involved in LNN model, $\theta = (\mu_0, \sigma^2, \tau_0)$.

The optimal procedure to classify genes into certain expression pattern is according to the state favored by the posterior probabilities. In general, in a typical two conditions experiment, genes with posterior probability of differential expression pattern greater than 0.5 are identified as the most likely differentially expressed genes (Kendziorski *et al.*, 2007). For more details on the methodology, see Newton *et al.* (2001), Kendziorski *et al.* (2003) and Newton and Kendziorski (2003).

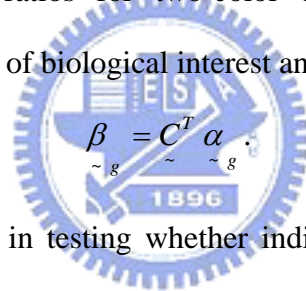
limma

limma package is implemented in the Bioconductor for differential expression

analysis of data arising from single channel such as Affymetrix and long-oligos or two channel such as cDNA microarray experiments. The central idea is to fit a linear model to the expression data for each gene (Smyth, 2004). The linear model for gene g is:

$$E(\tilde{y}_g) = X \tilde{\alpha}_g,$$

where \tilde{y}_g contains the expression data for the gene g , X is the design matrix, and $\tilde{\alpha}_g$ is a vector of coefficients. This model is specified by the design matrix X . If we have a set of I microarrays in our experiment, the response vector of the linear model is $y_g^T = (y_{g1}, \dots, y_{gI})$ for gene g . The responses will usually be log-intensities for single channel data or log-ratios for two-color data. Certain contrasts of the coefficients are assumed to be of biological interest and these are defined by



$$\tilde{\beta}_g = C^T \tilde{\alpha}_g.$$

In general, we are interested in testing whether individual contrast values β_{gj} are equal to zero. For example, with a three conditions experiment, if we concern whether there are difference between condition 1 and 2 and between condition 2 and 3 respectively, we may set the design matrix X and the contrast matrix C as follows

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{bmatrix}.$$

Then, test the hypotheses $\beta_{g1} = 0$ and $\beta_{g2} = 0$ individually.

The basic statistic used for hypothesis test with respect to a certain contrast β_{gj} is the moderated t-statistic in which posterior residual standard deviations are used in

place of ordinary standard deviations. They use the empirical Bayes approach to shrink the estimated sample variances towards a common value, resulting in far more stable inference for small numbers of arrays. Additionally, they proposed an alternative statistic, called B-statistic which is log posterior odds that the gene is differentially expressed. The posterior odds are in terms of a moderated t-statistic. The B-statistic is monotonic increasing in the moderated t-statistic under some conditions. Even if these conditions do not hold, the two statistics will rank the genes in very similar order. To test hypotheses about all contrasts simultaneously, a moderated F-statistic which is appropriate quadratic forms of moderated t-statistic is used.

2.6 Datasets

Our purpose is to evaluate which combination of preprocessing and differential expression methods performs well. We attempt to evaluate both validity and reliability of these combinations. To properly compare the combinations in terms of validity, we request that the truth differentially expressed genes of the dataset must be known. One kind of microarray experiment is called “spike-in experiments”, that is, some gene fragments have been added at known concentrations. These genes are called spike-in genes. To evaluate the validity, we choose three spike-in datasets, human genome U95 dataset from Affymetrix, human genome U133 dataset from Affymetrix, and a wholly defined control spike-in dataset (Choe *et al.*, 2005). To properly compare the method combinations in terms of reliability, we use a dataset which was generated using samples from rats and these samples are averagely distributed to different test sites (Guo *et al.*, 2006). We use four datasets in total, and describe all briefly as follow.

Affymetrix human genome U95 dataset (HGU95)

The human data set with array type HG-U95A consist of a series of genes spiked-in at known concentrations and arrayed in a format analogous to cyclic Latin

Square format. But there is still a little different from cyclic Latin Square. They represent a subset of the data used to develop and validate the Affymetrix Microarray Suite (MAS) 5.0 algorithm.

A standard 14×14 cyclic Latin Square design must consist of 14 gene groups in 14 experimental groups. Each gene group contains only one spike-in gene, and each experimental group contains the same 14 spiked-in gene groups but spiked-in at different concentrations. For example, the concentration of the 14 gene groups in the first experimental group is 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1024pM. Each subsequent experimental group rotates the spike-in concentrations by one group; i.e. experimental group 2 begins with 0.25pM and ends at 0pM, on up to experimental group 14, which begins with 1024pM and ends with 512pM. Except for the 14 spike-in genes, a common background cRNA have been added at all arrays.

The Affymetrix human genome U95 dataset contains 14 human genes in each of 14 experimental groups. Most groups contain 1 gene. Exceptions are group 1, which contains 2 genes, and group 12, which is empty. Specifically, transcript 407_at listed as present in group 12 is actually included in group 1 (together with 37777_at). For more comprehensible, we show the details in Table 3. The columns represent the 14 spiked-in gene groups and the rows represent the 14 experimental groups. The first row shows the gene name in each gene group.

Most experimental groups contain 3 replicates, except that the 3rd experimental group contain only 2 replicates and both the 13th and 14th experimental group contain 12 replicates. Replicates within each group result in a total of 59 arrays. This dataset is available at http://www.affymetrix.com/support/technical/sample_data/datasets.affx

Some researchers reported that there are 16 spike-in probesets in this dataset as opposed to the 14 originally described by Affymetrix (Cope *et al.*, 2004). The two additional genes are "33818_at" and "546_at". They claimed that "33818_at" has the

pattern of gene group 12 missing from the Latin Square, agreed by three methods of calculating expression (RMA, MAS 5.0, dChip). Wolfinger and Chu (2002) identified this as well. They also claimed "546_at" should be considered with the same concentration as "36202_at" in gene group 9, since it has pattern the same as "36202_at", as shown by three methods. Wolfinger and Chu (2002) identified this as well. Due to the competitive preprocessing methods we choose are not merely the three methods, recognizing the two genes as spike-in genes maybe not advisable. For this reason, we recognize the 14 genes originally described by Affymetrix as the entire spike-in genes.

Affymetrix human genome U133 dataset (HGU133)

This dataset with a particular array type HG-U133A_tag consist of more genes spiked-in at known concentrations and arrayed in a cyclic Latin Square format. The dataset is expected to be useful for the development and comparison of expression analysis methods. Distinct from the HGU95 dataset above, this data set includes many more spikes, and a smaller concentration spike (0.125pM).

This dataset consists of 14 spiked-in gene groups in 14 experimental groups. Distinct from the HGU95 dataset above, each gene group contains three spike-in genes. Thus there are 42 spike-in genes in total in this dataset. Each experimental group contains the same 42 spiked-in genes, but the genes in different gene group are spiked-in at different concentrations. For example, the concentration of the 14 gene groups in the first experimental group is 0, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, and 512pM. Each subsequent experimental group rotates the spike-in concentrations by one group; i.e. experimental group 2 begins with 0.125pM and ends at 0pM, on up to experimental group 14, which begins with 512pM and ends with 256pM. For more comprehensible, we show the details in Table 4.

The same as HGU95 dataset, all arrays have a common background cRNA except for the 42 spike-in genes. Each experimental group contains 3 replicates, and replicates within each group result in a total of 42 arrays. This dataset is available at http://www.affymetrix.com/support/technical/sample_data/datasets.affx .

A wholly defined control spike-in dataset

Due to the vast numbers of genes interrogated in a microarray experiment, only a relatively small fraction of gene expression differences tend to be validated in any given study. Choe *et al.* (2005) generated a new control dataset for the purpose of evaluating methods for identifying differentially expressed genes between two sets of triplicated hybridizations to Affymetrix GeneChips. The two sets are called spike-in samples and control samples, resulting in a total of 6 arrays. This dataset has three main features to facilitate the relative assessment of different analysis options. First, this experiment has 1331 spike-in genes spiked-in at known relative concentrations between the spike-in and control samples. The dataset has a larger fraction of gene expression differences than the general spike-in datasets. Second, this experiment used a defined background sample of 2535 genes presented at identical concentrations in both spike-in and control samples, rather than a biological RNA sample of unknown composition. Third, this dataset includes lower fold changes, beginning at only a 1.2-fold concentration difference to 4-fold concentration difference. This dataset is available at <http://www.ccr.buffalo.edu/halfon/spike/index.html>.

Here, we give a summary table for the three spike-in datasets in Table 2.

Rat dataset

The dataset we used is just a part of the complete dataset from a rat toxicogenomic study, which is one of the reference datasets of MAQC (MicroArray Quality Control) project

(<http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/>). The purpose of

the MAQC project is to provide quality control tools to the microarray community in order to avoid procedural failures and to develop guidelines for microarray data analysis by providing the public with large reference datasets along with readily accessible reference RNA samples. The rat toxicogenomic dataset was generated using 36 RNA samples from rats treated with three chemicals (aristolochic acid, riddelliine and comfrey). In total there were six treatment/tissue groups: kidney from aristolochic acid-treated rats (K_AA), kidney from vehicle control (K_CTRL), liver from aristolochic acid-treated rats (L_AA), liver from riddelliine-treated rats (L_RDL), liver from comfrey-treated rats (L_CFY) and liver from vehicle control (L_CTRL). Within each treatment/tissue group there were six biological replicates. Aliquots of these samples were prepared and distributed to each of the test sites for gene expression profiling using microarrays from four different platforms (Affymetrix, Agilent, Applied Biosystems and GE Healthcare). There are two test sites using Affymetrix platform, and we adopt only the data from the two test sites. Each test site generated 36 arrays respectively. In this paper, when we refer to the Rat dataset, it denotes the 72 arrays in all which were generated from the two sites using Affymetrix platform. This dataset is available at

<http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/> .

3 Materials and Methods

Our purpose is to evaluate which combinations of preprocessing and differential expression methods perform well. Specifically, we compare the combinations according to two main criteria, the validity and the reliability of the combinations. First, we select datasets having some particular properties in terms of different criteria and some interested preprocessing method is used on the datasets to summary the probe set measurements. Then, these measurements of genes are performed by some interested differential expression method, and the likely differentially expressed genes chosen by certain combination of preprocessing and differential expression method are listed. Based on the list of differentially expressed genes, we can evaluate the validity and the reliability of the combination. We divide the assessment of validity and reliability into two Sections 3.2 and 3.3 respectively in detail.

3.1 Implementation of methods selected

There are four preprocessing methods and five differential expression methods applied to each of the datasets we selected. Three statistical models, MAS 5.0, dChip and RMA, and one physical model, PDNN, are considered. The five differential expression methods are fold-change, two sample t-test, SAM, EBarrays and limma. A total of 35 combinations are resulted. We may regard each criterion of methods as “score” to express the level of significance. The higher the score, the more significant the result.

3.1.1 Four preprocessing methods used

MAS 5.0

MAS 5.0 (Microarray Suite software, Version 5.0) is offered by Affymetrix

(Affymetrix, 2002). Each probe including PM and MM must be preprocessed for background adjustment according to its location on the array. To avoid obtaining a negative value when subtracting MM from PM, MAS 5.0 introduces the concept of an Ideal Mismatch (IM) derived from MM and never bigger than its PM. The expression level is defined as the anti-log of a robust average (Tukey biweight) of the value $\{\log_2(PM_{jg} - IM_{jg})\}$ where PM_{jg} and IM_{jg} represent the PM and IM intensities for j-th probe pair of gene g. Finally, the expression level is scaled using a trimmed mean. We apply the absolute analysis of MAS 5.0 in R.

dChip (including dChip(PM-MM) and dChip(PM-only))

Li and Wong (2001a) proposed a Model Based Expression Index model (MBEI) where multiple arrays are available to estimate the expression levels. For any given gene, the model is defined as follows:

$$\begin{aligned} MM_{ij} &= \nu_j + \theta_i \alpha_j + \varepsilon \\ PM_{ij} &= \nu_j + \theta_i \alpha_j + \theta_i \phi_j + \varepsilon = \nu_j + \theta_i (\alpha_j + \phi_j) + \varepsilon \end{aligned} \quad (1)$$

where PM_{ij} and MM_{ij} denote the PM and MM intensity values from the i-th array and the j-th probe pair for this gene. θ_i denotes the expression index for this gene in the i-th array. α_j and ϕ_j represent the increasing rate of intensity value of the MM_{ij} probe and the additional increasing rate in the corresponding PM_{ij} probe respectively. ν_j is the baseline response of the j-th probe pair due to nonspecific hybridization, and ε are assumed to be independent normally distributed errors. Two methods based on the model above are developed: (1) subtracting MM from PM intensities (Li and Wong, 2001a) (2) using PM intensities only (Li and Wong, 2001b). Li and Wong's measure is defined as the maximum likelihood estimates of the expression index θ_i and the estimation procedure includes rules for outlier removal.

“Invariant Set” normalization method is used to normalize arrays at PM and MM probe levels.

RMA

Irizarry *et al.* (2003a) developed a log scale linear additive model using only PM probes, it is also known as RMA (Robust Multi-array Analysis). For any given gene, it is described as $T(PM_{ij}) = e_i + a_j + \varepsilon_{ij}$, where PM_{ij} is the PM intensity of array i and probe pair j for this gene. $T(\cdot)$ represents the transformation that background corrects, normalizes by quantile normalization, and logs the PM intensities. The three terms on the right represent the log₂ scale expression value for this gene of array i , the log scale affinity effects for probe j , and error respectively (Irizarry *et al.*, 2003b). To protect against outlier probes, a robust procedure such as median polish is used to estimate model parameters and the log scale measure of expression level e_i .

PDNN

Zhang *et al.* (2003) proposed a simply free energy model, called “position-dependent nearest-neighbor (PDNN) model”. Different from most methods focused on statistical models such as the methods introduced above, it is a physical model taking into account the sequence of nearest-neighbors (adjacent two bases) and the position of these nucleotide pairs. In the PDNN model, the signal of a probe is divided into three components: gene-specific binding, non-specific binding, and uniform background. And the free energy of the two bindings of a probe can be expressed as a weighted sum of its stacking energies (Sugimoto *et al.*, 1995), where the stacking energies depend on the sequence of nearest-neighbors and the weights depend on the position along the probe. Further technical details can be found in Zhang *et al.* (2003).

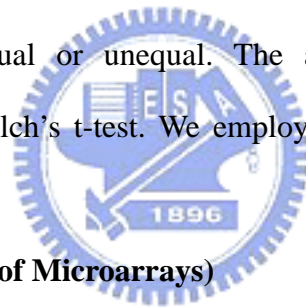
3.1.2 Five differential expression methods used

Fold-change (FC)

Fold-change is the most commonly used method of detecting differentially expressed gene between two compared condition samples. For any given gene, fold-change is calculated by the probe set intensity ratio of two compared condition samples. If there are replicates, we usually average across the samples for each condition in advance. Then the ratio of these averaged values is referred as fold-change. Fold-change is employed as the score of significance.

Two sample t-test (including t-test and Welch t-test)

The simplest statistical method for comparing means between two groups is two sample t-test. When carrying out a two sample t-test, the variances of the two samples may be assumed to be equal or unequal. The approach of unequal variance assumption is also called Welch's t-test. We employ minus p-value as the score of significance.



SAM (Significance Analysis of Microarrays)

It was proposed by Tusher, Tibshirani and Chu (2001). The method is based on a modified version of the standard t-statistic to adjust the high variance probably caused by a low expression level. For each gene g , the “relative difference” d_g in gene expression is defined as the form which adds an exchangeability factor to the denominator of the standard two sample t-statistic for equal variance. Exchangeability factor is added to ensure that the variance of d_g is independent of gene expression level. Rank all genes by the observed relative difference d_g and denote the new arrangements as $d_{(g)}$. B sets of permutations of the samples are taken to obtain the expected relative difference $\bar{d}_{(g)}^*$ by a similar way (For more details, see Tusher *et*

al., 2001 and Chu *et al.*). A scatter plot of $d_{(g)}$ vs. $\bar{d}_{(g)}^*$ is used and the genes apart from the $d_{(g)} = \bar{d}_{(g)}^*$ line by a distance greater than the threshold Δ are regarded as differentially expressed genes.

Using the samr package in R, the differentially expressed genes can be identified by giving a threshold Δ . But the number of genes selected is determined by the given threshold Δ , we can not set at will. And further filtering criteria which are not mentioned in the original paper (Tusher *et al.*, 2001) are carried out. Thus, we give up using the samr package, and employ the difference between $d_{(g)}$ and $\bar{d}_{(g)}^*$ as the score of significance according to the methodology referred in Tusher *et al.*(2001).

EBarrays (including of EBarrays(GG) and EBarrays(LNN))

An empirical Bayes analysis, implemented in Bioconductor EBarrays package, attempt to describe the probability distribution of expression levels for gene g and select differentially expressed genes by posterior probability of differential expression. Two mixture models, Gamma-Gamma model and lognormal-normal model, are considered according to their sampling and prior distributions. For more details on the methodology, see Newton *et al.* (2001), Kendzierski *et al.* (2003), and Newton and Kendzierski (2003). We employ the posterior probability of differential expression as the score of significance.

limma

Smyth (2004) proposed a method of linear models and empirical Bayes methods which is implemented in the Bioconductor limma package (Smyth, 2005). The linear model for gene g is $E(y_{\sim g}) = X_{\sim g} \alpha_{\sim g}$, where $y_{\sim g}$ is the expression level vector of I arrays in total for this gene, $X_{\sim g}$ is the design matrix, and $\alpha_{\sim g}$ is a vector of coefficients. Certain contrasts of the coefficients are assumed to be of biological

interest and these are defined by $\beta_{\sim g} = C^T \alpha_{\sim g}$. In general, we are interested in testing whether individual contrast values β_{gj} are equal to zero. The basic statistic with respect to a certain contrast β_{gj} is the moderated t-statistic in which posterior residual standard deviations are used in place of ordinary standard deviations by empirical Bayes approach. Alternative statistic, called B-statistic, represents log posterior odds that the gene is differentially expressed. The default argument in limma package is B-statistic and we employ it as the score of significance.

3.2 Assessment of validity

To properly compare the combinations in terms of validity, we request that the true differentially expressed genes of the dataset must be known. Thus, we choose three datasets which provide the results of spike-in experiments where gene fragments have been added at known concentrations. The three datasets are human genome U95 dataset from Affymetrix, human genome U133 dataset from Affymetrix, and a wholly defined control spike-in dataset (Choe *et al.*, 2005). The three datasets provide various number of spike-in genes. ROC curves are used for the evaluation. We describe the three datasets briefly as follow.

Affymetrix human genome U95 dataset (HGU95)

This dataset was used to develop and validate MAS 5.0 algorithm. It consists of 59 arrays, where 14 different cRNA gene fragments have been spiked-in at various known concentrations ranging from 0.25 to 1024pM. Except for the 14 spike-in genes, a common background cRNA have been added at all arrays. The 14 spike-in genes are arranged in the format similar to a 14×14 cyclic Latin square design with each concentration appearing once in each row and column. The difference from a 14×14 cyclic Latin square design is that there are two out of the 14 spike-in genes spiked-in

at the same concentrations across arrays (Table 3). Most experimental groups contain 3 replicates, except that the 3rd experimental group contains only 2 replicates and both the 13th and 14th experimental group contain 12 replicates. For more details, see Affymetrix website and this dataset is available here

http://www.affymetrix.com/support/technical/sample_data/datasets.affx.

Affymetrix human genome U133 dataset (HGU133)

Distinct from the HGU95 dataset above, this dataset includes many more spikes, and a smaller concentration spike (0.125pM). This dataset consists of 14 gene groups in 14 experimental groups. A cyclic Latin Square format is designed for each gene group and experimental group. Each gene group containing three spike-in genes and each experimental group containing 3 replicates result in a total of 42 spike-in genes and 42 arrays. For more details, see Affymetrix website and this dataset is available here

http://www.affymetrix.com/support/technical/sample_data/datasets.affx

A wholly defined control spike-in dataset (Golden Spike)

Choe *et al.* (2005) generated a new control dataset which contains two sets of triplicated hybridizations to Affymetrix GeneChips. The two sets are called spike-in samples and control samples respectively, resulting in a total of 6 arrays. This dataset has three main features: (1) 1331 spike-in genes spiked-in at known relative concentrations between the spike-in and control samples, a larger fraction of gene expression differences. (2) a defined background sample of 2535 genes presented at identical concentrations in both spike-in and control samples, rather than a biological RNA sample of unknown composition. (3) a lower fold changes beginning at only a 1.2-fold concentration difference. This dataset is available at

<http://www.ccr.buffalo.edu/halfon/spike/index.html> .

Methodology of comparison

In order to evaluate the validity of these combinations, a receiver operating characteristic curve (simply called ROC curve) is used. ROC curve, which is widely used to evaluate the differential expression methods in microarray analysis, is a graphical plot of the sensitivity versus 1-specificity for a binary classifier system as its discrimination threshold is varied. Sensitivity and specificity are statistical measurements of how well a binary classification test correctly identifies the truth. Sensitivity is defined as the probability that the test lead to make positive decision given that the truth is actually a positive case. This is also known as the true positive rate (TPR). And specificity is defined as the probability that a negative decision is made when the truth is negative. In other words, 1-specificity represents that the probability that the positive decision is made when the truth is negative, and the meaning is equivalent to the false positive rate (FPR). For most differential expression methods, null hypothesis is usually defined as gene expressed equally under two different conditions. The four outcomes of a test can be formulated as the following table.

TP : true positive FP : false positive
 FN : false negative TN : true negative

TPR : true positive rate (sensitivity)
 FPR : false positive rate (1-specificity)

	Null hypothesis H_0 (non-differentially expressed)	
	False	True
Reject H_0 (Called significant)	TP ($1 - \beta$)	FP (α)
Not reject H_0 (Not called significant)	FN	TN

Thus, the ROC curve is represented equivalently as a plot of the false positive (FP) rate as the x coordinate versus the true positive (TP) rate as the y coordinate. It provides tools to select possibly optimal methods by comparing the area under ROC curve (simply called AUC). The area measures discrimination, that is, the ability of the test to correctly classify those positive case and negative case in fact. The range of AUC is from 0 to 1 since both the x and y axes have values ranging from 0 to 1. The bigger its AUC is, the better overall performance of this test. We take advantage of ROC curve and AUC as criteria to assess the validity of different combinations.

Here we make a brief description of how to accomplish an average ROC curve for a selected combination of some dataset, preprocessing method, and differential expression method. For each spike-in dataset, spike-in genes are considered as true positives and non-spike-in genes as true negatives. For each dataset, different experimental groups imply that the spike-in genes are spiked-in at different concentrations. Thus, only replicates are regarded as being in the same experimental group. For each pair of experimental groups, we compute the number of true positive (TP) and false positive (FP) for a large range of thresholds. To form an average ROC curve, we compute the average TP according to each FP value. An average ROC curve is created by plotting the FP versus its average TP. And the area under average ROC curve is the measure of this combination (Cope *et al.*, 2004).

3.3 Assessment of reliability

To properly compare the method combinations in terms of reliability, we use a particular dataset which was generated using samples from rats and these samples are averagely distributed to different test sites (Guo *et al.*, 2006). Overlap rates between two test sites using Affymetrix platform are compared.

Rat dataset

The dataset we used is just a part of the complete dataset from a rat toxicogenomic study (Choe *et al.*, 2005), which was generated using 36 RNA samples from rats treated with three chemicals (aristolochic acid, riddelliine and comfrey). In total there are six treatment/tissue groups: kidney from aristolochic acid-treated rats (K_AA), kidney from vehicle control (K_CTRL), liver from aristolochic acid-treated rats (L_AA), liver from riddelliine-treated rats (L_RDL), liver from comfrey-treated rats (L_CFY) and liver from vehicle control (L_CTRL). Within each treatment/tissue group, there are six biological replicates. Aliquots of these samples were prepared and distributed to each of five test sites. Each test site generated 36 arrays respectively. We adopt only the partly data which come from the two test sites using Affymetrix platform. In this paper, when we refer to the Rat dataset, it denotes the 72 arrays in all which were generated from the two sites using Affymetrix platform. This dataset is available at <http://www.fda.gov/nctr/science/centers/toxicoinformatics/magc/>.

Methodology of comparison

In order to compare the reliability of these combinations, we plot graphs where x-axis represents the number of genes selected as differentially expressed genes and y-axis represents the overlap rate of two gene lists for a given number of differentially expressed genes. For example, when we employ two sample t-test as differential expression method and use p-value 0.05 as threshold, two gene lists according to the two test sites are produced respectively by collecting the genes which have p-value smaller than 0.05. The numerator of overlap rate is defined as the number of overlapping genes for both two gene lists, and the denominator of overlap rate is defined as the total number of genes in the union of two gene lists. Thus, if there are genes $\{a, b, c, d, e\}$ have p-value smaller than 0.05 for the first test site and genes $\{c, d, e, f\}$ have p-value smaller than 0.05 for the second test site. Overlap rate is

calculated by $\frac{3}{6} = 0.5$.

Four graphs are created respectively according to each of the four tissues suffering different treatments versus their controls, and an average graph is created for summarizing the four conditions. For given data from the same test site and employed competitive combination, genes are ranked by the “score” of significance of the employed competitive combination referred in Section 3.1.2. For a fixed threshold of the “score”, two significant gene lists from the two test sites are produced respectively. The overlap rates of the two gene lists are computed for a large range of threshold of the “score”. Competitive combinations are showed as lines in the graph and their overlap rates between two test sites are compared. The higher overlap rate, the better performance in reliability.



4 Results

4.1 Assessment of validity by ROC curve

In practice, we rarely validate more than 100 genes as differentially expressed genes (Cope *et al.*, 2004). Under both HGU95 and HGU133 datasets, the growth in TP of most combinations has already become flat gradually after $FP > 100$ (Figure 1-1 and Figure 1-3). Furthermore, in these two datasets, true positive rates of the best performance have reached about 95% when $FP < 100$. As for these combinations performing worst, their true positive rates increase after $FP > 100$; even so, their performances still can not catch up to these combinations which perform well before $FP < 100$ (Figure 1-2 and Figure 1-4). Thus, we focus on the part of $FP < 100$ and report the summary statistic AUC up to 100 FP in both HGU95 and HGU133 datasets.

The other spike-in dataset, Golden spike dataset, unlike most microarray experiments assuming a small percentage of genes are differentially expressed, has nearly 10% genes differentially expressed. Considering the situation of FP less than 100 to evaluate performance is not suitable for this dataset. The patterns of all combinations where false positive rate larger than 0.1 are similar to the patterns where false positive rate close to 0.1 (Figure 1-5 and Figure 1-6). Thus, we recommend a conservative choice 0.1 as a cutoff of false positive rate.

Two algorithms of all combinations can not be executed in R and we have no information about their performance. That are HGU133 + dChip(PM-MM) + EBarrays(GG) and HGU133 + PDNN + EBarrays(GG). Thus, there are 35 combinations for both HGU95 and Golden Spike datasets, and only 33 combinations for HGU133 dataset.

We use AUC up to 100 FP in HGU95 and HGU133 datasets and up to 0.1 false positive rate in Golden spike dataset as assistants. For each dataset, all combinations

are ranked based on AUC. If there is distinct difference of AUC between two continuously ranked combinations, combinations are apart from there and divided into two groups. By this way, total combinations are clustered in four groups for HGU95 dataset, four groups for HGU133, and three for Golden spike dataset shown in Table 5, 6, 7 respectively.

For HGU95 dataset

Under HGU95 dataset, (1)RMA or PDNN cooperated with most differential expression methods have excellent performances, except for Welch t-test employed as differential expression method (Figure 1-2). (2)Conversely, the combinations of preprocessing method using MAS 5.0 or dChip(PM-MM) are inferior to other compared combinations (Figure 1-2), and the combinations in the group with smallest AUC is entirely composed by MAS 5.0 and dChip(PM-MM) as preprocessing method (Table 5). (3)As long as using Welch t-test as differential expression method, the performance is not good enough even if cooperated with RMA or PDNN (Figure 2-1). (4)For a fixed differential expressed method, performances vary largely by employing different preprocessing methods, except for t-test and Welch t-test (Figure 3-1). And all combinations using t-test outperform than using Welch t-test.

For HGU133 dataset

Results in HGU133 are very similar to HGU95. (1)~(3) conclusions are shown in HGU133 as well (Figure 1-4 , Table 6 and Figure 2-2). The different result is that the performances vary largely by employing different preprocessing methods for each differential expressed method.

For Golden Spike dataset

Results in Golden Spike dataset are unlikeness to two datasets above. (1)Instead of RMA and PDNN, dChip have outstanding performances applied to this dataset. Through viewing Figure 1-6, all combinations are divided into three groups clearly.

There are 11 combinations classified into the outstanding group, and all of them combine with dChip, especially for dChip(PM-only) that cooperated with every differential expression methods are contained. But dChip(PM-MM) has extreme performance. When it is cooperated with fitting differential expression method, such as t-test, Welch t-test, limma and SAM, the performance will be outstanding. On the contrary, it will perform disappointingly (Figure 2-3). (2)The following 7 combinations are the worst, MAS5.0+SAM, MAS5.0+FC, MAS5.0+ EBarrays(GG), MAS5.0+ EBarrays(LNN), dChip(PM-MM)+FC, dChip(PM-MM)+ EBarrays(GG), and dChip(PM-MM)+ EBarrays(LNN) (Table 7). Notice that, for all of the three datasets, the five combinations, MAS5.0+FC, MAS5.0+ EBarrays(GG), MAS5.0+ EBarrays(LNN), dChip(PM-MM)+FC, and dChip(PM-MM)+ EBarrays(LNN), are classified into the worst group clustered by AUC.

4.2 Assessment of reliability by overlap rate

For this dataset, the true number of differentially expressed genes is unknown. We show the patterns of all combinations in log scale in Figure 4, and find that the trend of most of combinations is similar when the number of genes selected as differentially expressed is less than 10000. Moreover, if there are too many genes identified as differentially expressed genes, a much lower threshold of “score” of significance of the differential expression method must be set. But it is not a practical threshold. Thus, our comparison in reliability focuses on the value of x-axis less than 10000. Here, the four tissues suffering different treatments versus their controls are simply called as K_AA, L_AA, L_CFY, and L_RDL.

Low overlap rate for MAS 5.0 and dChip(PM-MM)

For each condition, K_AA, L_AA, L_CYF, and L_RDL, combinations are divided into five small graphs by preprocessing method such as Figure 5-1~5-4.

Figure 5-1~5-3 show that the overlap rates across two sites are lower than 0.6 when using MAS 5.0 or dChip(PM-MM), but higher overlap rates occur for three other preprocessing methods, RMA, PDNN and dChip(PM-only). For L_RDL, overlap rates exceed 0.6 when using MAS 5.0 or dChip(PM-MM), but that is caused by overall improvement of overlap rate for L_RDL, not for MAS 5.0 or dChip(PM-MM) only (Figure 5-4).

Performances for EBarrays

For each preprocessing method cooperated with EBarrays, very similar patterns under Gamma-Gamma model and Lognormal-Normal model are shown (Figure 6). Usually, when using EBarrays, there is no overlap gene when small genes selected as differentially expressed but a rapidly increment in overlap rate happens when differentially expressed genes increase to some level. The level varies with different preprocessing method, usually MAS 5.0 and dChip(PM-MM) have lower level and the others have a higher level. However, even if a rapidly increment happens, the performance is still not good enough when compared to other combinations that perform well. The feature above can be saw by Figure 5-1~5-4.

Top 2 combinations

Now we assign the same color to combinations using the same differential expression method in Figure 7, most lines are clustered by color obviously. The performance is worse when using t-test (green) or Welch t-test (blue), and is better when using FC (black). SAM and limma perform well when fewer genes selected as differentially expressed.

Because of the poor performances with MAS5.0, dChip(PM-MM), t-test, Welch t-test and EBarrays, we consider totally 9 permutations with RMA, dChip(PM-only), PDNN as preprocessing method and FC, SAM, limma as differential expression method in Figure 8. Figure 8 shows that the two combinations, RMA+FC (blue) and

PDNN+FC (yellow) have the highest overlap rate and nearly equal. Thus the top two combinations in reliability are RMA+FC and PDNN+FC.



5 Conclusions and Discussion

5.1 Conclusions

Validity

Notice that, the 15 combinations of the first two groups clustered by AUC for HGU95 are also contained in the first two groups for HGU133, except that the algorithm of PDNN+EBarrays(GG) can not be executed for HGU133 (Table 5, 6). Ranks of combinations are similar when using HGU95 and HGU133 with an approximate proportion of genes expressed differentially, but very different when using Golden spike dataset which has a large proportion of genes expressed differentially. Top combinations seem to be substituted according to the amount of spike-in genes in the dataset. If a high validity is required when considering an experiment with a few differentially expressed genes, we recommend RMA or PDNN as preprocessing method but are sure to avoid collocating with Welch t-test. Nevertheless for an experiment with a larger proportion of genes expressed differentially, dChip(PM-only) are recommended as preprocessing method, or dChip(PM-PM) collocated with t-test, Welch t-test, limma and SAM are recommended. No matter what dataset is used, the same five combinations have the lowest validity, that is, MAS5.0+FC, MAS5.0+EBarrays(GG), MAS5.0+EBarrays(LNN), dChip(PM-MM)+FC, and dChip(PM-MM)+EBarrays(LNN).

Giving an overview of the three spike-in datasets, we assign the same color by preprocessing method in Figure 1-2, Figure 1-4, and Figure 1-6. Combinations of the same color are slightly clustered together. But when we assign the same color by differential expression method in Figure 3-1 ~ Figure 3-3, colors are in a disorderly behavior. It seems that preprocessing method influences the validity more than differential expression method.

Reliability

Actually, the patterns of the four conditions are similar, so we introduce an average graph that facilitates to compare all combinations. We assign the same color by preprocessing method in Figure 9-1 and by differential expression method in Figure 9-2, and find that differential expression method influences the reliability much more than preprocessing method because the combinations are clustered by differential expression methods. When employing FC as differential expression method, combinations have the highest overlap rates, especially for cooperated with RMA or PDNN.

Consideration to both validity and reliability

We give an overview of both validity and reliability, validity is influenced more by preprocessing method, but reliability is influenced more by differential expression method. To give consideration to both validity and reliability, six combinations are recommended when differentially expressed genes are less, RMA+FC, RMA+ SAM, RMA+ limma, PDNN+FC, PDNN+SAM, and PDNN+limma. Three combinations are recommended when differentially expressed genes are more, dChip(PM-only)+FC, dChip(PM-only)+SAM, and dChip(PM-only)+limma. However, four combinations lead to both low validity and low reliability. That are MAS5.0+ EBarrays(GG), MAS5.0+ EBarrays(LNN), dChip(PM-MM)+ EBarrays(LNN), and dChip(PM-MM)+ EBarrays(LNN). If you only focus on the simple t-test as differential expression method, the assumption of equal variance is advised because of higher accuracy and precision result.

5.2 Discussion

The strange pattern of EBarrays in Figure 6 is caused by too many genes having posterior probability of differential expression equal to 1. When ranking genes by

posterior probability, too many equal values make the order meaningless. For example, more than one thousand genes have posterior probability equal to 1 when using PDNN+EBarrays(GG) for L_CFY treatment/control. We can not select only 100 genes as differentially expressed genes in this situation. Even if using spike-in datasets, there are still too many genes having posterior probability of differential expression equal to 1. That is one disadvantage of EBarrays. And we can not find the best way to deal with genes having the equal values of score of significance.



Reference

Affymetrix. (2002) Statistical algorithms description document.

http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf

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

Choe,S.E., Boutros,M., Michelson,A.M., Church,G.M. and Halfon,M.S. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, **6**:R16

Chu,G., Narasimhan,B., Tibshirani,R. and Tusher,V. SAM “Significance Analysis of Microarrays”–Users guide and technical document. *Technical Report, Stanford University*. <http://www-stat.stanford.edu/~tibs/SAM/sam.pdf>

Cope,L.M., Irizarry,R.A., Jaffee,H.A., Wu,Z. and Speed,T.P. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323-331.

Guo,L., Lobenhofer,E.K., Wang,C., Shippy,R., Harris,S.C., Zhang,L., Mei,N., Chen,T., Herman,D., Goodsaid,F.M., Hurban,P., Phillips,K.L., Xu,J., Deng,X., Sun,Y.A., Tong,W., Dragan,Y.P. and Shi,L. (2006) Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature Biotechnology*, **24**, 1162-1169.

Huber,W. , Irizarry,R.A. and Gentleman,R. (2005) Preprocessing overview. In Gentleman,R., Irizarry,R.A., Carey,V.J., Dudoit,S. and Huber,W. (eds), *Bioinformatics and Computational Biology Solutions using R and Bioconductor*,

Springer, New York, **Chapter 1**, pp. 3-12.

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

Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003b) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, **31**, e15.

Kendziorski,C., Sarkar,D., Chen,M. and Newton,M. (2007) The vignette of EBarrays package in Bioconductor.

<http://bioconductor.org/packages/2.0/bioc/vignettes/EBarrays/inst/doc/vignette.pdf>

Kendziorski,C.M., Newton,M.A., Lan,H. and Gould,M.N. (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, **22**, 3899-3914.

Li,C. and Wong,W. (2001a) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Science*, **98**, 31-36.

Li,C. and Wong,W. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, **2(8)**, research 0032.1-0032.11.

Newton,M.A. and Kendziorski,C.M. (2003) Parametric Empirical Bayes Methods for Microarrays. In Parmigiani,G., Garrett,E.S., Irizarry,R.A. and Zeger,S.L. (eds),

The Analysis of Gene Expression Data: Methods and Software. Springer,
Chapter 11, pp. 254-271.

Newton,M.A., Kendzierski,C.M., Richmond,C.S., Blattner,F.R. and Tsui,K.W. (2001)
On differential variability of expression ratios: improving statistical inference
about gene expression changes from microarray data. *Journal of Computational
Biology*, **8**, 37-52.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing
differential expression in microarray experiments. *Statistical Applications in
Genetics and Molecular Biology*, **3**, Article 3.

Smyth,G.K. (2005) Limma: linear models for microarray data. In Gentleman,R.,
Carey,V., Dudoit,S., Irizarry,R. and Huber,W. (eds), *Bioinformatics and
Computational Biology Solutions using R and Bioconductor*, Springer, New
York, **Chapter 23**, pp. 397–420.

Sugimoto,N. *et al.* (1995) Thermodynamic parameters to predict stability of
RNA/DNA hybrid duplexes. *Biochemistry*, **34**, 11211-11216.

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 Science*, **98**, 5116–5121.

Wolfinger,R. and Chu,T.-M. (2002) Who are those strangers in the latin square?
Critical Assessment of Microarray Data Analysis ‘CAMDA 02’.

Zhang,L., Miles,M.F. and Aldape,K.D. (2003) A model of molecular interactions on
short oligonucleotide microarrays. *Nature Biotechnology*, **21**, 818–821.

Table 1. Summary of the four preprocessing methods used.

Model	Method	Background adjustment	Normalization	Summarization	Reference
Statistical model	MAS5.0	Locational adjustment & MM subtracted	Scale normalization	Tukey biweight average	Affymetrix, 2002
	dChip (PM-MM)	MM intensities are subtracted	Invariant set	Fit a model based expression index	Li and Wong, 2001a
	dChip (PM only)	PM only	Invariant set	Fit a model based expression index	Li and Wong, 2001b
	RMA	Convolution background correction	Quantile normalization	A robust linear model is fitted (median polish)	Irizarry <i>et al.</i> , 2003
Physical model	PDNN	PM only	Quantile normalization	A free energy model accounts for background and signal.	Zhang <i>et al.</i> , 2003

Table 2. Summary of the three spike-in datasets used.

Dataset	Spike-in genes / Total genes in array	Conditions	Total arrays	Replicates (conditions)	Fold change range	Reference
HGU95	14 / 12626	14	59	2 (1) , 3 (11) , 12 (2)	$2 \sim 2^{12}$	Affymetrix
HGU133	42 / 22300	14	42	3 (14)	$2 \sim 2^{12}$	Affymetrix
Golden Spike	1331 / 14010	2	6	3 (2)	1.2 ~ 4.0	Choe <i>et al.</i> , 2005

Table 3. Affymetrix human genome U95 dataset contains 14 spike-in gene groups in each of 14 experimental groups. This table shows the spiked-in concentrations (pM).

HGU95		Spike-in Gene Groups \longrightarrow													
		37777_at	684_at	1597_at	38734_at	39058_at	36311_at	36889_at	1024_at	36202_at	36085_at	40322_at	407_at	1091_at	1708_at
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Experimental Groups \downarrow	A	0	0.25	0.5	1	2	4	8	16	32	64	128	0	512	1024
	B	0.25	0.5	1	2	4	8	16	32	64	128	256	0.25	1024	0
	C	0.5	1	2	4	8	16	32	64	128	256	512	0.5	0	0.25
	D	1	2	4	8	16	32	64	128	256	512	1024	1	0.25	0.5
	E	2	4	8	16	32	64	128	256	512	1024	0	2	0.5	1
	F	4	8	16	32	64	128	256	512	1024	0	0.25	4	1	2
	G	8	16	32	64	128	256	512	1024	0	0.25	0.5	8	2	4
	H	16	32	64	128	256	512	1024	0	0.25	0.5	1	16	4	8
	I	32	64	128	256	512	1024	0	0.25	0.5	1	2	32	8	16
	J	64	128	256	512	1024	0	0.25	0.5	1	2	4	64	16	32
	K	128	256	512	1024	0	0.25	0.5	1	2	4	8	128	32	64
	L	256	512	1024	0	0.25	0.5	1	2	4	8	16	256	64	128
	M, N, O, P.	512	1024	0	0.25	0.5	1	2	4	8	16	32	512	128	256
	Q, R, S, T.	1024	0	0.25	0.5	1	2	4	8	16	32	64	1024	256	512

Table 4. Affymetrix human genome U133 dataset contains 14 spike-in gene groups in each of 14 experimental groups. This table shows the spiked-in concentrations (pM).

		Spike-in Gene Groups \longrightarrow													
		203508_at	204205_at	204836_at	207777_s_at	207160_at	209606_at	205398_s_at	206060_s_at	207641_at	203471_s_at	AFFX-r2-TagA_at	AFFX-r2-TagD_at	AFFX-r2-TagG_at	AFFX-LysX-3_at
		204563_at	204959_at	205291_at	204912_at	205692_s_at	205267_at	209734_at	205790_at	207540_s_at	204951_at	AFFX-r2-TagB_at	AFFX-r2-TagE_at	AFFX-r2-TagH_at	AFFX-PheX-3_at
		204513_s_at	207655_s_at	209795_at	205569_at	212827_at	204417_at	209354_at	200665_s_at	204430_s_at	207968_s_at	AFFX-r2-TagC_at	AFFX-r2-TagF_at	AFFX-DapX-3_at	AFFX-ThrX-3_at
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Experimental Groups \downarrow	1	0	0.125	0.25	0.5	1	2	4	8	16	32	64	128	256	512
	2	0.125	0.25	0.5	1	2	4	8	16	32	64	128	256	512	0
	3	0.25	0.5	1	2	4	8	16	32	64	128	256	512	0	0.125
	4	0.5	1	2	4	8	16	32	64	128	256	512	0	0.125	0.25
	5	1	2	4	8	16	32	64	128	256	512	0	0.125	0.25	0.5
	6	2	4	8	16	32	64	128	256	512	0	0.125	0.25	0.5	1
	7	4	8	16	32	64	128	256	512	0	0.125	0.25	0.5	1	2
	8	8	16	32	64	128	256	512	0	0.125	0.25	0.5	1	2	4
	9	16	32	64	128	256	512	0	0.125	0.25	0.5	1	2	4	8
	10	32	64	128	256	512	0	0.125	0.25	0.5	1	2	4	8	16
	11	64	128	256	512	0	0.125	0.25	0.5	1	2	4	8	16	32
	12	128	256	512	0	0.125	0.25	0.5	1	2	4	8	16	32	64
	13	256	512	0	0.125	0.25	0.5	1	2	4	8	16	32	64	128
	14	512	0	0.125	0.25	0.5	1	2	4	8	16	32	64	128	256

Table 5. Area under ROC curve (FP<100) for HGU95 dataset.

HGU95	Preprocessing	Differential expression	AUC (FP<100)
1	PDNN	limma	0.948579
2	RMA	limma	0.94818
3	PDNN	FC	0.944115
4	PDNN	SAM	0.944046
5	RMA	FC	0.943009
6	RMA	SAM	0.942039
7	RMA	EBarrays(GG)	0.927794
8	RMA	EBarrays(LNN)	0.923941
9	dChip(PM-only)	limma	0.905235
10	PDNN	EBarrays(GG)	0.902541
11	PDNN	EBarrays(LNN)	0.90131
12	PDNN	t.test	0.898442
13	RMA	t.test	0.886426
14	dChip(PM-only)	t.test	0.88254
15	dChip(PM-only)	SAM	0.880197
16	dChip(PM-only)	FC	0.846546
17	dChip(PM-MM)	t.test	0.841166
18	dChip(PM-MM)	limma	0.835926
19	dChip(PM-MM)	SAM	0.825455
20	dChip(PM-only)	EBarrays(GG)	0.824395
21	dChip(PM-only)	EBarrays(LNN)	0.820898
22	MAS5.0	t.test	0.815033
23	MAS5.0	limma	0.799162
24	MAS5.0	SAM	0.794531
25	PDNN	Welch.t	0.767155
26	RMA	Welch.t	0.7576
27	dChip(PM-only)	Welch.t	0.742685
28	dChip(PM-MM)	Welch.t	0.701568
29	dChip(PM-MM)	FC	0.668716
30	dChip(PM-MM)	EBarrays(LNN)	0.647701
31	dChip(PM-MM)	EBarrays(GG)	0.645769
32	MAS5.0	Welch.t	0.644588
33	MAS5.0	FC	0.615917
34	MAS5.0	EBarrays(GG)	0.612341
35	MAS5.0	EBarrays(LNN)	0.587304

Table 6. Area under ROC curve (FP<100) for HGU133 dataset

HGU133	Preprocessing	Differential expression	AUC (FP<100)
1	RMA	EBarrays(GG)	0.863092
2	RMA	EBarrays(LNN)	0.862798
3	RMA	FC	0.817002
4	RMA	limma	0.81548
5	RMA	SAM	0.815347
6	PDNN	SAM	0.81162
7	PDNN	limma	0.809847
8	PDNN	FC	0.797237
9	dChip(PM-only)	limma	0.786985
10	dChip(PM-only)	SAM	0.785353
11	PDNN	EBarrays(LNN)	0.779613
12	PDNN	t.test	0.777446
13	dChip(PM-MM)	SAM	0.771588
14	dChip(PM-only)	t.test	0.770554
15	dChip(PM-MM)	limma	0.764034
16	RMA	t.test	0.752983
17	dChip(PM-MM)	t.test	0.752711
18	MAS5.0	SAM	0.720726
19	PDNN	Welch.t	0.720642
20	dChip(PM-only)	FC	0.718709
21	MAS5.0	limma	0.706744
22	dChip(PM-only)	Welch.t	0.706271
23	RMA	Welch.t	0.699885
24	dChip(PM-only)	EBarrays(GG)	0.684316
25	MAS5.0	t.test	0.670529
26	dChip(PM-only)	EBarrays(LNN)	0.669003
27	dChip(PM-MM)	Welch.t	0.668742
28	dChip(PM-MM)	FC	0.577278
29	MAS5.0	Welch.t	0.553659
30	MAS5.0	EBarrays(GG)	0.552828
31	MAS5.0	EBarrays(LNN)	0.549571
32	dChip(PM-MM)	EBarrays(LNN)	0.54558
33	MAS5.0	FC	0.535097

Table 7. Area under ROC curve (FPR<0.1) for Golden Spike dataset.

GoldenS	Preprocessing	Differential expression	AUC (FP<100)
1	dChip(PM-only)	limma	0.56372
2	dChip(PM-only)	SAM	0.559223
3	dChip(PM-only)	t.test	0.547767
4	dChip(PM-only)	Welch.t	0.535408
5	dChip(PM-MM)	t.test	0.521514
6	dChip(PM-MM)	Welch.t	0.512999
7	dChip(PM-only)	FC	0.507993
8	dChip(PM-MM)	limma	0.501604
9	dChip(PM-only)	EBarrays(GG)	0.496914
10	dChip(PM-only)	EBarrays(LNN)	0.493245
11	dChip(PM-MM)	SAM	0.481958
12	PDNN	FC	0.36528
13	PDNN	limma	0.345729
14	RMA	limma	0.338737
15	RMA	SAM	0.336246
16	MAS5.0	t.test	0.335354
17	RMA	FC	0.334257
18	RMA	EBarrays(GG)	0.328945
19	RMA	EBarrays(LNN)	0.32708
20	PDNN	SAM	0.321463
21	MAS5.0	Welch.t	0.314948
22	PDNN	EBarrays(GG)	0.312507
23	PDNN	EBarrays(LNN)	0.312131
24	RMA	t.test	0.307509
25	RMA	Welch.t	0.295258
26	PDNN	t.test	0.292082
27	MAS5.0	limma	0.282148
28	PDNN	Welch.t	0.260143
29	MAS5.0	SAM	0.108924
30	dChip(PM-MM)	FC	0.058929
31	dChip(PM-MM)	EBarrays(LNN)	0.034358
32	dChip(PM-MM)	EBarrays(GG)	0.016425
33	MAS5.0	FC	0.00642
34	MAS5.0	EBarrays(LNN)	0.004662
35	MAS5.0	EBarrays(GG)	0.004136

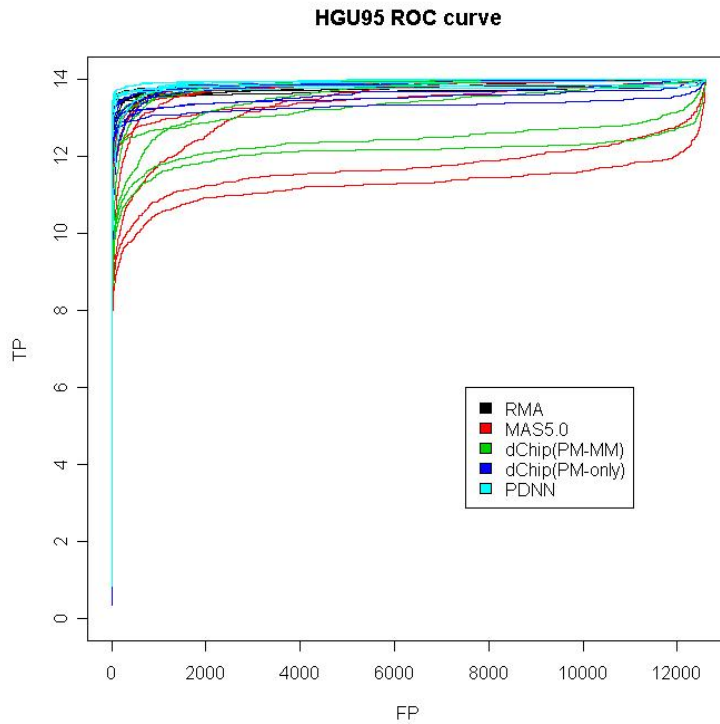


Figure 1-1. ROC curves for all combinations using HGU95 dataset (35 in total). Combinations using the same preprocessing method are assigned to the same color as shown in the legend.

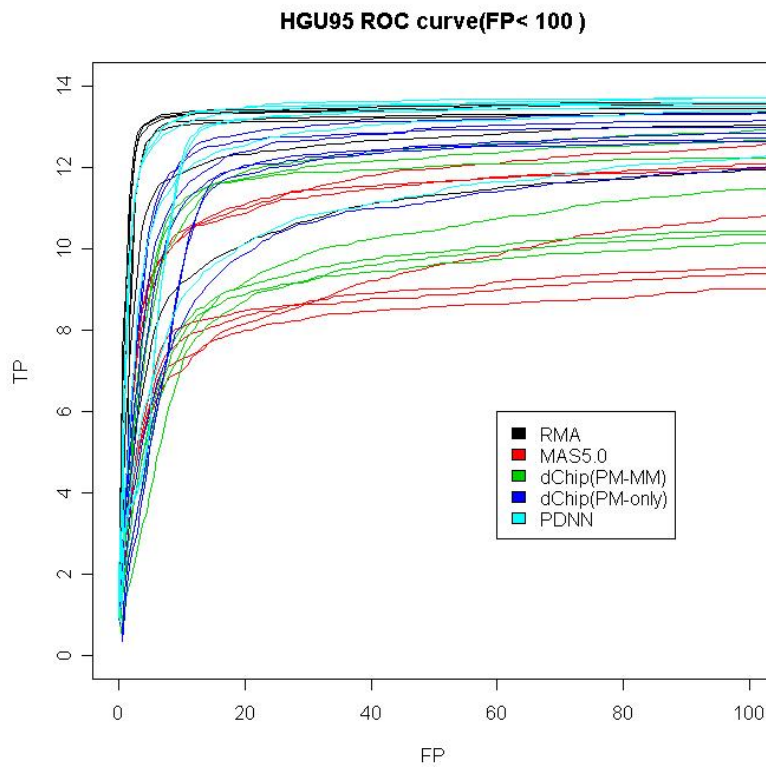


Figure 1-2. ROC curves for all combinations using HGU95 dataset (35 in total) but $FP < 100$.

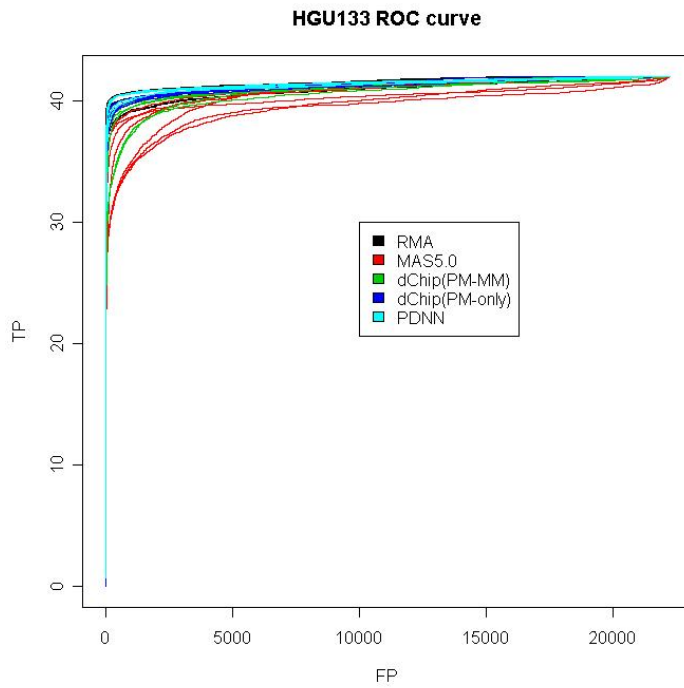


Figure 1-3. ROC curves for all combinations using HGU133 dataset (33 in total). Combinations using the same preprocessing method are assigned to the same color as shown in the legend.

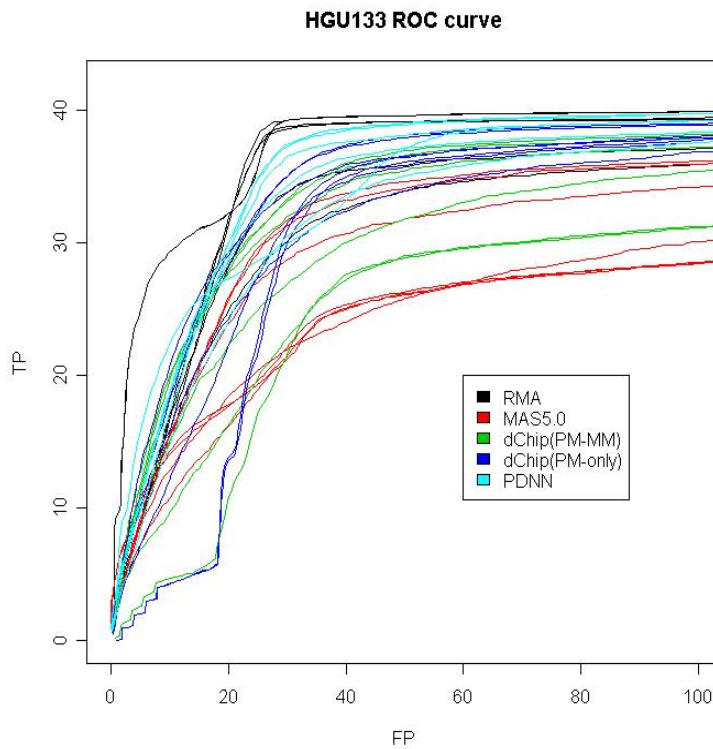


Figure 1-4. ROC curves for all combinations using HGU133 dataset but FP<100 (33 in total).

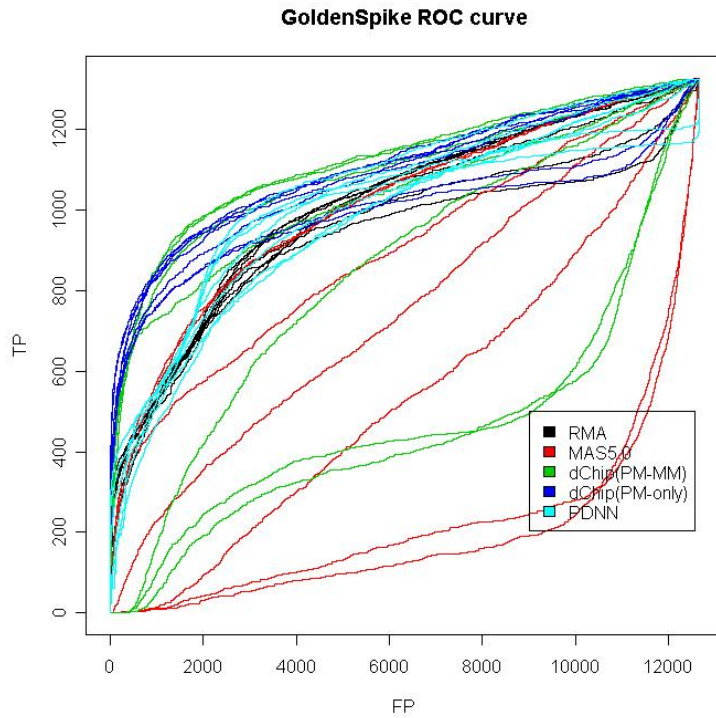


Figure 1-5. ROC curves for all combinations using Golden Spike dataset (35 in total). Combinations using the same preprocessing method are assigned to the same color as shown in the legend.

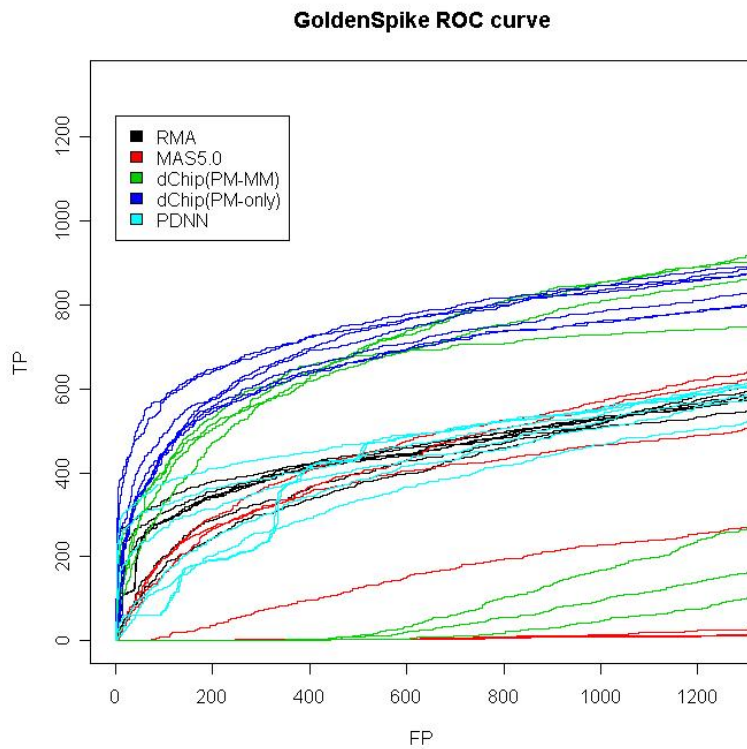


Figure 1-6. ROC curves for all combinations using Golden Spike dataset (35 in total) but false positive rate < 0.1.

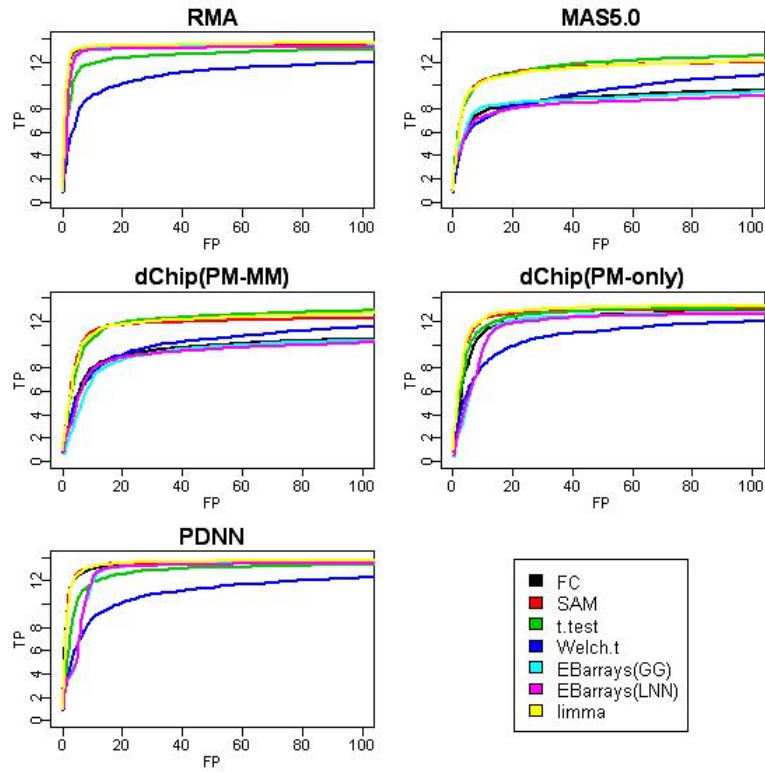


Figure 2-1. For HGU95 dataset, ROC curves of all combinations are divided by preprocessing method. Combinations using the same differential expression method are assigned to the same color as shown in the legend.

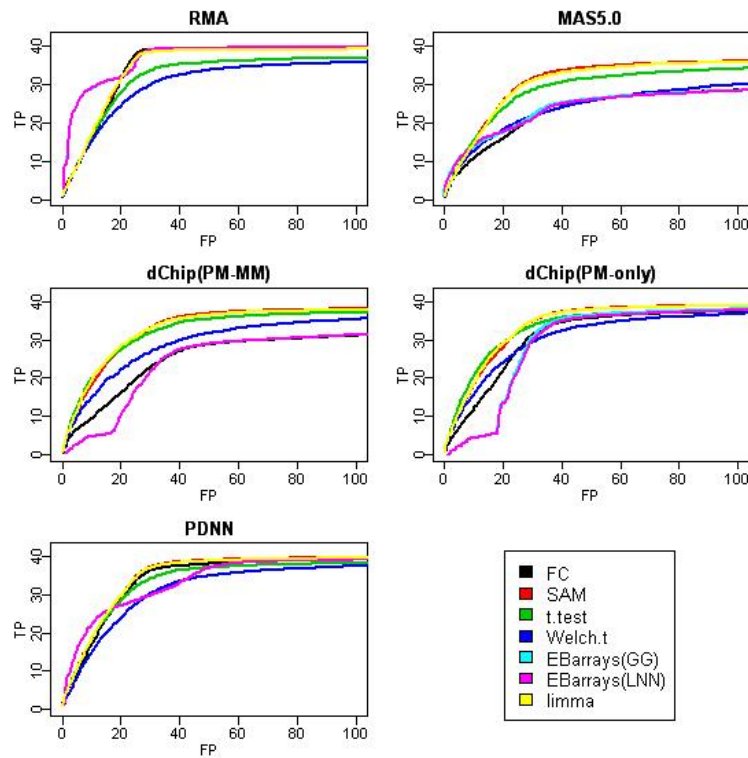


Figure 2-2. For HGU133 dataset, ROC curves of all combinations are divided by preprocessing method.

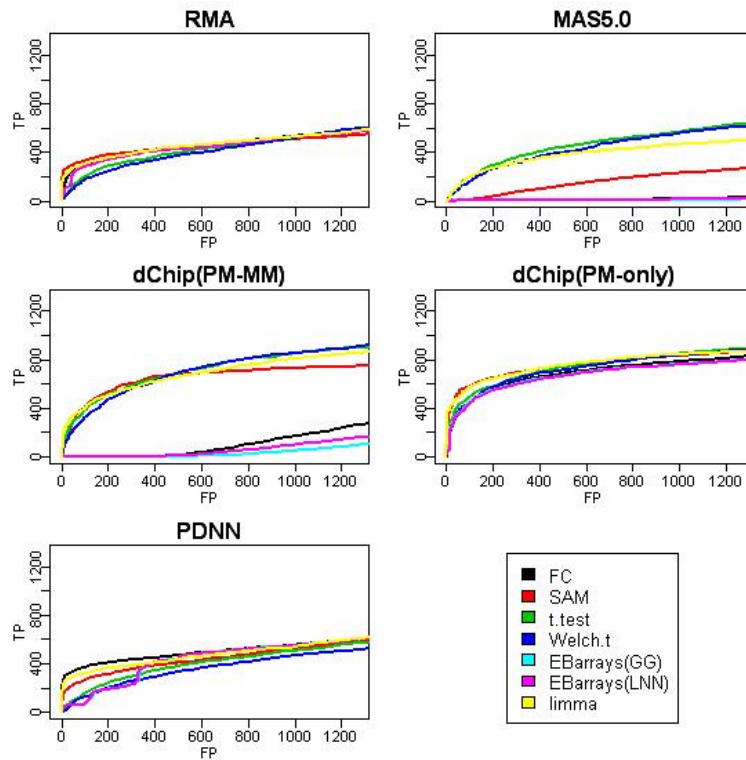


Figure 2-3. For Golden Spike dataset, ROC curves of all combinations are divided by preprocessing method.

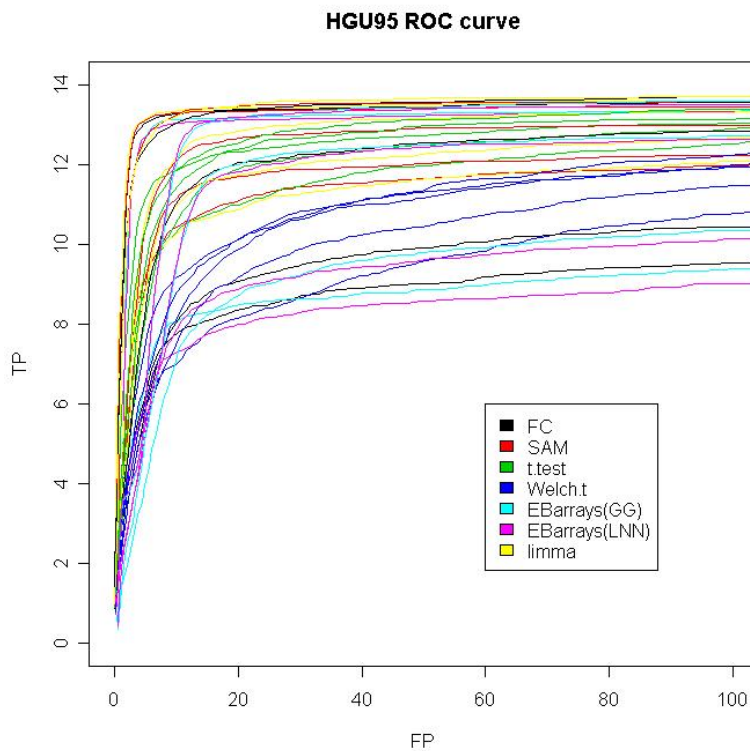


Figure 3-1. ROC curves for all combinations using HGU95 dataset. Combinations using the same differential expression method are assigned to the same color as shown in the legend.

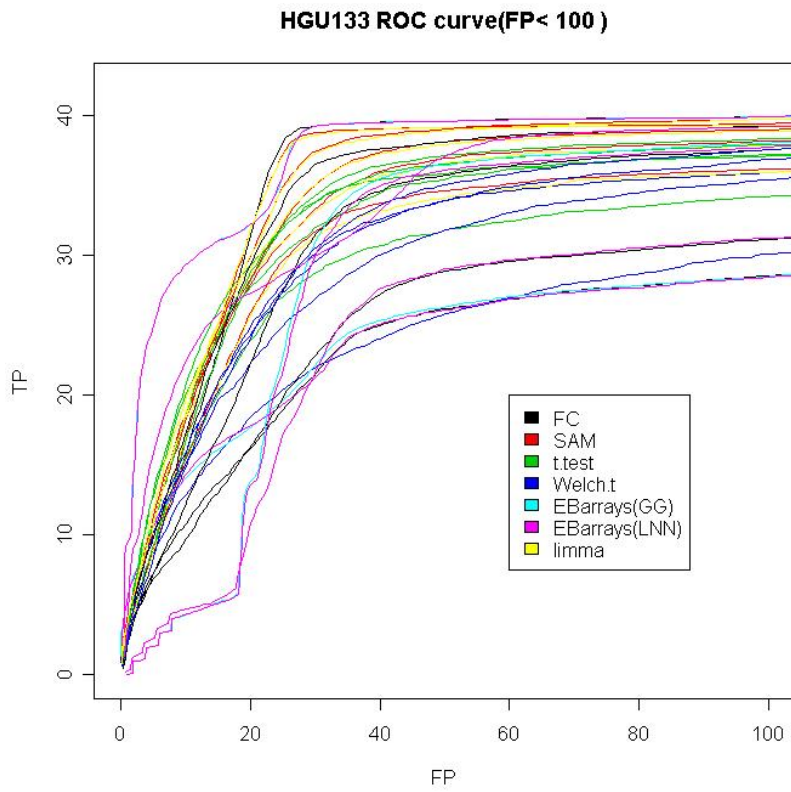


Figure 3-2. ROC curves for all combinations using HGU133 dataset.

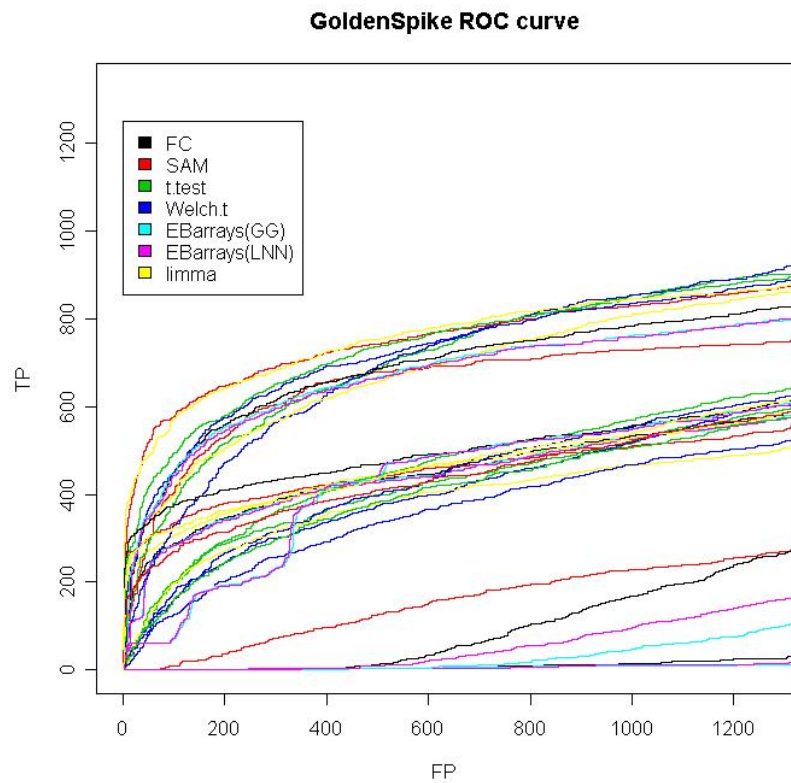


Figure 3-3. ROC curves for all combinations using Golden Spike dataset.

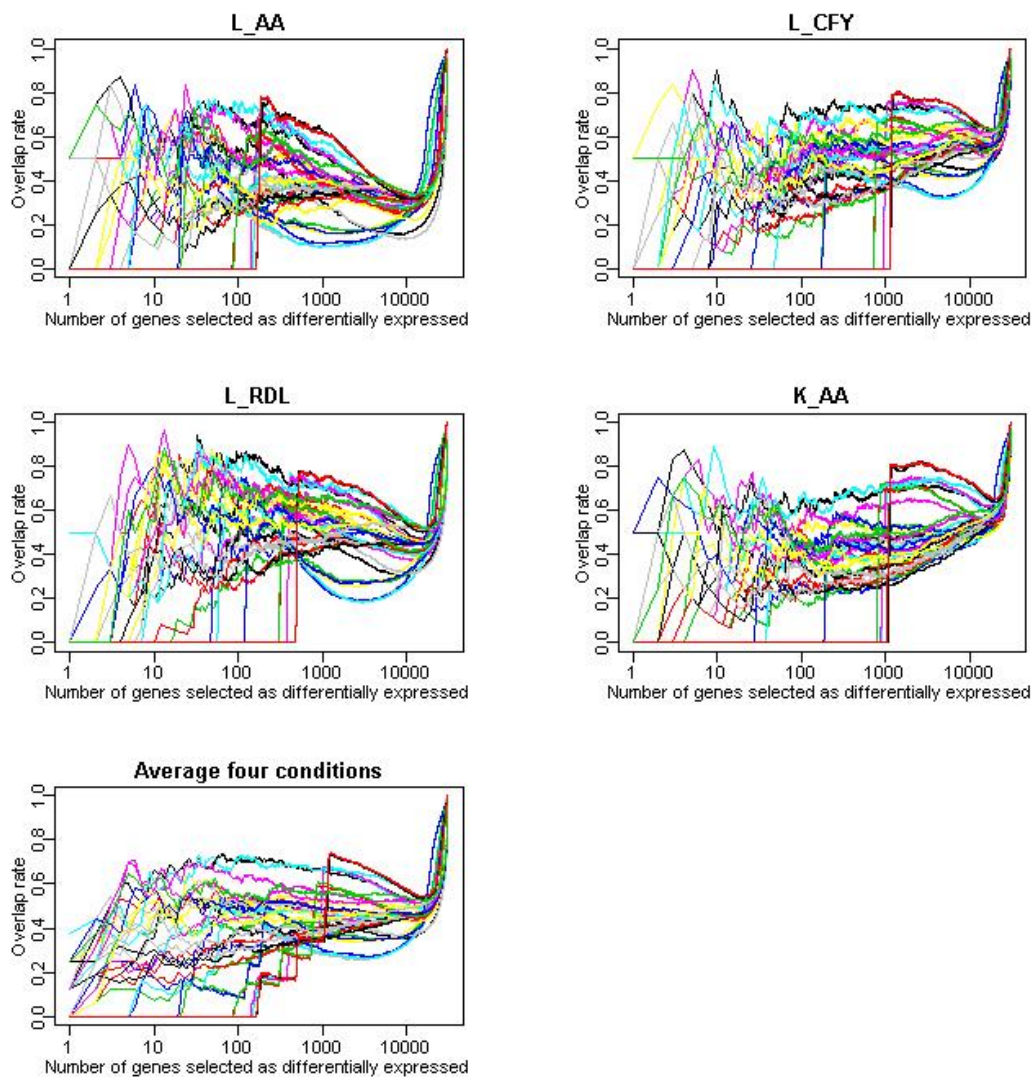


Figure 4. Overlap rate of two differentially expressed gene lists generated using different combinations. The x-axis represents the number of genes selected as differentially expressed, and the y-axis is the overlap rate of two gene lists for a given number of differentially expressed genes. The four tissues suffering different treatments versus their controls are simply called as K_AA, L_AA, L_CFY, and L_RDL. The fifth graph shows an average plot across the four conditions. x-axis is in log scale. A line represents one kind of combinations and there are 36 combinations in total. This graph shows the overall patterns.

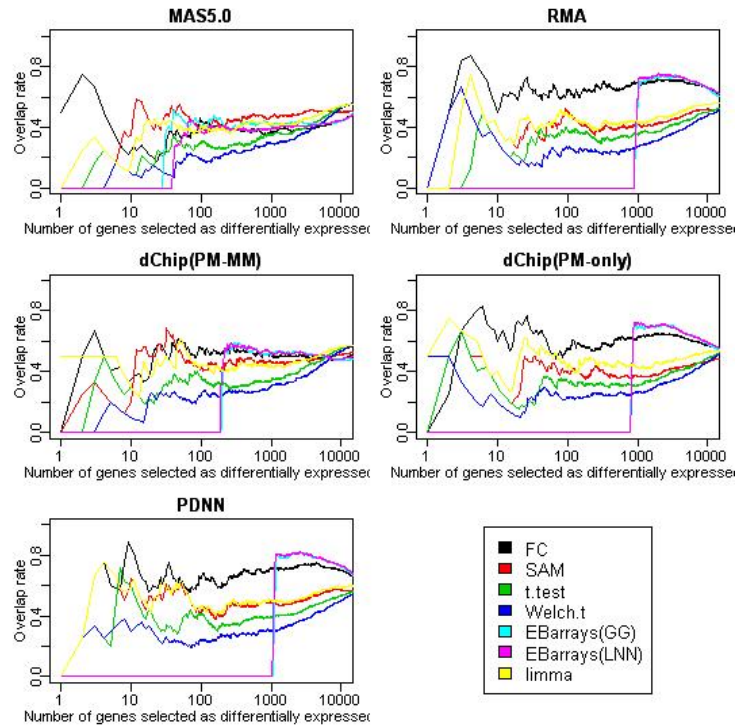


Figure 5-1. Overlap rate of two differentially expressed gene lists generated using different combinations for K_AA treatment/control. All combinations are divided by preprocessing method. Combinations using the same differential expression method are assigned to the same color as shown in the legend.

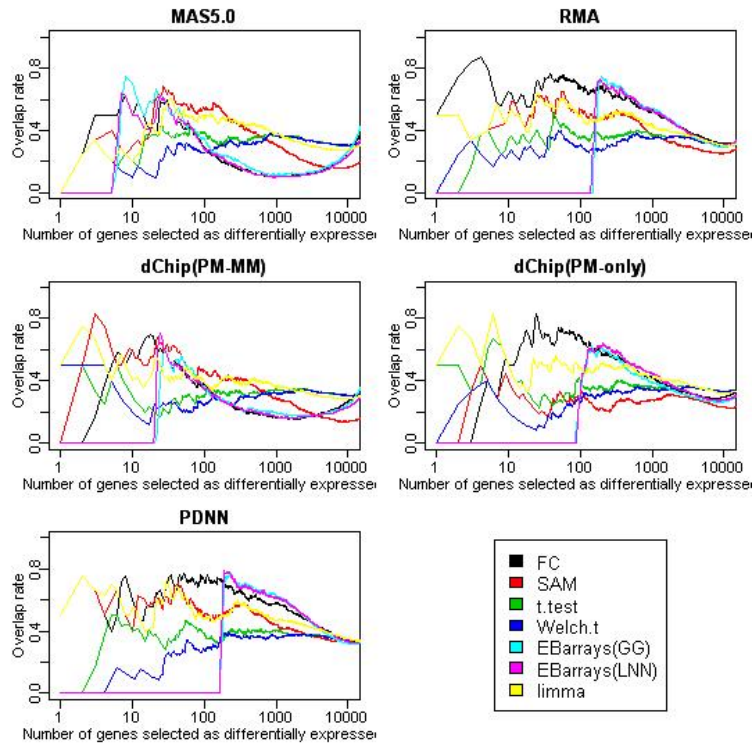


Figure 5-2. Overlap rate of two differentially expressed gene lists generated using different combinations for L_AA treatment/control.

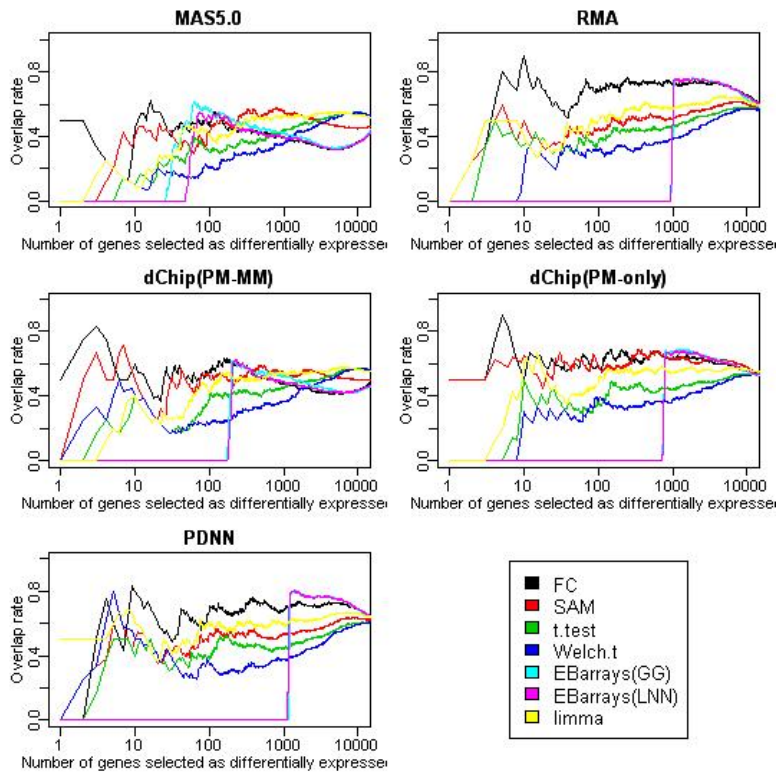


Figure 5-3. Overlap rate of two differentially expressed gene lists generated using different combinations for L_CFY treatment/control.

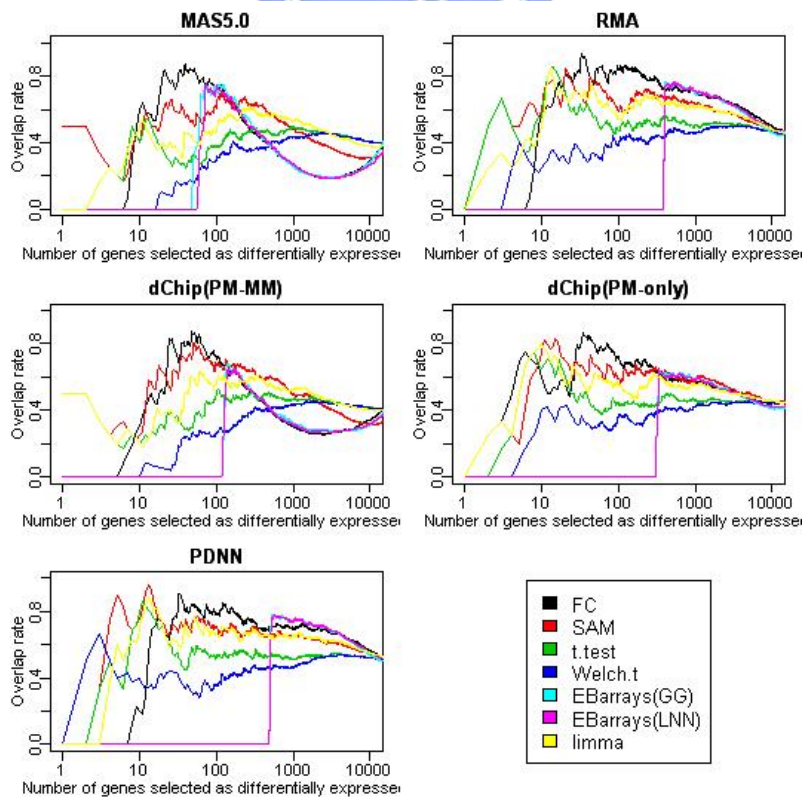


Figure 5-4. Overlap rate of two differentially expressed gene lists generated using different combinations for L_RDL treatment/control.

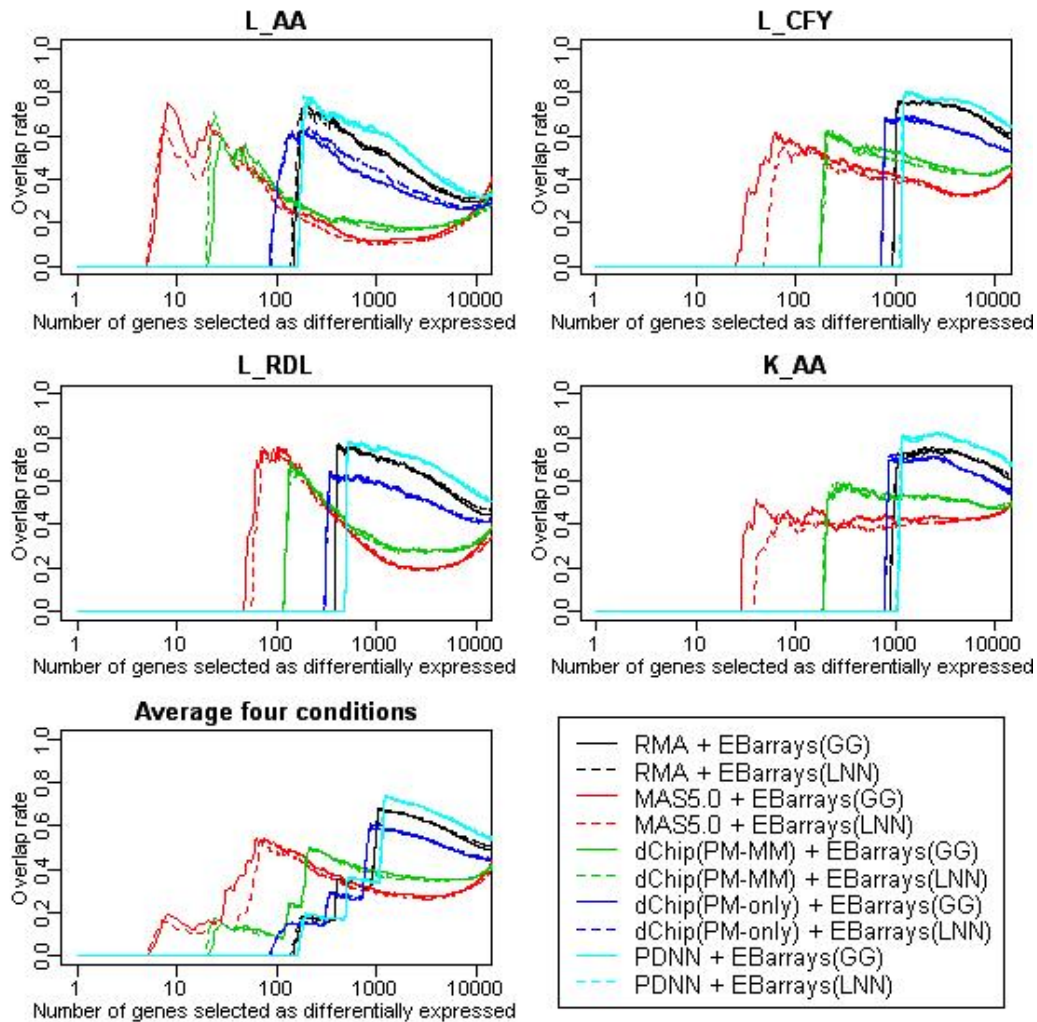


Figure 6. Overlap rate of two differentially expressed gene lists generated using different combinations with EBarrays as differential expression method. Ten combinations in total are shown in the legend.

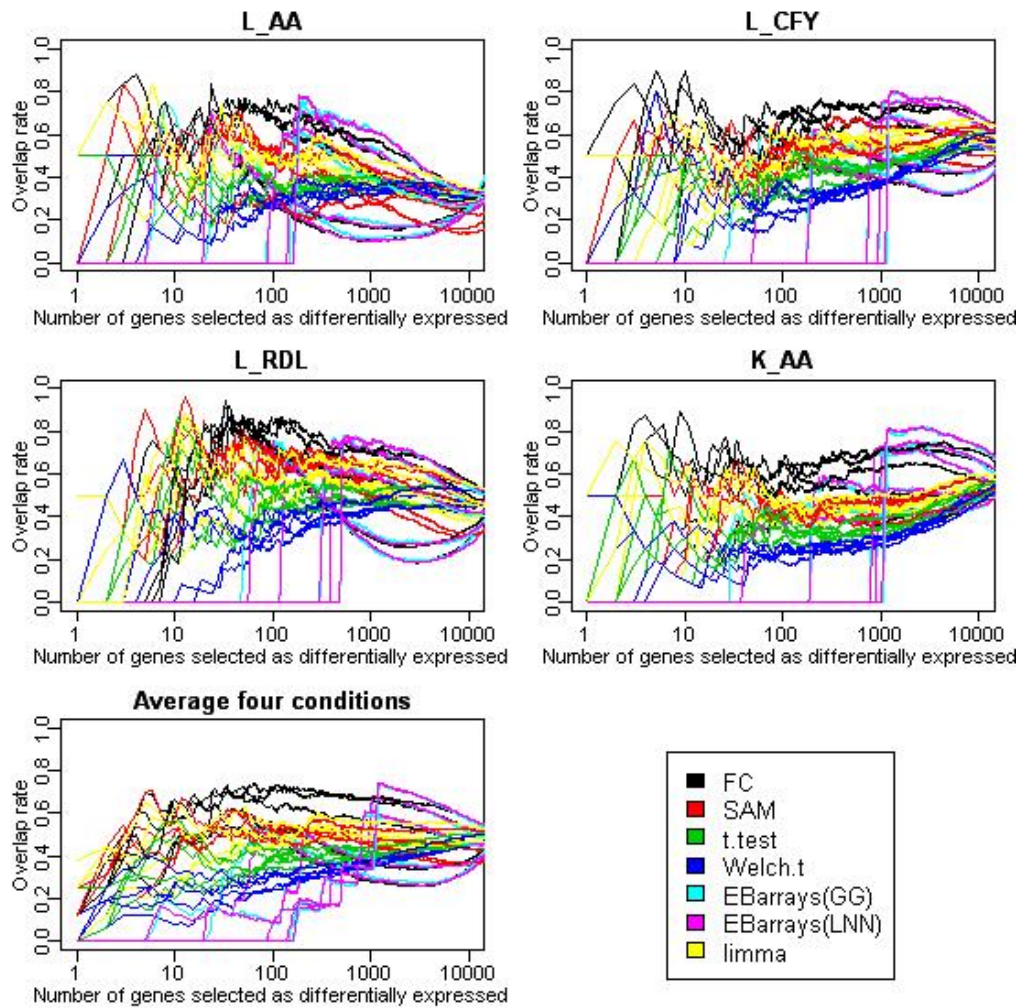


Figure 7. Overlap rate of two differentially expressed gene lists generated using different combinations. Combinations using the same differential expression method are assigned to the same color as shown in the legend. All combinations are included.

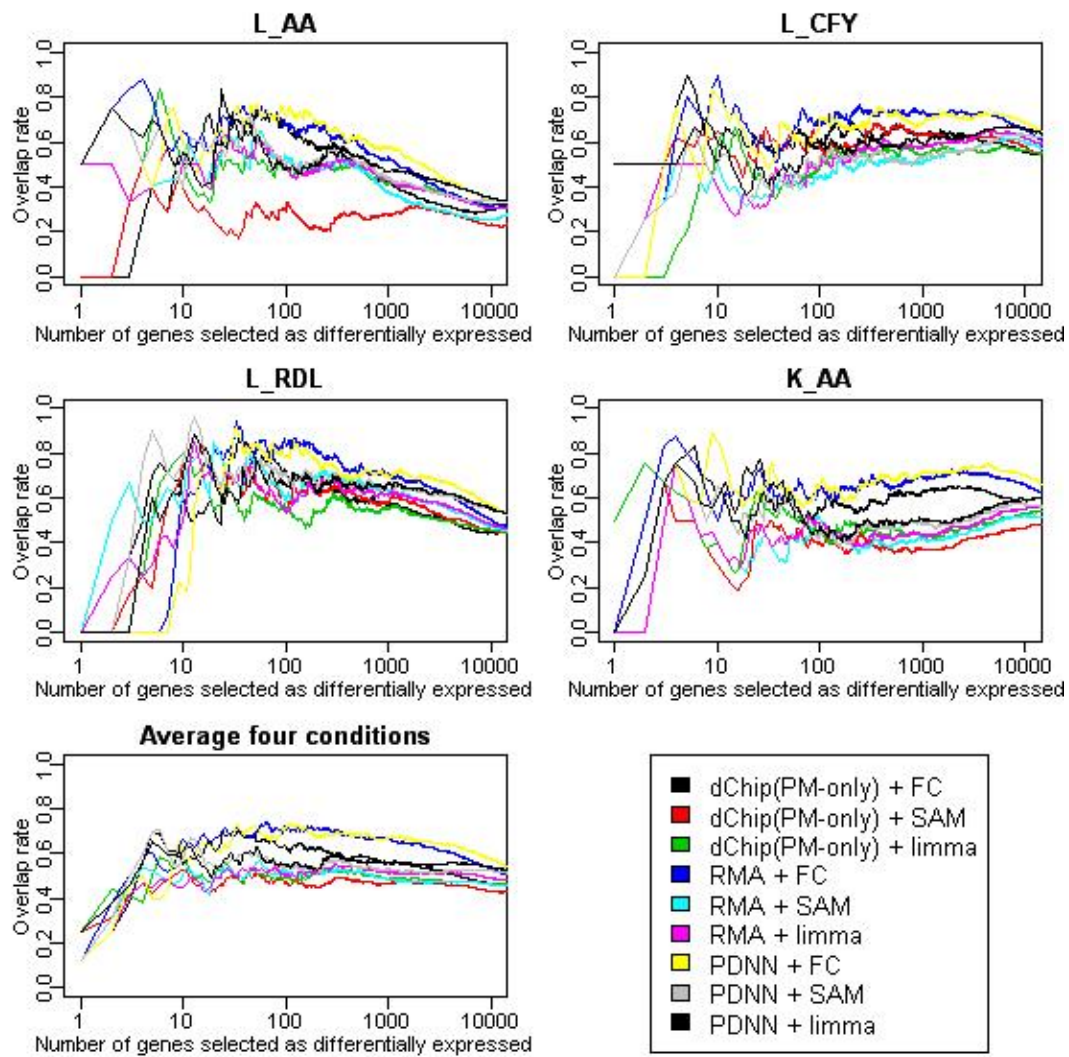


Figure 8. Overlap rate of two differentially expressed gene lists generated using different combinations. Only the nine permutations with RMA, dChip(PM-only), PDNN as preprocessing method and FC, SAM, limma as differential expression method are plotted.

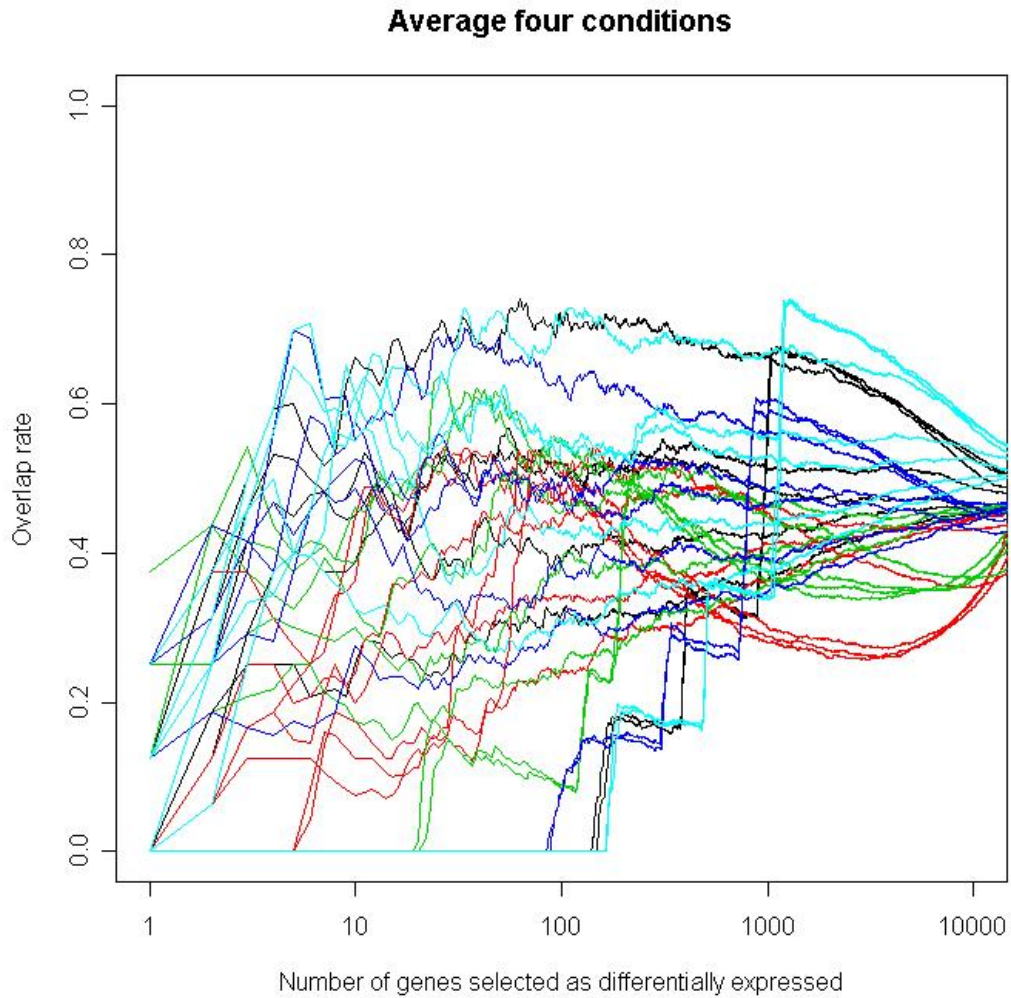


Figure 9-1. Average overlap rate of two differentially expressed gene lists generated using different combinations. Combinations using the same preprocessing method are assigned to the same color. All combinations are included. Black for RMA, red for MAS5.0, green for dChip(PM-MM), blue-black for dChip(PM-only), and baby blue for PDNN.

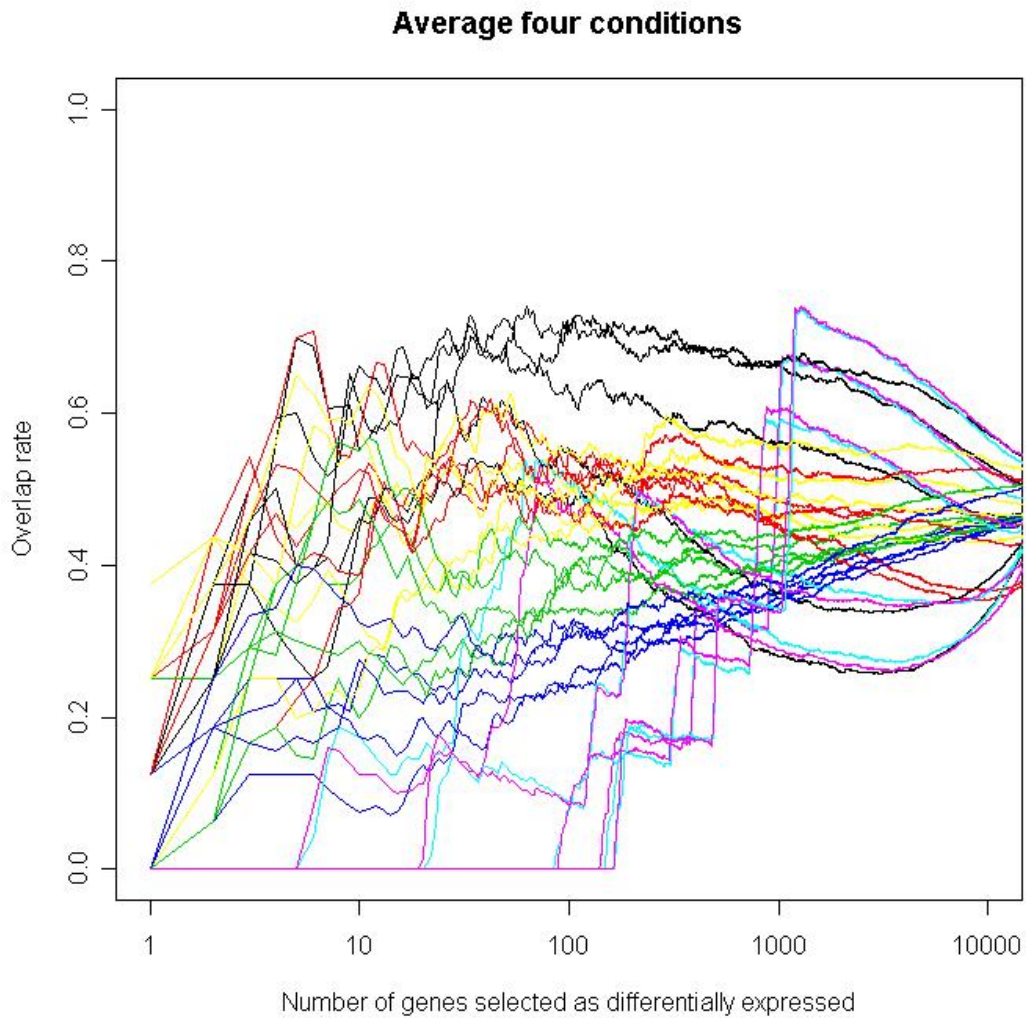


Figure 9-2. Average overlap rate of two differentially expressed gene lists generated using different combinations. Combinations using the same differential expression method are assigned to the same color. All combinations are included. Black for FC, red for SAM, green for t-test, blue-black for Welch t-test, baby-blue for EBarraays(GG), pink for EBarraays(LNN), and yellow for limma.