MIXED EFFECTS MODELS FOR TIME SERIES GENE EXPRESSION DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

İBRAHİM ERKAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
STATISTICS

DECEMBER 2011

Approval of the thesis:

MIXED EFFECTS MODELS FOR TIME SERIES GENE EXPRESSION DATA

submitted by **İBRAHİM ERKAN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Statistics Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Öztaş Ayhan _____
Head of Department, **Statistics**

Assist. Prof. Dr. Özlem İlk _____
Supervisor, **Statistics Dept., METU**

Assoc. Prof. Dr. İnci Batmaz _____
Co-Supervisor, **Statistics Dept., METU**

Examining Committee Members:

Prof. Dr. Öztaş Ayhan _____
Statistics Dept., METU

Assist. Prof. Dr. Özlem İlk _____
Statistics Dept., METU

Assoc. Prof. Dr. Özlen Konu _____
Dept. Of Molecular Biology and Genetics, Bilkent Uni.

Assist. Prof. Dr. Ceylan Yozgatlıgil _____
Statistics Dept., METU

Assist. Prof. Dr. Zeynep Kalaylıoğlu _____
Statistics Dept., METU

Date: 02.12.2011

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : İbrahim Erkan

Signature :

ABSTRACT

MIXED EFFECTS MODELS FOR TIME SERIES GENE EXPRESSION DATA

Erkan, İbrahim

Ph.D., Statistics Department

Supervisor : Assist. Prof. Dr. Özlem İlk

Co-Supervisor : Assoc. Prof. Dr. İnci Batmaz

December 2011, 129 pages

The experimental factors such as the cell type and the treatment may have different impact on expression levels of individual genes which are quantitative measurements from microarrays. The measurements can be collected at a few unevenly spaced time points with replicates. The aim of this study is to consider cell type, treatment and short time series attributes and to infer about their effects on individual genes. A mixed effects model (LME) was proposed to model the gene expression data and the performance of the model was validated by a simulation study. Realistic data sets were generated preserving the structure of the sample real life data studied by Nymark et al. (2007). Predictive performance of the model was evaluated by performance measures, such as accuracy, sensitivity and specificity, as well as compared to the competing method by Smyth (2004), namely Limma. Both methods were also compared on real life data. Simulation results showed that the predictive performance of LME is as high as 99%, and it produces False Discovery Rate (FDR) as low as 0.4% whereas Limma has an FDR value of at least 32%. Moreover, LME has almost 99% predictive capability on the continuous time parameter where Limma has only about 67% and even it cannot handle continuous independent variables.

Keywords: Microarray Data, Unevenly Spaced Time Points, Subject-wise Testing

ÖZ

ZAMAN SERİSİ GEN İFADE VERİLERİ İÇİN KARMA ETKİLİ MODELLER

Erkan, İbrahim

Doktora, İstatistik Bölümü

Tez Yöneticisi : Yrd. Doç. Dr. Özlem İlk

Ortak Tez Yöneticisi : Doç. Dr. İnci Batmaz

Aralık 2011, 129 sayfa

Hücre türü ve ilaç gibi deneysel faktörler, genlerin mikrodizinlerden elde edilen nicel ölçümleri olan bireysel ifade düzeylerini etkilemektedir. Yapılan ölçümler eşit aralıklı olmayan birkaç zaman noktasında ve tekrarlı şekilde toplanabilir. Bu çalışmanın amacı hücre türü, ilaç etkisi ve kısa zaman serisi gibi unsurları inceleyip, bunların genler üzerindeki bireysel etkileri üzerine çıkarsama yapmaktır. Gen ifade verisinin hiyerarşik yapısını modellemek üzere karma etkili bir model (LME) önerilmiş ve bir benzetim çalışmasıyla modelin başarımı doğrulanmıştır. Benzetim verileri, Nymark vd. (2007) tarafından yapılan çalışmada kullanılan gerçek verinin yapısına uygun olarak türetilmiştir. Modelin tahmin edici başarımı doğruluk, hassasiyet ve özgüllük ölçüleri ile değerlendirilmiş ve Smyth (2004) tarafından önerilen Limma isimli seçenek yöntemle karşılaştırılmıştır. Ayrıca her iki yöntem gerçek veriler üzerinde de karşılaştırılmıştır. Benzetim sonuçları önerilen modelin tahmin edici başarımının %99 gibi çok yüksek bir düzeyde olduğunu, hatta Hatalı Keşif Oranı (FDR) değerlerinin %0.4 kadar düşük olup, aynı değerin Limma'da en az %32 kadar olduğunu göstermiştir. Dahası, LME'nin sürekli bir bağımsız değişken olan zaman parametresi üzerindeki tahmin edici başarımı %99 düzeyinde iken, Limma sadece %67 düzeyinde kalmış olup sürekli bağımsız değişkenlerin kullanımına bile uygun değildir.

Anahtar Kelimeler: Mikrodizin Verileri, Eşit Aralıklı Olmayan Zaman Noktaları, Nesne Bazında Test

To my family

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my patient advisor, Assist. Prof. Dr. Özlem İlk for her endless support and trust in me.

I am thankful to Assoc. Prof. Dr. İnci Batmaz that she has been very encouraging me during my graduate studies and she has also been my co-advisor. She has always been an invaluable mentor and a familial to me.

I would like to thank to Fulbright Association that they awarded me as a PhD scholar and provided funding for my studies at Statistics Department of the Wharton School of Business at University of Pennsylvania.

I would like to thank to The Scientific and Technological Research Council of Turkey (TÜBİTAK) that they provided me a PhD program support funding for four years. Also The Turkish Academic Network and Information Centre (ULAKBIM), a R&D Facility Institute of TÜBİTAK, provided me access to TR-GRID computing system so that I could run all the computations throughout my thesis.

I would like to thank to Assoc. Prof. Dr. Özlen Konu for providing me access to the server of Molecular Biology and Genetics Department of Bilkent University. She also contributed a lot to my thesis especially in biological essentials of the study.

Although it is not directly funded, this thesis was initiated as part of TUBITAK-TBAG-106T548 project.

I would like to thank all the members of my thesis supervising committee and thesis jury that are Prof. Dr. Öztaş Ayhan, Assoc. Prof. Dr. İnci Batmaz, Assoc. Prof. Dr. Özlen Konu, Assist. Prof. Dr. Özlem İlk, Assist. Prof. Dr. Zeynep Kalaylıoğlu and Assist. Prof. Dr. Ceylan Yozgatlıgil.

I cordially thank to my beloved wife Dr. B. Burçak Başbuğ Erkan who spent lots of her time and nerves to back me up when I was studying and she never for a second got to thinking I was not able to make it.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ACC	Accuracy
AD	Anderson – Darling
ANOVA	Analysis of Variance
DNA	Deoxyribonucleic Acid
EM	Expectation-Maximization
FDR	False Discovery Rate
FPR	False Positive Rate
gcRMA	Gene Chip Robust Multiarray Averaging
LME	Linear Mixed Effects
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
mRNA	Messenger Ribonucleic Acid
nlme	Non-linear Mixed Effects
NPV	Negative Predictive Value
PPV	Positive Predictive Value
REML	Restricted Maximum Likelihood
RMA	Robust Multiarray Averaging
RNA	Ribonucleic Acid
SAM	Significance Analysis of Microarrays
SMD	Stanford Microarray Database
SOM	Self Organizing Maps
SPC	Specificity
STEM	Short time Series Expression Miner
TPR	True Positive Rate

CHAPTER 1

INTRODUCTION

The change in the nature in recent years with pollution all over the world has created many direct and side effects to our health. Especially, cancer and similar diseases started to threaten us more. As a result, the treatment of such diseases has become more important and also elaborate because of increasing number of uncontrollable sources of variation. In this context genetics has become one of the most popular research areas.

1.1 Microarrays

Scientists have invented microarrays to display protein activity at genes. It is a breakthrough when scientists are able to see the change in gene activities under the presence and absence of conditions of interest. This knowledge would help them very much such as in finding a coherent certain treatment, inheritance of genetic diseases and many more. For example, researchers may want to investigate the effect of asbestos on breast cancer. Therefore, gene activities can be observed on both asbestos exposed cells and non-exposed cells and the genes which change activity under different conditions can be determined. Then, required actions can be taken accordingly.

1.2 Scope of the Study

This thesis includes the analysis of short course time series microarray gene expression data. Short course time series data are observed in the course of time (where time points may be unevenly spaced) when microarray experiments are used to study the behaviour of genes and their expression levels are investigated. There can be more than one observations per time point and the number of observations per time point may vary through the series because of the nature of the experiment. Figure 1.1 illustrates the structure of the data that is studied in this thesis.

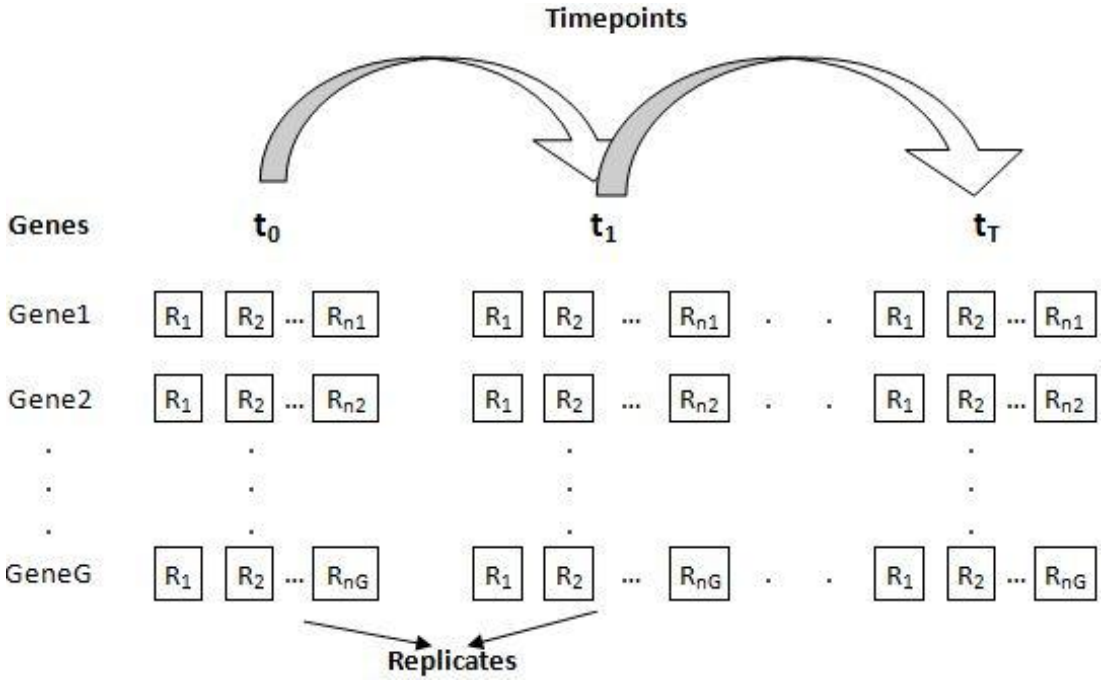


Figure 1.1 Structure of time series gene expression data with replicates

The analysis of such data has some challenges for researchers. The first challenge is that gene expressions across time points may have a dependence structure which should not be ignored during the analyses. The probe measure which represents the relative expression level of an individual gene has more than one sampling points over time. Therefore, the measurements obtained over time belong to the gene creating a dependent sequence of measurements.

The second challenge is that the number of time points is very few (generally less than or equal to eight) compared to classical time series data which usually have more than 50 observations for a convenient time series modeling. There are a number of reasons behind this. The first and the foremost one is that the microarray chips are very expensive for both an extensive use and many repetitions of the same experiment. The second reason is that sometimes, depending on the structure and the donor of the experiment made, it is impossible to repeat the experiment many times. For example, it is hard to make people attend the experiment 50 times and provide blood cells, or in an experiment that you investigate the behaviour of a poison, the rat can die and it becomes impossible to repeat the experiment many times. Figure 1.2 skecthes a sample of size six from a single probe measured over five time points that are 0h, 1h, 6h, 24h, 48h and 168h.

As the number of time points may vary, the number of replicates per time point may vary as well. The less the number of replicates, the harder to fit models because estimation of the variance components gets harder or impossible. Sometimes, the data is unbalanced that cause another challenge for researchers.

The third challenge is unevenly spaced time points. Unevenly spaced time points indicate that the amount of time between consecutive measurements is not the same across all time points. The time elapsed after an observation may vary. In this case, monitoring the process over time becomes very difficult also making it very hard to express the reason behind the change in measurements. This is unusual in classical time series approach.

In addition to the challenges that were mentioned above, it is methodologically and computationally very extensive and demanding that short time series microarray gene expression data may contain replicates changing in number at every time point and gene/probe set. For example, the sample in Figure 1.2 has 5 timepoints and it has 1 replicate only at 48h where other measurements are singletons at all other time points. Moreover, there may be factors such as cell type as well as treatment, one or both of which might have more than two levels. Time as a source of variation in microarray experiments is a continuous independent variable rather than a qualitative factor most of the time. Biologists are very keen on finding out whether a treatment has an acute or chronic effect on the subject of interest.

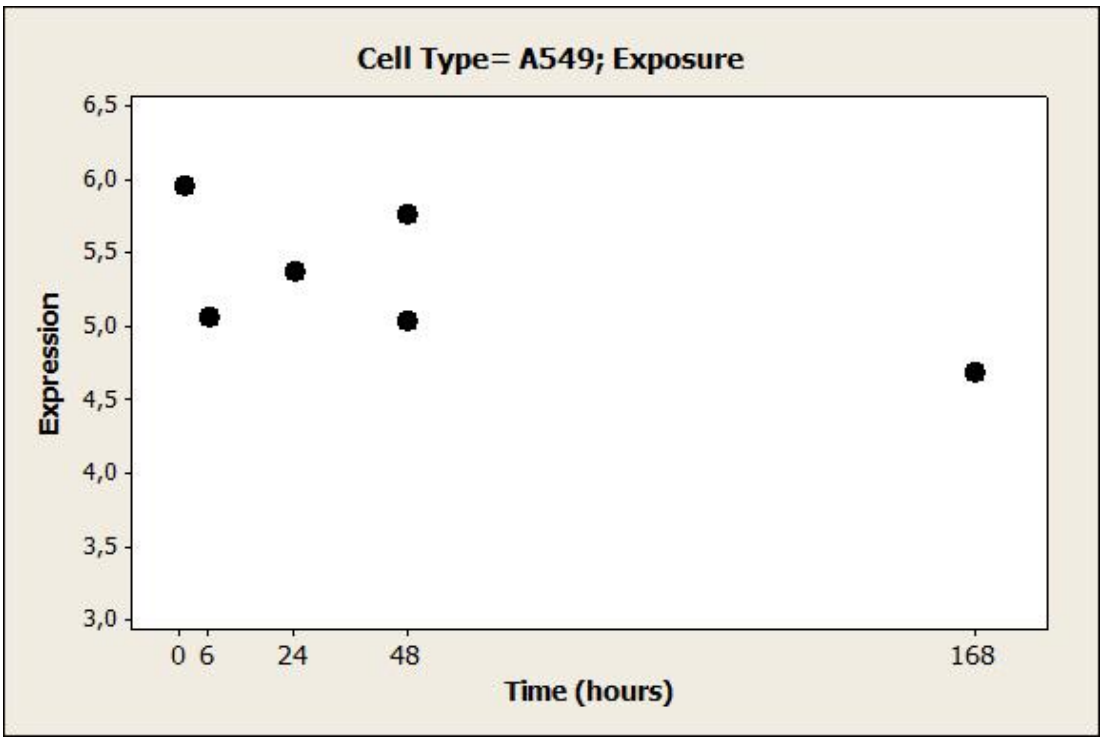


Figure 1.2 Sample observations from "238743_at" probe measured under A549 cell and exposure group from Asbestos Dataset (Nymark et al., 2007)

Summing up the challenges and the complexities above, biologists want to infer about the changes in gene expressions for individual genes. Qualitative factors are well handled by means of t-tests, ANOVA, clustering, splines and linear modeling. However, the challenging part here is to infer about the subject specific gene expression profile under the existence of qualitative experimental conditions. A profile is the change in the gene expression level across the time panel when the measurements are collected. Biologists would like to know if there is any significant change in the short course of time as well as the long course of time. A careful investigation of the literature indicates that the proposed methods are insufficient to resolve this problem, and this study aims to propose a plausible model to solve it.

The model in this dissertation on the other hand, first clusters the genes using a two-step clustering scheme thus breaking the correlation structure and creating gene-sets and then applies a linear mixed model to estimate effects of group, treatment and time points for short time series datasets with or without replication.

1.3 Contributions

- Proposing the use of LME method at every other individual time interval of a short time series microarray data, providing modeling and testing a short time series profile and subjectwise testing.
- Comparing it with Limma, the competing and most widely used alternative method. It was shown through a comprehensive simulation study that the proposed methodology outperformed Limma in accuracy, specificity, positive predictive value, negative predictive value, false discovery rate and F1 value performance measures in overall results.
- Fitting the random effects together with the fixed effects produce more unbiased results compared to when they are fit only as fixed effects. Existence of the random effects compensates the shrinkage of the fixed effects towards to the mean value. This also helps to avoid any over and under estimation of parameters that occur by chance. This is where Limma method fell behind and produced false discoveries.
- Handling repeated measures and unbalanced data via LME for short time series microarray data.
- Producing more appropriate results when the data of interest has hierarchical levels with many factors such as cell type, treatment, short time series and the clusters that contain the probe sets with similar expression profiles.
- Providing great flexibility whenever additional factors or terms such as covariates or categorical factors are to be added to the LME model.
- Detecting acute and chronic effects of a treatment via modeling the short time series microarray gene expression profiles.
- Handling the differing time lags by incorporating time as an independent variable into the LME model as well as testing the time change effect.
- Increasing predictive capabilities and F1 value short time series microarray gene expression profile analysis for both factors and independent variables such as continuous time parameter.
- Proposing a two-stage clustering algorithm for detection of time series gene expression profiles.
- Proposing a real like data simulation algorithm for short time series microarray gene expression data with differing number of replicates per time point as well as incorporating cell type, exposure and other required effects.

- Providing a computational framework for analyzing short time series microarray gene expression data proposing a very comprehensive simulation and short time series microarray gene expression data fitting algorithm along with an R code. For the simulation part, the code first generates realistic gene expression profiles. The profiles can be modified such as changing the number of cell types, treatment groups and time points. The code then generates realistic initial set of data by making use of a mixture normal distribution. The parameters of the initial data can be adjusted as well. According to the profiles that are created the code can simulate the short time series, replications in accordance with the structure of the profile. Realistic noise and experimental factors are also incorporated to the simulated data. For the real data fitting part, researchers who would like to utilize the code should only rename the column names of their dataset and run the code. On the overall, the code is very user friendly, easy to use and allows customizations. The code is free to access and can be downloaded from www.metu.edu.tr/~oilk/LME_code.zip.

An overview of microarrays, the structure of the short time series microarray gene expression data analyzed in this study, scope of the study and the contribuions to the literature were included in this chapter. In the second chapter, a comprehensive review of the literature will be presented. Alternative studies, previous work in this area and their findings will be summarized. In the third chapter, methodology that is used will be introduced step by step. The forth chapter will include all of the details of the simulation study and the fifth chapter will provide findings, results, illustrative examples and remarks on both simulated data and real life data. Conclusions and possible future studies will take place in the sixth chapter. For ease of reading and understanding most of the figures and tables on performance measures that were obtained in the simulation study are given in appendix.

CHAPTER 2

LITERATURE REVIEW

Analysis of microarray data is difficult because they come in large sizes and there are variations result from different sources. These sources can be the measurement system, the experimenter and environmental factors such as temperature, pressure, array reading, image noise and translation errors. These cause increased measurement variation and bias. Studying with time series microarray data is even more difficult because of the correlation between sequential measurements and possible missing values.

Although there are challenges in analyzing short time series gene expression data, they are frequently used. Ernst et al. (2005) provided a statistic about the number of time points of the microarray datasets from Stanford Microarray Database (SMD). This study indicates that more than 80% of all time series datasets contain less than or equal to eight time points. Therefore, researchers working on microarray experiments are highly in need of statistical techniques that help determining the patterns or modeling the behaviour of genes in the short course of time.

An insightful investigation of the literature in the field of short time series microarray data analysis direct us to mainly split them into two common approaches as model based and non-model based methods. In the proceeding two sections of this chapter, representative and frequently cited studies from both approaches are summarized. The methodology that researchers used to analyse the microarray data also differed according to the main interest such as grouping the genes that show similar behaviour under certain conditions or differentially expressed gene or gene groups.

2.1 Non-model Based Methods

Recently, microarray studies are becoming more and more popular. In such studies, not only the dimensions of the data sets are large but also the number of factors manipulating the microarray measurements are large. Cell type, genes or gene groups, time effect, array type and treatment type can be good examples as sources of variation to the microarray measurements. Accordingly, the number of required inferences are also large. Therefore, reducing the dimension is a plausible way for the simplification of the problem. For example, k-means clustering method (Tavazoie et al., 1999), hierarchical clustering (Eisen et al., 1998) and analysis of variance (Kerr et al., 2000) have been used in many researches. Matsui et al. (2008) suggested that clustering microarray data is very useful in reducing dimension as well as understanding co-regulated genes that behave similarly under diseases or certain treatment. They also pointed that clustering genes and treating them as a group improve the predictive variance. Their approach maked use of a logrank test that is used for multivariate permutation procedure seeking for an optimum cut-off point for the p-value to decide whether a cluster has differentially expressed genes or not while the test does not sacrifice from false discovery rate. As a non-model based alternative, Watson (2006) studied a clustering method that helped finding coexpressed gene sets and utilized an R package for that. The software can identify groups of genes that are expressed similarly. Their algorithm avoids parametric modeling or testing. Another nonparametric approach was proposed by Shah & Corbeil (2011). They used tensor analysis in order to transform data and without clustering the data explicitly they were able to identify groups of differentially expressed genes in a short time-series data. The downsides of their approach are that it cannot distinctly determine the source of differential expression such as cell type, exposure or time, and the time interval where genes are differentially expressed cannot be identified.

Complexity and duration of the computations during the microarray analyses are important attributes for a microarray study. In opposition to the most methods Qin et al. (2008) proposed a computationally less intensive clustering algorithm for detecting differentially expressed genes from simple microarray experiments. They also performed comprehensive simulation studies which showed that their method is substantially more powerful and also more robust than well-known SAM and eBayes approaches. Another study which compares its performance to SAM was done by Sinha & Markatou (2011). They developed a computer software package, the main advantage of which is that it is capable of doing both

significance analysis and clustering. It is user friendly as well. On the other hand, it does not return any models and incapable of determining the effects of experimental factors such as cell type and exposure.

Some clustering methods were developed for time series microarray data (Bar-Joseph et al. (2002); Ramoni et al. (2002)). Bar-Joseph et al. (2002) based his approaches on statistical models for clustering purposes. He proposed representing every set by a spline curve as a solution to problems raising from missing values and unevenly spaced time points. However, this approach requires long time series for convenient results. Androulakis et al. (2007), on the other hand, discussed clustering methods for analyzing short time series gene expression data. They also pointed out some challenges, opportunites and also the quality of clustering methods. However, they did not mention any model based approach or testing. Another study for achieving a similar goal is the difference-based clustering algorithm that was given in Kim & Kim (2007). They claimed that their algorithm outperforms the competing alternatives namely k-means, Self Organizing Maps (SOM) and Short Time-series Expression Miner (STEM) methods in terms of clustering short time series gene expression data with replicates. STEM application assumed replicates at each time point but, the data structure studied in this thesis may not even have replicates per time point.

Although clustering is very helpful tool in understanding the structure of the data, determining similar gene behaviour, reducing the dimension lessening the number of the parameters to be estimated, the method has still some downsides. Clustering methods cannot test statistical significance and will probably detect clusters even if they don't exist (Xu et al., 2002). Moreover, Xu pointed out that clustering methods are sensitive to data transformations and units of measure. In addition, Park et al. (2003) noted that clustering methods cannot produce stable results as the number of genes increase. In an unpublished work Kuenzel (2010) compared most of the clustering methods and pointed that clustering make microarray data sensible. However, he concluded that there are so many clustering methods which makes it hard to choose. Nevertheless, although clustering is usually needed, it is clear that clustering as the only method to analyze the microarray data may not be sufficient in all studies especially for short time series microarray data.

2.2 Model Based Methods

2.2.1 Hierarchical and Mixture Models

Two main approaches dominate the field of microarray analyses. First is hierarchical models, which most of the time utilize Bayesian tools and mixture modeling. Some leading examples can be given as Qiu et al. (2008) study which proposed bayesian hierarchical model where the marginal distribution of different gene clusters is a three mixture of a multivariate normal distribution. This way they made it possible to assign different marginal means and variances for different gene clusters to detect the differentially expressed genes. Likewise, Efron et al. (2001) proposed that the distribution of the gene profiles represents a mixture distribution and hence there is no need for a multiple testing correction. In the mixture distribution, one component represents the differentially expressed and the other component represents the suppressed set of genes. Pan (2002), He (2004), Do et al. (2005), McLachlan et al. (2006) and Broët et al. (2002) can be reviewed for more detailed discussions on this approach. Another foregoing study on mixture modeling is Najarian et al. (2004) study which suggested a nonparametric mixture model method improving over the classical nonparametric mixture model method by increasing the repeatability of the output obtaining similar results on different fits of the model as well as reducing the sensitivity of the output on the parameters. In addition, Moser et al. (2004) used a mixed model approach that clusters the *gene expression* X *immunological status* interactions by a mixture of normal distributions for a short time series gene expression data. Therefore, differentially expressed genes and others took place in different components of the mixing model. The last study that worthed mentioning here as an example for utilizing mixture distributions is Celeux et al. (2005) study which implemented a mixture of mixed models in order to cluster gene expression profiles. Rather than testing gene expression profile significancies over time, they focused more on model based clustering of profiles trying to detect the number of components of mixing models. In terms of classification of probe sets, mixture modeling can be a very plausible application such that it may help assigning probe sets to different mixing components according to their expression values.

In addition to the mixture modeling studies mentioned above, Bayesian hierarchical modeling for assessing the the level of gene expression was applied by Broët et al. (2002). In contrast with Efron et al. (2001), they especially stated that the representation of genes with components of mixture distributions is a binary fashion and the level of expression

should be incorporated. They also compared their results with classical t-test approach and showed their improvement. They also stated downsides of the classical approaches particularly when it comes to testing. The points mentioned in Broët et al. (2002) were taken into careful consideration and some results on multiple testing corrected results were also discussed in this study. They concluded that the data were eligible for further analyses. RMA method was proposed as the preprocessing method. The work by Broët et al. (2002) lacks the analysis of time series microarray data and specific time profiles.

2.2.2 Mixed Effects Models

The second approach that dominates the the field of microarray analyses is the mixed-effects modeling. Wernisch et al. (2003) proposed a mixed-effects modeling approach especially designed for taking the correlation structure among replicates. They used an ANOVA based method and showed the advantages of the proposed model over the t-test at gene level. Wolfinger et al. (2001) used mixed models approach to directly control over the rate of true positives and claims an improvement on false negatives compared to the existing methods. Unfortunately, they did not focus on short time series gene expression profile analysis. They used genewise t-test approach and made use of Bonferroni multiple testing correction which is outdated in favor of Benjamini-Hochberg multiple testing approach. All the studies in this paragraph lacks modeling of short time series microarray data and focus more on comparison studies with replicated microarray data.

There can be found many linear mixed-effects procedures in the literature, some of which focus more on clustering. A technical report of this kind is prepared by Eng et al. (2008). They propose a mixed effects model in order to cluster genes according to the relative likelihood ratios for grouping parameters. Although they propose the model for time course microarray data, their main point of interest is not the short course experiments. Besides, they assume non-diagonal covariance matrices for grouped gene sets. On the other hand, rather than gcRMA, they quantile normalize within group data and then assume normality. Their performance criteria is misspecification performance rather than profile testing in opposition to the method proposed in this dissertation. They treat the unknown parameters as missing values and apply EM algorithm for estimation. They also test the robustness of their proposal by testing against candidate models. Since, gcRMA technique is applied to the raw data in this study and the distribution of each array is equalized, robustness is not a big issue.

Mixed-effects models are very flexible to use and can be used in a vast environment of applications. An example as an alternative application in the gene expression pathway analysis is Wang et al. (2008) study. They perform tests in order to detect differential gene expression pathways. However, their null hypothesis is that differential expression between two particular groups of genes does not differ significantly from the genes in the pathway compared with the rest of the genes. In contrast with their hypothesis, the model proposed in this thesis compares control groups to treatment groups, consecutive time points and different cell lines with each other and tests whether they are differentially expressed in consecutive time intervals.

Another remarkable study on time course microarray data is given in Wang et al. (2009). They proposed a mixed effects model in order to model the variability around the mean gene expression profile. They are testing for the changes over a pathway and testing for the null hypothesis that the average gene expression of a gene group is not differentially expressed over time. They included independent random variables for array and differing covariance effects between genes. The model used in their study seems to be similar to that is used in this study however the application is different. Their method do not involve clustering and unable to detect acute and chronic changes over time as well as testing for the differential expression at a specific time interval.

2.2.3 Other Models and Methods

Alternative to the clustering based methods, model based microarray data analysis methods depend their inferences on either statistical tests or statistical models. Recent studies based on statistical tests and models has become more common. For example, Xu et al. (2002) tried to model the gene expressions by regression models using variables such as time, dose, cell/tissue type. Park et al. (2003) introduced a new statistical test procedure based on repeated measures analysis of variance to identify differentially expressed genes in time series experiments. Hong & Li (2006) calculated probabilities for each gene by hierarchical models that use information from each gene, and proposed identifying genes whose expressions change over time. However, this approach is also based on splines and require long time series data. Hidden Markov Models (Schliep et al. (2003); Zeng & Garcia-Frias (2006)) could not overcome the general restrictions of clustering methods although these are model based clustering methods that are frequently used in time series analyses. Not necessarily, model based methods may still include clustering schemes, however,

clustering may be used as a preparatory tool or for simplification purposes as in this study. Inferences though are based on statistical tests or model significancies.

As an example, Ng et al. (2006) focuses on random-effects model to cluster genes from a time-series with or without replication. They studied an extension of normal mixture models in order to model correlated and replicated measurements. They made use of linear mixed-effects model for the mixture components to be able to incorporate the covariance structure. They offered a general framework for clustering of genes with correlated measurements with replicates that can also be time-course data. Ramoni et al. (2002) used a model based clustering approach based on constant coefficient autoregressive curves. Autoregressive curves require evenly spaced regular series, and this method unfortunately is useless for unevenly spaced microarray data which is quite common in microarray studies. On the other hand, Bar-Joseph et al. (2003) proposed spline estimation model for estimating missing data in time series gene expression datasets. Their model accounted for unevenly spaced time points as well. The spline method provided in their study incorporated some spline coefficients for the gene sets in the same cluster as the cluster covariates as well as subject-specific parameters. They were more focused on unobserved time points. More recently, Furlotte et al. (2011) proposed using linear mixed models to estimate confounding effects and also to measure pairwise correlation of genes. Their proposal however does not directly related to the analysis of short time series data. Yet another study focused on short microarray time-series data analysis using gene-specific linear mixed models to test group effects together in experiments involving two color microarrays is Passos et al. (2011). They modified the design matrix in order to be able to handle two color property by the proposed mixed-effects model. Their aim is to make the comparison of one or two color microarrays cost efficient. They used a linear mixed model for analyzing time series gene expression data and tried to find out the effects of premises on the cost of comparing different arrays.

Handling the unevenly spaced time points and proposing a more plausible model fitting alternative, a quadratic regression modeling method was proposed by Liu et al. (2005) for detecting differentially expressed genes in a short time series microarray data. Their work is one of the rarest works that accounted for the time as a contiuous variable rather than treating it as a factor measured in sequential equally spaced timepoints. They pointed out that taking time as a continuous variable preserves actual time information. They incorporated time effect as a second order term in the model, and fit the quadratic

regression for every gene testing the significance of time at each step. The method is very useful for detecting significant gene expression patterns over time and similar patterns, however, unable to take into account the differing experimental conditions such as different tissues and different treatments.

The common and straightforward approach while testing the significance from many genes from a microarray is to conduct a standard t-test for every single gene or gene group. If the null hypothesis is rejected in the favor of a significant change in the gene activity, the researcher concludes that the gene is differentially expressed. As stated in Qiu et al. (2008), one downside of this approach is that the test statistic requires a variance component and accordingly a standard error estimate for the mean estimator on each gene. Especially, in short time series microarray data the available sample size for each gene is very small for a good variance estimation which is also the case in this study. Although there are some approaches proposed for a stabilized variance estimation for the t-test (e.g. Significance Analysis of Microarrays (SAM), proposed by Tusher et al. (2001)), multiple testing problem occurs during testing many hypotheses which is a common problem also for all other testing procedures. Multiple testing correction is used as a remedy for this problem and helps controlling False Discovery Rate (FDR), but it also brings new problems aside (like uncontrolled true positive rate).

Methodologies and studies on microarray analyses are growing rapidly and sometimes hard to keep up. Tai & Speed (2005) summarized statistical analysis techniques of short time series microarray data. They mentioned especially downsides of applying methods for cross-sectional data to the longitudinal data and emphasized on the effect of underestimating the subject specific variance. They compared classical F-test, moderated F-statistic, B-splines and clustering. However, they did not present solid results. Although their study did not go beyond a literature review, it is useful to look up for a collection of methods. Mutarelli et al. (2007) also used a B-spline basis to model genewise expression pattern and hence performed a classical F-test for a single gene. However, their approach did not take time as a continuous variable but rather as a factor.

Other than the classical testing procedures as well as handling time series microarray data, Sasik et al. (2002) described a model that first analyzes the time-course raw data. They therefore, reduced the dimension of the data representing it by only vital components that characterize the gene expression profiles. They then superficially clustered the components

to visually analyze the differentially expressed profiles. They neither offered a parametric model that accounts for treatment, array type, time effect nor for a single probe set.

Another point of view in the analyses of microarray gene expression data is to select a convenient comparison procedure. He (2004) dicussed the advantages and downsides of parametric and nonparametric test procedures for detecting differentially expressed genes for especially comparing two groups (e.g. tissues) and proposed a weakly parametric model namely a spline function approach to characterize the distributions of differentially and non-differentially expressed genes. In another study published in the same year, Broët et al. (2004) proposed a new strategy for comparing more than two groups based on a flexible mixture model for the marginal distribution of a modified F-statistic. Their model utilizes a combination of false positive and false negative discovery rates in order to select the differentially expressed genes.

Some of the studies that have been mentioned so far have focused on cross-sectional data where the methods used are not suitable for time series microarray analyses since they do not account for the factors that change over time and do not deal with the correlation between measurements (Xu et al., 2002). Nymark et al. (2007) provided a special algorithm including canonical correlation and gene onthology analyses to differentiate the profiles of short time series gene expressions from three different types of cell lines that were exposed to Asbestos. They used permutation tests to identify differentially expressed genes in short time series clusters. Using Nymark's reference Asbestos dataset in his studies, Korpela (2006) studied the short time series microarray data in terms of data quality and clustering. The study puts some insights on the reference dataset by exploring the clustering algorithm and data quality control studies. Another study that put insight to correlational analysis is He & Zeng (2006) which presented a new method namely trend correlation for identifying functional linkages between genes. The method is a two-step method for comparing gene expression profiles over time. Their method does not involve short time series and the exampled series consisted of 17 timepoints.

Trend testing is another approach for detecting the change in gene expression profiles. In a methodological study, Chen (2005) proposed C&G statistic in order to test the significance of individual gene expression profiles. C&G statistic combined Bartlett's C-statistic given in Bartlett (1966), that is used to test for the existence of trends, and fisher's G statistic given in Fisher (1929), for testing the significance of harmonic series. The method was useful for

testing the significance for genewise time series expression data. However, his method did not account for unevenly spaced time points and different tissue cells.

The last but not the least method to identify differentially expressed genes is to fit a fixed effects linear model independently for every subject in the data. One of the most cited articles in the analyses of data from microarray experiments is Smyth (2004) which is the reference paper for microarray gene expression data analyses package for R, namely the Limma package. Smyth (2004) handled the problem as a multiple testing problem of many genes. Every set of genes was treated as a single set of data and then a hybrid of classical and bayesian approach was used and a linear model was fitted to the expression data for each gene. Actually, prior distributions were defined in order to correct on the estimates of the parameters that are called empirical Bayesian estimates. Therefore, moderated t-statistics were obtained for every single gene. Empirical Bayesian approach empowers Limma even when testing with small sample sizes. Although the model is computationally straightforward compared to other methods, it is only possible to treat a timepoint as a factor level with Limma. Limma cannot handle continuous covariates. Therefore, it loses time information and suffers modeling the expression profile over time. Another major disadvantage of Limma is that it cannot handle unbalanced designs. The number of replicates has to be the same at all levels of the experiment for a gene. Mixed modeling cannot be incorporated to Limma either.

CHAPTER 3

METHODOLOGY

A multiple stage data processing and analysis were proposed and used throught this study. The data used in all analyses were preprocessed after being scanned from a microarray device. The processes applied and presented in sequence in this chapter are preprocessing, especially normalization as part of it, filtering, clustering and model fitting. The applications of the proposed methods were done on the real life dataset studied and referred by Nymark et al. (2007) and will be named as "asbestos dataset" hereafter. The asbestos dataset have 54,675 probe sets, measured from two cell types namely A549 and Beas2B. Measurements on each cells have both control and exposure groups that were collected on six time points.

3.1 Preprocessing

Although microarrays are very high technology devices, the process of reading gene activities as an image from the device and changing this image to an appropriate dataset in meaningful metrics for further analyses require some preparation steps. That is because, the raw data is very likely to be perturbed by environmental effects, such as the noise on the obtained image, biological noise due to organic activities and the noise that occurs during experimenting for replicates. Besides, all organic cells may not be identical even if they belong to the same structure. For a fair comparison, microarray data should be corrected to reduce any possible noise effect. On the other hand, the data collected from the device are not in the suitable domain and scale for further statistical analyses. All these factors lead researchers to microarray data preprocessing. The common steps and operations are given in fair detail as follows.

3.1.1 Data Preparation

The data obtained at first stage from a microarray are the measures of the intensity of light. Before a microarray experiment is conducted, it is set up. Setting up of a microarray experiment can refer to both preparing physical conditions for the experiment and statistical design of the experiment such as deciding on the number of replicates. The best way of designing an experiment is attained by biologists and/or geneticists working together with statisticians. A biologist describes the biological nature of the experiment where the statistician describes the statistical nature.

Microarray experiments and accordingly the data collected from the experiment are mostly studied for determining different sources of variation. The researcher can be interested in comparing two donors, different animals or organisms of the same type. Likewise, treatment of different drugs, or treatment of a single drug can be subject to interest. A researcher may want to compare even completely different tissues. Therefore, the experiment and the experimental conditions are set up in this direction.

Controllable sources of variation in these studies such as gender, age, duration of treatment and many others are selected such that they will not create bias throughout the study. However, unfortunately, the experiment and the data collected thereafter are also affected by uncontrollable sources of variation from environmental variables such as the measurement system and sampling errors due to technical or biological replication.

Although experimental conditions can be set up such that the resulting measurements are evenly affected by the uncontrolled variation, most of the time it is very difficult to do this in real life because of financial and experimental restrictions. Also the measurement devices incorporate some noise to the system which is inevitable. All these effects are to be reduced or eliminated in the microarray data preprocessing steps to obtain the realistic dataset.

3.1.2 Normalization

Normalization enables the researchers compare two or more arrays from different populations such as case and control, cell type 1 and cell type 2, within cell type and between cell types. Normalization also reduces the effect of the variation from external sources during the experimentation and collection of the data. Some examples for the external variation are the chemical substances on the surface of the microarray, the way

that the spots on the microarray are prepared, methods for labeling, hybridization techniques, image analysis and isolation of the RNA (Bilban et al. (2002); Claverie (1999); Schuchhardt et al. (2000); Lou et al. (2001); Tseng et al. (2001); Yue et al. (2001)).

Further processing is applied to the raw data. As a preprocessing procedure gcRMA technique by Wu et al. (2004) has been used in this study. gcRMA is an array normalization method which also does background noise correction.

Different normalization methods have different data processing capabilities. Studies so far have shown that RMA is a very successful method in normalization (e.g. Bolstad et al. (2003)). Irizarry et al. (2003) has shown that RMA method is better than other normalization methods. As an extension to RMA method, gcRMA has produced very accurate results without any loss of precision (Irizarry et al. (2006), page 793). A comparative study for the normalization procedures are studied in Lim et al. (2007) and Shedden et al. (2005).

Main advantages and common properties of gcRMA can be listed as follows:

- Corrects the background noise using the mismatched gene sequences.
- Equalizes the distribution of each array.
- Uses robust median polishing procedure.
- Makes use of quantile normalization.
- Proven to be better than competing alternatives such as RMA and MAS5 (see Wu et al. (2004)).
- Returns expression values on $\log_2$ scale.

The main difference and major improvement of gcRMA technique over the RMA is that it uses a linear model to represent the summarized gene expression values. This is the major improvement of gcRMA over RMA. Wu et al. (2004) proposes below statistical model for background adjustment:

$$PM = O_{PM} + N_{PM} + S \tag{3.1}$$

$$MM = O_{MM} + N_{MM} + \varphi S \tag{3.2}$$

where

PM Perfect match (stands for probe pairs all of which have correct nucleotide matchings)

O_{PM} Optical noise for perfect match

N_{PM} Non specific binding effect for perfect match

S A quantity proportional to RNA expression (the quantity of interest)

MM Mismatch (stands for mismatching probe pairs)

O_{MM} Optical noise for perfect match

N_{MM} Non specific binding effect for mismatch

φ A coefficient that lies between 0 and 1. The φ proportion of the mismatched probe pairs is assumed to be true signal.

$O \sim LogN$ and $\log(N_{PM}), \log(N_{MM}) \sim Bivariate\ Normal$ where $\mu = \begin{bmatrix} \mu_{PM} \\ \mu_{MM} \end{bmatrix}$ and $var[\log(N_{PM})] = var[\log(N_{PM})] \equiv \sigma^2$ and there is also a constant correlation ρ between probes. Means μ_{MM} and μ_{PM} are smooth functions of the linear combinations of the means of probes and their bases. Because non specific binding is expected not to affect the optical noise, *O* and *N* are assumed to be independent.

Above parameters are estimated from the data and the problem then changed into the prediction of *S*. Estimation is not a big deal since generally microarray datasets are large enough for such a purpose.

Wu et al. (2004) makes two important assumptions about the above model such that $\varphi = 0$ and *O* is an array dependent constant. Then they offer both frequentist (MLE) and Bayesian alternatives for estimating the *PM* and *MM* parameters and finally they end up with summarized gene expression levels by the use of the following model:

$$\begin{aligned} Y_{gij} &= O_{gij} + N_{gij} + S_{gij} \\ &= O_{gij} + \exp(\mu_{gij} + \epsilon_{gij}) + \exp(s_g + \delta_g X_i + a_{gij} + b_i + \xi_{gij}) \end{aligned} \tag{3.3}$$

Y_{gij} Probe intensity for the probe *j* in the probe set *g* on the array *i*.

ϵ_{gij} Non-specific binding error term which has a normal distribution. It accounts for the noise of the same probe that behaves differently in different arrays.

s_g The baseline log expression level for probe set *g*.

a_{gij} The effect of probe *j* in gene *g* on array *i*.

b_i The array affect that requires normalization.

δ_g The coefficient of covariate *X* to be estimated. The emphasis of normalization process is on this parameter.

Above model requires elaborate computation for the parameter estimation for both MLE and Bayesian methods.

Most of the microarray gene expression analysis studies indicate that majoriy of the genes remain relatively less active during the experiments. Only a smaller portion of the genes generally show changing expression values. Therefore, those less active, namely underexpressed or supressed genes can be taken out of the analyses. In this study, all of the genes process were included in the normalization, which is called *global normalization*, but inactive genes were filtered by *kOverA* function (see section 3.3) afterwards just before the statistical analyses (Bilban et al., 2002).

3.1.3 Quantile Normalization

gcRMA technique is similar to the RMA other than the gene expression summarization procedure. It makes use of quantile normalization after summarization. Quantile normalization is used to make the distribution of each array identical. It is called "quantile normalization" because this method equalizes the quantiles of gene expression measures from each array.

Assume that there are *n* arrays and *p* probe sets from each array. The method is applied as follows:

1. Calculate *k*th quantiles for each array such as $q_k = (q_{k1}, q_{k2}, \dots, q_{kn})$ for $k = 1,2, \dots, n$. Practically, order all arrays in ascending order of magnitude. Therefore, *i*th order statistics are found for each array. Note the original ordering of each array for later use in step 4.
2. Find the means of every *i*th order statistics across all arrays. Therefore, *n* means are obtained from *n* arrays.
3. Substitute every *i*th order statistic from all arrays with the *i*th mean calculated in step 2. Therefore, the *i*th order statistic of each array is equal to the *i*th mean.
4. Reorder each array to its original ordering.

Therefore, all quantiles of each array are equalized. One downside of this procedure is that it may cause replicated gene expression values on the tails of the array distribution. However, Wu et al., (2004) stated that this is not a problem since probeset values are calculated by using more than one probe.

An example of the distribution of a sample dataset before and after the normalization are shown in below graphs. Box plots for randomly selected 10 arrays from Asbestos dataset is given in Figure 3.1 indicate very heavy tailed skewed distribution for different arrays which reduced the readability of the graph. Even the median and the left tail is impossible to see. In order to increase the readability, graph was trimmed on the Y axis for better visualization and given in Figure 3.2. The distributions of the arrays after normalization is sketched in Figure 3.3 that indicates the distributions of arrays became almost identical.

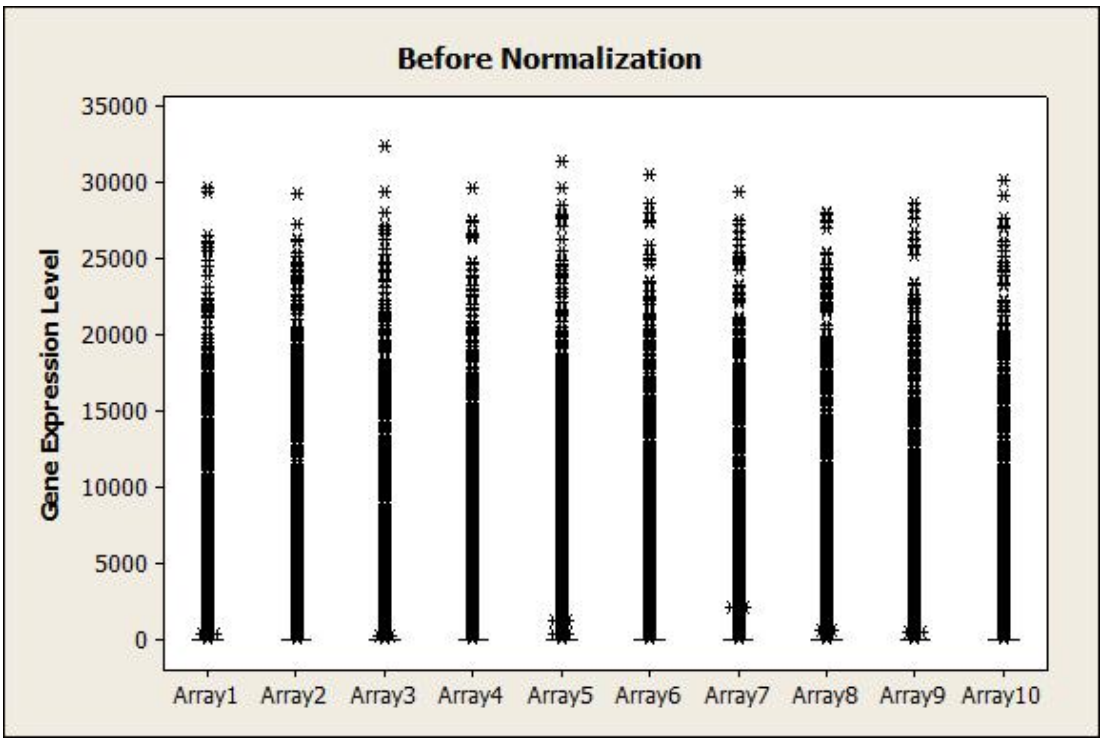


Figure 3.1 Boxplots of some arrays from Asbestos dataset before normalization

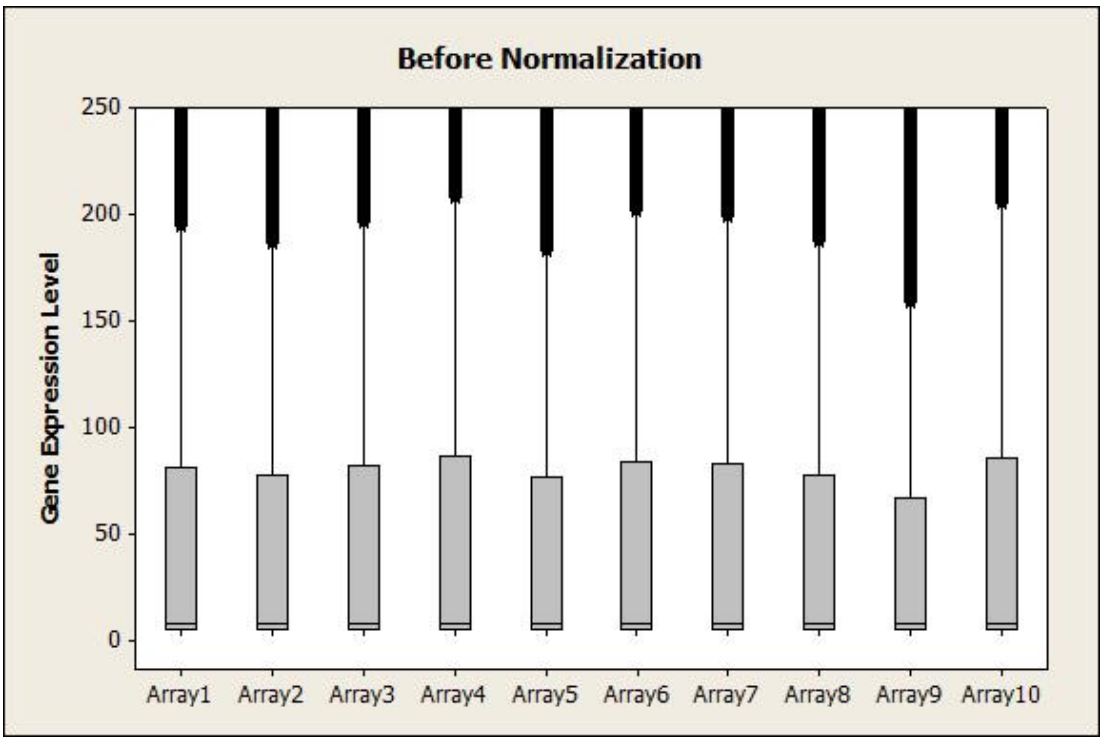


Figure 3.2 Some arrays from Asbestos dataset before normalization (trimmed)

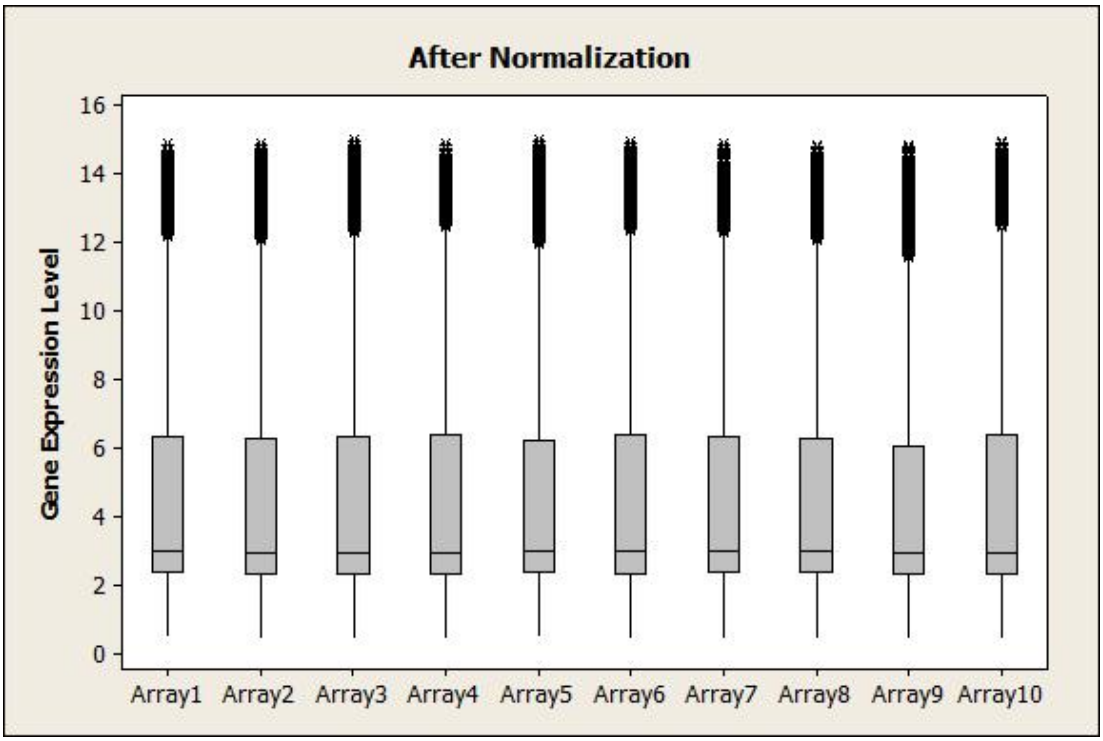


Figure 3.3 Some arrays from Asbestos dataset after normalization

3.2 What do the data look like?

When a microarray is scanned, the resulting image is saved as a raw DAT file. The raw image file then is translated to numerical values and saved into a .CEL file which is done in the quantification step. This process is basicly the reading of the pixel intensity values and changing them into real numbers. However, a single pixel may not represent a single feature such as a probe or a spot. Instead, more than one pixel or a group of pixels represent a spot. The content of a CEL file can be seen in Figure 3.4.

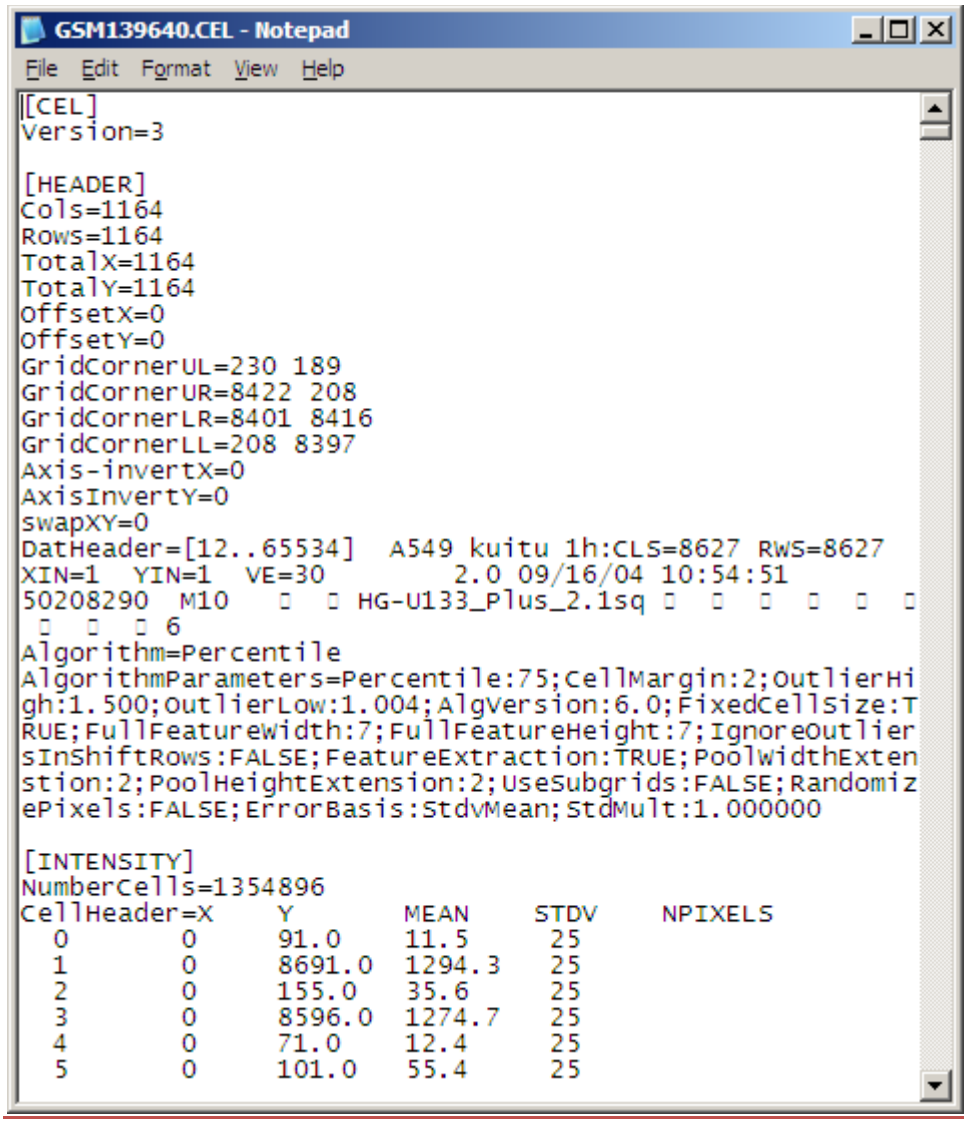


Figure 3.4 CELL file content

The real values of the pixel intensities start after the [INTENSITY] line. There are also some summary information about the intensity data. This data belong to a single array or a single

cell line. However, in a microarray experiment there can be more than one array or cell line. As a result of this fact, datasets from each CEL files need to be combined as a microarray data matrix and further processes are required.

3.3 kOverA Filtering

Statistical studies over gene expression data analysis show that statistical methods for determining differentially expressed genes are more successful when we omit low expressed genes. Low expressed genes are most of the time indicators of no activity. kOverA function in R by Gentleman et al. (2009) removes genes that have expression values lower than a specified gene expression level, which is the threshold. By incorporating a condition number, it is optional to select the least number of arrays with expression values higher than the threshold to decide whether or not to remove the gene. For example, if the condition number is selected as 1 and the threshold as 5, kOverA would not remove any gene that has gene expression level above 5 on at least 1 array. If the condition number was chosen as 2, the function will require at least 2 measurements with expression levels higher than 5 not to remove that gene. The threshold depends on the selection of the analyst. In this study, the threshold value was selected as 3.5 throughout the gene expression profile. Therefore, a gene is removed from the analyses if it has an expression value below 3.5 at every point of measurement.

3.4 Clustering

Clustering is a way of splitting the data into groups according to a predefined criteria. As there can be many number of different clustering criteria, there are also many different data structures, e.g. longitudinal, cross-sectional, etc.. Clustering is one of the major solutions to dimension reduction problem. This study required clustering in order to reduce the dimension of the data for further statistical analyses. One other main reason is that grouping the similar probesets help biological interpretation of the results. Scientists would like to identify genes or groups of genes that show similar behaviour under similar conditions.

The probe sets which are the subjects in our study were grouped together according to both their gene expression levels and gene expression profiles. All of the measurements from each probe set were treated as a vector of observations and were grouped in the same cluster. Therefore, there is no chance that any two measurements from the same

probe set belong to different clusters. Genes for which the expression levels over time are close to each other have been grouped as a first step. Then, within these groups the gene expression profiles over time are examined. Genes, whose expression profiles sketch a very similar pattern over time are regrouped. Therefore, the grouping procedure is a two step process.

There are two very well known data partitioning methods, namely supervised learning and unsupervised learning. *Supervised learning* helps to assign objects to predefined groups, also referred as *classification*, whereas *unsupervised learning* help to assign the groups to the objects, also referred as *clustering*. Unsupervised learning bases the determination of the groups on the data.

Genes (or probe sets as their representatives) that have similar gene expression profiles over time are to be grouped for a better understanding and statistical purposes. However, there is no predefined groups or group categories that we can place the probe sets into. Therefore, grouping must be done due to the behaviour of the probe sets. This fact makes the context of this study take place in the unsupervised learning part of above methods, namely clustering.

Furthermore, clustering methods are also mainly grouped into two major categories such as *partitioning methods* and *hierarchical methods*. Partitioning methods require the pre-determination of the number of clusters. On the other hand, hierarchical methods require the pre-determination of a clustering criterion and the number of clusters are the count of the clusters satisfying that criterion (Dudoit & Gentleman, 2002).

3.4.1 K-means Clustering

K-means clustering is used to cluster observations according to their magnitudes. At the end of a k-means cluster analysis a researcher should expect to obtain clusters, in each of which observations have relatively similar values. K-means clustering is an unsupervised learning method and therefore, it does not require the clusters predefined. However, it requires a prior knowledge, or at least, a prior idea about the number and centers of the clusters. Namely initial clusters must be defined in order to be able to obtain the final clusters. This is the main drawback of this method. On the other hand, final clusters obtained by k-means don't have any hierarchy amongst them. K-means is a partitioning method in clustering.

k-means algorithm tries to minimize the sum of squared distances between a cluster center and the observations in that cluster. This is called minimizing the within cluster sum of squares. (Tou & Gonzalez, 1974) describe the k-means algorithm as follows:

1. Choose *k* initial cluster centers (also called centroids) from the set $S = S_1(1), S_2(1), \dots, S_k(1)$.
2. Distribute the sample $x_1, x_2, \dots, x_n$ to the *k* clusters so as to minimize the within cluster sum of squares:

$$\underset{S}{argmin} \sum_{i=1}^{k} \sum_{\substack{j=1 \\ x_j \in S_i}}^{n} \left\| x_j - \mu_i \right\|^2 \tag{3.4}$$

   Note that the clustering can also be applied to vectors instead of individual observations. In that case the sum of squares measure is simply the euclidean distance since the 2-norm is used.

3. Recalculate the *k* cluster centroids.
4. Repeat steps 2 and 3 until convergence.

The k-means clustering algorith is a special case of the well known Expectation – Maximization (EM) algorithm. The step 2 above is the E step where step 3 is the M step. Assignment of observations to the clusters in step 2 at any iteration must satisfy the criteria

$$x_j \in S_i \; if \; \left\| x_j - \mu_i \right\| \leq \left\| x_j - \mu_{i^*} \right\| \; for \; all \; i^* = 1, 2, \dots, k \tag{3.5}$$

and the new cluster centroids in the M step is calculated as

$$\mu_i^{next} = \frac{1}{|S_i|} \sum_{\substack{j=1 \\ x_j \in S_i}}^{n} x_j \tag{3.6}$$

The number of final clusters is though is very controversial. It is one of the major problems of nonparametric estimation in statistics. The *hist* function in R which is based on the procedure given in Becker et al. (1988) and Venables & Ripley (2002) was used to estimate the number of k-means clusters.

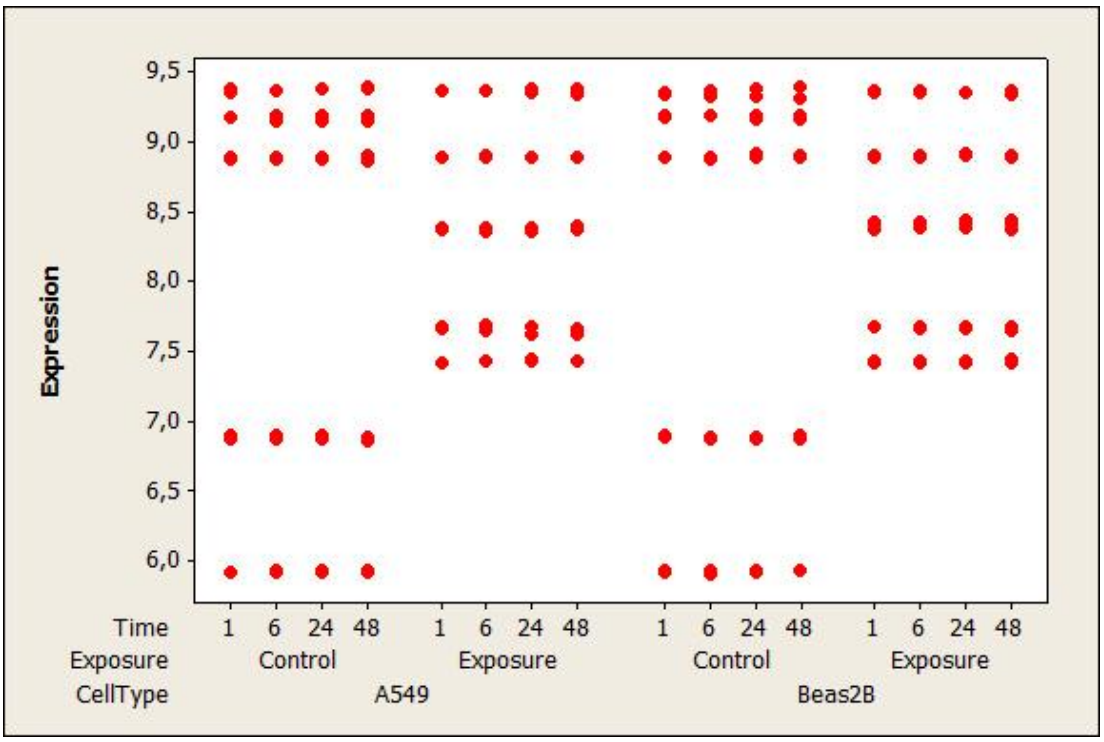


Figure 3.5 Sample gene expression profiles from Asbestos dataset before k-means clustering

The idea behind k-means clustering is to group genes that have similar expression profiles in terms of magnitude. An example of clustered version of Figure 3.5 is given in Figure 3.6 where two of the k-means clusters are shown. The first graph in Figure 3.6 to the left contains measurements from cell type 1 and the second graph to the right contains measurements from cell type 2. The panels of each graph contains measurements from control and exposure groups respectively. The members of the first cluster are shown in circles and the members of the second cluster are shown in filled squares. Probes are enumarated and shown on the graph as data labels. According to the k-means clustering results probe sets 1, 5 and 6 were clustered in the second cluster and probe sets 2, 3 and 4 were clustered in the first cluster. k-means clustering was applied in such an algorithm that all of the measurements of a single probe set were represented in the same k-means cluster regardless of the cell type, exposure, time point and replicates.

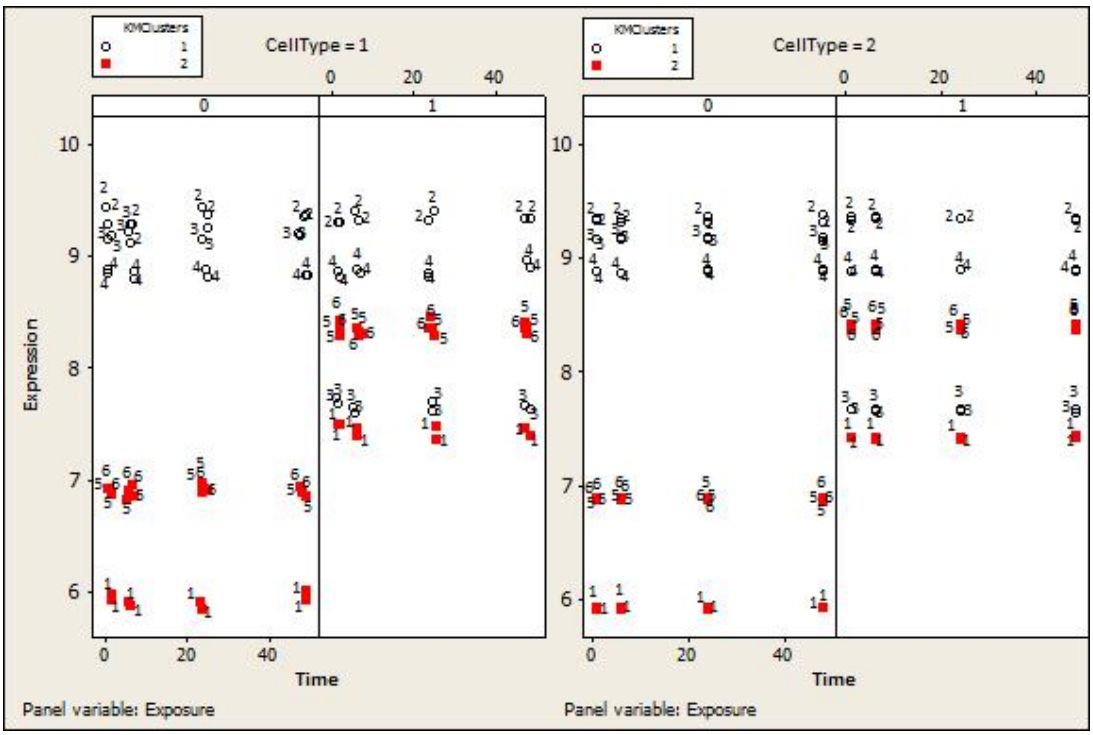


Figure 3.6 Sample gene expression profiles from Asbestos dataset after k-means clustering

3.4.2 Hierarchical Clustering

Hierarchical clustering is another clustering approach in unsupervised learning. Hierarchical clustering was applied to the k-means clustered data as the second stage of clustering in this study. Unlike the k-means clustering, the number of final clusters is not predefined. Instead, a criteria is defined for the similarity or the distance (dissimilarity) between the members of each cluster. Besides, a distance measure between cluster centroids, so called *linkage method*, is also used to distinguish and reflect the shape of the clusters. The more members satisfy the similarity criteria and join into a cluster, the less the number of clusters in general.

There is a tree-like hierarchy between the clusters as a result of the fact that items satisfying a similarity criteria group into the same cluster. This hierarchical structure can be formed by either **divisive** methods, so called *top-down* methods, or **agglomerative** methods, so called *bottom-up* methods. Both methods are summarized below:

3.4.2.1 Divisive (top-down) algorithm

1. Select a similarity measure within clusters and between clusters (linkage method). Define a threshold (criteria for similarity) for within cluster similarity measure.

2. Start with a single cluster and assign all items into this cluster.
3. Continue splitting clusters until every cluster satisfies the criteria in step 1.

Sustaining and tracking the main structure of the data is an advantage for this method. On the other hand, considering all the possible divisions of the groups is a computational disadvantage. An illustration for divisive algorithm can be found in Figure 3.7.

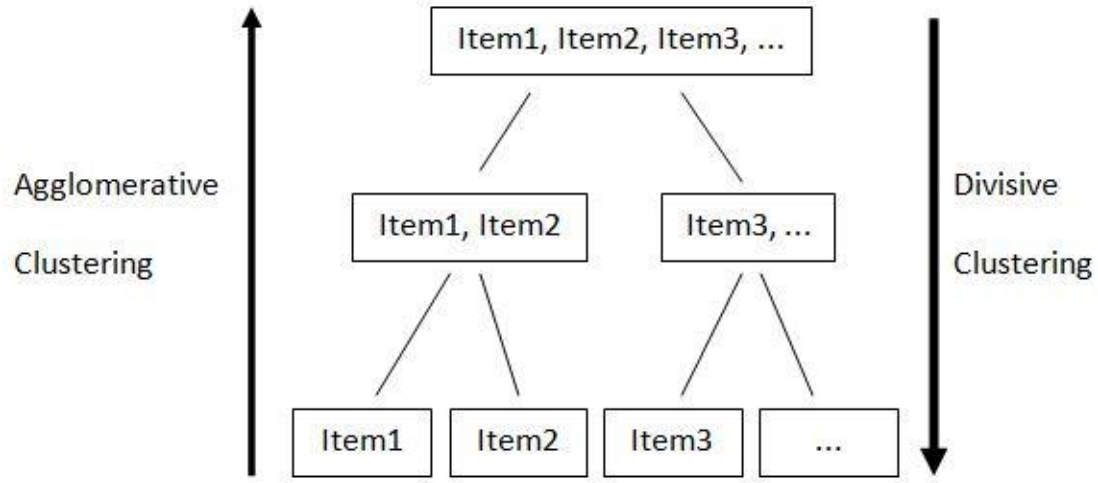


Figure 3.7 Illustration of divisive and agglomerative clustering

3.4.2.2 Agglomerative (bottom-up) algorithm

1. Select a similarity measure within clusters and between clusters (linkage method). Define a threshold (criteria for similarity) for within cluster similarity measure.
2. Start with as many clusters as the number of items (such that every single item is a single cluster).
3. Continue amalgamating clusters until every cluster satisfies the criteria in step 1.

Dendrograms are graphs that are used to illustrate the hierarchical clusters. A sample dendrogram is given in Figure 3.8. Items that have smaller distance values are grouped in the same cluster (e.g. Probe_2 and Probe_27). The grouping depends on the cutoff value for the dendrogram tree. More detail on this will be given in section 3.8.

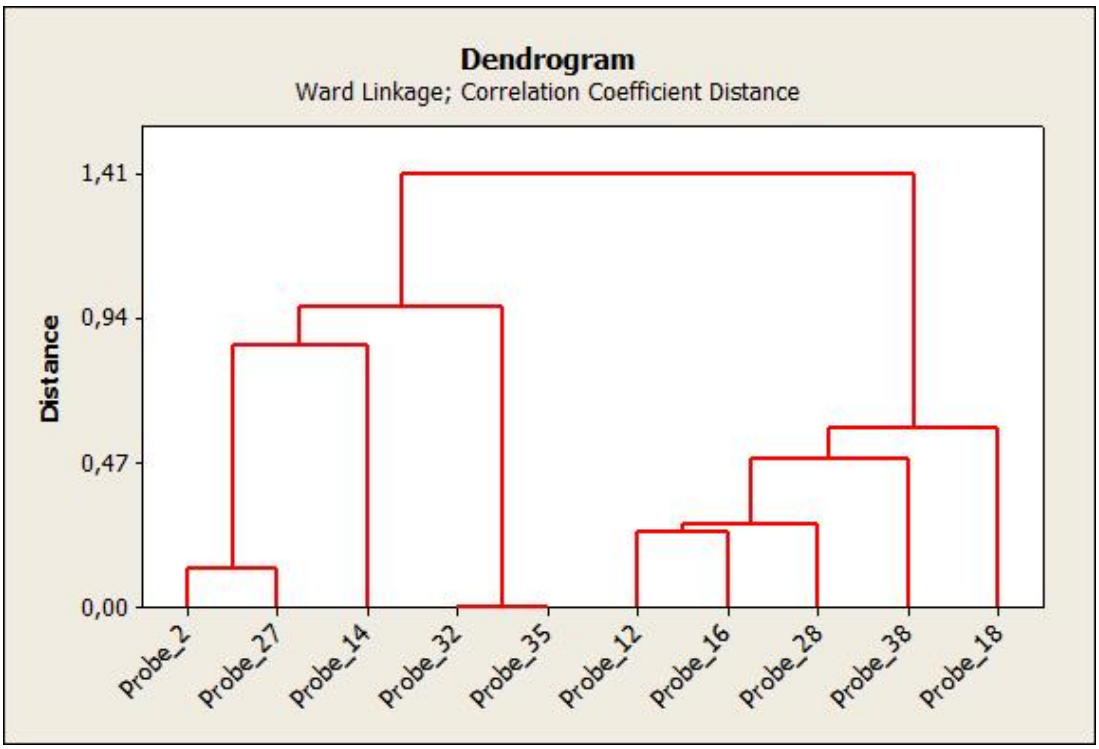


Figure 3.8 Dendrogram of clusters with Pearson correlation distance using Ward's linkage

3.4.2.3 Distance (Dissimilarity) Metrics For Vectors in a Cluster

Distance is a measure of how far two items are. Items can be points or vectors. There is a very close relationship between the distance and similarity. As the distance between two items decrease, their similarity increase. If the distance between the i*th* and the j*th* items is d_{ij} and the largest $d_{ij}\ is\ d_{max}\ for\ all\ (i,j)$ then the similarity can be defined as $s_{ij} = 100(1 - \frac{d_{ij}}{d_{max}})$. There can be found different measures for different purposes in the literature. Most commonly used ones are Manhattan (also known as City-Block Distance, L1 Norm), Euclidean (also known as L2 Norm), Mahalanobis, Pearson correlation, Spearman rank correlation and Absolute or squared correlation. These distance metrics for two vectors $X = (x_1,\ x_2, \dots, x_n)$ and $Y = (y_1,\ y_2, \dots, y_n)$ are defined as follows.

Manhattan distance:

$$d = \sum_{i=1}^{n} |x_i - y_i| \tag{3.7}$$

Euclidean distance:

$$d = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{3.8}$$

Mahalanobis distance:

$$d = \sqrt{(X - Y)^T \Sigma^{-1} (\mathrm{X} - \mathrm{Y})} \tag{3.9}$$

where Σ is the variance-covariance matrix of *X* and *Y*.

Spearman Rank Correlation distance; $d = 1 - r_{XY}$ where

$$r_{XY} = \frac{\sum_{i=1}^{n} \left(x_{(i)} - \bar{x}\right)\left(y_{(i)} - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(x_{(i)} - \bar{x}\right)^2 \left(y_{(i)} - \bar{y}\right)^2}} \tag{3.10}$$

with centered sample moments and $x_{(i)}$ and $y_{(i)}$ being ranks.

Absolute or squared correlation:

$$d = 1 - |r_{XY}| \quad \text{or} \quad d = 1 - r_{XY}^2 \tag{3.11}$$

Pearson correlation was used as the distance metric in this study because it is the most suitable method to distinguish between expression profiles over time. Ernst et al., (2005) suggested the use of correlation distance as it has certain advantages for clustering similar gene expression profiles. Moreover, Eisen et al., (1998) pointed the correlation coefficient as a very successful measure for clustering purposes.

3.4.2.4 Pearson Correlation Distance

Pearson correlation is a commonly used coefficient that measures the strength of the linear relationship between two variables. It is the standardized covariance between two variables. The correlation distance between two vectors $X' = [x_1, x_2, \dots, x_n]$ and $Y' = [y_1, y_2, \dots, y_n]$ is as follows:

$d_{XY} = 1 - r_{XY}$ where $r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$ with centered sample moments.

Please, note that the correlation can be estimated without centering the moments, i.e., by removing the sample mean terms from the equation. In this case, it is called uncentered correlation (also known as angular separation, cosine angle) distance.

3.4.2.5 Distance Between Clusters

As the items are grouped into the clusters according to their similarities, hierarchical algorithms must decide how to split or combine the clusters. This is done by measuring the cluster distance.

3.4.2.6 Cluster centroid

A cluster centroid is the center or the midpoint of a cluster. If there is more than one vector in a cluster, the cluster centroid is the mean of the means of those vectors. If there is another location measure like median for vector representations, then centroid can be calculated by using the median as well. Different applications are possible and an illustration of distance between clusters and cluster centroid is given in Figure 3.9.

For a given cluster, the average of the distances between observations and the centroid is the average distance from the centroid. Likewise, the maximum of these distances is the maximum distance from the centroid.

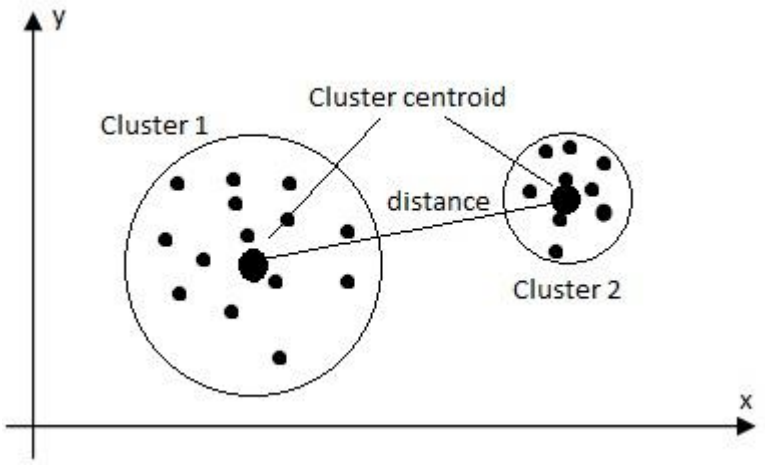


Figure 3.9 Illustration of distance between clusters and cluster centroids

3.4.2.7 Linkage methods

A proper linkage method should be used in order to define how the distance between two clusters will be determined. Most widely used methods are Single, Average, Centroid, Complete, Median, McQuitty's and Ward's linkages. Ward's linkage was used in this study to link the clusters as it tends to minimize within cluster sum of squares.

Ward's distance between two vectors is defined as follows.

$$d = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{2n}(z_i - \bar{z})^2 \tag{3.12}$$

where z_i belong to the combined dataset.

This way the method provides similar profiles grouped in the same cluster. The downside of this method is that it is very sensitive to the outliers. However, this downside is an advantage in this study as it is much preferable. Because, any perturbed measure can be an indication of supressed or overexpressed gene in the given time.

3.5 Mixed Effects Models

Longitudinal data occurs when repeated measurements are observed from the same subject over time. In other words, it is used to model and investigate the change of a feature which is measured repeatedly from subjects in the course of time. Especially in medical studies, the feature that is subject to measurement can be blood pressure, lung volume, cholesterol level, or serum glucose. Likewise in microarray experiments gene expression level is a characteristic that can be measured over time. Each subject can be measured repeatedly at successive times in experimental studies where levels of the factors are controlled by the experimenter. Even if some factors are controlled by the experimenter, there are many uncontrollable factors that affect the measurement variation. The resulting data structure cannot be easily modelled and inferred as there must be some assumptions and restrictions on the covariance matrix.

In the mixed effects models, the distribution of the measurements from every subject is assumed to be identical with varying stochastic parameters. The distribution of the measurements constitutes a stage and the distribution of the parameters is another stage on the mixed effects analyses. Therefore, researchers must account for a multivariate distribution combining together repeated measures from individuals as well as the random

parameters. Laird & Ware, (1982) stated that the marginal distribution of the repeated measurements from subjects is multivariate normal with a special covariance matrix where the linear regression model is fitted for each subject that are conditional on subjects' individual parameters. EM algorithm or Bayesian methods are widely used to estimate such a model that combines random parameters as well as random regression coefficients. One main advantage of this approach is that it doesn't require balanced designs. In other words, approach can also be used when the number of replications from each subjects are not necessarily the same. Laird & Ware, (1982) were influenced from the ideas of Harville, (1977) and they defined the statistical hierarchical model as in Equation (3.13). Hierarchical name describes the two stage of the estimation where the first stage is the estimation of the population parameters, individual effects and within-subject variation, and the second stage is the between subject variation.

General form for the mixed model employed for this study is as follows:

Let $\boldsymbol{\beta}$ denote a $p \times 1$ vector of unknown population parameters and $\boldsymbol{X_i}$ be a known $n_i \times p$ design matrix. Let $\boldsymbol{b_i}$ denote a $k \times 1$ vector of unknown individual effects and $\boldsymbol{Z_i}$ a known $n_i \times k$ design matrix. Usually, Z_i is taken as a subset of X_i and the following model is proposed:

Level 1: For each individual unit, i (individual units or namely subjects in applications of this study are the clusters containing probe sets that have similar gene expression profiles),

$$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{e_i}, \tag{3.13}$$

where $\boldsymbol{e_i}$ is distributed as multivariate normal with mean vector $\mathbf{0}$ and $n_i \times n_i$ positive definite covariance matrix $\sigma^2\boldsymbol{\Lambda_i}$**.** That is shown as $N(\mathbf{0}, \sigma^2\boldsymbol{\Lambda_i})$. $\boldsymbol{\Lambda_i}$ depends on i because it is n_i dimensional, however, the parameters in $\boldsymbol{\Lambda_i}$ do not depend on i (independent from subject) and $\boldsymbol{\Lambda_i}$ is taken as identity matrix in general. At this level $\boldsymbol{\beta}$ and $\boldsymbol{b_i}$ are considered fixed and $\boldsymbol{e_i}$ are assumed to be independent. A representation of this model was given by Lindstrom and Bates (1988) as:

$$\boldsymbol{y_i}|\boldsymbol{b_i} \sim N(\boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i}, \sigma^2\boldsymbol{\Lambda}_i) \tag{3.14}$$

where $i = 1,2,\ldots,C$ represents each subject and C is the total number of subjects (clusters).

Level 2: The $\boldsymbol{b_i}$ are distributed as $N(\mathbf{0}, \sigma^2 \boldsymbol{D})$, independently of each other and of the $\boldsymbol{e_i}$ where $\sigma^2 \boldsymbol{D}$ is a positive-definite covariance matrix. The population parameters, **β** are treated as fixed effects. Therefore, the marginal distributions of $\boldsymbol{y_i}$ are independent multivariate normal with mean $\boldsymbol{X_i \beta}$ and covariance matrix $\boldsymbol{\Sigma_i} = \sigma^2\left(\boldsymbol{\Lambda_i} + \boldsymbol{Z_i D Z_i^T}\right)$. The structure of the data in matrix form is as follows:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_C} \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \\ \vdots \\ \boldsymbol{X_C} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \vdots \\ \boldsymbol{\beta_p} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b_1} \\ \boldsymbol{b_2} \\ \vdots \\ \boldsymbol{b_C} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_1} & \mathbf{0} \\ \ddots & \\ \mathbf{0} & \boldsymbol{\Sigma_C} \end{bmatrix}, \quad \vec{\mathbf{D}} = \begin{bmatrix} \boldsymbol{D} & \mathbf{0} \\ \ddots & \\ \mathbf{0} & \boldsymbol{D} \end{bmatrix},$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z_1} & \mathbf{0} \\ \ddots & \\ \mathbf{0} & \mathbf{Z_C} \end{bmatrix}$$

where $\boldsymbol{y_i}$ are the response vectors, $\boldsymbol{X_i}$ are the fixed effect design matrix, $\boldsymbol{\beta_i}$ are the fixed effect parameters, $\mathbf{Z_i}$ are the random effect design matrix, $\boldsymbol{b_i}$ are the random effect parameters, $\boldsymbol{\Sigma_i}$ are the covariance matrices, $\boldsymbol{D}$ are the covariance matrix components for every *i*th subject.

Therefore, the entire model can be written as $\boldsymbol{y}|\boldsymbol{b} \sim N(\boldsymbol{X\beta} + \boldsymbol{Zb},\ \boldsymbol{\sigma^2 \Lambda})$ where $\boldsymbol{b} \sim N(\mathbf{0}, \boldsymbol{\sigma^2 \vec{D}})$ and the marginal distribution of $\boldsymbol{y}$ is $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \sigma^2\left(\boldsymbol{\Lambda} + \boldsymbol{Z\vec{D}Z^T}\right) = \sigma^2 \boldsymbol{V}$.

The computational part of the applications for fitting mixed effects model to the microarray gene expression data uses the structure above. The general framework of Lindstrom & Bates, (1988) and the model formulation that is described in Laird & Ware, (1982) are theoretical bases. The variance-covariance parametrizations are given in Pinheiro, (1996). These references belong to R *nlme* library which is the software package used throughout the calculations (Pinheiro et al., 2011).

3.5.1 Estimation

The computation and the estimation of the parameters in a linear mixed effects model is very intensive. The ordinary least squares estimates are not a plausible alternative as they are biased although they are very straightforward to handle. The normalization procedures before the gene expression analyses prepare a very applicable basis to the analyses. On the other hand, the hierarchical complexity and the number of parameters to estimate in the model can be described in terms of conditional likelihood functions.

Computational advances in solving such complex models allow us to handle high dimensional data. There are different methods to maximize the likelihood function. Every method has its own estimation procedure and standard errors of the estimates. The computational stages of solving a linear mixed effects model is generally split into three parts: estimation of the fixed effects (i.e. $\boldsymbol{\beta}$), estimation of random effects (i.e. $\boldsymbol{b_i}$), and estimation of variance parameters (i.e. $\boldsymbol{\Lambda_i}$, $\boldsymbol{D}$ that are variance components or covariance terms).

3.5.2 Estimation of the Fixed and Mixed Effects

We can fit the mixed effects model by maximizing the likelihood function which is conditional on the data. In other words, the likelihood function provides the information on how likely the model parameters are given the data and it is defined by using the density function of the observations.

In the classical approach where measurements are independent of each other, the likelihood function is simply the product of density functions of every individual observation. However, in a mixed model setting, measurements are not assumed to be independent of each other and hence the likelihood function cannot be the product of individual densities. The likelihood is the multivariate distribution of the measurements. It is the multivariate normal distribution of $\boldsymbol{y}$ incorporating all the variance parameters and the fixed effects. The variance parameters here cover all the parameters to be estimated in $\vec{\boldsymbol{D}}$ and $\boldsymbol{\Lambda}$. As the expected value of the random effects is $\mathbf{0}$ vector (recall that the $\boldsymbol{b_i}$ are distributed as $N(\mathbf{0}, \sigma^2\boldsymbol{D})$), automatically the expected value reduces to $\boldsymbol{X\beta}$ with covariance matrix $\boldsymbol{\Sigma} = \sigma^2(\boldsymbol{\Lambda} + \boldsymbol{Z\vec{D}Z^T})$. The regular likelihood function based on the multivariate normal density function is then

$$L = \frac{exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})\right]}{(2\pi)^{(1/2)n}|\Sigma|^{1/2}} \tag{3.15}$$

and therefore, the log-likelihood can be written as

$$log(L) = -\frac{1}{2}nlog(2\pi) - \frac{1}{2}[log|\Sigma| + (\boldsymbol{y} - \boldsymbol{X\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})] \tag{3.16}$$

The above likelihood can be used for the estimation of the model parameters. Partial derivatives of the likelihood function yield the normal equations and hence the maximum likelihood (ML) estimators. However, one downside of this technique is that the variance

parameter estimates tend to have downward bias especially for small samples ((Lindstrom & Bates, 1988); (Brown & Prescott, 2006)). The bias is simply caused by the nature of the maximum likelihood method since it does not take into account the loss of degrees of freedom for the estimation of fixed effect coefficients (i.e. $\boldsymbol{\beta}$).

In order to correct the downward bias, Restricted Maximum Likelihood (REML) (sometimes referred as Residual Maximum Likelihood Method) is proposed by Patterson & Thompson, (1971). The method simply eliminates the $\boldsymbol{\beta}$ parameter from the log-likelihood. As a result, the log-likelihood function is a function of variance component parameters. The likelihood function is obtained in terms of the residual terms, that are $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, as it is used in the above likelihood equation. There is a slight difference between a regular likelihood approach and REML in the way the residuals are defined. Clearly, residual term for REML, $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, does not contain the random effects regressor $\boldsymbol{Z}\widehat{\boldsymbol{b}}$ and therefore, it is not $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\widehat{\boldsymbol{b}})$. Excluding the random coefficient term from the residual definition is not erroneous since the residuals contain all sources of the random variation.

In linear regression, estimation space is orthogonal to the residuals and therefore, it can be shown that $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ and $\widehat{\boldsymbol{\beta}}$ are independent ((Diggle et al., 1994), Section 4.5). This provides us that the joint likelihood for $\boldsymbol{\beta}$ and the variance parameters, $\boldsymbol{\theta} = \{\sigma^2, \boldsymbol{\Lambda}, \boldsymbol{D}\}$, can be written as the product of the likelihoods based on $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ and $\widehat{\boldsymbol{\beta}}$ as follows:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y}) = L\big(\boldsymbol{\theta}\big|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\big)L(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \boldsymbol{\theta}) \tag{3.17}$$

thus rearranging the terms yields the likelihood function of the variance parameters, $\boldsymbol{\theta}$, given the residuals as

$$L\big(\boldsymbol{\theta}\big|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\big) = \frac{L(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y})}{L(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \boldsymbol{\theta})} \tag{3.18}$$

where we already have the numerator of the above ratio as in Equation (3.17). For the denominator, we need the ML estimate of the $\boldsymbol{\beta}$. This is very straightforward since the log-likelihood can be differentiated with respect to $\boldsymbol{\beta}$ and then equated to 0.

$$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0 \tag{3.19}$$

and rearrangement of above equation gives the fixed effects estimate as:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1}\boldsymbol{X^T\Sigma^{-1}y} \tag{3.20}$$

and the variance of $\widehat{\boldsymbol{\beta}}$ is

$$\begin{aligned} var\left(\widehat{\boldsymbol{\beta}}\right) &= \left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1}\boldsymbol{X^T\Sigma^{-1}}\,\boldsymbol{var}(\boldsymbol{y})\,\boldsymbol{\Sigma^{-1}X}\left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1} \\ var\left(\widehat{\boldsymbol{\beta}}\right) &= \left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1}\boldsymbol{X^T\Sigma^{-1}}\,\boldsymbol{\Sigma}\,\boldsymbol{\Sigma^{-1}X}\left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1} \\ var\left(\widehat{\boldsymbol{\beta}}\right) &= \left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1} \end{aligned} \tag{3.21}$$

The result assumes that $\boldsymbol{\Sigma}$ is known. However, it is almost impossible to know $\boldsymbol{\Sigma}$ and it should be estimated. When $\boldsymbol{\Sigma}$ is estimated, it causes a downward bias in $var\left(\widehat{\boldsymbol{\beta}}\right)$. An unbiased estimator for the variance of $\widehat{\boldsymbol{\beta}}$ was suggested by Liang & Zeger, (1986) by using the observed correlations between residuals which is known as the "empirical variance estimator".

$$\begin{aligned} var\left(\widehat{\boldsymbol{\beta}}\right) &= \left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1}\boldsymbol{X^T\Sigma^{-1}}\,\boldsymbol{cov}(\boldsymbol{y})\,\boldsymbol{\Sigma^{-1}X}\left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1} \\ var\left(\widehat{\boldsymbol{\beta}}\right) &= \left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1}\boldsymbol{X^T\Sigma^{-1}}\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^{\boldsymbol{T}}\boldsymbol{\Sigma^{-1}X}\left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1} \end{aligned} \tag{3.22}$$

Although empirical variance estimator reduces the bias for small samples, Long & Ervin, (2000) stated that it causes a lack of modeled covariance by reflecting the observed covariance in the data.

$\widehat{\boldsymbol{\beta}}$ has a multivariate normal distribution with mean $\boldsymbol{\beta}$ and variance $\left(\boldsymbol{X^T\Sigma^{-1}X}\right)^{-1}$. Hence the likelihood in the denominator of Equation (3.18) can be written as

$$L\left(\boldsymbol{\beta}\middle|\widehat{\boldsymbol{\beta}},\boldsymbol{\theta}\right) \propto \frac{exp\left[-\frac{1}{2}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\boldsymbol{T}}\boldsymbol{X^T\Sigma^{-1}X}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right]}{|\boldsymbol{X^T\Sigma^{-1}X}|^{1/2}} \tag{3.23}$$

Dividing the numerator by the denominator in Equation (3.18) returns the restricted likelihood equation

$$L\left(\boldsymbol{\theta}\middle|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) \propto \frac{exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^{T}\boldsymbol{\Sigma^{-1}}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})\right]}{|\boldsymbol{X^T\Sigma^{-1}X}|^{1/2}|\boldsymbol{\Sigma}|^{1/2}} \tag{3.24}$$

and the restricted log-likelihood is obtained as

$$log\left[L\left(\boldsymbol{\theta}\middle|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\right] = K-\frac{1}{2}\left[log|\Sigma|-log\left|\boldsymbol{X^T}\boldsymbol{\Sigma^{-1}}\boldsymbol{X}\right|^{-1}+(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})\right] \tag{3.25}$$

The restricted log-likelihood does not contain the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ as if they were integrated out. It is why this function can be called the marginal likelihood.

The estimators for the random effects require the likelihood function of parameters $\boldsymbol{\beta}, \boldsymbol{b}$ and $\boldsymbol{\theta}$ which we can define as follows:

$$L(\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{\theta}|\boldsymbol{y}) = L(\boldsymbol{\beta},\sigma^2,\boldsymbol{\Lambda}|\boldsymbol{\theta},\boldsymbol{b})\,L(\sigma^2,\boldsymbol{D}|\boldsymbol{b}) \tag{3.26}$$

where $(\sigma^2,\boldsymbol{\Lambda})$ and $(\sigma^2,\boldsymbol{D})$ are the variance components for $\boldsymbol{y}|\boldsymbol{b}$ and $\boldsymbol{b}$ respectively. We can obtain the following likelihood function using the multvariate normal distributions for $\boldsymbol{y}|\boldsymbol{b}$ and $\boldsymbol{b}$.

$$L(\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T(\sigma^2\boldsymbol{\Lambda})^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})-\frac{1}{2}\boldsymbol{b^T}\left(\sigma^2\overrightarrow{\boldsymbol{D}}\right)^{-1}\boldsymbol{b}\right]}{|\sigma^2\boldsymbol{\Lambda}|^{1/2}\left|\sigma^2\overrightarrow{\boldsymbol{D}}\right|^{1/2}} \tag{3.27}$$

and the log-likelihood is obtained as

$$log[L(\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{\theta}|\boldsymbol{y})] = -\frac{1}{2}\begin{bmatrix} log|\sigma^2\boldsymbol{\Lambda}| + (\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T(\sigma^2\boldsymbol{\Lambda})^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb}) \\ +log\left|\sigma^2\overrightarrow{\boldsymbol{D}}\right| + \boldsymbol{b^T}\left(\sigma^2\overrightarrow{\boldsymbol{D}}\right)^{-1}\boldsymbol{b} \end{bmatrix} + K \tag{3.28}$$

Differentiation of above log-likelihood with respect to $\boldsymbol{b}$ and then setting it equal to zero provides the $\widehat{\boldsymbol{b}}$.

$$\begin{aligned} \widehat{\boldsymbol{b}} &= \left(\boldsymbol{Z}^T(\sigma^2\boldsymbol{\Lambda})^{-1}\boldsymbol{Z}+\left(\sigma^2\overrightarrow{\boldsymbol{D}}\right)^{-1}\right)^{-1}\boldsymbol{Z}^T(\sigma^2\boldsymbol{\Lambda})^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}) \\ \widehat{\boldsymbol{b}} &= \sigma^2\overrightarrow{\boldsymbol{D}}\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}) \end{aligned} \tag{3.29}$$

and the variance of $\widehat{\boldsymbol{b}}$ can be found as follows:

$$var(\widehat{\boldsymbol{b}}) = \sigma^2\overrightarrow{\boldsymbol{D}}\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\sigma^2\overrightarrow{\boldsymbol{D}} - \sigma^2\overrightarrow{\boldsymbol{D}}\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\sigma^2\overrightarrow{\boldsymbol{D}} \quad (3.30)$$

Equation (3.17) through Equation (3.30) were adopted from Brown & Prescott (2006). They also point that there is a shrinkage in the estimator compared to what it would be if it were fixed. Again in here, $\boldsymbol{\Sigma}$ is assumed to be known and needs to be estimated when it is unknown. Estimate of $\boldsymbol{\Sigma}$ has a slight downward bias since it is sample based. Bayesian approach can be used to get rid of the bias that is introduced by the nature of ML method.

3.5.3 Estimation of the Variance Parameters

The estimation of the variance parameters are not as straightforward as the fixed and random effect coefficients because the derivative of the log-likelihood function for the variance parameters is not linear. Therefore, the solution of the derivatives require numerical methods. The literature presents many solutions to the estimation of the variance components of the mixed models including Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) utilizing iterative numerical methods or Expectation-Maximization (EM) algorithm. The EM algorithm which is utilized in this thesis for estimating the variance components is given by Laird & Ware (1982). The algorithm is described as follows.

The closed form of the ML estimates of the components of $\boldsymbol{\theta}$ can be found based on the quadratic forms in $\boldsymbol{b_i}$ and $\boldsymbol{e_i}$. Therefore we can easily obtain the following equations.

$$\hat{\sigma}^2 = \sum_{i=1}^{C} e_i^T e_i \Big/ \sum_{i=1}^{C} n_i = t_1 / \sum_{i=1}^{C} n_i \quad (3.31)$$

and

$$\widehat{D} = C^{-1}\sum_{i=1}^{C} b_i b_i^T = t_2/C \quad (3.32)$$

The t_1 and the $\frac{1}{2}k(k+1)$ components of t_2 therefore are the sufficient statistics for $\boldsymbol{\theta}$ where $k = Rank(D)$.

The estimates of the sufficient statistics, t_1 and t_2, can be calculated as follows:

$$\begin{aligned}\hat{t}_1 &= E\left[\sum_{i=1}^{C} e_i^T e_i \mid y_i, \hat{\beta}, \hat{\theta}\right] \\ &= \sum_{i=1}^{C}\left[\hat{e}_i^T \hat{e}_i + tr\left(var\{e_i|y_i, \hat{\beta}, \hat{\theta}\}\right)\right]\end{aligned} \tag{3.33}$$

and

$$\begin{aligned}\hat{t}_2 &= E\left[\sum_{i=1}^{C} b_i b_i^T \,|y_i, \hat{\beta}, \hat{\theta}\right] \\ &= \sum_{i=1}^{C}\left[b_i b_i^T + var\{b_i|y_i, \hat{\beta}, \hat{\theta}\}\right]\end{aligned} \tag{3.34}$$

where

$$\hat{e}_i = E\left[e_i|y_i, \hat{\beta}, \hat{\theta}\right] = y_i - X_i\hat{\beta} - Z_i\hat{b}_i \tag{3.35}$$

A preliminary estimate of $\boldsymbol{\theta}$ and thus $\boldsymbol{\beta}$ can be used to start the iterations between Equation (3.33) and Equation (3.34) which is defined as the "E" (Expectation) step and the iterations between Equation (3.31) and Equation (3.32) which is defined as the "M" (Maximization) step until convergence. At the time of convergence, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}$ are also obtained.

There are also various applications of generalized least squares estimation available. Bayesian alternatives to frequentist approach to the problem are also available.

Brown & Prescott, (2006) described the implicit forms of the Newton-Raphson iterative solution as well as the iterative generalized least squares estimation that is based on the full residual likelihood. They also presented posterior densities of the parameters for Bayesian framework approach. The work by Laird & Ware, (1982) is extensively used for applications especially in software packages. They offered ML and REML methods for estimating the variance components when the covariance matrix is unknown. As the solutions for both methods are not explicit for variance components they described how EM algorithm was implemented. The computations throughout this study was done on R software platform and its *nlme* package that utilizes the procedures mainly in Laird & Ware, (1982), Lindstrom & Bates, (1988) and Dempster et al., (1981). Computational details of the

EM algorithm applied to the mixed models discussed will not be covered here as they are too complex and out of interest. One important additional point that should be made here is that ML estimates create a bias in the estimates of random effects because the method does not take the loss of degrees of freedom in estimating the fixed effects. Usage of REML is required to overcome this drawback. Especially for small sample sizes REML has less bias outperforming ML.

3.6 The Reasons to Use Mixed Models

In the context of this study, the complicated structure of the data requires a powerful and comprehensive model for a detailed statistical inference. Common regression models lack handling random coefficients together with fixed effects, subjectwise analysis, time trend fitting and many other requirements. Especially the need for taking the correlation between measurements or consecutive time points into consideration is one of the most compelling part of the analyses. Factors that affect the inference on the results are based highly on the data. Some common advantages that favors the implementation of the mixed models are very well summarized and itemized in Brown & Prescott, (2006) as follows:

- Incorporating the covariance between measurements can be done with mixed models, and it improves the fitting appropriateness of the fixed effects estimates and standard errors.
- Handling repeated measures, unbalanced data and missing values is available via mixed models.
- If the data of interest has hierarchical levels with many factors such as cell type, treatment, short time series and the clusters that contain the probe sets with similar expression profiles in this study, mixed models produce more appropriate results. For example, cell type effects are allowed to vary randomly across treatment and control groups and across different time points.
- Fitting the random effects together with the fixed effects produce more unbiased results compared to when they are fit only as fixed effects. Existence of the random effects compensates the shrinkage of the fixed effects towards to the mean value. This also helps to avoid any over and under estimation of parameter estimates that occur by chance.

In addition to above, mixed models provide a great flexibility whenever additional factors or terms such as covariates or categorical factors are to be added to the model. They can

model data structure very well. Sometimes, because of the nature of the experiment, factors must be incorporated to the model as random factors which mixed models allow.

As the whole data can be analyzed and common means can be calculated, subjectwise (subject-specific in some sources) inference is available. One fascinating feature of mixed models is that they can separately estimate the fixed and the random effect slopes of an individual's change over a longitudinal period as well as its group-level mean slope.

Moreover, Baayen et al., (2008) summarized the advantages of the mixed effects modeling as follows:

Mixed-effects can handle covariates successfully even if they do not fold all the time during the experiment. The longitudinal effects differ in some sense that sometimes the effect of a factor can show up for a short period of time in the warm up period of the experiment. Biologists call this effect acute effect of a treatment in medical studies. On the contrary, sometimes the effect of a factor becomes more significant in the course of time while the acute one lessens. That is namely the chronic effect and it is more durable. Mixed-effects models allow us to distinguish between these two.

Especially clinical studies require careful investigation of the change in the response under certain experimental conditions. However, the change does not occur in the same time lag for every section of experimental period. Therefore, the classical signal-to-noise ratio approach is useless in the sense that it cannot take into account the change in time. A clear sketch of this situation is given in Figure 3.10 which consists of two panels. The one on top has unevenly spaced time points where the first time lag is 6 hours from 0 hour to 6 hour and the second time lag is 42 hours from 6 hour to 48 hours. The change in the gene expression level in the second interval is given as *h*. On the bottom panel of Figure 3.10 the sketched trend has evenly spaced time points and the change in the gene expression level in the second interval is the same as in the upper panel as *h*. Even if the change in the responses *(h)* for both graphs are the same, the time lags between consecutive measurements for both models are different. An insightful analysis has to take the continuous time effect into account that can be easily done by mixed models. A random slope parameter for time effect was used in this study to handle the change in gene expression level over unevenly spaced time points.

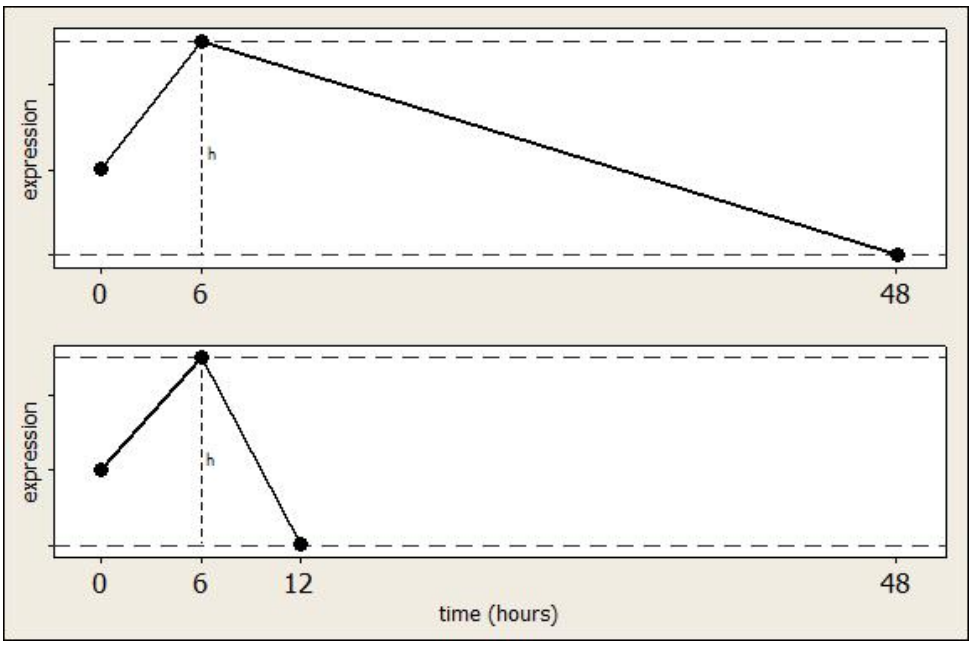


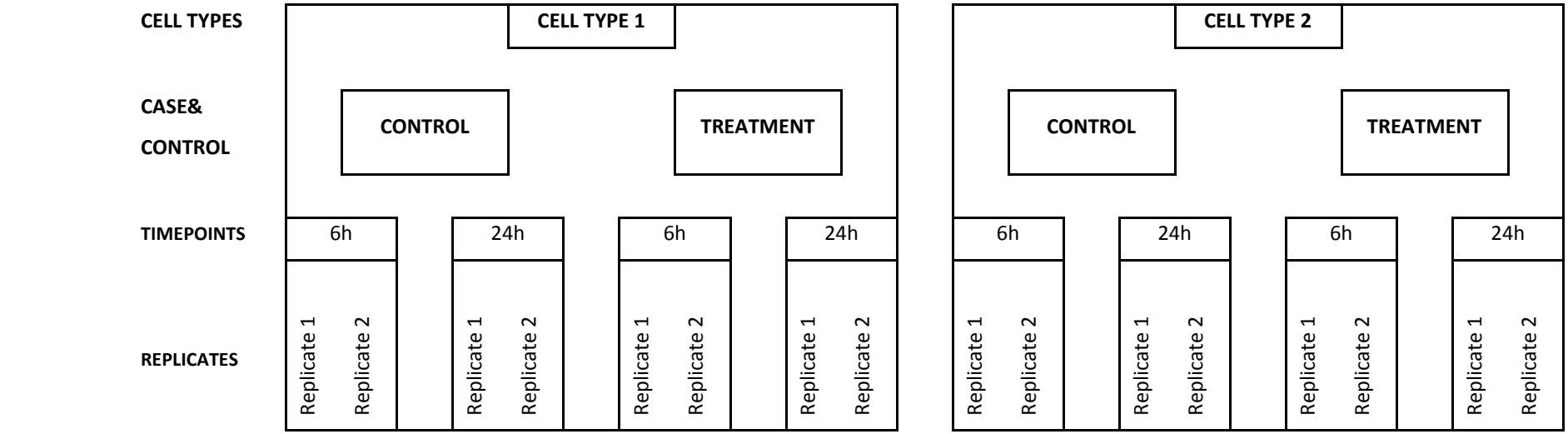
Figure 3.10 The same amount of change in the response in different time lags

Nevertheless the total change in the same interval in the main response is the same for both cases, the slope of the trend is different because of the different lags. Therefore, the time effect must be directly incorporated to the model to be able to detect the time trend. Incorporation of the time parameter to the mixed model as a continuous covariate can handle the situation as desired. On the contrary, some methods like Limma as the competing alternative of LME can only handle qualitative factors and lacks incorporating the real time effect. In the bottom line, mixed effects models are useful for modeling unevenly spaced time trends.

Mixed-effects models are able to handle many kinds of longitudinal effects straightforwardly into the statistical model and do not require prior averaging (Baayen et al., 2008). In addition, experimental conditions prior to the measurements can also be incorporated to the model and the analyses. Especially, qualitative properties of initial trials should be under statistical control (Baayen et al., 2008).

3.7 Data Structure

A representative structure of the data set that is subject to this study can be seen in Figure 3.11. The notation, e.g. $y_{1,109,1}^{1,0,6}$ will be introduced in the next section.



CLUSTERS	PROBE SETS																			
1	Probe_109	$y_{1,109}^{1,0,6}$	$y_{1,109}^{1,0,6}$	$y_{1,109}^{1,0,2}$	$y_{1,109}^{1,0,2}$	$y_{1,109}^{1,1,6}$	$y_{1,109}^{1,1,6}$	$y_{1,109}^{1,1,2}$	$y_{1,109}^{1,1,2}$	$y_{1,109}^{2,0,6}$	$y_{1,109}^{2,0,6}$	$y_{1,109}^{2,0,2}$	$y_{1,109}^{2,0,2}$	$y_{1,109}^{2,1,6}$	$y_{1,109}^{2,1,6}$	$y_{1,109}^{2,1,2}$	$y_{1,109}^{2,1,2}$
1	Probe_192	$y_{1,192}^{1,0,6}$	$y_{1,192}^{1,0,6}$	$y_{1,192}^{1,0,2}$	$y_{1,192}^{1,0,2}$	$y_{1,192}^{1,1,6}$	$y_{1,192}^{1,1,6}$	$y_{1,192}^{1,1,2}$	$y_{1,192}^{1,1,2}$	$y_{1,192}^{2,0,6}$	$y_{1,192}^{2,0,6}$	$y_{1,192}^{2,0,2}$	$y_{1,192}^{2,0,2}$	$y_{1,192}^{2,1,6}$	$y_{1,192}^{2,1,6}$	$y_{1,192}^{2,1,2}$	$y_{1,192}^{2,1,2}$
2	Probe_182	$y_{2,182}^{1,0,6}$	$y_{2,182}^{1,0,6}$	$y_{2,182}^{1,0,2}$	$y_{2,182}^{1,0,2}$	$y_{2,182}^{1,1,6}$	$y_{2,182}^{1,1,6}$	$y_{2,182}^{1,1,2}$	$y_{2,182}^{1,1,2}$	$y_{2,182}^{2,0,6}$	$y_{2,182}^{2,0,6}$	$y_{2,182}^{2,0,2}$	$y_{2,182}^{2,0,2}$	$y_{2,182}^{2,1,6}$	$y_{2,182}^{2,1,6}$	$y_{2,182}^{2,1,2}$	$y_{2,182}^{2,1,2}$
3	Probe_191	$y_{3,191}^{1,0,6}$	$y_{3,191}^{1,0,6}$	$y_{3,191}^{1,0,2}$	$y_{3,191}^{1,0,2}$	$y_{3,191}^{1,1,6}$	$y_{3,191}^{1,1,6}$	$y_{3,191}^{1,1,2}$	$y_{3,191}^{1,1,2}$	$y_{3,191}^{2,0,6}$	$y_{3,191}^{2,0,6}$	$y_{3,191}^{2,0,2}$	$y_{3,191}^{2,0,2}$	$y_{3,191}^{2,1,6}$	$y_{3,191}^{2,1,6}$	$y_{3,191}^{2,1,2}$	$y_{3,191}^{2,1,2}$
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 3.11 An example of the structure of short time series microarray data with replicates

The structure of the data represented in Figure 3.11 shows that the first factor is the cell type. Given the cell type, a measurement can be exposed to the treatment or be the control group which is the second factor namely the exposure. Given the cell type and the exposure a measurement or its replicates can be taken over unevenly spaced continuous time points that is the third factor. At the bottom of the structure there are probe sets that are grouped into clusters and also referred as subjects in this study.

3.8 The Model

The model proposed in this study to detect differentially expressed short time series gene clusters over time is as in Equation (3.36) below.

$$Y_{i,p,j}^{a,h,t_m(k)} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})I_h + \sum_{a=1}^{A-1}(\beta_{2a} + b_{2ai})I_a + (\beta_3 + b_{3i})t_m(k) + \varepsilon_{i,p,j}^{a,h,t_m(k)} \quad (3.36)$$

where

$$I_h = \begin{Bmatrix} 1\ if\ h = 1\ (if\ treatment) \\ 0\ otherwise\ (if\ control) \end{Bmatrix}$$

$$I_a = \begin{Bmatrix} 1\ if\ cell\ line\ a \\ 0\ otherwise \end{Bmatrix}$$

- $i = 1,2, \dots, C$ represents each cluster where *C* is the total number of clusters.
- $p = 1,2, \dots, N$ stands for probe sets where there are *N* probe sets.
- $j = 1,2, \dots, n_t$ stands for replicates where there are n_t replicates in the t^{th} timepoint.
- $a = 1, 2, \dots, A - 1$ stands for cell types where there are *A* cell types.
- $h = 0$ for control groups, and $h = 1$ for treatment groups.
- $m = 1,2, \dots, M$ stands for the model to be fitted to the data at consecutive couples of time points (e.g. *m*=1 for the model fitted to the data in the first time interval, *m*=2 for the model fitted to the data in the second time interval, and so on). Therefore, there are *M* time intervals.
- $k = 1,2$ stands for the order of the timepoint in a two-timepoint interval (e.g. *k*=2 indicates 6h at 1h – 6h period).

- $t_m(k)$ is the regressor variable (the timepoint value) for m^{th} model and k^{th} time (e.g. $t_1(2) = 1, or\ t_3(1) = 6$ where timepoints are {0, 1, 6, 24, 48, 168} respectively).
- $\beta_0, \beta_1, \beta_{2a}\ and\ \beta_3$ are the fixed effect coefficients that are same for all clusters.
- $b_{0i}, b_{1i}, b_{2ai}\ and\ b_{3i}$ are random effect coefficients that is specific to each of *C* clusters and *A* treatments.
- $Y_{i,p,j}^{a,h,t_m(2)} = Y_{i,p,j}^{a,h,t_{(m+1)}(1)}$, *e.g.* $Y_{1,5,2}^{2,0,t_3(2)} = Y_{1,5,2}^{2,0,t_4(1)} = Y_{1,5,2}^{2,0,24}$
- $\varepsilon_{i,p,j}^{a,h,t_m(k)} \sim N(0, \sigma^2)$ are the random noises.

The parameters in Equation (3.36) were estimated separately for each time interval for short time series as if the time series were consisted of only two time points. For example, if the short time series is consisted of 5 time points, the model is fitted for each of the 4 time intervals separately.

There are two main points on the response variable *Y*. First one is that it is in log_2 scale in all the analyses throughout the study. log_2 transformation is a common approach in preprocessing of microarray data for normalizing the expression values and also for equalizing their variances for modeling and testing purposes.

The second point is that the change in the gene expression level for a specific gene or a probe set corresponds to a fold change, since it is measured in comparison to a reference group. Biologists also prefer to talk in terms of fold change during analyses. It is also more convenient for a better understanding. Let Y_1 and Y_2 be raw gene expression values of the same gene measured on two different states. The fold change from state 1 to state 2, namely FC_{12}, can be written as $FC_{12} = \frac{Y_2}{Y_1}$. Hence, taking the log in base 2 returns the difference of the two as the fold change in $\log_2$ scale as in Equation (3.37).

$$\begin{aligned} FC'_{12} &= \log_2 Y_2 - \log_2 Y_1 \\ &= Y'_2 - Y'_1 \end{aligned} \tag{3.37}$$

This fact will be based upon during the simulations explained in the next chapter.

3.9 Replication

Likewise almost all of the statistical analyses, replication is a very important aspect also for design and analysis of microarray experiments. Yang & Speed, (2003) describes the three types of replicates that are most common in microarray experiments. The first one is the ***biological replicate*** where mRNA samples for microarray are collected from different experimental units (e.g. multiple cell lines, multiple biopsies, multiple patients, etc). The purpose of the biological replicate is to control the biological variability. The second one is the ***technical replicate*** where mRNA samples for microarray are collected from the same experimental unit but hybridized (and accordingly measured) on different microarrays. Technical replicate is done to control the technical variability within an experiment (e.g. array to array variation, reagent variation, dye incorporation, etc.). The third one is the ***within-array replicate*** where the same probes of an experimental unit are spotted and the same microarray are used for hybridization and analyses.

Tai & Speed, (2005) stated that replication is useful for detecting the change in the genes that happen in a limited time. They also suggested the biological replicate as the most preferable replication type because it makes the inference more convenient for larger populations rather than that of the experimental unit. Although they recommended at least three biological replicates per time point, the circumstances that the experiment is designed may not allow to do so at every analysis. Tai & Speed, (2005) also indicated that when there are only technical replicates, the experimenter lacks of calculating the pure error. On the other hand, the biological replicates allow calculating the variation between replicates and incorporating it to the analyses. The analyses and the simulations in this study was based on the biological replication as in the asbestos study (Nymark et al., 2007).

CHAPTER 4

APPLICATION

4.1 Simulation

A simulation study has been applied considering the nature of the data structure. The model proposed in this study is to be applicable for any short time series microarray data. Moreover, the success of the proposed model should be comparable to previously proposed ones in the literature. Therefore, the simulations did not depend on any particular statistical model. Accordingly, a non-model based simulation study including as many sources of variation as possible on the data was performed.

Nykter et al., (2006) stated that all microarray data simulation algorithms are based on a mathematical model and it is almost impossible to simulate an exact replica of a real life microarray data since the data are collected by the means of a measurement system. They also provided a very complicated data simulation algorithm having claimed to incorporate all the possible sources of variation to create a realistic data set. However they were unable to provide the explicit algorithm. The proposed algorithm is also far from being applicable. One important point that they make is that the "ground truth" for the start up as the initial data should be realistic. Then the resulting simulated data is much favorable for validation purposes. That is exactly what was done in this study as well. The asbestos data set were used as the ground truth at the first time point of the short time series.

Wang et al. (2008) also used a non-model based simulation algorithm in their study where they used a mixed-effects model for analyzing pathways. They generated different scenarios where proportion of genes with treatment effect, the proportion of up-regulated and down-regulated genes among the genes with treatment effect varied.

There are three main sources of variation in the microarray data of interest in this study. Cell type, exposure and time are the main effects. Every cell type was measured for exposed and control groups. Both exposure and control groups are measured in time course (in several timepoints so as to form a short time series). At each time point, there are n_i replicates. Figure 3.11 stands for the layout of the structure of the data. The number of replicates at each time point, the number of cell lines and the number of time points for each cell type may vary in applications. These kinds of variations do not create any problems in our model.

Clustering is a part of our study and it is applied during analyses and applied to the simulated data. Therefore, clusters were not simulated, instead they were calculated from the simulated data in complete accordance with the real life data application.

Besides the three main effects in the model, different probe sets and random noise that is highly observed in real life experimentation were also incorporated to the simulations. The outline of the simulation study with steps followed are as follows:

4.1.1 Generating the Initial Data

Short time series microarray data require simulating initial data as the first time point to start the series. In general time series data is simulated by using an autoregressive coefficient, namely the correlation is multiplied with the current state of the data and a reasonable amount of noise is added on top. In a similar fashion, having switched to the log scale the use of fold change can introduce correlation in microarray data simulations.

Fold changes due to the effects on the data was incorporated to the simulations. As explained in the previous chapter, fold change is the multiplicative amount of change in the original scaled data if there is a significant effect of the parameters. For example; a two fold change correnspponds to 100% increase or 50% decrease; three fold change means 200% increase or 66.7% decrease. It can be generalized as follows:

2-fold change

$$x \rightarrow 2x \Longrightarrow \left(\frac{2x-x}{x}\right) \times 100 = 100\% \uparrow$$

$$2x \rightarrow x \Longrightarrow \left(\frac{x-2x}{2x}\right) \times 100 = 50\% \downarrow$$

3-fold change

$$x \rightarrow 3x \Longrightarrow \left(\frac{3x-x}{x}\right) \times 100 = 200\% \uparrow$$

$$3x \rightarrow x \Longrightarrow \left(\frac{x-3x}{3x}\right) \times 100 = 66.7\% \downarrow$$

Representation of fold change effects to consecutive measurements in the original scale can be given as follows:

$$x_t = (1 + d)x_{t-1} + \varepsilon \quad (4.1)$$

d : The fold changing effect

x_t : The value of the random variable *X* at state *t*

x_{t-1} : The value of the random variable *X* at state *t-1*

ε : Random error, generally taken as $N(0, \sigma^2)$

Therefore, the amount of change in the data is *d**100%. If *d*=1, then we have (1+1)=2 fold change which is equal to 100% change or in other words the expression level is doubled. However the above representation is not valid for the data in $\log_2$ scale. Therefore, the equation was modified as follows:

$$x'_t = x'_{t-1} + d + \varepsilon \quad (4.2)$$

The asbestos data set were used as the representative initial data and it was analyzed. The histogram on Figure 4.1 sketches the distribution of an array from Asbestos dataset. Although the selected array is A549 cell line under asbestos exposure at the first hour, all other arrays would have sketched the same distribution since their quantiles were normalized.

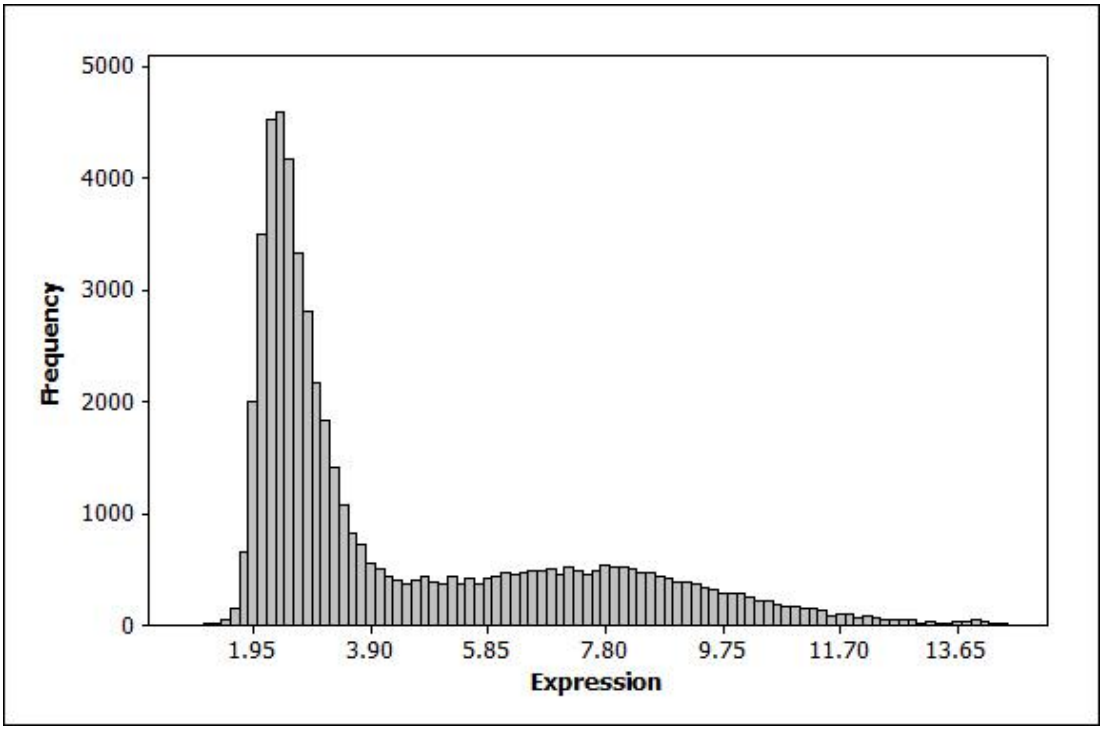


Figure 4.1 The distribution of asbestos data (Cell type:A549, asbestos exposed, observed at 1 hour)

Actually, regardless of the environmental effects, significant differential expressions are likely to be observed in a cell. Genes are distinct parts of the DNA sequence and different genes are expressed to form different organic structures. As a result, although *Gamma(3,2)* seems to represent the above distribution, it is still not a reasonable alternative. A careful review of the distribution gives a clue about the modality. It is clear to state that there are more than one locations that data centralizes. This required that the complexity of the representing distribution must be increased.

Expectation-Maximization (EM) algorithm helped to fit a mixture of normals to the data as follows.

$$f(x; p_i, \mu_i, \sigma_i) = \sum_{i=1}^{3} p_i \, g(x; \mu_i, \sigma_i) \quad for \quad x, \mu_i, \sigma_i \in \mathbb{R} \; and \; \sigma_i > 0 \tag{4.3}$$

where $g(x; \mu_i, \sigma_i)$ is a normal density. R package has "*mixtools*" library and "*normalmixEM*" function to apply the EM algorithm. The fit that is obtained by resulting mixture is much more satisfying. The maximum likelihood estimates of the parameters from above mixture distribution are as follows:

Table 4.1 MLE estimates of mixing proportions, location and scale parameters of Asbestos data measured under exposure at 1 hour

Proportions	**Means**	**Standard Deviations**
$\widehat{p_1} = 0.34$	$\widehat{\mu_1} = 2.34$	$\widehat{\sigma_1} = 0.27$
$\widehat{p_2} = 0.27$	$\widehat{\mu_2} = 2.97$	$\widehat{\sigma_2} = 0.50$
$\widehat{p_3} = 0.39$	$\widehat{\mu_3} = 7.28$	$\widehat{\sigma_3} = 2.44$

Resketching the fitted mixture of normal distributions with the parameters given in Table 4.1 yields Figure 4.2:

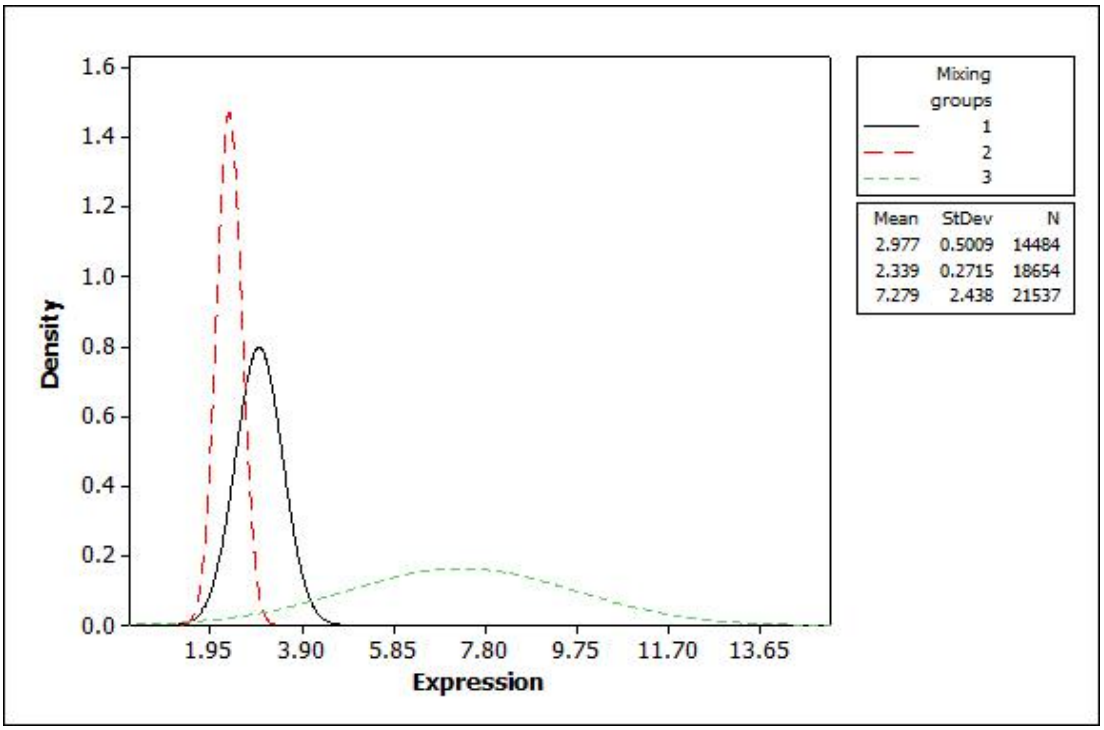


Figure 4.2 Densities of estimated mixing normal distributions by EM algorithm

4.1.2 Cell Type Effect

The first main effect cell line or "cell type" has an effect on all observations regardless of exposure, time, replicate and probe set. As a matter of fact that time series data is needed and in a time series every other measurement is affected from a previous measurement sequentially. If cell effect is applied at every time interval it would cause a cumulative effect and create a bias in the mean cell effect. Therefore, in order to apply the cell effect once and not to let it cumulate through the time steps, the fold change for cell effect was added once to the initial data. The differential expression for the cell type factor is generated as the contrast between different cell types. As a rule of thumb in linear models, a qualitative factor is represented with the contrasts generated by its levels. For example, if there are two cell types in the experiment, they are represented with only one parameter, that is the effect of the second minus the first level, which is the reference level.

The number of cell types may vary due to the concept of the experiment. Therefore, there may be a number of combinations for the contrasts that are possible to be observed in the real life data. Firstly, all possible orderings of cell significances were generated. That is 3^A possible orderings of significances where *A* is the number of cell types as stated in Equation (3.36). An example of possible orderings when there are two cell types is given in Table 4.2.

All possible combinations of cell type significances were ordered and were also ranked, namely the profile ranks to be used for sampling purposes in proceeding stages of the simulation.

Table 4.2 Possible significance orderings for two cell types

Profile Rank	Cell Type 1	Cell Type 2
1	-1	-1
2	-1	0
3	-1	1
4	0	-1
5	0	0
6	0	1
7	1	-1
8	1	0
9	1	1

-1 : Cell type has a reducing effect on the gene expression level

0 : Cell type has no effect on the gene expression level

1 : Cell type has an increasing effect on the gene expression level

Secondly, a contrast table was calculated for modeling purposes. The case given in Table 4.2 can be represented by a single parameter in the model which is the contrast between cell type 2 and cell type 1 ($e.g. \beta_{cell} = \beta_{Cell2} - \beta_{Cell1}$). Therefore, a contrast table was prepared as in Table 4.3.

Table 4.3 Contrast table for two cell types

Profile Rank	Cell Type 1	Cell Type 2	$I_{(Cell2-Cell1)}$
1	-1	-1	0
2	-1	0	1
3	-1	1	1
4	0	-1	-1
5	0	0	0
6	0	1	1
7	1	-1	-1
8	1	0	-1
9	1	1	0

In this table, $I_{(Cell2-Cell1)}$ is the indicator function. As a result:

-1 : Contrast parameter is significant, cell type 1 has larger effect

0 : Contrast parameter is non-significant

1 : Contrast parameter is significant, cell type 2 has larger effect

Therefore, the system is able to handle whether a specific gene or a probe set is significantly affected by the type of the cell as well as the type of the effect such as which cell type has larger or smaller effect.

The true cell type effects for every single probe set were randomly selected from the contrast table that were generated for every combination due to the number of cell types similar to Table 4.3. However, the proportion of the significant contrast parameters were selected around 10% where 90% of the probe sets simulated have a non-significant cell type parameter. Many real life data applications showed that on the average 10% of the genes are differentially expressed and the rest of the genes as the majority are supressed or not active.

4.1.3 Exposure Effect

Like the cell type, the exposure effect which means the effect of the treatment to observed expression levels must be applied once and before the time effect. Exposure can have an increasing effect, decreasing effect or an insignificant effect on the gene expression levels in constrast with control genes that are not exposed to the treatment.

There assumed to be one exposure and one control group in this thesis. However, if required, additional exposure and control groups can be incorporated to the study like the number of cell types are two or more. The number of possible combinations for the contrasts that are possible to be observed in the real life data is 3 as follows.

-1 : Exposure to the treatment has a reducing effect on the gene expression level

0 : Exposure to the treatment has no significant effect on the gene expression level

1 : Exposure to the treatment has an increasing effect on the gene expression level

Therefore, the system is able to handle whether a specific gene or a probe set is significantly affected by exposure to the treatment as well as the type of the effect such as increasing, decreasing or not changing. The exposure parameter is a natural contrast parameter in contrast to the control group effect.

The true exposure effects for every single probe set were randomly generated. However, the proportion of the significant exposure parameters were selected around 10% where 90% of the probe sets simulated have a non-significant exposure parameter.

4.1.4 Time Effect

Having generated the first set of measurements for first timepoint for all probe sets, the effect of the next timepoint can be considered and incorporated into the simulation. All the cell type effects were applied first and then all the exposure effects were applied and then finally the time effects were applied during the simulations. This is because of the convenience of creation of the simulated data. For example, both control and exposure groups in a cell must have the cell effect. Therefore, the cell effect was incorporated firstly. Accordingly, exposure effects and time effects were introduced to the data.

The significant changes in gene expression level from the first timepoint to the second was taken as interval effect. Basically, like the cell type and exposure effects, there are three possible effects or namely trends for this effect. The amount of gene expression may increase, decrease or not change during the time interval between consecutive timepoints (see Table 4.4). Therefore, as a first action, time trends or formally gene expression profiles need to be generated. If "T" is the number of timepoints, then, there are (T-1) intervals and $3^{(T-1)}$ possible gene expression profiles. The time parameter stands as the continuous covariate in the model and it represents the slope of the fitted line in a particular time interval. Therefore, the slope can be as follows:

-1 : Significantly negative (decreasing gene expression level by time)
0 : Non-significantly zero (no change in the gene expression level by time)
1 : Significantly positive (increasing gene expression level by time)

An example of possible orderings is given in Table 4.4 in the case of 3 timepoints and 2 time intervals.

Table 4.4 Possible orderings of interval significances with 3 time points

Timepoints:	t_0	$\xrightarrow{Interval\ 1}$	t_1	$\xrightarrow{Interval\ 2}$	t_2
		increase: +1		increase: +1	
Effects:		decrease: -1		decrease: -1	
		no change: 0		no change: 0	

According to Table 4.4, all possible time profiles and corresponding profile ranks can be tabulated as Table 4.5.

Table 4.5 All possible time (interval) significances and their profile ranks with 3 time points

Profile Rank	Interval 1	Interval 2
1	-1	-1
2	-1	0
3	-1	1
4	0	-1
5	0	0
6	0	1
7	1	-1
8	1	0
9	1	1

4.1.5 The Algorithm

All possible gene expression profiles were created beforehand and a probabilistic sampling scheme was applied in order that the simulation system creates less significant changes in the gene expression levels over the time.

The simulation algorithm in the concept of the cell type, exposure and time parameter effects as explained above is as follows:

1) Define the number of cell types to generate (A) (selected as 2 and 3).
2) Define the number of probe sets to generate (N) (selected as 500 or 1000 for ease of computation).
3) Define the number of time points to generate (T) (selected as 2, 3 and 4).

4) Define the number of replicates per time point (n_t replicates in the t^{th} timepoint. Although the number of replicates may differ and the model proposed in this study can easily handle this, for the ease of simulations the number of replicates per time point was selected constant. Another reason of selecting balanced design is that the competing alternative procedure Limma cannot handle unbalanced designs. Number of replicates were selected as 2 and 3).
5) Define the maximum number of EM iterations (selected as 100).
6) Define the convergence criteria for EM iterations (selected as the difference of consecutive estimates should be less than or equal to 10^{-5})
7) Define the number of simulation runs for MCMC estimation purposes (selected as 500 runs).
8) Define the Type I error rate (α) (selected as 0.05, 0.10, 0.20 and 0.40).
9) Define the fold change amount

   eff_C: the amount of fold change when there is significant cell type effect

   eff_E: the amount of fold change when there is significant exposure effect

   eff_T: the amount of fold change when there is significant time (interval) effect

   Throughout all the simulations fold change effects for significant changes were selected as the same such as 1.5, 2 and 3 ($eff_C = eff_E = eff_T$). Any significantly decreasing change resulted as the multiplication of the fold change with -1 (*e.g.* $-eff_T$). Any insignificant change resulted as a 0 fold change.
10) Generate the contrast table of cell type parameter(s) including profile ranks and the true significances.
11) Generate the contrast table of time parameter including profile ranks and the true significances on every interval.
12) Sample the cell type parameter(s) significances for every single probe set by sampling a profile rank from the below discrete distribution given in (3.24).

$$P(Q=q)=\begin{cases} q\dfrac{0.05}{\sum_{i=1}^{M-1} i} & for\ \ q=1,2,\dots,(M-1) \\ 0.9\ for\ q=M & \\ (2M-q)\dfrac{0.05}{\sum_{i=1}^{M-1} i} & for\ \ q=(M+1),\dots,3^A \end{cases} \tag{4.4}$$

where $M=\frac{3^A+1}{2}$ being the median profile rank standing for the non-significant profile; *A* is the number of cell types and Q is the profile rank when all possible profiles were ordered in ascending order of magnitude. Therefore, proportionally 90% of the probe sets will have non-significant cell type effect. As an example,

Table 4.6 shows all the possible profile ranks along with the cell type contrasts in a 2 cell type experiment and the selection probabilities calculated by using Equation (4.4) for each profile to be selected.

Table 4.6 Profile ranks and their selection probabilities for two cell type case

Profile Rank	**Cell Type 1**	**Cell Type 2**	$I_{(Cell2-Cell1)}$	**Selection Probability**
1	-1	-1	0	0.005
2	-1	0	1	0.010
3	-1	1	1	0.015
4	0	-1	-1	0.020
5	0	0	0	0.900
6	0	1	1	0.020
7	1	-1	-1	0.015
8	1	0	-1	0.010
9	1	1	0	0.005

13) Sample the exposure parameter significances for every single probe set by sampling from the below discrete distribution given in (4.5).

$$P(Z = z) = \begin{cases} 0.10\ for\ z = -1 \\ 0.90\ for\ z = 0 \\ 0.10\ for\ z = 1 \end{cases} \tag{4.5}$$

where *Z* is the significance indicator. Therefore, proportionally 90% of the probe sets will have non-significant exposure effect. In other words, 90% of the probe sets that were exposed to the treatment will not be affected from the treatment.

14) Sample the time parameter (interval) significances for every single probe set by sampling a profile rank from the below discrete distribution given in (4.6).

$$P(Q = q) = \begin{cases} q\dfrac{0.05}{\sum_{i=1}^{M-1} i}\ \ for\ \ q = 1,2,\dots,(M-1) \\ 0.9\ for\ q = M \\ (2M-q)\dfrac{0.05}{\sum_{i=1}^{M-1} i}\ \ for\ \ q = (M+1),\dots,3^{(T-1)} \end{cases} \tag{4.6}$$

where $M = \frac{3^{(T-1)}+1}{2}$ being the median profile rank standing for the non-significant profile; *T* is the number of timepoints and Q is the profile rank when all possible profiles were ordered in ascending order of magnitude. Therefore, proportionally 90% of the probe sets will have non-significant time effect. As an example, Table 4.7 shows all the possible profile ranks along with the time (interval) significances in

a 3 time point experiment and the selection probabilities calculated by using Equation (4.26) for each profile to be selected.

Table 4.7 Profile ranks and their selection probabilities for time (interval) significances

Profile Rank	Interval 1	Interval 2	Selection Probability
1	-1	-1	0.005
2	-1	0	0.010
3	-1	1	0.015
4	0	-1	0.020
5	0	0	0.900
6	0	1	0.020
7	1	-1	0.015
8	1	0	0.010
9	1	1	0.005

Table 4.7 indicates that proportionally 2% of the probe sets will be assigned profile rank 4 and those will not have a significant time effect in the first interval where the same probe sets will be differentially down regulated in the second time interval.

Through the steps 12, 13 and 14, the true significances for cell type, exposure and time parameters are generated. The data generation part starts on step 15.

15) Generate the initial set of data from the distribution given by Equation (4.3) with parameters given on Table 4.1.
16) Repeat step 15 as many replications as required.
17) Generate data for all time points for the *1*st probe set (*p=1*) incorporating significance effects according to the generated true significances whether the probe set is differentially expressed in the given interval:

$$y_{i,p,j}^{a,h,t_m(k)} = y_{i,p,j}^{a,h,t_m(k-1)} + eff_C(I_{a,p}) + eff_E(I_{h,p}) + eff_T(I_{m,p}) + \varepsilon \qquad (3.7)$$

where $y_{i,p,j}^{a,h,t_m(0)}$ is the initial response created in step 15 when *k=1* (where $k \in \{1,2\}$). $\varepsilon \sim N(0,\sigma^2)$ and σ^2=0.1. $I_{a,p}$, $I_{h,p}$ and $I_{m,p}$ are indicator functions as follows:

$$I_{a,p} = \begin{Bmatrix} -1 \; if \; I_{(Cell_a - Cell_1)} = -1 \\ 0 \; if \; I_{(Cell_a - Cell_1)} = 0 \\ 1 \; if \; I_{(Cell_a - Cell_1)} = 1 \end{Bmatrix} for \; cell \; type = a, probe \; set = p$$

$$I_{h,p} = \begin{Bmatrix} -1\ if\ I_{(Exp-Cont)} = -1 \\ 0\ if\ I_{(Exp-Cont)} = 0 \\ 1\ if\ I_{(Exp-Cont)} = 1 \end{Bmatrix} for\ exposure = h, probe\ set = p$$

$$I_{m,p} = \begin{Bmatrix} -1\ if\ I_{Time} = -1 \\ 0\ if\ I_{Time} = 0 \\ 1\ if\ I_{Time} = 1 \end{Bmatrix} where\ interval = m, probe\ set = p$$

Repeat this step until data is generated for all time points $k = 1,2, \dots, K$ with all replicates $j = 1,2, \dots, n_t$.

18) Repeat step 17 for all other probe sets $p = 2, \dots, N$.

4.1.6 Exemplary Simulation Study

4.1.6.1 Example 1.

A simulation study was performed with the parameters given below:

- N = 500 (Total number of probe sets)
- A = 2 (Number of different cell types)
- n = 2 (number of replicates at each time point)
- $t \in \{1, 6, 24, 48\}$ are the time points
- Fold change is 2
- The standard deviation of replicates per time point is 0.5

Probe 459 was picked up randomly, true significance table is given in Table 4.8 and the simulated expression values for this probe set is given in Figure 4.3.

Table 4.8 True significance profile for probe set 459

Probes	$\boldsymbol{I_{(Cell2-Cell1)}}$	**Exposure**	**Interval 1**	**Interval 2**	**Interval 3**
Probe_459	0	1	0	0	0

According to Table 4.8, there is no significant difference between cell type 2 and cell type 1 and there is no significant change by time over the intervals. Exposure has a significant increasing effect on the expression levels. Therefore, cell type and time parameter is not truly significant but exposure parameter is.

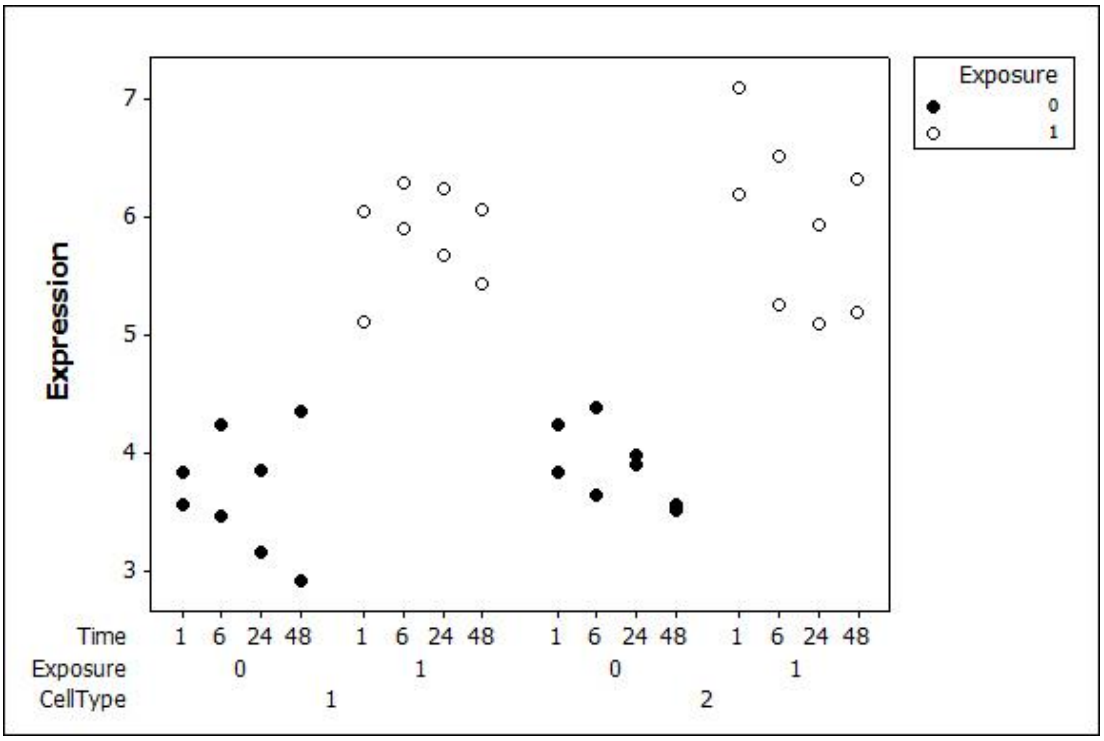


Figure 4.3 Simulated expression values for probe set 459

According to Figure 4.3 which shows the simulated data for probe set 459, there are two replicates per time point. It is clear that the mean expression level does not change across the time points on all groups (e.g. cell type 1 and 2, control and exposure groups). On the other hand, the exposure groups (empty circles) have a significant increasing effect on the gene expression levels. Exposure group has an expression level average of approximately 6 whereas control group has an average expression level around 4. There is a 2-fold change due to exposure effect.

4.1.6.2 Example 2.

A simulation study was performed with the parameters given below:

- N = 500 (Total number of probe sets)
- A = 2 (Number of different cell types)
- n = 3 (number of replicates at each time point)
- $t \in \{1, 6, 24\}$ are the time points
- Fold change is 2
- The standard deviation of replicates per time point is 0.1

Probe 151 was picked up randomly, true significance table is given in Table 4.9 and the simulated expression values for this probe set is given in Figure 4.4.

Table 4.9 True significance profile for probe set 151

Probes	$I_{(Cell2-Cell1)}$	**Exposure**	**Interval 1**	**Interval 2**
Probe_151	-1	0	1	-1

According to Table 4.9, the difference between cell type 2 and cell type 1 is significant and change by time over the intervals are also significant (e.g. significant increase in the gene expression level at first interval and significant decrease in the gene expression level at second interval). Exposure does not have a significant effect on the expression levels. Therefore, simulated significance profile indicates that cell type and time parameters are significant but exposure parameter is not.

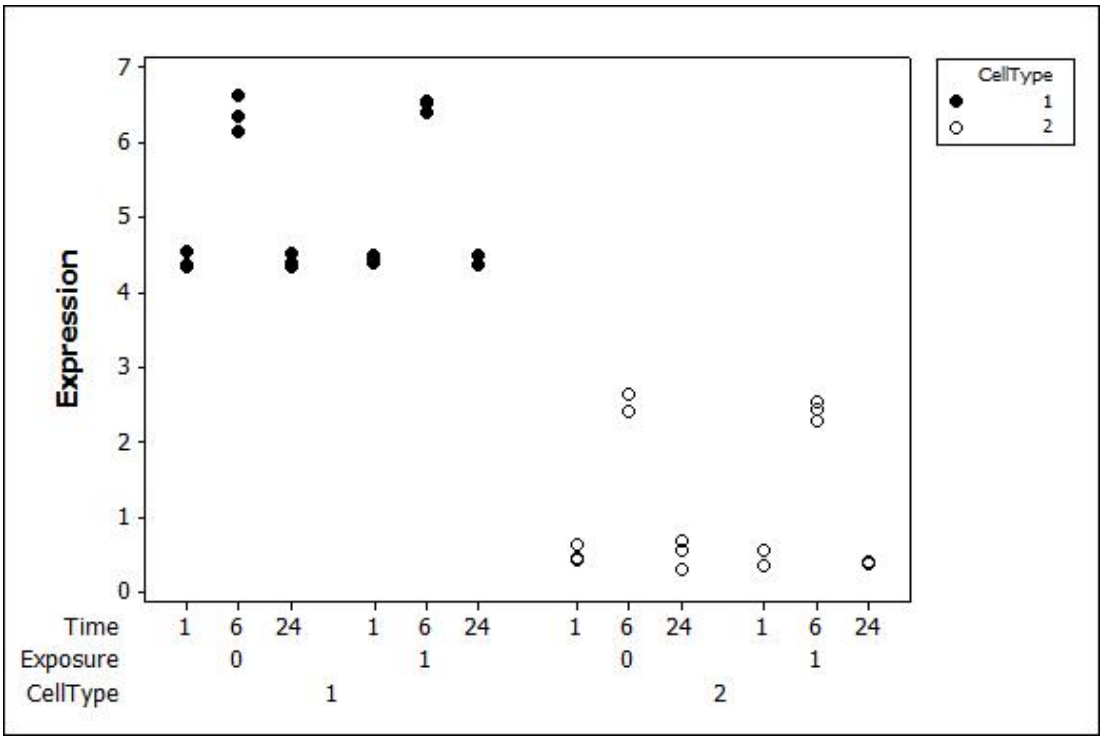


Figure 4.4 Simulated expression values for probe set 149

According to Figure 4.4 which shows the simulated data for probe set 149, there are three replicates per time point. Even though the data points on the graph were jittered by 0.025 over the y axis, there are still some overlapping points because the standard deviation for replicates was selected as 0.1. It is clear that the mean expression level increases during the

first time interval and decreases during the second time interval. On the other hand, it is clear that cell type 2 (empty circles) have significantly lower gene expression levels. Exposure on the other hand does not seem to have a significant effect on gene expression levels.

4.2 Essentials of the Simulation Study

Applications through the proposed methodology was realized in two stages as simulation study and real life data modeling. Simulations was performed for comparison and performance evaluation purposes. Real life data fitting was applied to both representative set of 520 probe sets from asbestos data which Nymark et al., (2007) used and to the full asbestos data which has 54,675 probe sets. The data firstly filtered by using *qvalue* and *kOverA* functions. K-means clustering was applied to the filtered data. Every single k-means cluster was then applied hierarchical clustering and similar gene expression profiles on each k-means cluster were detected. Finally, different hierarchical clusters on different k-means clusters formed the groups which are modeling units (subjects but not the experimental units). Proposed Linear Mixed Effects (LME) model was fit on every time interval independently for a short time series microarray data. Likewise, the competing alternative Limma was also fitted the same way in order to have a fair comparison. The simulation study was performed on TUBITAK ULAKBIM GRID Computer in Ankara where it took almost two weeks to finalize all the runs with massive data. Simulations were repeated 250 times for all combinations of below parameter settings:

- Maximum number of iterations for the LME optimization algorithm was 100 (default setting on nlme package is 50).
- Maximum number of iterations for the nlm optimization step inside the LME optimization was 100 (default is 50).
- Number of iterations for the EM algorithm used to refine the initial estimates of the random effects variance-covariance matrix was 1000 (Default is 25. Purposefully selected as very larger than the default number because the quality of initial estimates affect the success of convergence).
- The tolerance value to decide convergence for iterations for both EM algorithm and LME optimization is 10^{-5}.
- The numbers of probe sets in simulated datasets were 500 and 1000. On the real life asbestos data there are 54675 probe sets.

- The numbers of cell types in simulated datasets were 2 and 3.
- The numbers of time points were {1, 6, 24}, {1, 6, 24, 48} and {1, 6, 24, 48, 168} respectively. Therefore, the number of time intervals to fit the proposed model was 2, 3 and 4 respectively.
- The number of replicates per time point was selected as 2 and 3 respectively.
- The fold change value that was used for generating simulated data was 1.5, 2 and 3 respectively.
- The p-value cut-off points for significance testing were 0.05, 0.1, 0.2, 0.3 and 0.4.

Considering all the possible number of probe sets (2), number of cell types (2), number of time points (3), number of replicates (2) and number of fold change settings (3) for the simulations, all the possible combinations of these settings were 72. Therefore, simulations were run on 72 different settings. On every single simulation, performance measures for both LME and Limma methods were calculated based on the same simulated data. Moreover, the number of subject-wise test results were reported from both methods that match or do not match in terms of significance such as significant and non-significant. The performance measures on the simulations were calculated upon the values from Table 4.10 as follows:

Table 4.10 Ground Truth vs. Model Results

		Ground Truth		
		Positive	**Negative**	TOTAL
Model result	**Positive**	True Positive (TP)	False Positive (FP)	P'
	Negative	False Negative (FN)	True Negative (TN)	N'
	TOTAL	P	N	

True Positive Rate (TPR): This measure is also known as the “sensitivity” and is equivalent with hit rate and recall. It is the proportion of positive test results when the ground truth are positive. It corresponds to the power of the test in hypothesis testing. It is better when it’s close to 1, and worse when close to 0.

$$TPR = TP / P = TP / (TP + FN)$$

False Positive Rate (FPR): This measure is equivalent with fall-out. It is the proportion of positive test results when the ground truth are negative. It corresponds to the Type I error in hypothesis testing. It is expected to be less than or equal to the significance level.

$$FPR = FP / N = FP / (FP + TN)$$

Accuracy (ACC): Accuracy is the proportion of correctly classified test results. It is the sum of true positives and true negatives. It is better when it’s close to 1, and worse when close to 0.

$$ACC = (TP + TN) / (P + N)$$

Specificity (SPC): This measure is also known as the “True Negative Rate”. It measures the ability of the test to result as negative when the ground truth is negative. It is better when it’s close to 1, and worse when close to 0.

$$SPC = TN / N = TN / (FP + TN) = 1 − FPR$$

Positive Predictive Value (PPV): This measure is equivalent with precision. It is the proportion of correct positive test results among all positive test results. It gets closer to 1 as the number of false positives lessen. Therefore, it is better when it’s close to 1, and worse when close to 0. It depends on the number of positives in the ground truth.

$$PPV = TP / (TP + FP) = 1 – FDR$$

Negative Predictive Value (NPV): It is the proportion of correct negative test results among all negative test results. It gets closer to 1 as the number of false negatives lessen. Therefore, it is better when it’s close to 1, and worse when close to 0. It depends on the number of negatives in the ground truth.

$$NPV = TN / (TN + FN)$$

False Discovery Rate (FDR): It is controls the Type I error rate in a multiple hypothesis testing environment. It is the proportion of false positives among all positive test results.

FDR = FP / (TP + FP)

F_1 score: F_1 score is a measure of a test's accuracy that considers both precision and recall. It is better when it's close to 1, and worse when close to 0.

$$F_1 = 2\frac{precision * recall}{precision + recall} = 2\frac{PPV * TPR}{PPV + TPR}$$

4.3 Implementation of clustering methods

In this study, a two step clustering procedure was proposed and applied to both simulated data and to asbestos data. All probe sets were first clustered by k-means algorithm. Next, each of the k-means clusters were clustered by hierarchical clustering. Therefore, two independent clustering schemes were obtained successively. The final clusters were obtained such that they contain the probe sets that have the same k-means and the hierarchical cluster number. Resulting clusters have similar gene expression levels and patterns over time. The procedure was formerly used by Chen et al. (2005). They concluded that applying divisive hierarchical clustering to the k-means performs very well for similar demands. In a noteworthy study by Möller-Levet et al. (2005) introduced a clustering procedure which focused on clustering unevenly spaced time series gene expression data. They defined a distance measure for short time-series, and developed a fuzzy short time-series algorithm by utilizing the standard fuzzy c-means algorithm. The algorithm, however, computationally very complicated, intensive and hard to understand.

The sensitivity of the two-step clustering algorithm used in this study can be adjusted by changing the cutting level of the tree produced in hierarchical clusters in the second step. The tree in Figure 4.5 can be cut at height 0.8 or 0.4 according to define the final number of clusters. The height represents the distance for that special figure. The higher the height is the less the number of final clusters. However, it is often not very easy to decide at what point to cut the tree and create final clusters. The *cutree* function of R *stats* package by Development Core Team (2010) uses y-axis as the distance instead of similarity. The *cutree* function can calculate the heights of the dendrogram but the applicator has to decide at what point to cut the tree. To select an optimal point to cut the tree, the heights were sorted in ascending order of magnitude and the 5*th* percentile point was selected as the

cutting point. The lower the percentile point, the less distance between probe sets in the cluster and the similar the gene expression profiles in the final clusters. However, lowering the cutoff point increases the number of clusters to analyze and affect the modeling part by changing the number of subjects. To decide which quantile to select as the cutoff point for the hierarchical clustering tree is a trade off between the number of clusters and the similarity of gene expression profiles inside the final clusters. An illustration of this is given in Figure 4.5 which is based on the representative set of 520 probe sets from asbestos data. If the tree was cut at 0.4, it would return 6 clusters. On the other hand, if the tree was cut at 0.8, it would return 4 clusters.

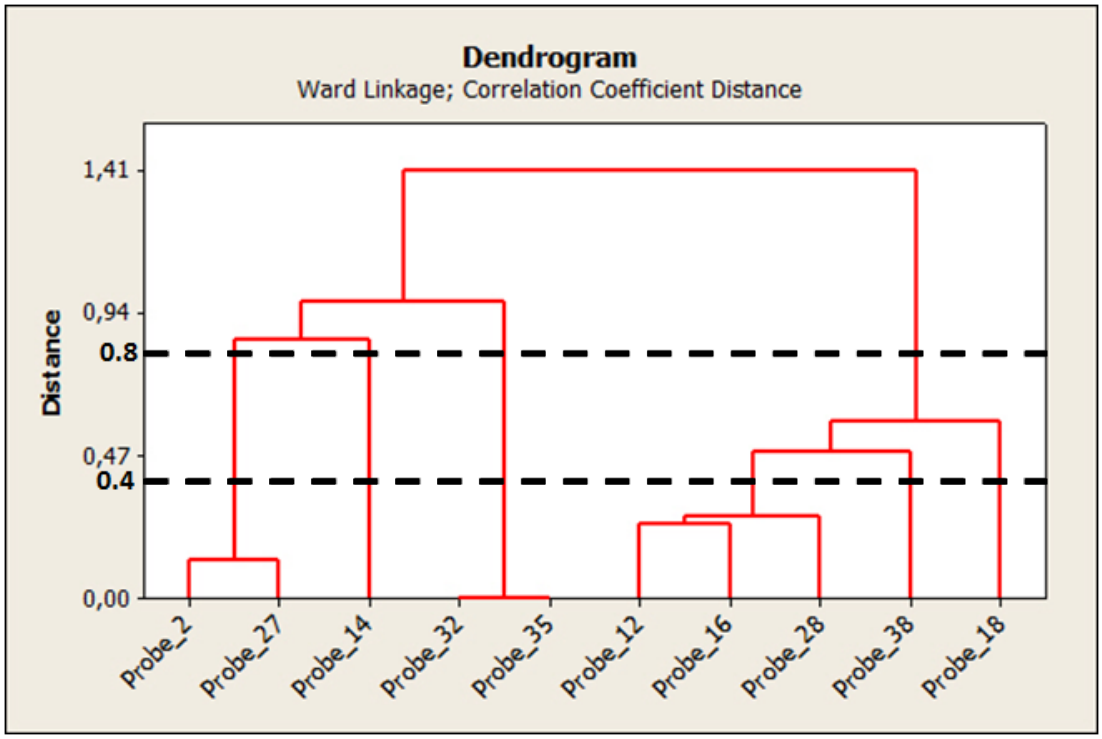


Figure 4.5 Sample dendrogram illustrating cutoff points

Grouping due to gene expression level over time was done with K-Means Clustering method (Macqueen, 1967). K-means algorithm can be utilized in R by *kmeans* command in *stats* package. We have selected the original method of founder, ***MacQueen's algorithm***, whereas Hartigan & Wong (1979), Lloyd (1982) and Forgy (1965) methods can also be applied. Each algorithm can be tested against each other in order to compare the resulting clusters, however it is not in the scope of this thesis. Clustering the data in regards with the means is one of the most important and hard-to-solve problems of statistics. Because of

the fact that distributional assumptions are very difficult to make, non-parametric and empirical methods can be used for such purposes.

Grouping due to gene expression profiles over time was done with Hierarchical Clustering method. In this method, the similarity measure was selected as ***Pearson's correlation coefficient*** and the linkage method to define the distance between two clusters was selected as ***Ward's distance***. A sample of 4 probe sets is given in Figure 4.6. A clustered version of Figure 4.6 is given in Figure 4.7. Solid lines stand for the measurements from hierarchical cluster 1 and dashed lines for hierarchical cluster 2. Both clusters have 2 probe sets.

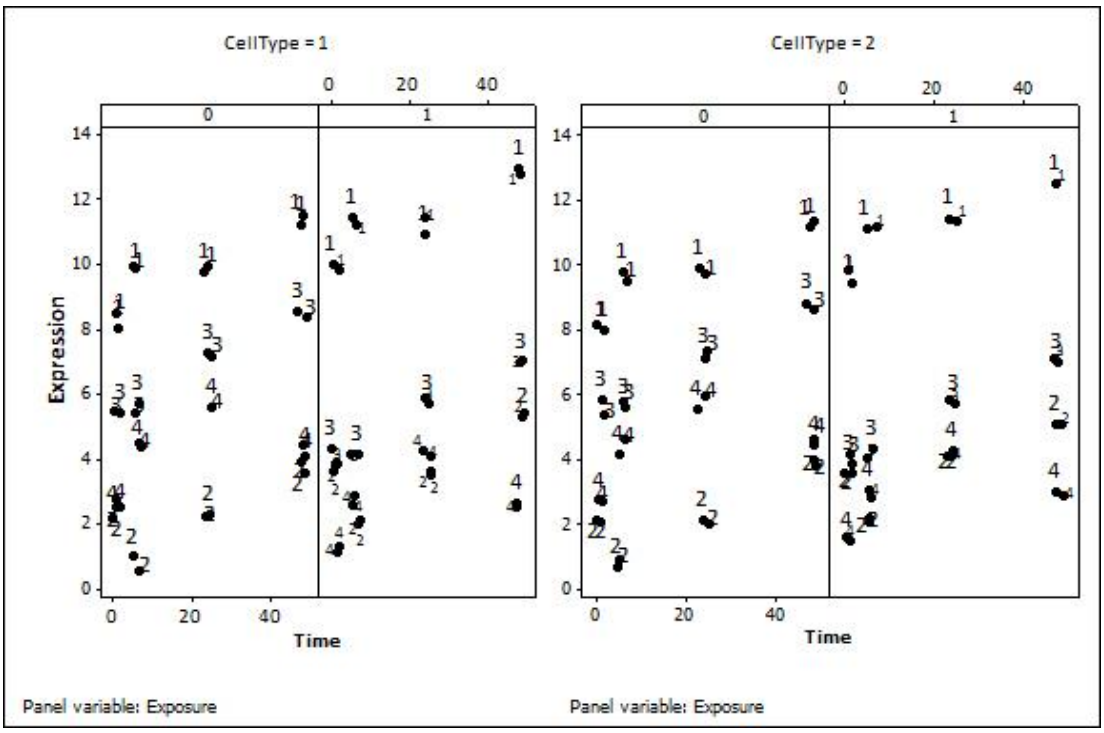


Figure 4.6 Sample probe sets from Asbestos dataset

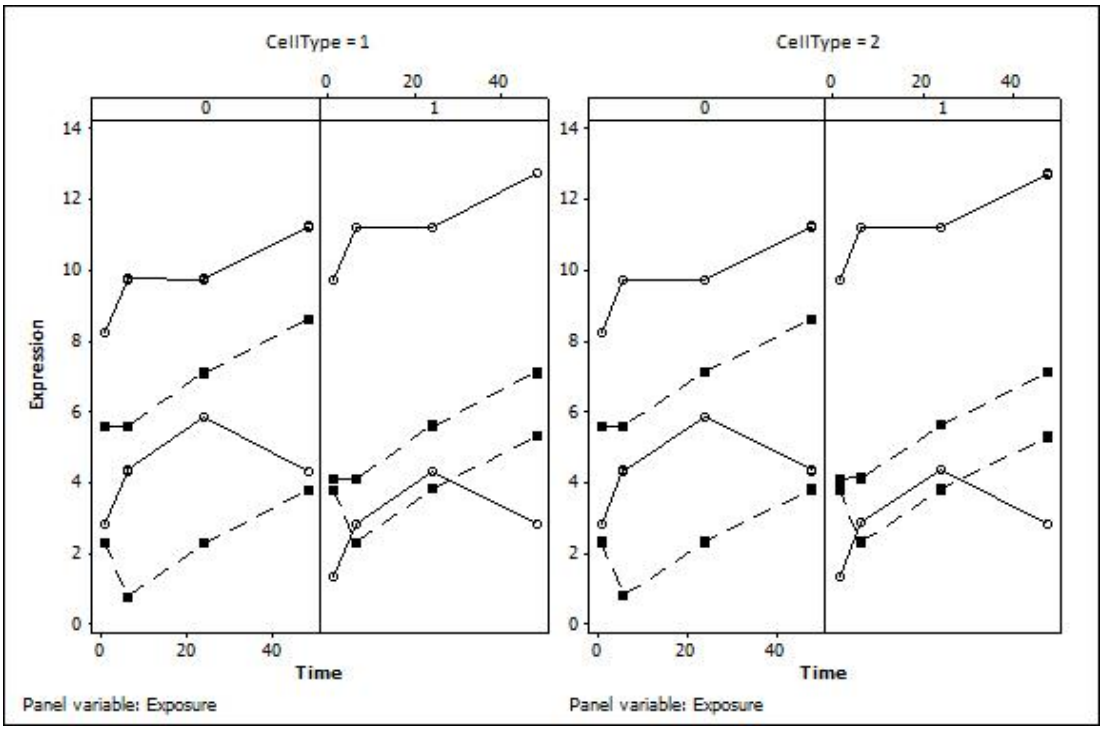


Figure 4.7 Sample hierarchical clusters from Asbestos dataset (with probe sets lined)

4.4 Essentials of the Asbestos Data

The asbestos full dataset has 54,675 probe sets measured at time points as in Figure 4.8. On A549 cell type, there are 6 time points and at only 48h there is a second replicate for the exposure group. On Beas2B cell type, there are 5 time points and at only 24h there is a second replicate for the exposure group. There are no replicates measured for the control group. The resulting data matrix is of 54675x22 dimension.

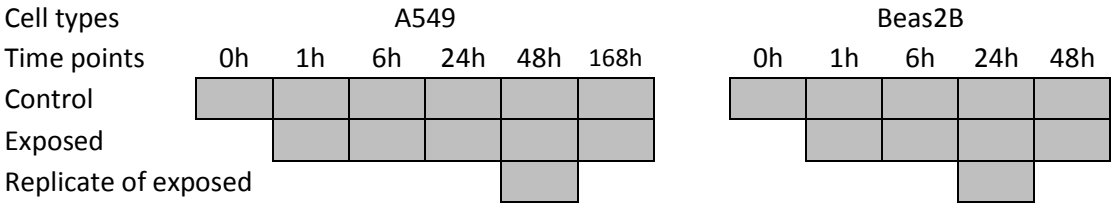


Figure 4.8 Structure of Asbestos Dataset

The design matrix that was used after filtering and clustering was of dimension 19771x22. After applying the two-stage clustering algorithm to the asbestos data 19771 probe sets were represented by 18771 final clusters where probe sets with similar expression profiles are grouped in the same cluster. Among all clusters, 17903 clusters contained only one single probe set, and accordingly, there were 779 clusters with 2 probe sets, 64 clusters

with 3 probe sets, 15 clusters with 4 probe sets, 4 clusters with 5 probe sets, 5 clusters with 6 probe sets and finally only 1 cluster contained 8 probe sets. As the clustering methods used in this study are iterative methods and this results differing number of final clusters. Nevertheless, the number of final clusters does not change remarkably from time to time.

Some probe sets from k-means clusters 1 and 2 on cell type 1, control group at 0*th* timepoint are given in Table 4.11.

Table 4.11 A part of LME design matrix

Probe Set	Expression	Cell Type	Exposure	Time	Clusters	KM Clusters	HC Clusters
1553764_a_at	9.45053825	1	0	0	1	1	1
1553979_at	9.027150307	1	0	0	2	1	2
1554241_at	10.79145014	1	0	0	3	1	3
1555758_a_at	11.07361497	1	0	0	4	1	4
209714_s_at	10.72810803	1	0	0	4	1	4
1555832_s_at	9.380775524	1	0	0	5	1	5
1053_at	8.297427203	1	0	0	501	2	1
203696_s_at	8.698420986	1	0	0	501	2	1
1552257_a_at	8.305801219	1	0	0	502	2	2
1552287_s_at	7.264148058	1	0	0	503	2	3
1552347_at	6.759981143	1	0	0	504	2	4

For a visual representation of clustering on real data, a representative set of 520 probe sets of the asbestos dataset which was also used in Nymark et al. (2007) was used. Four probe sets that are 209202_s_at, 218609_s_at, 230327_at and 236296_x_at were randomly selected and their expression values measured at 0h, 1h, 6h, 24h, 48h and 168h were sketched in Figure 4.9. Cell type and exposure were not indicated on the graph for the ease of understanding.

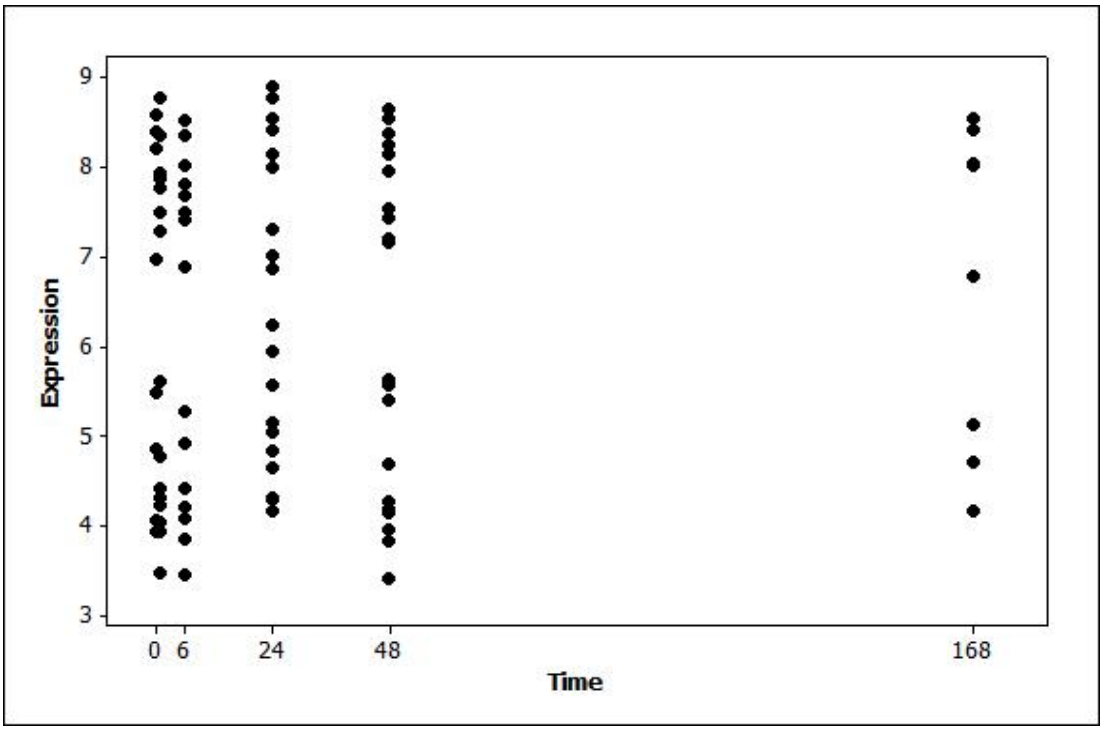


Figure 4.9 Expression values from randomly selected four probe sets

The k-means clustering of the selected probe sets is given in Figure 4.10. The k-means clusters split the four probe sets into two groups according to their gene expression levels. The first k-means cluster is composed of observations around an expression value of 8 and the second cluster is composed of observations around an expression value of 4.5. K-means clustering clustered the data by only using expression levels. The second stage of the clustering is to apply hierarchical clustering to every single k-means cluster. The resulting clusters are shown in Figure 4.11.

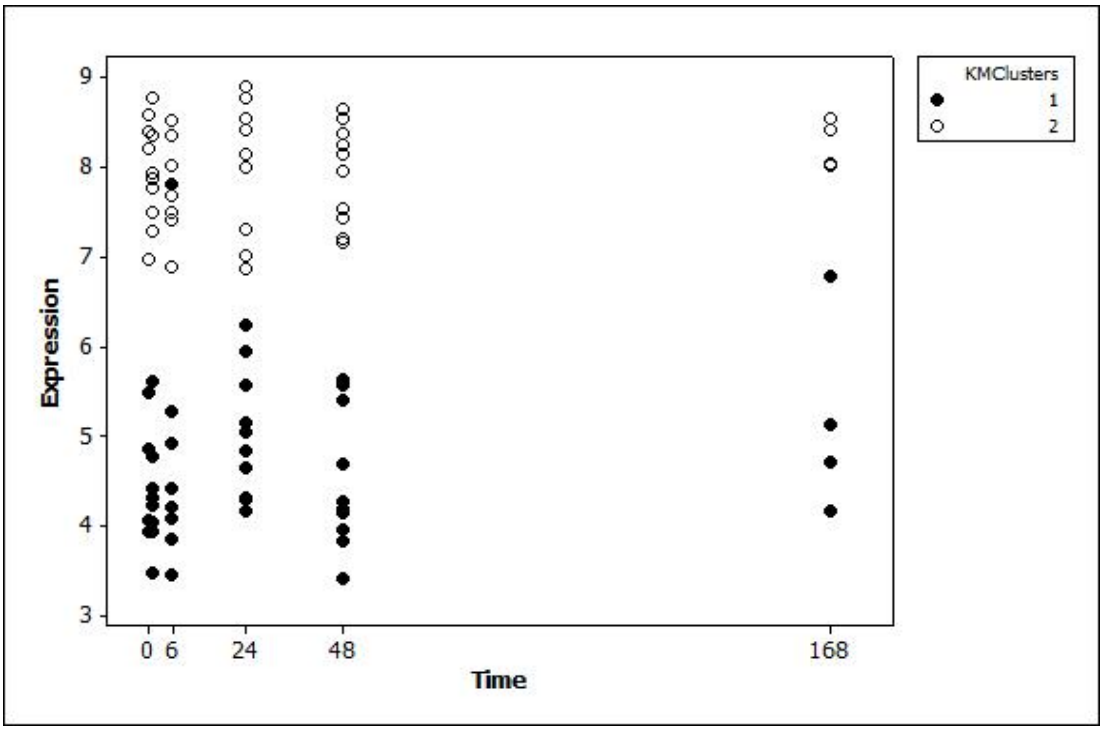


Figure 4.10 K-means clustering results of randomly selected four probe sets

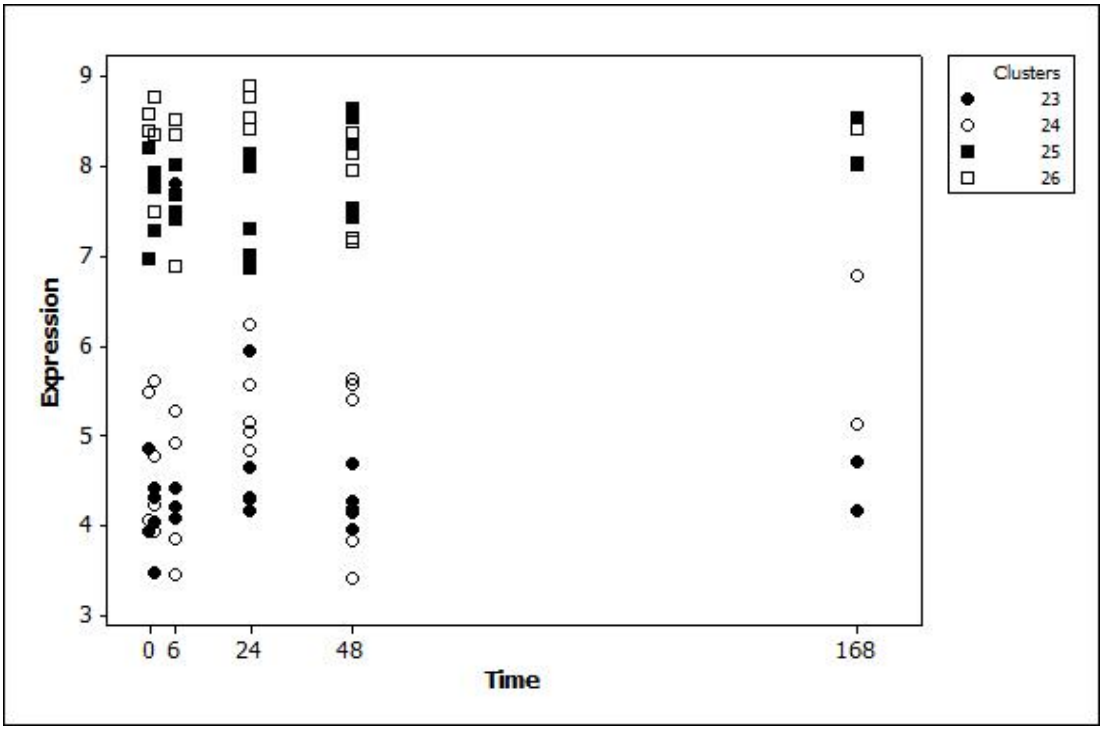


Figure 4.11 Hierarchical clustering within k-means clusters of randomly selected four probe sets

For a better understanding of cluster profiles over time, the mean expression values at each time point were connected and are given in Figure 4.12.

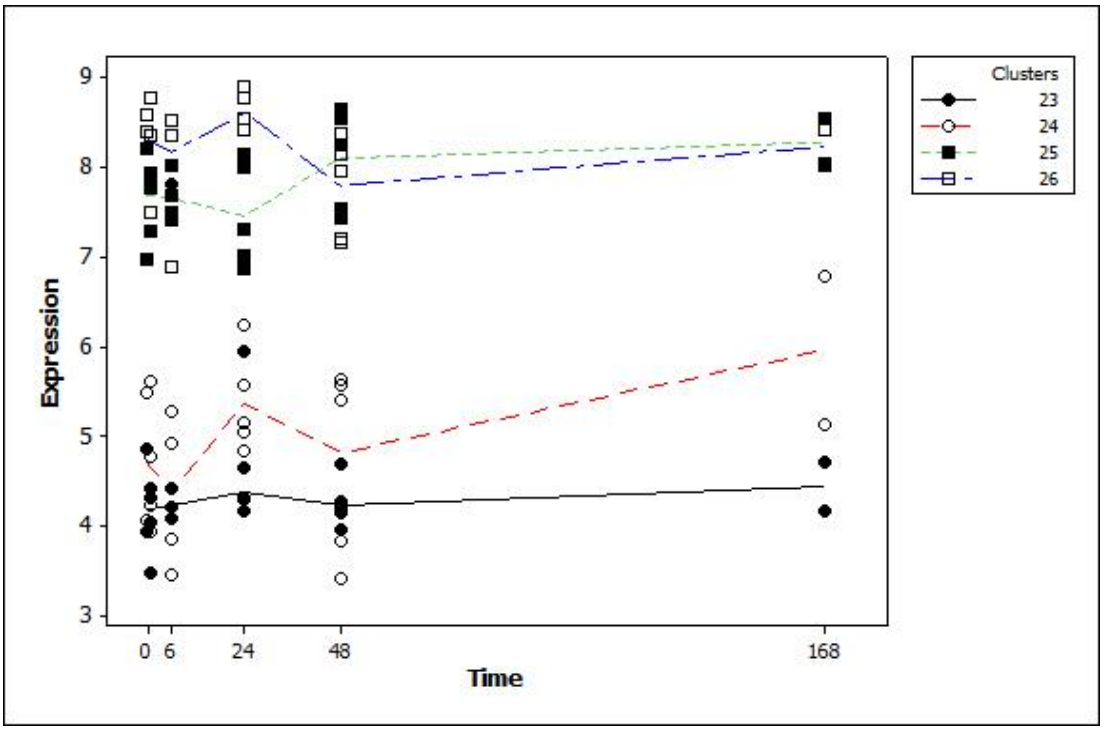


Figure 4.12 Representation of gene expression profiles of four randomly selected probe sets

LME results indicate that exposure parameter is significant at 10% type-I error level and time parameter is significant at 20% Type-I error level for cluster 23 (e.g. probe set 209202_s_at) over the second time interval. The p-values for the exposure and time parameters are 0.073 and 0.126 respectively. The measured and the fitted data for cell types 1 and 2 are given in Figure 4.13 and Figure 4.14.

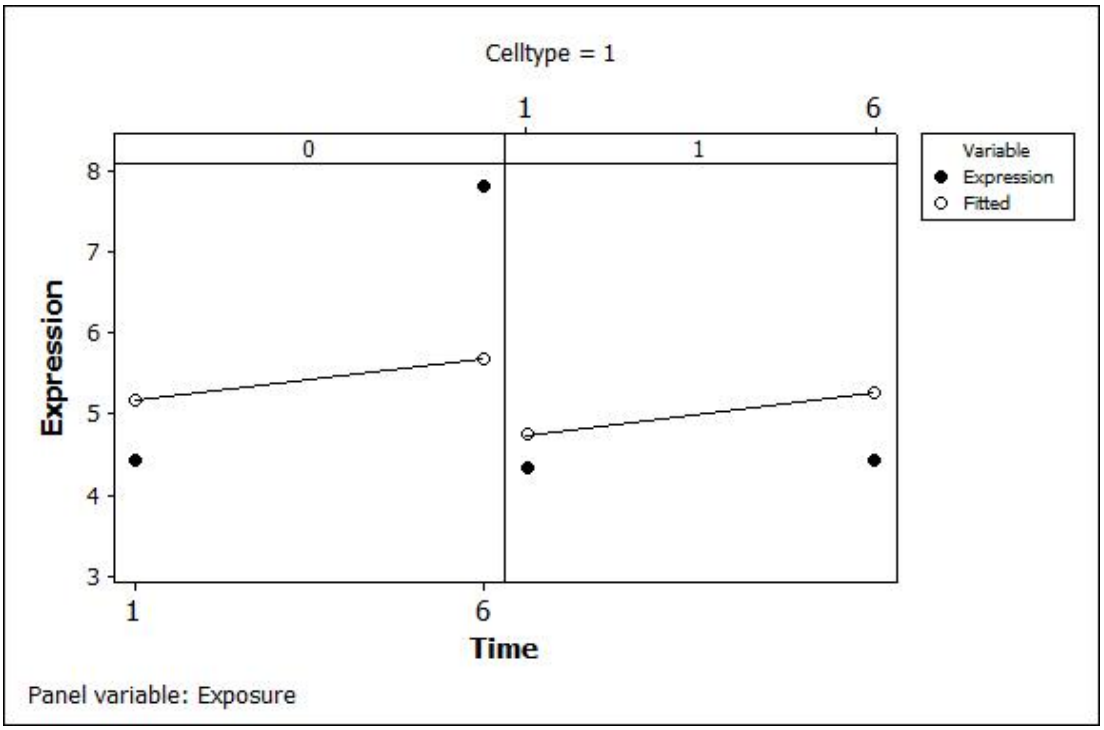


Figure 4.13 Measured and fitted data for cluster 23 with significant exposure and time parameters (cell type I)

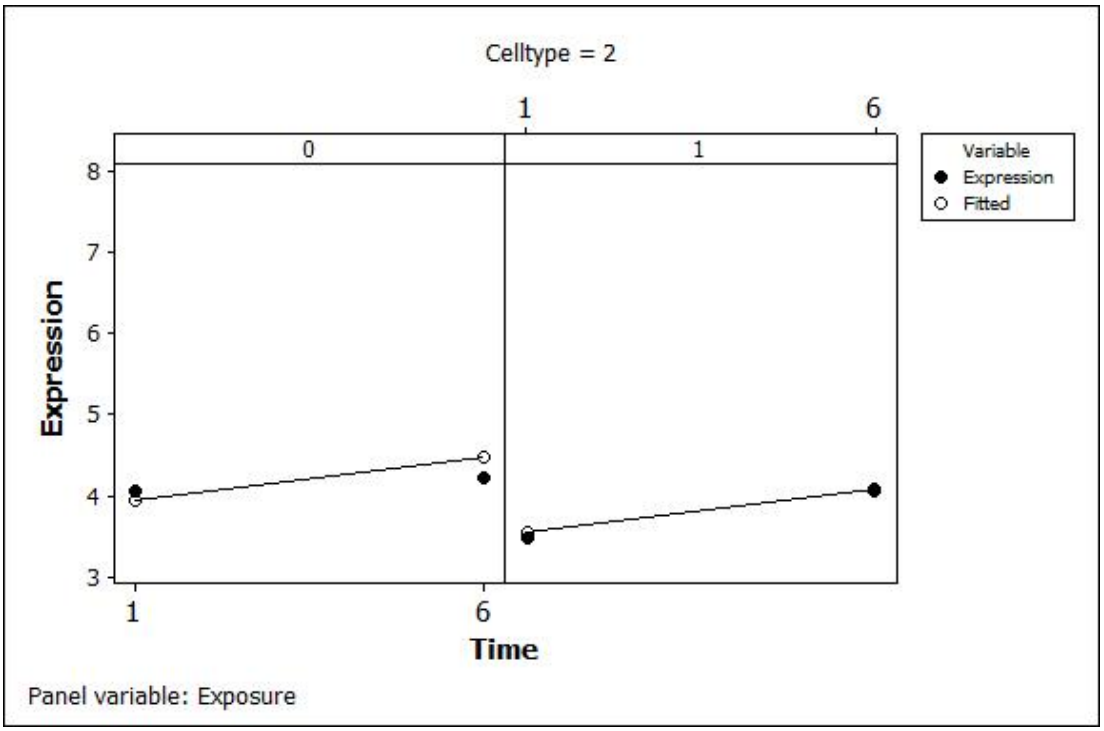


Figure 4.14 Measured and fitted data for cluster 23 with significant exposure and time parameters (cell type II)

The normality assumption in the model given in Section 3.8 was also tested for the majority of the genes accross ararys, control and exposure groups. Anderson-Darling (AD) normality test was used to test for the normality. The results indicated that the normality assumption is valid. There are two examples of normality tests for randomly selected genes, ACADVL and HNRNPM, in Figure 4.15 and Figure 4.16 respectively.

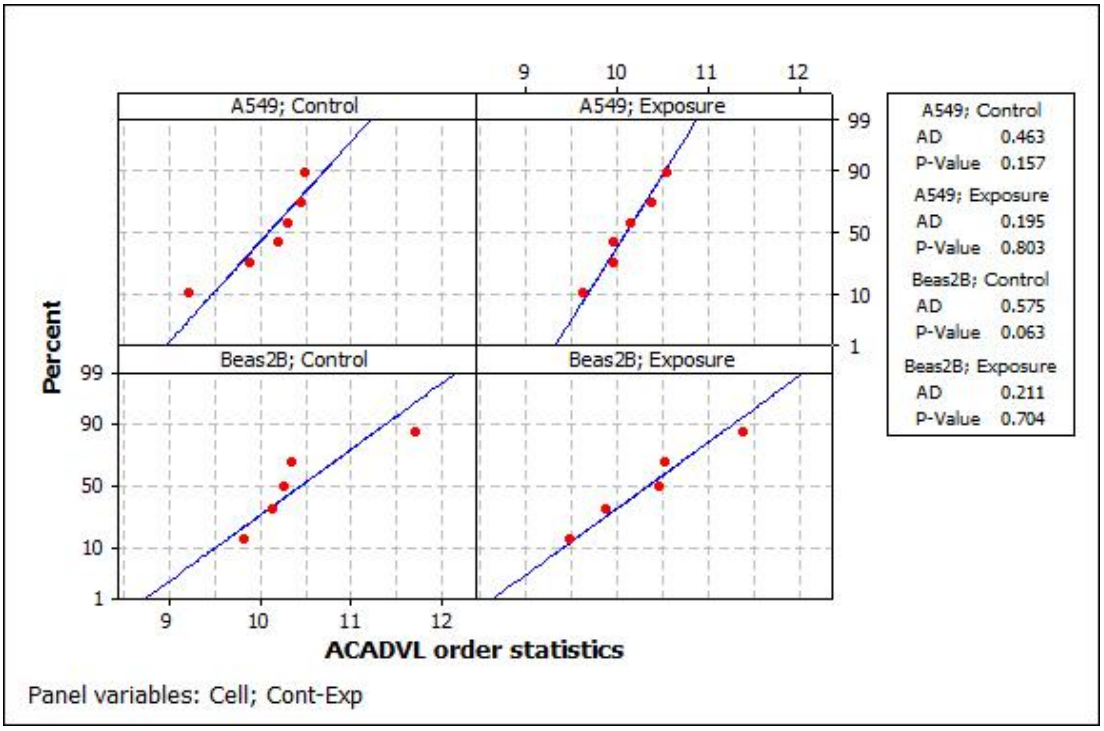


Figure 4.15 Probability plot and normality test for ACADVL gene

AD test, failed to reject the null hypothesis that the underlying distributions are normal for A549 control group, A549 exposure group, Beas2B control group and Beas2B exposure group. The p-values of AD test are 0.157, 0.803, 0.063 and 0.704 respectively. Likewise, similar results were observed also for HNRNPM gene where p-values of Ad test were 0.144, 0.671, 0.073 and 0.318 respectively.

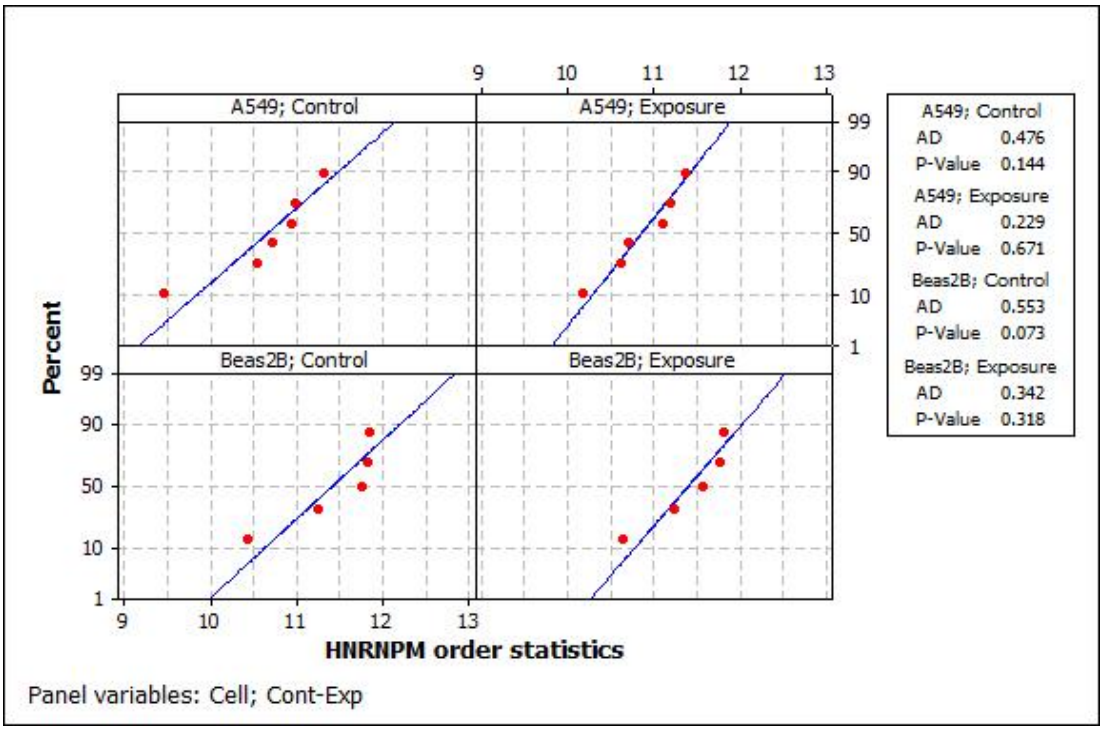


Figure 4.16 Probability plot and normality test for HNRNPM gene

CHAPTER 5

FINDINGS AND RESULTS

5.1 Simulation Results

Simulations were run for both ML and REML estimation methods for LME in comparison to Limma. Expected values of performance rates TPR, FPR, ACC, SPC, PPV, NPV, FDR and F_1 score were all calculated for LME and Limma. Besides, the number and proportion of probe sets that were found significant or non-significant by both LME and Limma for cell type, exposure and time parameters were reported.

Simulation results make a long table since there are many parameter settings. Therefore, only representative tables and figures will be displayed in this chapter. Full list of tables can be found in appendix.

5.1.1 Results Based on Cell Type Parameter

Results based on REML estimation in LME and Limma results were tabulated in Table 5.1, Table 5.2 and Table 5.3 respectively. Both results were obtained by using the same simulated datasets that contain 500 probe sets, parameter is cell type, number of cell types is 2, number of replicates is 2, number of time points is 3.

Table 5.1 Simulation results based on REML estimation for LME (foldchange=1.5)

p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
0.05	LME	0.725	0.001	0.959	0.999	0.993	0.955	0.007	0.838
	Limma	0.999	0.075	0.935	0.925	0.683	1.000	0.317	0.811
0.10	LME	0.952	0.001	0.992	0.999	0.992	0.991	0.008	0.971
	Limma	0.999	0.121	0.896	0.879	0.572	1.000	0.428	0.727
0.20	LME	0.986	0.002	0.997	0.998	0.990	0.998	0.010	0.988
	Limma	0.999	0.205	0.823	0.795	0.439	1.000	0.561	0.610
0.30	LME	0.993	0.003	0.997	0.997	0.983	0.999	0.017	0.988
	Limma	0.999	0.285	0.754	0.715	0.361	1.000	0.639	0.530
0.40	LME	0.998	0.003	0.997	0.997	0.979	1.000	0.021	0.988
	Limma	0.999	0.366	0.684	0.634	0.305	1.000	0.695	0.467

Table 5.2 Simulation results based on REML estimation for LME (foldchange=2)

p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
0.05	LME	0.693	0.001	0.953	0.999	0.992	0.950	0.008	0.816
	Limma	0.999	0.075	0.935	0.925	0.687	1.000	0.313	0.814
0.10	LME	0.953	0.001	0.992	0.999	0.993	0.991	0.007	0.972
	Limma	0.999	0.123	0.894	0.877	0.573	1.000	0.427	0.728
0.20	LME	0.985	0.001	0.997	0.999	0.992	0.998	0.008	0.988
	Limma	0.999	0.207	0.822	0.793	0.443	1.000	0.557	0.614
0.30	LME	0.993	0.002	0.997	0.998	0.985	0.999	0.015	0.989
	Limma	0.999	0.284	0.756	0.716	0.366	1.000	0.634	0.536
0.40	LME	0.998	0.003	0.997	0.997	0.979	1.000	0.021	0.989
	Limma	0.999	0.367	0.685	0.633	0.309	1.000	0.691	0.472

Table 5.3 Simulation results based on REML estimation for LME (foldchange=3)

p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
0.05	LME	0.697	0.000	0.954	1.000	0.996	0.951	0.004	0.820
	Limma	1.000	0.075	0.936	0.925	0.684	1.000	0.316	0.812
0.10	LME	0.962	0.001	0.994	0.999	0.996	0.993	0.004	0.979
	Limma	1.000	0.124	0.893	0.876	0.565	1.000	0.435	0.722
0.20	LME	0.990	0.001	0.998	0.999	0.995	0.998	0.005	0.993
	Limma	1.000	0.209	0.820	0.791	0.436	1.000	0.564	0.607
0.30	LME	0.994	0.001	0.998	0.999	0.991	0.999	0.009	0.993
	Limma	1.000	0.289	0.751	0.711	0.358	1.000	0.642	0.527
0.40	LME	0.999	0.002	0.998	0.998	0.986	1.000	0.014	0.993
	Limma	1.000	0.373	0.679	0.627	0.301	1.000	0.699	0.463

Limma performed better than LME only when TPR is the main concern. Limma produced almost perfect TPR for the given parameter settings against very high TPR levels of LME. The difference is only noticable at 0.05 p-value cutoff level. Limma tends to produce very small p-values for cell type parameter leading to very high FDR. Its FDR turned out to be very high. LME is concretely superior in FDR. Both LME and Limma did not produce differing results in different time intervals since both models were fit independently in every interval. In order to lessen the number of figures the results given in the following tables were based only on interval 1. However, results showing all intervals as well as the TPR, FPR, ACC, SPC, PPV, NPV and F1 value were sketched in the appendix.

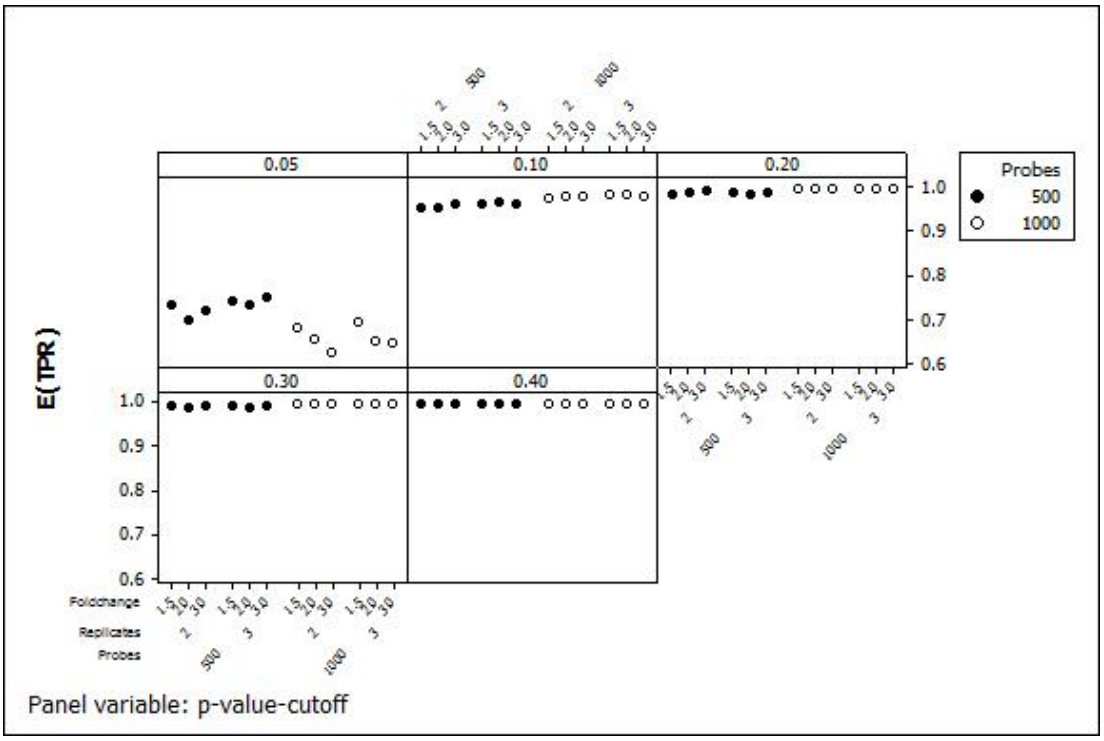


Figure 5.1 Expected TPR of LME for cell type parameter

TPR performance of LME become 0.95 and get close to 1 when p-value cutoff was selected 0.10 and larger. TPR became almost 1 as the cutoff value of p-value was selected 0.20 or higher as given in Figure 5.1. As the p-value cutoff was increased, the expected false discovery rate increased very slightly as in Figure 5.2. Considering both figures, the optimum value for the p-value cutoff for LME was found to be 0.20. However, even for 0.40 cutoff value, expected FDR was found to be around 0.015.

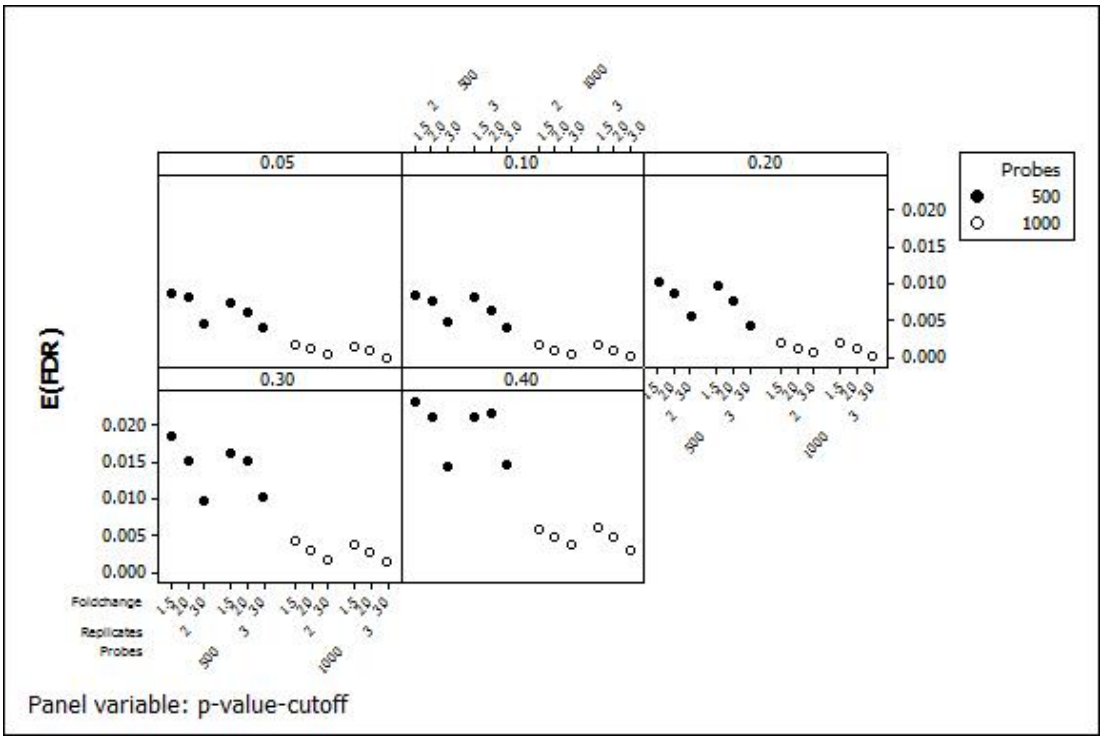


Figure 5.2 Expected FDR of LME for cell type parameter

Almost all the expected TPR values given in Figure 5.3 are equal to 1. On the other hand, FDR values are unacceptable and increase dramatically as the p-value cutoff were increased (Figure 5.4). The most reasonable p-value cutoff selection for Limma was 0.05. One should note that Limma results were corrected by Benjamini-Hochberg multiple testing procedure and multiple testing corrections lose efficiency as the number of test items increase. Both methods performed better for higher values of foldchange as expected.

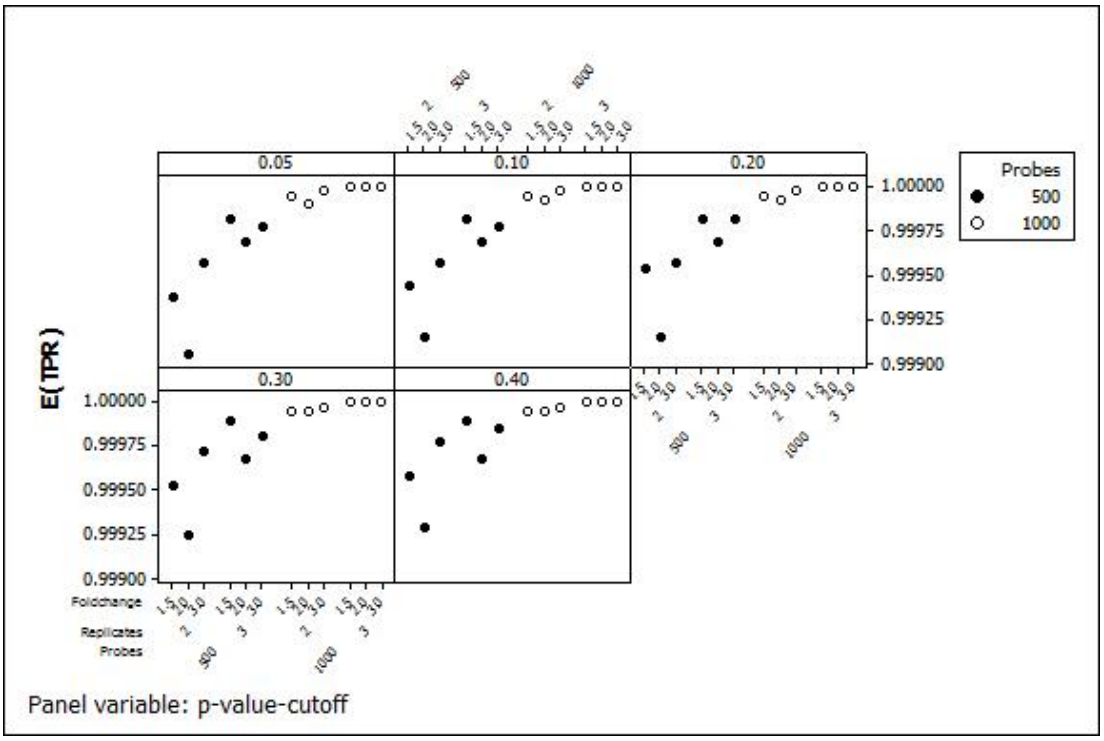


Figure 5.3 Expected TPR of Limma for cell type parameter

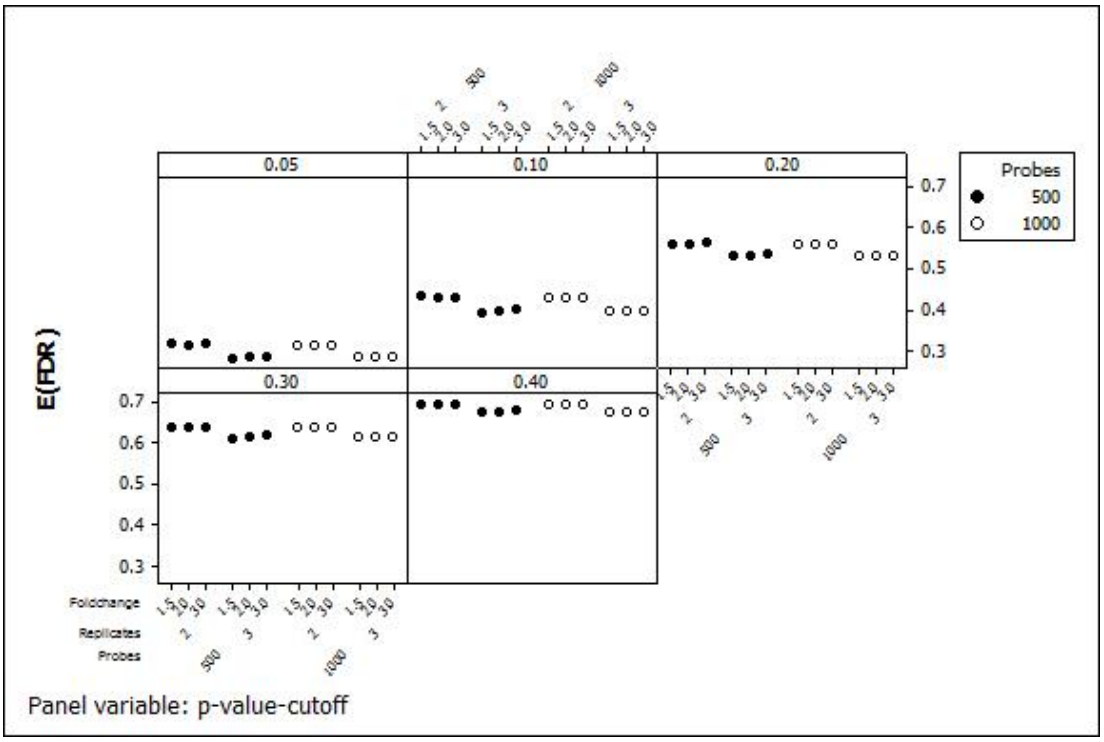


Figure 5.4 Expected FDR of Limma for cell type parameter

Simulation results were also compared in terms of test results from both methods. The category axis of the graph (x-axis) in Figure 5.5, Figure 5.6 and Figure 5.7 has definitions for a side by side comparison such as “Both Not” which means both methods did not detect any differential expression, “Both Sig” which means both methods found the probe sets as differentially expressed, “Limma Sig” which means LME did not detect any differential expression but Limma did and finally “LME Sig” which means LME detected differential expression but Limma did not. Resulting proportions by fold change (1.5, 2 and 3) are given in Figure 5.5, Figure 5.6 and Figure 5.7 respectively for cell type parameter.

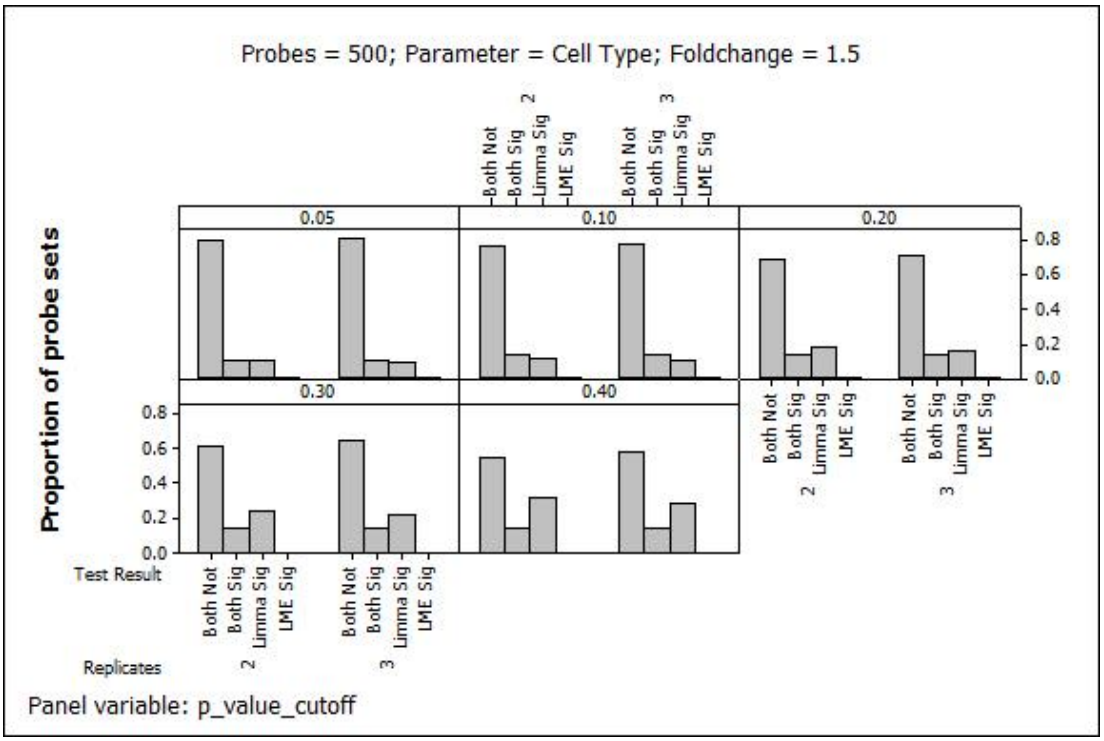


Figure 5.5 Significance test results of probe sets (foldchange=1.5) for cell type parameter

According to the significance test results, in the test of cell type parameter, Limma produced very high number of significant results that were not found to be significant by LME. However, the vice versa did not happen. This may be due to the high number of FDR of Limma. As expected, the number of significant detected probe sets increased as the p-value cutoff value was increased. Change in the fold change did not differ the results significantly.

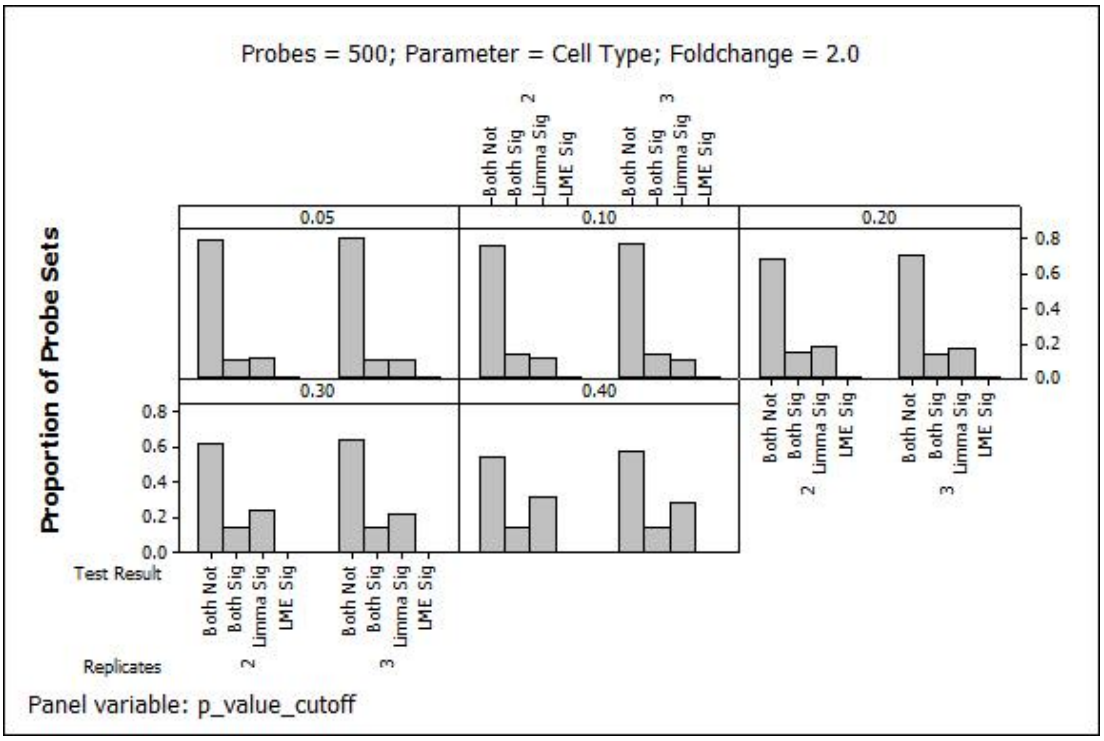


Figure 5.6 Significance test results of probe sets (foldchange=2.0) for cell type parameter

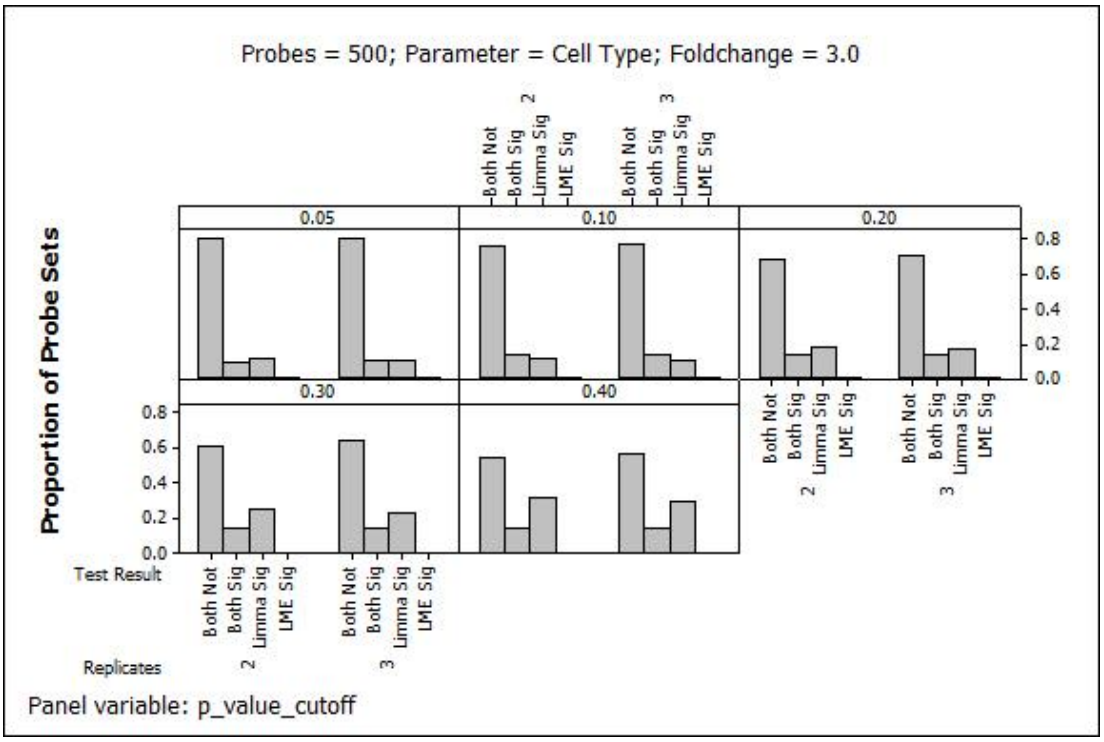


Figure 5.7 Significance test results of probe sets (foldchange=3.0) for cell type parameter

5.1.2 Results Based on Exposure Parameter

TPR results for different probe numbers, replicates and p-value cutoff values can be seen in Figure 5.8. LME tended to produce slightly better TPR results for exposure parameter as the number of simulated probe sets were increased from 500 to 1000. At the 0.20 significance level, almost all the TPR values were close to 1.

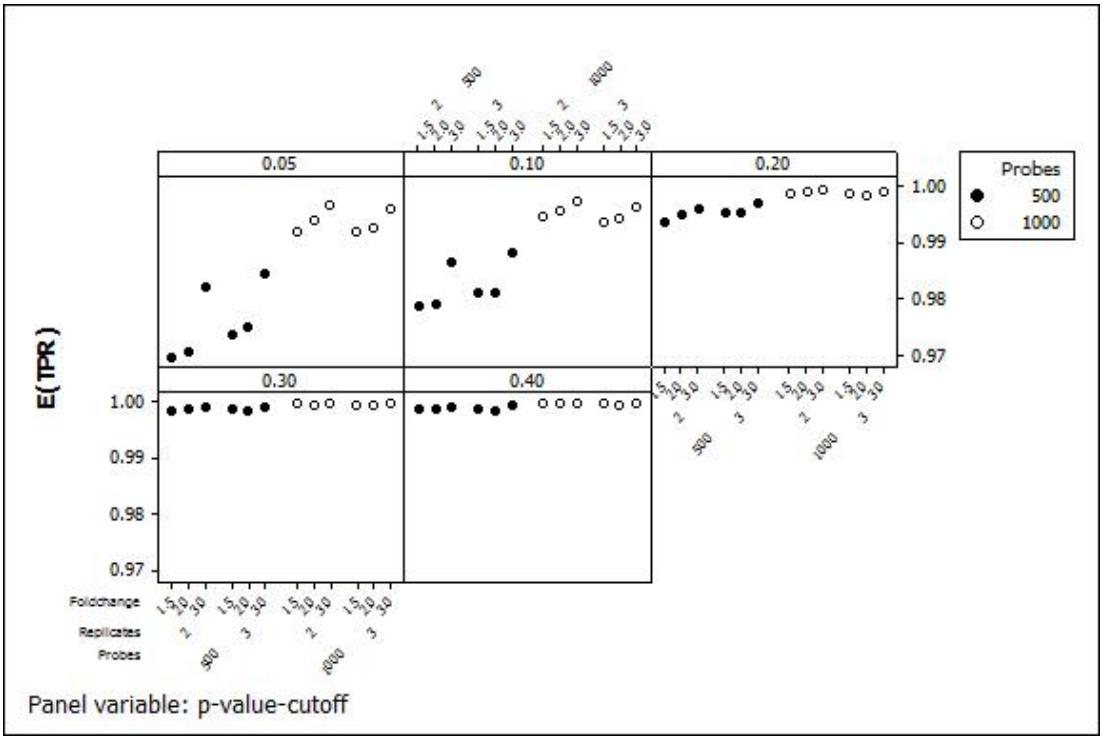


Figure 5.8 Expected TPR of LME for exposure parameter

TPR values must be evaluated together with the FDR values where LME definitely outperformed Limma. No matter what the significance level was, LME produced very low FDR values as given in Figure 5.9.

Limma also performed well for TPR as given in Figure 5.10. However, the difference came in with the FDR values of Limma. 30% to 70% of the probe sets were falsely discovered as significant by Limma (Figure 5.11) which is very high compared to those of LME where FDR values changed from 0.5% to 3%.

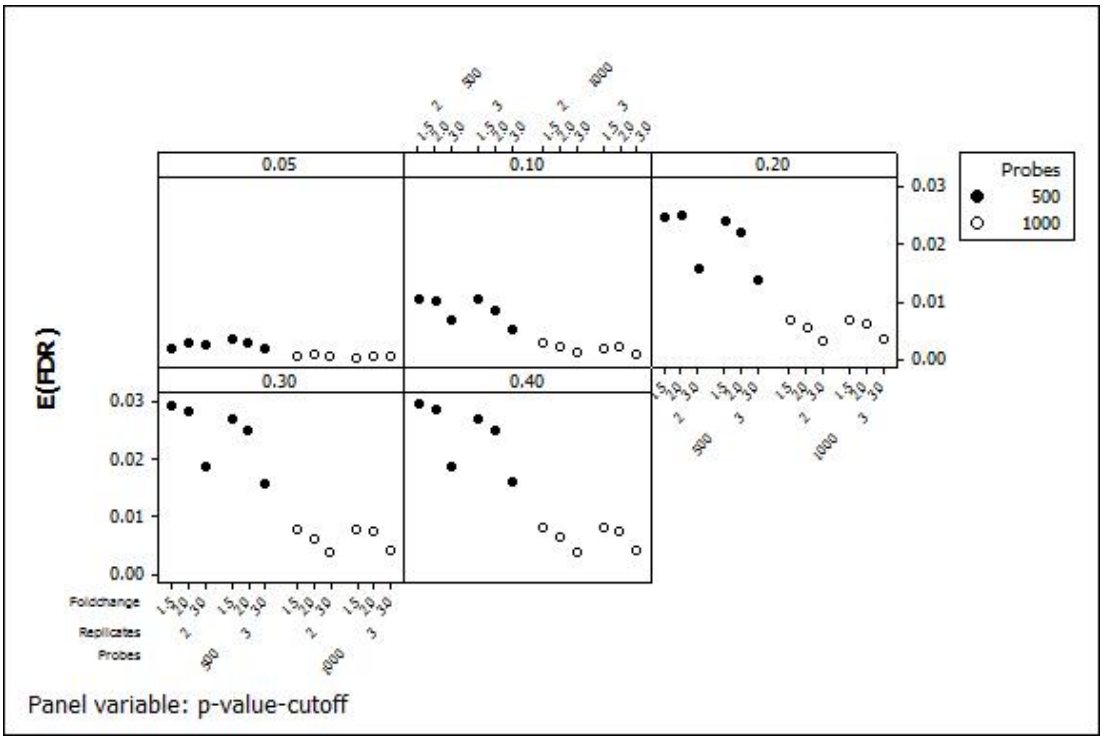


Figure 5.9 Expected FDR of LME for exposure parameter

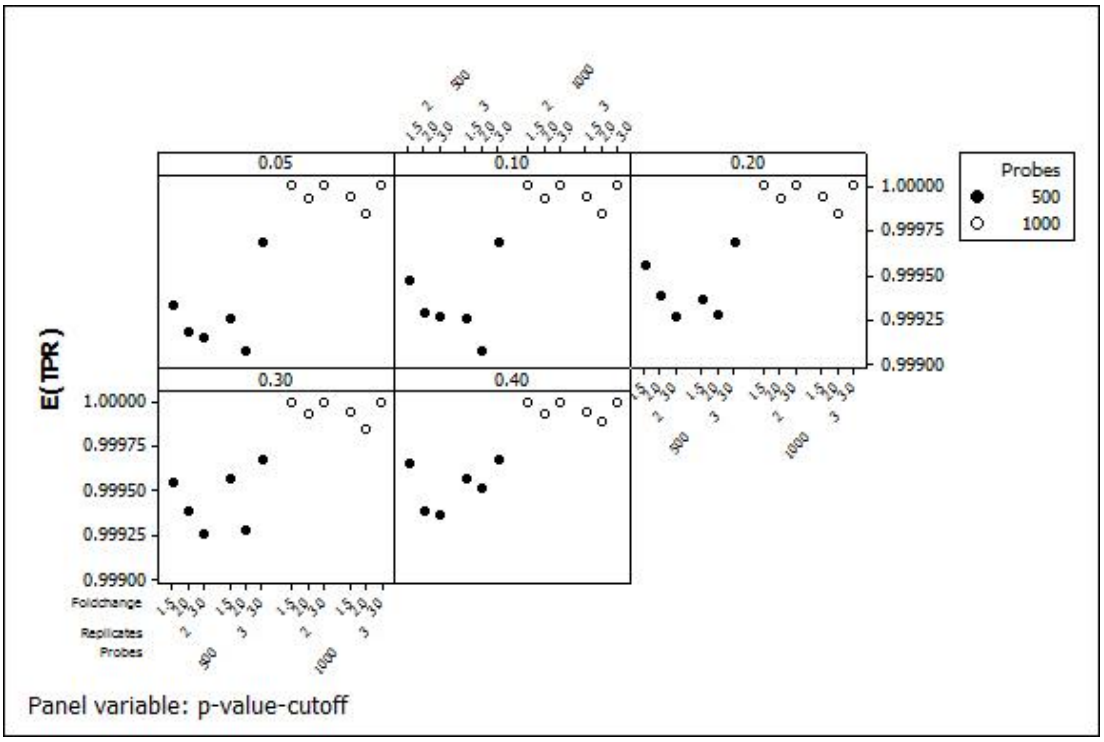


Figure 5.10 Expected TPR of Limma for exposure parameter

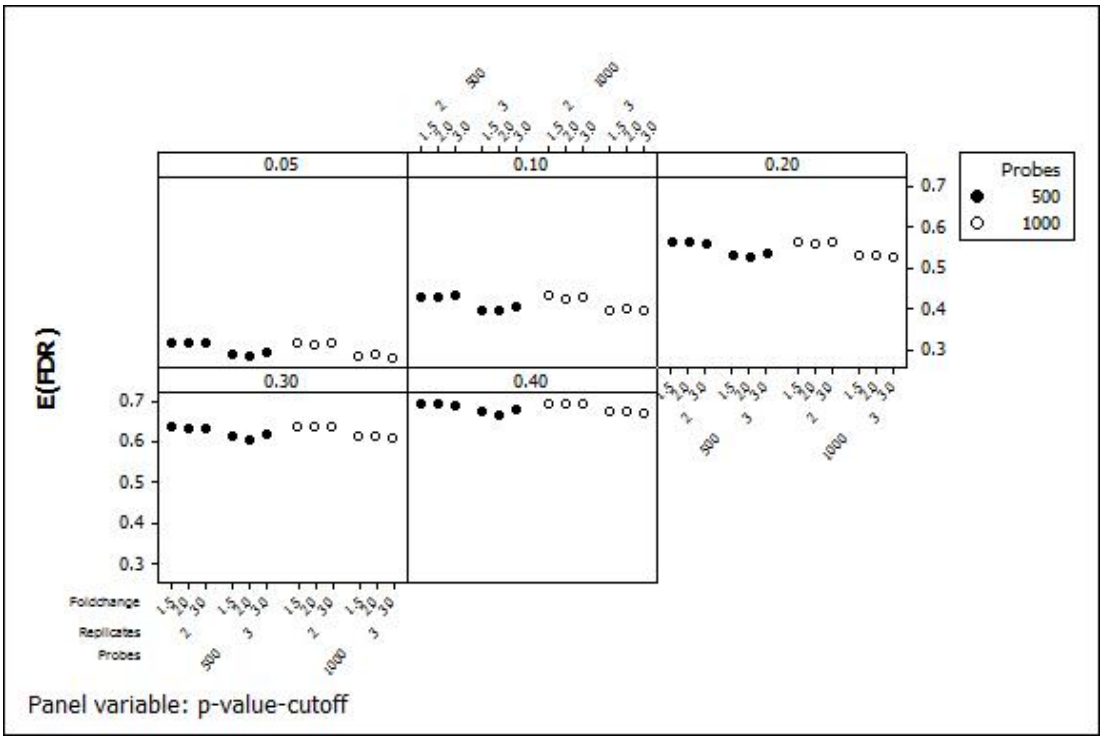


Figure 5.11 Expected FDR of Limma for exposure parameter

According to the significance test results, a similar result were observed in the test of cell type parameter. Specifically, Limma produced very high number of significant results that were not found to be significant by LME for the exposure parameter. As expected, the number of significant detected probe sets increased as the p-value cutoff value was increased. Change in the fold change did not differ the results significantly. Resulting proportions by the number of replicates and fold change are given in Figure 5.12, Figure 5.13 and Figure 5.14 for exposure parameter.

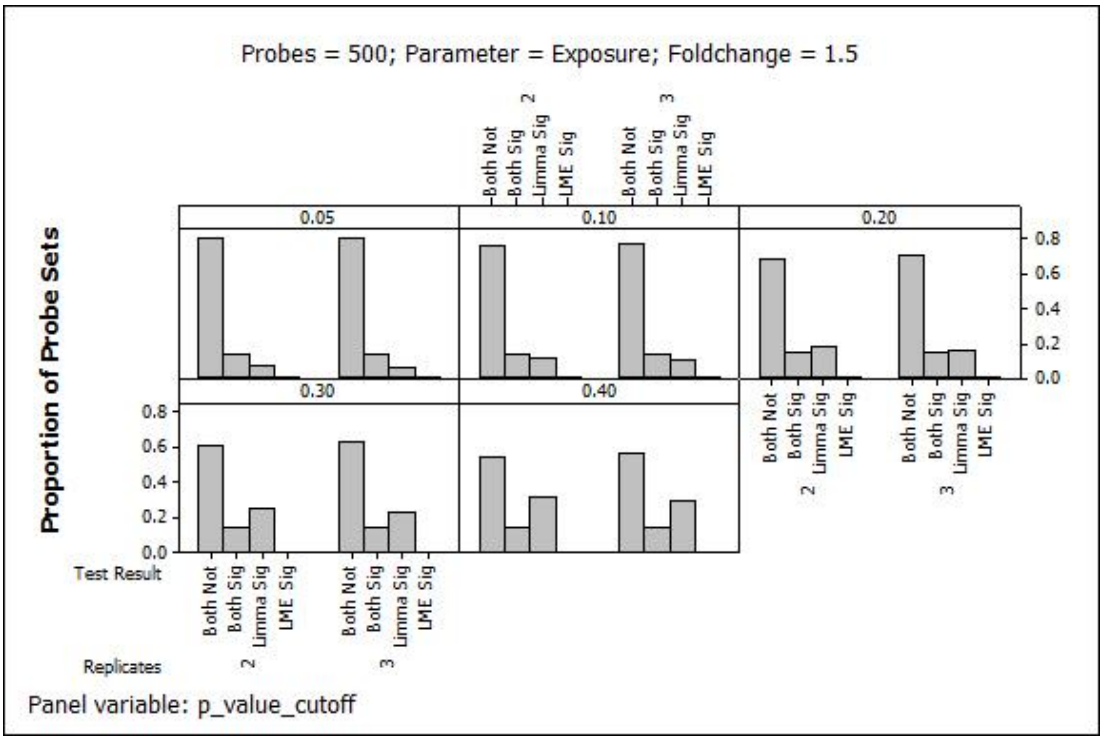


Figure 5.12 Significance test results of probe sets (foldchange=1.5) for exposure parameter

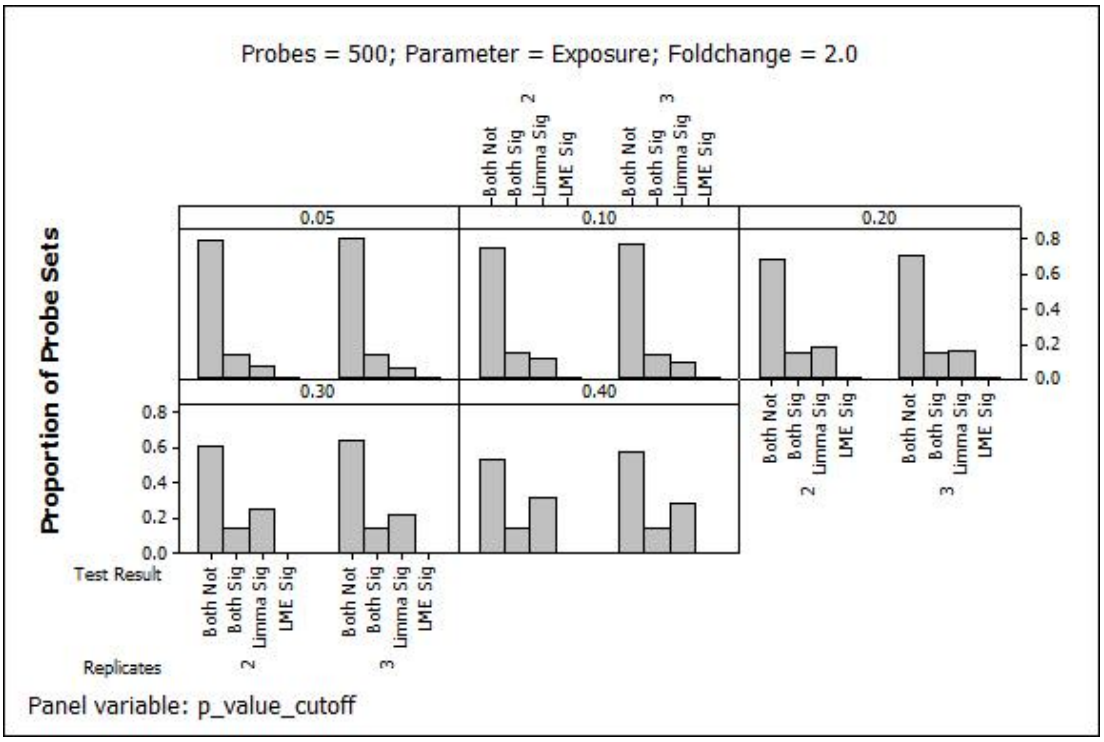


Figure 5.13 Significance test results of probe sets (foldchange=2.0) for exposure parameter

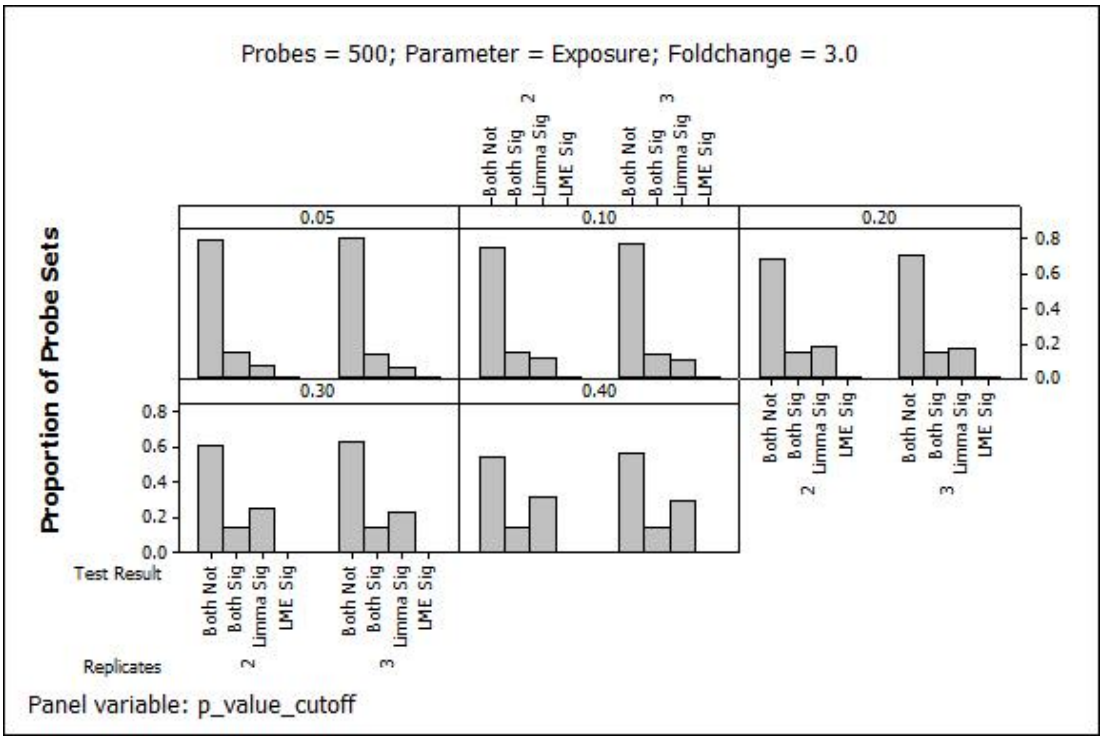


Figure 5.14 Significance test results of probe sets (foldchange=3.0) for exposure parameter

5.1.3 Results Based on Time Parameter

TPR and FDR results on time parameter for LME and Limma for different probe numbers, replicates and p-value cutoff values can be seen in Figure 5.15 through Figure 5.18. LME also poduced superior results for time parameter in both TPR and FDR. Time parameter is an independent explanatory variable in the model and as explained in Chapter 3 and especially in Figure 3.10, time lag is accounted for in LME. On the other hand, the time parameter can only be treated as factor by Limma for which drastical decreases in TPR values of LME were observed. For the TPR values of the time parameter, Limma could not go over 67.5% but still produced very high FDR values (Figure 5.17 and Figure 5.18).

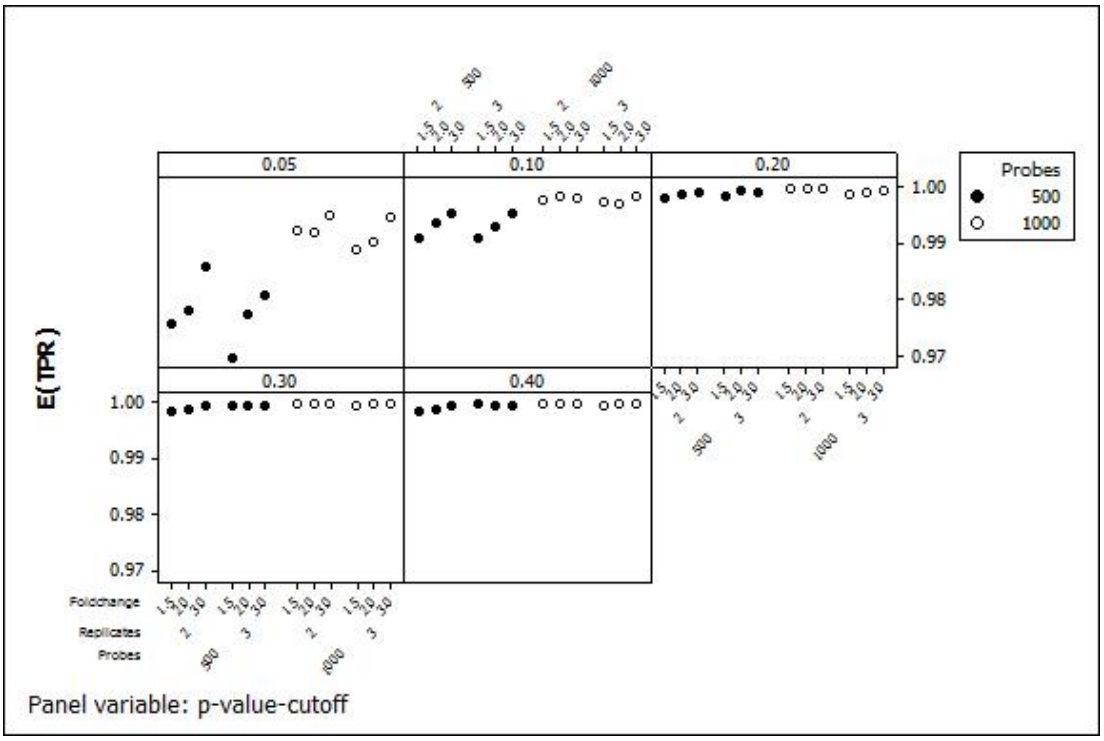


Figure 5.15 Expected TPR of LME for time parameter

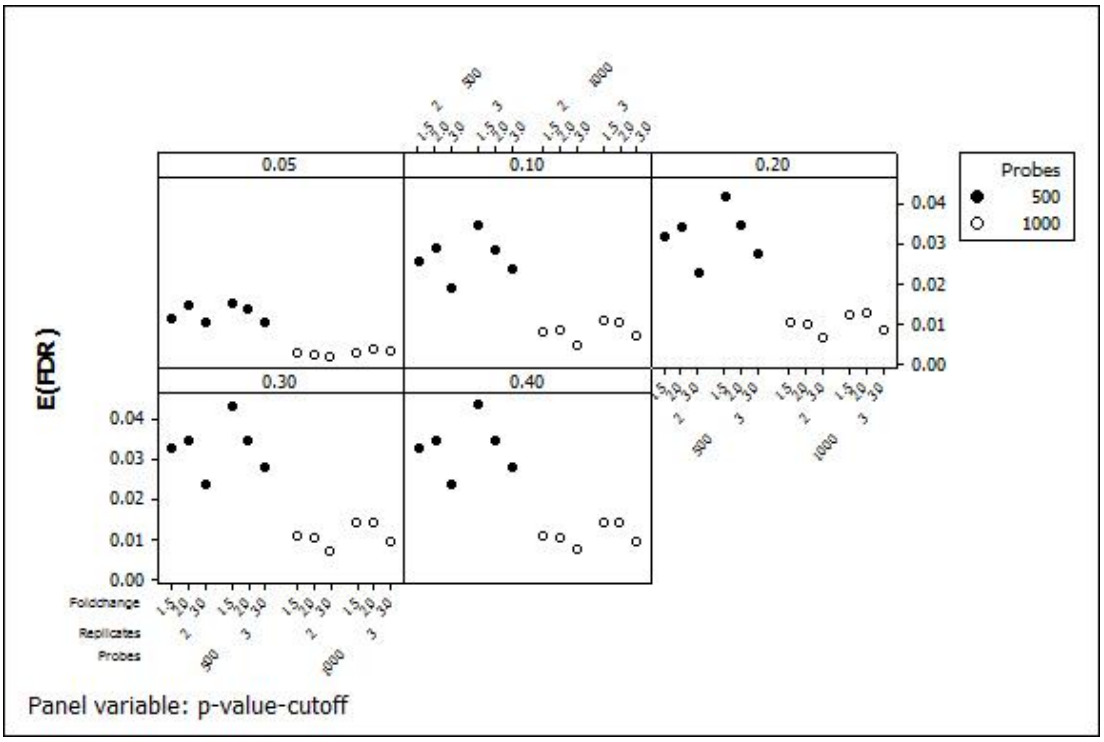


Figure 5.16 Expected FDR of LME for time parameter

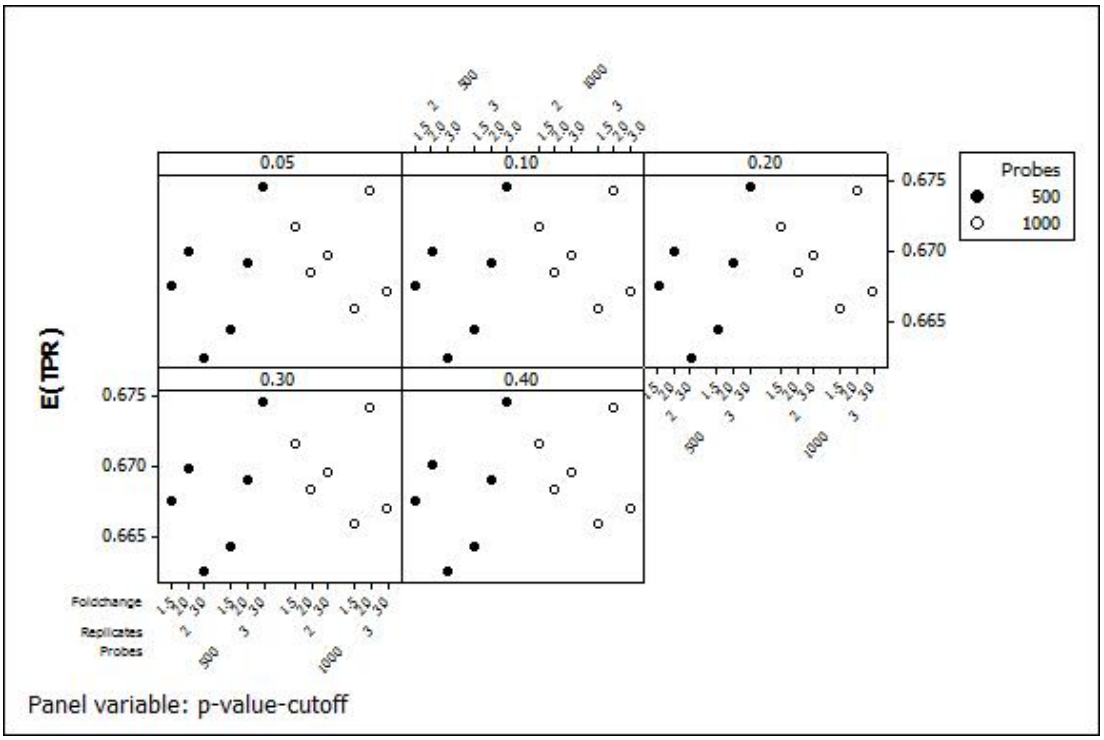


Figure 5.17 Expected TPR of Limma for time parameter

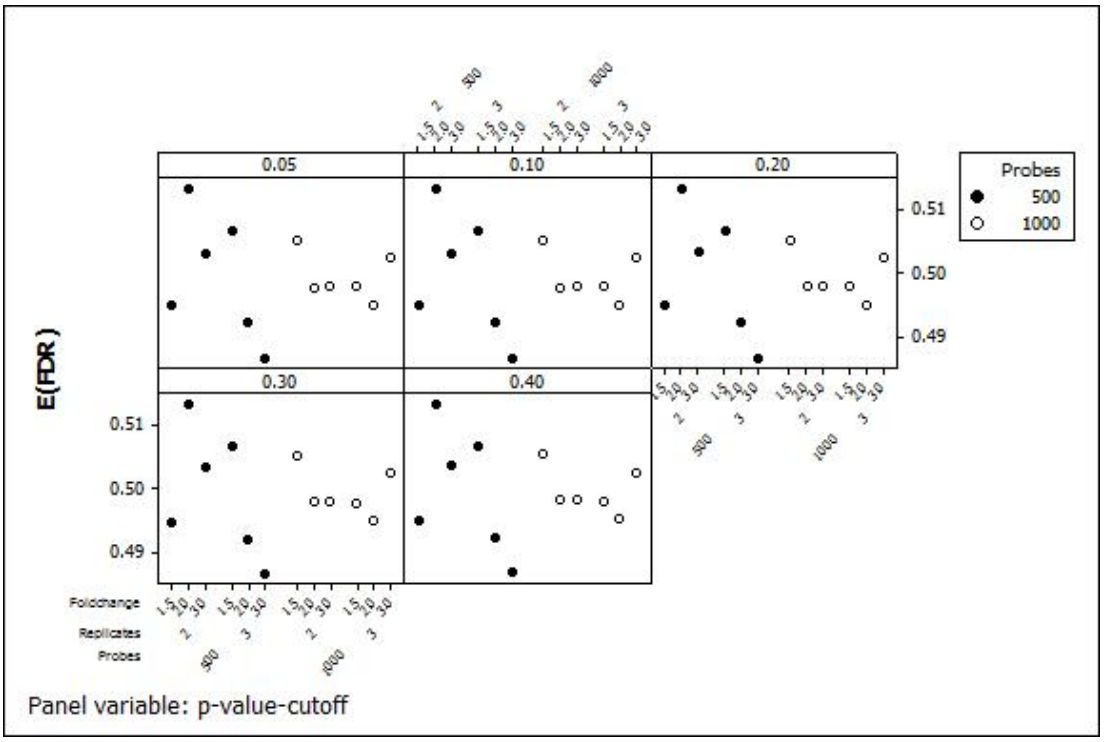


Figure 5.18 Expected FDR of Limma for time parameter

According to the significance test results from both methods, 10% to 15% of the probe sets returned conflicting results. Time parameter was the only parameter for which some proportion of probe sets was found to be differentially expressed only by LME but not by Limma which can be clearly seen in Figure 5.19, Figure 5.20 and Figure 5.21 in increasing order of fold change from 1.5 to 3.0. The results, however, did not differ by fold change.

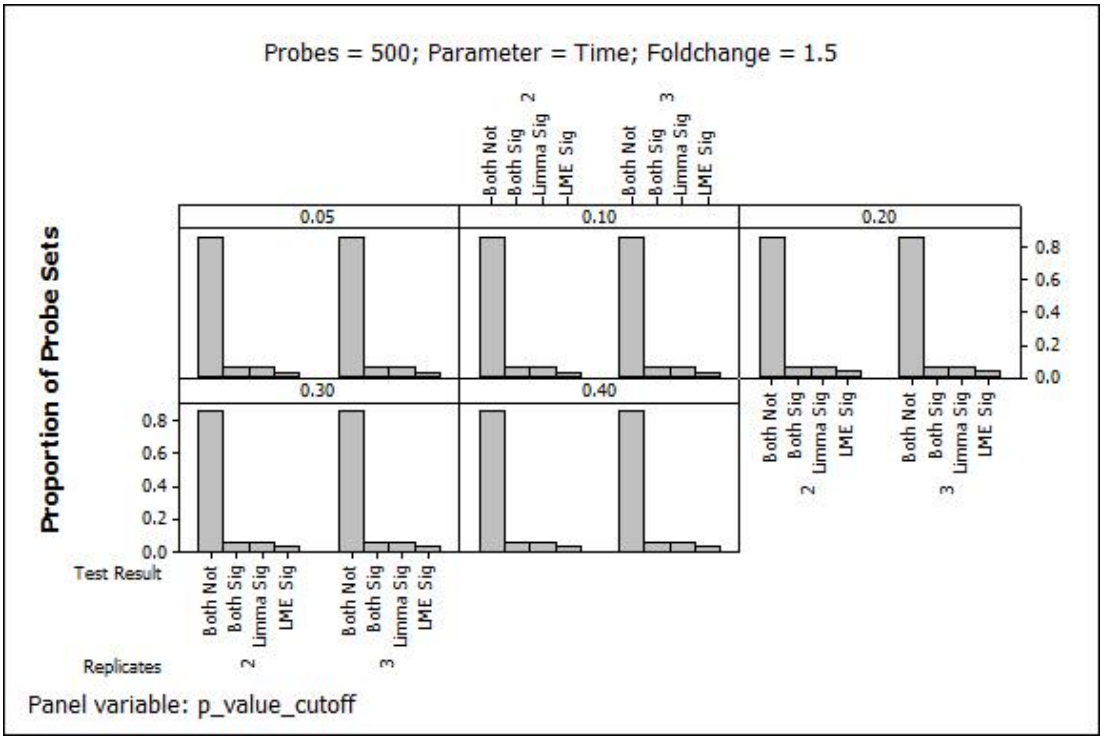


Figure 5.19 Significance test results of probe sets (foldchange=1.5) for time parameter

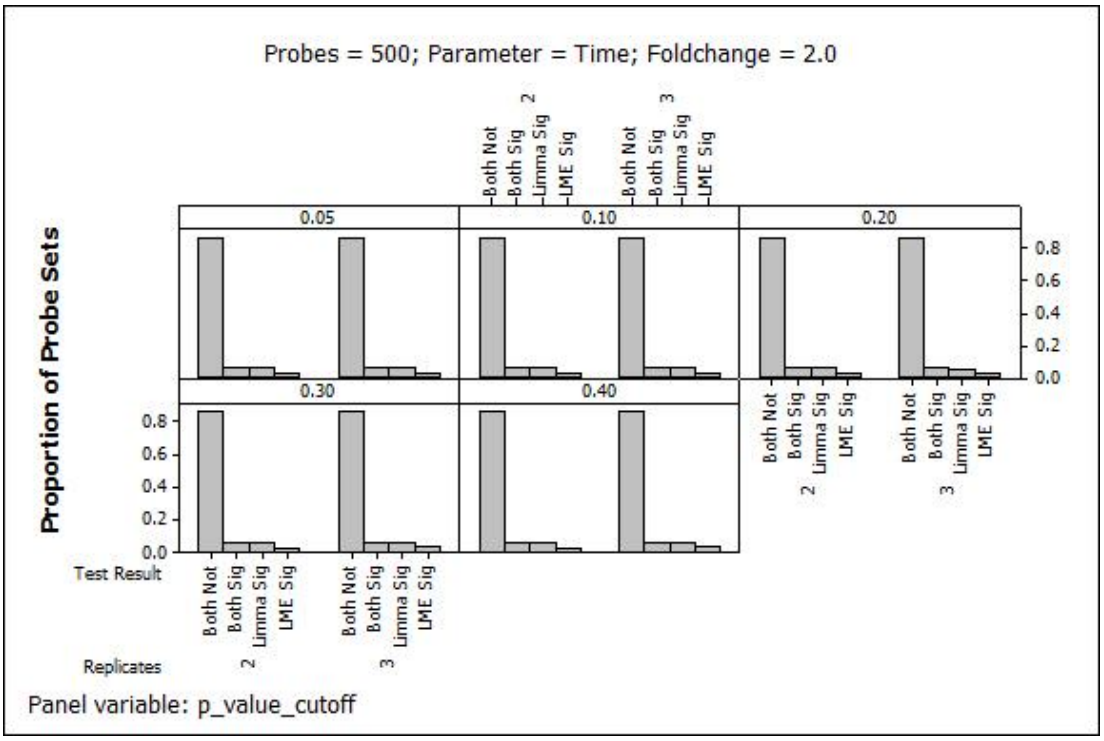


Figure 5.20 Significance test results of probe sets (foldchange=2.0) for time parameter

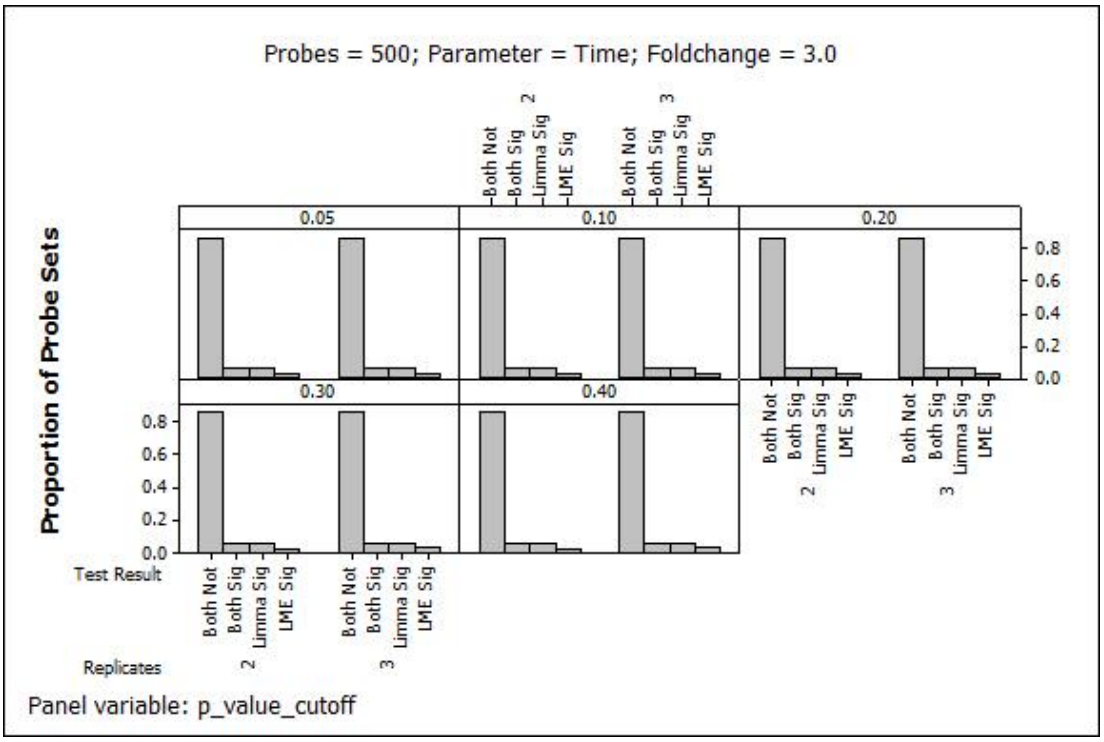


Figure 5.21 Significance test results of probe sets (foldchange=3.0) for time parameter

5.2 Results on Asbestos Dataset

The significance testing results for filtered and clustered asbestos dataset (19,771 probe sets in 18,771 clusters) were done for cell type, exposure and time parameters on 5 time intervals (e.g. 0h-1h, 1h-6h, 6h-24h, 24h-48h and 48h-168h). Although statistical analyses were based on clusters, the final performance measurements and comparisons were held on probe sets for a fair evaluation. The number of significant probes from filtered asbestos dataset at 20% significance level are given in Figure 5.22 for both LME and Limma methods. In accordance with the simulation results, the number of probe sets detected as significant by Limma is obviously larger than those of LME. Simulation results support that the reason for the large number of significantly detected probe sets and high TPR values is the false discoveries by Limma. At this point, the results of Limma are definitely incomparable to those of LME. For example, at the first time interval, LME detected 13% of the probes that are differentially expressed in the Beas2B cell type in comparison to A549 cell type. However, Limma detected 62% of the probes as differentially expressed in Beas2B in comparison to A549. The results are similar in all other time intervals.

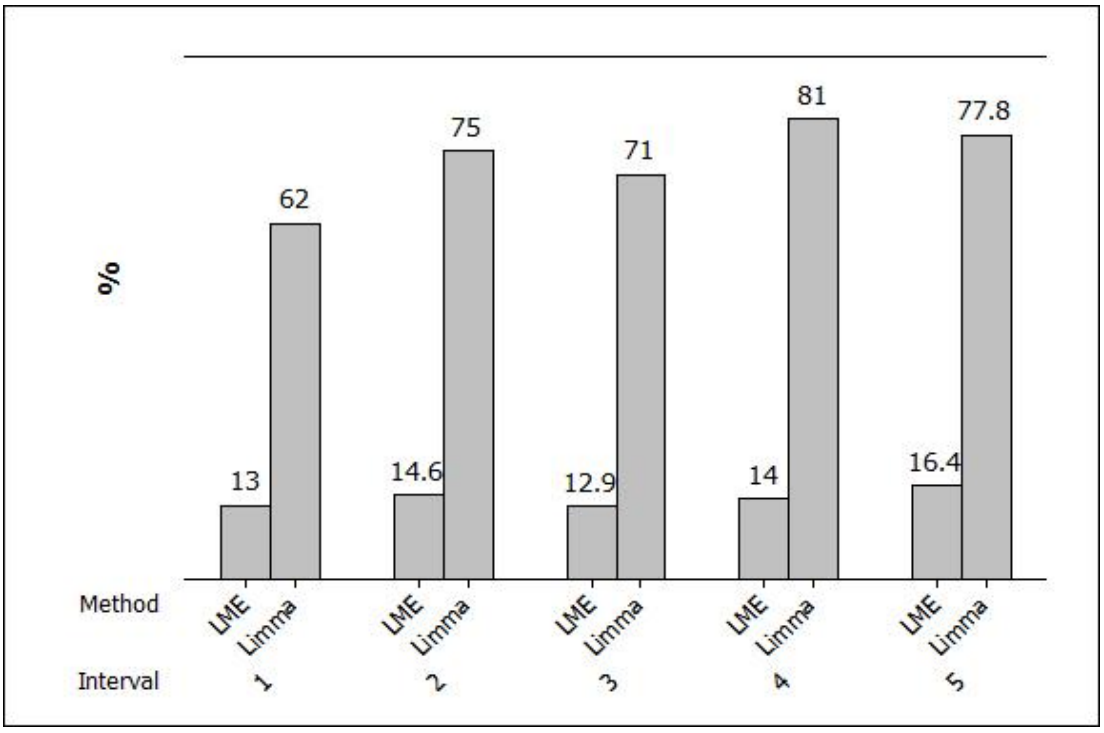


Figure 5.22 Proportion of probe sets found to be differentially expressed by cell type in each time interval

Figure 5.23 shows the proportion of the probe sets among all the significant probes that are significant in one or more time intervals. For example, among all the significant probe sets, 31% of the probes were found to be differentially expressed in only one interval by LME. Likewise, 22% of the probes were found to be differentially expressed in two time intervals by LME. On the other hand, 6% of the probe sets were detected as differentially expressed in only 1 time interval among the all probe sets that were found to be differentially expressed by Limma. This indicates that Limma tends to find more significantly expressed profiles on more intervals. According to Limma results, 46% of the probe sets were differentially expressed by cell type at all 5 intervals.

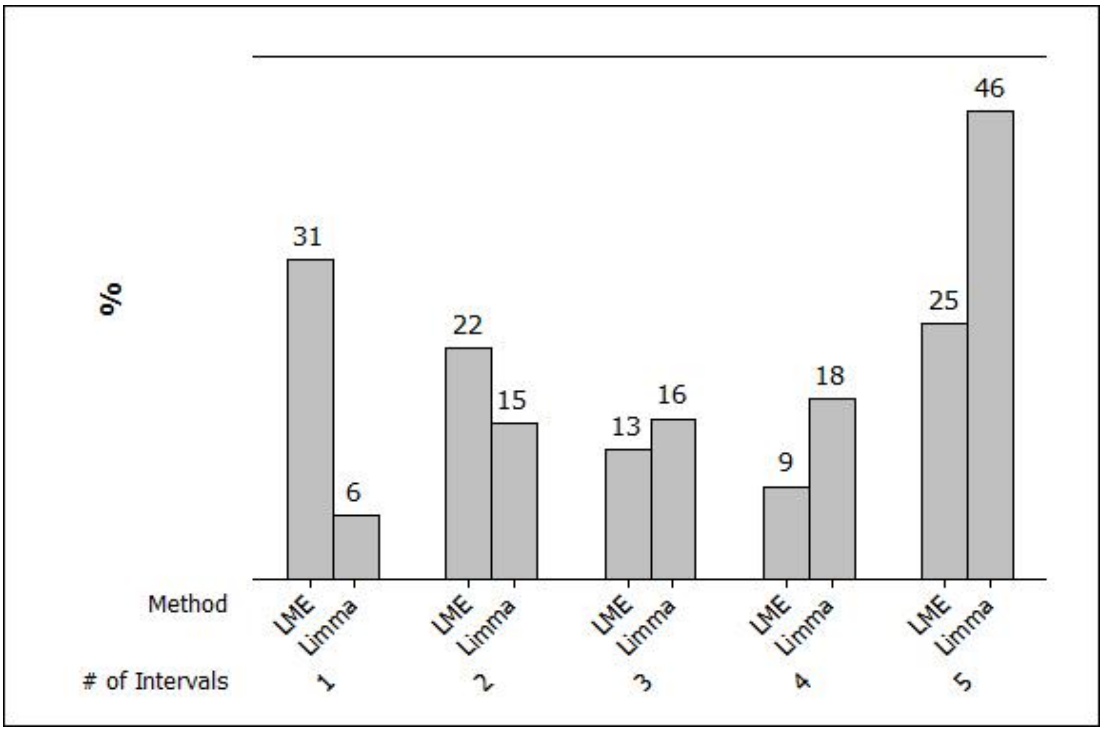


Figure 5.23 Proportion of probe sets found to be differentially expressed by cell type in one or more intervals

There was a very interesting result sketched in Figure 5.24 that Limma was unable to detect any probe sets as differentially expressed by the exposure in contrast with the LME's 18%, 7%, 17%, 3% and 3% significant probe set detection in subsequent time intervals. Even though Limma produced much larger FDRs compared to those of LME, interestingly it failed to detect any differential gene expression due to exposure.

Another interesting result from exposure effect that can be seen in Figure 5.25 is that 63% of the differentially expressed probe sets that were detected by LME were significant only

in one time interval, 32% were significant on two intervals. The exposure effect therefore, seem to be acute and its effect does not seem to last very long. Only 3% of the probe sets were differentially expressed on all intervals. There is always a doubt towards a false discovery though.

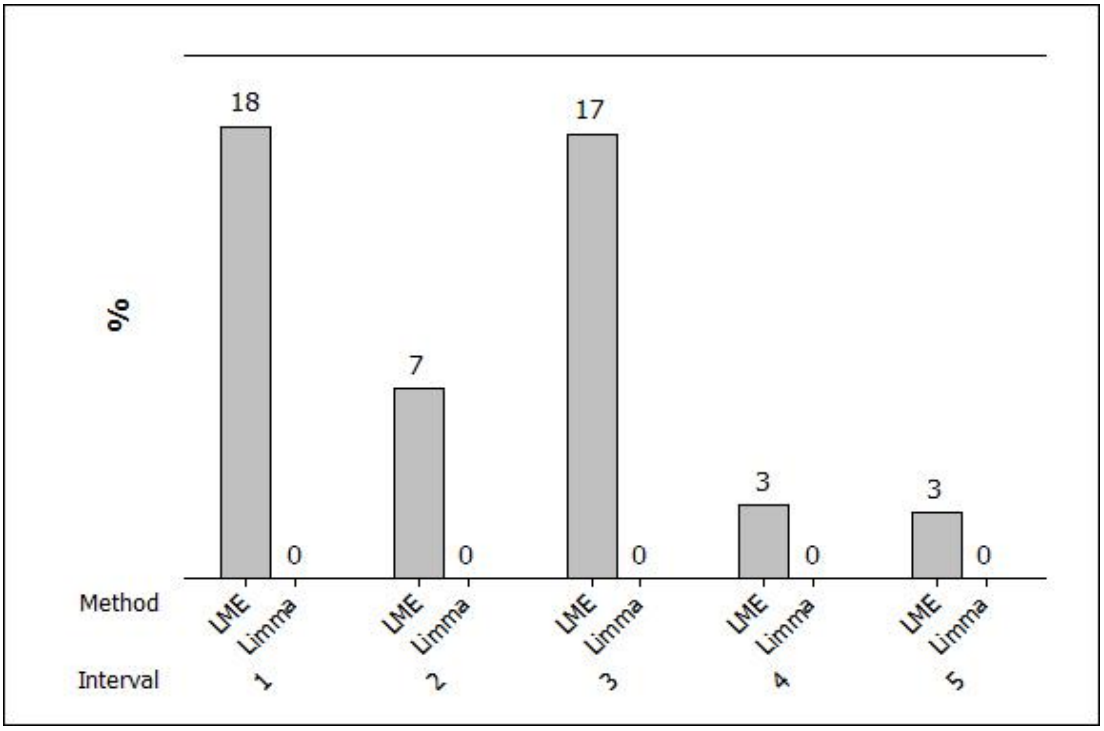


Figure 5.24 Proportion of probe sets found to be differentially expressed by exposure in each time interval

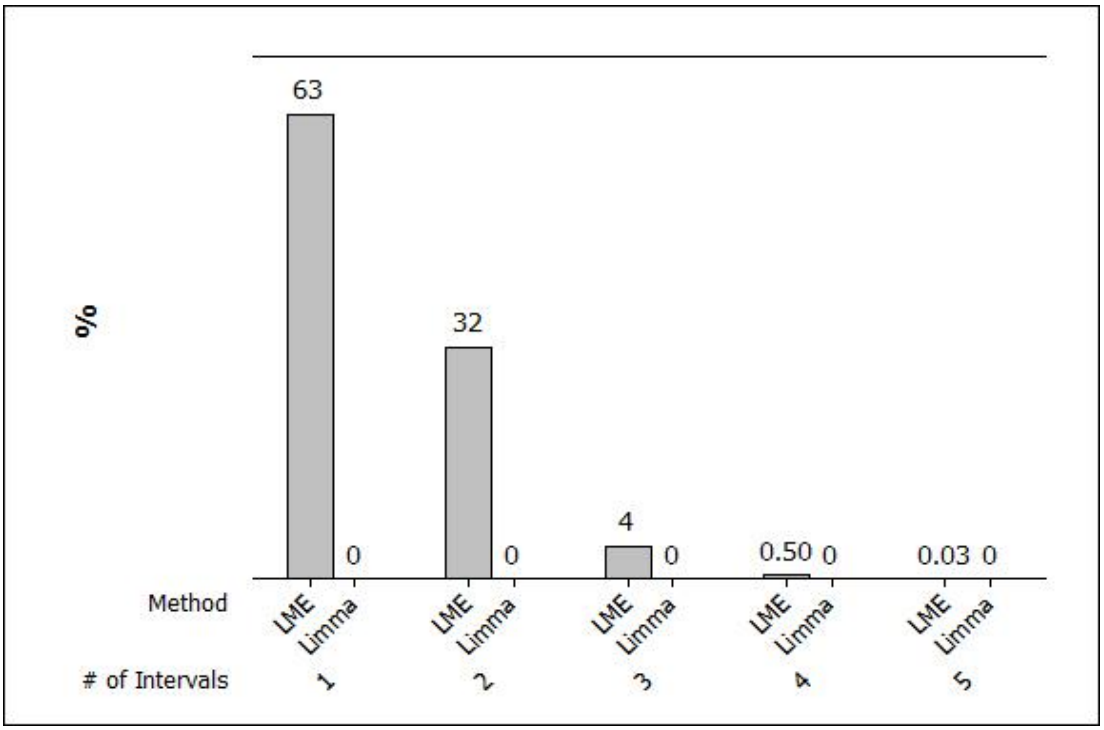


Figure 5.25 Proportion of probe sets found to be differentially expressed by exposure in one or more intervals

Time effect on probe sets seemed to be prominent in later intervals as it can be seen in Figure 5.26. The proportion of probe sets that were found to be differentially expressed by both methods due to time effect increase by time except for the last time interval. Limma again detected larger number of significant probe sets where majority is expected to be false discoveries. LME was able to detect around 15% of probes as differentially expressed especially in the last three intervals.

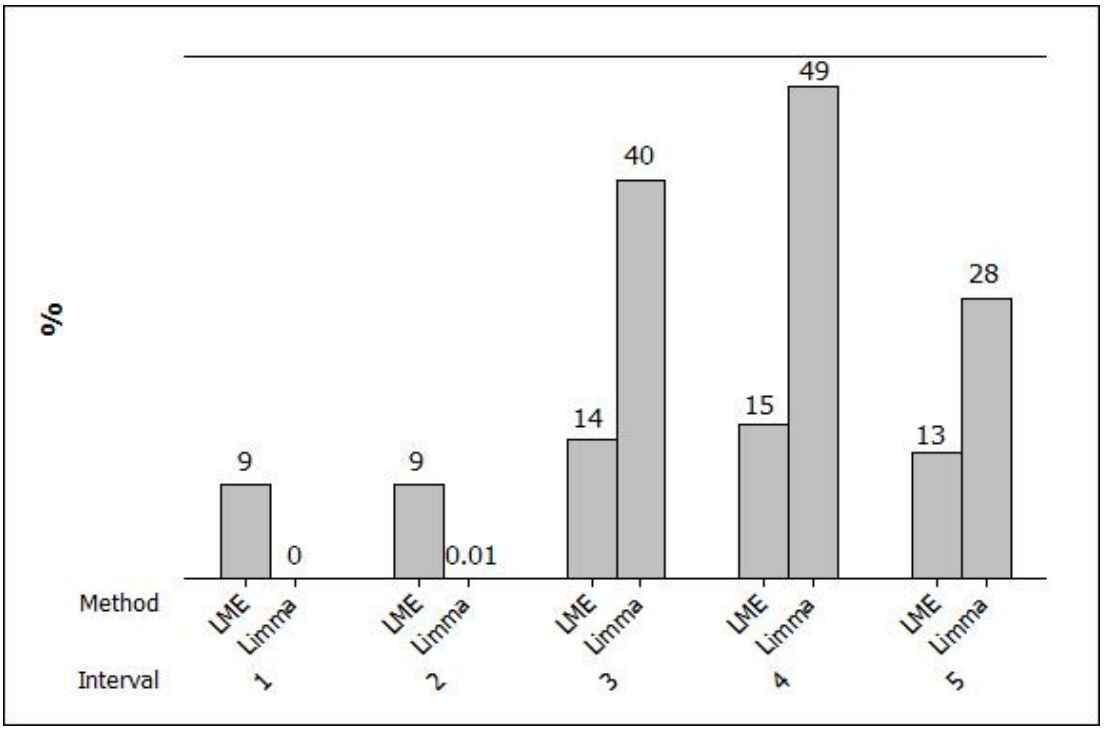


Figure 5.26 Proportion of probe sets found to be differentially expressed by time in each time interval

Among all the probe sets that were detected as differentially expressed by either methods, about 60% of the probe sets were found to be differentially expressed at only one time interval, about 30% at two intervals and about 10% at three intervals. At this point, LME and Limma returned similar detection patterns in terms of dispersion through the intervals which can be seen in Figure 5.27.

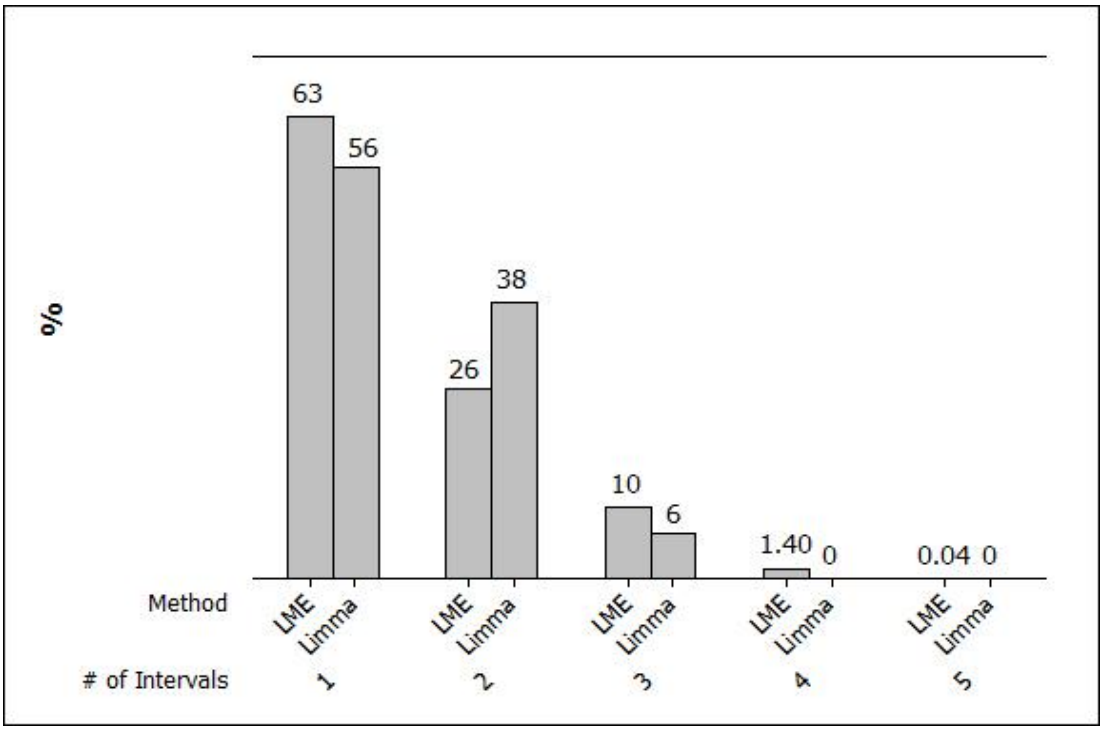


Figure 5.27 Proportion of probe sets found to be differentially expressed by time in one or more intervals

In order to compare the test results for cell type produced by both methods probe by probe yielded proportions of differentially expressed probe sets as in Figure 5.28. The category axis of the graph (x-axis) has definitions for a side by side comparison such as “Both Not” which means both methods did not detect any differential expression, “Both Sig” which means both methods found the probe sets as differentially expressed, “Limma Sig” which means LME did not detect any differential expression but Limma did and finally “LME Sig” which means LME detected differential expression but Limma did not. The graph is fair enough to see that Limma produced more positive test results compared to that of LME throughout all intervals. However, simulations in this study indicated that Limma tends to produce remarkably more false discoveries. Around 13% to 16% of the probe sets in different intervals were found to be differentially expressed by both methods.

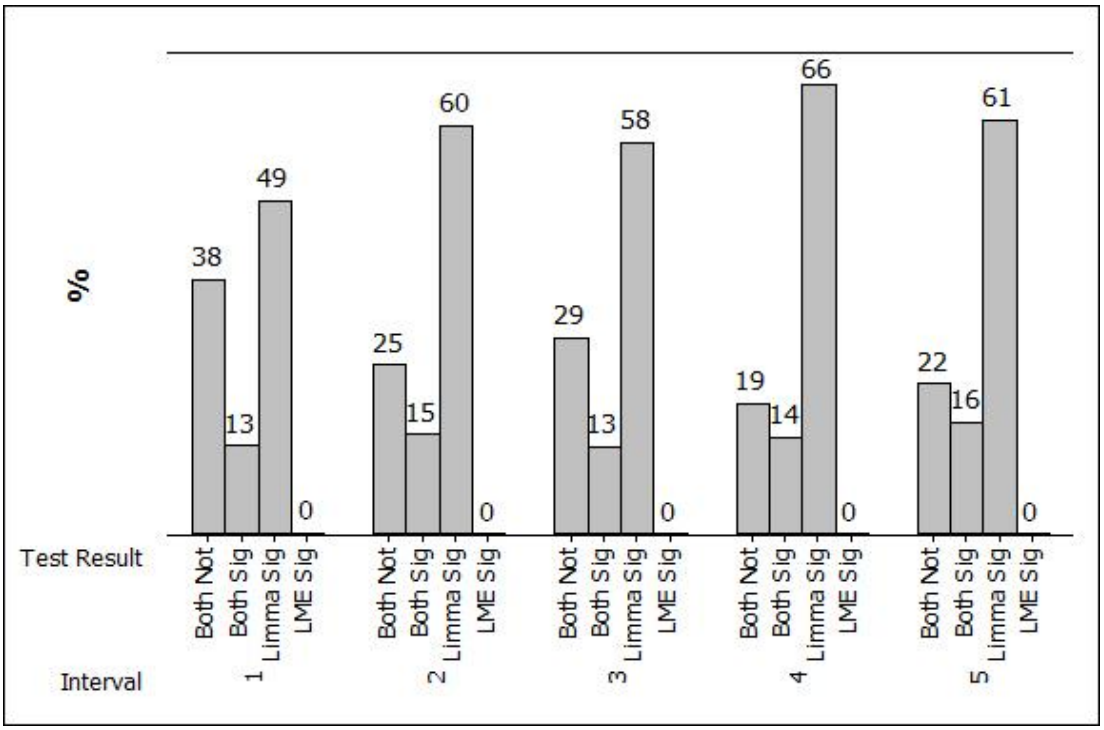


Figure 5.28 A probe by probe comparison of the test results for cell type effect

In order to illustrate some testing result from the LME model, some clusters and their significance test results are also presented. For example, Cluster 5001 contains six genes and for the first four time intervals there is no significant gene activity detected by LME model, but the increase in the gene expression level in the fifth interval was found to be significant. The cluster is sketched in Figure 5.29.

Another illustration as an example result by the LME model is Cluster 14283. According to the LME test result for this cluster, there is a significant exposure effect in the second, fourth and fifth time interval as well as there is a significant time effect in the second and the third intervals. For a better visualization of the exposure and time effects, the cluster is represented in two separate graphics as in Figure 5.31 and Figure 5.32 respectively in addition to the graph in Figure 5.30.

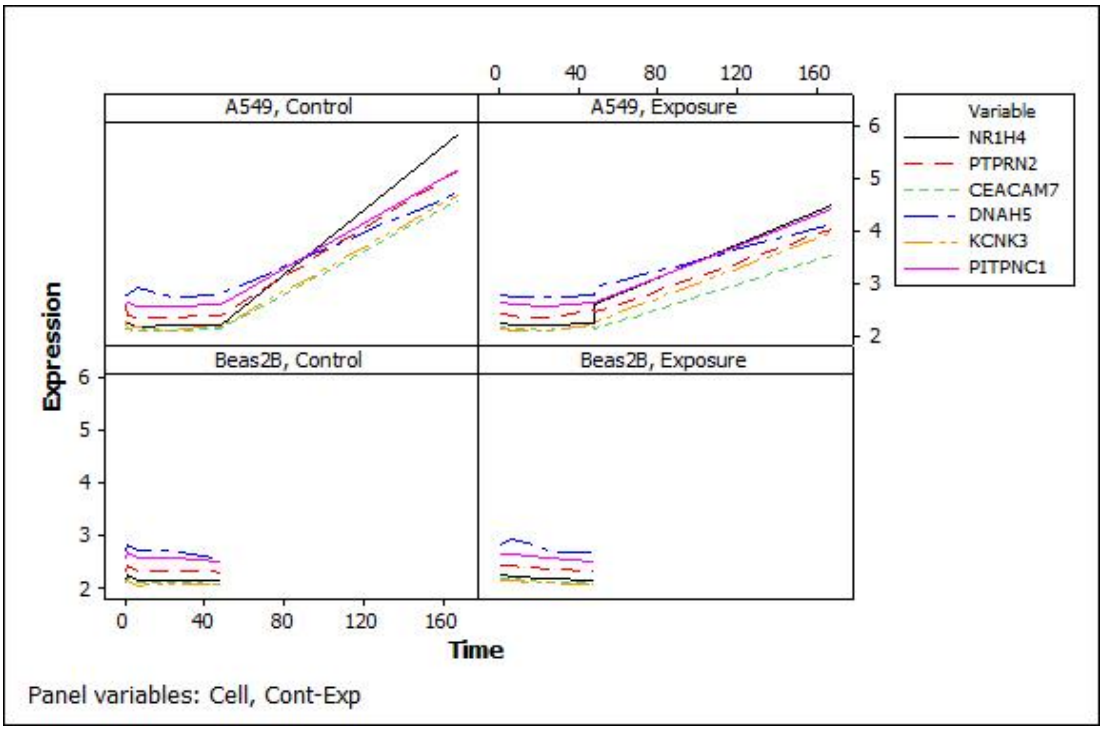


Figure 5.29 Gene expression profile of Cluster 5001

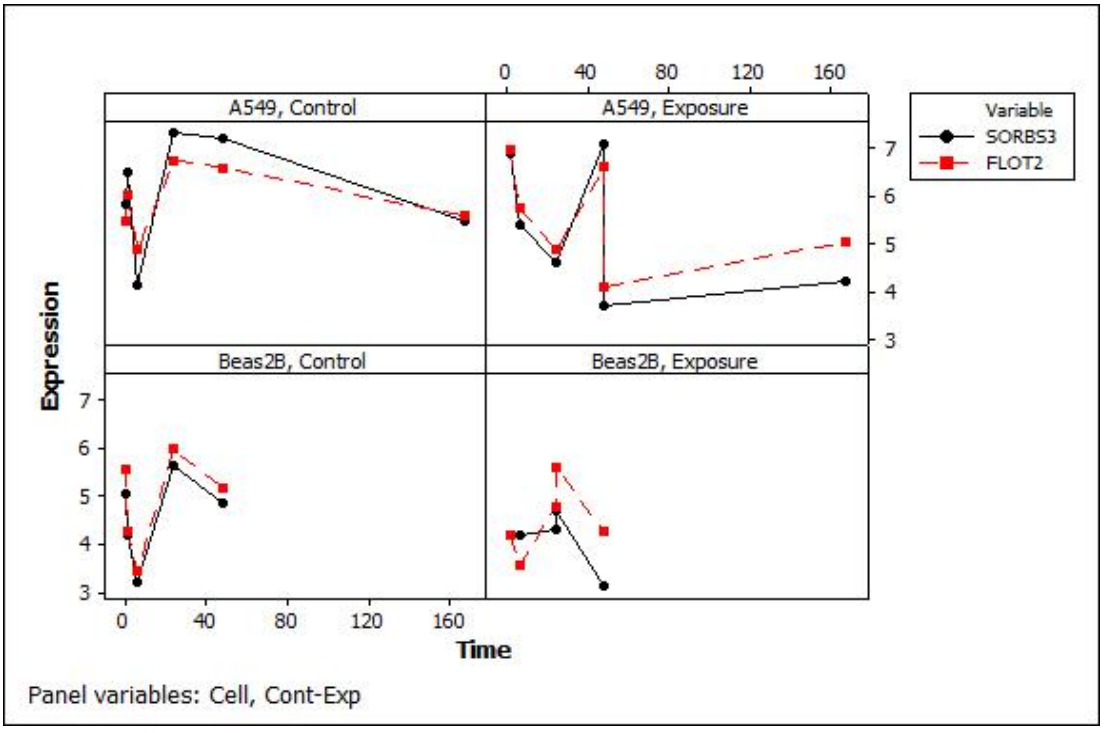


Figure 5.30 Gene expression profile of Cluster 14283

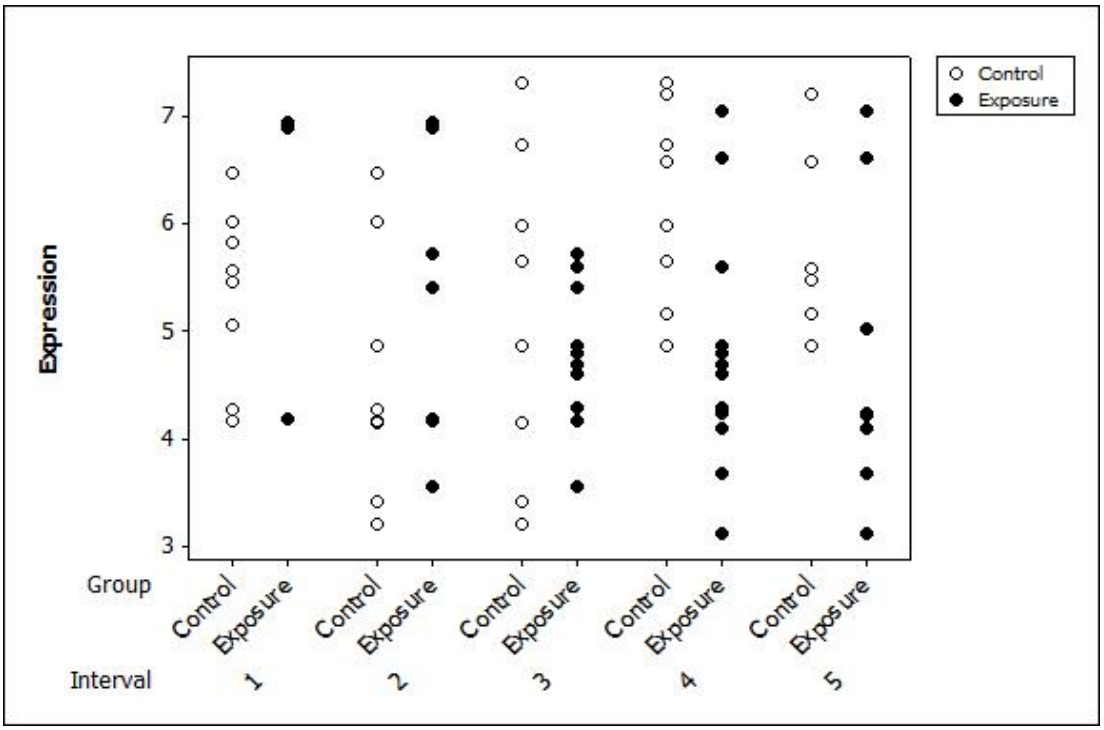


Figure 5.31 Control and exposure groups expression levels across the time intervals from Cluster 14283

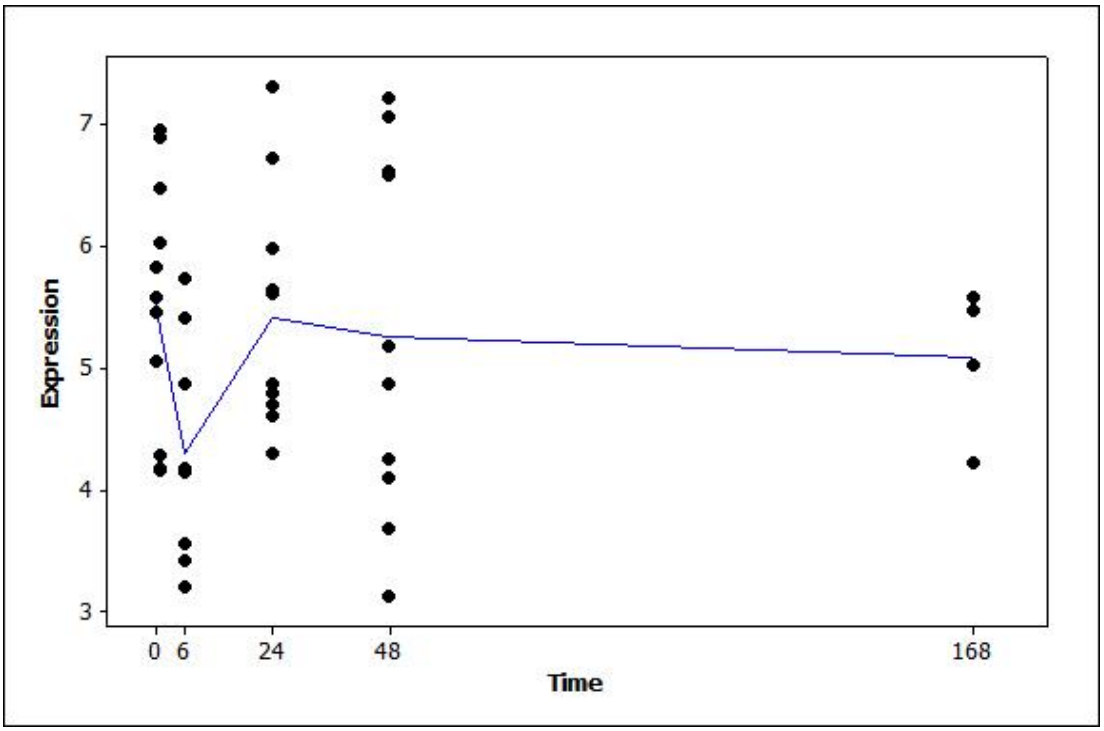


Figure 5.32 Gene expression levels across the time points from Cluster 14283

CHAPTER 6

CONCLUSION

This thesis includes the analysis of short course time series microarray gene expression data. Short course time series data are observed in the course of time (where time points may be unevenly spaced) when microarray experiments are used to study the behaviour of genes and their expression levels are investigated. There can be more than one observations per time point and the number of observations per time point may vary through the series because of the nature of the experiment.

The analysis of such data has some challenges for researchers:

- Gene expressions across time points may have a dependence structure which should not be ignored during the analyses.
- The probe measure which represents the relative expression level of an individual gene may have more than one sampling points (replicates) over time. Therefore, the measurements obtained over time belong to the gene creating a dependent sequence of measurements.
- The number of time points is very few (generally less than or equal to 8) compared to classical time series data which usually have more than 50 observations for a convenient time series modeling. As the number of time points in the short time series may vary, the number of replicates per time point may vary as well. The less the number of replicates the harder to fit models because estimation of the variance components gets harder or impossible. Sometimes, the data is unbalanced that cause another challenge for researchers.
- Unevenly spaced time points indicate that the amount of time between consecutive measurements is not the same across all time points. The time elapsed after an observation may vary. This is unusual in classical time series approach.

- Short time series modelling is methodologically and computationally very extensive and demanding. The data may contain changing number of replicates per time point. Moreover, there may be factors such as cell type as well as treatment, one or both of which might have more than two levels. Time as a source of variation in microarray experiments is a continuous independent variable rather than a qualitative factor most of the time. Biologists are very keen on finding out whether a treatment has an acute or chronic effect on the subject of interest.
- Subject-wise or gene-wise inference over the short time profile is required for researchers which drastically increases the number of simultaneous hypothesis tests.
- As an alternative method, Limma, was found to be more appropriate for experimental design models containin only qualitative factors rather than continuous independent variables.

6.1 The contributions of this thesis to the literature

- Proposing the use of LME method at every individual time interval of a short time series microarray data provides modeling and testing a short time series profile and subjectwise testing.
- Comparing it with Limma the competing and most widely used alternative method, namely Limma. It was shown through a comprehensive simulation study that proposed methodology outperformed Limma in true positive rate, accuracy, specificity, positive predictive value, negative predictive value, false discovery rate and F1 value performance parameters in overall results.
- Providing a detailed statistical inference for the complicated structure of the data of interest which requires a powerful and comprehensive model to handle.
- Providing subjectwise analysis, time trend fitting and many other requirements for short time series microarray data.
- Fitting the random effects together with the fixed effects produce more unbiased results compared to when they are fit only as fixed effects. Existence of the random effects compensates the shrinkage of the fixed effects towards to the mean value. This also helps to avoid any over and under estimation of parameter estimates that occur by chance. This is where Limma method fell behind and produced false discoveries.

- Handling repeated measures, unbalanced data and missing values via LME for short time series microarray data.
- Producing more appropriate results when the data of interest has hierarchical levels with many factors such as cell type, treatment, short time series and the clusters that contain the probe sets with similar expression profiles.
- Providing great flexibility whenever additional factors or terms such as covariates or categorical factors are to be added to the LME model.
- Detecting acute and chronic effects of a treatment via modeling the short time series microarray gene expression profiles.
- Handling the differing time lags by incorporating time as an independent variable into the model by LME as well as testing the time change effect.
- Increasing predictive capabilities and F1 value in short time series microarray gene expression profile analysis for both factors and independent variables such as continuous time parameter.
- Proposing a two stage clustering algorithm for the detection of time series gene expression profiles.
- Proposing a real like data simulation algorithm for short time series microarray gene expression data with differing number of replicates per time point as well as incorporating cell type, exposure and other required effects.
- Proposing a very comprehensive simulation and short time series microarray gene expression data fitting R code. For the simulation part, the code first generates realistic gene expression profiles. The profiles can be modified such as changing the number of cell types, treatment groups and time points. The code then generates realistic initial set of data by making use of a mixture normal distribution. The parameters of the initial data can be adjusted as well. According to the profiles that are created the code can simulate the short time series, replications in accordance with the structure of the profile. Realistic noise and experimental factors are also incorporated to the simulated data. For the real data fitting part, researchers who would like to utilize the code should only rename the column names of their dataset and run the code. On the overall, the code is very user friendly, easy to use and allows customizations. The code is free to access and can be downloaded from www.metu.edu.tr/~oilk/LME_code.zip.

According to the results and the findings of overall analyses, two stage clustering of the microarray time series data and then the application of LME in each time interval of probe sets are very plausible alternatives for subjectwise testing in short time series microarray data. The methodology proposed in this study was also compared to the competing and most widely used alternative method, namely Limma. It was shown through a comprehensive simulation study that proposed methodology outperformed Limma in true positive rate, accuracy, specificity, positive predictive value, negative predictive value, false discovery rate and F1 value performance parameters in overall results. Besides, other outperforming properties of the proposed methodology along with LME can be itemized as follows:

- The complicated structure of the data of interest requires a powerful and comprehensive model to handle for a detailed statistical inference. Common regression models lack handling random coefficients together with fixed effects, subjectwise analysis, time trend fitting and many other requirements.
- When the data of interest has hierarchical levels with many factors such as cell type, treatment, short time series and the clusters that contain the probe sets with similar expression profiles, LME produce more appropriate results.
- Fitting the random effects together with the fixed effects produce more unbiased results compared to when they are fit only as fixed effects. Existence of the random effects compensates the shrinkage of the fixed effects towards to the mean value. This also helps to avoid any over and under estimation of parameter estimates that occur by chance. This is where Limma method fell behind and produced false discoveries.
- Differing time lags can be easily handled by incorporating time as an independent variable into the model by LME.
- Predictive capabilities and F1 value of LME is superior for both factors and independent variables such as continuous time parameter.

There are two drawbacks of the proposed methodology. The first one is the computational complexity of the model and the second one is that sometimes the iterative estimation procedure in LME may not converge. However, changing the optimization method or the number of initial simulations for EM estimation helped almost all of the time. Limma, on the other hand, is computationally much easier to handle.

In addition to above remarks, it is worth mentioning that changing the parameters of the clustering algorithm such as the number of k-means, the cutoff value in the hierarchical clustering and the number of filtered genes may affect the performance of the model as well as the convergence performance of the LME iterations.

6.2 Future Work

There are still so many problems to solve in the analysis of short time series gene expression data. Foremost ones can be listed as follows:

- Gene expression profile clustering requires optimization such that the number of profile groups is a very difficult to identify.
- Splitting gene expression profiles according to their expression levels requires further research. A gene expression profile is a vector of random measurements and may have very different expression levels at any time point.
- Creation of initial dataset and generating realistic simulation data is another study to proceed. The short time series is based upon the initial data and it is very crucial to the gene expression profiles over time. In this study EM algorithm was used to model initial data column from a realistic dataset and short time series gene expression profiles were generated from a discrete distribution. There is still so much work to be done on this topic. Especially, incorporation of noise factors requires a comprehensive study. In addition, existence of changing number of replicates per time points makes the simulation of expression profiles more difficult.

REFERENCES

Androulakis, I.P., Yang, E. & Almon, R.R. (2007). Analysis of Time-Series Gene Expression Data: Methods, Challenges, and Opportunities. *Annual Review of Biomedical Engineering,* **9,** 3.1-3.24.

Baayen, R.H., Davidson, D.J. & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language,* **59,** 390-412.

Bar-Joseph, Z., Gerber, G., Gifford, D., Jaakkola, T. & Simon, I. (2002). A new approach to analyzing gene expression time series data. in *Proceedings of The Sixth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*.

Bar-Joseph, Z., Gerber, G., Gifford, D.K., Jaakkola, T.S. & Simon, I. (2003). Continuous Representations of Time Series Gene Expression Data. *Journal of Computational Biology,* **10,** 341-356.

Bartlett, M.S. (1966). *An Introduction to Stochastic Processes with Special Reference to Methods and Applications*, 2nd Edition ed. Cambridge: Cambridge University Press.

Becker, R.A., Chambers, J.M. & Wilks, A.R. (1988). *The new S language*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Bilban, M., Buehler, L.K., Head, S., Desoye, G. & Quaranta, V. (2002). Normalizing DNA Microarray Data. *Current Issues in Molecular Biology,* **4,** 57-64.

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

Broët, P., Lewin, A., Richardson, S., Dalmasso, C. & Magdelenat, H. (2004). A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics,* **20,** 2562-2571.

Broët, P., Richardson, S. & Radvanyi, F. (2002). Bayesian Hierarchical Model for Identifying Changes in Gene Expression from Microarray Experiments. *Journal of Computational Biology,* **9,** 671-683.

Brown, H. & Prescott, R. (2006). *Applied Mixed Models in Medicine*, 2nd ed.: John Wiley & Sons, Ltd.

Celeux, G., Martin, O. & Lavergne, C. (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling,* **5,** 243-267.

Chen, J. (2005). Identification of significant periodic genes in microarray gene expression data. *BMC Bioinformatics,* **6**:286.

Chen, T.-S., Tsai, T.-H., Chen, Y.-T., Lin, C.-C., Chen, R.-C., Li, S.-Y. & Chenl, H.-Y. (2005). A Combined K-Means And Hierarchical Clustering Method For Improving The Clustering Efficiency Of Microarray. *Proceedings of 2005 International Symposium on Intelligent Signal Processing and Communication Systems***,** 405-408.

Claverie, J.-M. (1999). Computational Methods for the Identification of Differential and Coordinated Gene Expression. *Human Molecular Genetics,* **8,** 1821-1832.

Dempster, A.P., Rubin, D.B. & Tsutakawa, R.K. (1981). Estimation in Covariance Components Models. *Journal of the American Statistical Association,* **76,** 341-353.

Diggle, P.J., Liang, K. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

Do, K.-A., Müller, P. & Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Applied Statistics,* **54,** 627-644.

Dudoit, S. & Gentleman, R. (2002). Cluster Analysis in DNA Microarray Experiments. in *Bioconductor Short Course, Winter 2002*.

Efron, B., Tibshirani, R., Storey, J.D. & Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association,* **96,** 1151-1160.

Eisen, M.B., Spellman, P., Brown, P. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences,* **95,** 14863-14868.

Eng, K.H., Keleş, S. & Wahba, G. (2008). A Linear Mixed Effects Clustering Model for Multi-species Time Course Gene Expression Data. Madison, WI: University of Wisconsin.

Ernst, J., Nau, G.J. & Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics,* **21,** i159-i168.

Fisher, R.A. (1929). Tests of Significance in Harmonic Analysis. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character,* **125,** 54-59.

Forgy, E.W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics,* **21,** 768-769.

Furlotte, N.A., Kang, H.M., Ye, C. & Eskin, E. (2011). Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity. *Bioinformatics,* **27,** i288–i294.

Gentleman, R., Carey, V., Huber, W. & Hahne, F. (2009). genefilter: Methods for filtering genes from microarray experiments. in *R package version 1.30.0*.

Hartigan, J.A. & Wong, M.A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* **28,** 100-108.

Harville, D. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association,* **72,** 320-338.

He, F. & Zeng, A.-P. (2006). In search of functional association from time-series microarray data based on the change trend and level of gene expression. *BMC Bioinformatics,* **7**:69.

He, W. (2004). A spline function approach for detecting differentially expressed genes in microarray data analysis. *Bioinformatics,* **20,** 2954-2963.

Hong, F. & Li, H. (2006). Functional Hierarchical Models for Identifying Genes with Different Time-Course Expression Profiles. *Biometrics,* **62,** 534-544.

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

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

Jung, S.-H. & Jang, W. (2006). How accurately can we control the FDR in analyzing microarray data? *Bioinformatics,* **22,** 1730-1736.

Kerr, M.K., Martin, M. & Churchill, G.A. (2000). Analysis of Variance for Gene Expression Microarray Data. *Journal of Computational Biology,* **7,** 819-837.

Kim, J. & Kim, J.H. (2007). Difference-based clustering of short time-course microarray data with replicates. *BMC Bioinformatics,* **8**:253.

Korpela, M. (2006). Analysis of changes in gene expression time series data. in *Department of Computer Science and Engineering* Espoo: Helsinki University of Technology.

Kuenzel, L. (2010). Gene clustering methods for time series microarray data. in *Biochemistry 218*.

Laird, N.M. & Ware, J.H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics,* **38,** 963-974.

Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika,* **73,** 13-22.

Lim, W.K., Wang, K., Lefebvre, C. & Califano, A. (2007). Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics,* **23,** i282-i288.

Lindstrom, M.J. & Bates, D.M. (1988). Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association,* **83,** 1014-1022.

Liu, H., Tarima, S., Borders, A.S., Getchell, T.V., Getchell, M.L. & Stromberg, A.J. (2005). Quadratic regression analysis for gene discovery and pattern recognition for non-cyclic short time-course microarray experiments. *BMC Bioinformatics,* **6**:106.

Lloyd., S.P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory,* **28,** 129-137.

Long, J.S. & Ervin, L.H. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician,* **54,** 217-224.

Lou, X.J., Schena, M., Horrigan, F.T., Lawn, R.M. & Davis, R.W. (2001). Expression monitoring using cDNA microarrays. A general protocol. *Methods in Molecular Biology,* **175,** 323-340.

Macqueen, J.B. (1967). Some methods of classification and analysis of multivariate observations. in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*.

Matsui, S., Yamanaka, T., Barlogie, B., Shaughnessy Jr, J.D. & Crowley, J. (2008). Clustering of significant genes in prognostic studies with microarrays: Application to a clinical study for multiple myeloma. *Statistics in Medicine,* **27,** 1106-1120.

Mclachlan, G.J., Bean, R.W. & Jones, L.B.-T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics,* **22,** 1608-1615.

Möller-Levet, C.S., Klawonn, F., Cho, K.-H., Yin, H. & Wolkenhauer, O. (2005). Clustering of unevenly sampled gene expression time-series data. *Fuzzy Sets and Systems,* **152,** 49-66.

Moser, R.J., Reverter, A., Kerr, C.A., Beh, K.J. & Lehnert, S.A. (2004). A mixed-model approach for the analysis of cDNA microarray gene expression data from extreme-performing pigs after infection with Actinobacillus pleuropneumoniae. *Journal of Animal Science,* **82,** 1261-1271.

Mutarelli, M., Cicatiello, L., Ferraro, L., Grober, O.M., Ravo, M., Facchiano, A.M., Angelini, C. & Weisz, A. (2007). Time-course analysis of genome-wide gene expression data from hormone-responsive human breast cancer cells. *BMC Bioinformatics,* **9:12**.

Najarian, K., Zaheri, M., Rad, A.A., Najarian, S. & Dargahi, J. (2004). A novel Mixture Model Method for identification of differentially expressed genes from DNA microarray data. *BMC Bioinformatics,* **5**:201.

Nykter, M., Aho, T., Ahdesmäki, M., Ruusuvuori, P., Lehmussola, A. & Yli-Harja, O. (2006). Simulation of microarray data with realistic characteristics. *BMC Bioinformatics,* **7**:349.

Nymark, P., Lindholm, P.M., Korpela, M.V., Lahti, L., Ruosaari, S., Kaski, S., Hollmén, J., Anttila, S., Kinnula, V.L. & Knuutila, S. (2007). Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC Genomics,* **8**:62.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics,* **18,** 546-554.

Park, T., Yi, S.-G., Lee, S., Lee, S.Y., Yoo, D.-H., Ahn, J.-I. & Lee, Y.-S. (2003). Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics,* **19,** 694–703.

Passos, V.L., Tan, F.E.S. & Berger, M.P.F. (2011). Cost-efficiency considerations in the choice of a microarray platform for time course experimental designs. *Computational Statistics and Data Analysis,* **55,** 944-954.

Patterson, H.D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika,* **58,** 545-554.

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D. & The R Development Core Team (2011). nlme: Linear and Nonlinear Mixed Effects Models. in *R package version 3.1-102*.

Pinheiro, J.C. & Bates, D.M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing,* **6,** 289-296.

Qin, H., Feng, T., Harding, S., A., Tsai, C.-J. & Zhang, S. (2008). An efficient method to identify differentially expressed genes in microarray experiments. *Bioinformatics,* **24,** 1583-1589.

Qiu, W., He, W., Wang, X. & Lazarus, R. (2008). A Marginal Mixture Model for Selecting Differentially Expressed Genes across Two Types of Tissue Samples. *The International Journal of Biostatistics,* **4**.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramoni, M.F., Sebastiani, P. & Kohane, I.S. (2002). Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences,* **99,** 9121-9126.

Sasik, R., Iranfar, N., Hwa, T. & Loomis, W.F. (2002). Extracting transcriptional events from temporal gene expression patterns during Dictyostelium development. *Bioinformatics,* **18,** 61-66.

Schliep, A., Schönhuth, A. & Steinhoff, C. (2003). Using hidden Markov models to analyze gene expression time course data. *Bioinformatics,* **19,** i255-i263.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H. & Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Research,* **28,** e47.

Shah, M. & Corbeil, J. (2011). A General Framework for Analyzing Data from Two Short Time-Series Microarray Experiments. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on,* **8,** 14-26.

Shedden, K., Chen, W., Kuick, R., Ghosh, D., Macdonald, J., Cho, K., Giordano, T., Gruber, S., Fearon, E., Taylor, J. & Hanash, S. (2005). Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. *BMC Bioinformatics,* **6**:26.

Sinha, A. & Markatou, M. (2011). A Platform for Processing Expression of Short Time Series (PESTS). *BMC Bioinformatics,* **12**:13.

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**.

Tai, Y.C. & Speed, T.P. (2005). Statistical Analysis of Microarray Time Course Data. in *DNA Microarrays* Ed. U. Nuber. Taylor and Francis.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. (1999). Systematic determination of genetic network architecture. *Nature Genetics,* **22,** 281-285.

Tou, J.T. & Gonzalez, R.C. (1974). *Pattern Recognition Principles*. Addison-Wesley.

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

Tusher, V.G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences,* **98,** 5116-5121.

Venables, W.N. & Ripley, B.D. (2002). *Modern Applied Statistics with S*. Springer.

Wang, L., Chen, X., Wolfinger, R.D., Franklin, J.L., Coffey, R.J. & Zhang, B. (2009). A Unified Mixed Effects Model for Gene Set Analysis of Time Course Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology,* **8**:47.

Wang, L., Zhang, B., Wolfinger, R.D. & Chen, X. (2008). An Integrated Approach for the Analysis of Biological Pathways using Mixed Models. *PLoS Genet,* **4**(7):e1000115.

Watson, M. (2006). CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics,* **7**:509.

Wernisch, L., Kendall, S.L., Soneji, S., Wietzorrek, A., Parish, T., Hinds, J., Butcher, P.D. & Stoker, N.G. (2003). Analysis of whole-genome microarray replicates using mixed models. *Bioinformatics,* **19,** 53-61.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. & Paules, R.S. (2001). Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models. *Journal of Computational Biology,* **8,** 625-637.

Wu, Z., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F. & Spencer, F. (2004). A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association,* **99,** 909-917.

Xu, Y., Olman, V. & Xu, D. (2002). Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics,* **18,** 536-545.

Yang, Y.H. & Speed, T.P. (2003). Design and analysis of comparative microarray experiments. in *Statistical Analysis of Gene Expression Microarray Data* Ed. T. Speed. Chapman & Hall/CRC Press.

Yue, H., Eastman, P.S., Wang, B.B., Minor, J., Doctolero, M.H., Nuttall, R.L., Stack, R., Becker, J.W., Montgomery, J.R., Vainer, M. & Johnston, R. (2001). An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Research,* **29**.

Zeng, Y. & Garcia-Frias, J. (2006). A novel HMM-based clustering algorithm for the analysis of gene expression time-course data. *Computational Statistics & Data Analysis,* **50,** 2472-2494.

APPENDIX

A. TABLES OF PERFORMANCE MEASURES IN SIMULATIONS

Table A.1 Expected performance measures of cell type parameter for 500 probe sets

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
2	1.5	0.05	LME	0.725	0.001	0.959	0.999	0.993	0.955	0.007	0.838
			Limma	0.999	0.075	0.935	0.925	0.683	1.000	0.317	0.811
2	1.5	0.10	LME	0.952	0.001	0.992	0.999	0.992	0.991	0.008	0.971
			Limma	0.999	0.121	0.896	0.879	0.572	1.000	0.428	0.727
2	1.5	0.20	LME	0.986	0.002	0.997	0.998	0.990	0.998	0.010	0.988
			Limma	0.999	0.205	0.823	0.795	0.439	1.000	0.561	0.610
2	1.5	0.30	LME	0.993	0.003	0.997	0.997	0.983	0.999	0.017	0.988
			Limma	0.999	0.285	0.754	0.715	0.361	1.000	0.639	0.530
2	1.5	0.40	LME	0.998	0.003	0.997	0.997	0.979	1.000	0.021	0.988
			Limma	0.999	0.366	0.684	0.634	0.305	1.000	0.695	0.467
2	2	0.05	LME	0.693	0.001	0.953	0.999	0.992	0.950	0.008	0.816
			Limma	0.999	0.075	0.935	0.925	0.687	1.000	0.313	0.814
2	2	0.10	LME	0.953	0.001	0.992	0.999	0.993	0.991	0.007	0.972
			Limma	0.999	0.123	0.894	0.877	0.573	1.000	0.427	0.728
2	2	0.20	LME	0.985	0.001	0.997	0.999	0.992	0.998	0.008	0.988
			Limma	0.999	0.207	0.822	0.793	0.443	1.000	0.557	0.614
2	2	0.30	LME	0.993	0.002	0.997	0.998	0.985	0.999	0.015	0.989
			Limma	0.999	0.284	0.756	0.716	0.366	1.000	0.634	0.536
2	2	0.40	LME	0.998	0.003	0.997	0.997	0.979	1.000	0.021	0.989
			Limma	0.999	0.367	0.685	0.633	0.309	1.000	0.691	0.472
2	3	0.05	LME	0.697	0.000	0.954	1.000	0.996	0.951	0.004	0.820
			Limma	1.000	0.075	0.936	0.925	0.684	1.000	0.316	0.812
2	3	0.10	LME	0.962	0.001	0.994	0.999	0.996	0.993	0.004	0.979
			Limma	1.000	0.124	0.893	0.876	0.565	1.000	0.435	0.722

Table A.1 Expected performance measures of cell type parameter for 500 probe sets (continued)

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
2	3	0.20	LME	0.990	0.001	0.998	0.999	0.995	0.998	0.005	0.993
			Limma	1.000	0.209	0.820	0.791	0.436	1.000	0.564	0.607
2	3	0.30	LME	0.994	0.001	0.998	0.999	0.991	0.999	0.009	0.993
			Limma	1.000	0.289	0.751	0.711	0.358	1.000	0.642	0.527
2	3	0.40	LME	0.999	0.002	0.998	0.998	0.986	1.000	0.014	0.993
			Limma	1.000	0.373	0.679	0.627	0.301	1.000	0.699	0.463
3	1.5	0.05	LME	0.734	0.001	0.960	0.999	0.993	0.957	0.007	0.845
			Limma	0.999	0.064	0.945	0.936	0.716	1.000	0.284	0.834
3	1.5	0.10	LME	0.955	0.001	0.992	0.999	0.993	0.992	0.007	0.974
			Limma	0.999	0.106	0.909	0.894	0.602	1.000	0.398	0.751
3	1.5	0.20	LME	0.981	0.001	0.996	0.999	0.991	0.997	0.009	0.986
			Limma	0.999	0.183	0.842	0.817	0.467	1.000	0.533	0.637
3	1.5	0.30	LME	0.992	0.003	0.996	0.997	0.981	0.999	0.019	0.986
			Limma	0.999	0.256	0.779	0.744	0.384	1.000	0.616	0.555
3	1.5	0.40	LME	0.998	0.004	0.996	0.996	0.975	1.000	0.025	0.987
			Limma	1.000	0.334	0.711	0.666	0.323	1.000	0.677	0.488
3	2	0.05	LME	0.707	0.001	0.955	0.999	0.994	0.952	0.006	0.827
			Limma	1.000	0.066	0.943	0.934	0.713	1.000	0.287	0.832
3	2	0.10	LME	0.952	0.001	0.992	0.999	0.994	0.992	0.006	0.973
			Limma	1.000	0.108	0.907	0.892	0.601	1.000	0.399	0.751
3	2	0.20	LME	0.983	0.001	0.997	0.999	0.993	0.997	0.007	0.988
			Limma	1.000	0.185	0.841	0.815	0.468	1.000	0.532	0.637
3	2	0.30	LME	0.991	0.002	0.997	0.998	0.986	0.999	0.014	0.988
			Limma	1.000	0.259	0.777	0.741	0.385	1.000	0.615	0.556
3	2	0.40	LME	0.998	0.003	0.997	0.997	0.979	1.000	0.021	0.988
			Limma	1.000	0.336	0.711	0.664	0.325	1.000	0.675	0.491
3	3	0.05	LME	0.723	0.000	0.957	1.000	0.996	0.954	0.004	0.837
			Limma	1.000	0.067	0.943	0.933	0.708	1.000	0.292	0.829
3	3	0.10	LME	0.960	0.001	0.993	0.999	0.996	0.993	0.004	0.978
			Limma	1.000	0.108	0.907	0.892	0.600	1.000	0.400	0.750
3	3	0.20	LME	0.988	0.001	0.998	0.999	0.995	0.998	0.005	0.992
			Limma	1.000	0.184	0.841	0.816	0.467	1.000	0.533	0.637
3	3	0.30	LME	0.994	0.001	0.998	0.999	0.990	0.999	0.010	0.992
			Limma	1.000	0.262	0.774	0.738	0.381	1.000	0.619	0.552
3	3	0.40	LME	0.998	0.002	0.998	0.998	0.986	1.000	0.014	0.992
			Limma	1.000	0.339	0.708	0.661	0.322	1.000	0.678	0.487

Table A.2 Expected performance measures of cell type parameter for 1000 probe sets

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
2	1.5	0.05	LME	0.695	0.000	0.956	1.000	0.998	0.952	0.002	0.819
			Limma	1.000	0.074	0.936	0.926	0.684	1.000	0.316	0.813
2	1.5	0.10	LME	0.976	0.000	0.996	1.000	0.998	0.996	0.002	0.987
			Limma	1.000	0.122	0.895	0.878	0.568	1.000	0.432	0.724
2	1.5	0.20	LME	0.996	0.000	0.999	1.000	0.998	0.999	0.002	0.997
			Limma	1.000	0.205	0.823	0.795	0.437	1.000	0.563	0.608
2	1.5	0.30	LME	0.998	0.001	0.999	0.999	0.996	1.000	0.004	0.997
			Limma	1.000	0.287	0.752	0.713	0.356	1.000	0.644	0.525
2	1.5	0.40	LME	1.000	0.001	0.999	0.999	0.994	1.000	0.006	0.997
			Limma	1.000	0.370	0.681	0.630	0.300	1.000	0.700	0.462
2	2	0.05	LME	0.624	0.000	0.945	1.000	0.999	0.941	0.001	0.768
			Limma	1.000	0.076	0.935	0.924	0.682	1.000	0.318	0.811
2	2	0.10	LME	0.969	0.000	0.995	1.000	0.999	0.995	0.001	0.984
			Limma	1.000	0.124	0.893	0.876	0.567	1.000	0.433	0.724
2	2	0.20	LME	0.996	0.000	0.999	1.000	0.999	0.999	0.001	0.998
			Limma	1.000	0.209	0.820	0.791	0.437	1.000	0.563	0.608
2	2	0.30	LME	0.998	0.000	0.999	1.000	0.998	1.000	0.002	0.998
			Limma	1.000	0.289	0.751	0.711	0.360	1.000	0.640	0.529
2	2	0.40	LME	1.000	0.001	0.999	0.999	0.996	1.000	0.004	0.998
			Limma	1.000	0.370	0.682	0.630	0.305	1.000	0.695	0.468
2	3	0.05	LME	0.598	0.000	0.942	1.000	1.000	0.937	0.000	0.748
			Limma	1.000	0.073	0.937	0.927	0.691	1.000	0.309	0.817
2	3	0.10	LME	0.976	0.000	0.996	1.000	0.999	0.996	0.001	0.988
			Limma	1.000	0.121	0.896	0.879	0.575	1.000	0.425	0.730
2	3	0.20	LME	0.997	0.000	1.000	1.000	0.999	1.000	0.001	0.998
			Limma	1.000	0.206	0.823	0.794	0.442	1.000	0.558	0.613
2	3	0.30	LME	0.998	0.000	1.000	1.000	0.999	1.000	0.001	0.998
			Limma	1.000	0.291	0.750	0.709	0.359	1.000	0.641	0.529
2	3	0.40	LME	1.000	0.000	1.000	1.000	0.997	1.000	0.003	0.998
			Limma	1.000	0.372	0.680	0.628	0.304	1.000	0.696	0.467
3	1.5	0.05	LME	0.720	0.000	0.960	1.000	0.999	0.956	0.001	0.837
			Limma	1.000	0.064	0.944	0.936	0.712	1.000	0.288	0.831
3	1.5	0.10	LME	0.977	0.000	0.996	1.000	0.999	0.996	0.001	0.988
			Limma	1.000	0.104	0.910	0.896	0.604	1.000	0.396	0.753
3	1.5	0.20	LME	0.995	0.000	0.999	1.000	0.999	0.999	0.001	0.997
			Limma	1.000	0.180	0.845	0.820	0.469	1.000	0.531	0.638

Table A.2 Expected performance measures of cell type parameter for 1000 probe sets (continued)

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
3	1.5	0.30	LME	0.997	0.000	0.999	1.000	0.997	1.000	0.003	0.997
			Limma	1.000	0.256	0.779	0.744	0.382	1.000	0.618	0.553
3	1.5	0.40	LME	1.000	0.001	0.999	0.999	0.995	1.000	0.005	0.997
			Limma	1.000	0.334	0.712	0.666	0.321	1.000	0.679	0.487
3	2	0.05	LME	0.666	0.000	0.952	1.000	0.999	0.948	0.001	0.799
			Limma	1.000	0.065	0.944	0.935	0.713	1.000	0.287	0.832
3	2	0.10	LME	0.981	0.000	0.997	1.000	0.999	0.997	0.001	0.990
			Limma	1.000	0.105	0.909	0.895	0.604	1.000	0.396	0.753
3	2	0.20	LME	0.995	0.000	0.999	1.000	0.999	0.999	0.001	0.997
			Limma	1.000	0.182	0.843	0.818	0.469	1.000	0.531	0.638
3	2	0.30	LME	0.997	0.000	0.999	1.000	0.997	1.000	0.003	0.997
			Limma	1.000	0.257	0.779	0.743	0.384	1.000	0.616	0.555
3	2	0.40	LME	1.000	0.001	0.999	0.999	0.994	1.000	0.006	0.997
			Limma	1.000	0.335	0.711	0.665	0.323	1.000	0.677	0.489
3	3	0.05	LME	0.605	0.000	0.942	1.000	1.000	0.938	0.000	0.754
			Limma	1.000	0.064	0.945	0.936	0.721	1.000	0.279	0.838
3	3	0.10	LME	0.976	0.000	0.996	1.000	1.000	0.996	0.000	0.988
			Limma	1.000	0.105	0.909	0.895	0.610	1.000	0.390	0.758
3	3	0.20	LME	0.998	0.000	1.000	1.000	1.000	1.000	0.000	0.999
			Limma	1.000	0.183	0.843	0.817	0.474	1.000	0.526	0.643
3	3	0.30	LME	0.999	0.000	1.000	1.000	0.999	1.000	0.001	0.999
			Limma	1.000	0.259	0.778	0.741	0.388	1.000	0.612	0.559
3	3	0.40	LME	1.000	0.000	1.000	1.000	0.998	1.000	0.002	0.999
			Limma	1.000	0.337	0.711	0.663	0.328	1.000	0.672	0.494

Table A.3 Expected performance measures of exposure parameter for 500 probe sets

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
2	1.5	0.05	LME	0.970	0.000	0.996	1.000	0.998	0.995	0.002	0.984
			Limma	0.999	0.075	0.935	0.925	0.683	1.000	0.317	0.812
2	1.5	0.10	LME	0.979	0.001	0.996	0.999	0.990	0.996	0.010	0.984
			Limma	0.999	0.122	0.895	0.878	0.571	1.000	0.429	0.727
2	1.5	0.20	LME	0.994	0.004	0.996	0.996	0.975	0.999	0.025	0.984
			Limma	1.000	0.206	0.822	0.794	0.439	1.000	0.561	0.611
2	1.5	0.30	LME	0.999	0.005	0.996	0.995	0.971	1.000	0.029	0.984
			Limma	1.000	0.290	0.750	0.710	0.358	1.000	0.642	0.527
2	1.5	0.40	LME	0.999	0.005	0.996	0.995	0.970	1.000	0.030	0.984
			Limma	1.000	0.371	0.680	0.629	0.303	1.000	0.697	0.465
2	2	0.05	LME	0.971	0.000	0.996	1.000	0.997	0.995	0.003	0.984
			Limma	0.999	0.076	0.934	0.924	0.685	1.000	0.315	0.812
2	2	0.10	LME	0.979	0.001	0.996	0.999	0.990	0.997	0.010	0.985
			Limma	0.999	0.124	0.893	0.876	0.571	1.000	0.429	0.727
2	2	0.20	LME	0.995	0.004	0.996	0.996	0.975	0.999	0.025	0.985
			Limma	0.999	0.211	0.819	0.789	0.439	1.000	0.561	0.610
2	2	0.30	LME	0.999	0.005	0.996	0.995	0.971	1.000	0.029	0.985
			Limma	0.999	0.289	0.752	0.711	0.363	1.000	0.637	0.533
2	2	0.40	LME	0.999	0.005	0.996	0.995	0.971	1.000	0.029	0.985
			Limma	0.999	0.373	0.679	0.627	0.306	1.000	0.694	0.468
2	3	0.05	LME	0.982	0.000	0.997	1.000	0.997	0.997	0.003	0.990
			Limma	0.999	0.076	0.935	0.924	0.686	1.000	0.314	0.814
2	3	0.10	LME	0.987	0.001	0.997	0.999	0.993	0.998	0.007	0.990
			Limma	0.999	0.125	0.893	0.875	0.570	1.000	0.430	0.726
2	3	0.20	LME	0.996	0.002	0.997	0.998	0.984	0.999	0.016	0.990
			Limma	0.999	0.209	0.821	0.791	0.441	1.000	0.559	0.612
2	3	0.30	LME	0.999	0.003	0.997	0.997	0.981	1.000	0.019	0.990
			Limma	0.999	0.289	0.752	0.711	0.363	1.000	0.637	0.532
2	3	0.40	LME	0.999	0.003	0.997	0.997	0.981	1.000	0.019	0.990
			Limma	0.999	0.369	0.682	0.631	0.308	1.000	0.692	0.470
3	1.5	0.05	LME	0.974	0.001	0.996	0.999	0.996	0.996	0.004	0.985
			Limma	0.999	0.066	0.943	0.934	0.712	1.000	0.288	0.832
3	1.5	0.10	LME	0.981	0.002	0.996	0.998	0.990	0.997	0.010	0.985
			Limma	0.999	0.107	0.908	0.893	0.604	1.000	0.396	0.753
3	1.5	0.20	LME	0.995	0.004	0.996	0.996	0.976	0.999	0.024	0.986
			Limma	0.999	0.186	0.840	0.814	0.468	1.000	0.532	0.638

Table A.3 Expected performance measures of exposure parameter for 500 probe sets (continued)

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
3	1.5	0.30	LME	0.999	0.004	0.996	0.996	0.973	1.000	0.027	0.986
			Limma	1.000	0.262	0.774	0.738	0.384	1.000	0.616	0.555
3	1.5	0.40	LME	0.999	0.005	0.996	0.995	0.973	1.000	0.027	0.986
			Limma	1.000	0.342	0.706	0.658	0.323	1.000	0.677	0.488
3	2	0.05	LME	0.975	0.000	0.996	1.000	0.997	0.996	0.003	0.986
			Limma	0.999	0.065	0.944	0.935	0.717	1.000	0.283	0.835
3	2	0.10	LME	0.981	0.001	0.996	0.999	0.991	0.997	0.009	0.986
			Limma	0.999	0.106	0.909	0.894	0.606	1.000	0.394	0.755
3	2	0.20	LME	0.995	0.004	0.996	0.996	0.978	0.999	0.022	0.987
			Limma	0.999	0.179	0.846	0.821	0.476	1.000	0.524	0.645
3	2	0.30	LME	0.999	0.004	0.996	0.996	0.975	1.000	0.025	0.987
			Limma	0.999	0.253	0.783	0.747	0.392	1.000	0.608	0.563
3	2	0.40	LME	0.999	0.004	0.996	0.996	0.975	1.000	0.025	0.987
			Limma	1.000	0.331	0.715	0.669	0.330	1.000	0.670	0.496
3	3	0.05	LME	0.985	0.000	0.998	1.000	0.998	0.998	0.002	0.991
			Limma	1.000	0.067	0.942	0.933	0.707	1.000	0.293	0.828
3	3	0.10	LME	0.988	0.001	0.998	0.999	0.995	0.998	0.005	0.991
			Limma	1.000	0.109	0.906	0.891	0.598	1.000	0.402	0.748
3	3	0.20	LME	0.997	0.002	0.998	0.998	0.986	0.999	0.014	0.992
			Limma	1.000	0.187	0.839	0.813	0.464	1.000	0.536	0.634
3	3	0.30	LME	0.999	0.003	0.998	0.997	0.984	1.000	0.016	0.992
			Limma	1.000	0.263	0.773	0.737	0.381	1.000	0.619	0.551
3	3	0.40	LME	0.999	0.003	0.998	0.997	0.984	1.000	0.016	0.992
			Limma	1.000	0.343	0.704	0.657	0.320	1.000	0.680	0.484

Table A.4 Expected performance measures of exposure parameter for 1000 probe sets

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
2	1.5	0.05	LME	0.992	0.000	0.999	1.000	1.000	0.999	0.000	0.996
			Limma	1.000	0.074	0.936	0.926	0.685	1.000	0.315	0.813
2	1.5	0.10	LME	0.995	0.000	0.999	1.000	0.997	0.999	0.003	0.996
			Limma	1.000	0.122	0.894	0.878	0.569	1.000	0.431	0.725
2	1.5	0.20	LME	0.999	0.001	0.999	0.999	0.993	1.000	0.007	0.996
			Limma	1.000	0.208	0.820	0.792	0.436	1.000	0.564	0.607
2	1.5	0.30	LME	1.000	0.001	0.999	0.999	0.992	1.000	0.008	0.996
			Limma	1.000	0.288	0.752	0.712	0.359	1.000	0.641	0.528
2	1.5	0.40	LME	1.000	0.001	0.999	0.999	0.992	1.000	0.008	0.996
			Limma	1.000	0.371	0.681	0.629	0.303	1.000	0.697	0.465
2	2	0.05	LME	0.994	0.000	0.999	1.000	0.999	0.999	0.001	0.997
			Limma	1.000	0.074	0.936	0.926	0.690	1.000	0.310	0.816
2	2	0.10	LME	0.995	0.000	0.999	1.000	0.998	0.999	0.002	0.997
			Limma	1.000	0.121	0.896	0.879	0.576	1.000	0.424	0.731
2	2	0.20	LME	0.999	0.001	0.999	0.999	0.994	1.000	0.006	0.997
			Limma	1.000	0.208	0.821	0.792	0.442	1.000	0.558	0.613
2	2	0.30	LME	1.000	0.001	0.999	0.999	0.994	1.000	0.006	0.997
			Limma	1.000	0.291	0.750	0.709	0.361	1.000	0.639	0.531
2	2	0.40	LME	1.000	0.001	0.999	0.999	0.994	1.000	0.006	0.997
			Limma	1.000	0.376	0.677	0.624	0.304	1.000	0.696	0.466
2	3	0.05	LME	0.996	0.000	0.999	1.000	1.000	0.999	0.000	0.998
			Limma	1.000	0.074	0.936	0.926	0.685	1.000	0.315	0.813
2	3	0.10	LME	0.997	0.000	0.999	1.000	0.999	1.000	0.001	0.998
			Limma	1.000	0.122	0.895	0.878	0.571	1.000	0.429	0.727
2	3	0.20	LME	0.999	0.001	0.999	0.999	0.997	1.000	0.003	0.998
			Limma	1.000	0.206	0.822	0.794	0.439	1.000	0.561	0.610
2	3	0.30	LME	1.000	0.001	0.999	0.999	0.996	1.000	0.004	0.998
			Limma	1.000	0.287	0.752	0.713	0.360	1.000	0.640	0.529
2	3	0.40	LME	1.000	0.001	0.999	0.999	0.996	1.000	0.004	0.998
			Limma	1.000	0.370	0.682	0.630	0.304	1.000	0.696	0.466
3	1.5	0.05	LME	0.992	0.000	0.999	1.000	1.000	0.999	0.000	0.996
			Limma	1.000	0.064	0.945	0.936	0.716	1.000	0.284	0.834
3	1.5	0.10	LME	0.994	0.000	0.999	1.000	0.998	0.999	0.002	0.996
			Limma	1.000	0.106	0.909	0.894	0.604	1.000	0.396	0.753
3	1.5	0.20	LME	0.999	0.001	0.999	0.999	0.993	1.000	0.007	0.996
			Limma	1.000	0.183	0.843	0.817	0.470	1.000	0.530	0.639

Table A.4 Expected performance measures of exposure parameter for 1000 probe sets (continued)

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
3	1.5	0.30	LME	1.000	0.001	0.999	0.999	0.992	1.000	0.008	0.996
			Limma	1.000	0.260	0.776	0.740	0.384	1.000	0.616	0.555
3	1.5	0.40	LME	1.000	0.001	0.999	0.999	0.992	1.000	0.008	0.996
			Limma	1.000	0.340	0.707	0.660	0.322	1.000	0.678	0.487
3	2	0.05	LME	0.992	0.000	0.999	1.000	1.000	0.999	0.000	0.996
			Limma	1.000	0.065	0.944	0.935	0.714	1.000	0.286	0.833
3	2	0.10	LME	0.994	0.000	0.999	1.000	0.998	0.999	0.002	0.996
			Limma	1.000	0.107	0.908	0.893	0.602	1.000	0.398	0.752
3	2	0.20	LME	0.998	0.001	0.999	0.999	0.994	1.000	0.006	0.996
			Limma	1.000	0.183	0.842	0.817	0.469	1.000	0.531	0.638
3	2	0.30	LME	1.000	0.001	0.999	0.999	0.992	1.000	0.008	0.996
			Limma	1.000	0.259	0.777	0.741	0.384	1.000	0.616	0.555
3	2	0.40	LME	1.000	0.001	0.999	0.999	0.992	1.000	0.008	0.996
			Limma	1.000	0.338	0.709	0.662	0.324	1.000	0.676	0.489
3	3	0.05	LME	0.996	0.000	0.999	1.000	1.000	0.999	0.000	0.998
			Limma	1.000	0.063	0.946	0.937	0.720	1.000	0.280	0.837
3	3	0.10	LME	0.996	0.000	0.999	1.000	0.999	0.999	0.001	0.998
			Limma	1.000	0.105	0.910	0.895	0.607	1.000	0.393	0.755
3	3	0.20	LME	0.999	0.001	0.999	0.999	0.997	1.000	0.003	0.998
			Limma	1.000	0.180	0.845	0.820	0.473	1.000	0.527	0.643
3	3	0.30	LME	1.000	0.001	0.999	0.999	0.996	1.000	0.004	0.998
			Limma	1.000	0.255	0.780	0.745	0.388	1.000	0.612	0.559
3	3	0.40	LME	1.000	0.001	0.999	0.999	0.996	1.000	0.004	0.998
			Limma	1.000	0.335	0.711	0.665	0.325	1.000	0.675	0.490

Table A.5 Expected performance measures of time parameter for 500 probe sets

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
2	1.5	0.05	LME	0.976	0.001	0.997	0.999	0.989	0.998	0.011	0.982
			Limma	0.668	0.061	0.916	0.939	0.505	0.968	0.495	0.575
2	1.5	0.10	LME	0.991	0.002	0.997	0.998	0.975	0.999	0.025	0.983
			Limma	0.668	0.061	0.916	0.939	0.505	0.968	0.495	0.575
2	1.5	0.20	LME	0.998	0.003	0.997	0.997	0.968	1.000	0.032	0.983
			Limma	0.668	0.061	0.916	0.939	0.505	0.968	0.495	0.575
2	1.5	0.30	LME	0.999	0.003	0.997	0.997	0.967	1.000	0.033	0.983
			Limma	0.668	0.061	0.916	0.939	0.505	0.968	0.495	0.575
2	1.5	0.40	LME	0.999	0.003	0.997	0.997	0.967	1.000	0.033	0.983
			Limma	0.668	0.061	0.916	0.939	0.505	0.968	0.495	0.575
2	2	0.05	LME	0.978	0.001	0.997	0.999	0.985	0.998	0.015	0.982
			Limma	0.670	0.062	0.917	0.938	0.487	0.970	0.513	0.564
2	2	0.10	LME	0.993	0.003	0.997	0.997	0.971	0.999	0.029	0.982
			Limma	0.670	0.062	0.917	0.938	0.487	0.970	0.513	0.564
2	2	0.20	LME	0.999	0.003	0.997	0.997	0.966	1.000	0.034	0.982
			Limma	0.670	0.062	0.917	0.938	0.487	0.970	0.513	0.564
2	2	0.30	LME	0.999	0.003	0.997	0.997	0.965	1.000	0.035	0.982
			Limma	0.670	0.062	0.917	0.938	0.487	0.970	0.513	0.564
2	2	0.40	LME	0.999	0.003	0.997	0.997	0.965	1.000	0.035	0.982
			Limma	0.670	0.062	0.917	0.938	0.487	0.970	0.513	0.564
2	3	0.05	LME	0.986	0.001	0.998	0.999	0.990	0.999	0.010	0.988
			Limma	0.662	0.061	0.916	0.939	0.497	0.968	0.503	0.568
2	3	0.10	LME	0.995	0.002	0.998	0.998	0.981	0.999	0.019	0.988
			Limma	0.662	0.061	0.916	0.939	0.497	0.968	0.503	0.568
2	3	0.20	LME	0.999	0.002	0.998	0.998	0.977	1.000	0.023	0.988
			Limma	0.662	0.061	0.916	0.939	0.497	0.968	0.503	0.568
2	3	0.30	LME	1.000	0.002	0.998	0.998	0.976	1.000	0.024	0.988
			Limma	0.663	0.061	0.916	0.939	0.497	0.968	0.503	0.568
2	3	0.40	LME	1.000	0.002	0.998	0.998	0.976	1.000	0.024	0.988
			Limma	0.663	0.061	0.916	0.939	0.496	0.968	0.504	0.567
3	1.5	0.05	LME	0.970	0.001	0.996	0.999	0.985	0.997	0.015	0.977
			Limma	0.664	0.062	0.915	0.938	0.493	0.969	0.507	0.566
3	1.5	0.10	LME	0.991	0.003	0.996	0.997	0.965	0.999	0.035	0.978
			Limma	0.664	0.062	0.915	0.938	0.493	0.969	0.507	0.566
3	1.5	0.20	LME	0.998	0.004	0.996	0.996	0.959	1.000	0.041	0.978
			Limma	0.664	0.062	0.915	0.938	0.493	0.969	0.507	0.566

Table A.5 Expected performance measures of time parameter for 500 probe sets (continued)

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
3	1.5	0.30	LME	1.000	0.004	0.996	0.996	0.957	1.000	0.043	0.978
			Limma	0.664	0.062	0.915	0.938	0.493	0.969	0.507	0.566
3	1.5	0.40	LME	1.000	0.004	0.996	0.996	0.956	1.000	0.044	0.977
			Limma	0.664	0.062	0.915	0.938	0.493	0.969	0.507	0.566
3	2	0.05	LME	0.977	0.001	0.997	0.999	0.986	0.998	0.014	0.982
			Limma	0.669	0.059	0.918	0.941	0.508	0.969	0.492	0.577
3	2	0.10	LME	0.993	0.003	0.997	0.997	0.972	0.999	0.028	0.982
			Limma	0.669	0.059	0.918	0.941	0.508	0.969	0.492	0.577
3	2	0.20	LME	0.999	0.003	0.997	0.997	0.966	1.000	0.034	0.982
			Limma	0.669	0.059	0.918	0.941	0.508	0.969	0.492	0.577
3	2	0.30	LME	1.000	0.003	0.997	0.997	0.965	1.000	0.035	0.982
			Limma	0.669	0.059	0.918	0.941	0.508	0.969	0.492	0.577
3	2	0.40	LME	1.000	0.003	0.997	0.997	0.965	1.000	0.035	0.982
			Limma	0.669	0.060	0.918	0.940	0.508	0.969	0.492	0.577
3	3	0.05	LME	0.981	0.001	0.997	0.999	0.989	0.998	0.011	0.985
			Limma	0.675	0.061	0.916	0.939	0.513	0.968	0.487	0.583
3	3	0.10	LME	0.995	0.002	0.997	0.998	0.976	0.999	0.024	0.986
			Limma	0.675	0.061	0.916	0.939	0.513	0.968	0.487	0.583
3	3	0.20	LME	0.999	0.003	0.997	0.997	0.973	1.000	0.027	0.986
			Limma	0.675	0.061	0.916	0.939	0.513	0.968	0.487	0.583
3	3	0.30	LME	1.000	0.003	0.997	0.997	0.972	1.000	0.028	0.986
			Limma	0.675	0.061	0.916	0.939	0.513	0.968	0.487	0.583
3	3	0.40	LME	1.000	0.003	0.997	0.997	0.972	1.000	0.028	0.986
			Limma	0.675	0.061	0.916	0.939	0.513	0.968	0.487	0.583

Table A.6 Expected performance measures of time parameter for 1000 probe sets

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
2	1.5	0.05	LME	0.992	0.000	0.999	1.000	0.997	0.999	0.003	0.995
			Limma	0.672	0.061	0.918	0.939	0.495	0.970	0.505	0.570
2	1.5	0.10	LME	0.998	0.001	0.999	0.999	0.992	1.000	0.008	0.995
			Limma	0.672	0.061	0.918	0.939	0.495	0.970	0.505	0.570
2	1.5	0.20	LME	1.000	0.001	0.999	0.999	0.989	1.000	0.011	0.995
			Limma	0.672	0.061	0.918	0.939	0.495	0.970	0.505	0.570
2	1.5	0.30	LME	1.000	0.001	0.999	0.999	0.989	1.000	0.011	0.994
			Limma	0.672	0.061	0.918	0.939	0.495	0.970	0.505	0.570
2	1.5	0.40	LME	1.000	0.001	0.999	0.999	0.989	1.000	0.011	0.994
			Limma	0.672	0.061	0.917	0.939	0.495	0.970	0.505	0.570
2	2	0.05	LME	0.992	0.000	0.999	1.000	0.998	0.999	0.002	0.995
			Limma	0.668	0.061	0.916	0.939	0.502	0.969	0.498	0.574
2	2	0.10	LME	0.998	0.001	0.999	0.999	0.991	1.000	0.009	0.995
			Limma	0.668	0.061	0.916	0.939	0.502	0.969	0.498	0.574
2	2	0.20	LME	1.000	0.001	0.999	0.999	0.990	1.000	0.010	0.995
			Limma	0.668	0.061	0.916	0.939	0.502	0.969	0.498	0.573
2	2	0.30	LME	1.000	0.001	0.999	0.999	0.989	1.000	0.011	0.995
			Limma	0.668	0.061	0.916	0.939	0.502	0.969	0.498	0.573
2	2	0.40	LME	1.000	0.001	0.999	0.999	0.989	1.000	0.011	0.995
			Limma	0.668	0.061	0.916	0.939	0.502	0.969	0.498	0.573
2	3	0.05	LME	0.995	0.000	0.999	1.000	0.998	1.000	0.002	0.997
			Limma	0.670	0.060	0.917	0.940	0.502	0.969	0.498	0.574
2	3	0.10	LME	0.998	0.000	0.999	1.000	0.995	1.000	0.005	0.997
			Limma	0.670	0.060	0.917	0.940	0.502	0.969	0.498	0.574
2	3	0.20	LME	1.000	0.001	0.999	0.999	0.993	1.000	0.007	0.996
			Limma	0.670	0.060	0.917	0.940	0.502	0.969	0.498	0.574
2	3	0.30	LME	1.000	0.001	0.999	0.999	0.993	1.000	0.007	0.996
			Limma	0.670	0.060	0.917	0.940	0.502	0.969	0.498	0.574
2	3	0.40	LME	1.000	0.001	0.999	0.999	0.992	1.000	0.008	0.996
			Limma	0.670	0.060	0.917	0.940	0.502	0.969	0.498	0.574
3	1.5	0.05	LME	0.989	0.000	0.999	1.000	0.997	0.999	0.003	0.993
			Limma	0.666	0.061	0.916	0.939	0.502	0.968	0.498	0.573
3	1.5	0.10	LME	0.997	0.001	0.999	0.999	0.989	1.000	0.011	0.993
			Limma	0.666	0.061	0.916	0.939	0.502	0.968	0.498	0.573
3	1.5	0.20	LME	0.999	0.001	0.999	0.999	0.987	1.000	0.013	0.993
			Limma	0.666	0.061	0.916	0.939	0.502	0.968	0.498	0.572

Table A.6 Expected performance measures of time parameter for 1000 probe sets (continued)

Replicates	Foldchange	p-value-cutoff	Method	E(TPR) (Power)	E(FPR) (Type I Error)	E(ACC)	E(SPC)	E(PPV)	E(NPV)	E(FDR)	F1value
3	1.5	0.30	LME	0.999	0.001	0.999	0.999	0.986	1.000	0.014	0.993
			Limma	0.666	0.061	0.916	0.939	0.502	0.968	0.498	0.572
3	1.5	0.40	LME	1.000	0.001	0.999	0.999	0.985	1.000	0.015	0.993
			Limma	0.666	0.061	0.916	0.939	0.502	0.968	0.498	0.572
3	2	0.05	LME	0.990	0.000	0.999	1.000	0.996	0.999	0.004	0.993
			Limma	0.674	0.060	0.918	0.940	0.505	0.970	0.495	0.578
3	2	0.10	LME	0.997	0.001	0.999	0.999	0.989	1.000	0.011	0.993
			Limma	0.674	0.060	0.918	0.940	0.505	0.970	0.495	0.578
3	2	0.20	LME	0.999	0.001	0.999	0.999	0.987	1.000	0.013	0.993
			Limma	0.674	0.060	0.918	0.940	0.505	0.970	0.495	0.577
3	2	0.30	LME	1.000	0.001	0.999	0.999	0.986	1.000	0.014	0.993
			Limma	0.674	0.060	0.918	0.940	0.505	0.970	0.495	0.577
3	2	0.40	LME	1.000	0.001	0.999	0.999	0.986	1.000	0.014	0.993
			Limma	0.674	0.060	0.918	0.940	0.505	0.970	0.495	0.577
3	3	0.05	LME	0.994	0.000	0.999	1.000	0.997	0.999	0.003	0.996
			Limma	0.667	0.061	0.916	0.939	0.498	0.969	0.502	0.570
3	3	0.10	LME	0.998	0.001	0.999	0.999	0.993	1.000	0.007	0.996
			Limma	0.667	0.061	0.916	0.939	0.498	0.969	0.502	0.570
3	3	0.20	LME	0.999	0.001	0.999	0.999	0.992	1.000	0.008	0.995
			Limma	0.667	0.061	0.916	0.939	0.498	0.969	0.502	0.570
3	3	0.30	LME	1.000	0.001	0.999	0.999	0.990	1.000	0.010	0.995
			Limma	0.667	0.061	0.916	0.939	0.497	0.969	0.503	0.570
3	3	0.40	LME	1.000	0.001	0.999	0.999	0.990	1.000	0.010	0.995
			Limma	0.667	0.061	0.916	0.939	0.497	0.969	0.503	0.570

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Erkan, İbrahim

Nationality: Turkish (TC)

Date and Place of Birth: 12 December 1980, İzmir

Marital Status: Married

Phone: +90 533 763 29 25

Fax: +90 312 210 29 59

email: ierkantr@yahoo.com

EDUCATION

Degree	Institution	Year of Graduation
PhD	METU Statistics	2011
BS	Ege University Statistics	2003
High School	İzmir Çınarlı A.T.H.S. Electr.	1998

WORK EXPERIENCE

Year	Place	Enrollment
2010 – Present	Ankara Development Agency	Development Expert
2007 – 2010	İnova Yazılım Ltd. Şti.	Expert Statistician
2004 – 2007	METU Statistics Department	Research Assistant

FOREIGN LANGUAGES

Fluent English

PUBLICATIONS

Jensen, S.T., Erkan, I., Arnardottir, E.S. and Small, D.S. (2009). Bayesian testing of many hypotheses X many genes: a study of sleep apnea. *Annals of Applied Statistics,* 3:1080-1101.