Mathematical Statistics
Stockholm University

Application of the concept of False Discovery Rate on predicted cancer outcome with microarrays

Sally Salih

Examensarbete 2006:1

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Examensarbete **2006:1**,
http://www.math.su.se/matstat

Application of the concept of False Discovery Rate on predicted cancer outcome with microarrays

Sally Salih*

February 2006

Abstract

There have been a lot of studies dealing with DNA microarrays; most of the studies not only are very optimistic but the researchers even claim that their works result in good predicted prognoses. Unfortunately, most of these claims are not correct; the reason is mainly due to misclassifications and to the fact that they have not properly filtered genes with high false positive in their expression datasets. The key question is that how many genes from the experiment are false positive but declared significances.

In this Master Thesis we apply the FDR concept on seven studies that have tried to predict prognosis of cancer patients using DNA microarray analysis. We would like to investigate if the genes that are used in the classification are truly differentially expressed, or at least with low FDR. The FDR will be estimated for each expression dataset from the above mentioned studies in a mixture model that contains DE and non-DE genes. The methodology, assumptions and theories are described and explained. The estimated values and produced figures are presented, discussed and assessed. By applying the FDR concept, we have been able to conclude that some researchers ignore the fact that there are false positive in the datasets. Moreover, we have detected some connections between FDR and misclassifications.

*Postal address: Dept. of Mathematical Statistics, Stockholm University, SE–106 91 Stockholm, Sweden. E-mail: sallysalih@yahoo.se. Supervisor: Åke Svensson.

Acknowledgement

I would like to thank my supervisor Yudi Pawitan for all his constructive instructions and valuable advises that I have received during the time at MEB, KI. I would like also to thank Alex Ploner for his support and help and Anna Johansson for introducing me to the MEB group. My thanks go to my supervisor Åke Svensson at SU for his advices and encouragements during the project and to my siblings for their support. I delicate this thesis to my parents and my wonderful daughter 'Ferida'.

Sally Salih
Stockholm, 2006

Contents

1 Introduction ……………………………………………………..................................1

1.1 Objectives ………………………………………………………………………………..1

1.2 The structure of the report ……………………………………………………………..1

2 Literature review …………………………………………………………………………3

2.1 Gene Expression ………………………………………………………………………..3

2.2 DNA arrays technology …………………………………………………………………4

2.3 Types of DNA arrays …………………………………………………………………...5

2.3.1 cDNA microarrays …………………………...…………………………………………………....5

2.3.2 Oligonucleotide microarrays ……………………………………………………………………....6

2.4 Data normalization ……………………………………………………………………...6

2.5 Supervised and unsupervised learning ……………………………………………….6

2.5.1 Clustering ………………………………………………………………………………………….7

2.5.2 Classification ……………………………………………………………………………………….7

3 Statistical background ………………………………………………………………...8

3.1 Statistical Tests …………………………………………………………………………8

3.1.1 Single testing ……………………………………………………………………………………….8

3.1.2 Multiple testing ……………………………………………………………………………………..8

3.2 False Discovery Rate ………………………………………………………………….10

3.2.1 FDR definition ……………………………………………………………………………………....10

4 Technical Details ……………………………………………………………………...13

4.1 Data sources …………………………………………………………………………..13

5 Methodology …………………………………………………………………………..17

5.1 Preprocessing ………………………………………………………………………….17

6 Statistical Data Analysis and Results ………………………………………………...20

6.1 FDR ……………………………………………………………………………………..20

6.2 Misclassifications ……………………………………………………………………….24

7 Previous studies ……………………………………………………………………….....25

8 Discussion ………………………………………………………………………………..26

9 Conclusion ……………………………………………………………………………….28

Appendix A ………………………………………………………………………………….29

Appendix B ………………………………………………………………………………….30

Appendix C …………………………………………………………………………………36

References ………………………………………………………………………………….39

Chapter 1

Introduction

DNA microarrays are a new area of research that has evolved since the last decade. We have gone from gene-by-gene study to tens of thousands of genes simultaneously studies. DNA microarrays have been applied to screen changing level in gene expression during essential biological processes through investigating the alteration in gene expression across collections of related cases. Microarray data analysis of gene expression measurements launches various stimulating chances and obstacles. The most commonly challenging tasks face the researchers are how they can easily and effectively detect the differentially expressed genes between control and treatment samples and how they can control the risk of accepting false genes among accepted genes. Samples can be two groups, diseases, conditions, tissues or patients. The concern of multiple testing regarding null hypothesis testing has long been a problem for researchers. In recent published papers there are many established approaches for correcting the significance levels to guarantee that the possibility of committing any false positive is the same to or below the preferred significance level. Such approaches are controlling the Family Wise Error Rate FWER and controlling the False Discovery Rate FDR. These approaches are faced by another major problem because as they control the false positive rate they become more likely to gain false negatives. An alternative method for solving this problem has introduced in a recent paper by (Pawitan et al., 2005). The uniqueness with this approach is that the method controls the probability of making any false positives in the study, while controlling the false negative as well.

1.1 Objectives

In this thesis we apply the FDR concept to the seven studies that have tried to predict prognosis of cancer patients using DNA microarray analysis. We would like to investigate if the genes that are used in the classification are all truly differentially expressed, or at least some are with low FDR, under two different conditions. The FDR will be estimated for each expression dataset from the above mentioned studies in a mixture model that contains Differentially Expressed and non-differentially expressed genes. The methodology, assumptions and theories that are used will be described and explained. Finally the estimated values and produced figures will be presented, discussed and assessed.

1.2 The structure of the report

The report is organized as follow: In chapter 1 we begin with a short introduction to DNA microarray analysis concern, an explanation of the objective of the thesis and a description of the structure of the report. Chapter 2 gives literature review about gene expression and detailed information about DNA arrays types and technology. Chapter 3 presents and explains briefly the statistical backgrounds. Chapter 4 presents an overview of technical aspects including data sources. In chapter 5 we describe the methodology, preprocessing and the assumptions that we have made regarding the expression datasets. The concept of FDR is applied to the datasets and the graphical presentation is illustrated in chapter 6. Previous study is summarized in chapter 7. The results are discussed in chapter 8. In chapter 9 we conclude with giving some recommendation for further application.

Chapter 2

Literature review

2.1 Gene Expression

According to the Watson - Crick Model, proposed in 1953 and confirmed in the succeeding decades, the DNA molecule is a double helix composed of two antiparallel strands of nucleotides; each nucleotide consists of on of four nitrogenous bases (adenine (A), thymine (T), guanine (G), or cytosine (C)), a deoxyribose sugar, and a phosphate (Leland et al., 2000). Figure 2 displays the structure of the DNA double helix and shows how the chemical bases are bind. A always binds with T and C binds G; each strand is bind to a complementary strand by the paired bases. The two chains are held together by hydrogen bonds between nitrogen bases (Dudiot, 2003).

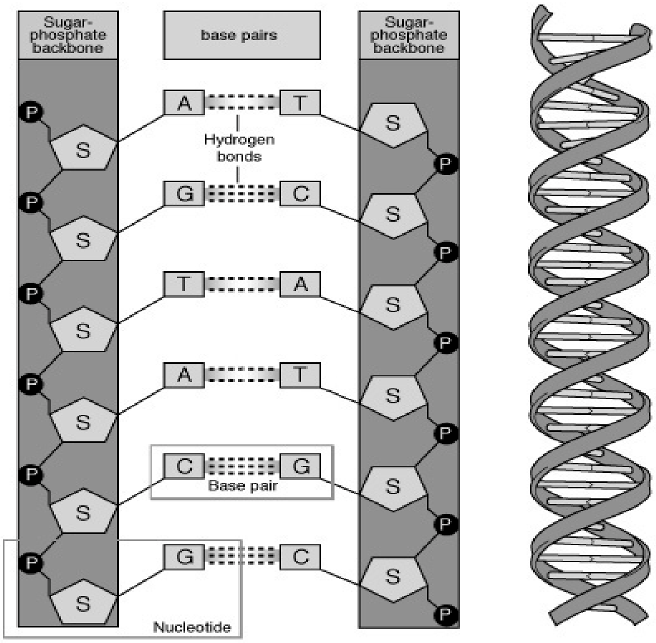


Fig.2.1 DNA - A More Detailed Description
Adapted in from Graphics Gallery

A gene is a part of DNA that contains the instruction for constructing a particular protein. When a gene in the state of being active in a cell we say that the gene is expressed (Hardin, 2005, P 4-5). When genes have their expression levels differ under different conditions, we say that the genes are differentially expressed. If various genes are being differentially expressed then diverse proteins will be manufactured in a certain cell. RNA molecule is quite similar to DNA except for some slight differences; the most important types of RNA for gene expression are mRNA and cDNA. By measuring the quantity of mRNA in the cell corresponding to some genes one can detect which genes are expressed. Gene expression occurs in two steps, the first is the transcription during which DNA is transcribed into mRNA. The second step is the translation, during which mRNA is translated to produce a protein (Carey et al., 2003, P 4). Se fig.2.2

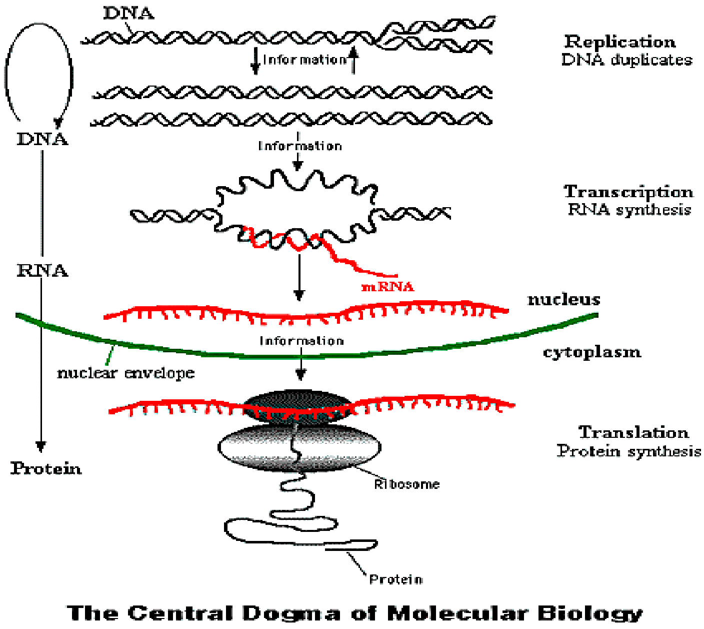


Fig.2.2: Adapted in from Graphics Gallery

2.2 DNA arrays technology

An array is an orderly arrangement of samples. Data format in DNA arrays is organized in an expression data matrix form, where the first column represent gene annotations and the first row represent sample annotations. Each cell in the gene expression matrix represents gene expression levels. Fig.2.3 demonstrates how the data is arranged if we have n genes and m samples. Microarrays can be used to compare the gene expressions across different conditions for the same sample. Normally, only an extremely tiny proportion of the genes present on any given array are identified as differentially expressed. We are very interested in monitoring these expression levels simultaneously and to be able to identify those genes that are differentially expressed. Microarray data analysis can help in predicting of some diseases' outcomes, discovering set of genes that function together, uncovering molecular similarities among subsets of samples and examining if there are any genetic differences between healthy and diseased genes. Microarray experiments have promising future to improve detection and medication for the disease, by detecting the genes affected by treatments. There are some problems accompany DNA microarrays, namely the huge amount of data that produced in the microarray experiment are quite complex and contain noisy genes with no good or clear expression.

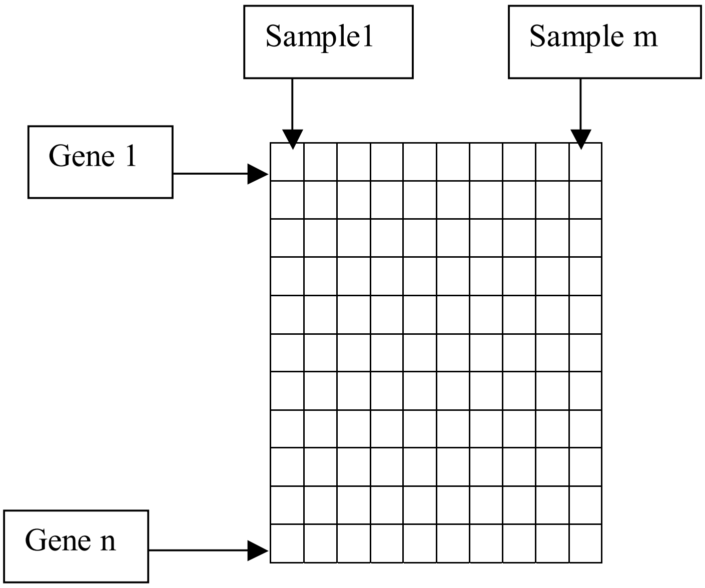


Fig.2.3 Expression data matrix

2.3 Types of DNA arrays

The DNA array technology has been developed to produce thousands of genes of interest simultaneously. In this section we will present the two most famous types of DNA arrays, the spotted cDNA arrays and Oligonucleotide microarrays.

2.3.1 cDNA microarrays

The spotted cDNA array has been generally developed and advanced by Stanford researchers. They utilized the procedure of dumping cDNA tags onto a glass slide with precise robotic printer (David et al., 2001). Identified cDNA fragments are then hybridized to the cDNA probes on a chip. In microarray experiments the reference samples are usually labeled with green dye and test samples with red dye. By reference samples we mean that the samples are single DNA strands in normal condition compared to the samples that will be tested. A green spots on the microarray indicate that the genes in reference sample are expressed at higher level than genes in test samples and vice versa. The following are the main character of this type of DNA microarrays:

1. One probe per gene.
2. Probes of varying length, which can provides more specificity if we are using longer probes.
3. Two target samples per array.
4. Flexible.
5. One microarray can contain only the genes of interest.
6. One microarray can contains less than 15,000 DNA spots.

2.3.2 Oligonucleotide microarrays

The producer of this type of microarrays is Affymetrix. There are about 16-20 probe-pairs per gene; one probe-pair is a (Perfect Match PM, MissMatch MM) pair. The perfect match is constructed to meet a tiny subseries of the gene about 25 bases long. The mismatch is a control probe, which is identical to pm except with the middle base turned over to its complement. The probes are synthesized directly onto a glass slide using photolithographic masks. Sample processing includes the production of labeled cDNA (David et al., 2001). The samples are then hybridized on the chip and the RNA expression of each of the genes is estimated by the difference in signal pm-mm averaged over the 20 probe pairs from the matching sign got after laser scanning. The following are the main character of this type of DNA microarrays:

1. Short oligonucleotide arrays.
2. DNA spots are uniform and extremely close together.
3. One target sample per array (Probes are 25-mer).
4. One microarray can contains more than 400,000 DNA spots.
5. It is user friendly.

In spite of the DNA microarrays being used, the produced microarrays can be with no trouble transferred into different commercially accessible data analysis programs.

2.4 Data normalization

Normalization is needed to ensure that differences in intensities are due to differential expression, and not some printing, hybridization, or scanning. It is curial to do this procedure prior to any statistical analysis to be able to compare gene expression results. The process identifies and removes the effects of systematic variation other than differential expression in the measurements. There are two procedures of doing the normalization, the first one is called Per-chip normalization adjusts the total or average intensity of each array to be roughly the same, which reduce minor differences in example probe preparations and hybridizations. The second procedure is per-gene normalization, which compare the results for a single gene across all the samples to reveal genes that have same expression pattern. Here we assume that the gene expression data are normalized accurately.

2.5 Supervised and unsupervised learning

There are three types of statistical issues that are related particularly with tumor classification, namely identification of new tumor classes using gene expression profiles (unsupervised learning), classification of malignancies into known classes (supervised learning), and identification of marker genes that characterize the different tumor classes (feature selection) (Speed et al., 2003, P94). The first two types are explained in the next sections.

2.5.1 Clustering

Clustering is the process of grouping data with unknown a priori class membership. The cluster analysis includes estimating the number of classes and assigning genes to these classes (Speed, T 2003, and P94). All data have to be normalized before clustering; this leads to the fact that clustering is dependent on how the data are normalized and on what data are included in the analyses. Clustering can be valuable in detecting subcategories of tumor, and revealing previously unrecognized similarities.

2.5.2 Classification

The aim of supervised learning is class prediction of different genes. The first step is to understand the basis for class prediction from a learning set and second is to find out what discriminates the different groups. This will be done by selecting the most informative genes to be used in classifications. Moreover, by having clear information about the classes the different genes belong to, we can employ the classification to perform an effective feature selection. Feature selection is performed to prevent including irrelevant features by removing variables that are noise with respect to the outcome, from the dataset. The feature selection can be performed explicitly before constructing the classifier or implicitly.

Chapter 3

Statistical background

3.1 Statistical Tests

There are many statistical techniques and procedures that are developed to deal with microarray data. The major aim of these tools is to supply us with a prepared list of good candidate genes to follow up. Here we will go through basic facts about statistical tests and introduce the common definition and theories of the False Discovery Rate.

3.1.1 Single testing

In general, in testing any single hypothesis, we specify an acceptable maximum probability of committing a Type I error. This is explained by the following scenario:

- True hypothesis

  H_0: gene is not differentially expressed

  H_1: gene is differentially expressed

- Test declaration

  H_0 rejected, which means the gene is differentially expressed

  H_0 not rejected, which implies that the gene is not differentially expressed

	Test declaration	
True hypothesis	H_0 not rejected	H_0 rejected
H_0 true	true negative	false positive
H_0 false	false negative	true positive

Table 3.1 possible outcomes from single testing.

If we choose, for example a p-value of 0.05, the value will refer to a 5 % risk of committing a false positive error. This implies that any difference in expression with a p-value less than .05 will be considered significant. When many hypotheses are tested, and each test has a specified Type I error probability, the probability that at least some Type I errors are committed increases with the number of hypotheses. This may have serious consequences if the set of conclusions must be evaluated as a whole.

3.1.2 Multiple testing

We faces by the multiplicity problem because thousands of hypotheses are tested simultaneously, which includes multiple estimation as well as testing. To elucidate the problem of multiple hypothesis testing we continue with the example from section 3.1.1, but we will be performing two or more independent statistical tests at the same p-value level (0.05). In the case of two genes if neither gene is differentially expressed in truth, then the risk of declaring that one or both of them is differentially expressed is

1-0.95*0.95 = 0.0975.

If we have 5 genes and none of them is truly differentially expressed, then the risk of committing at least one false positive error is

1-0.95*0.95*0.95*095*0.95 = 0.2262.

If we have 200 genes and none of them is truly differentially expressed, then the risk of committing at least one false positive error is approximately equal to 1. This indicates that small p-values imply significance in single testing, but in multiple hypothesis testing they do not necessarily means significance. We can distinguish between two kinds of Type 1 error rates:

Family–wise error rate (FWER)

The FWER is defined as the probability of at least one Type I error (false positive) among the genes selected as significant.

FWER = Pr(V> 0)

False discovery rate (FDR)

The FDR (Benjamini & Hochberg 1995) is the expected proportion of Type I errors (false positives) among the rejected hypotheses.

FDR = E(Q),

With $Q = \begin{cases} V/R, R > 0 \\ 0, R = 0 \end{cases}$

If V = number of false positives,

R = number of rejections of null hypotheses

A large number of controlling criteria of Type 1error rates have been proposed in recent years. The two well known approaches are Bonferroni correction and the estimation of FDR.

Bonferroni correction

It divides the preferred level by the number of the tests. Every gene test is then performed at the Bonferroni adjusted level. In spite of its simplicity, the Bonferroni correction is too conservative; by controlling the group-wise error rate each individual test is held to an unreasonably high standard. This increases the probability of a Type II error, and makes it likely significant results will fail to be detected.

Estimation of the FDR

The estimation method that we have adapted in this thesis has the property of controlling the probability of making any false positives in the study, while controlling the false negative as well. This procedure retains higher power than Bonferroni correction. In the rest of this chapter we will be describing only the estimation procedure of FDR.

3.2 False Discovery Rate

This section summaries the basic assumptions, definitions and theories of FDR that we have used to perform the estimation.

3.2.1 FDR definition

The approach of FDR was introduced by Benjamini and Hochberg 1995 and defined as the expected proportion of false positives among the declared significant results (Benjamini and Hochberg, 1995). The FDR approach has become well known and popular among researchers in these fields because of its directly useful interpretation and its simplicity in application.

We have decided to give here a summary of the basic theoretical aspects of the FDR as introduced by the developers of the OC-plus package, as we have used OC-plus package to estimate the FDR. In the paper (Pawitan et al., 2005), they focus on the common problem of comparison between two independent groups with equal variance. Thus the assumption is a two-group comparison problem with *n* arrays per group, using the standard t-test with pooled variance estimate. The expression values in log2-scale are assumed normally distributed. The standard deviation σ for each gene is assumed to be equal to one. This is corresponding to standardizing the expression measurements by their standard deviation, so the fold changes have a universal scale in standard deviation units. In reality the expression variance varies between genes, but by standardizing the variance it will be assumed that all genes have equal variance. Table 3.2 illustrates a simple two-by-two table presenting how genes are classified according to their true status and test results

	Test result		
	non-DE	DE	Total
True			
non-DE	*A*	*B*	m0
DE	*C*	*D*	m-m0
Total			m

Table 3.2 Adapted table from (Pawitan et al., 2005), which illustrates a simple two-by-two table

A is the number of non-differentially expressed genes that are correctly classified, B is the number of non-differentially expressed genes that are incorrectly classified, and similarly for *C* and *D*. In the analysis, it is assumed that the genes are independent, although the results can be expected to hold for weakly dependent genes. From the table above, the following are fined the:

The false positive rate (significance level) is B/(A+ B),

The sensitivity of the test is D/(C + D),

The FDR is B/ (B + D) and

The False Negative Rate (FNR) is C/(C + D).

The following is a list of definitions and assumptions that are used to estimate the FDR:

- FDR is the proportion of false positives among the declared differentially expressed genes.
- t-statistic is the standard two-sample t-statistic with pooled variance.
- Significant result or differentially expressed call is declared for $|t$ -statistics$| > c$. The critical value c is allowed to vary.
- Significance level α of a test is the same as the false positive rate, which is the proportion of false positives among truly non- differentially expressed genes.
- Sensitivity is the proportion of truly differentially expressed genes which are declared significant and corresponds to the power of the design or 1 minus the FNR.
- n is the sample size per group
- p_0 is the proportion of truly non- differentially expressed genes.
- $p_1 = 1 - p_0$ is the proportion of truly differentially expressed genes. This is assumed to be equally split between down-regulated and up-regulated genes, and the differential expression is assumed to be concentrated at some fold changes.
- Log-fold change is the mean difference in log2-scale and in standard deviation units.
- Distribution of the true differences shows the log-fold changes of the truly differentially expressed genes. The following scenario are used:
  log-fold changes at -1 and +1 (with equal proportions of 0.5 * p_1 each)

Under the null hypothesis of no differential expression, the t -statistic is distributed according to central t-distribution. In the presence of differential expressed genes the distribution F of the observed t -statistics is a mixture of the form

$F(t) = p_0F_0(t) + p_1F_1(t),$

$F_1(t) = 0.5\{G_1(t) + G_2(t)\},$

Where $F_0(t)$ is the central t-distribution with degrees of freedom df $= 2n-2$, and $F_1(t)$ is the distribution of the statistics for non- differentially expressed genes. $G_1(t)$ and $G_2(t)$ are non-central t-distributions with df =2n -2 and non-centrality parameters $\sqrt{n/2}\,D/\sigma$ and $-\sqrt{n/2}\,D/\sigma$, respectively. The parameter D/σ is assumed non-zero log-fold change. In the computation, D/σ is -1 or 1. Given a critical value c>0, the proportion of differentially expressed genes can be computed as

$\{F(-c) + 1 - F(c)\}$

If F is symmetric, this is equal to

$2\{1 - F(c)\}$ (1).

The significance level is computed as

$\{F_0(-c) + 1 - F_0(c)\}$

This is equivalent to

$2\{1 - F_0(c)$ (2)

Applying (1) and (2), the FDR as a function of the c is then given by:

$$FDR = \frac{\rho_0\{1 - F_0(c)\}}{1 - F(c)}$$

The sensitivity of the test is computed as

$2\{1 - F_1(c)\}$

The computation of F is immediate from the observed statistics, using the empirical distribution function. In the following calculations the methods that suggested by (Storey et al., 2002) is used to estimate p_0:

p_0 = {number of P-values > λ} / {m (1- λ)}

For a certain choice of λ, and the P-values are associated with the t- statistics. This nonparametric formula has been justified intuitively with the following argument (Pawitan et al., 2005)

- The largest P-values come from non-DE genes,
- For these genes the P-values are uniformly distributed between 0 and 1,and

  E{ Number of P-values > λ } =mπ_0 (1-λ),

Thus the estimation of p_0 looks reasonable. Simple choices of l such as 0.5 or 0.75 are often used.

Chapter 4

Technical Details

4.1 Data sources

The microarray studies of cancer prognosis that published between 995 and 2003 were reviewed in 2003 by (Ntzani et al., 2003). From the review, we have selected studies on survival-related outcomes. These studies include at least 60 patients. Various classification methods are used in the studies such as linear discriminant analysis and support vector machines. The replicates sizes are between 60 and 240. This chapter includes a summary table and important findings for each study. For further information, turn to appendix B

Cancer type	Filtration	ceiling	floor	Number of genes before filtration	Number of genes after filtration	Type of DNA
Diffuse Large-B-Cell Lymphoma	Various quality controls	16 000	20	12 196	7 399	Lymphochip
ALL[1]	Genes with max/min<3 are excluded	16 000	20	12 625	12236	Affymetrix oligonucleotide
Breast cancer	Genes with more than two-fold regulation and significance of regulation p<0.01 in more than 5 experiments were kept	16 000	20	24 481	4546	Agilent
Lung adenocarcinoma	Genes with max/min<3 & max-min<130 are excluded	16 000	20	12 600	6532	Affymetrix Gene chip
Lung adenocarcinoma	Genes with max/min<3 & max-min<104 are excluded	16 000	20	12 600	5562	Affymetrix Gene chip
Medulloblastoma	Genes with max/min<3 & max-min<100 are excluded	16 000	20	7 129	6777	Affymetrix Gene chip
Hepatocellular carcinoma	Genes with max/min<5& max-min<500 are excluded	16 000	20	7 129	4863	Affymetrix oligonucleotide

Table 4.1 A summary table of the seven studies

Diffuse large B cell lymphoma, Rosenwald et al., 2002

The authors used the gene-expression profiles of the diffuse large-B-cell lymphoma to develop a molecular predictor of survival. The median follow-up was 2.8 years overall and about 57% of the patients died during this time. The statistical that were used in this study were A two-sides t-test to identify subgroups, a permutation test to formulate training and validation groups, univariate Cox

[1] Acute lymphoblastic leukemia

model and one-sided Wald test to identify significant variables, a two-sided Wald test for the proportional-hazards model to formulate the Gene-Expression-Based Outcome Predictor, the chi-square test or Fisher's exact to calculate all p-values related to the difference between the subgroups. The study resulted in the identification of three gene expression subgroups germinal center B-cell like, activated B-cell like, and type 3 diffuse large-B-cell lymphoma. In addition to that, they claimed that they were able to detect in the germinal-center B-cell–like subgroup two common oncogenic events in diffuse large-B-cell lymphoma, bcl-2 translocation and c-rel amplification. They were convinced that patients in this subgroup had the highest five-year survival rate. They concluded that DNA microarrays could be used to construct a molecular predictor of survival after chemotherapy for diffuse large-B-cell lymphoma.

Acute lymphoblastic leukemia, Yeoh et al., 2002

In order to determine if gene expression profiles could identify the important leukemia subtypes, the authors analyzed gene expression form patients with lymphoblastic leukemia. There were no patients lost to follow-up during the treatment. The samples that were collected from 1991 to 1994, their median follow-up of the event-free-survival was 8.09. The samples that were collected from 1994 to 1998, their median follow-up of the event-free-survival was 4.52. They performed unsupervised hierarchical clustering, principal component analysis, discriminant analysis with variance and self-organizing map to identify the classes of the genes. K-nearest neighbors, support vector machine, prediction by collective likelihood of emerging patterns, an artificial neural network and weighted voting were used to build classifiers. Discriminating genes were selected by Chi-square, t-statistics, Wilkins' and Correlation-based Feature Selection (CFC). They claimed that their results had demonstrated that expression profiling could not only correctly identify the known prognostically important leukemia subtypes, but could further improve their capacity to evaluate a patient's risk of failing therapy. Moreover, they stated that the analysis resulted in the identification of a new leukemia subtype.

Breast cancer, van't Veer et al., 2002

The goal of the study was to identify a gene expression signature that predicted distant metastases in patients without tumor cells in local lymph nodes at diagnosis. In this study all patients were followed at least once a year for a period of at least 5 years. They applied agglomerative hierarchical clustering in unsupervised two-dimensional clustering. Discriminating analysis, leave-one-out cross validation procedures was used in the supervised classification. Finally, they utilized the odds ratio to examine the development of metastases. The p-values associated with the odds ratio were calculated by Fisher exact test. In their results they declared that the prognostic profile provided a strong device to tailor systemic treatment that could reduce the cost of breast cancer treatment. Beside that, they stated that the signature that revealed BRCA1 status could further enhance the diagnosis of hereditary breast cancer. In conclusion, they claimed that genes that were over expressed with a poor prognosis profile were important goal for developing new cancer drugs.

Lung adenocarcinoma, Beer et al., 2002

To identify specific genes that predict survival among patients with lung adenocarcinoma, the authors correlated gene-expression profiles with clinical outcome in a cohort of patients with in early stage of the disease. They applied t-tests to identify differences in mean expression levels and agglomerative hierarchical clustering in the unsupervised clustering. To examine if cluster membership was associated with physical and genetic characteristics of the tumors Pearson, chi-square and Fisher's exact test were used. Training and validation approach and cross-validation approach were used to determine if gene-expression profiles were associated with variability in survival times. The main result of the study was that the detection of a set of genes predicting survival in early stage lung adenocarcinoma permitted the discovery of a high-risk group.

Lung adenocarcinoma, Bhattacharjee et al., 2001

To define distinct subclasses of lung adenocarcinoma the authors applied the cluster analysis of expression data. To validate the discovery of the classes they used probabilistic model-based clustering. They constructed a supervised classifier by using k-nearest neighbor classifiers based on the signal-to-noise statistic. After that a Kaplan-Meier curves was generated for each cluster. In this study they stated that they had identified distinct lung adenocarcinoma subclasses that were reproducibly generated across different cluster methods. Furthermore, they asserted that they had discovered putative metastases of extrapulmonary origin with non-lung expression signatures among presumed lung adenocarcinomas. The conclusion from this study was that gene expression analysis could serve as a diagnostic.

Ramaswamy et al., 2003

The objective of the paper was to study the molecular nature of metastasis. In the supervised prediction they applied the following methods, the signal-to-noise statistic, signal-to-noise metric, weighted-voting classification algorithm. In Hierarchical clustering they applied a weighted centered correlation for arrays. For the selection and permutation of the genes signature associated with metastasis, they used the signal-to-noise metric and Kaplan-Meier survival analysis. Mantel-Haenszel log-rank test was used to calculate the statistical significance of differences between survival curves. Finally, they performed two-tailed t-test to determine the correlation between individual genes in the signature associated with metastasis and clinical outcome. The main finding in this study was detecting a gene expression signature which distinguished primary from metastatic adenocarcinomas. The authors confirmed this detecting by employing the expression signature to data on 279 primary solid tumors of distinct types. Moreover, they declared that solid tumors carrying the gene-expression signature were most likely to be associated with metastasis and poor clinical outcome.

Medulloblastoma, Pomeroy et al., 2002

The main reason of the study was to employ gene expression to predict the central nervous system embryonal tumor outcome. In the unsupervised clustering they applied hierarchical clustering .in the supervised learning they applied the signal-to-noise statistic, permutation of the column labels, a modification of k-nearest neighbor. The study resulted in demonstrating that medulloblastomas were molecularly distinct from other brain tumors. Moreover, they declared that they had revealed previously unrecognized evidence supporting the derivation of medulloblastomas from cerebellar granule cells. Finally, they claimed that the clinical outcome of children with medulloblastomas was very well predictable on the basis of the gene expression profiles of their tumors at diagnosis.

Hepatocellular carcinoma, Iizuka et al., 2003

The aim of the study was to construct a predictive system to predict early intrahepatic recurrence of hepatocellular carcinoma. Within one year the follow-up was at least once every three months after surgery. They developed a predictive system with the Fisher Linear Classifier. To utilize the system with the available data they adopted a cross-validation approach. They also used the chi-square, Fisher's exact test and t-test to evaluate differences in clinicological factors between recurrence and non-recurrence. They performed multivariate analysis using the stepwise logistic regression model to assess independent factors for early intrahepatic recurrence. They stated that their system predicted early intrahepatic recurrence or non-recurrence for patients with hepatocellular carcinoma much more accurately than the support vector machine based system. In addition, they claimed that this result proposed that their system possibly would operate as a new procedure for illustrating the metastatic potential of hepatocellular carcinoma.

Chapter 5

Methodology

It is crucial to have the raw datasets formatted carefully and correctly prior to more specific statistical analysis methods. Preprocessing usually consists of thresholds, filtration and transformation steps. In this chapter we describe the adjustment procedures that we have performed, and the assumptions that we have made prior to any analysis. A summary table of the preprocessing is also provided in this chapter.

5.1 Preprocessing

The following are the preprocessing operations that are used on the chosen datasets.

- **Thresholds** are the upper and lower numerical limits employed to values that can produce noise such as negative values.
- **Filtrations** applied to eliminate certain genes from the dataset that produce noise and to be able to focus on the most important changes by producing reasonable sample size. One problem with filtrations is the risk of eliminating some informative genes, who are expressed at very low levels or that are induced to a lesser extent.
- **Transformations** are now commonly applied to microarray data especially a base 2 logarithmic transform. The log transformation removes much of the relationship between the standard deviation and the mean, beside that it is easy to convert and more intuitive.

Cancer type	Event	Total # of patients	# of events	Complementary to # of events
Diffuse Large-B-Cell Lymphoma	Non-survival	240	138	102
ALL[2]	Non-relapse	233	32	201
Breast cancer	Developed metastases	97	46	51
Lung adenocarcinoma	Non-survival	86	24	62
Lung adenocarcinoma	Non-survival	62	31	31
Medulloblastoma	Non-survival	60	21	39
Hepatocellular carcinoma	Recurrence within 1 year	60	20	40

Table 5.1 A summary over preprocessing of the seven studies

[2] Acute lymphoblastic leukemia

Diffuse large B cell lymphoma, A Rosenwald et al. (2002)

The datasets have already been normalized before I start working with the project, so we have decided to avoid renormalizing the datasets. The rows and columns numbers are 7 399 and 240 respectively.

Acute lymphoblastic leukemia, Yeoh et al., 2002

The samples that we have used in the analysis are 201 CCR cases, 27 Heme Relapse cases and 5 additional relapse cases. A ceiling of 16 000 and a floor of 20 is chosen. Genes whose expression does not vary across the dataset are removed. After filtration the numbers of genes and samples are 12236 and 233 respectively.

Breast cancer, van't Veer et al., 2002

The authors have already done the ceiling and floor to the datasets. We have applied the same filtration as in the original paper and only genes with following criteria are kept. If genes with at least a twofold difference and P-value of less than 0.01 in more than five tumors. The filtration has resulted in 4 546 genes and 97 patients.

Lung adenocarcinoma, Beer et al., 2002

For the upper threshold, a ceiling of 16,000 units is chosen for this study. We also use a threshold of 20 units for low expression values below this level. We apply the same filtration as in Stefan's paper (Michiels et al., 2005). Gene expression values are subjected to variation filters that exclude genes showing minimal variation across the samples being analyzed. After filtration the numbers of cases and genes are 86 and 6 532 respectively.

Lung adenocarcinoma, Bhattacharjee et al., 2001 and Ramaswamy et al., 2003

A ceiling of 16000 and a floor of 20 is chosen. For this study, genes with little variation in expression are excluded. Genes with the following criteria are excluded using the variation filter, for example if max/min < 3 and max – min < 104. After filtration the numbers of genes and of samples are 5 562 and 62 respectively.

Medulloblastoma, Pomeroy et al., 2002

In this study dataset C, which consists of 39 medulloblastomas survivors and 21 treatment failures (non-survivors), is used. We have applied the same preprocessing as in the original paper. A ceiling of 16 00 is chosen and the floor is set to 20 units. If genes show the following criteria the variation filter will exclude them; max/min < 5 and max – min < 500. After filtration we have got 6777 genes and 60 samples

Hepatocellular carcinoma, N Iizuka, et al. (2003)

In this study the datasets consist of 33 Training sets and 27 Blinded sets. A ceiling of 16 00 units is chosen and the floor is set to 20 units. If max/min < 5 and max – min < 500 then variation filter excludes the genes with these criteria. The number of genes is 4 863 and the number of samples is 60

Chapter 6

Statistical Data Analysis and Results

Before doing any statistical analysis, we have preprocessed the datasets as described in chapter 3. Prior to estimating the FDR for each study, we have plotted some Q-Q plot to informally assess the distribution of the genes (See Appendix C). The standard reference is normal distribution. If there are no differentially expressed genes we expect the majority of the statistics to look like a sample from normal distribution. The plots have been quite helpful, as the t-statistics do look like from normal distribution except at the tails; which indicates that the genes at the tail are differentially expressed. This shows that we have some useful gene expression among the selected genes. For van't Veer et al. (2002), the plot indicates that the genes exhibit strong differential expression.

6.1 FDR

To estimate the FDR in the expression datasets that we are working with, we have chosen to present them in figures. The computation of FDR follows the same mixture model as in chapter 3.

$F(t) = p_0F_0(t) + p_1F_1(t),$

We have used EOC function in OCplus –package to compute the estimated operating characteristic curves, for FDR and sensitivity as a function of critical values. Each FDR and sensitivity is plotted against the cutoff level on the t-statistic. The FDR and sensitivity are estimated using the formulae in Chapter 3, and they are constrained to be monotone. They are estimated at the same time because we do not want to drop some truly differentially expressed genes by appointing the FDR very low. For each study we have tried to choose some suitable critical value to estimate the FDR.

Diffuse large B cell lymphoma, A Rosenwald et al. (2002)

To estimate the FDR we compare the non-survival =138 arrays with the survival = 102 arrays. The FDR is < 40% and the sensitivity is ~ 50% of critical value ~ 2 (se Figure 6.1). We observe that both the FDR and the sensitivity are moderately acceptable.

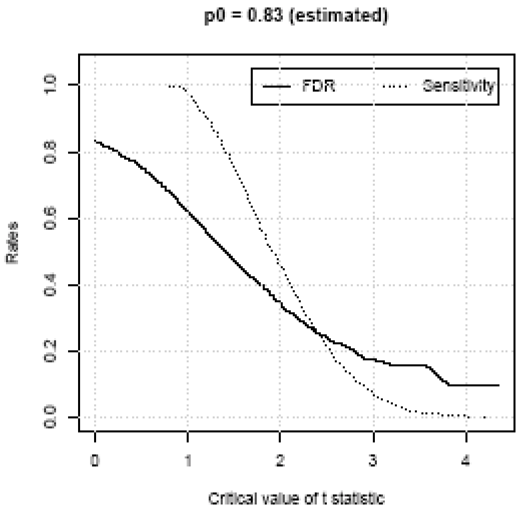


Figure 6.1 The FDR and sensitivity curves for the datasets from Rosenwald et al. (2002)

Acute lymphoblastic leukemia, Yeoh et al., 2002

To estimate the FDR we compare the non-relapse = 32 arrays with the relapse = 201 arrays. The FDR is ~ 70% and sensitivity is > 20% of critical value = 2 (se figure 6.2). In this study the FDR is quite high while the sensitivity is very low.

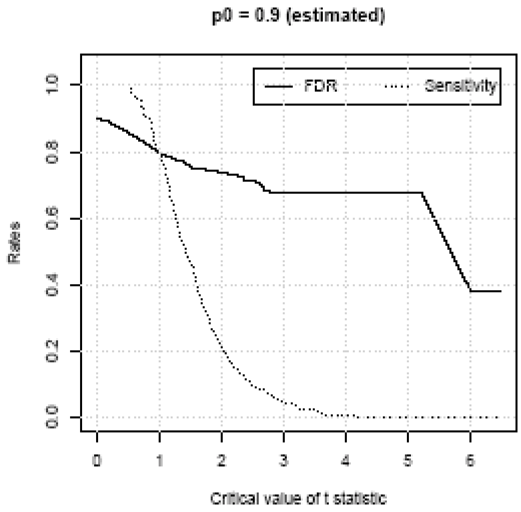


Figure 6.2 The FDR and sensitivity curves for the datasets from EJ Yeoh et al. (2002)

Breast cancer, van't Veer et al., 2002

To estimate the FDR we compare patients who developed metastases = 46 arrays with metastases-free survival = 51 arrays. The FDR is ~ 10% and sensitivity is ~ 60% of critical value = 2 (se figure 6.3). At this cutoff level, the FDR is relatively low and the sensitivity is acceptable.

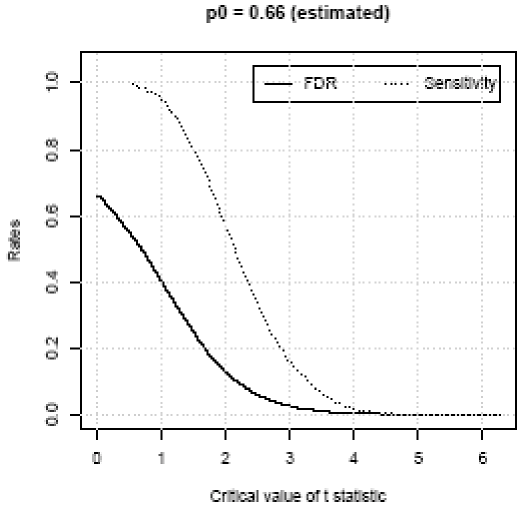


Figure 6.3 The FDR and sensitivity curves for the datasets from LJ vant't Veer et al. (2002)

Lung adenocarcinoma, Beer et al., 2002

To estimate the FDR we compare the non-survival =24 arrays with the survival = 62 arrays. The FDR is ~ 70% and sensitivity is ~ 40% of critical value = 2 (se figure 6.4). We see that the FDR is relatively high and the sensitivity is quite low.

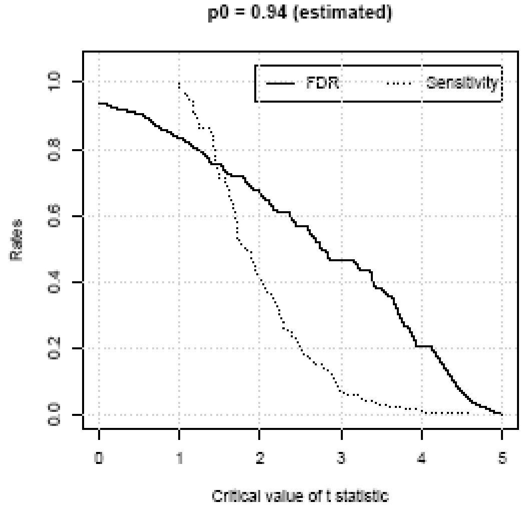


Figure 6.4 The FDR and sensitivity curves for the datasets from DG Beer, et al. (2002)

Lung adenocarcinoma, Bhattacharjee et al., 2001 and Ramaswamy et al., 2003

To estimate the FDR we compare the non-survival =31 arrays with at least 4-year survival = 31 arrays. The FDR is > 80% and sensitivity is < 20% of critical value = 2 (se fig. 6.5). We notice that the FDR is very high while the sensitivity is low.

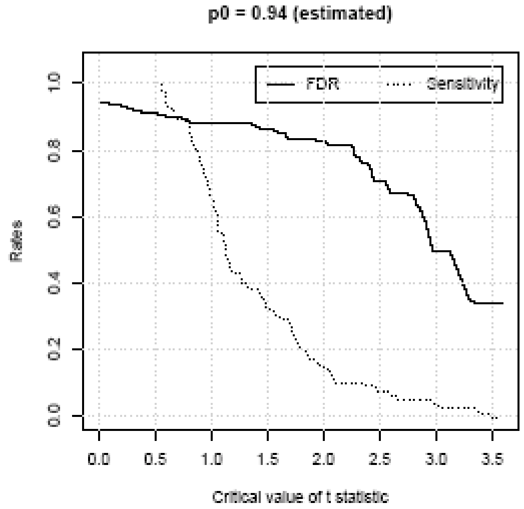


Figure 6.5 The FDR and sensitivity curves for the datasets from A Bhattacharjee, et al. (2001) and S Ramaswamy, et al. (2003)

Medulloblastoma, Pomeroy et al., 2002

To estimate the FDR we compare the non-survival =21 arrays with the survival = 39 arrays. The FDR is ~ 70% and sensitivity is > 20% of critical value = 2 (se fig. 6.6). We notice that the FDR is quite high and the sensitivity is low.

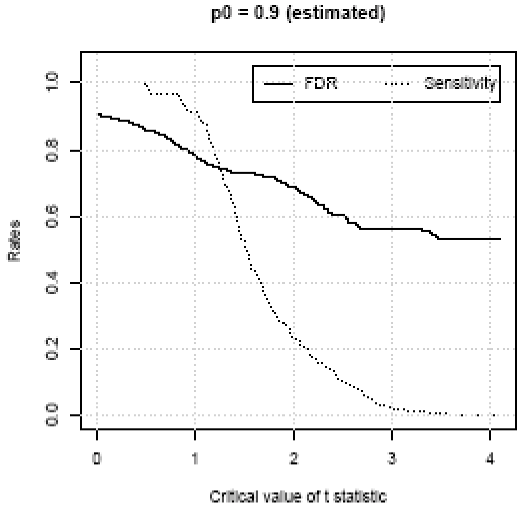


Figure 6.6 The FDR and sensitivity curves for the datasets from SL Pomeroy, et al. (2002)

Hepatocellular carcinoma, Iizuka et al., 2003

To estimate the FDR we compare patients with recurrence =20 arrays with recurrence-free survival within one year = 40 arrays. The FDR is > 30% and sensitivity is > 40% of critical value = 2 (se fig. 6.7). We see that both the FDR and the sensitivity are below the average.

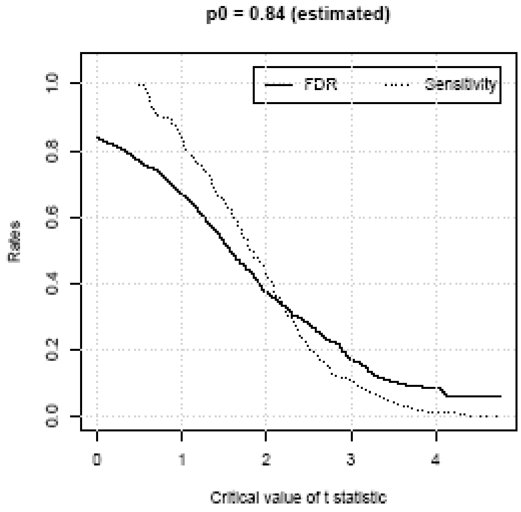


Figure 6.7 The FDR and sensitivity curves for the datasets from N Iizuka et al. (2003)

6.2 Misclassifications

To investigate the proportion of misclassifications, we have reanalyzed the expression datasets from each study and have got similar figures to the figures that produced by (Michiels et al., 2005) se fig.6.8. Generally, the assessment of the studies reveals that the studies have demonstrated relatively high misclassifications rate. The misclassifications are due to random error and biasing factors as are explained by (Michiels et al., 2005).

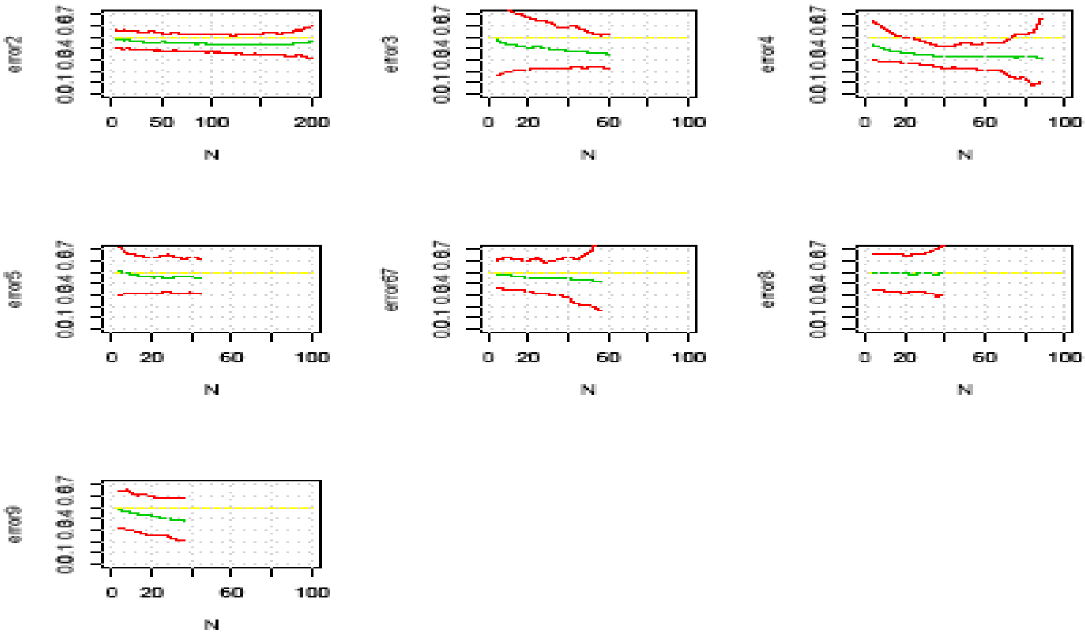


Fig 6.2 Reproduced graphs of the rate of misclassifications: The upper and the lower lines represent the 95% confidence interval, the middle line represent the misclassification rate

Chapter 7

Previous studies

Michiels et al., 2005, reanalyzed and studied these seven studies (Technical details and appendix B for further information); their approach was to identify a molecular signature in training and validation set of patients. They extended this approach by using multiple random sets to investigate the stability of the molecular signature and the proportion of misclassifications. First, they eliminated genes that showed little or no variation across samples. Second, they classified patients in the validation set by applying the nearest-centroid prediction rule; they assume in applying this method that all genes expressions have the same variability. They found that the list of genes identified as predictors of prognosis was extremely variable due to the fact that molecular signature strongly depended on the selection of patients in the training sets. The study concluded that five of the seven studies did not classify patients better than chance. Fig.7.1 shows their results from running the prediction rule.

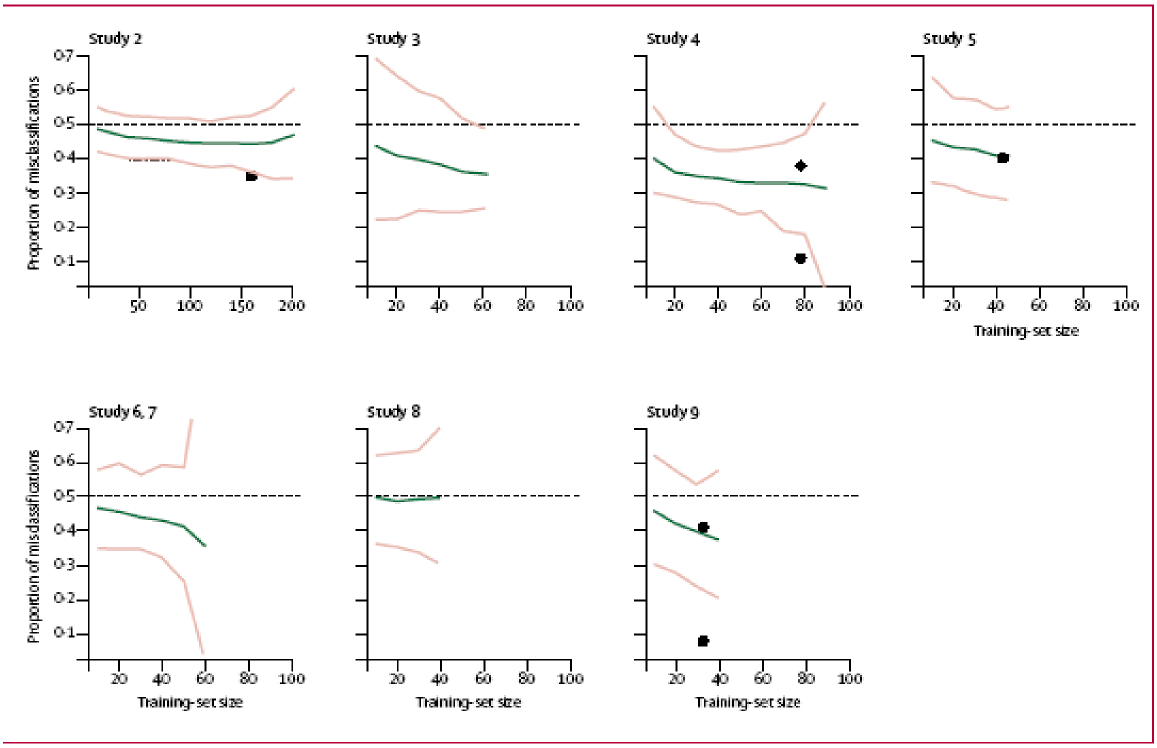


Figure 2: **Proportion of misclassifications in validation sets as a function of corresponding training-set sizes in the seven studies**[2-9]
Green lines=mean proportion of misclassifications obtained from 500 random training-validation sets as a function of the training-set size. Pale red lines=95% CIs. Dots=misclassification rates in original publications. Iizuka and colleagues[8] published two misclassification rates by two different methods on the same validation set. Diamond=second misclassification rate on a larger independent validation set[16] from the van 't Veer study.[4]

Fig. 7.1 Adapted graph form Michiels et al., 2005

Chapter 8

Discussion

In this thesis we have estimated the FDR for each study. The following are the result from computing FDR and the results from the reproduced graph of the misclassification rate.

Study	**Misclassifications' range**
Rosenwald et al (2002)	<40% - <50%
Yeoh et al, (2002)	35% - 45%
Van't Veer et al (2002)	30 % - 40 %
Beer, et al (2002)	40 % - 45 %
Bhattacharjee, et al (2001) **and** **Ramaswamy, et al (2003)**	35% - >45%
Pomeroy, et al (2002)	About 50%
Iizuka, et al (2003)	35% - 45%

Table 8.1 The table summarizes the misclassification range from the seven datasets

Study	**FDR**	**Sensitivity**
Rosenwald et al (2002).	< 40%	~50%
Yeoh et al, (2002)	~ 70%	> 20%
Van't Veer et al (2002)	~ 10%	~60%
Beer, et al (2002)	~ 70%	~40%
Bhattacharjee, et al (2001) **and** **Ramaswamy, et al (2003)**	> 80%	< 20%
Pomeroy, et al (2002)	~ 70%	> 20%
Iizuka, et al (2003)	> 30%	> 40%

Table 8.2 The table summarizes the results of estimating FDR and sensitivity from the datasets

From table 8.2, we see that in all studies there are genes with quite high FDR, except for van't Veer et al. (2002). In addition, we observe that Yeoh et al (2002), Beer et al (2002), Bhattacharjee et al (2001) and Ramaswamy et al (2003), have produced the highest FDR and the lowest sensitivity. When it comes to misclassification, we see that Rosenwald et al (2002), Beer et al (2002) and Pomeroy et al (2002) have produced the highest misclassification. Comparing the tables we can conclude that there are detectable connections between misclassifications and FDR. Examining the connections is beyond the scope of the thesis.

Chapter 9

Conclusions

This thesis evaluates the FDR in expression datasets from the seven most famous studies that have predicted prognoses of cancer patients. Before doing the project, (Michiels et al., 2005) reanalyzed the same datasets and revealed the high rate of misclassification in the performed classifications. So we suspect that there is a possibility of having high false positive genes. Applying the FDR concept on the datasets has revealed that there are genes with high FDR. Moreover it has indicated that there are some connection between misclassifications and FDR. The high FDR in these studies could have avoided if they have had applied some reliable approaches to filter data prior to data mining. More precisely, pre-selection of genes that pass an initial FDR testing may improve more specific analyses such as gene clustering.

Appendix A

Glossary

The following are the abbreviations that we used in this paper:

cDNA	complementary deoxyribonucleic acid
df	degrees of freedom
DNA	deoxyribonucleic acid molecule is a double-stranded polymer composed of four basic molecular units called nucleotides.
EOC	Empirical Operating Characteristics
FDR	False Discovery Rate
FNR	The False Negative Rate
FWER	Family Wise Error Rate
MM	mismatch.
mRNA	messenger ribonucleic acid
non-DE	non-differentially expressed
OC	Operating Characteristic
PM	stands for perfect match and
RNA	ribonucleic acid
Q-Q plot	Quantile-Quantile Plot

Appendix B

Detailed information about the data sources

Diffuse large B cell lymphoma, A Rosenwald et al. (2002)

Data Collection

The paper, the gene expression datasets as well as the supplementary information were downloaded from

http://llmpp.nih.gov/DLBCL/

The study consisted of two datasets. Both the expression datasets, which contained 12 196 genes and the clinical datasets, which included information from 240 patients were used in the original paper to formulate and test the gene expression-based outcome predictor. The clinical datasets were divided into preliminary group with 160 patients and validation group with 80 patients. In this study the authors used the gene-expression profiles of the diffuse large-B-cell lymphoma to build up a molecular predictor of survival. The methods used the biopsy samples of the lymphomas that obtained from 240 patients. The DNA microarrays were used to examine the lymphomas for gene expression and then analyzed for genomic abnormalities. Hierarchical clustering was used to define distinguishing gene-expression profiles into subsets. A molecular predictor of risk was created and then checked in a validation group. Finally they compared the precision of the predictor with that of the international prognostic index.

Acute lymphoblastic leukemia, Yeoh et al., 2002

Data Collection

The raw and clinical values were obtained from

http://www.stjuderesearch.org/data/ALL1/all_datafiles.html

The authors of the original paper had divided up the files that contained expression datasets into

1- The six diagnostic groups (BCR-ABL, E2A-PBX1, Hyperdiploid >50, MLL, T-ALL, TEL-AML1)
2- Others which included (Hyperdiploid 47_50, Hypodip, Normal, Pseudodip)
3- Group which contained diagnostic samples that did not fit into one of the above groups.

   There were 233 diagnostic samples .These samples included

   201 CCR cases.

   27 Heme Relapse cases.

   5 additional relapse cases

Here were some abbreviations and explanations that were used by the authors

Subtype Name-C#	Dx Sample of patient in CCR

Subtype Name-R#	Dx Sample of patient who developed a hematologic relapse
Subtype Name-#	Dx Sample used for subgroup classification only
Subtype Name-2M#	Dx Sample of patient who later developed 2nd AML
Subtype Name-N	Dx Sample in novel group

Protocol- Protocol that patient was treated on%Outcome-

CCR	Continuous complete remission
Heme Relapse	Hematologic relapse
Other Relapse	Extramedullary relapse
2nd AML	Diagnostic samples of patients who later developed 2nd AML
Censored	Censored due to BM transplant, treated off protocol, or died in CR
NA	Not applicable, primarily because the patient was not treated on total 13, and thus is excluded

The authors analyzed 327 diagnostic bone marrow (BM) samples with Affymetrix oligonucleotide microarrays containing 12,600 probe sets to find out if gene expression profiling of leukemic cells could be used to detect known biologic ALL subsets. They used first, an unsupervised two dimensional hierarchical clustering algorithm to group genes on the basis of similarity in their pattern of expression over the samples. They utilized the same procedure to group the leukemia samples on the basis of similarities in their pattern of genes expressed. Data reduction was achieved by applying a selection of metrics to define the genes most useful in class distinction. The selection of the genes was used in supervised learning procedures to construct classifiers that could identify the specific genetic or prognostic subsets. Assessment of the performance of each model was done by leave-one-out cross validation on a randomly selected stratified training set.

Breast cancer, van't Veer et al., 2002

Data Collection

Expression and clinical datasets were downloaded from:

http://www.nature.com/index.html

&

http://www.rii.com/publications/2002/vantveer.htm

The expression datasets contained information about log10 (intensity) log10 (ratio) p-value for each sample. There were 117 patients where only 97 of them were sporadic patients. The clinical datasets included the following samples

* 1-44 are 5 years survival

* 45-79 developed distant metastases

* 80-100 from patients with BRCA. 94 and 99 were from patients with BRCA2 and the rest were from BRCA1

* 102-113 developed distant metastases

* 114-120 were 5 yr_survival

The authors chose to perform gene clustering and sample clustering independently without interfering between the two dimensions using an agglomerative hierarchical clustering algorithm. They calculated the correlation between the prognostic category and the logarithmic expression ratio across the selected samples for each individual gene in the significant genes. To estimate the significance of each correlation they used a permutation technique to create Monte-Carlo data. They found that 231 genes satisfied this criterion in the real data set. The significance for each of the 231 genes as a prognostic reporter was estimated by a metric similar to the "Fisher" statistic. They used the method of "leave-one-out" for cross validation. The odds ratio and the p-value related to the odds were calculated. In the multivariate analysis a logistic model was performed with outcome of disease as the dependent variable, and the p-value for the relevant parameter was derived from the likelihood ratio test in the model.

Lung adenocarcinoma, Beer et al., 2002

Data Collection

The expression and clinical datasets were downloaded from:

http://www.nature.com/cgi-taf/DynaPage.taf?file=/nm/journal/v8/n8/abs/nm733.html

There were 86 patients in the clinical datasets; 67 stage 1 and 19 stage III. These patients included 24 non survival patients and 62 survival patients. In this study the author associated gene expression profiles with clinical outcome in a cohort of patients with lung adenocarcinoma and detect specific genes that predict survival among patients with stage I disease. For further validation, they also showed that the risk index predicted survival in an independent cohort of stage I lung adenocarcinoma. They studied the bonds between cluster and patient and tumor characteristics. They claimed that there are relationships between cluster and stage and between cluster and differentiation. T-tests were used to detect variations in mean gene expression levels between comparison groups. Moreover, hierarchical clustering was performed. Pearson, $\chi 2$ and Fisher's exact tests were used to evaluate whether cluster membership was associated with physical and genetic characteristics of the tumors. A leave-one-out cross-validation procedure in which tumors was used to detect genes that were univariately related to survival. The risk index was described as a linear combination of the gene-expression values for the top genes.

Lung adenocarcinoma, Bhattacharjee et al., 2001

Data Collection

The expression and clinical datasets were downloaded from

http://www.broad.mit.edu/cgibin/cancer/publications/pub_paper.cgi?mode=view&paper_id=80

The clinical datasets in this study consisted of 62 patients and was with the following criteria; 31 patients were four years survival and 31 of the patients were non-survival. For Dataset A, they used a standard deviation threshold of 50 expression units to pick the 3,312 mainly variable transcript sequences. For Dataset B, 52 pairs of replicates were used to reveal the quality of the dataset, and 45 pairs were used to choose 675 features whose expression varied the almost across all samples. They used the CLUSTER and TREEVIEW programs for hierarchical clustering and viewing of both Datasets A and B. Hierarchical clustering was done following median centering and normalization. To validate the classes discovered by hierarchical clustering, they carried out probabilistic clustering on 200 bootstrap datasets that were resampled with replacements from the original number of samples in Dataset B. They constructed a supervised classifier by identifying subclasses and chose marker genes classifiers. They produced Kaplan–Meier (K-M) curves for the clusters and weighed survival within the cluster against all other samples.

Lung adenocarcinoma, Ramaswamy et al., 2003

Data Collection

The authors aim was to analyze the datasets that contained the primary and metastatic adenocarcinomas. They performed the initial primary tumor versus metastases comparison on two distinct Affymetrix oligonucleotide microarrays. They applied cross-platform mapping of genes and took into account all genes that fell into the same UniGene clusters from both microarrays to be mapped genes. They compared 64 primary adenocarcinomas to 12 metastatic adenocarcinomas. By sorting all the genes in the dataset, they were able to identify genes correlated with particular class distinctions. Graded quantities of these marker genes were then used to construct a weighted-voting classifier. They used the Cluster and TreeView software to carry out average linkage clustering. They used a weighted centered correlation for arrays to carry out clustering. They used the signal-to-noise metric to determine the individual correlation for each of the probe sets with the two primary lung tumor clusters defined through hierarchical clustering. They chose the top 21 probe sets with. They picked random sets of genes from the extremely unstable genes; they then used these gene sets to carry out independent clusterings of the primary lung adenocarcinomas, and used Kaplan–Meier survival analysis to each clustering. They performed the Mantel–Haenszel log-rank test to evaluate the statistical significance of variations between survival curves. They used two-tailed t-tests to clarify and find out the correlation between individual genes in the signature associated with metastasis and clinical outcome in each solid-tumor data set.

Medulloblastoma, Pomeroy et al., 2002

Data Collection

The papers as well as the supplementary information were downloaded from
http://www.nature.com/nature/index.html

The gene expression data sets were downloaded:
http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
The complete study consisted of
* 68 children with medulloblastomas.
* 10 young adults with malignant gliomas
* 5 children with AT/RT
* 5 with renal/external rhabdoid tumors
* 8 children with supratentorial PNETs
In this study dataset C was used which consisted of:
- 39 medulloblastomas survivors
- 21 treatment failures (non-survivors)

In this study the authors wanted to verify if the various kinds of tumors are molecularly distinguishable. They resolved the problem by constructing a classification system based on DNA microarray gene expression data. First, they applied a principal component analysis technique to reduce the high dimensional data to three inspectable dimensions corresponding to linear combinations of genes that explained most of the variance in the original data set. Then the correlation in genes with specific class distinctions were identified by sorting the genes using the array Permutation technique. They were able to develop a modification of the k-NN algorithm which predicts the class of a new data. The k-NN models were assessed by leave-one-out cross-validation.

Hepatocellular carcinoma, Iizuka et al., 2003

Data Collection

The paper, the gene expression datasets as well as the supplementary information were downloaded from
http://www.sciencedirect.com/
The study consisted of the following datasets:
training sets (n=33)
recurrence = 12 (group A)
non-recurrence = 21 (group B)
blinded sets (n=27)
recurrence = 8 (group A)
non-recurrence = 19 (group B)

The writers investigated mRNA expression profiles in tissue specimens from a training set, comprising patients with hepatocellular carcinoma, with high-density Oligonucleotide microarrays. They used a training set in a supervised learning method to build up a predictive system, containing 12 genes, with the Fisher linear classifier. They then compared the predictive ability of their system with one of a support vector machine based system on a blinded set of samples from 27 newly registered patients.

Appendix C

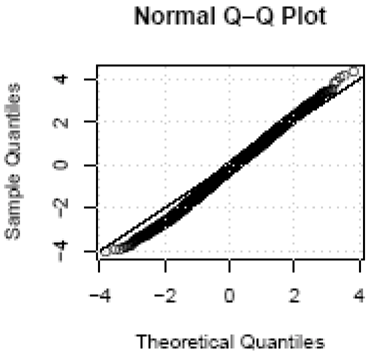
Figures

Some Q-Q plot from each study

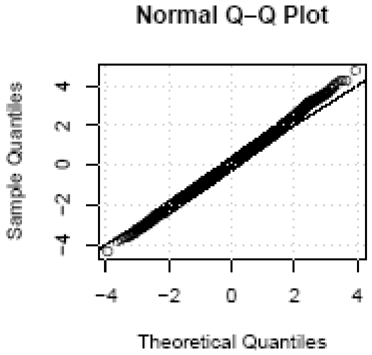
The following explanations apply to all studies

1. All figures are after transforming to log2.
2. The observed t-statistics are plotted against the expected t-statistics.

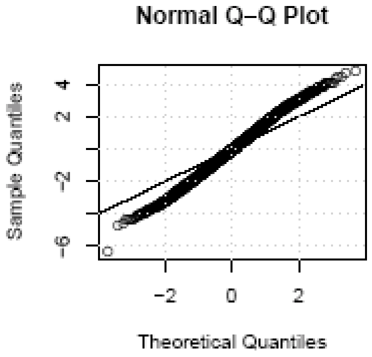
Diffuse large B cell lymphoma, (Rosenwald et al., (2002)



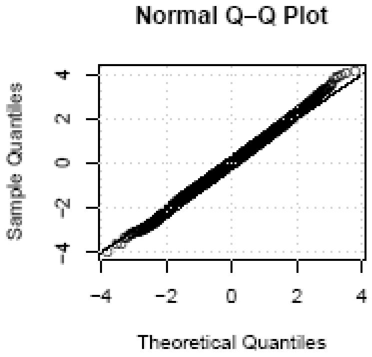
Acute lymphoblastic leukemia, Yeoh et al., 2002



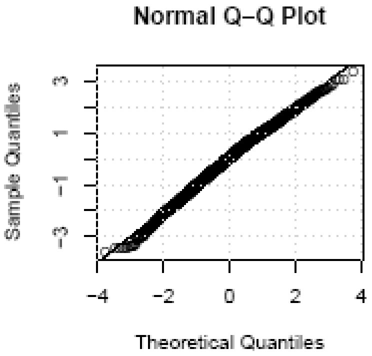
Breast cancer, van't Veer et al., 2002



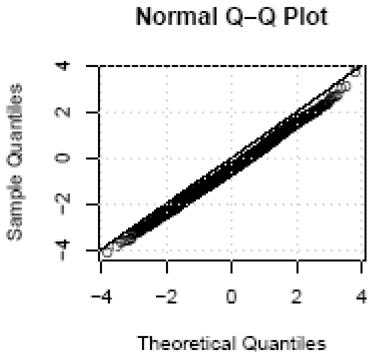
Lung adenocarcinoma, Beer et al., 2002



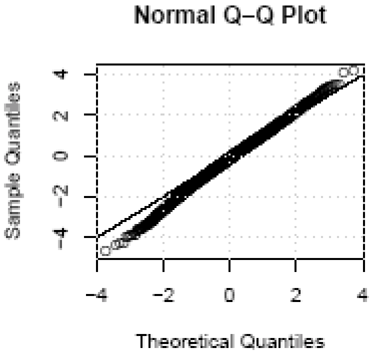
Lung adenocarcinoma, Bhattacharjee et al., 2001 and Ramaswamy et al., 2003



Medulloblastoma, Pomeroy et al., 2002



Hepatocellular carcinoma, Iizuka et al., 2003



References

A Bhattacharjee, WG Richards and J Staunton *et al.*, *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses*, *Proc Natl Acad Sci USA* **98** (2001), pp. 13790–13795.

A Rosenwald, G Wright and WC Chan *et al.*, *The use of molecular profiling to predict survival after chemotherapy for diffuse large B cell lymphoma*, *N Engl J Med* **346** (2002), pp. 1937–1947.

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

David B. et al., *Two Testing in Microarray Analysis: What is Gained?*Department of Biostatistics, Ryals Bldg, Suite 327, 1665 University Blvd, Birmingham, Alabama 35294

David M Mutch et al., *Microarray data analysis: a practical approach for selecting differentially expressed genes*, *Genome* **Biology 2(12)**

DG Beer, SL Kardia and CC Huang *et al.*, *Gene-expression profiles predict survival of patients with lung adenocarcinoma*, *Nat Med* **8** (2002), pp. 816–824.

Dudoit Sandrine *Primer on Genetics and Molecular Biology*, KolleKolle, Denmark 2003

Efron B, Tibshirani R, Storey JD, Tusher V. (2001) *Empirical Bayes Analysis of a Microarray Experiment*. _JASA_, 96(456), p. 1151-60.

EJ Yeoh, ME Ross and SA Shurtleff *et al.*, *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling*, *Cancer Cell* **1** (2002), pp. 133–143.

Hardin, J., 2005, *Microarray Data from a Statistician's Point of View*, ASA

Leland H. Hartwell et al (2000). *Genetics from Genes to Genomes,* McGraw-Hill Companies, Boston.

LJ van't Veer, H Dai and MJ van de Vijver *et al.*, *Gene expression profiling predicts clinical outcome of breast cancer*, *Nature* **415** (2002), pp. 530–536.

Michiels, S. et al. (2005) *Prediction of cancer outcome with microarrays: a multiple random validation strategy*. Lancet 2005; 365: 488–92

N Iizuka, M Oka and H Yamada-Okabe *et al.*, *Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection*, Lancet **361** (2003), pp. 923–929.

Ntzani EE, Ioannidis JP. *Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment.* Lancet 2003; 362: 1439–44.

Pawitan, Y., K. R. K. Murthy, S. Michiels, and A. Ploner. 2005. *Bias in the estimation of false discovery rate in microarray studies*. Bioinformatics **21:**3865-3872

Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A (2005) *False discovery rate, sensitivity and sample size for microarray studies*. Bioinformatics. **21**, 3017–3024.

S.Dudoit, J.P. Shaffer, and J.C. boldrick "Multiple Hypothsis testing in microarray experiments" Technical Report #10

SL Pomeroy, P Tamayo and M Gaasenbeek *et al.*, *Prediction of central nervous system embryonal tumour outcome based on gene expression*, *Nature* **415** (2002), pp. 436–442.

S Ramaswamy, KN Ross, ES Lander and TR Golub, *A molecular signature of metastasis in primary solid tumors*, *Nat Genet* **33** (2003), pp. 49–54.

Speed, Terry (2003). *STATISTICAL ANALYSIS of GENE EXPRESSION MICROARRAY DATA*, First Edition, CHAPMAN & HALL/CRC.

Storey J. D (2002) *A direct approach to false discovery rates.J. R. Statist. Soc. B.,* ***64****,479-498.*

Yekutieli,Dl and benjamini Y.(1999) "*Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics*" J.Stat.Plan Infer., 82,171-196