2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Joint modeling of phenotipic variables and gene expression data in early drug development experiments

Promotor :
Prof. dr. Ziv SHKEDY

Nolen Joy Perualila

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Joint modeling of phenotipic variables and gene expression data in early drug development experiments

Promotor :
Prof. dr. Ziv SHKEDY

Nolen Joy Perualila

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Acknowledgments

I wish to thank several people for being with me through my ups and downs towards the realization of this thesis.

My sincere appreciation goes to my thesis supervisor, Prof. Ziv Shkedy, for providing me with my thesis subject and for stirring up my interest in the field of bioinformatics. I am fortunate to have your kind words and advice concerning my work. Also, I would like to thank Prof. Dan Lin for several helpful discussions that provided me a deeper understanding of my topic.

To all my Censtat professors, thank you for your guidance and teaching expertise in giving us quality education. To Mrs. Martine Machiels, thank you for being with us in every step of the way!

I am also grateful to all my classmates and friends for making my stay in Belgium easier and more worthwhile. Special thanks to Atiq, Patwary, Veronica, Angelica, Moira and Mekdes for the great friendship that I would surely treasure forever. To my very responsible group mates, Saka and Leandro, I was amazed by how dynamic a group we were and how well we handled everything. Thank you for the very nice working environment you have shared with me!

To Ate Chella, Nang Gemma and Guido, I cant thank you enough for being my surrogate family here in Belgium. Thank you for your unending care and attention! Also, to Prof. Tina Sotto for all her encouragements and advice, I am deeply grateful!

All these experiences have been possible through the financial support of VLIR-UOS, which I gratefully acknowledge. Thank you for this great opportunity!

I would like to extend my appreciation to Prof. Geraldine Garcia, who told me about the masters programme in Biostatistics and advised me on the application for the VLIR scholarship.

Finally, I am forever indebted to my parents and Cesar for their understanding, endless patience and encouragement when it was most required. I am happy to share all that I have accomplished to my brothers and sisters, Novie, Njlo, Neil John, Nuary Jan, and Nejema. Thank you for your continuing love and support!

And to everyone, whom I have forgotten to mention, thank you for helping me, in one way or another, in reaching another milestone in my life.

Nolen Joy Perualila

12 September 2011

JOINT MODELLING OF PHENOTYPIC VARIABLES AND GENE EXPRESSION DATA IN EARLY DRUG EXPERIMENTS

by

Nolen Joy Perualila

Hasselt University,2011
Under the Supervision of Prof. dr. Ziv Shkedy

Abstract

Drug development benefits enormously from a microarray experiment, a tool that allows accurate and relatively inexpensive collection of gene expression information for thousands of genes at a time. Recently, it has become a commonplace in biomedical research to monitor gene expression levels associated with different phenotypes. It is the aim of the investigator to determine which genes or combination of of genes could serve as biomarker for the IC_{50}. The joint modeling approach that allows the investigation of the relationship between the gene expression and the IC_{50} after adjusting for the treatment effect was used in the selection and evaluation of genomic biomarkers. Depending on their intended use, biomarkers are further classified as prognostic and therapeutic. In the hope of achieving information gain, Supervised Principal Components Analysis (SPCA) was also conducted to construct a joint biomarker profile. Of the 7722 genes, 288 and 900 genes can serve as therapeutic and prognostic biomarkers for the response, respectively. Thirty (30) are identified to be potential prognostic/therapeutic genes.The top 2 therapeutic, top 8 prognostic and top 5 therapeutic/prognostic genomic biomarkers were used to construct their respective joint biomarker profile.

Keywords: *Biomarkers; Gene expression; Genomic Biomarker; Joint Biomarker; Joint modeling; Microarray experiment; Phenotype; Prognostic; Therapeutic; SPCA*

Prof. dr. Ziv Shkedy 12 September 2011

Contents

List of Figures

List of Tables

1 Introduction

Today's drug discovery process is time consuming and tremendously expensive. It has evolved into an extremely complex procedure and can take up to fifteen years to develop one new medicine from the earliest stages of discovery to the time it is available for treating patients. Also, it costs millions of dollars to bring a single drug candidate to market. In early drug experiments, each candidate drug is administered at different dosages and is tested against target cells. Compounds that retard the growth of the cells are recommended for the next level of testing. Phenotyping is recognised as providing a quantified measurement of drug resistance, It involves direct quantification of drug sensitivity. The outcome of a phenotypic test may be expressed as IC_{50}, IC_{90} or IC_{95} value which expresses the concentration of a particular drug required to inhibit the growth of the virus by 50%, 90% or 95%, respectively. These results are easily interpreted but often prove to be time consuming, expensive and labour intensive.

Drug development may benefit enormously from a tool that allows accurate and relatively inexpensive collection of gene expression information for thousands of genes at a time (Drăghici, 2003). Gene expression data measuring the absolute or relative transcript abundance of potentially every gene in the cell provide invaluable insights into the global functioning of organisms (Parmigiani *et al.*, 2006). Recently, microarray experiments have become commonplace in biomedical research to monitor gene expression levels associated with different phenotypes. Microarray experiments involve quantitative analysis of the expression levels of many thousands of genes in parallel. A typical analysis of DNA microarrays allows monitoring expression levels of thousands of genes simultaneously, and identifying genes whose expression levels are significantly changed under different experimental conditions (Azuaje, 2006). As comprehensive analyses of genes became available, interest in developing reliable sets of measurable genetic characteristics that correlate with specific clinical outcomes (phenotypes), such as physiological processes, pharmacological responses to a therapeutic intervention (compound efficacy), toxicological measures, *etc.*, is gaining attention over the years.

Each gene represented in microarrays can be considered a potential biomarker. A consensus definition of a biomarker is a factor that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Biomarkers Definitions Working Group, 2001).The use of biomarkers has the potential to facilitate the availability of safer and more effective drug or biotechnology products, to guide dose selection and to enhance their benefit-risk profile (EMEA, 2009). Appropriate biomarkers can provide critical feedback once an agent has reached clinical testing. Perhaps most importantly, biomarker-based studies may provide early evaluations of the key question of mechanistic success or failure (hitting the target). Lack of mechanistic activity in early clinical trials can help to curtail further costly clinical testing and redirect efforts toward additional pre-

clinical studies (Park, 2004). Hence, in both pre-clinical and clinical trials, biomarkers have the potential to encourage innovation, improve efficiency, save costs, and gain research organizations a valuable advantage over their competitors.

The selection and evaluation of genomic biomarkers play a vital role in drug discovery and development, motivating the use of apt statistical techniques to understand the complex nature of the relationship between genes and clinical outcomes (Lin,*et al.*, 2011). An incorrect or suboptimal method might lead to a great loss in terms of identifying genes that otherwise could have lead to a substantial improvement in understanding the properties of the relevant clinical outcome. New developments in the possibility of predicting the clinical outcome or phenotypic variable is not just limited to the use of a single gene expression level but also combinations thereof (Lin *et al.*, 2010). The former is called genomic biomarker, while the later is termed a joint biomarker. This, however, is challenging due to high dimensionality of data and relatively small number of observations. Moreover, although the number of genes assayed is large, there may be only a small number of biomarkers that are associated with variations of phenotypes. Biomarkers have been classied as prognostic and/or therapeutic depending on their relationship with the clinical endpoint and their response to treatments.

Therapeutic genes are those that respond to treatment and possibly aid in understanding the treatment effect on the gene-expression which can be predictive for the treatment effect on the response and they are not linearly associated with the response.

The prognostic genes, on the other hand, are the ones that are related to the response irrespective of treatment. These genes enable us to learn a great deal about the biological pathways between them and the response of interest.

In the clinical trial setting: there is a true end point and (usually) one candidate to be a surrogate - a biomarker that is intended to substitute for the true endpoint. Prentice (1989) defined a surrogate endpoint as a variable for which a test of the null hypothesis of no relationship to the treatment group under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint.It is also important to highlight that a microarray experiment is equivalent to the single trial setting and hence, the gene specic models used to compute the association measures should be tuned to reflect the single trial setting. For a microarray with m genes, there are m candidates that can be used as biomarker. The analysis presented in this paper aims at finding a subset of genes, that are associated with the response, IC_{50}, while accounting for the treatment effects and can be used as biomarkers. The joint modeling of the phenotypic variable/response and the gene expression facilitates the evaluation of their association. Figure 1 illustrates the joint model for the surrogacy and microarray experiments. This follows similar lines as the one presented in the case studies of Lin *et al.* (2010) and Bair *et al.*(2006). In addition, it is also of interest to construct a joint biomarker where relevant information from a subset of genes is combined to predict the response.

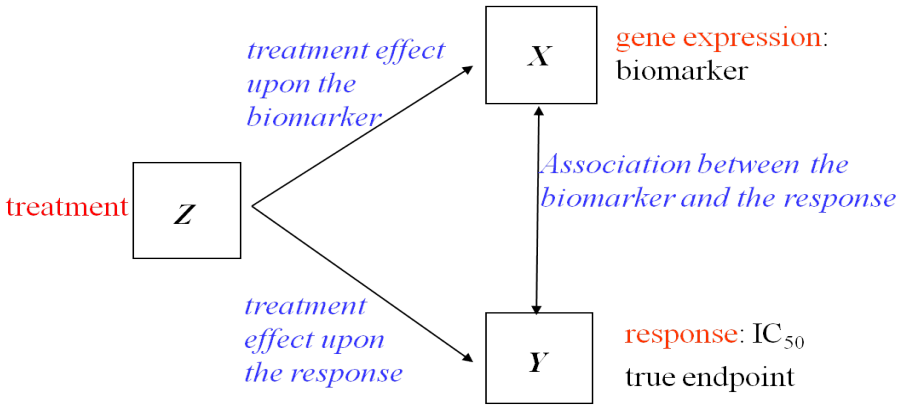

Figure 1: *The Biomarker Setting*

This thesis is organized as follows: Section 2 presents a brief description of the data. Section 3 offers the methods used to answer the objectives. It covers a detailed introduction of the joint model to evaluate the association between the gene expression and the response after adjusting for the treatment effect in microarray experiments. It also tackles issues on multiple hypothesis testing or multiplicity, cross-validation and the construction of joint biomarker profile using SPCA technique. In Section 4, the results obtained from applying the methods to the data are presented. The paper ends with some discussions and conclusions given in Section 5.

2 Data

In this paper, we focus on the microarray setting (Figure 2), in which data are available from a single trial. For each cell line, microarray data consisting of 7722 genes (X) together with a (clinical/experimental) response variable (Y) are available under 2 treatments (Z). Cell lines that have a link to the target disease were chosen. A dose-response experiment with 47 cell lines was conducted. Of the 47 cell lines, 32 were randomized to treatment 14 and the rest to treatment 29. Each cell line is then treated with a compound that is known to belong to different clusters. Unlike treatment 14 which is composed of compounds coming from one cluster, treatment 29 is a mixture of clusters wherein 7 of the 15 cell lines are treated with compounds that belong to cluster 181, 3 to the reference group and the rest to 5 other different clusters. There was no information, however, on how the clustering of compounds and the treatment groups were done.

Cell lines can respond or not respond to the compound. The IC_{50} represents the activity of the cell line (responder/non responder). The IC_{50} phenotype is the dose level for which the response is half way to the maximum effects. The estimated IC_{50} is an important measure of compound efficacy.

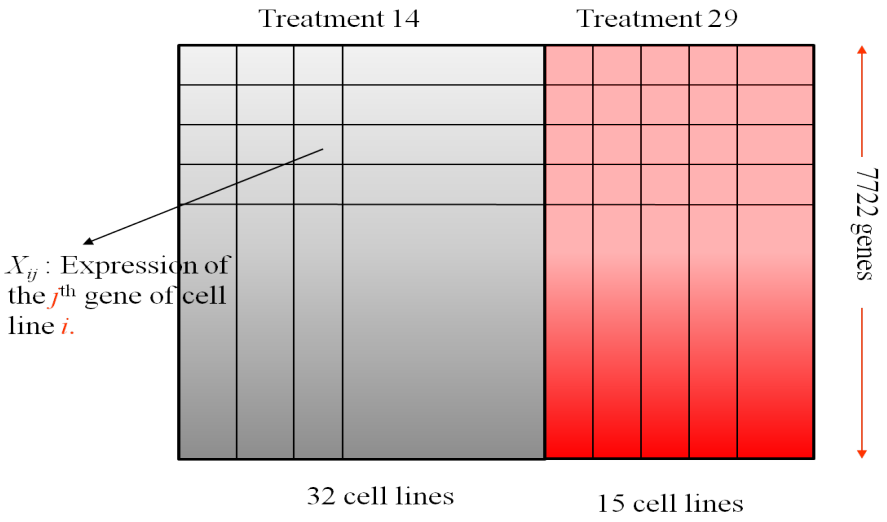

Figure 2: *The Microarray Setting with Two Treatment groups*

3 Methodology

This section provides methods to identify genes that can serve as therapeutic and/or prognostic biomarkers. The most commonly used method in gene expression studies is typically done for each gene separately. These tests are then aggregated again by ranking genes based on a statistic. This simple approach ignores the multivariate nature of gene expression and the omnipresent dependencies between genes. This, in turn, misses trends or interactions that exist between different genes and fails to discover interesting gene combinations. However, because there are typically so many groups and subgroups of correlated and interacting genes present in a single microarray, a multivariate approach that tries to incorporate most or all covariances becomes highly complex that consequently impede interpretability and generalizability (Ghölmann and Talloen, 2009). Thus, the gene-by-gene analysis is often the option (Ge *et al.*,2003; Kerr *et al.*, 2000), providing a simple solution that is very helpful in extracting relevant information from the high dimensional and complex microarray data.

3.1 Selection of Gene-Specific Biomarkers

3.1.1 Fold-change

The simplest and most intuitive approach to finding genes that are differentially expressed is to consider the fold change which is just the difference per gene between the averages of the two treatment groups.

Genes showing fold change above 2 (or another arbitrary cut-off) were regarded as potentially

regulated and were chosen for further investigation. However, this method has important disadvantages. One is related to the fact that microarray technology tends to have a bad signal/noise ratio for genes with low expression levels. Genes with high fold change may also be highly variable and thus with low significance of the regulation. A small difference between means will be hard to detect if there is lots of variability or noise. This method should only be done if one prefers to focus on the absolute difference between groups, thereby deliberately ignoring the information of the variation within the groups (Ghölmann and Talloen, 2009).

3.1.2 Gene-specific Joint Model

Analysis of variance (ANOVA) is a general analysis approach that can provide information able to discern significant difference in expression levels of a gene exposed to several treatment conditions. An ANOVA basically partitions the observed variation in gene expression between the samples into components due to different groups and unexplained variation (the residual noise). It determines the significance of the group effect by comparing the differences between the groups to the variation within the group. Moreover, since selecting genes as possible biomarkers for a particular response requires quantifying the degree of association between the response of interest (Y_i) and the gene expression (X_{ij}), after correcting for treatment (Z_i), a joint ANOVA model is fitted for these outcomes of interest.

Following Buyse *et al.* (2000), the gene-specific joint model that therefore allows testing for which gene is differentially expressed and which gene can serve as a biomarker is specified as follows

$$\begin{aligned} X_{ij} &= \mu_j + \alpha_j Z_i + \epsilon_{ij} \\ Y_i &= \mu_Y + \beta Z_i + \epsilon_i \end{aligned} \tag{1}$$

where the error terms have a joint zero-mean normal distribution with covariance matrix.

$$\Sigma_j = \begin{pmatrix} \sigma_{jj} & \sigma_{jY} \\ \sigma_{jY} & \sigma_{YY} \end{pmatrix} \tag{2}$$

or equivalently formulated as

$$\begin{pmatrix} X_{ij} \\ Y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_j + \alpha_j Z_i \\ \mu_Y + \beta Z_i \end{pmatrix}, \Sigma_j \right] \tag{3}$$

and Z_i is an indicator variable which takes a value of 1 if subject was randomized to treatment 14 and 0 to treatment 29. The parameters α_j and β are the treatment effects upon the j^{th} gene and the response, respectively, and μ_j and μ_Y are gene-specific and the response-related intercepts, respectively. Furthermore, both outcomes are assumed to be normally distributed.

The model parameters' estimation can be implemented in R using the *gls()* function in the nlme library. This would give $\hat{\alpha_j}$ and $\hat{\beta}$, the maximum likelihood estimates of the treatment effect for the jth gene expression and IC_{50}, respectively.

In the context of surrogate-marker evaluation in randomized clinical trials, Buyse and Molenberghs (1998) proposed the adjusted association as a measure of association, a coefficient derived from the covariance matrix (Eq.2) of gene-specific joint model (Eq. 1).

$$\rho_j = \frac{\sigma_{jY}}{\sqrt{\sigma_{jj}\sigma_{YY}}} \tag{4}$$

Additionally, one would expect a good prognostic biomarker to have a strong association with the true endpoint. A large value of ρ_j would provide evidence that the IC_{50} would then be largely determined by the j^{th} gene regardless of any treatment effect. Thus, this follows that ρ_j=1 indicates a deterministic relationship between the gene expression and the response, in the sense that, given gene expression, a perfect prediction of the IC_{50} score is possible. It is also possible to get a ρ_j equal to 1 even if the gene is not differentially expressed.

3.1.3 Information-Theory Approach

The main rationale for explicitly accommodating for the two outcomes' correlation is to allow for estimation of their association (Lin *et al.*,2011). The information-theoretical approach proposed by Alonso and Molenberghs (2007), which is elegant and computationally simple, can also be considered. In this case with normally distributed bivariate outcomes, consider the following linear models:

$$E(Y_i|Z_i) = \breve{\mu} + \breve{\beta} \times Z_i \tag{5}$$

$$E(Y_i|Z_i, X_{ij}) = \breve{\mu} + \breve{\beta} Z_i + \breve{\alpha} X_{ij} \tag{6}$$

where $\breve{\alpha}$ is the gene-specific effect upon the outcome. Model 5 relates the expected value of the true endpoint to the treatment only while 6 relates it to surrogate endpoint as well. Upon fitting models 5 and 6, the individual-level association can be measured by:

$$R^2_{hj} = 1 - exp\left(\frac{-G^2}{n}\right) \tag{7}$$

where G^2 denotes the likelihood ratio statistics to compare models 5 and 6, and n is the sample size. Note that for continuous outcomes,R^2_{hj} and the squared adjusted association $R^2_j = {\rho_j}^2$, known as the coefficient of determination from the linear model, give identical results since model 9 is implied by the gene specific joint model with the following conditional distribution

$$Y_i|Z_i, X_{ij} \sim N\left(\breve{\mu} + \breve{\beta}Z_i + \breve{\alpha}X_{ij}, \sigma^2\right) \tag{8}$$

where $\breve{\mu} = \mu_Y - \sigma_{jY}\left({\sigma_{jj}}^{-1}\mu_j\right), \breve{\beta} = \beta - \sigma_{jY}\left({\sigma_{jj}}^{-1}\right)\alpha_j, \breve{\alpha} = \sigma_{jY}\left({\sigma_{jj}}^{-1}\right)$
and $\sigma^2 = \sigma_{YY} - \sigma_{jY}\left({\sigma_{jj}}^{-1}\right)$

The conditional model can be extended by adding an interaction term between X and Z.

$$E(Y_i|Z_i, X_{ij}) = \breve{\mu} + \breve{\beta}Z_i + \breve{\alpha}X_{ij} + \delta_{ij}Z_iX_{ij}. \tag{9}$$

3.2 Testing and Evaluation of surrogate/biomarker genes in microarray experiments

3.2.1 Testing for prognostic biomarkers

The associations between the gene expression and the response after adjusting for treatment effects could be of linear or non-linear type. If a linear relationship between the gene expression and the response, after accounting for the treatment effect, can be possibly assumed, the gene can then serve as a prognostic biomarker, which can be used to predict the response.

To determine whether or not the gene can serve as a prognostic biomarker, the joint model is fitted twice, assuming different covariance matrix. The likelihood ratio test is used in order to test the null hypotheses:

$$\begin{aligned} H_{0j} : \Sigma_j &= \begin{pmatrix} \sigma_{jj} & 0 \\ 0 & \sigma_{YY} \end{pmatrix} \longrightarrow \rho_j = 0 \\ H_{1j} : \Sigma_j &= \begin{pmatrix} \sigma_{jj} & \sigma_{jY} \\ \sigma_{jY} & \sigma_{YY} \end{pmatrix} \longrightarrow \rho_j \neq 0 \end{aligned} \tag{10}$$

A gene is declared an up-regulated prognostic biomarker if the null hypothesis in (10) is rejected and $\rho_j > 0$, and a down-regulated prognostic biomarker when $\rho_j < 0$.

An alternative to the parametric adjusted correlation,ρ_j, is the nonparametric Spearman correlation given below, computed based on relative ranks and not on the quantitative values of the residuals. This is a more robust measure of association in the presence of outliers. Hence, it can also be used to measure the association between the response and gene expression for each gene.

$$r_{sj} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{11}$$

where d_i is the rank difference between the residuals from the joint model and n is the sample size.

3.2.2 Testing for therapeutic biomarkers

Using the joint model to identify therapeutic biomarkers or genes which are differentially expressed, the following null hypotheses are tested for each gene.

$$\begin{aligned} H_{0j} &: \alpha_j = 0 \\ H_{1j} &: \alpha_j \neq 0 \end{aligned} \tag{12}$$

Testing the treatment effect upon the response consists of testing $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$. Note that for the case, in which both: $H_{oj} : \alpha_j = 0$ and $H_0 : \beta = 0$ are rejected, implies that the gene is a potential therapeutic biomarker. In case that, $H_{oj} : \alpha_j = 0$, $H_0 : \beta = 0$, and $H_0 : \rho_j = 0$ are rejected, the gene is declared as a potential prognostic/therapeutic biomarker.

For therapeutic biomarkers, the adjusted association is not applicable since the association between the gene expression and the response is not linear. Lin *et al.* (2010) proposed a measure of association for therapeutic biomarker using the regression tree. Regression tree approach (Venables and Ripley, 1994), is a widely used technique to model the relationship between a response and a predictor without prior knowledge of the relationship between them. It is primarily used to construct a set of decision rules on the predictor variables by recursively partitioning the data into successively smaller groups with binary splits based on a single predictor variable. The optimum split is often chosen based on a split that maximizes the heterogeneity of the two resulting groups with respect the response variable (Adetayo, K., 2010).

For a therapeutic biomarker, because gene-expression is differentially expressed, the tree can be restricted to have only two terminal nodes (two final homogenous groups of the response), in which the cutoff point (or the split point) is determined only by the gene-expression level. An example of the cutoff point is shown as the vertical line in Figure 3. The total variability of the response, the deviance, without any information about the gene expression level can be measured by

$$D(Y) = \sum_{i=1}^{n} (Y_i - \hat{\mu})^2 \tag{13}$$

where $\hat{\mu} = \sum_{i=1}^{n} \frac{Y_i}{n}$ and $i = 1, \ldots, n$ indexes the arrays.

Let $D_1(Y|X)$ and $D_2(Y|X)$ denote the deviance in each of the terminal nodes and $D(Y|X)$ be their sum. The deviance reduction,$D(Y) - D(Y|X)$, measures the gain in prediction of the response level using gene-expression, as compared to the case where the gene-expression is not used. In other words, the reduction in deviance measures whether information about the gene-expression is relevant for predicting the response level. The relative deviance reduction, R_D^2, is given by

$$R_D^2 = \frac{D(Y) - D(Y|X)}{DY} \tag{14}$$

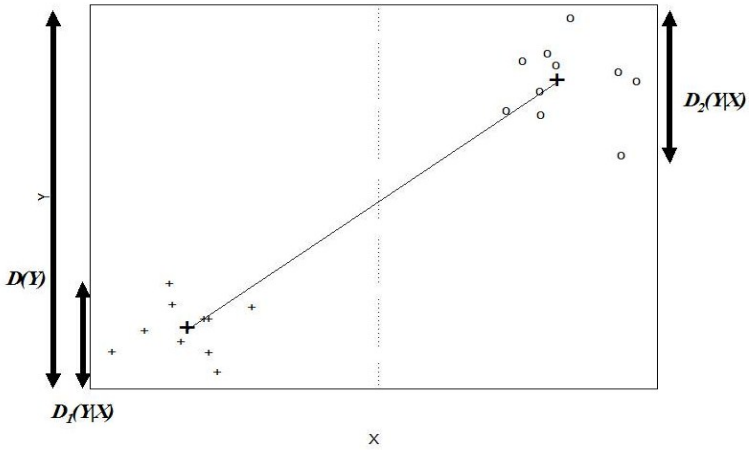

Figure 3: *An illustration of a regression tree model for a hypothetical example with two terminal nodes. The blue line in the plot indicates the split point in the regression tree. In $D(Y)$ represents the total variability in the response Y, while $D_1(Y|X)$ and $D_2(Y|X)$ represent the variability within each of the terminal nodes.*

which measures the proportion of variability explained by the regression tree model.

Although prognostic and therapeutic biomarkers are evaluated using different validity measures, R^2_D and R^2_j, respectively, they both, however, belong to the family of information theoretic association measures. This implies that both measures can be equally interpreted as tools that measure the proportion of information in the response captured by using the gene expression (Lin *et al.*, 2010).

3.3 Cross-Validation

Too much variability or presence of several outlying values especially for small datasets can distort the true relationship between the response and the gene expression. Leave-one-out cross validation was performed to assess the quality of the resulting measures of association for each type of biomarker. Each learning set is created by taking all the samples except one, the test set being the sample left out. Thus, for 47 samples, we have 47 different learning sets and 47 different tests set A comparable results of the measure of association using the full data and the 47 cross validation data sets suggests a good estimate of R^2 values.

3.4 Multiplicity

All the tests mentioned in the preceding subsections are repeated as many times as there are genes in the microarray dataset. This increases the number of false positive results, i.e genes are found to be statistically different between conditions or is linearly associated with the response,

but are not in reality. This problem of multiple testing or multiplicity needs to be corrected by calculating an adjusted p-value that takes into account the number of tests that have been carried out. Every multiple testing correction procedure uses some error rate to measure the occurrence of incorrect conclusions. Two types of error rates are the Family-wise error rate (FWER) and false discovery rate (FDR).

FWER refers to the expected occurrence of false positives among all tested genes, assuming that the null hypothesis is true. Here, the term family refers to the collection of hypotheses $H_1, \ldots, H_m$ that is being considered for joint testing. Once the family has been defined, strong control of the FWER (at a joint level α) requires that FWER$\leqslant \alpha$ for all possible constellations of true and false hypotheses (Lehmann and Romano, 2005).

On the other hand, the FDR is the expected occurrence of false positives among the genes call significant assuming that the null hypothesis is true. (Ghölmann and Talloen, 2009). While FWER methods only allow very few occurrences of false positives, FDR methods allow a percentage of positives to be false positives. In multiple testing, strong control of the Family-Wise error rate (FWER) can be unnecessarily stringent in microarray settings (Xu and Hsu, 2007). Traditional approaches to control FWER are too conservative when applied to microarray data. Approaches based on the control of the FDR have gained their popularity in the microarray setting, because they lead to a higher power as compared to the methods that control the FWER.

Various procedures to control FDR have been proposed (Benjamini, Hochberg, 1995; Benajamini and Yekuteli, 2001; Yuketili and Benjamini,1999). The Benjamini and Hochberg (BH) is only theoretically valid when the genes are not correlated. The Benjamini and Yuketieli (BY) method is valid for any level of correlation between the genes but is so conservative that almost no one uses it. Simulation suggests that the BH method is unlikely to fail for realistic scenarios and is not too conservative, which is therefore the preferred choice in the microarray setting (Ghölmann and Talloen, 2009).

The FDR correction is applied after ordering the p-values from largest to smallest. The BH-FDR is computed as follow:

$$p_j^{BH} = p_j \frac{m}{order(p_j)} \; with \quad j = 1, \ldots, m \;\; genes \tag{15}$$

Note that the multiplicity correction affects only the arbitrary threshold choice and does not change the ranking of the genes. All the tests make use of an FDR level of 0.05.

3.5 Joint Biomarker Profile

3.5.1 Supervised Principal Component Analysis (SPCA)

The gene-specific approach allows identifying individual genes as biomarkers. However, information gain might be achieved if a joint biomarker could be constructed. This joint biomarker profile combines the information from expression levels of individual genes into one score and use this score as a biomarker. In the microarray setting, the number of predictors (genes) is large compared to the number of observations and the design matrix (X) is likely to be singular which makes linear regression to summarize information into one linear predictor no longer feasible. One way out is to eliminate the multicollinearity problem by performing principal component analysis (PCA). Then, the first principal component (i.e the gene profile/signature) is used to predict the response. In gene expression, the first principal component can be interpreted as the weighted average across the selected genes that explain the largest amount of variation in the data (Parmigiani, 2003). However, as mentioned by H.M. Bøvelstad(2007), a drawback of PCA is that there is no guarantee that the principal component is associated with the response. That is, directions with high variability in the gene expressions can be due to effects that are not related to the response. Bair *et al.* (2004) then proposed the supervised principal components (SPCA) method, to construct a gene profile that can be used to predict a quantitative response. SPCA is similar to conventional principal components analysis except that one can use only those genes with the strongest estimated correlation with the response. This supervised analysis (i.e., supervised gene screening step) reduces the dimension of the expression matrix ($\mathbf{X}$) and greatly increases the likelihood that the resulting principal components are associated with the outcome of interest. The SPCA, therefore, relies on the underlying assumption that there is a latent variable U($\mathbf{X}$) (the gene profile), which is maximally associated with the response variable Y. This is in contrast to the setting considered in the individual genomic biomarkers, in which the biomarker (gene-expression) is not latent but observed (Tilahun *et al.*., 2010).

3.5.2 Joint prognostic biomarker

Upon identification of the top k potential prognostic biomarkers using the gene specific joint model (1), PCA was used to summarize them into a single value, denoted by $U(\mathbf{X_I})_i$. The surrogacy measure can be obtained by fitting the joint model for the outcome and the joint biomarker profile

$$\begin{aligned} U(\mathbf{X_I})_\mathbf{i} &= \acute{\gamma} + \acute{\beta}\mathbf{Z_i} + \acute{\epsilon}_\mathbf{i} \\ \mathbf{Y_i} &= \tilde{\gamma} + \tilde{\beta}\mathbf{Z_i} + \tilde{\epsilon}_\mathbf{i} \end{aligned} \tag{16}$$

which implies that condition on the gene profile, the IC_{50} follows the following regression model

$$\mathbf{Y_i} = \gamma + \delta\mathbf{U}(\mathbf{X_I})_\mathbf{i} + \beta\mathbf{Z_i} + \epsilon_\mathbf{i} \tag{17}$$

Once the joint biomarker,$U(\mathbf{X_I})_\mathbf{i}$ is computed, the same measures presented for gene specific joint model, ρ or R^2, can be used to quantify its association with the response, with the spearmans correlation, r_s as an alternative. The final joint biomarker profile is composed of k number of top genes that maximizes R^2. Similarly, genes that are classified as prognostic/therapeutic biomarkers are also used to construct a joint profile.

3.5.3 Joint therapeutic biomarker

For therapeutic biomarker genes, similar to the single-gene case, the regression tree with two terminal nodes is used, that can be expressed as

$$\mathbf{Y_i} = \gamma_\mathbf{0} + \gamma_\mathbf{1}\mathbf{I_i}\left[\mathbf{U}(\mathbf{X_{II}})_\mathbf{i}\right] + \epsilon_\mathbf{i}, \qquad \mathbf{I_i}\left[\mathbf{U}(\mathbf{X_{II}})_\mathbf{i}\right] = \left\{ \begin{array}{ll} 1 & U(\mathbf{X_{II}}) > \eta \\ 0 & U(\mathbf{X_{II}}) \leq \eta \end{array} \right\} \tag{18}$$

where $\mathbf{I_i}\left[\mathbf{U}(\mathbf{X_{II}})_\mathbf{i}\right]$ is an indicator variable, that depends on the split point (η) dening the two terminal nodes in the regression tree. The number of top genes, k, is determined based on the gene profile that maximizes the relative deviance reduction, R^2_D.

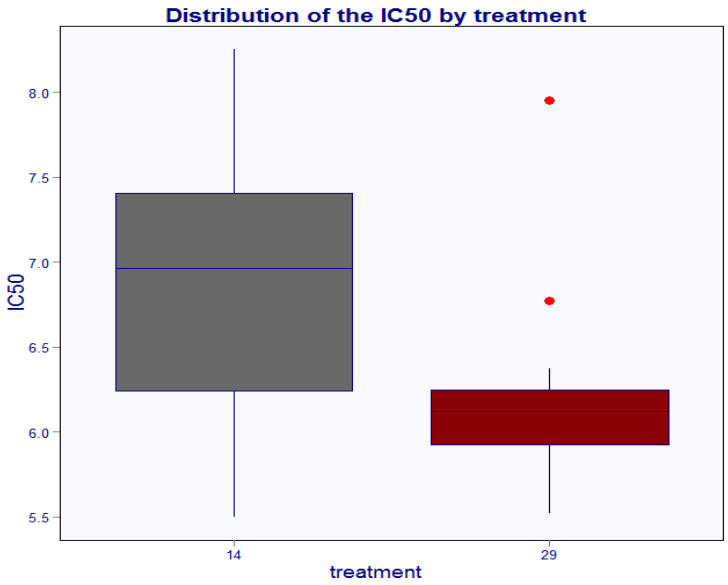

Figure 4: *Box plot of IC_{50} in treatments 14 and 29.*

4 Results

The box plot in figure 4 shows the distribution of the phenotypic variable, IC_{50}, by treatment. Two extreme upper values are observed for treatment 29. The group means for treatment 14 and 29 are respectively 6.8661and 6.1693, giving a difference of 0.6968. This suggests that treatment 29 has better inhibition. Testing the treatment effect upon the response using the joint model produced a test-statistic equal to -3.167 with corresponding p-value of 0.0028, an evidence of a significant treatment effect.

4.1 Testing and Evaluation of surrogate/biomarker genes

Given that the treatment effect upon the response is present, methods discussed in section 3 to identify differentially expressed genes that can serve as therapeutic biomarkers are applied. A maximum fold change of 1.28 was observed. The intensity plot showing the expression levels of the top three genes with the highest treatment effect are depicted in figure 5. The average group intensities are indicated by a horizontal line. In the figure, notice that treatment 29 is a mixture of different shapes. This is to differentiate major clusters from which the compound used in treating the cell lines came from. In addition, the expression levels of the genes using a reference compound and a placebo, Dimethyl sulfoxide (DMSO), which can both serve as control, are presented. Note, however, that this paper investigates only the difference between the two treatment groups and the control versus treatment effect can be another focus of ones research. Interestingly, it is generally observed that treatment 29 has the same mean level with

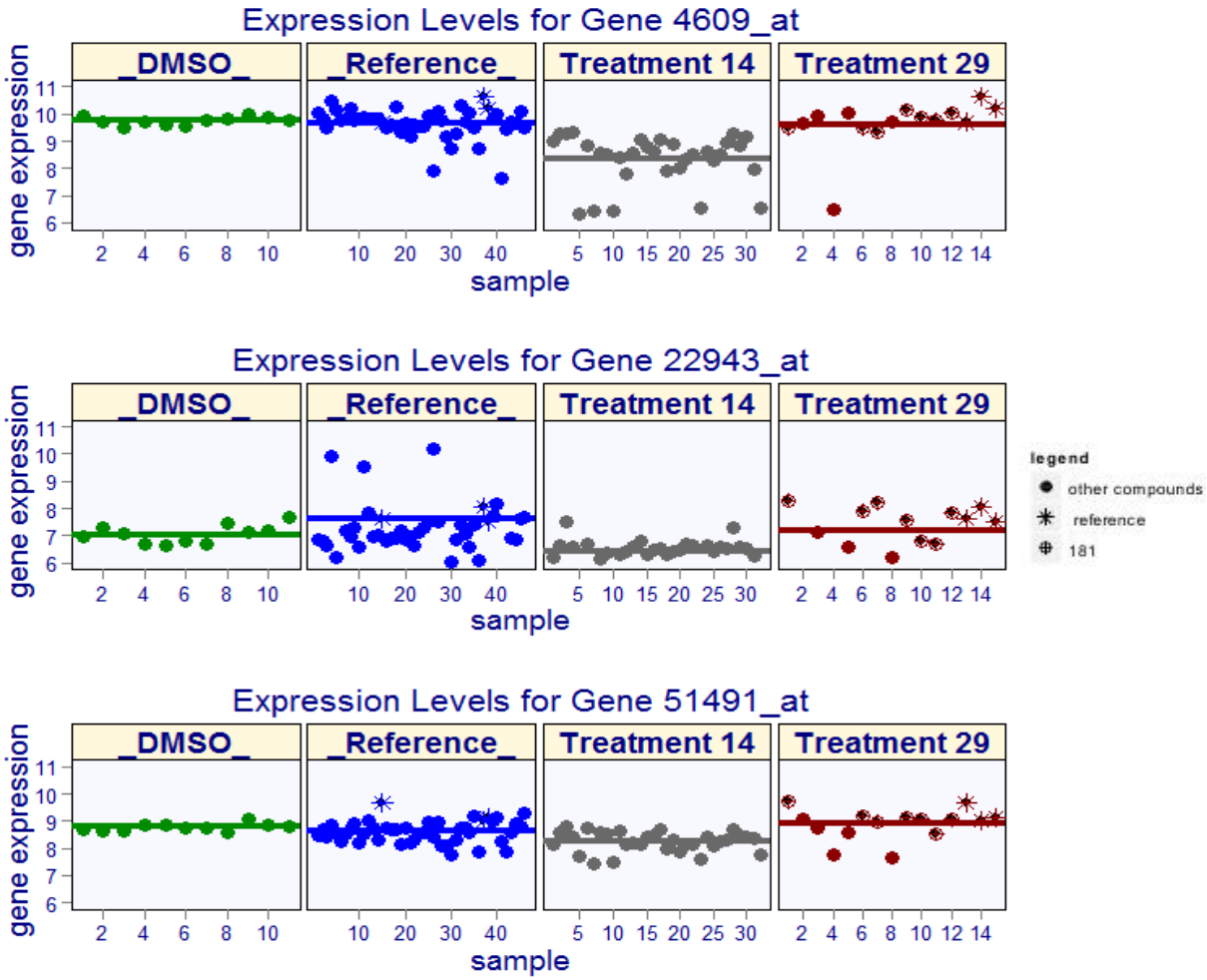

Figure 5: *Intensity plots of top 3 genes ranked based on Fold change.*

the control groups. Figure 5 also reveals that the genes have different variation across the replicates within a treatment. This implies that ranking genes based on the fold change will not produce an accurate gene list that can serve as most potential therapeutic biomarkers. Indeed a gene with the high fold change may just be highly variable and thus with low significance of the difference.

Hence, it is of advantage to use the gene-specific joint model that allows the test for the significance of the treatment effects upon the gene expression ($Ho : \alpha_j = 0$) while incorporating the information of variation within the treatment. Of the 7722 genes, 288 were found to be differentially expressed at FDR level of 0.05. Figure 6 shows the expression levels for the top three differentially expressed genes. Observe that, although the fold change exhibited by the genes is small, it can be seen that the variation within the group is also small which led to a significant treatment effect. This can be clearly visualized by using a volcano plot in figure 7

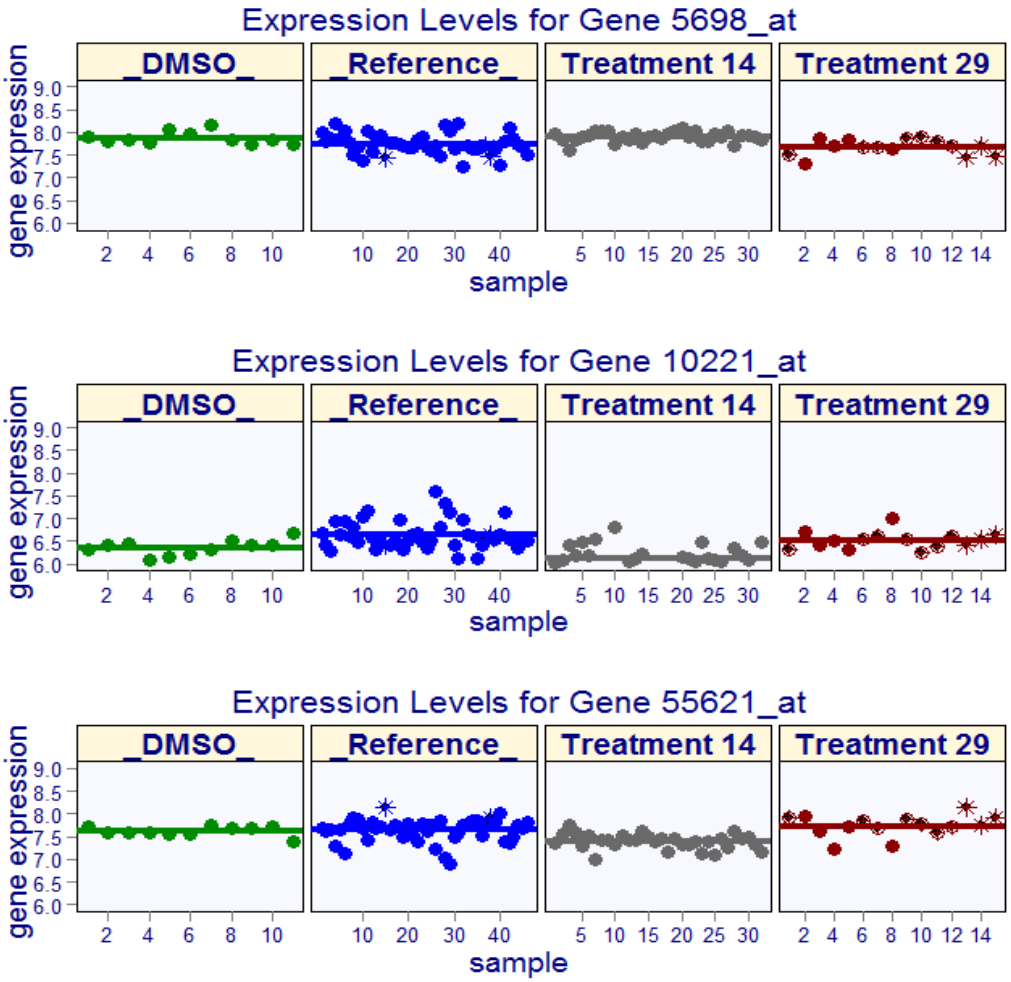

Figure 6: *Intensity Plots of Top 3 genes ranked on the basis of statistical significance*

that plots the fold change versus the log_{10}(p-value), where a number of genes are significant (green points) at FDR level of 0.01 despite a small fold change.

Figure 8 shows a heatmap of the 50 most differentially expressed gene where every row represents a gene expression profile of many genes across several samples. The abundance is depicted by color coded squares where red refers to up-regulated genes, white depicts genes that do not change, and blue depicts down regulated genes. It is readily noticeable that the genes nicely separate the samples treated with treatment 14 (gray-colored columns) and 29 (red-colored columns), which is as expected.

The list of the top 15 differentially expressed genes which are not linearly related with the response after adjusting for treatment effect (graphically supported in lower panel of figure 8) that can serve as therapeutic biomarkers are presented in table 1. The relative deviance reduction, R^2_D using the regression tree, is used to measure the quality of the therapeutic biomarkers.

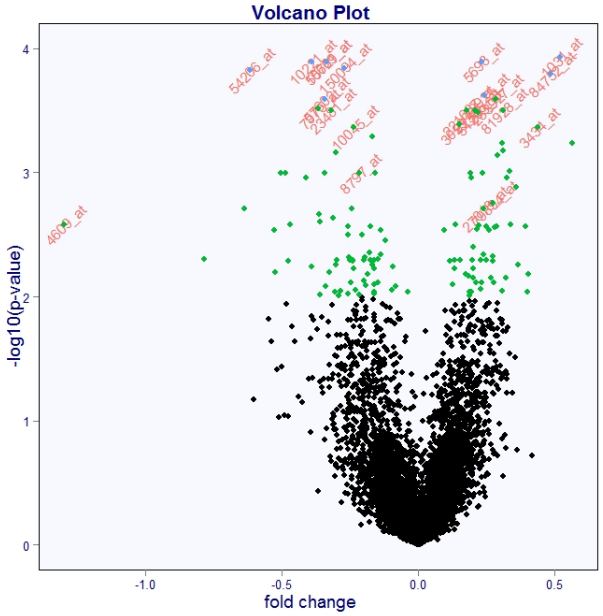

Figure 7: *Volcano plot*

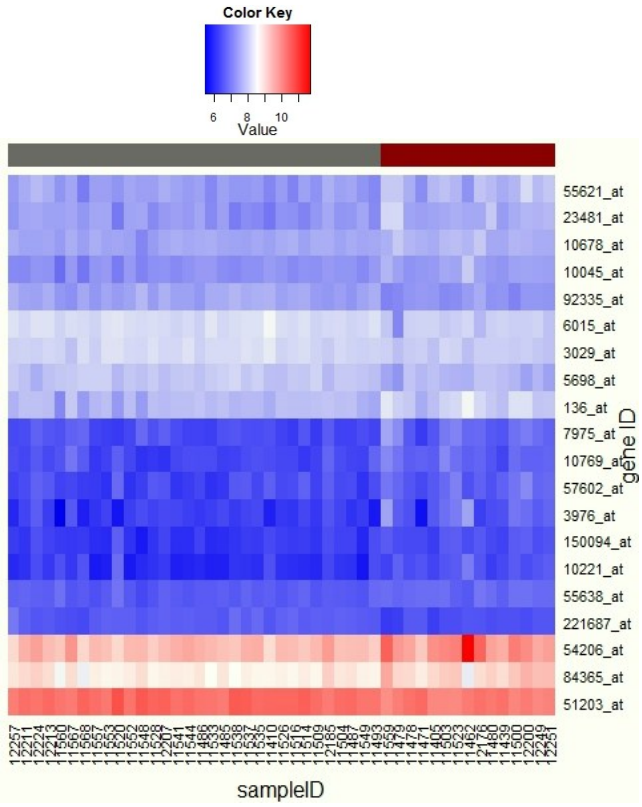

Figure 8: *Heatmap of top 50 differentially expressed genes*

Inspecting the R^2_D of the top 20 potential therapeutic biomarkers reveals that the values are very low with 0.4760 as the highest. This may be explained by examining the upper panel of figure 9 wherein the cut-off point does not clearly separate the data into two homogenous groups that corresponds to the treatment groups. Moreover, the variability within each of the terminal nodes $D(Y|X)$, is too high. This suggests that the total variability in the response is not much reduced by forming these two groups in the gene expression. The leave-one out cross validation data yield comparable results with the original data which might give comfort to the validity of the measures.

Table 1: *Results of top 15 differentially expressed genes with their significant treatment effects alpha, raw p-values, BH-FDR adjusted p-values and Relative Deviance Reduction (full and leave-one-out cross validation) using the regression tree*

ID	α	**t-stat**	**p-value**	**Adj-pval**	R^2_D	R^2_{Dcv}
5698_at	-0.2329	-5.5701	<0.0001	0.0001	0.4081	0.4084
10221_at	0.3911	5.5698	<0.0001	0.0001	0.1983	0.1999
55621_at	0.3361	5.5644	<0.0001	0.0001	0.3086	0.3089
10769_at	0.3358	5.5531	<0.0001	0.0001	0.2182	0.2157
150094_at	0.2715	5.5122	<0.0001	0.0001	0.4350	0.4358
54206_at	0.6147	5.4897	<0.0001	0.0001	0.1070	0.1085
57602_at	0.3414	5.3225	<0.0001	0.0003	0.4759	0.4760
7975_at	0.3671	5.2661	<0.0001	0.0003	0.0675	0.0694
92335_at	-0.2090	-5.2389	<0.0001	0.0003	0.1473	0.1555
221687_at	-0.1789	-5.2371	<0.0001	0.0003	0.2282	0.2234
23481_at	0.3177	5.2178	<0.0001	0.0003	0.1199	0.1245
51203_at	-0.2185	-5.2042	<0.0001	0.0003	0.2440	0.2395
3029_at	-0.1509	-5.1399	<0.0001	0.0004	0.1243	0.1274
10045_at	0.2341	5.1133	<0.0001	0.0004	0.1586	0.1637
10678_at	0.1676	5.0611	<0.0001	0.0005	0.1177	0.1577

4.2 The adjusted Association ρ_j: Testing for prognostic Biomarkers

For a gene to be identified as potential prognostic biomarkers, it must be shown to be linearly associated with the response after accounting for treatment effect. The gene-specific joint model is fitted twice to all genes and the likelihood ratio test is performed in order to test the null hypothesis that the correlation between the response and gene expression is zero ($H_{0j} : \rho_j = 0$) using an FDR level of 0.05. Figure 10 shows that the genes are not differentially expressed

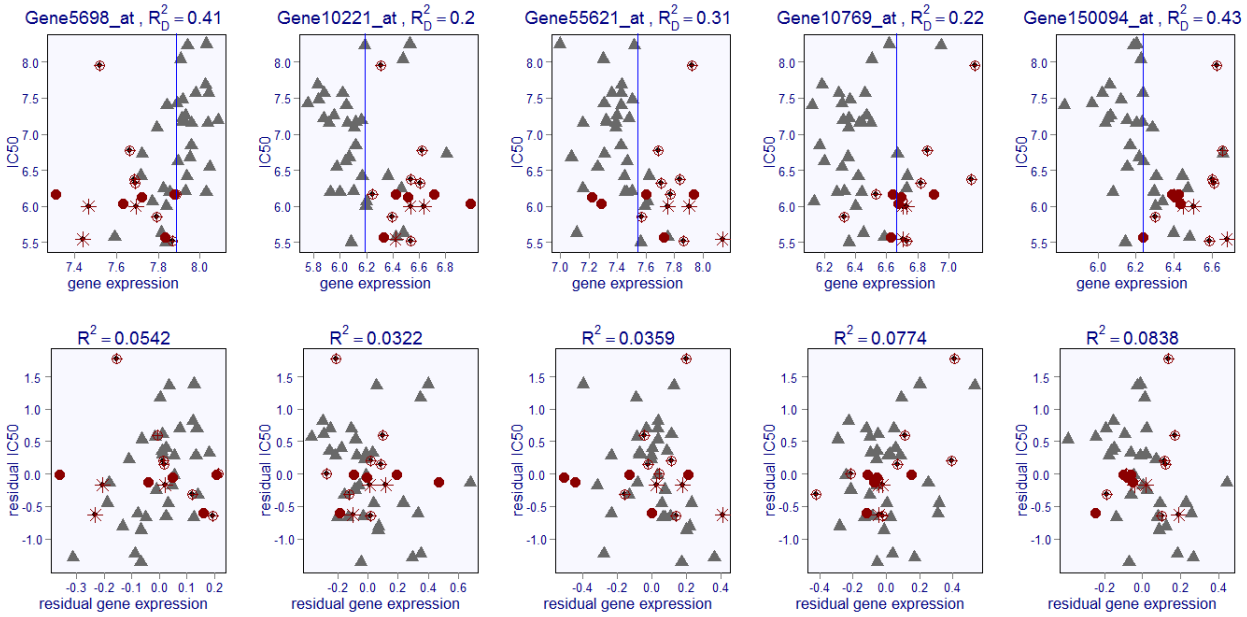

Figure 9: *Scatterplot of gene expression and response (upper panel) and their corresponding residuals from the joint model (lower panel) of the top 5 differentially expressed genes, where the vertical lines are the cut-off values of the regression tree.*

(upper panel) and the linear pattern between the gene expression and the response, remains after adjusting for the treatment effect, i.e. the residuals resulting from the joint model (lower panel). Table 2 lists the results for the top 15 genes with the highest R^2 using both the full data and the leave-one-out cross validation datasets, The R^2 for both datasets are similar which supports the validity of the obtained measures.

4.3 Therapeutic/Prognostic: Type I/II

There are 30 genes classified as either prognostic/therapeutic gene after multiplicity adjustment. The list of the top 10 genes with the highest adjusted association is presented in table 3. Scatter plots are produced for the top 5 genes which are displayed in figure 11. It can be observed that the two treatment groups are quite separated with respect to gene expression and response (upper panel) and their association can be summarized by a straight line (lower panel). To check the influence of outlying case(s), leave-one out cross validation was performed and the R^2_{cv} gives an indication of a reasonable R^2, having similar values. Most of the genes of this type are down-regulated, i.e $\rho < 0$.

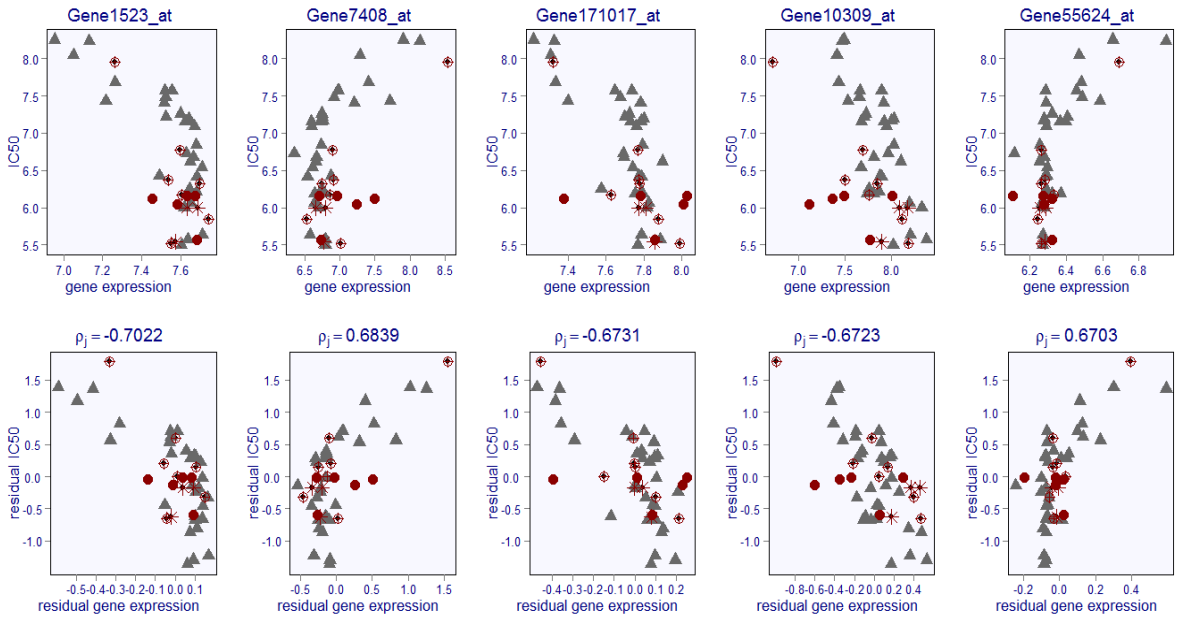

Figure 10: *Scatterplot of gene expression and response (upper panel) and their corresponding residuals from the joint model (lower panel) of the top 5 potential prognostic biomarkers.*

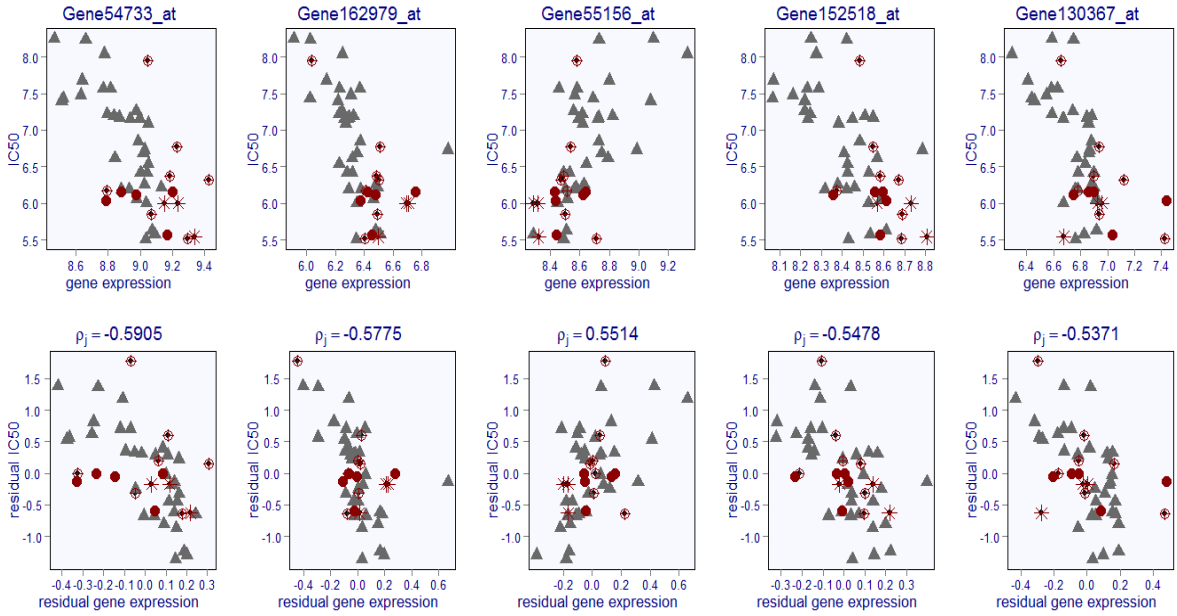

Figure 11: *Scatterplot of gene expression and response (upper panel) and their corresponding residuals from the joint model (lower panel) of the top 5 potential prognostic/therapeutic biomarkers.*

Table 2: *Results of top 15 genes with the highest significant adjusted association, raw p-values (LRT), BH-FDR adjusted p-values, Spearman correlation*

ID	ρ	rs	LRT p-value	Adj-pval	R^2	R^2_{cv}
1523_at	-0.7022	-0.5872	<0.0001	0.00012	0.4931	0.4928
7408_at	0.6839	0.5210	<0.0001	0.00016	0.4677	0.4672
171017_at	-0.6731	-0.6087	<0.0001	0.00016	0.4530	0.4525
10309_at	-0.6723	-0.5888	<0.0001	0.00016	0.4520	0.4658
55624_at	0.6703	0.5574	<0.0001	0.00016	0.4493	0.4496
6242_at	-0.6699	-0.5478	<0.0001	0.00016	0.4487	0.4510
23623_at	-0.6605	-0.5203	<0.0001	0.00021	0.4362	0.4577
23647_at	-0.6587	-0.5796	<0.0001	0.00021	0.4339	0.4343
4900_at	-0.6574	-0.4552	<0.0001	0.00021	0.4322	0.4319
81542_at	0.6523	0.5628	<0.0001	0.00024	0.4256	0.4248
388552_at	-0.6518	-0.4751	<0.0001	0.00024	0.4248	0.4246
57037_at	-0.6488	-0.6105	<0.0001	0.00026	0.4210	0.4237
1594_at	-0.6475	-0.6278	<0.0001	0.00026	0.4193	0.4207
4731_at	-0.6460	-0.5173	<0.0001	0.00026	0.4174	0.4167
57583_at	0.6421	0.4612	<0.0001	0.00030	0.4123	0.4138

4.4 Testing for interaction between group and gene expression: δ_j

The conditional model is extended by adding the interaction between group and gene expression. It is then fitted to each gene and the null hypothesis of no interaction effect ($\delta_j = 0$) is tested. The adjusted association based on the information theory was calculated. Thirty-three genes were found to have significant interaction effects after adjusting for multiplicity at the FDR of 0.05.

The top 5 genes with highest adjusted association are shown in figure 12. It can be readily seen that the significant interaction effect for Gene 8884 is only attained due to the presence of one

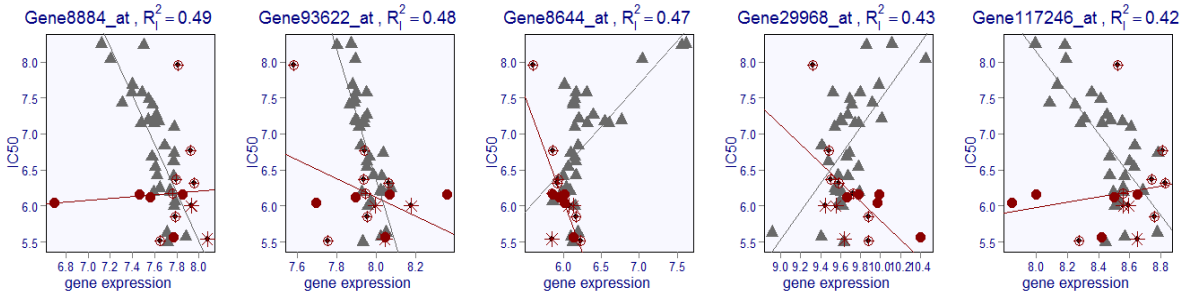

Figure 12: *Scatterplot of gene expression and response displaying interaction effects*

Table 3: *Results of top 15 prognostic/therapeutic biomarkers with the highest significant adjusted association*

ID	α	**t-stat**	**p-value**	**Adj-pval**	ρ	**rs**	R^2	R^2_{cv}
54733_at	0.2298	3.9552	0.0002	0.0088	-0.5905	-0.6490	0.3487	0.3482
162979_at	0.1698	3.0530	0.0030	0.0481	-0.5775	-0.5760	0.3336	0.3387
55156_at	-0.1821	-3.0515	0.0030	0.0482	0.5514	0.5068	0.3041	0.3046
152518_at	0.1998	4.1940	0.0001	0.0051	-0.5478	-0.6195	0.3001	0.3008
130367_at	0.2260	3.8256	0.0002	0.0116	-0.5371	-0.5768	0.2885	0.2887
8139_at	0.2304	3.7375	0.0003	0.0139	-0.5369	-0.4947	0.2883	0.2875
6195_at	0.1626	3.0457	0.0030	0.0486	-0.5176	-0.4265	0.2679	0.2718
10322_at	0.0979	3.3242	0.0013	0.0303	-0.5110	-0.4720	0.2611	0.2630
23108_at	0.1921	3.7717	0.0003	0.0129	-0.4939	-0.4675	0.2439	0.2429
57630_at	0.0996	3.6593	0.0004	0.0157	0.4921	0.4195	0.2422	0.2430
730101_at	-0.1748	-3.2590	0.0016	0.0343	-0.4793	-0.4107	0.2298	0.2308
285958_at	0.3246	3.2385	0.0017	0.0356	-0.4609	-0.5390	0.2124	0.2115
6662_at	0.2715	3.0803	0.0027	0.0460	-0.4578	-0.4236	0.2096	0.2087
54475_at	0.2836	3.5266	0.0007	0.0203	-0.4529	-0.4904	0.2051	0.2044
84800_at	0.4824	3.8359	0.0002	0.0114	-0.4495	-0.4772	0.2020	0.2014

point that influences the line. Table 4 presents the list of the top 15 genes with the highest R^2_h.

4.5 Joint Biomarker Profile

Several genes were identified as individual biomarker in the previous section. However, combining information about expression levels from a number of potential biomarkers into one variable, collectively termed as joint biomarker, might improve the quality of the obtained measures of association. SPCA was conducted for each type of biomarker and prediction of the response is done using the first principal components. The scatter plots of the first principal components against the response is shown in figures 13, 15, and 17.

For the construction of potential joint therapeutic biomarker, the top 2 therapeutic genes were used since it gave the highest R^2_D of 0.435 (Refer to figure 14). This is lower than the highest observed R^2_D among the individual biomarkers of 0.47. Therefore the gain in constructing a joint therapeutic profile is not evident in this case. The splitting of the groups with respect to the gene expression resulted to a lot of misclassified observations, i.e., there is no clear separation of the two treatment groups. Regression Tree approach, in this case, is not a good approach to evaluate a nonlinear relationship between a biomarker and a response.

Table 4: *Results for top 15 Genes that exhibited treatment-gene interaction effects: test-statistic, raw-pvalue, BH-FDR adjusted p-values, measure of association without interaction and adjusted association with interaction*

ID	**t-stat**	**p-value**	**Adj-pval**	$R^2_{hw—o}$	$R^2_{hw—interaction}$
8884_at	5.1758	<0.0001	<0.0001	0.1673	0.4869
93622_at	4.6566	<0.0001	<0.0001	0.2194	0.4811
8644_at	-4.4728	<0.0001	0.0340	0.2186	0.4667
29968_at	-4.9865	<0.0001	<0.0001	0.0934	0.4255
117246_at	4.4164	0.0001	0.0340	0.1598	0.4220
6047_at	4.8110	<0.0000	<0.0001	0.1075	0.4198
26073_at	4.4520	0.0001	0.0340	0.1523	0.4197
284184_at	4.8484	<0.0001	<0.0001	0.0840	0.4078
79230_at	4.5574	<0.0001	<0.0001	0.1151	0.4033
28991_at	4.8875	<0.0001	<0.0001	0.0703	0.4023
57019_at	4.5028	0.0001	0.0340	0.1066	0.3929
4214_at	-4.5335	<0.0001	0.0000	0.0979	0.3896
224_at	-4.2379	0.0001	0.0340	0.1275	0.3846
54807_at	4.4326	0.0001	0.0340	0.0857	0.3724

The number of potential prognostic genes that can together serve as joint prognostic biomarker is determined based on the joint profile that gives the highest R^2. The top 8 genes yielded an R^2 of 0.5995 when used together as a potential joint prognostic profile. This is much better than when using a single gene as biomarker in this case. However, the significance of the association measure of the joint profile is not tested. The same observation holds for the potential/ therapeutic biomarkers, a large improvement in the R^2 is observed when using the top 5 genes as joint biomarker.

5 Discussion and Conclusion

This study aims to identify and evaluate gene-specific biomarkers for IC_{50}, a summary of cell lines activity in a dose-response experiment. Moreover, a joint biomarker is also constructed using information from several genes simultaneously. The selection and evaluation of biomarkers uses the same set of methods that have been devised to validate surrogate endpoints.The analysis in this paper is similar to the case study presented by Tilahun *et al.* (2010) where they also look for potential gene-specific and joint biomarkers for the response. The purpose of finding biomarkers is not just limited to classify microarray samples into groups, but to predict the clinical outcome (Lin *et al.*, 2010). Biomarkers have been classied as prognostic and/or therapeutic

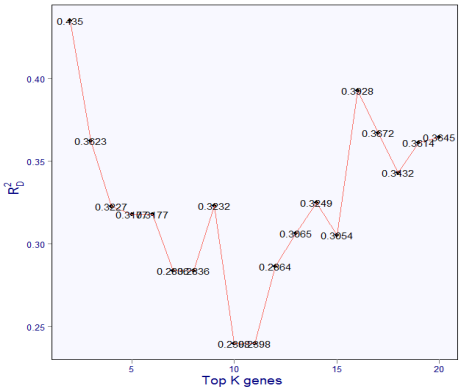

Figure 13: *Plot of R^2_D using the top k genes as potential joint therapeutic biomarkers*

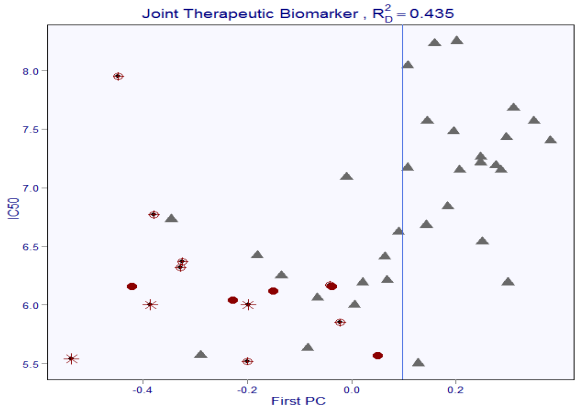

Figure 14: *Joint biomarker profile using the top 2 potential therapeutic biomarkers*

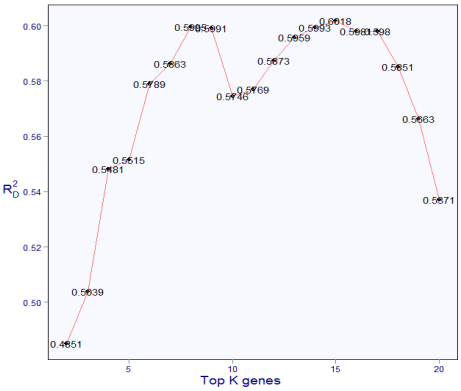

Figure 15: *Plot of R^2 using the top k genes as potential joint prognostic biomarkers*

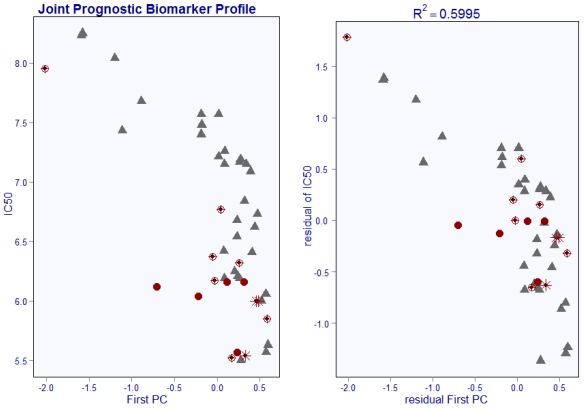

Figure 16: *Joint biomarker profile using the top 8 potential prognostic biomarkers*

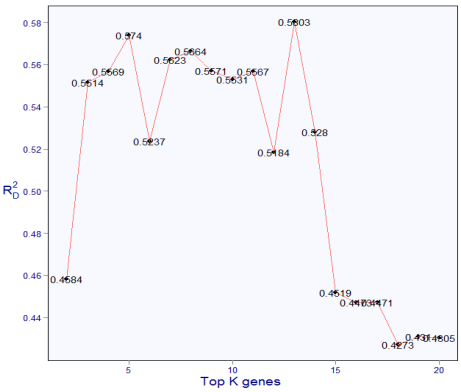

Figure 17: *Plot of R^2 using the top k genes as potential joint therapeutic/prognostic biomarkers*

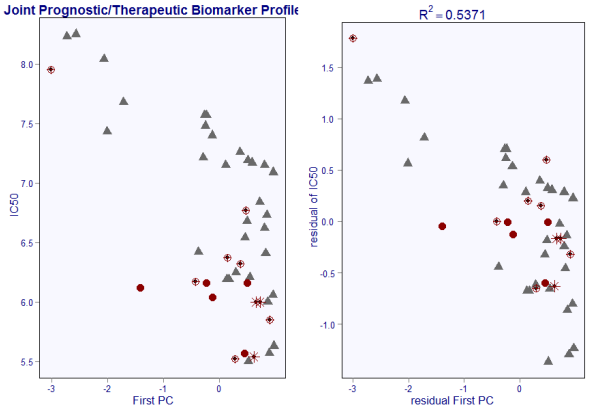

Figure 18: *Joint biomarker profile using the top 5 potential prognostic biomarkers*

depending on their relationship with the clinical endpoint and their response to treatments.

A gene-by-gene analysis was carried out to tests whether a gene can be used as potential prognostic/therapeutic biomarker. The joint modeling for gene expression and response was adopted to facilitate the identification of these two types of biomarkers. For the first type, the gene expression can be used to predict the level of the response through a linear association adjusting for the treatment effect whereas for the second type, a non-linear association with the response is observed but the gene is differentially expressed, given a significant treatment effect upon the response. Here, the treatment effect on the gene-expression can be predictive for the treatment effect on the response.

Of the 7722 genes, 288 and 900 genes can serve as therapeutic and prognostic biomarkers for the response, respectively. Thirty (30) are identified to be potential prognostic/therapeutic genes. All hypothesis involved in the identification is tested using an FDR level of 0.05.

An R^2- type measure of association was used to evaluate the quality of a certain biomarker. For prognostic biomarkers, since linear association is evident between the gene expression and the response after correcting for treatment effect, the adjusted association proposed by Buyse and Molenberghs (1998), that is a widely used measure of association in the surrogate marker literature, is estimated. In this case, where the response is of continuous type and is assumed to follow a normal distribution, the information-theoretical approach (Alonso and Molenberghs, 2007) to model the correlation between the gene expression and the response adjusting for treatment effect provides similar results with the joint modeling approach from where the adjusted association is based. But the information theoretic approach proves to be advantageous since, as outlined by Tilahun *et al.* (2010), it involves less computation time given the number of potential biomarkers available, which amounts to the number of models that need to be fitted. Additionally, it can be also applied to non-normal setting. Nevertheless, the joint model also allows estimating a non-linear association (using Spearmans correlation) between the response and the gene-expression after adjusting for the treatment effects for the prognostic biomarkers. The potential prognostic biomarkers are then ranked based on the adjusted association.

On the other hand, the association with the response of the second type of biomarker can not be captured by the linear association using a linear regression model. One of the possible measures proposed to quantify the amount of information in the response captured by the gene expression is by using the relative deviance reduction, obtained from fitting a regression tree. The splitting of the groups with respect to the gene expression resulted to a lot of misclassified observations, i.e., there is no clear separation of the two treatment groups. Regression Tree approach, in this case, is not a good approach to evaluate the association between differentially expressed genes and a response. Hence, the need to find measures that could optimally capture this possibly non-linear relationship can be another focus of research.

Also, the possible influence of an outlying observation in the measures of association is also a concern. In some cases a reasonably linear relationship appears to be nonlinear due to a few number of outlying observations. It is therefore worthwhile to thoroughly investigate the type of genes that have been selected before making a decision to promote a gene as a possible biomarker.In this case, the leave-one out cross-validation was performed to ensure a reliable estimate of R^2 values.

For the construction of the joint biomarkers, the supervised principal component analysis (SPCA) was employed. Following the approach of Tilahun *et al.* (2010), a joint biomarker consisting of a subset of the top 20 genes are constructed and each is evaluated using the respective measure of association for each type of biomarker. The joint biomarker that maximized the association is then chosen. Further, it was checked whether the magnitude of the measure has increased when using a joint biomarker than the gene-specific biomarker. It was found out that among the measures; only the relative deviance reduction of the potential joint therapeutic biomarker that is composed of the top 2 genes has lower value than the maximum value obtained by a gene-specific therapeutic biomarker. In this case, there is no gain in constructing a joint biomarker. But for the other two sets of biomarkers, the joint biomarkers provide an improvement in the measure of association. The top 8 and top 5 genes are used to construct the joint prognostic and joint prognostic/therapeutic biomarkers, respectively. Take note however that significance testing of the resulting measures was not carried out here. Arguably, the evaluation of the biomarkers can be an important component in the decision making process, but at least equally important is experts' opinion coming in from pharmacological, biological, clinical, economical considerations. In addition, SPCA is not the only proposed method to identify joint biomarkers in the literature. In fact, constructing a joint biomarker profile is still the topic of ongoing research. Moreover, before using these genes as biomarkers, validation procedure should be carried out, either using independent experiments or biological validation.All the analyses presented in this paper are done using the R 2.11.0 software.

As a conclusion, joint modeling of phenotypic variables and gene expression data permits the selection and evaluation of genomic biomarkers in early drug development experiments. It is also the first step in the construction of a joint biomarker. Furthermore, genomic biomarkers can be classified as prognostic and/or therapeutic ones depending on their intended use. Once target genes are identified as potentially good biomarkers, their selection and relevance are still to be validated.

References

[1] Adetayo, K. (2010). Statistical Methods for Affymetrix Microarray Experiments in Early Drug Development Studies: Gene Signatures, Dose-response study, and Probe Level Analysis. [PhD Dissertation,UHasselt,BE].

[2] Alonso, A. and Molenberghs, G. (2007). Surrogate marker evaluation from an information theory perspective. *Biometrics*, **63**, 180186.

[3] Azuaje F. and Dopazo J. (2006). Data Analysis and Visualisation in Genomics and Proteomics. John Wiley and Sons.

[4] Bair, E., Hastie, T., Paul, D., Tibshirani, R. (2006). Prediction by supervised principal components. *J. Am. Stat. Assoc.*,**101**, 119-137.

[5] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Soceity, Biostatistics*, **57**, 289300.

[6] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annal of Statistics*, **29(4)**, 11651188.

[7] Biomarkers Denitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred denitions and conceptual framework. *Clinical Pharmacology and Therapy*, **69**, 89-95.

[8] Bøvelstad H.M. (2007). Predicting survival from microarray data-a comparative study. *Bioinformatics*, **23(16)**, 2080-87.

[9] Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiment. *Biometrics*, **54**, 186-201.

[10] Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (2000) The validation of surrogate endpoints in meta-analysis of randomized experiments. *Biostatistics*, **1**, 49-67.

[11] Drăghici, S. (2003). Data Analysis Tools for DNA Microarrays. Chapman & Hall.

[12] EMEA, 2009 (EMEA/CHMP/ICH/380636/2009). ICH Topic E16. Genomic Biomarkers Related to Drug Response: Context, Structure and Format of Qualification Submissions.

[13] Ge, Y., S. Dudoit, and T. Speed (2003). Resampling-based multiple testing for microarray data analysis. *TEST* **12**,1-77.

[14] Gölhmann, H. and Talloen, W. (2009). Gene expression studies using Aymetrix microarrays. Chapman & Hall.

[15] Kerr, M.K., MArtin, M., Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. *J Comput Biol*, **7(6)**,819-37.

[16] Lehmann, E.L. and Romano, J.P. (2005). Generalizations of the familywise error rate. *Ann. Statist.*, **33(3)**, 1138-1154.

[17] Lin, D., Shkedy, Z., Molenberghs, G., Talloen, W., Gölhmann, H., Bijnens, L. (2010). Selection and evaluation of gene-specific biomarkers in pre-clinical and clinical microarray experiments. *Online Journal of Bioinformatics*, **11(1)**,106-107.

[18] Lin, D., Tilahun, A., Cortinas, J., Shkedy, Z., Molenberghs, G., Gölhmann, H., Talloen, W., Bijnens, L. (2011). Comparison of Methods for the Selection of Genomic Biomarkers. *International Journal of Data Mining and Bioinformatics*[accepted for publication].

[19] Park, J., Kerbel, R., Kellof, G., Barrett, J., Chabner, B., Parkinson D., Peck J., Ruddon, R., Sigman C. and Slamon., D. (2004). Rationale for Biomarkers and Surrogate Endpoints in Mechanism-Driven Oncology Drug Development, *Clinical Cancer Research*, **10**, 3885.

[20] Parmigiani, G., Garrett, R., Irizarry, S., and Zeger, S. (2006). The Analysis of Gene Expression Data: Methods and Software. Springer NY.

[21] Prentice, R.L. (1989). Surrogate endpoints in clinical trials: denitions and operational criteria. *Statistics in Medicine*, **8**, 431440.

[22] Tilahun, A., Lin, D., Shkedy, Z., Geys, H., Alonso, A., Peeters, P., Talloen, W., Drinkenburg, W., Gölhmann, H., Gorden, E., Bijnens, L., Molenberghs, G. (2010). Genomic biomarkers for depression: Feature-specic and joint biomarkers.*Statistics in Biopharmaceutical Research.***2(3)**, 419-434.

[23] Welsch M., Mangravite L., Medina, M., Tantisira, K., Zhang, W., Huang S., McLeod, H., Dolan, M. (2009). Pharmacogenomic Discovery Using Cell-BAsed Models. *Pharmacological Reviews*, **61**, 413-429.

[24] Xu, H. and Hsu, J. C. (2007). Using the Partitioning Principle to control the generalized Family Error Rate. *Biometrical Journal*, **49**, 5267.

[25] Yekutieli, D. and Benjamini, Y. (1999). Resampling-Based False Discovery Rate Controlling Multiple Test Procedures for Correlated Test Statistics. *Journal of Statistical Planning and Inference*, **82**, 171196.

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Joint modeling of phenotipic variables and gene expression data in early drug development experiments

Richting: **Master of Statistics-Biostatistics**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Perualila, Nolen Joy

Datum: **12/09/2011**