Genomic Biomarkers for Depression: Feature-Specific and Joint Biomarkers

Peer-reviewed author version

TILAHUN ESHETE, Abel; LIN, Dan; SHKEDY, Ziv; GEYS, Helena; ALONSO ABAD, Ariel; Peeters, Pieter; Talloen, Willem; Drinkenburg, Wilhelmus; Goehlmann, Hinrich; Gorden, Evian; BIJNENS, Luc & MOLENBERGHS, Geert (2010) Genomic Biomarkers for Depression: Feature-Specific and Joint Biomarkers. In: STATISTICS IN BIOPHARMACEUTICAL RESEARCH, 2(3). p. 419-434.

DOI: 10.1198/sbr.2009.08091
Handle: http://hdl.handle.net/1942/12117

Selection and Evaluation of Gene-specific Biomarkers in Pre-clinical and Clinical Microarray Experiments

Summary

Biomarker discovery has become one of the major drivers of pharmaceutical research and drug development. Over the last five years, microarray experiments have become an increasingly common laboratory tool allowing investigation of the activity of thousands of genes simultaneously (Amaratunga and Cabera 2004). This enables the determination of genomic biomarkers using microarray experiments. In these experiments, some responses are measured indicating the outcome of the treatments. In such situations, the primary question of the study is whether the gene expression can serve as a biomarker for the responses or not. In this paper, we distinguish between two types of biomarkers: in the first type, the association between the gene expression and the response with adjustment for treatment effect can be captured by a straight line, while in the second type, the treatment effect both on the gene expression and the response plays a central role. We propose a joint model for the gene expression and the response, which allows the investigator (1) to detect differentially expressed genes as biomarkers and (2) to identify genes associated with the response.

Keywords: Joint Model; Microarray Experiments; Biomarkers; Differentially Expressed Genes.

1 Introduction

Biomarkers play an increasingly important role in improving the effectiveness of drug research and development in pharmaceutical industries. 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. In this paper, we focus on the microarray setting, in which data are available from a single trial. For each

subject, microarray data (X) together with a (clinical/experimental) response variable (Y) are available under various conditions/treatments (Z).

Several authors have discussed the issue of using gene expression data to predict a specific response. In particular, Nguyen and Rocke (2002) discussed the case, in which gene expression data are used as predictors in a Cox proportional hazard model. Dimension reduction of gene expression data is done by applying the partial least squares method which maximizes the covariance between the response Y and a linear combination of the gene expression data. Tan *et al.* (2006) similarly used the partial least squares method, for data reduction in the context of cytotoxicity experiments. More recently, Bair *et al.* (2006) used supervised principal components analysis to predict a survival response using gene expression as predictors. The analysis presented in this paper aims at finding a subset of genes, that are associated with the response and can be used as biomarkers, which follows similar lines as the one presented in the the case studies of Nguyen and Rocke (2002) and Bair *et al.* (2006). However, we should take into account that both gene expression and the response are influenced by the treatment that was administrated to the subjects before and after the experiment.

In what follows, we focus on continuous responses and we employ two case studies to illustrate two types of genes, that can serve as biomarker for the response. The first one was carried out in a clinical trial with 19 depressed patients followed up by anti-depressant treatments; the Hamilton Depression (HAMD) score of each patient was recorded as the primary clinical outcome. The second study involved 24 rats in two treatment groups and the response of primary of interest is the distance traveled by the rat on the experimental surface. In both examples, we address two questions of primary interest (1) which genes are differentially expressed (i.e., the treatment has a significant effect on the gene expression) and (2) which genes are associated with the response (i.e., whether the investigator can draw any conclusions about the distance traveled by the rats and the HAMD scores, based on the gene expression data).

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 pre-clinical and clinical microarray experiments is given in Section 3.1. A graphical interpretation of the relationship between biomarker genes and the response is given in Section 3.2. We distinguish between two types of relationship: in the first type (*prognostic biomarker*) the correlation between the

gene expression and the response can be summarized by a straight line, while in the second type (*therapeutic biomarker*) the focus is on the effect of treatments on the response and the gene expression. As the gene-specific joint model for the gene expression and the response of primary interest is fitted, the adjusted association, proposed by Buyse and Molenberghs (1998), which can be derived from the covariance matrix of the error component, is used to test for the first type of association. Moreover, for the case that the association between gene expression and the response is not linear, we propose two new measures, (1) the relative deviance reduction, which is used, in analogy with the adjusted association, to evaluate the quality of a second type of biomarker with therapeutic treatment effect and (2) R^2 by using the support vector regression methodology in Section 3.3. In Section 4, the results obtained using the methods above to the case studies are presented. The paper ends with some discussions and conclusions in Section 5.

2 Data

2.1 Case Study I: Clinical Study of Depression

The first of two case studies is a clinical depression study involving one hundred participants (66% females) with major depression (50 from Sydney, New South Wales and 50 from Adelaide, South Australia). Participants have been referred from the practicing general practitioners and psychiatrists. In addition, 31 patients have been followed up 4-6 weeks after commencement of treatment with antidepressants. However, out of the 31 depressed patients which had measurements after treatment, there are only 19 patients available for both the gene expression and HAMD score. There were a total of 17,502 genes measured for each patient using microarray Affymetrix chips. In addition to the gene expression, storage time of the samples, age, gender, and season when the samples were collected and wether or not the subjects fasted were recorded for each patient. The response of primary interest is the change from the baseline HAMD score.

Figure 1*a* shows the change in HAMD score before and after the anti-depressant treatments. In Figure 1*b* we have genes with strong association with the HAMD score after correcting for the different confounding effects as can be seen from the linear pattern in the plot. Figure 1*c* gives an example of a gene that has shown weak association with the response as the points

form a cloud with no apparent pattern. As stated earlier the objective is to find genes which have strong association irrespective of the effect of treatment and other confounding variables.

Figure 1 about here.

2.2 Case Study II: Behavioral Experiment of Distance Travel by Rats

The second case study was obtained from a behavioral experiment, in which 24 male, experimentally naive Long-Evans rats obtained from Janvier (France), weighing 300–370 g at the start of the experiment were randomized into two treatment groups (12 rats in each group). Quinpirole hydrochloride (Sigma-Aldrich) was dissolved in physiological saline and administered at a dose of 0.5 mg/kg administered s.c. (the method used by Szechtman *et al.* 1998). Equivalent volumes of saline were used in solvent injections. Animals were tested in a large open field. Two data sets also encompassed behavioral data; in addition, microarray data were collected. Rat behavior data parameters, suggested by Szechtman *et al.* for the definition of compulsive checking, were recorded systematically. In particular, the parameter of primary interest is defined as the distance traveled by the rats. The active response to the treatment doses is expected to increase this distance. After dose administration, rat microarray chips were taken through cutting the rat's brain into the frontal, striatum, and thalamus parts. Thus, there are three microarray chips for each animal, and thousands of gene expressions are measured within each chip. Each chip measured the expression levels of 5644 genes for each rat. All microarray related steps, including the amplification of total RNAs, labeling, hybridization, and scanning were carried out as described in the GeneChip Expression Analysis Technical Manual, Rev.4 (Affymetrix 2004). For the illustration purpose, we use microarray data of the thalamus part for analysis.

Figure 2*a* shows the boxplot for the total distance traveled by rats in each treatment group, Figure 2*b* shows a boxplot of gene expression (for gene 345), and Figure 2*c*, a scatterplot for the response and the gene expression (for gene 345), where the response, distance traveled by rats under two treatment conditions, is statistically significant ($p < 0.0001$), as obtained from a two-sided t-test. Note that, after adjusting for the treatment effect, the association

between the residual response and residual gene expression does not seem to be linear in Figure 2*d*. This is in contrast with the pattern revealed by gene 331, shown in Figure 2*e*, whose association seems to be linear in the scatterplot after adjusting for the treatment effect, as shown in Figure 2*f*.

Figure 2 about here.

3 Methods

3.1 A Gene-specific Joint Model

In this section, we discuss a joint model for the gene expression and the response, which allows us to test the two questions of primary interest, namely, which gene is differentially expressed and which gene can serve as a biomarker. Following Buyse *et al.* (2000), we define a gene-specific joint model, in which the linear predictors of the response and the gene expression are given by

$$\begin{array}{rcl} E(X_{ij}|Z_i) & = & Z_i\alpha_j, \quad j = 1, \ldots, m \; ; \;\; i = 1, \ldots, n, \\ E(Y_i|Z_i) & = & Z_i\beta. \end{array} \tag{1}$$

Here, Z_i is a known design matrix of intercept and treatments, α_j is a gene-specific parameter vector for gene j, and β is the parameter denoting the treatment effect upon the response.

Note that (1) is a gene-specific model and, in practice, is fitted for each gene separately, a procedure often termed "gene-by-gene" analysis. It is further assumed that the two outcomes are normally distributed:

$$\begin{pmatrix} X_{ij} \\ Y_i \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} Z_i\alpha_j \\ Z_i\beta \end{pmatrix}, \Sigma_j = \begin{pmatrix} \sigma_{_{jj}} & \sigma_{_{jY}} \\ \sigma_{_{jY}} & \sigma_{_{YY}} \end{pmatrix} \right). \tag{2}$$

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 of gene-specific joint model (2):

$$\rho_j = \frac{\sigma_{_{jY}}}{\sqrt{\sigma_{_{jj}}\sigma_{_{YY}}}}. \tag{3}$$

Indeed, $\rho_j = 1$ indicates a deterministic relationship between the gene expression and the response, in the sense that, given gene expression, a perfect prediction of the HAMD score

is possible. Note that, ρ_j can be equal to 1 even if the gene is not differentially expressed as we will illustrate in the next section. For the special case that the only covariate included in the model is a treatment variable, as in the behavioral experiment, joint model (1) can be rewritten as

$$\begin{array}{rcl} E(X_{ij}|Z_i) & = & \mu_j + \alpha_j Z_i, \quad j = 1, \ldots, m\,; \;\; i = 1, \ldots, n, \\ E(Y_i|Z_i) & = & \mu_Y + \beta Z_i. \end{array} \tag{4}$$

Here, α_j and β are the gene-specific and the outcome treatment effects, respectively. Conditioning on a specific gene, (4) is equivalent to the surrogacy model for a single trial, proposed by Burzykowski *et al.* (2005).

Alonso and Molenberghs (2007) derived an appealing expression, generalizing R^2, as will be described next. In the continuous-outcome case, it derives from the models:

$$E(Y_i) = Z_i\beta, \tag{5}$$

$$E(Y_i|X_{ij}) = Z_i\beta + \gamma_j X_{ij}. \tag{6}$$

Here, γ_j is the gene-specific effect upon the outcome. Note that the gene specific model (6) is similar to the underlying model presented in Bair *et al.* (2006) and it is implied by the joint model. For the special case that the only covariate in the model is the treatment, the joint distribution of X and Y and the mean structure in (6) implies the following conditional model:

$$Y_i|Z_i, X_{ij} \sim N(\delta_0 + \delta_1 Z_i + \delta_2 X_{ij}, \sigma^2)$$

where $\delta_0 = \mu_Y - \sigma_{jY}\sigma_{jj}^{-1}\mu_j$, $\delta_1 = \beta - \sigma_{jY}\sigma_{jj}^{-1}\alpha_j$, $\delta_2 = \sigma_{jY}\sigma_{jj}^{-1}$ and $\sigma^2 = \sigma_{YY} - \sigma_{jY}^2\sigma_{jj}^{-1}$.

Further, Alonso and Molenberghs (2007) and Tilahun *et al.* (2009) showed that R_h^2 can be obtained from

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

where G^2 denotes the likelihood ratio statistics to compare models (5) and (6), and n is the sample size.

3.2 Graphical Interpretation of Association Between Biomarker and Response

Figure 3 shows scatter plots of three hypothetical examples of the relationship between gene-expression (X) and response (Y). Circles represent values for one treatment group and pluses depict measurements for the other treatment group. In all examples, the treatment effect upon response is significant. The upper three panels present the scatter plot of gene-expression versus the response, while the lower three panels show the scatter plot of the residuals (after adjusting for treatment effects) for both response and gene-expression. In panel a, the gene is not differentially expressed, but there is a linear association of gene-expression with response. Note that the linear pattern remains after adjusting for the treatment effect, as shown in panel d. We term a gene with this pattern a *prognostic biomarker*. Panel b shows an example in which the gene is differentially expressed, the two treatment groups are clearly separated, but the association between gene-expression and response does not have a linear appearance, which can be seen also in panel e. We term a gene with this type of relationship a *therapeutic biomarker*. Panel c shows a combination of both preceding patterns. A gene is differentially expressed and the treatment effect upon response is significant, the two treatment groups are clearly separated with respect to gene-expression and response, and the association between gene-expression and response can be summarized by a straight line. This can also be seen from panel f, which shows the same example after adjusting for treatment effects. We term a gene with this type of relationship a *prognostic/therapeutic biomarker*. It is expected that the adjusted association will capture the linear association in panel c and f, but will not apply to the *therapeutic biomarker* in panels b and d.

Figure 3 about here.

3.3 Inference and Evaluation of Association Measures for Both Types of Biomarkers

3.3.1 Adjusted Association for Prognostic Biomarkers

Within the microarray setting, we test whether or not a gene can then serve as a *prognostic biomarker*, which can be used to predict the response. Thus, one needs to test the hypotheses

$$\begin{aligned} H_{0j}^{B} &: \rho_j = 0, \\ H_{1j}^{B} &: \rho_j \neq 0, \end{aligned} \tag{8}$$

where ρ_j is defined in (3). A gene is declared an up-regulated *prognostic biomarker* if the null hypothesis in (8) is rejected and $\widehat{\rho}_j > 0$, and a down-regulated *prognostic biomarker* when $\widehat{\rho}_j < 0$.

A prognostic biomarker, where ρ_j is found to be significant, can be evaluated using an estimate $\widehat{\rho}_j$, which measures the linear association between gene-expression and response, after accounting for treatment effects. It is the square root of the R^2-measure based on the joint model in (4).

3.3.2 R^2-type Measures for Therapeutical Biomarkers

A typical analysis of DNA microarrays allows monitoring expression levels of thousands of genes simultaneously, and identifying differentially expressed genes. This type of genes has potential to be identified as *therapeutic biomarkers*, for which the treatment effect on the gene-expression can be predictive for the treatment effect on the response. For this purpose, we test which genes are differentially expressed using (4). Hence, for each gene, we test the hypotheses

$$\begin{aligned} H_{0j}^A &: \alpha_j = 0, \\ H_{1j}^A &: \alpha_j \neq 0. \end{aligned} \tag{9}$$

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

The R^2 measure based on the linear regression models discussed in the previous section is used to quantify the level of association between the response and the gene expression for prognostic biomarker. In what follows, we use different predictive models, namely, the regression tree and support vector regression in order to evaluate the quality of a candidate biomarker in the case that the association can not be summarized by a straight line.

3.3.3 Evaluation Using Tree-based Method

In order to evaluate the quality of therapeutic biomarkers, the adjusted association ρ_j is not applicable. Thus, we follow the information-theory approach of Alonso and Molenberghs (2007) and propose a measure for therapeutic biomarker, i.e., the relative deviance reduction. The total variability of the response, the deviance, without any information about the gene-expression level can be measured by

$$D(Y) = \sum_{i=1}^{n} (Y_i - \widehat{\mu})^2 , \tag{10}$$

where $\widehat{\mu} = 1/n \sum_{i=1}^{n} Y_i$ and $i = 1, \ldots, n$ indexes the arrays. For a therapeutic biomarker, because gene-expression is differentially expressed, one can use the gene-expression level to predict the response level. While a linear regression model is not an appropriate model for this type of biomarker, a regression tree model (Venables and Ripley 1994), in which the gene-expression is the only predictor, can capture the structure of the data, as shown in Figure 4.

Figure 4 about here.

Moreover, because the gene is differentially expressed, we can restrict the tree 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 4. Let k denote the number of terminal nodes in the tree and let $D(Y|X, k = 2)$ denote the sum of deviances for the terminal nodes,

$$\begin{aligned} D(Y|X, k = 2) &= D_1(Y|X) + D_2(Y|X) \\ &= \textstyle\sum_{Y_i \in k_1} (Y_i - \widehat{\mu}_1)^2 + \sum_{Y_i \in k_2} (Y_i - \widehat{\mu}_2)^2 , \end{aligned} \tag{11}$$

where $D_1(Y|X)$ and $D_2(Y|X)$ denote the deviance in each of the terminal nodes, k_1 and k_2 denote the sets of subject indices corresponding to the two terminal nodes, and $\widehat{\mu}_1$ and $\widehat{\mu}_2$ are the mean response in the two terminal nodes. The deviance reduction, $D(Y) - D(Y|X, k = 2)$, 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^2_D, is given by

$$\begin{aligned} R^2_D &= \frac{D(Y) - D(Y|X)}{D(Y)} = \frac{D(Y) - D_1(Y|X) - D_2(Y|X)}{D(Y)}, \\ &= \frac{\sum_{i=1}^{n} (Y_i - \widehat{\mu})^2 - \left[\sum_{Y_i \in k_1} (Y_i - \widehat{\mu}_1)^2 + \sum_{Y_i \in k_2} (Y_i - \widehat{\mu}_2)^2\right]}{\sum_{i=1}^{n} (Y_i - \widehat{\mu})^2}. \end{aligned} \tag{12}$$

Similar to R^2 in linear regression, R^2_D measures the proportion of variability explained by the regression tree model. It is easy to see that R^2_D, as a measure of association, is equivalent to the variance reduction factor discussed by Alonso *et al.* (2003). Moreover, in the case where the model has two terminal nodes and we fit a regression model $Y_i = \beta_0 + \beta_1 X_{ij}$, one can easily see that the R^2 of the regression model and the R^2_D of the regression tree are equal (i.e., $R^2 = R^2_D$) for gene j.

Following Alonso and Molenberghs' (2007) information-theory approach, it is easy to see that both R^2_{indiv} (i.e., the square of ρ (3)) and R^2_D belong to the family of information-theoretic association measures. This is a crucial point, as it implies that, although prognostic and therapeutic biomarkers are evaluated using different validity measures (i.e., ρ and R^2_D, respectively), both measures can be interpreted in the same way. Both R^2_{indiv} and R^2_D measure the proportion of information in the response captured by using the gene expression.

3.3.4 Evaluation Using Support Vector Regression

The term support vector machines (SVM) refers to a family of learning algorithms that is nowadays considered as one of the most efficient methods throughout a variety of applications. In particular, in regression and time-series prediction applications, excellent performance has been obtained (Drucker *et al.*, 1997; Müller *et al.*, 1997; Stitson *et al.*, 1999; Mattera and Haykin, 1999). SVM is a supervised learning technique for classification and regression. The SVM algorithm is a non-linear generalization of the so-called *generalized portrait algorithm* developed in the sixties by Vapnik and Lerner (1963) and Vapnik and Chervonenkis (1964), but the first practical implementation was only published in the early nineties. Ever since, the popularity of the method has been growing among the machine learning and statistical communities. SVM can also be applied to regression problems by the introduction of an alternative loss function (Smola, 1996). The loss function must be modified to include a distance measure. SVM regression uses the ε-insensitive loss function.

If the deviation between the predicted and actual values is less than ε, then the regression function is considered good, which can be mathematically expressed as: $-\varepsilon \leq \omega \cdot X_{ij} - b - Y_i$. From a geometric point of view, it can be seen as a band of size 2ε around the hypothesis function and any point outside this band is considered a training error. Suppose the outcome can be explained by a linear model; the goal is to find a fitting hyperplane $\langle \omega, X_{ij} \rangle + b = 0$. Formally, we need to minimize $\|\omega\|^2/2$, subject to the constraints $Y_i - \langle \omega, X_{ij} \rangle - b \leq \varepsilon$ and $\langle \omega, X_{ij} \rangle - Y_i \geq \varepsilon$. To account for training errors and the possibility of handling non-linearity, we can map the input data X_{ij} into a, possibly higher-dimensional, so-called feature space $\Phi(X_{ij})$ and introduce some weights into our optimization problem, which now becomes:

$$\min \frac{\|\omega\|^2}{2} + \mathrm{C} \cdot \sum_{i}^{N} (\xi_i + \widehat{\xi}_i),$$

subject to the constraints:

$$\begin{aligned} Y_i - \langle \omega, \Phi(X_{ij}) \rangle - b &\leq \varepsilon + \xi_i, \\ \langle \omega, \Phi(X_{ij}) \rangle - Y_i &\geq \varepsilon + \xi_i, \\ \xi_i, \widehat{\xi}_i &\geq 0. \end{aligned}$$

We thus need to solve a constrained optimization problem. It turns out that, in most cases, it can be solved more easily in its dual formulation. Moreover, the dual formulation provides the key for extending SVM to non-linear functions. Hence, we will use a standard dualization method using Lagrange multipliers, as described in Fletcher (1989). For more details, we refer to Vapnik (1995). Several kernels can be used. We focus on: (1) polynomial: $(\gamma(\langle X_{ij}, X_{kj} \rangle + \delta)^d$; (2) radial basic function (RBF): $exp(\gamma \|X_{ij}, X_{kj}\|^2)$; and (3) sigmoid: $tanh(\gamma(\langle X_{ij}, X_{kj} \rangle + \delta)$.

To select a kernel, all three kernels were tuned using cross-validation for a subset of genes with high R_D^2, and finally, the kernel, together with the set of parameters that produces the smallest mean squared error is retained for all the genes. In this way, we controlled for the risk of over-fitting, given that the set of parameters used to obtain the final model are selected using a cross-validation procedure. We then go on to evaluate the model performance for each of the observations left out in the cross-validated samples and thus the ability of the model to generalize beyond the fitting data. In this paper, as Hsu *et al.* (2001) pointed out, our choice is for the RBF kernel, which can handle the non-linear mapping and has few

parameters to be controlled (C between 0.25 and 6, with step of 0.25 and γ between 0.5 and 50 with step of 0.5). The parameters C and γ obtained from the tuning process were then used to estimate the measure of association. Similar to the case of regression trees, the association measure can be computed using the ratio between the portion of the variability not explained by the model and the total variability of the residuals from the response:

$$R^2_{SVM} = \frac{D(Y) - DSVMR(Y \mid X)}{D(Y)}.$$

$D(Y)$ can be calculated as in (12), and $DSVMR(Y \mid X)$ is the sum of the squares of the differences between the actual value (Y_i) and their estimated value obtained when the SVM regression model is employed.

4 Application to the Data

4.1 Case Study I: Clinical Study of Depression

As mentioned in Section 2, 17,502 genes and the HAMD scores were measured before and after the anti-depressant treatment for the 19 patients. Now, denoting the difference in HAMD score before and after treatment by Y_i, and the gene differences by X_{ij}, the adjusted association ρ and R^2 values were computed for all genes after correcting for some other variables, such as storage time, gender, and age of the patient using the models discussed in the preceding sections. These values were also obtained from leave-one-out cross validation data. Note, however, because all patients are treated, that there is no need to adjust for treatment effect given that the treatment effect is accounted for by taking the difference from the baseline. The results are summarized in Table 1 and Figure 5. Table 1 lists the results for the top 20 genes with the highest R^2 (i.e., the square of the adjusted association ρ) using both the full data set and with 19 leave-one-out cross validation data sets, and their raw p-values obtained from permutations, and BH-FDR adjusted p-values. After the adjustment for multiplicity at the FDR of 0.05, two genes remain as significant, which are potential prognostic biomarkers. Figure 5 depicts the scatterplot of the residuals from the top two genes and HAMD score. It is possible to observe that there appears to be a linear association between the HAMD score and the gene after adjusting for treatment and other covariates. For a detailed analysis of the clinical depression study, we refer to Tilahun *et al.* (2009).

However, there are no genes found to be differentially expressed after multiplicity adjustment (BH-FDR procedure; Benjamini and Hochberg, 1995). For an illustration of genes with significant treatment effect, we refer to the second case study, i.e., the animal behavior experiment.

Table 1 about here.

Figure 5 about here.

4.2 Case Study II: Animal Behavior Study

The gene-specific joint model (4) is fitted to all genes and the likelihood ratio (LR) test is performed to test the null hypothesis $H_0 : \rho_j = 0$. Figure 6 shows the top two genes with highest adjusted associations. Table 2 presents the list of the top ten genes. After adjusting for multiplicity controlling the FDR at 0.05 (using the BH-FDR procedure; Benjamini and Hochberg, 1995), none of the null hypotheses is rejected.

Figure 6 about here.

Table 2 about here.

Next, we test the null hypothesis of $H_0 : \alpha_j = 0$ for 5644 genes, among them 14 genes are found to be differentially expressed at the FDR level of 0.05. Table 3 gives the list of 14 genes with their treatment effects α, the test statistics, the p-values and their BH-FDR adjusted p-values, as well the R^2_D using the regression tree and R^2_{SVM} using the support vector regression.

Table 3 about here.

Figure 7 about here.

Figure 7 shows a spectral map (i.e., a biplot of principal component analysis with special weights on genes, Wouters *et al.* ()) of the top 14 differentially expressed genes. The samples

(in squares) from the two treatment groups seem to cluster at the left (treatment group) and right side (control group) of the plot, respectively. Genes (in circles) appearing at the left hand side are down-regulated, while genes shown at the right hand side are up-regulated. This result confirms the conclusion obtained from Table 3.

However, due to the small sample size, the normality assumption of test for α in model (6) can be problematic in this setting. Therefore, a more robust test, namely, Kolmogorov-Smirnov (K-S) tests for two sample problem can be used instead. As a result, the number of differentially expressed genes reduces to eight, with p-values given in Table 3. In Table 3, we also present the results of R^2_D and R^2_{SVM} using the leave-one-out cross-validation data. It is easy to note that the R^2_D and R^2_{SVM} decrease for all these 14 genes. In particular the decrease in R^2 is more substantial for genes which are not found significant by K-S test. This may indicate some violation of assumption in the distribution or outlying samples inducing the treatment effect, or small variability of expression levels in either of the treatment groups. They are examined in the following paragraphs.

Figure 8 shows the R^2_D vs. R^2_{SVM} without (panel a) and with leave-one-out cross validation (panel b). The 14 differentially expressed genes identified by the significance test of coefficients are indicated by pluses (differentially expressed by using the K-S test) and stars (non-differentially expressed by using the K-S test). We can observed that among 14 genes, five genes (shown as stars in panel b) have their R^2_D and R^2_{SVM} decrease below 0.4 after cross validation correction. These five genes are no longer called differentially expressed, by using Kolmogorov-Smirnov tests. It is important to note that, in our dataset, as most of genes have no association between the gene expression and response induced by the treatment effects, it is reasonable to observe in Figure 8 that the R^2 values of a large number of genes shrinks towards to zero. Thus, this shrinkage effect of cross validation enhances the credibility of differentially expressed genes as potential therapeutic biomarkers.

We also notice that R^2_D and R^2_{SVM} are not exactly in agreement. This is because the regression tree minimizes the reduction in variance by using one cut-off value in the gene expression, while the support vector machine maximizes the distance between two groups of subjects through a kernel function. Nevertheless, both methods identify most of potential therapeutic biomarker genes in the list of Table 3. Moreover, R^2_D and R^2_{SVM} seem to be in agreement for genes with relatively high values of R^2_D and R^2_{SVM}. Note that for this group of

genes, the regression tree tends to produces slightly better values of association as compared to the SVM ($R^2_{SVM} \leq R^2_D$). This is could be due to the fact that some therapeutic genes with large treatment effects on gene expression and the response lacks of functional relationship to be described by the SVM.

Figure 8 about here.

Figure 9 shows the scatter plots of treatment effects upon the gene expression versus the R^2_D using the regression tree and the R^2_{SVM} using the SVM regression with leave-one-out cross validation data, respectively. It is easy to note that genes with significant treatment effects by using the K-S test (in pluses) are found to have high values of both the R^2_D and R^2_{SVM} and relatively large treatment effects. Genes in stars are no longer differentially expressed by using the K-S test and show small treatment effects.

Figure 9 about here.

Figure 10 shows examples of the top two genes with the highest R^2_D by using the regression tree, where the vertical lines are the cut-off values of the regression tree. Genes in the first row are identified as differentially expressed by K-S test, listed in Table 3. Two genes in the second row are no longer significant using K-S test. Due to the variability in the data, the cross-validation data yield poor results for the prediction of the response using the gene expression. However, there are some genes with small treatment effects, for examples two genes shown in the third row of Figure 10, which seem to have high R^2_D.

Figure 10 about here.

Figure 11 shows examples of the top two genes with the highest R^2_{SVM} by using the support vector regression with the radial kernel. Genes in the first row are identified as differentially expressed by K-S test, as listed in Table 3. The two genes in the second row are no longer significant using K-S test,thus yield low values of R^2_{SVM}. The variability in the gene expression is large and causes substantial decrease in R^2_{SVM} using the cross validation correction. There are some genes, as we can observe from the plots of gene 2105 and 2104 (in the third row of Figure 11), which show high value of R^2_{SVM}.

In order to overcome the fitting with some outlying sample in our case study with small sample sizes, we use the cross-validation correction with the regression tree and support vector regression. As a result, the prediction of the response using these two approaches achieves reasonable good results.

Figure 11 about here.

5 Conclusion and Discussion

In this paper, we have discussed methods for identifying biomarkers in drug-discovery microarray experiments. The applied approach focused on modeling the association between gene-expression and response, after adjusting for the treatment effect. The purpose of finding biomarkers is not just limited to classify microarray samples into groups, but to predict the clinical outcome, either continuous, categorical, or of survival type.

Because the methods for this analysis vary between type of the clinical outcomes/responses, we have considered the response to be continuous and defined two types of biomarkers, namely, prognostic and therapeutic 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 or other potential confounding variables; and hence they were evaluated using the adjusted association proposed by Buyse and Molenberghs (1998). Using the clinical depression data, we have identified two prognostic biomarkers with significant adjusted association after multiplicity adjustment.

On the other hand, the second type of biomarker can not be captured by the linear association using a linear regression model. We proposed two types of R^2 measures to quantify the amount of information in the response captured by the gene expression using the regression tree and support vector regression. These two methods were similar in the sense that they both made attempts at classifying subjects into two groups by maximizing/minimizing a certain gain/loss function, and they both succeeded in finding genes with large treatment effects to yield high values of R^2. However, support vector regression is more flexible than the regression tree. Because it can accommodate other type of associations. It is also unadvantageous to some genes with large treatment but lack of functional relationship with

the response. In the animal behavior study, we evaluated eight therapeutic biomarkers using the regression tree and support vector machine regressions after using K-S test. The R^2_D yields slightly higher values than R^2_{SVM} for the eight differentially expressed genes. More importantly, the validity of these potential genes needs to be confirmed further biologically.

For both types of biomarkers, we have discussed the testing of biomarkers using the joint modeling approach and evaluation of biomarkers using R^2-type measures. Especially for the therapeutic biomarkers, we have found that the differentially expressed genes generally lead to high R^2 for evaluation, while the non-differentially expressed genes are sometimes possible to show high R^2, because of small variability in the gene expression and/or some outlying samples in the context of small sampling size study. In this case, the cross-validation correction is needed to ensure a good estimate of R^2 values. However, before using these genes as biomarkers, validation procedure should be carried out, either using independent experiments or biological validation.

In this paper, we have only considered methods for selecting individual biomarkers. Constructing a joint biomarker profile is the topic of ongoing research.

Tables and Figures

Table 1: *Clinical study of depression: results for top 20 Genes.* R^2*: adjusted association measure based on the full data; and* R^2_{cv}*: adjusted association using leave-one-out cross validation; Raw-p: Raw p-values obtained using permutations; Adj-p: BH-FDR adjusted p-values.*

Gene ID	R^2	R^2_{cv}	Raw-p	Adj-p
736	0.75799	0.75417	< 0.0001	0.0365
2419	0.72954	0.72432	< 0.0001	0.0426
3455	0.65364	0.64771	< 0.0001	0.1553
9859	0.65074	0.64600	< 0.0001	0.1553
8427	0.59101	0.59063	0.0001	0.3142
1954	0.58815	0.58298	0.0001	0.3142
13988	0.57995	0.57829	0.0002	0.3142
6342	0.57866	0.57280	0.0002	0.3142
6119	0.57713	0.57238	0.0002	0.3142
16073	0.56329	0.55755	0.0002	0.3142
16501	0.54474	0.53802	0.0003	0.3142
16415	0.53941	0.53452	0.0003	0.3142
5543	0.53819	0.53281	0.0003	0.3142
14657	0.53768	0.53557	0.0003	0.3142
9635	0.52764	0.52479	0.0004	0.3142
6195	0.52365	0.51872	0.0004	0.3142
4900	0.51947	0.51930	0.0005	0.3142
12791	0.51780	0.51261	0.0005	0.3142
15294	0.51463	0.51233	0.0005	0.3142
4375	0.50908	0.50180	0.0006	0.3142

Table 2: *Animal behavioral experiment: results of top ten genes with highest adjusted association, raw p-values from permutations and BH-FDR adjusted p-values.*

Gene	ρ	Raw-p	Adj p-value (BH-FDR)
4955	-0.7493	0.0005	0.1129
2841	-0.7346	0.0007	0.1129
4909	-0.7218	0.0009	0.1129
1785	0.7161	0.0009	0.1129
331	0.7028	0.0009	0.1129
2596	-0.7019	0.0009	0.1129
1844	0.6999	0.0009	0.1129
3200	0.6972	0.0010	0.1129
1796	0.697	0.0010	0.1129
554	-0.6947	0.0015	0.1129

Table 3: *Animal behavioral experiment: results of 14 genes with their significant treatment effects α, the test statistics, the p-values and their BH-FDR adjusted p-values, K-S BH-FDR adjusted p-values, as well the R^2_D and $R^2_D - cv$ with leave-one-out cross validation using the regression tree and R^2_{SVM} and $R^2_{SVM} - cv$ with leave-one-out cross validation using the support vector regression.*

ID	α	t stat	p-value	adj p-val	K-S p-val	R^2_D	$R^2_D - cv$	R^2_{SVM}	$R^2_{SVM} - cv$
1962	-4.3425	-11.017	0.0000	0.0000	0.0125	0.7565	0.4446	0.6607	0.5224
345	-3.3194	-10.248	0.0000	0.0000	0.0021	0.7565	0.4829	0.6115	0.4519
4447	-0.9	-9.041	0.0000	0.0000	0.0021	0.7565	0.6318	0.7309	0.6574
60	-3.8216	-8.06	0.0000	0.0000	0.0125	0.5548	0.4620	0.5508	0.4058
662	-0.7587	-8.049	0.0000	0.0000	0.0125	0.6441	0.3417	0.6176	0.4071
486	-2.415	-7.891	0.0000	0.0000	0.0125	0.728	0.2428	0.5369	0.4513
59	-2.429	-6.957	0.0000	0.0002	0.0125	0.5548	0.4620	0.4896	0.3739
214	-0.8303	-6.416	0.0000	0.0007	0.0125	0.5548	0.4620	0.4019	0.2541
5216	0.6936	6.348	0.0000	0.0008	0.0720	0.5578	0.1705	0.6092	0.3465
2247	-0.3606	-5.344	0.0000	0.0088	0.0720	0.6122	0.5447	0.4806	0.3521
5614	1.0157	5.152	0.0000	0.0131	0.1596	0.5769	0.3360	0.4081	0.0863
158	-0.8809	-4.884	0.0001	0.0240	0.3017	0.4309	0.1812	0.2702	0.1234
4297	0.4651	4.65	0.0001	0.0405	0.3776	0.4015	0.0284	0.5414	0.2841
1316	0.6048	4.565	0.0001	0.0468	0.7272	0.4181	0.0802	0.4938	0.2652

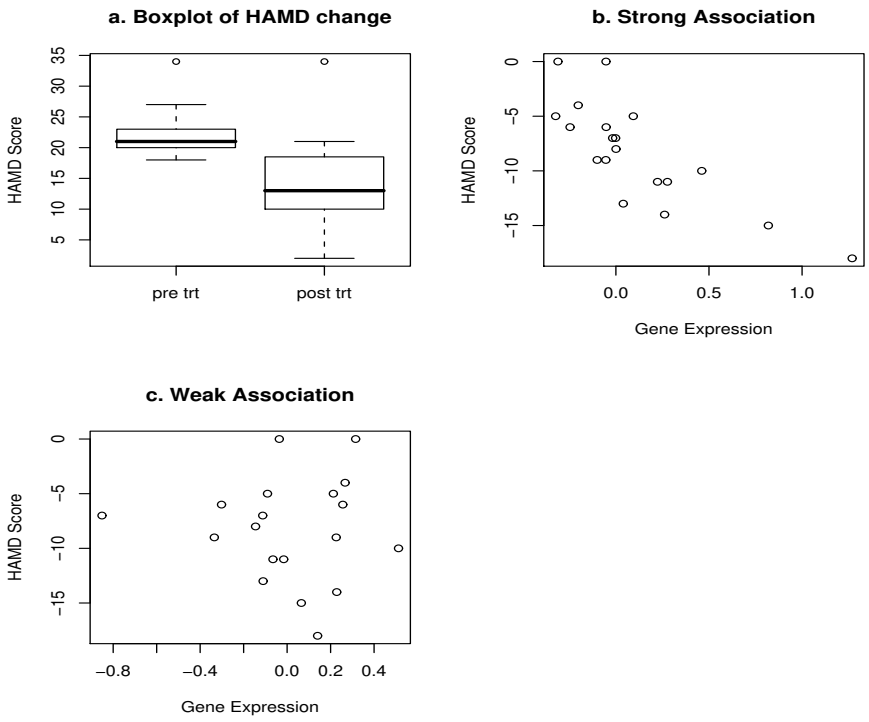

Figure 1: *Boxplot of the HAMD score (panel a). Panel b shows a gene, whose expression levels are strongly associated with the HAMD score, while Panel c shows a gene, with weak association with the HAMD score after adjusting for the effect of treatments and other possible confounding variables.*

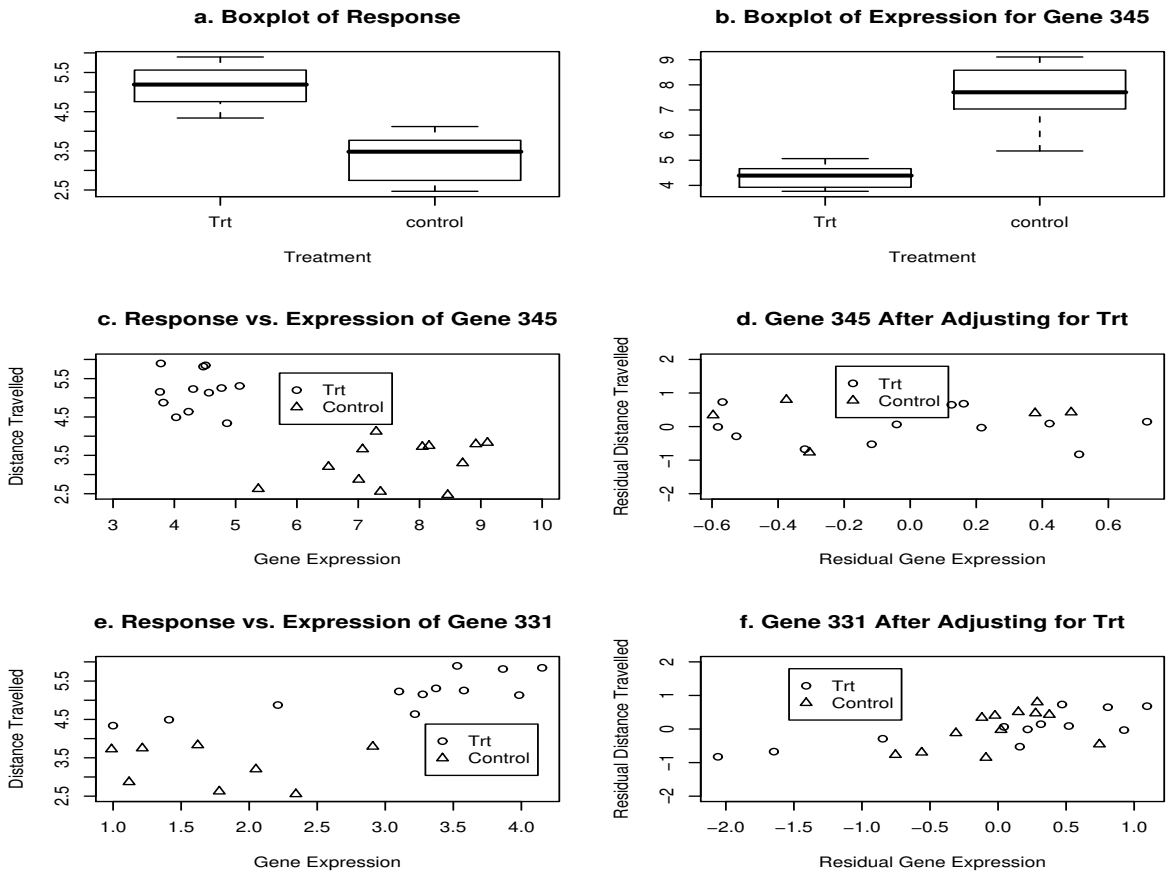

Figure 2: *Boxplot of the response (panel a) and expression of gene 345 (panel b) in the treatment groups, scatterplot of the response and expression of gene 345 (panel c), scatterplot of the response and expression of gene 345 showing no linear association after adjusting for the treatment effect (panel d), scatterplot of the response and expression of gene 331 (panel e), and scatterplot of the response and expression gene 331 showing linear association after adjusting for the treatment effect (panel f) .*

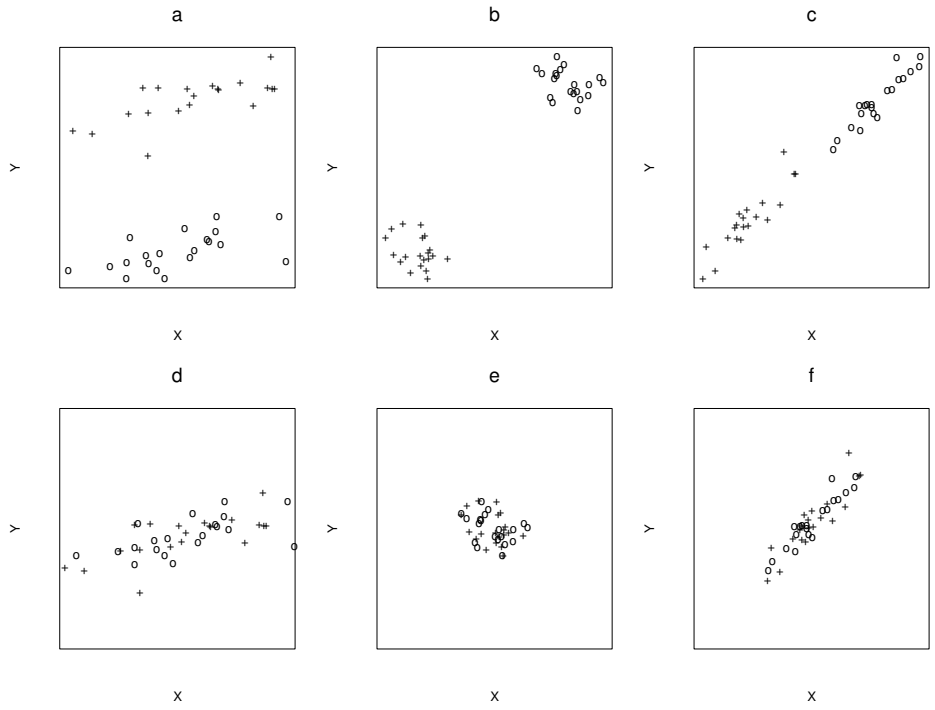

Figure 3: *Biomarker types in microarray experiment, when the response is differentially expressed. Pluses and circles represent the two treatment groups, respectively. Upper row (panel a, b, and c): scatterplots for the response (Y) versus gene-expression (X). Lower row (panel d, e, and f): scatterplots for the residuals after adjusting for treatment effects. Column 1 (panel a and d): an example of a prognostic biomarker. Column 2 (panel b and d): an example of a therapeutic biomarker. Column 3 (panel c and f): an example of a prognostic/therapeutic biomarker.*

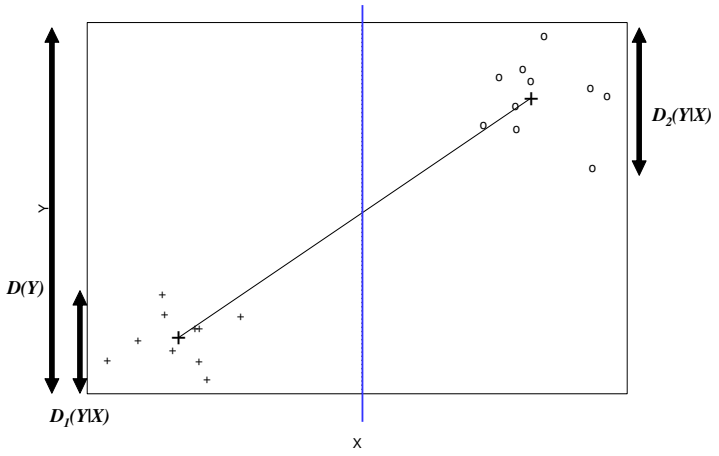

Figure 4: *An illustration of a regression tree model for a hypothetical example with two terminal nodes. The blue line in the plot indicates the split point in the regression tree.* $D(Y)$ *represents the total variability in the response* Y*, while* $D_1(Y|X)$ *and* $D_2(Y|X)$ *represent the variability within each of the terminal nodes.*

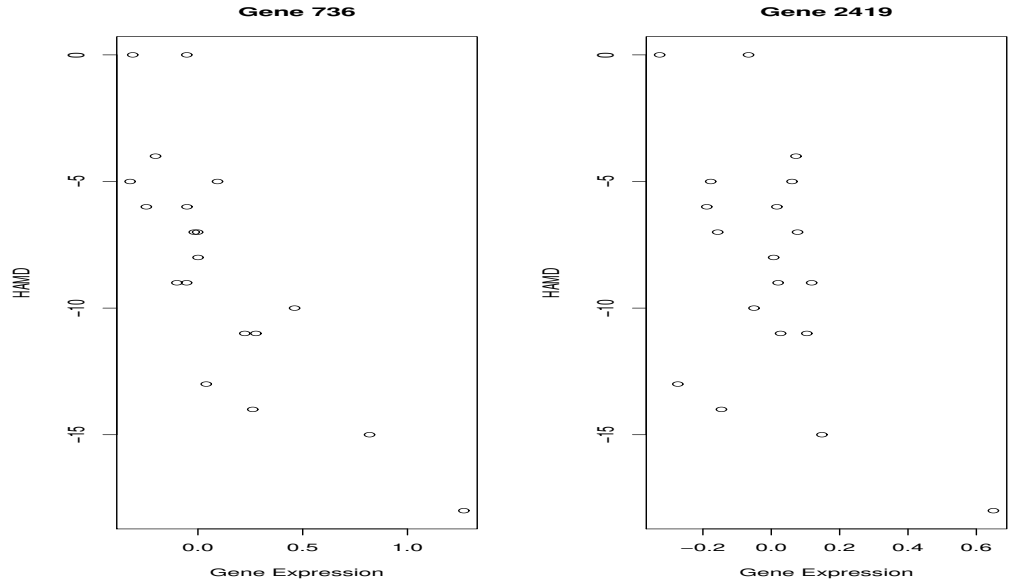

Figure 5: *Top two genes with highest adjusted association obtained from the joint model.*

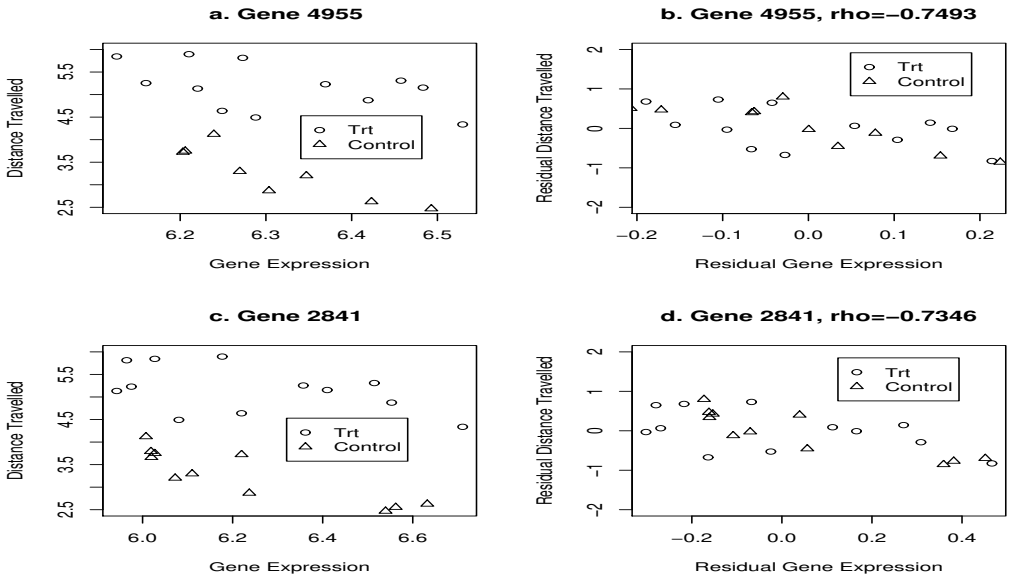

Figure 6: *Scatterplot of the response versus the gene expression for gene 4955 and 2841 (panel a and c). Scatterplot of the response versus the gene expression after adjusting for the treatment effect, yielding the adjusted association of for gene 2955 and 2841, respectively (panel b and d).*

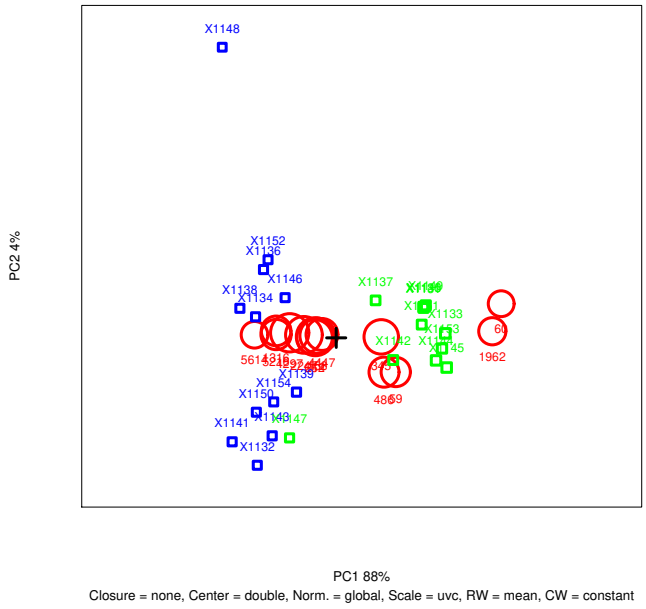

Figure 7: *Spectral map of animal behavior study using the top 14 differentially expressed genes. The squares are the samples and the circles are the genes, where the large circle, the higher expression levels of the genes have. The samples seem to be separated in two treatment groups: at the left side, they are from the treatment group except for one sample and at the right side, they are from the control group.*

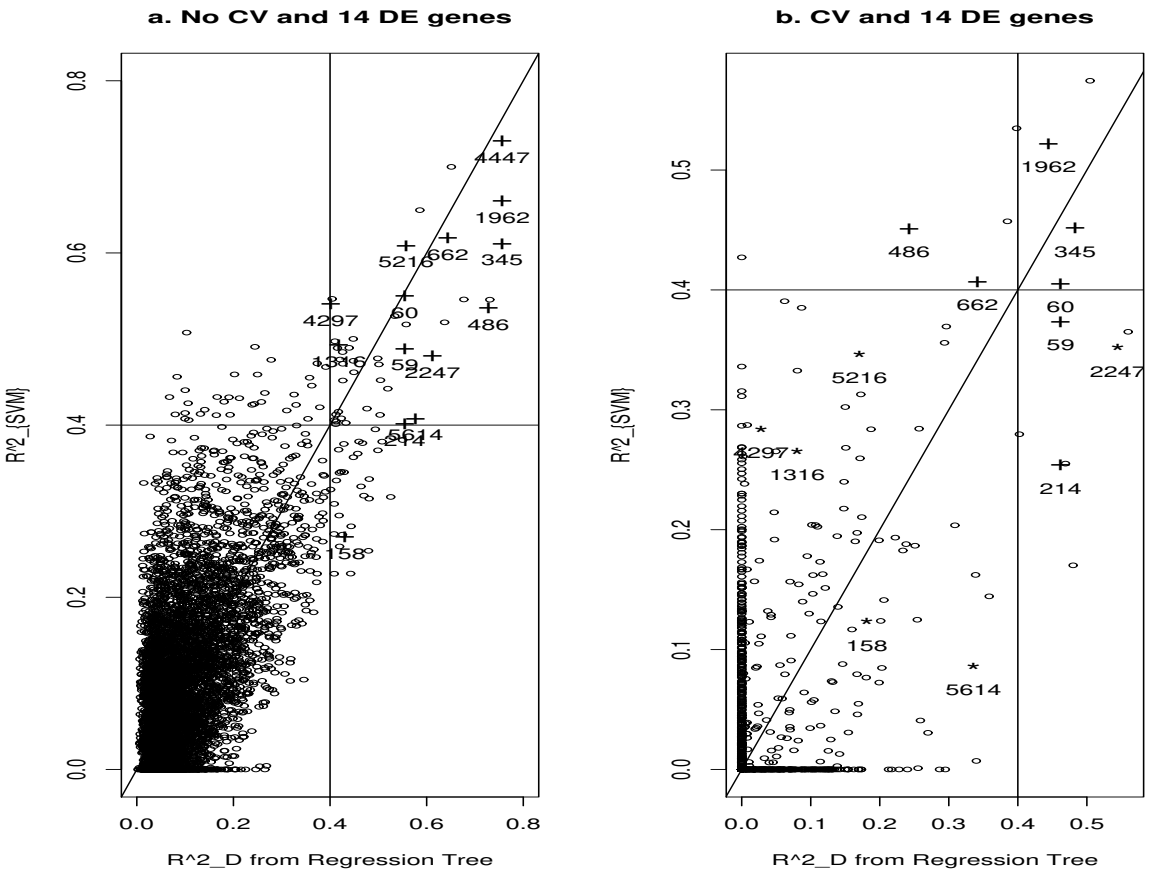

Figure 8: *Plot of R^2_D using the regression tree versus R^2_{SVM} using the support vector machine regression. The pluses in the plot are genes, which are differentially expressed, as listed in Table 3. Genes in stars in panel b are not differentially expressed by using the Kolmogorov-Smirnov tests and they are mostly having R^2_D and R^2_{SVM} below 0.4 after leave-one-out cross validation.*

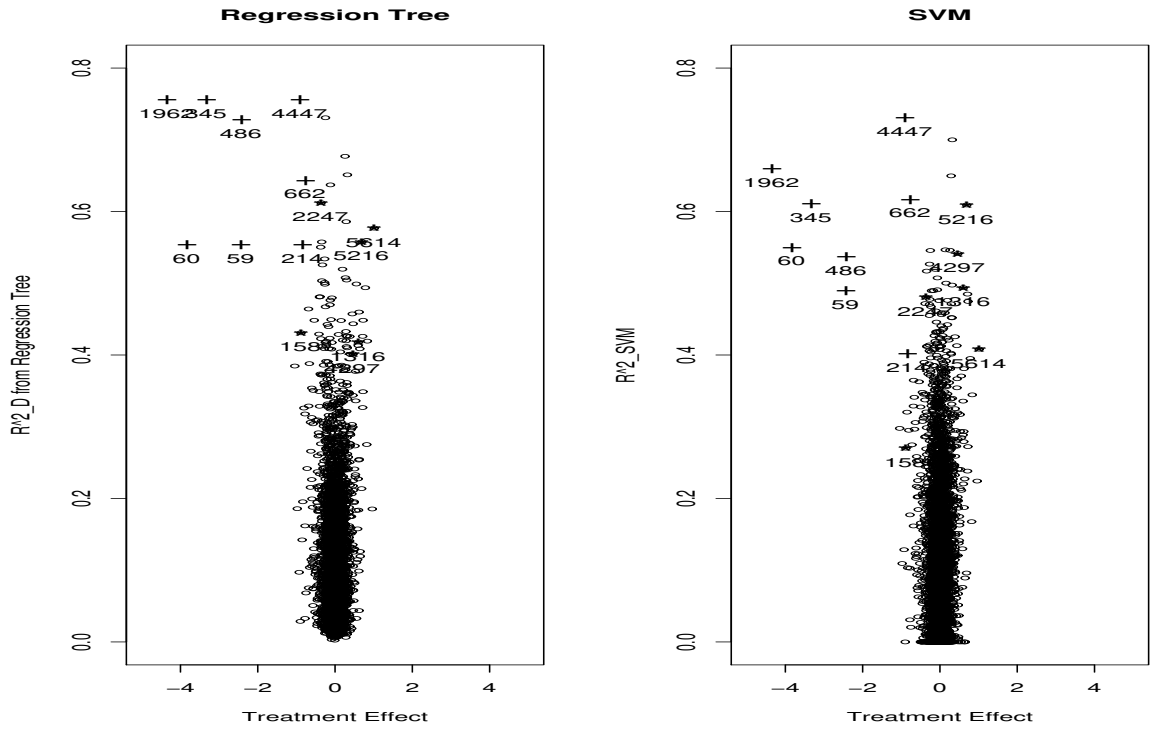

Figure 9: *Plot of treatment effects on gene expression versus the R^2_D using the regression tree and the R^2_{SVM} using the support vector machine regression. Pluses are genes with significant treatment effects upon the gene expression and they are showing relative high R^2_D and R^2_{SVM}.*

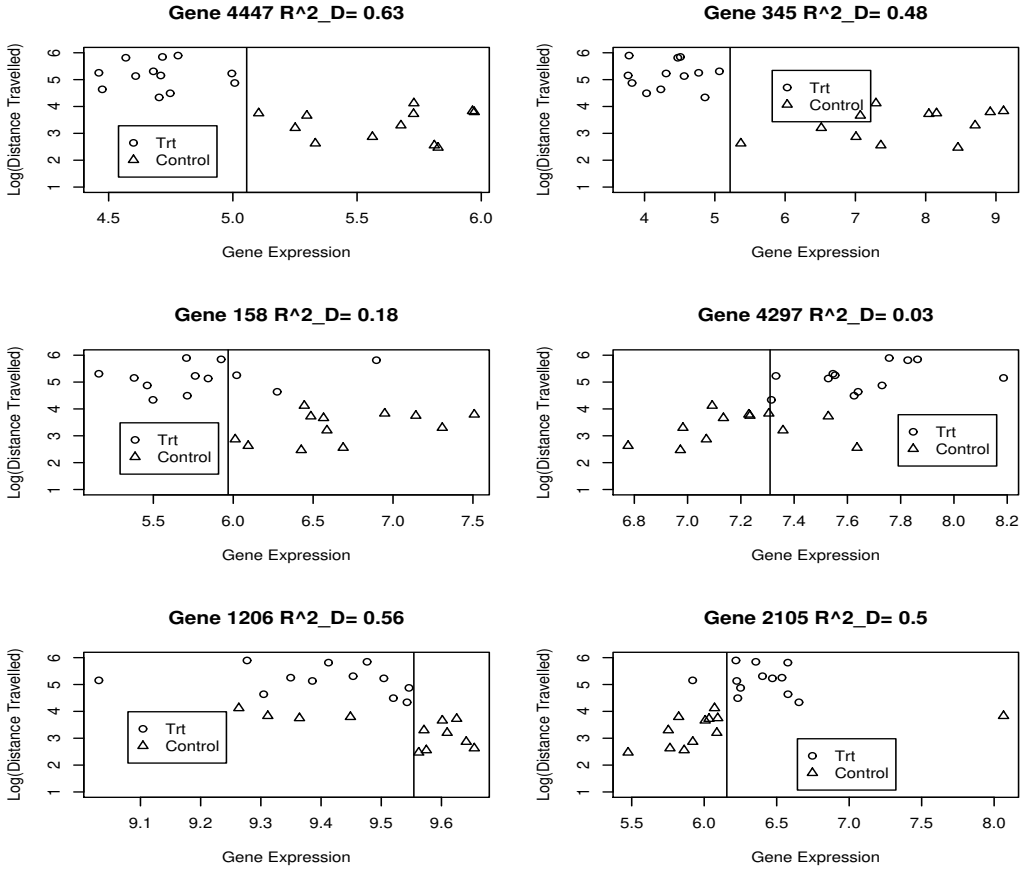

Figure 10: *Example of six genes using the regression tree, where the vertical lines are the cut-off values of the regression tree and* R_D^2 *values are obtained by using the leave-one-out cross validation data.*

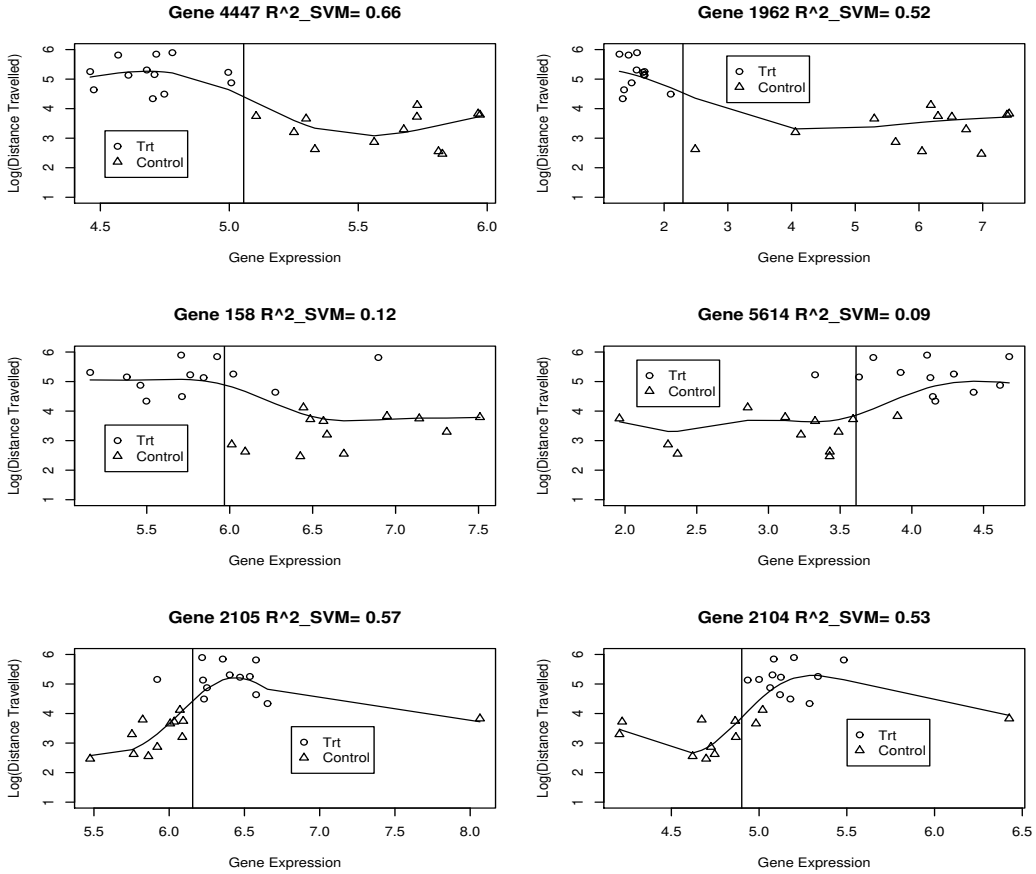

Figure 11: *Examples of six genes from the support vector machine regression using the radial kernel. The line is the fitted values of the support vector machine regression and* R^2_{SVM} *values are obtained by using the leave-one-out cross validation data.*

References

Affymetrix. GeneChip Expression Analysis Technical Manual, Rev.4. *Santa Clara, CA,* available at *http://www.affymetrix.com/support/technical/manual/expression_manual.affx.* 2004.

Alonso, A. and Molenberghs, G. Surrogate marker evaluation from an information theory perspective. *Biometrics*, 2007; **63**, 180–186.

Amaratunga, D. and Cabrera, J. *Exploration and Analysis of DNA Microarray and Protein Array Data.* New York: John Wiley & Sons. 2004.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. Prediction by supervised principal components. *Journal of the American Statistical Association*, 2006; **101**, 119–137.

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

Burzykowski, T. Molenberghs, G., and Buyse, M. *The Evaluation of Surrogate Endpoints.* New York: Springer. 2005.

Buyse, M. and Molenberghs, G. The validation of surrogate endpoints in randomized experiments. *Biometrics*, 1998; **54** 186–201.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 2000; **1**, 49–67.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., and Vapnik V. Support vector regression machines. In: Mozer M.C., Jordan M.I., and Petsche T. (Eds.), *Advances in Neural Information Processing Systems*, 1997; **9**, MIT Press, Cambridge, MA, 155–161.

Fletcher R. *Practical Methods of Optimization.* New York: John Wiley & Sons. 1989.

Hsu C.W., Chang C.C., Lin C.J. A Practical Guide to Support Vector Classification. Available at http://www.csie.ntu.edu.tw/ cjlin/libsvm/. 2001.

Mattera, D. and Haykin, S. Support vector machines for dynamic reconstruction of a chaotic system. In: B. Schölkopf, C.J.C. Burges, and A.J. Smola (Eds.), *Advances in Kernel Methods- Support Vector Learning, MIT Press*, 1999; Cambridge, MA, 211–242.

Müller, K.R., Smola, A., Rätsch, G., Schölkopf, B., Kohlmorgen, J., and Vapnik V. Predicting time series with support vector machines. In: W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud (Eds.), *Artificial Neural Networks ICANN'97*, 1997; Lecture Notes in Computer Science 1327, Berlin: Springer, 999–1004.

Nguyen, D.V. and Rocke, D.M. Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, 2002; **18**, 1625–1632.

Smola, A. Regression estimation with support vector learning machines. *Master thesis*. 1996; Technische Universität München, Munich, Germany.

Stitson, M., Gammerman, A., Vapnik, V., Vovk, V., Watkins, C., and Weston, J. Support vector regression with ANOVA decomposition kernels. In: Schölkopf B., Burges C.J.C., and Smola A.J. (Eds.), *Advances in Kernel Methods-Support Vector Learning*, 1999; MIT Press Cambridge, MA, 285–292.

Szechtman H, Sulis W, Eilam D. Quinpirole induces compulsive checking behaviour in rats: a potential animal model of obsessive-compulsive disorder (OCD). *Behavioral Neuroscience*, 1998; **112**: 1475–1485.

Tan, Y., Shi, L., Tong, W., Hwang, G.T., and Wang, C. Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Computational Biology and Chemistry*, 2004; **28**, 235-244.

Tilahun, A., Lin, D., Shkedy, Z., Geys, H., Alonso, A., Peeters, P., Tallone, W., Drinkenburg, P., Bijnens, L., and Molenberghs, G. Genomic Biomarkers for depression: Feature-specific and joint biomarkers. *Journal of Biopharmaceutical Research*. 2009; Accepted

Vapnik V. and Lerner A. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 1963; **24**, 774–780.

Vapnik V. and Chervonenkis A. A note on one class of perceptrons. *Automation and Remote Control*, 1964; **25**.

Vapnik V.. *The Nature of Statistical Learning Theory.* New York: Springer. 1995.

Wouters, L., Göhlmann H. W.H., Bijnens, L., Kass, S.U., Molenberghs, G., Lewi, P.J.. Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics*, 2003; **59**, 1131–1140.