2010
2011

FACULTY OF SCIENCES

Master of Statistics: Bioinformatics

Masterproef

Large scale prediction of phenotipic variables using gene expression data

Promotor :
Prof. dr. Ziv SHKEDY

Elvis Ndah

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

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

De transnationale Universiteit Limburg is een uniek samenwerkingsverband van twee universiteiten in twee landen: de Universiteit Hasselt en Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Maastricht University

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: Bioinformatics

Masterproef

Large scale prediction of phenotipic variables using gene expression data

Promotor :
Prof. dr. Ziv SHKEDY

Elvis Ndah

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

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

ACKNOWLEDGEMENT

I am grateful to God almighty for the good health and strength bestowed upon me to complete this work. I would like to express my gratitude to all those who gave me the possibility to complete this thesis. My deepest gratitude to my mother AZAH Odilia, who has been very instrumental in my life. I would like to express my sincere gratitude to Prof. Dr. Ziv Skhedy for his supervision and guidance. My deepest appreciation to Dr. Nji Abatih for his undying support and encouragements. In addition, I would to thank all my professors and classmates in helping me to broaden my view and knowledge. I would like to thank Mrs. Martine MACHIELS for her patience in handling all my complains.

SUMMARY

Motivation: Microarray technologies are increasingly being used in early drug development. When the gene expression data from microarray experiments also contains the IC50 values of a drug, it is of interest to predict the dosage based on the gene expression profiles. In this study, the supervised principal component analysis (SPCA) by Bair et al., (2006) was used to select and evaluate possible biomarkers and joint biomarker for the IC50 values of compound 352.

Results: The response of interest is a vector of IC50 values of compound 352 the corresponding genes expression contains 7722 genes and 32 samples. This expression matrix was separated to genes negatively correlated to the IC50 and those positively correlated. The approach used in the SPCA method for biomarker (gene) selection is to select genes only the top k genes that are associated to the IC50, using this genes construct a joint biomarker and assess statistical significance for prediction using the joint biomarker.

Key Words: Biomarker, Drug development, Gene expressions, Microarray experiments, Supervised principal component analysis.

CONTENTS

1. INTRODUCTION

Since the introduction of microarray technology into biomedical research it has been very useful in monitoring thousands of genes in a single experiment at the expression level across the genome and has been proven efficient in exploring the complex patterns in biological systems (Zhao and Simon, 2010) . However, this ability to measure thousands of genes has resulted to data sets with the number of genes p, far exceeding the number of samples N and is often represented by a $p \times N$ matrix. Each row of the matrix represents the expression levels of a gene across different biological samples and each column represents the gene expression levels of a genome under a sample. Although there are many genes in the microarray, only subsets of these genes have meaningful contributions to variations relating to a given phenotype (Chen et al., 2008). Statistical methods are often required to analyse the intensity of the genes across different conditions associated to an outcome such as, classification and prediction of the phenotypes based on the expression profiles. There are many challenges for the development and validation of predictive models in microarray settings due to the large number of predictors (Zhao and Simon, 2010).

In these statistical studies, a frequent objective is to identify a subset of genes whose expression profiles are significantly correlated with a given phenotype. There has been extensive attention on the study of the relationship between categorical or survival outcomes and gene expression profiles, on the other hand there are relatively few publications on prediction for continuous phenotypes (Segal et al., 2003). In the microarray settings with the large p and small N standard regression models are problematic because $X^T X$ may be singular, (where X is the design matrix). Of particular interest is when the gene expression profiles are measured alongside the IC50 (Inhibitory concentration 50%) value of a drug which is the concentration at which the drug (inhibitor) produces 50% inhibition. The genes in a microarray experiment are likely to have a strong and complex correlation structure between the expression levels of the genes due to biological pathway and gene network relationships (Segal, et al., 2003). Numerous techniques have been proposed to handle the problem of dimensionality such as; (a) the Least absolute shrinkage and selection operator (LASSO) (Tibshirani, R., 1996), Partial least squares, principal component regression and supervised principal component analysis (SPCA). The LASSO penalizes the regression coefficients by shrinking them towards zero. However, it is limited in that the number of non-zero coefficients in the solution is at most N for any choice of the shrinkage parameter. In situations where there are highly correlated variables, LASSO will randomly select one of them. Zhao and Simon (2010) used the LASSO to predict the Gleason score of human prostate cancers. Partial least squares down-weights genes not correlated to the response and maximizing the covariance between the response and a linear combination of the genes. Nguyen and Rocke (2002) used the Partial Least Squares

method to select relevant genes for prediction of survival time for cancer patients. The disadvantage of the Partial Least squares method is that it does not remove the noisy genes; as a result, a large number of noisy features can contaminate the predictions (Hastie et al., 2008). Principal components regression (PCR) uses principal components analysis to decompose the set of p genes, into k principal components, $k<p$. The k principal components selected are then included in a regression model that account for as much variation in the gene expressions as possible (Hastie et al., 2008). The disadvantage of this method is that the principal component aims at explaining the variability in the predictors rather than the outcome so it is possible to select genes that have little or no relationship with the outcome. Bair et al., (2006) proposed a modified version of the PCR called supervised principal component analysis (SPCA) that does variable selection based on the association of the genes to the response. It excludes genes that have a weak linear association to the response. SPCA is the mothod of choice for the analysis in this study. This is because it does variable selection based on the outcome information and it allows the possibility of including only those features that are highly correlated to the response into the model.

1.1 Objective

The main objective of this study was to select a relevant subset of genes (biomarkers) from the gene expression profile that are associated to the IC50 value of compound 352. From the selected genes, estimate a joint biomarker associated to the effect of compound 352. Also of importance was to assess the statistical significance of the association between the joint biomarker and the IC50. Genes that are negatively correlated to the IC50 are separated from those that are positively correlated and the 2 groups of genes are analysed separately.

The rest of this report is organized as follows; section 2 gives a brief background about gene expression profiles and drug development. Section 3 describes the data and methods used for the analysis and the results are outlined in section 4. The discussions and conclusion makes up section 5 of the report. The appendix (section 6) contains some tables of and graphs for model diagnostics.

2. BACKGROUND

The rapid development of the microarray technology has motivated its use in clinical and preclinical trials, diagnosis or predictions of phenotypes. Many studies have attempted to find relevant subset of genes associated to a phenotype. A relevant subset of genes, often referred to as 'biomarkers', may be useful in separating patients in diagnosis, prognosis and for appropriate therapeutic selection in clinical management.

The Food and Drug Administration (FDA) defined a biomarker as a characteristic objectively measured and evaluated as an indicator of normal biologic or pathogenic processes, or pharmacologic responses to a therapeutic intervention (Carroll, 2007). Current research has shown that biomarkers can provide indications of both the potential effectiveness and the potential hazards associated with a therapeutic intervention. Biomarkers are now playing an increasingly important role in the discovery and development of new drugs (Chau et al., 2008).

The use of biomarkers in drug discovery, development and post-approval has the potential to facilitate development of safer and more effective medicines (Lin et al, 2010) and guide dose selection as well as the understanding of the mechanism by which the drug works. It provides insights on decisions whether to continue with the development of the drug, to screen compounds for toxicity before they enter clinical trials, to monitor the development of toxicity during clinical trials, and to forecast adverse events resulting from wider exposure (Frank & Hargreaves, 2003). Microarray experiment makes it possible to search for genes that can serve as biomarkers; because of the dimension of the microarray there are many potential biomarkers for any given phenotype, thus there is a need for gene selection. The large p small N problem with highly correlated genes makes it very difficult to select the most powerful biomarker. Because of the large number of genes, it is easy to find biomarkers that perform excellently with the training data but achieves poor prediction outside the training data (Bovelstad et al., 2007) so one should be cautious in the statistical method used and the choice of the model.

No one biomarker is likely to have all of the characteristics necessary to provide a robust understanding of response, as a result, the use of multiple biomarkers is likely to improve prediction. However, the use of a combination of biomarkers (referred to as a joint biomarker) may introduce challenges, such as how to combine results, and the interpretations in different clinical contexts. Statistical techniques play a vital role for the selection and evaluation of biomarkers in drug development and for the understanding of the complex nature of the relationship between genes and phenotypes. The improper use of these techniques or interpretation of biomarkers can be detrimental to the research; it may lead to misdirecting therapy or research activities.

Lin et al., (2010) classified biomarker into two groups, therapeutic and/or prognostic genes depending on their relationship to a clinical outcome. Therapeutic biomarkers are genes that response to treatment and can enable the clinicians understand the effect of a treatment on the clinical outcome while prognostic biomarkers are related to the response with or without the treatment effect. The use of microarray experiments in clinical and preclinical trials has been extended beyond the level of subject classification into possible prediction of clinical outcomes using single genes (biomarker) or combination of genes (joint biomarker). In this study, the interest is on prognostic biomarkers.

3. METHODOLOGY

3.1 Data

The microarray data set used for the analysis in this study is from a drug development research of compound 352. In this experiment the expression level of *7722* genes where measured from *32* samples resulting to a gene expression matrix of *7722 x 32*, with *p=7722* genes and *N=32* samples. The outcome of this study is the IC50 value of compound 352. The microarray data is summarized in a *pxN* matrix $X = x_{ij}$ as shown below

$$\mathrm{X} = \begin{bmatrix} x_{11} & x_{21} & . & . & x_{N1} \\ x_{12} & x_{22} & . & . & x_{N2} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{1p} & x_{2p} & . & . & x_{Np} \end{bmatrix} \qquad \text{the IC50 vector} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_N \end{bmatrix}$$

N = 32 is the number of sample and *p = 7722* is the number of genes. The data set is divided into two, i.e. genes negatively correlated (X_{neg}) and those positively correlated (X_{pos}) to the IC50. *2794* genes are correlated to the IC50 while *4928* are negatively correlated to the IC50 thus two data sets are obtained X_{pos} a *2794 x 32* matrix of positively correlated genes and X_{neg} a *4928 x 32* matrix of negatively correlated.

3.2 Gene Specific Model

Let Y be the IC50 values of the 32 samples and x_i the expression vector of gene *i, (i=1, ..., p)* from either X_{pos} or X_{neg} expression matrices above. To test the association between the *ith* gene and the IC50 it suffices to test for independence (Sohn et al., 2011), thus the inference of interest would be to test the hypotheses:

H_0^i: gene *i* is associated with the IC50 against H_1^i: gene *i* is not associated with the IC50

To identify prognostic genes Lin et al., (2010) propose the use of a univariate regression model. Sohn et al., (2011), Sreekumar & Jose (2008) and Lin et al., (2010), used the linear regression coefficient to quantify the hypothesis of independence. Let x_i be a vector of gene expression values for gene *i, i=1,...,p, (p=2794/4928)* and Y the vector of the IC50 values with mean α, the gene-specific regression model is

$$E(Y) = \alpha + \beta_i x_i \qquad (1)$$

The gene-specific univariate regression coefficients β_i, of model (1) quantify the association between gene expressions and the IC50. In order to assess the significance of the association it suffices to test the null hypothesis

$$H_0^i:\ \beta_i = 0 \qquad versus \qquad H_1^i:\ \beta_i \neq 0 \qquad (2)$$

(Sohn et al., 2011 and Thomas et al., 2001) where H_0^i is the null hypothesis of gene *i*. In the marginal testing H_0^i is rejected in favour of H_a^i if $\beta_i \neq 0$ (Sreekumar & Jose, 2008), if the null hypothesis in (2) is rejected, genes with $\beta_i > 0$ are up-regulated prognostic biomarkers (positive biomarkers) and the gene with $\beta_i < 0$ are down regulated (negative biomarkers) (Lin et al., 2010). Simultaneous testing for association of thousands of genes would lead to the multiple testing problem (Benjamini & Hochberg, 1995). Thus, using a common significance level will generally result in a large number of false positives, so there is a need to control for type I error rate (Hastie et al., 2008). The false discovery rate of Benjamini and Hochberg (FDR-BH) was used to adjust for multiplicity. The False Discovery Rate (FDR) is the expected proportion of genes falsely declared significant, among the total number of genes that are significant (Hastie et al. 2008). The FDR procedure is based on table 1 below and is defined as

$$FDR = E\left(\frac{V}{R}\right)$$

Table 1: *Number of errors committed when simultaneously testing the null hypothesis.*

	Non Significant	Significant	Total
H_0 True	$m_0 - V$	V	m_0
H_0 False	$m_1 - S$	S	m_1
Total	m – R	R	m

Where m_0 is the total number of true null hypothesis, m_1 is the total number of false null hypothesis, V the true null hypothesis that are declared significant, S is the number of falsely declared null hypothesis declared significant and R the total number of null hypothesis that are significant. The Benjamini–Hochberg (BH) procedure is based on p-values obtained from an asymptotic approximation to the test statistic, or a permutation distribution (Benjamini & Hochberg, 1995). Benjamini and Hochberg (1995) show that under the assumption of independence and regardless of the distribution of the p-values and the number of true/false null hypotheses, the algorithm for the FDR-BH procedure is

- Order the p-values: $P_{(1)} \leq P_{(2)} \leq , \ldots , \leq P_{(i)}, \ldots , \leq P_{(p)}$ corresponding to the ordered hypothesis$H_{(1)} \leq H_{(2)} \leq , \ldots , \leq H_{(i)}, \ldots , \leq H_{(p)}$
- Define $L = \max\{ j: P_{(j)} \leq \alpha . \frac{j}{p}\}$
- Reject all hypothesis $H_{(j)}$ for which $P_{(j)} \leq P_{(L)}$the FDR-BH threshold.

The FDR-BH procedure was implemented using the multtest package of the R statistical software.

3.3 Supervised Principal Component Analysis (SPCA)

Bair et al (2006) proposed the SPCA that handles the problem of high-dimensionality and estimation of a latent variable by using only genes with the strongest correlation to the response. This method uses principal components (PC), estimated from a selected subset of genes to predict an outcome, it is a supervised process because gene selection is based on the outcome information. The main assumption of the SPCA method is that there is a latent variable *U(X)* that is associated with a given response. This assumption is based on the fact that only a subset of genes from the microarray experiment work together to bring about changes in cellular processes that is related to a phenotype (Chen et al, 2008 and Bair et al 2006). The latent variable *U(X)* is viewed as an aspect of a cellular process, which cannot be measured directly but may be estimated by combining the expression values of the relevant subset of genes. The latent variable estimated is the joint biomarker associated to the phenotype. The supervised principal components (PC) helps to uncover groups of genes that are expressed together (Bair et al 2006) and estimation of the PCs from the selected subset of genes improves prediction accuracy (Chen et al 2008). The main advantage of the SPCA is that it is consistent in the features it selects (Roberts & Martin, 2006 and Bair et al., 2006). Features that have little or no linear association to the response are excluded from the model.

Let X be a standardized $p \times N$ matrix of feature measurements and Y a response, X is standardized so that the slopes can be comparable across the genes. The SPCA algorithm is as follows,

1. Compute univariate standard regression coefficients for each feature.
2. Form a reduced data matrix consisting of features whose univariate coefficient exceeds a threshold θ (θ is estimated by cross-validation).
3. Compute the first (or first few) principal components of the reduced data matrix
4. Use these principal component(s) in a regression model to predict the outcome

3.3.1 Singular value Decomposition (SVD)

Let X be a one of the $N \times p$ matrix of gene expression in section 3.1, in which the p columns index the genes and the N rows index the samples, let x_i be a random variable for the gene expression

values of the *ith* sample. Let Σ be a *pxp* covariance matrix of *X* then the ordered eigenvalues of Σ, $(\lambda_1 \geq \lambda_2 \ldots \geq \lambda_N)$ are such that the vector of coefficient α_1 corresponds to the largest eigenvalue λ_1 and $var(\alpha_1^T x) = \lambda_1$ (Chen et al. 2008) then $\Sigma\alpha_i = \lambda_i\alpha_i$, $i = 1, \ldots, N$ (Johnson & Wichern, 2007). The principal component scores of the gene expression vector *x* is a linear combination of *x* and the eigenvector α defined as $\alpha_{k1}^T x = \alpha_{k1}x_1 + \alpha_{k2}x_2 + \ldots + \alpha_{kp}x_p$. The vectors $\{\alpha_{k1,}\alpha_{k2}, \ldots, \alpha_{kN}\}$ are referred to as the loadings of the *kth* principal components $k = 1, \ldots, N$.

Principal component analysis (PCA) is directly related to singular value decomposition when the principal components are calculated from the covariance matrix (Wall et al., 2001). Jolliffe (2002) estimated the eigenvectors $\{\alpha_{k1,}\alpha_{k2}, \ldots, \alpha_{kp}\}$, of *X* using SVD assuming the gene expression matrix *X* is standardized such that the gene expression values have mean 0 and variance 1. The mathematical definition for singular value decomposition of X is the following:

$$X = UDV^T$$

where *U* is a $p \times N$ matrix, its columns are the principal component scores and form an orthonormal basis for the gene expression profiles corresponding to the ordered singular values $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_N \geq 0$ of D, an *NxN* diagonal matrix. V^T is an *N x p* matrix whose rows are the right singular vectors (expression levels vector). SVD provides a means of estimating the coefficients of the principal components and the principal component scores of each sample simultaneously (Jolliffe, 2002).

3.3.2 Underlying Model

Let *R* be a list of genes related to a phenotype from a gene expression matrix $X = x_1, x_2, \ldots, x_p$, that explains the variation of a response Y. Suppose that Y is related to a an underlying latent variable *U(X)* by a linear model

$$Y = \beta_0 + \beta_1 U(X) + \varepsilon \qquad (3)$$

this latent variable *U(X)*, represents an underlying biological process associated with the genes in *R*. $\varepsilon \sim N(0,1)$ is the error term. Let X_r be the gene expression measurements for the genes in *R*,

$$X_r = \alpha_{0r} + \alpha_{1r} U(X) + \epsilon_r, \qquad r \in R \qquad (4)$$

where α_{0r} and α_{1r} are weights and $\epsilon_r \sim N(0,1)$. In addition to the genes in X_r, there are many additional genes X_j, $j \notin R$ that are independent of *U(X)* but included in the gene expression matrix *X*. The SPCA method aims at estimating the subset *R* of genes from the matrix *X* responsible for the cellular process, compute the latent variable *U(X)* fit model (3) and assess statistical significance

of association between *U(X)* and the outcome. From the prediction model (3), statistical significance of β_1 would indicate significant association between *U(X)* and the outcome (Chen et al, 2008).

3.3.3 Description

Bair et al.,(2006) published a technical report about the SPCA method, where they describe the the method, below is a summary of the model description. Assume X is a gene expression matrix centred to have mean 0 and variance 1, then the singular value decomposition of *X* as described in section 3.3.1 is $X = UDV^T$. The SPCA algorithm is made up of 4 main steps, the screening step that measures the association of the individual genes to the response, the second step estimates the subset of genes *R* that are associated to the response, then compute the latent variable *U(X)* and finally use *U(X)* for prediction. This procedure ensures that the latent variable constructed maximizes the association between the response and the gene expressions and is such that *R* is the best gene subset associated to the response (Tilahun, et al 2010).

Screening: this step eliminates genes that have little or no correlation to the response. The measure of association between the response and the individual genes is the univariate regression coefficient. Let S be a p-vector of these standardized regression coefficients then s_j is equivalent to β_i's in the gene-specific model in section 3.2,

$$s_j = \frac{x_j^T y}{\|x_j\|} \quad , \quad where \ \|x_j\| = \sqrt{x_j^T x_j} \qquad (5)$$

Estimating R: let θ be a threshold value, Bair et al (2006) recommends estimating θ by cross validation. Let X_θ be a reduced matrix consisting of only those genes whose univariate regression coefficients exceed the threshold value θ, i.e. $|s_j| > \theta$, the estimate of *R* ($\hat{R}$) is the subset of genes that pass the threshold test.

Compute *U*: X_θ is a reduced matrix of gene expression profiles made up of only the genes in R, the SVD of X_θ is given by

$$X_\theta = U_\theta D_\theta V_\theta^T \qquad (6)$$

where $U_\theta = u_{\theta,1}, \ldots, u_{\theta,k}$, termed the ordered supervised principal components, where k is the number of features that pass the threshold.

Prediction: The first supervised principal component $u_{\theta,1}$ is used as a predictor in a univariate regression model with response y.

$$\hat{y}^{spc,\theta} = \bar{y} + \hat{\gamma}.u_{\theta,1} \qquad (7)$$

$\hat{y}^{spc,\theta}$ is the estimate of y based on the reduced matrix X_θ, and $\bar{y}$ the mean of y, thus $\hat{\gamma} = u_{\theta,1}^T y$, since $u_{\theta,1}$ is the left singular vector of X_θ, it has mean 0 and unit norm.

This method encourages the use of the first (or first few) principal component (Bair et al. 2006) because it explains most of the variability of the original variable and it has the highest prediction accuracy (Niklas & Low, 2011). Equation (6) can be rearranged so that

$$U_\theta = X_\theta V_\theta D_\theta^{-1} = X_\theta W_\theta \qquad (8)$$

$u_{\theta,1}$ is a linear combination of the columns in X_θ thus it can be written as

$$u_{\theta,1} = X_\theta w_{\theta,1} \qquad (8')$$

where $w_{\theta,1}$ is the loading of the first principal component, hence from (7) we get

$$\hat{y}^{spc,\theta} = \bar{y} + \hat{\gamma}.X_\theta.w_{\theta,1} \qquad (9)$$

$$= \bar{y} + X_\theta.\hat{\beta}_\theta \qquad (10)$$

where $\hat{\beta}_\theta = \hat{\gamma}.w_{\theta,1}$, ($\hat{\gamma}$ is a scalar), the regression model in (10) is restricted to the genes that pass the threshold test.

The SPCA method was used to select a subset of genes corresponding to a biological process that is responsible for the changes in a phenotype resulting from the effect of compound 352; the joint biomarker for the prediction of the IC50 was estimated for the positive and negative genes. The gene specific model (1) is equivalent to the screening step. The expression levels of these genes are combined into one variable, the first supervised principal component representing the joint biomarker. Rather than estimating a threshold value θ, the top k genes that maximize the prediction of the IC50 values is estimated (Roberts & Martins, 2006).The expression matrices X_{pos} and X_{neg} in section 3.1 were standardized and the SPCA procedure was implemented as follows;

1. Fit a gene-specific simple regression model relating the IC50 values to the individual gene expressions i.e. for each gene fit the model

   $$Y_i = \alpha + \beta_i x_{ij} + \varepsilon_{ij} \qquad (11)$$

   where Y_j is the IC50 value of the *jth* subject and x_{ij} the gene expression value of the *ith* gene for the *jth* subject, β_i is the standardized regression coefficient and $\varepsilon_{ij} \sim N(0,1)$.

2. For each of the p models in step 1, rank the genes based on decreasing magnitude of the regression coefficients β_i such that $\beta_1 \geq \beta_2 \geq, \ldots \geq \beta_p$, *p=4928/2794.*
3. Form a reduced matrix X_k, of the top k genes according to the ranking above k *=2,3, ..., m,* where m is the number of significant genes. Let $R_k = \{x_i \subseteq X: i = 1, .., k\}$ consists of the top k genes. Compute first PC score (PC1) using only genes in R_k and fit the model

$$Y_j = \alpha + \beta U_j(X_k) + \varepsilon_j$$

where Y_j is the IC50 and $U_j(X_k)$ is the joint biomarker of the *jth* sample and $\varepsilon_j \sim N(0,1)$ the error term. Let T_k be the R-squared value of the model. From the set of regression R-squared values $\{T_2, T_3, \ldots, T_m\}$ choose the value of k corresponding to the highest R-squared value i.e. $k = \{T_k : T_k = \max_{2 \leq k \leq m} T_k\}$, select the corresponding subset of top genes in R_k to form the reduce matrix X_k.

4. Compute the first supervised principal component $U_{k,1}$ of X_k use it in model (3) for predicting the IC50 of compound 352.

3.4 Cross Validation

To assess the performance of any prediction model, it is necessary to train and test the model on separate data. Model building processes should take into account the variation in prediction performance that would result from using a different data (Efron, 1983).

Regression R-squared, can be an overly optimistic view for accuracy of a prediction model. The estimates obtained from a model are biased because the model is tuned for the maximum agreement with the training data (Snee, 1977). It is advisable therefore to validate the regression model by testing the model on data not used for training (Snee, 1977). In cross-validation, a different subset of observations is successively set aside (validation set) and used to assess the performance of the prediction model. This procedure eventually results in a complete set of predicted values, each of which was generated by a model independent of the validation set.

For microarray settings, cross validation is very problematic due to the large number of highly correlated predictors. In order to select good features, Refaeilzadeh et al. (2007) recommends using almost the entire data i.e. at least 90% for model building. But for the data described in section 3.1, there are *N=32* observations and thousands of correlated features hence too many competing models for each split, thus leave one out cross-validation (LOOCV) would be an appropriate choice for feature selection (Efron, 1983 and Nguyen &Rocke, 2002).

In the LOOCV, let (Y_k, x_k) be the *kth* record in the data set ($k = 1, \ldots, N$) where Y_k is the IC50 and x_k the gene expression of the *kth* sample. Temporarily remove (Y_k, x_k) from the data set. Train the

model on the remaining *N-1* samples, test on (Y_k, x_k) and obtain the estimate of the joint biomarker $\widehat{U}(X_k)$ of the left out data. Choose the top k genes that correspond to the model with maximum correlation between the estimated joint biomarkers $\widehat{U}(X)$ and the IC50 value.

4. RESULT

This section presents the basic exploratory data analysis, and the results obtained from the gene-specific model and the supervised principal component analysis.

4.1 Exploratory Analysis

The microarray data set described in section 3.1 consist of gene expressions of p = 7722 genes from N = 32 samples. Let the outcome vector Y denote the IC50 value for the 32 samples. Out of the 7722 genes, 4928 are negative correlated to the IC50 values (negative genes) while 2794 have positive correlations (positive genes) to the IC50. The positive genes (up regulated) are analysed separately from the negative gene (down regulated), thus 2 new expression matrix X_{neg} a *4928x32* matrix of negatively correlated genes and X_{pos} a *2794x32* matrix of positively correlated genes.

The correlation to the negative genes ranges from -1.67 x 10^{-5} to -0.79 while that for the positive genes ranges from 1.74 x 10^{-5} to 0.74. Histograms of the distributions of correlation between the negative and positive genes to the IC50 values are in Figure 4.1 panel A and B respectively.

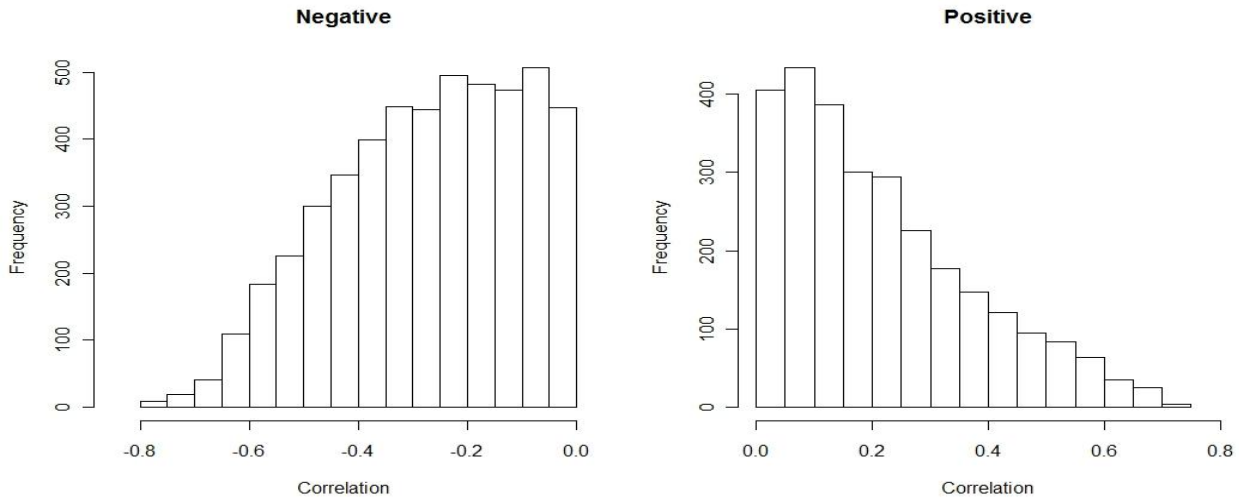

Figure 4.1: *Histogram of the correlation between the IC50 values and the gene expression of the 4928 negative (left panel) and the 2794 positive genes (right panel).*

The histogram shows that most of the genes have low correlations with the IC50 values. Gene 8884_at has the strongest negative correlation of -0.79 (Figure 4.2, left panel) and gene 222068_at has the strongest positive correlation of 0.74 (Figure 4.2, right panel) to the IC50.

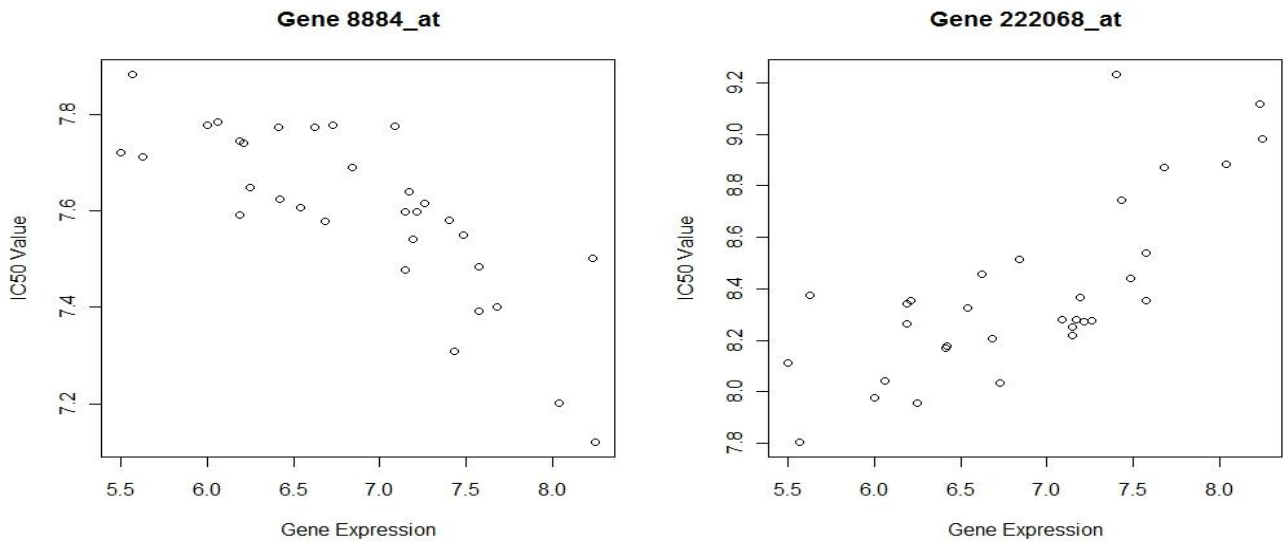

Figure 4.2: *Scatter plot of the genes expression against the IC50 values for the strongest negative correlation (gene 8884_at) and strongest positive correlation (gene 222068_at).*

Figure 4.3 shows the scatter plot of the gene with the weakest negative correlation of -1.67×10^{-5} (80135_at) and the gene with the weakest positive correlation of 7.43×10^{-5} (100289612_at)

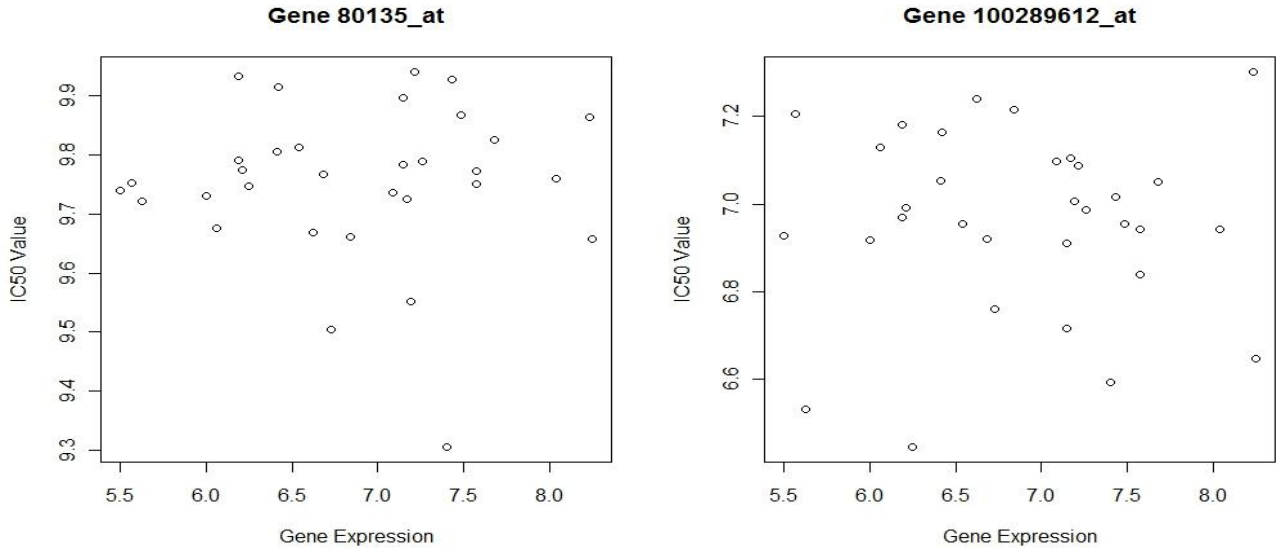

Figure 4.3: *Scatter plot of the genes expression against the IC50 values of the gene with the weakest negative correlation (gene 80135_at) and weakest positive correlation (gene 10029612_at).*

The scatter plots in Figure 4.3 shows no apparent relationship between the IC50 values and the gene profiles of the genes with very low correlation.

4.2 Gene-specific model

The gene specific model was used to identify and evaluate the association measure of each individual gene in the microarray as a possible biomarker for the IC50 values. Figure 4 C and D shows a histogram of the distribution of the regression coefficient β_i's of the gene-specific model; as expected majority of the genes have weak association to the IC50 values (Tilahun, et al., 2010). The same observation holds for the R-squared values. This is in line with the assumption that only a subset of relevant genes is associated to a phenotype. The distribution of the R-squared values (Figure 4 A and B) follows a t-distribution for both the positive and the negative genes.

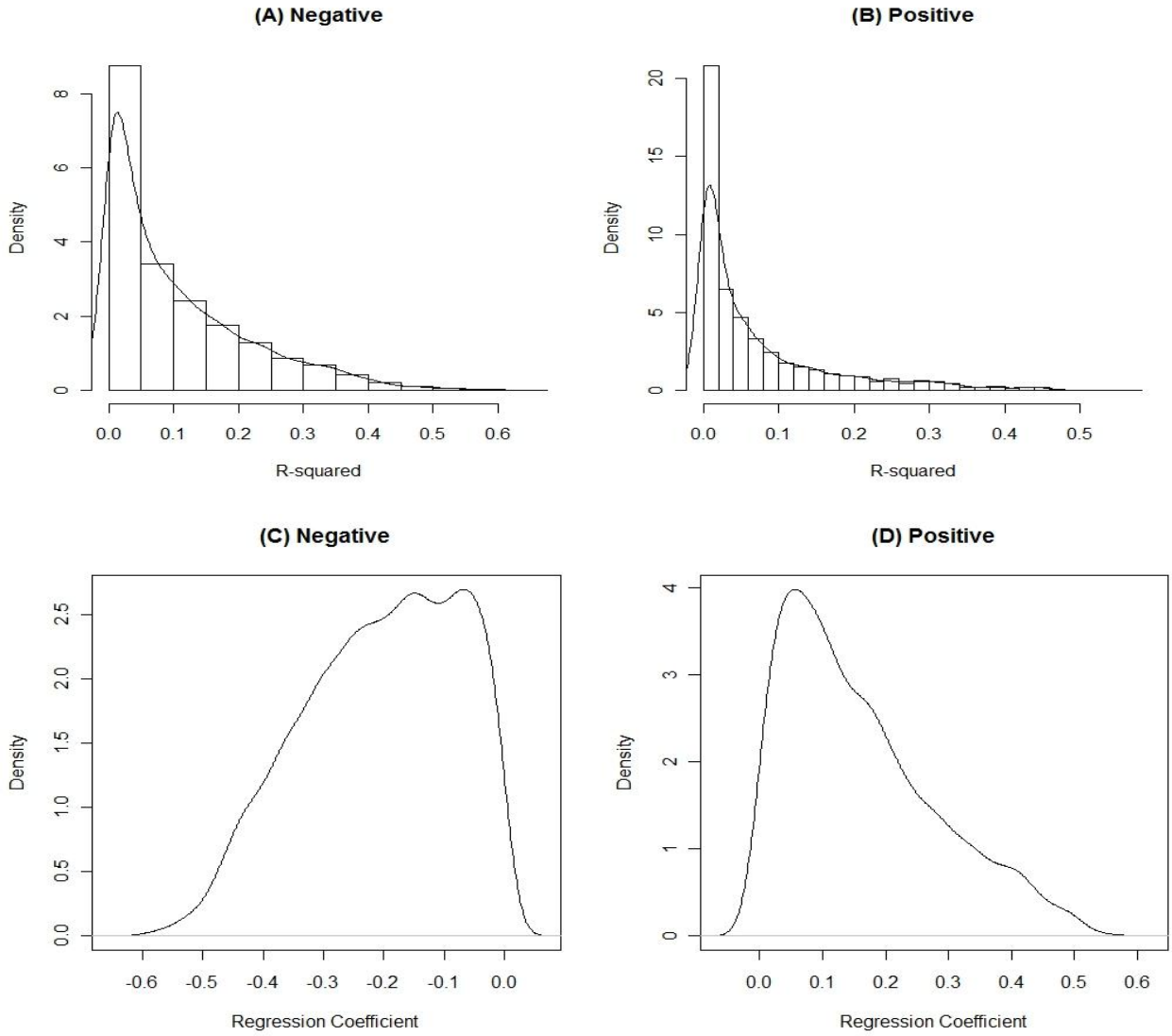

Figure 4.4: *histograms for the distributions of the regression coefficient and the R-squared values, for the gene-specific model.*

Table 4.1 summarizes the results of the top 15 genes, ranked in descending order of the regression coefficient. After multiplicity adjustment using the false discovery rate approach of Benjamini &

Hochberg (1995), 226 positive genes where found to be significantly associated with the IC50 values while 859 negative genes where significantly associated at the 5% significance level

Table 4.1: *List of the top 15 genes with their univariate regression coefficients, BH-FDR adjusted p-values (BH_{pvalue}) and R-squared.*

Positive Genes				Negative Genes			
Gene ID	β	R^2	BH_{pvalue}	**Gene ID**	β	R^2	BH_{pvalue}
222068_at	0.554	0.551	0.0025	8884_at	-0.591	0.625	0.0002
288_at	0.546	0.534	0.0025	113000_at	-0.588	0.619	0.0002
2729_at	0.542	0.526	0.0025	387338_at	-0.582	0.607	0.0002
255743_at	0.516	0.476	0.0051	573_at	-0.576	0.594	0.0003
57583_at	0.511	0.468	0.0051	55020_at	-0.572	0.586	0.0003
10553_at	0.509	0.465	0.0051	54733_at	-0.569	0.580	0.0003
51315_at	0.504	0.456	0.0051	1594_at	-0.565	0.572	0.0003
7690_at	0.503	0.453	0.0051	93622_at	-0.564	0.570	0.0003
57492_at	0.502	0.452	0.0051	26073_at	-0.549	0.540	0.0008
1604_at	0.502	0.452	0.0051	6242_at	-0.547	0.537	0.0008
4214_at	0.501	0.449	0.0051	23647_at	-0.547	0.536	0.0008
3304_at	0.500	0.449	0.0051	117246_at	-0.546	0.534	0.0008
2936_at	0.499	0.448	0.0051	10309_at	-0.544	0.531	0.0009
7408_at	0.499	0.447	0.0051	6047_at	-0.542	0.527	0.0009
81542_at	0.498	0.444	0.0051	285958_at	-0.542	0.526	0.0009

As expected, the genes with largest regression coefficients are the most significant after adjusting for multiplicity.

4.3 SPCA

The objective of this section is to obtain a joint biomarker for predicting the IC50 value of compound 352. A joint biomarker is more likely to explain more of the variability of the response than a particular gene obtained from the gene specific model. SPCA procedure was used to estimate the joint biomarker. The numbers of top genes used to construct the joint biomarker was selected by choosing the top *k* genes that maximize the R-squared value of model (3).

Figure 4.5 shows a plot of the R-squared values against the top *k* genes. The bolded point (in red) corresponds to the *k* with the maximum R-squared value. For the negative genes the maximum R-squared value of 0.76 was obtained when the joint biomarker *U(X)* included the top 15 genes, while

for the positive genes the best R-squared value of 0.75 was obtained when the top 3 genes was included to compute the joint biomarker.

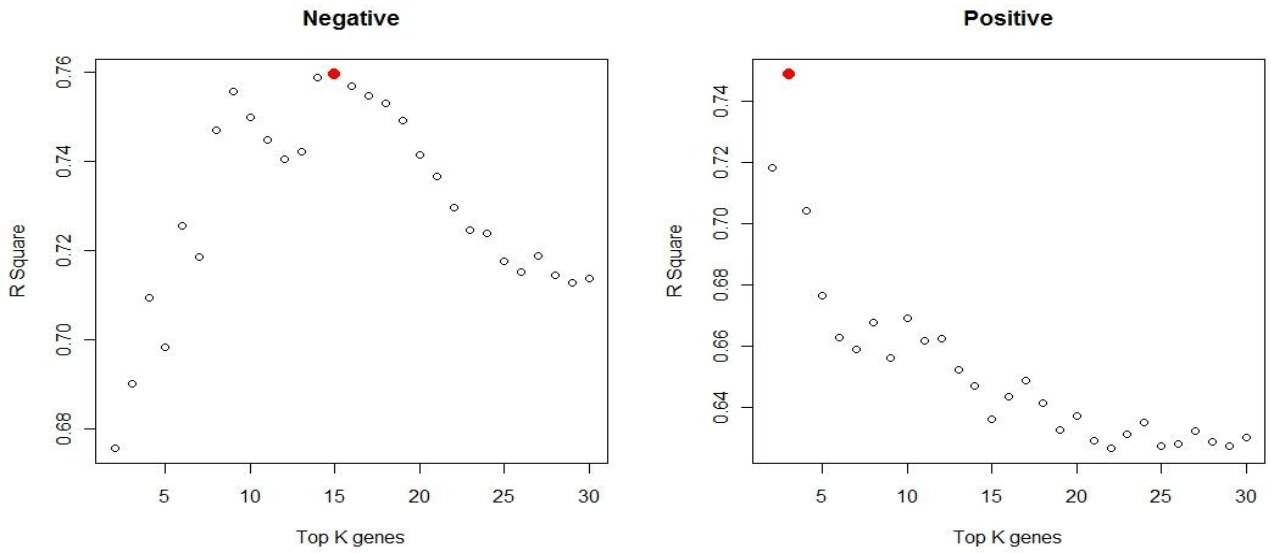

Figure 4.5: *Plot of the R-squared value against the top k genes.*

For the negative genes, after the top 15th gene was used to estimate the joint biomarker the inclusion of more genes in the model decreases in the R-squared value which implies that the addition of more genes only introduces noise to the model. The same observation holds for the positive gene (right panel). Table 6.1 in the appendix list the top 15 genes that show the highest association to the IC50 values.

Bair et al., (2006) and Tilahun et al. (2010) proposes some approaches to obtain a smaller but optimal list of genes from the top *k* genes that could best predict the IC50. One approach is to rank the genes according to their loadings on the first principal component and choose only genes with larger weights. Using this approach, the R-squared of the joint biomarker increases with the addition of the genes in the joint biomarker, but the R-squared value of 0.76 was achieved only when all the genes were included in the model. Another approach is to choose only those genes that increase the association between the joint biomarker and the IC50 values (Tilahun et al, 2010).

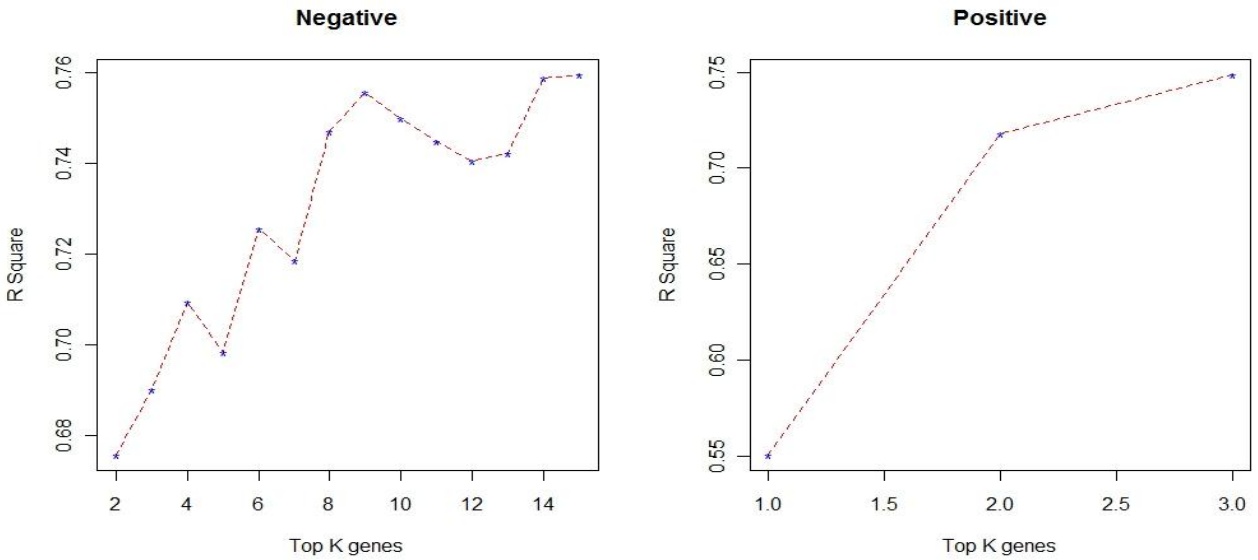

Figure 4.6: *plot of R-squared against the top k genes, showing the effect of each gene on the R-square value.*

Figure 4.6 shows that for the negative genes (left panel) when the 5th top gene was added to the joint biomarker the R-square decreases from 0.71 to 0.70; while the 6th gene increases it to 0.73, the 7th, 10th, 11th and 12th top genes also decreases the R-square value of the joint biomarker. Using this approach, the 9 genes that increases the value of the R-square were used as the joint biomarker, resulting to an R-squared of 0.79, which is higher than the 0.76 from the top 15 genes. The first principal component of the 9 genes explains 73% of their variability while the first principal components of the top 15 genes explain 74% of their variability. Table 4.2 shows the list of the Gene ID's of 9 genes that were finally used to construct joint biomarker, their loadings for the first supervised principal component, univariate R-squared values, and regression coefficients.

Table 4.2: *List of negative genes selected by SPCA for the joint biomarker with the loadings to the principal components, and univariate regression coefficients and R- squared.*

Gene ID	Loadings	Regression Coef	R-squared
8884_at	0.36	-0.591	0.62
113000_at	0.35	-0.587	0.62
387338_at	0.36	-0.582	0.61
573_at	0.36	-0.576	0.59
54733_at	0.33	-0.569	0.58
93622_at	0.31	-0.564	0.57
26073_at	0.31	-0.549	0.54
10309_at	0.33	-0.544	0.53
6047_at	0.30	-0.542	0.53

The model developed for the down-regulated (negative) joint biomarker using the 9 genes is

$$E(IC50) = 6.866 - 0.233\,U(X) \qquad (3.1)$$

where $U(X) = PC1^T \,.\, \underline{x}$,$PC1^T$ the transpose of the first supervised principal component loadings and $\underline{x}$ is a vector of gene expression values corresponding to the gene list in Table 4.2. For the positive genes, the first supervised principal component explains 72% of the variability of the reduced matrix for the top 3 genes. Table 4.3 shows the top 3 positive genes with the loadings of the first supervised principal components and their univariate R-squared and regression coefficients.

Table 4.3: *List of positive genes selected by SPCA for the joint biomarker with the loadings to the principal components, and their univariate regression coefficients and R- squared.*

Gene ID	Loadings	Regression Coef	R-squared
222068_at	0.541	0.554	0.55
288_at	0.592	0.546	0.53
2729_at	0.598	0.542	0.53

The model for the up-regulated (positive) joint biomarker is

$$E(IC50) = 6.866 + 0.441\, U(X) \qquad (3.2)$$

The relationship between the first supervised principal component and IC50 values are highly significant in the negative and positive models with p-values of 9.72×10^{-12} and 1.68×10^{-10} respectively. The model of the negative genes explains 79% of the variability of the IC50 values while the model for the positive genes explains 75%. The R-squared of model 3.1 and 3.2 may be overly optimistic since they have not been validated for prediction outside the training data, Tilahun et al., (2010) and Snee, (1977) cautioned on this.

4.3.1 Diagnostics

This section assesses models 3.1 and 3.2. The negative joint biomarker has a correlation of -0.89 with the observed IC50, while the positive joint biomarker has a correlation of 0.86. Figure 4.7 shows the scatter plot of the observed IC50 values against the estimated joint biomarker with an overlay regression lines of model 3.1 and 3.2.

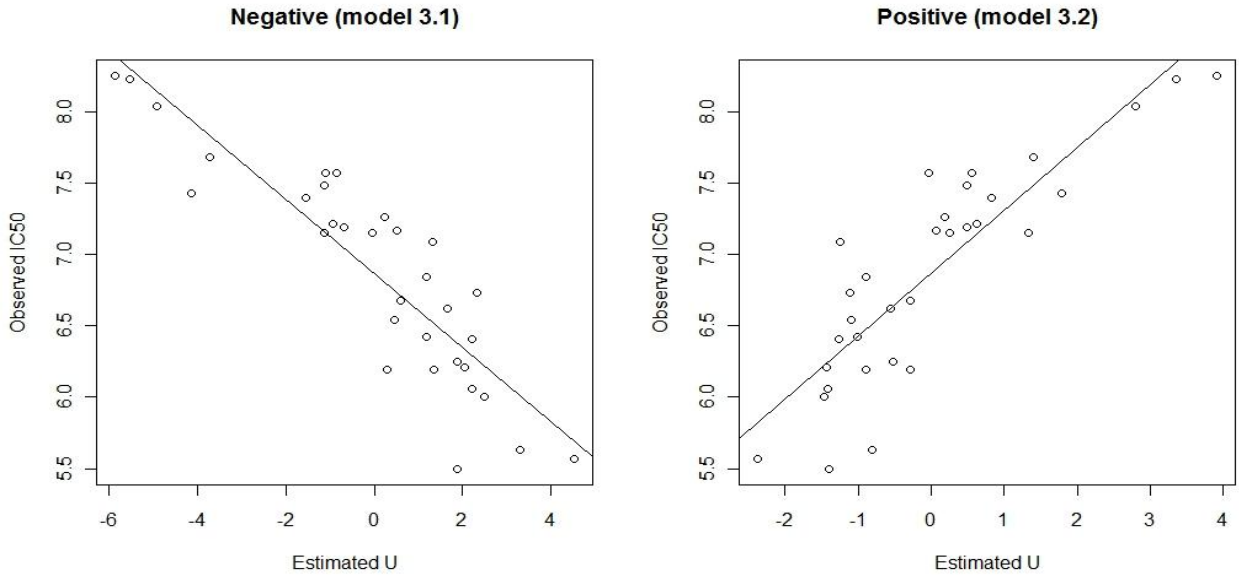

Figure 4.7: *scatter plot of the joint biomarker against the IC50 values with an overlay of the regression lines of models 3.1 and 3.2*

These scatter plots in figure 4.7 shows the point scattered around the regression line in a linear pattern thus the assumption of linearity between the joint biomarker and the IC50 can be reasonably assumed (Chattefuee & Hadi, 2006) for both models. Figure 6.1 in the appendix shows further diagnostic plots of models 3.1 and 3.2. In both models, there is no pattern to the plot of the residuals against the fitted values and they all lie within the interval -1.0 and 1.0 which is evidence of homogeneity of the residuals variance. The qq-plots do not show any mark deviation of the residuals from normality, Shapiro-wilks test confirms this assumption (Neter, et al. 2005).

4.4 Leave one out Cross-Validation (LOOCV)

LOOCV implementation of the gene-specific model produced results similar to that obtained without cross-validation. The top 15 genes were the same for both the positive and negative genes. The SPCA was implemented using the LOOCV to find the optimal number k ($k=2,\ldots, m$) of top genes. For each value of k, a LOOCV is applied to obtain an N-vector of estimated joint biomarker $\widehat{U}(X)$. The k whose estimates of the joint biomarkers $\widehat{U}(X)$has the highest correlation to the IC50 is chosen as the optimal k. Figure 4.8 shows a plot of the correlation between the first supervised principal component ($u_{k,1}$) and the observed IC50 values against the top k genes.

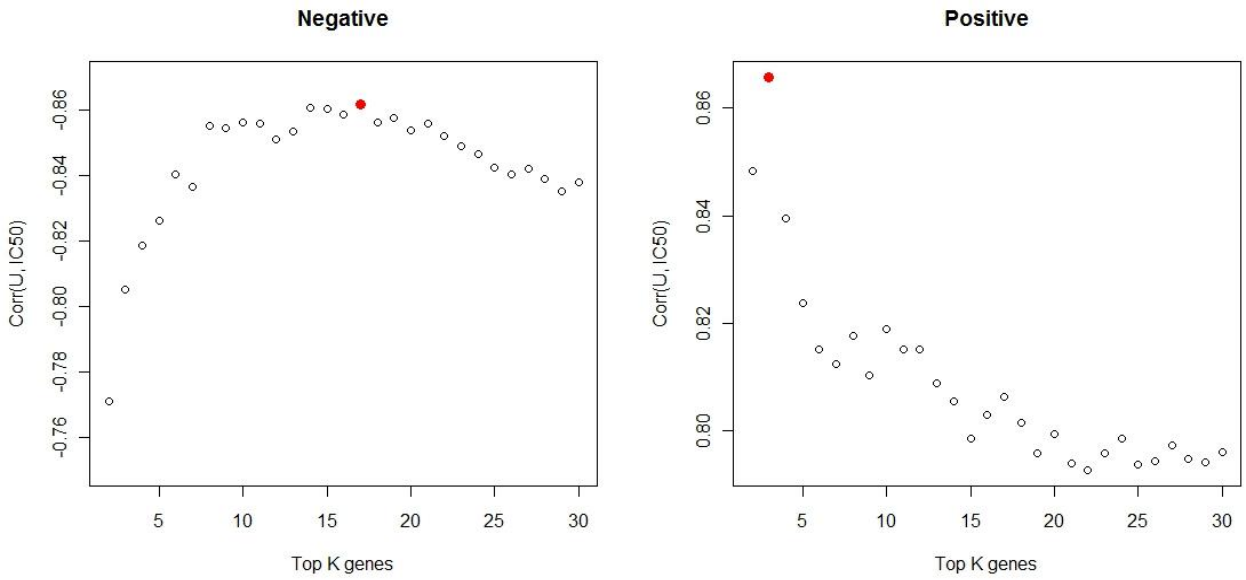

Figure 4.8: *Correlation between the First Supervised Principal Component U(X) and the IC50 values, against the top k genes. The red point is the k with the highest correlation.*

For the positive genes on implementing LOOCV with the SPCA model (3) the maximum correlation of 0.866 was obtained when the top 3 genes are used to construct the joint biomarker (Figure 4.8, right panel). This value coincides with the top 3 genes obtained in model 3.2 without the cross validation (list of genes in Table 4.3). The mean of the cross validation slope (0.441) coincides with the value estimated in 3.2. With the addition of more genes after the top 3 genes the correlation decreases rapidly, implying addition of noisy genes. While for the negative genes the maximum correlation of -0.861 was achieved when the top 17 genes were used to estimate the joint biomarker, a list of these genes is on Table 6.1 in the appendix.

Figure 4.9 (left panel) shows the scatter plot of the estimated joint biomarkers of the validation samples ($\hat{u}(x)_j = PC1^T . x_j$ where x_j is a vector of the gene expression for the *jth* sample) against the observed IC50 values with an overlay of the regression line.

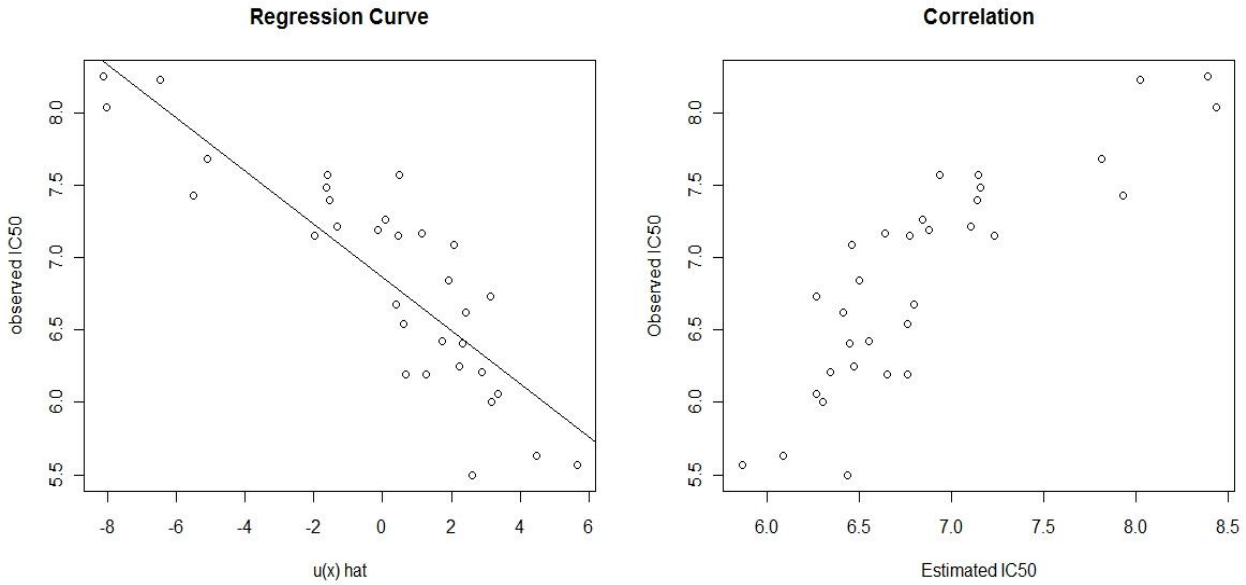

Figure 4.9*: Left panel, the scatter plot of the estimated joint biomarkers of the validation samples against the observed IC50 values. Right panel is the scatter plot of the observed IC50 values against the predicted IC50 using cross validation.*

The model obtained from the top 17 genes is,

$$E(IC_{50}) = 6.866 - 0.172U(X) \qquad (3.3)$$

where $U(X)$ is the joint biomarker of the top 17 negative genes, the slope (-0.172) is the mean of the cross validation estimates of the slopes. This model has a mean R-squared value of 0.75. For the plot in the right panel (figure 4.9), there is a linear relationship between the observed and predicted IC50 with a correlation of 0.84. Figure 6.2 in the appendix show the diagnostic plots for model 3.3. The scatter plot of the residuals against the fitted values all lie within the interval -1.0 and 1.0 thus homogeneity of the residual variance is assume satisfied. The qq-plot does not show a mark deviation from the normal line thus normality is assumed. The positive genes model with and without cross validation coincides while for the negative genes, the slope for the cross validation model is of smaller magnitude than that without cross validation. Shapiro-wilks test was used to formally confirmation the normality of the residuals.

Tilahun et al. (2010) observed that though the SPCA method aims at maximizing the association between the response and the joint biomarker, this association is always significant for different values of k. A permutation test was used to verify the association measure between the joint biomarker and the IC50 values for models 3.1, 3.2 and 3.3 and they were all found to be significant.

5. DISCUSSION AND CONCLUSION

The primary interest of this study is to find a subset of genes that can serve as a biomarker for the IC50 of compound 352. After identification and evaluation of the individual biomarkers the analysis was extended to prediction of the IC50 using a joint biomarker constructed from the selected genes. The gene specific model was used to identify and evaluate the association of individual genes to the IC50. This model assumes independence across the genes and a linear relationship between the IC50 and the gene expression profiles. Without considering the complex correlation structure genes that are univariately not associated with the IC50, but may be highly correlated to the IC50 when combined with other genes will be excluded from the model.

A linear regression model was used to quantify the measure of association between the IC50 and the genes. However, this results in loss of information because genes that are not linearly associated to the IC50 would show weak or no linear association. The SPCA method is based on linear association so the joint biomarker estimated will be constituted mainly of genes that are linearly associated to the response. Methods such as support vector regression and random forest measures non-linear association. The SPCA algorithm should be improved so that it can combine genes with linear and non-linear association when estimating the joint biomarker.

Leave one out cross validation was used for validation and model selection, otherwise model selection will be dependent on the data split due to the large number of highly correlated genes and small sample size. This will result to too many competing models. Model selected was based on the top k genes with the highest R-squared. The R-squared value is a measure of the fraction of variability around the mean of the response explained by the regression line. Using the R-squared value as ranking criteria imposes a ranking according to goodness of linear regression fit (Guyon and Elisseeff, 2003). Gene selection based on R-squared values increases the possibility of selecting the best subset of relevant genes. The use of principal component in the regression model ensures stability in the estimates of the coefficients, instability of the estimates may result due to correlation amongst the genes (Robert and Martins, 2006). The motivation for using only the first supervised principal component is because it is a linear combination with the maximum variance and contains most of the information of the reduced matrix X_k (Johnson and Wischern, 2005), 72% for the positive genes and 74% for the negative genes. The first few principal components capture almost all of the variability in the covariance space, and the use of only the first principal component would ease interpretation.

Bair et al. (2006), Robert and Martins (2006) and Chen et al., (2008) proved the consistency of the SPCA method in selecting the "correct" genes using both simulated and real data. SPCA assigns

principal components loadings to the selected genes without regards to the outcome variable. Looking at our results there are genes with higher weights but they pull down the R-squared value of the joint biomarker. The 7th top gene has loadings (0.274) greater than that for the 6th top gene (0.244), but yet when added to the model it pulls down the R-squared value of the joint biomarker model (3) while the 6th gene with lower loadings increases the R-squared (Figure 4.6, left panel)

There are some possible limitations of the SPCA method. The magnitude of the weights to the principal components does not take into account the association with the response. In a situation where the IC50 values are marginally independent of a subset of genes but jointly related to them, SPCA would fail to identify this subset of genes (Robert and Martin, 2006). SPCA does not take into account genes that are non-linearly associated to the outcome.

The subset of genes identified as biomarker for the IC50 values of compound 352 are those listed in Table 4.3 for the up-regulated (positive) genes and the top 17 listed in Table 6.1 for the down-regulated (negative) genes. The prediction model for the down-regulated (negative) gene is;

$$E(IC_{50}) = 6.866 - 0.172U(X)$$

Hence, for a unit increase of the expression value of the joint biomarker for the 17 down-regulated genes, the IC50 of compound 352 is decreased by -0.172. For the up-regulated (positive) gene model is

$$E(IC50)_{pos} = 6.866 + 0.441\,U(X)$$

Hence, for a unit increase of the expression value for the joint biomarker of the 3 up-regulated genes the IC50 of compound 352 is increased by 0.441. The joint biomarker is estimate by $U(X) = PC1^T.x$ where $PC1^T$ is the transpose of the principal component loadings and x is the gene expression profiles of the relevant subset of genes.

The prognostic biomarkers in Tables 6.1 and 4.3 were selected based on their association to the IC50 value of compound 352. The genes corresponding to the Gene ID's can be identified and the associated cellular process responsible for the phenotype determined and interpreted. Models 3.2 and 3.3 can be used for predictions.

6. APPENDIX

Table 6.1: *List of top 17 negative genes obtained using SPCA method with their univariate standardized regression coefficients and R-squared to the IC50 values using the gene specific model in section 3.2. The last 2 genes are the genes that where added when LOOCV was implemented in the gene selection procedure.*

Gene ID	Loadings (top 15)	Loadings (top 17)	Regression Coefficient	R-Squared
8884_at	0.279	0.259	-0.591	0.625
113000_at	0.270	0.253	-0.588	0.619
387338_at	0.277	0.263	-0.582	0.607
573_at	0.275	0.259	-0.576	0.594
55020_at	0.282	0.265	-0.572	0.586
54733_at	0.244	0.229	-0.569	0.580
1594_at	0.274	0.257	-0.565	0.572
93622_at	0.220	0.207	-0.564	0.570
26073_at	0.238	0.223	-0.549	0.540
6242_at	0.253	0.238	-0.547	0.537
23647_at	0.264	0.251	-0.547	0.536
117246_at	0.267	0.251	-0.546	0.534
10309_at	0.256	0.243	-0.544	0.531
6047_at	0.210	0.196	-0.542	0.527
285958_at	0.251	0.232	-0.542	0.526
84881_at		**0.242**	**-0.541**	**0.524**
79230_at		**0.242**	**-0.538**	**0.519**

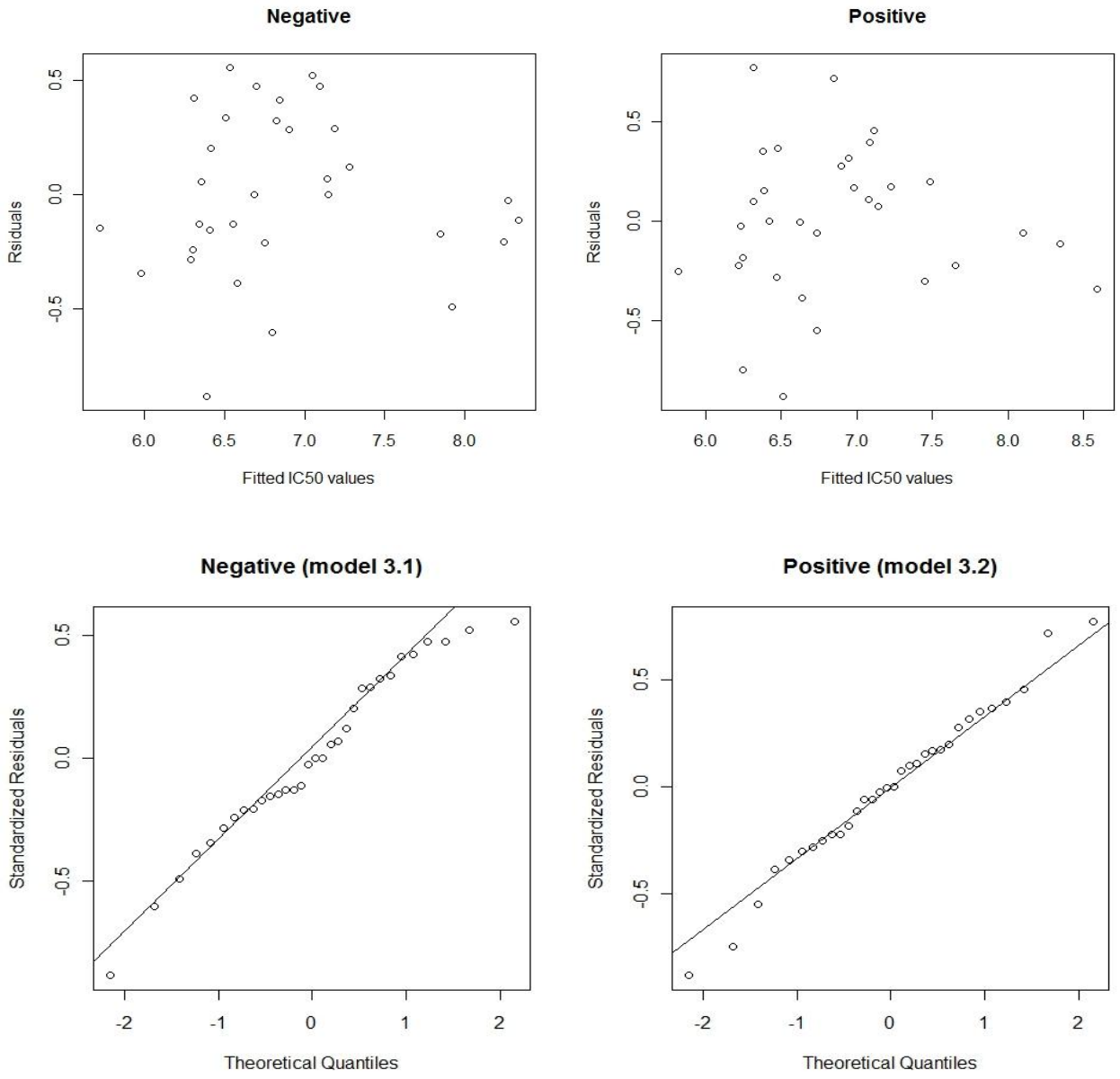

***Figure 6.1**: Diagnostic plots of the SPCA models 3.1 and 3.2*

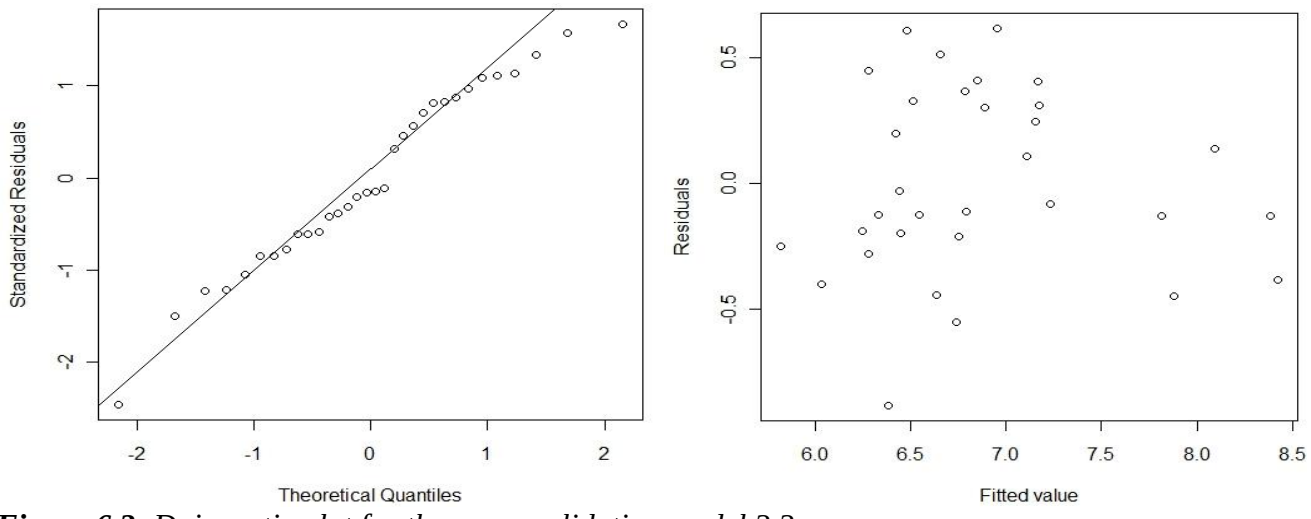

***Figure 6.2:** Daignostic plot for the cross-validation model 3.3*

7. REFERENCES

1. Bair, E., Hastie, T., Paul, D. and Tibshirani, R., (2006). Prediction by supervised principal components. *Journal of the American Statistical Association,* 101, 119-137.
2. Benjamini Y. & Hochberg Y., (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist Soc B,* 57(1), 289-300.
3. Bovelstad, H., Nygard, S., Storvold, H., Aldrin, M., Borgan, O., Frigessi, A. & Lingjaerde, O., (2007). Predicting survival from microarray data – a comparative study. *Oxford University Press*, 23(16), 2080-2087.
4. Carroll, K., (2007). Biomarkers in drug development: friend or foe? A personal reflection gained working within oncology. *Pharmaceut. Statist.*, 6, 253–260.
5. Chattefuee, S. &Hadi, A., (2006). *Regression Analysis by Example - 4th ed*. John Wiley & Sons, Hoboken, New Jersey.
6. Chau, C., Rixe, O., McLeod, H., &Figg, W., (2008). Validation of Analytic Methods for Biomarkers Used in Drug Development. *Clin Cancer Res,* 14(19), 5967- 5976.
7. Chen, Xi, Wang, L., Smith, J., and Zhang, B., (2008). Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Oxford University Press*, 24(21), 2474-2481.
8. Efron B., (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc*., 78(316), 331.
9. Frank, R. & Hargreaves, R., (2003). Clinical biomarkers in drug discovery and development. *Nature Publishing Group*, 2, 566-580.
10. Hastie, T., Tibshirani, R. & Friedman, J., (2008). *The elements of statistical learning - 2nd ed*. Springer series in statistics, Stanford, California.
11. Guyon,I., & Elisseeff, A., (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research.* 3, 1157-1182.
12. Johnson, D. & Wichern, W., (2007). *Applied Multivariate Statistical Analysis, 6th ed.* Pearson Prentice Hall: Upper Saddle River. New Jersey.
13. Jolliffe, I., (2002). *Principal Component Analysis – 2nd ed.* Springer, New York.
14. Lin, D., Shkedy, Z., Molenberghs, G., Talloen, W., Goelhmann, H. &Bijnens, L., (2010). Selection and evaluation of gene-specific biomarkers in pre-clinical and clinical microarray experiments. *Online Journal of Bioinformatics*, 11 (1): 106-127.
15. Neter, J., Kutner, M.&NachtsheimC., (2005). *Applied Linear Statistical Models, 5th Edition.* McGraw-Hill International Edition.

16. Nguyen, D. & Rocke, D., (2002). Partial least squares proportional hazard regression for application to DNA microarray survival data. *Oxford University Press*, 18(12), 1625-1632.
17. Nikas, J. & Low, W., (2011). ROC-supervised principal component analysis in connection with the diagnosis of diseases. *Am J Transl Res*, 3(2), 180-196.
18. Ransohoff, D., (2004). Rules of evidence for cancer molecular marker discovery and validation. *Nature Reviews/Cancer*,4, 309-313.
19. Refaeilzadeh P., Tang L., and Liu H., (2007). On comparison of feature selection algorithms. *In AAAI-07 Worshop on Evaluation Methods in Machine Learning II*.
20. Roberts, S. & Martins, M., (2006). Using supervised principal components analysis to assess multiple pollutant effects. *Environ Health Perspect*, 114(12), 1877-1882.
21. Segal,M., Dahlquist,K., & Conklin, B., (2003). Regression approaches for microarray data analysis. Journal of Computational Biology, 10, 961- 980.
22. Sohn I., Owzar K., Lim J., George S., Cushman S., & Jung S., (2011). Multiple testing for gene sets from microarray experiments. *BMC Bioinformatics*, 12(209).
23. Snee, R.,(1977). Validation of regression models: Methods and examples. *Technometrics*, 19(4), 415-428.
24. Sreekumar J., & Jose K., (2008). Statistical test for identification of differentially expressed genes in cDNA microarray experiments. *Indian Journal of Biotechnology*. 7, 423-436.
25. Tibshirani, R., (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1), 267-288.
26. Tilahun, A., Lin, D., Shkedy, Z., Geys, H., Alonso, A., Peeters, P., Talloen W., Drinkenburg, W., Göhlmann, H., Gorden, E., Bijnens, L., &Molenberghs, G., (2010). Genomic biomarkers for depression: feature-specific and joint biomarkers. *American Statistical Association Statistics in Biopharmaceutical Research*, 2(3), 419-434.
27. Wall, M., Dyck P., & Brettin, T., (2001*).* Singular value decomposition analysis of microarray data . *Oxford University press, 17(6), 566–568.*
28. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Marks, J., &Nevins, J.,(2001). Predicting the clinical status of human breast cancer using gene expression profiles. *PNAS*, 98, 11462-11467.
29. Zhao, Y. & Simon, R., (2010). Development and validation of predictive indices for continuous outcome using gene expression profiles. *Cancer informatics*, 9, 105 – 114.

8. CODES

The R statistical software package was used for the analysis in this report. The *multtest* package was used to implement the FDR-BH multiplicity correction, *prcomp()* was used for the singular value decomposition of the expression matrix. Codes for the implementation of the gene specific, SPCA methodology both with and without cross validation were self develop. Below is a description of the main functions:

*regression():*This function performs a univariate simple linear regression on a give gene expression dataset and response variable, it takes 2 arguments:

```
regression <- function(data, response) {
        beta<- r.sqr<-t.stats<- p.value<-rmse<-vector()
        for (i in 1:dim(data)[1]) {
                fit<-lm(y~data[i,])
                beta[i] <- fit$coef[[2]]
                r.sqr[i]<-summary(fit)$r.squared
                p.value[i]<-anova(fit)$"Pr(>F)"[1]
                t.stats[i]<-coef(summary(fit))[,"t value"][[2]]
        }
        result<-list(Beta=beta, R.square=r.sqr, P.value=p.value, T.stat<-t.stats)
        return(result)
}
```

Arguments:

- Data: a *p x N* matrix of expression values with p the number of genes and *N* the samples.
- Response: an *N*-vector of quantitative response associated to the *N* sample

The values returned are p-vectors of the results of the univariate regression of the response to the gene expression matrix Data:

- Beta: a vector of univariate regression coefficients
- R.square: a vector of the univarite R-squared values
- T.stat:a vector of the p test statistic for the univariate regression coefficients
- P.value: the vector of p-values.

topk():The*topk* function performs a simple linear regression of joint biomarker models given a number k corresponding to the number of top genes to use in the joint biomarker. For a given k, it constructs the first supervised principal component using SVD by the *prcomp()* function of the top k genes and fit a linear model using the *lm()* function.:

```
topk<- function(data, y, beta, k) {
        topk.genes<- beta[1:k,1]
        k.matrix<- data[topk.genes,]
```

```
        pc <- prcomp(t(k.matrix), cor = TRUE)$x[,1]
        fit <- lm(y~pc)
        r.sqr<- summary(fit)$r.squared
        p.value<-anova(fit)$"Pr(>F)"[1]
        result <- list(First.pc=pc, R.Square=r.sqr, P.value=p.value, k=k)
        return(result)
}
```

Arguments:

- Data: a *p x N* matrix of expression values with p the number of genes and *N* the samples.
- Y: an *N*-vector of quantitative response associated to the *N* sample
- beta: An ordered (descending) index of a measure of association obtained from a gene-specific model.
- K: a whole number representing the number of genes used to compute the first supervised principal component.

The values returned

- First.pc: the first supervised principal component used to fit the joint biomarker regression model
- R.Square: the R-squared value of the regression model
- P-value: the p-value of the slope (regression coefficient)
- K: a scalar, depicting the number of top genes used to compute the first principal components.

regression.cv():This function is the leave on out cross validation of *regression()* function above.

```
regression.cv <- function (data, y) {
        beta<-r.sqr<-vector()
        for (i in 1:dim(data)[1]) {
                beta.cv<-r.sqr.cv<-vector()
                x<-data[i,]
                for (j in 1:32) {
                        fit<-lm(y[-j]~x[-j])
                        beta.cv[j] <- fit$coef[[2]]
                        r.sqr.cv[j]<-summary(fit)$r.squared
                }
                r.sqr[i]<-mean(r.sqr.cv)
                beta[i]<-mean(beta.cv)
        }
        result<-list(Beta.cv=beta, R.square=r.sqr)
        return(result)
}
```

loocv(): This function performs a leave one out cross validation of the SPCA methodology. It fits a simple linear regression of joint biomarker models for a given number of top genes k, it computes the first principal component of the top k genes fits a linear model (*lm()*) while omitting one of the observation from the model and returns the regression summary results. It takes the same arguments like *topk()* function, but in this case k is a sequence of whole numbers rather than a scalar.

```
loocv<- function (matrix, genes, y, k) {
        y.hat<-u.x<-r.sqr<-p.value<-vector()
        y.hat.mat<- r.sqr.mat<-u.x.mat<-p.value.mat<-vector()
        for (i in 1:length(k)) {
                feat <- genes[1:k[i],1]
                mat <- matrix[feat,]
                for (j in 1:32) {
                        mat.cv<-mat[,-j]
                        y.cv<-y[-j]
                        pc <- prcomp(t(mat.cv), cor = TRUE, center=F)
                        x<-pc$x[,1]
                        fit <- lm(y.cv~x)
                        alpha <- fit$coef[[1]]
                        beta <- fit$coef[[2]]
                        r.sqr[j]<- summary(fit)$r.squared
                        p.value<- anova(fit)$'Pr(>F)'[1]
                        u.x[j] <- t(pc$rotation[,1]) %*% mat[,j]
                        y.hat[j] <- alpha + beta*u.x[j]
                }
                u.x.mat<-cbind(u.x.mat, u.x)
                y.hat.mat<-cbind(y.hat.mat, y.hat)
                r.sqr.mat<-cbind(r.sqr.mat, r.sqr)
                p.value.mat<-cbind(p.value.mat, p.value)
        }
        result<-list(Predict=y.hat.mat,R.square=r.sqr.mat,U.X=u.x.mat,P.value=p.value.mat)
        return(result)
}
```

Values returned

- Predict: a matrix of *N* x *K* consisting of the estimated IC50 value of the observation used as the test set for the various values of *k*, where *N* is the number of observations, and *m* the length of the sequence of top k genes to perform the leave one out cross validation.
- R.square: an *N x m* matrix of the R-squared values of the joint biomarker model, for each value of k and each omitted observation
- U.X: the value of the joint biomarker of the test (omitted) observation computed using the loadings of the first principal components calculated for each value of k

Gene-specific model

```
neg<-regression(gc14.neg, y)
pos<-regression(gc14.pos, y)
```

This snippet uses the *regression()* function to obtain the regression statistics for the *p* gene-specific models. Where y is the IC50 values and gc14.neg (gc14.pos) is the expression matrix of the negative (positive) genes.

Supervised principal component analysis

The script below uses the *topk()* function to obtain a sequence of R-squared for the negative genes values corresponding to a sequence of top k genes used to construct the first supervised principal component. Similar code snippet is applied for the positive genes.

```
top.genes<- seq(2, 30, 1)
r.sqr.neg.spca<- vector()
for (i in 1:length(top.genes)) {
        k <- top.genes[i]
        top <- topk(x=gc14.neg, y=y, beta=beta.neg, k=k)
        r.sqr.neg.spca[i] <- top$R.Square
}
```

Leave one out cross validation

The snippet below was used to perform the leave one out cross validation for the gene-specific model and the SPCA model.

```
neg.loocv<-regression.cv(gc14.neg, y)
pos.loocv<-regression.cv(gc14.pos, y)

neg.spca.loocv<-loocv(matrix=gc14.neg, genes=beta.neg, y=y, k=top.k)
pos.spca.loocv<-loocv(matrix=gc14.pos, genes=beta.pos, y=y, k=top.k)
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Large scale prediction of phenotipic variables using gene expression data

Richting: **Master of Statistics-Bioinformatics**
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,

Ndah, Elvis

Datum: **21/09/2011**