Predictive Gene Signature Selection for Adjuvant Chemotherapy in Non-Small Cell Lung Cancer Patients

by

Li Liu

A practicum report submitted to the
Department of Public Health Sciences
in conformity with the requirement for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
August 2014

Abstract

Objectives: BR.10 is a randomized controlled clinical trial that demonstrated the benefit of adjuvant chemotherapy in early stage non-small cell lung cancer patients with tumor completely resected. The purpose of this study is using microarray expression profiling to identify the subgroup of the patients who are truly benefited from the chemotherapy.

Methods: Gene expression profiling was conducted from the 133 prospectively collected frozen tumor samples (62 observations, and 71 chemo). The raw microarray data were preprocessed by normalization and adjusting batch effects. Gene probesets that predict of chemotherapy's benefit were selected using Cox's regression model fitted by modified covariates of chemotherapy and gene probesets interaction without main effect, and the elastic net regularization for variable selection. Bootstrap samples were fitted to record the frequencies of each probeset appeared in the model selection. The predictive signature was selected by combining the most appeared probesets (1^{st} principal component of the selected genes) in the multivariable model fitting that best separate patients between treatment arms. The signature's predictive effect was tested in the testing dataset.

Results: A 34-gene signature separates the patients in low predictive score group between two treatment arms, and the patients in the low predictive score group are beneficial to the chemotherapy in the testing set, where significant different survival outcomes (hazard ratio 0.13; 95% CI, 0.01 to 1.14; p-value = 0.03 of log-rank test) for the patients in chemo vs. observation treatment arms were observed. The selected gene signature and the purposed method were internal validated by the datasets of randomly selected 71 patients with replacement from all the samples. Among the 200 times validations, the log-rank test p-values are less than 0.05 in more than 90% validations, and the hazard ratios are less than 0.4 in almost all the validations.

Conclusion: 34-gene signature is able to predict the survival benefits from adjuvant chemotherapy for early stage non-small cell lung cancer patients, and save those patients who may not benefit from chemotherapy from suffering side effect that the treatment may induce.

Acknowledgement

Firstly, I would like to express my gratitude to my supervisor, Dr. Keyue Ding, for giving me the opportunity to study and work in the area I am interested in. Thank you for all the guidance, support, patience, and constant encouragement throughout my graduate program. Without your help, I would not have learned as much as I have here, and start my career in biostatistics, a complete new area for me.

I want to express my thanks to other professors, staffs and TAs in the Department of Public Health Science and the Department of Mathematics and Statistics, particularly to Dr. Wenyu Jiang, Dr. Devon Lin, Dr. Dongsheng Tu, Dr. Michael McIsaac, Dr. Paul Peng, Dr. Bingshu Chen, Jina Zhang, and Andrew Day for their help during my study and preparing the program.

Thank you my biostatistics fellows Nelson, Haoyu, and other students in the both departments. You give me memorable and joyful experiences during my intensive course study. I wish you all the best of your future careers.

I acknowledge the funding supports from Dr. Keyue Ding and Queen's University Graduate Scholarship for my Master's degree.

Last but not least, I also would like to give my special thanks to my husband and lovely kids, for their love and support. I wouldn't have been able to do this without you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Lung cancer is the leading cause of death from cancer and second most commonly diagnosed cancer. Based on Canadian cancer statistics (2014) [1], 20,500 Canadian will die from lung cancer, which represents 27% of all cancer death in 2014, and 26,100 Canadian will be diagnosed with lung cancer, which represents 14% of all new cancer cases in 2014. Non-small cell lung cancers (NSCLC) are accounts for about 85 to 90% of all lung cancers. The overall 10-year survival rate of patients with NSCLC is only 8-10%. About 25% to 30% of patients with NSCLC have stage I disease. Although the survival rate of those early stage lung cancer patients is relatively high, among them, 35% to 50% will relapse within 5 years even after an apparent complete resection [2, 3]. Several randomized trials had demonstrated modest benefits (4 – 15%) of 5-year survival with adjuvant chemotherapy (an additional chemotherapy treatment to lower the risk of cancer returning) in resected NSCLC [3, 4]. However, the conclusion of a randomized clinical trial is respect to the average treatment effect on the entire population. A treatment may only be beneficial for a group of patients but not for the others. Moreover, the response to standard chemo in lung cancer varies. For some patients, the treatment may cause serious adverse effects and potential detrimental effects [3, 4]. Therefore, it would be very helpful to

identify the subgroup of patients who are likely benefit from the treatment, and be spared the side effects of unnecessary treatment, so as to develop strategies for personalized medicine.

Tumor sample routinely collected accompanying cancer clinical trials, and pretreatment gene expression profiles of cancer possess the information about the disease and its sensitivity to therapy. Based on genome-wide measurement of expression levels, appropriate statistical analysis can extracted the information to predict patients' outcome and their response to treatment.

1.2 Objectives of this work

The primary objective of this project is using the 133 patients who had gene microarray data available from NCIC Clinical Trial Group BR.10 trial to identify a group of gene probesets (gene signature) that predicts the survival benefits from adjuvant chemotherapy on early stage non-small cell lung cancer patients.

1.3 Major analysis

In this study, gene expression profiling was conducted on mRNA from 133 frozen BR.10 tumor samples. Affymetrix HG-U133A microarray data are available for each patient. Raw microarray data are highly noisy when multiple arrays are involved. There exist systematic variations between RNA samples, which are not due to biological difference of samples, but technical reasons during microarray chips production. Microarray data preprocessing steps attempt to remove such variations that affect the measured gene expression level so that the biological difference among the samples can be distinguished. Two steps of microarray data preprocessing were involved in this work: normalization and batch effects removal. The

objective of the normalization is to adjust the gene expression values of all genes so that the ones are not really differentially expressed have similar values across the arrays [5]. Robust Multiarray Average (RMA) method [6] is one of the most popular normalization methods. Another step of microarray data preprocessing is adjusting the batch effects. The batch effects represent the systematic technical differences when samples are processed and measured in different batches, and which are also unrelated to any biological variation recorded during the experiment. The 133 microarray chips in this study were produced at different time, so we applied a few methods, including batch mean centering, gene standardization, empirical Bayes, and distance-weight discrimination, to adjust the batch effects.

After normalization and batch effect removal, microarray data sets were linked with BR. 10 clinical data for statistical analysis. The 133 patients in BR.10 were randomly split into training and testing sets. The probesets were firstly preselected by univariate survival analysis at Wald's test p-value < 0.05. Then predictive gene signature was selected from the training sets by bootstrap - Cox's regression models using elastic net regularization method for variable selection. Based on the frequency that the probesets been most selected in the bootstrapped samples, predictive score of each patient was developed by the interaction term of the treatment and the 1^{st} principal components of the selected gene expressions. Using the 1/3 and 2/3 cut-off points of the predictive scores, the patients were divided into low, middle, and high predictive score groups. The cut-off points and principal component analysis transformation matrix from the training set were applied to the testing set to predict the adjuvant chemotherapy treatment effects on the survival. Moreover, we also performed internal validation by using the selected predictive gene signature to the data sets of randomly selected a number of patients with replacement from all the 133 samples. All the analyses were performed using R language.

Chapter 2

Microarray Data Preprocessing

Before linking gene expression values to clinical outcome, the raw microarray data have to be preprocessed, i.e., normalization and adjusting batch effect. These preprocessing steps are necessary because there is likely to be observed intensity imbalance between array samples or between groups (batches) of array samples. This imbalance is not due to the biological difference between samples, but occurs for a variety of technical reasons during array chips production, such as different operating conditions and operation technicians during arrays production, different production time and so on. The purpose of preprocessing is to remove those variations due to technical reasons so that the genes which are biological differentially expressed can be explored.

2.1 Normalization

Differences in treatment of two microarray samples, especially in labeling and in hybridization, bias the relative measures on any two chips [5]. Normalization is the attempt to compensate for systematic technical difference between chips, to see more clearly the systematic biological differences between samples. In other words, the objective of normalization is to adjust the gene expression values of all genes so that the ones that are not really differentially expressed have similar values across the arrays.

2.1.1 Normalization methods

A typical microarray experiment involves the hybridization of an mRNA molecule to the DNA template from which it is originated. The amount of mRNA bound to each site on the array indicates the expression level of the various genes. Affymetrix Genechip® arrays used in this work are currently most widely used high-throughput technologies for the genome-wide measurement of expression profiling. The technology includes perfect match (PM) and mismatch (MM) probe pairs as well as multiple probes per gene to minimize mis-hybridization and/or cross-hybridization problems during array chips production. The probes that match a target sequence exactly, called the perfect match, and partner probes, which differ from the reference probes only by a single base in the center of the sequence, are called the mismatch probes. Normalization procedures combine multiple probe signals into a single absolute call. They usually involve three steps: background correction, normalization, and gene probe set summarization [7]. The background correction removes local artifacts and noise, so the measurements are not so affected by neighbouring measurement. Then certain linear or nonlinear normalization algorisms are applied, to remove array effects so the measurements from different arrays are comparable. Finally, the summarization step combines probe intensities across arrays so final measurement represents gene expression level. Various methods have been devised for each of these three steps and thus a great number of possible combinations exist. Among them, a few popular Affymetrix microarray normalization methods including Robust Multiarray Average (RMA), GeneChip Robust Multiarray Average (GC-RMA), and Affymetrix MicroArray Suite (MAS5) [8] were used in this study.

RMA normalization [6] is one of most popular and simplest normalization method. It performs background correction by using only PM information on each chip and inter-chip

quantile normalization. This method assumes that the distribution of gene abundance is nearly the same in all samples. For convenience the pooled distribution of probes on all chips are taken to normalize each chip. The original gene expression value is replaced by the quantile value of the reference chip. Then the probeset summarization combines probes for one probeset into a single number by Tukey's median polish method. GC-RMA method uses both PM and MM information for the normalization. MAS5 method normalizes each chip using weight (Tukey Biweight Estimate) based on its intensity difference from the mean. All these three normalization methods return expression levels on log2 scale. Comparing to MAS5, RMA/GC-RMA methods have less variance at low expression values and less false positive, but quality control after normalization is difficult, and quantile normalization has the risk of over-fitting and hiding real difference.

2.1.2 Analysis and results

BR.10 gene expression profiling was completed in 133 patients using U133A oligonucleotide microarray (Affemetrix, Santa Clara, CA) [4]. Of patients with microarray profile, 62 were in observation group, while 71 received adjuvant chemotherapy. The 133 raw microarray *.cel files can be download from the Gene Expression Omnibus website [9] (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse14814).

Two microarray analysis tools were used for microarray normalization in this work: R language based Bioconductor and BRB-Array Tools. Bioconductor is an open source, open development project to provide tools for the analysis of high-throughput genomic data, which contains numeric genome analysis packages. BRB-Array Tools is also free software developed by Biometric Research Branch of National Cancer Institute, which provides a set of tools for the

analysis of DNA microarray data. The 133 microarray raw data *.cel files were read in and normalized by RMA method using Bioconductor 'Affy' package. For comparison purpose, a data set using same background correction and summarization as RMA but without applying any normalization algorism was also implemented. The GC-RMA and MAS5 normalization methods were fulfilled in BRB-Array Tools.

Figures 1 and 2 show the distribution of the log intensities of all microarrays using boxplot and histograms/density plots. It can be seen that comparing to un-normalized data, the datasets generated by GC_RMA and MAS5 method are highly skewed to right hand side, while the one by RMA method is also slightly skewed. Figure 3 presents single array boxplots. Even after background correction, without applying any normalization method, the microarray distribution still show large variation across arrays, and after normalization, the distribution of the each array samples are relatively consistent.

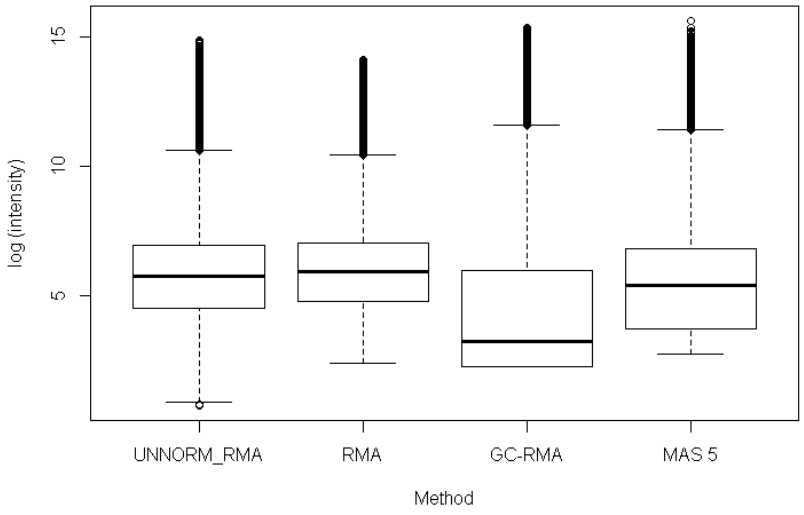


Figure 1 Boxplot of log intensity of all microarrays using different normalization methods (UNNORM_RMA: without normalization but using the same methods as RMA for background correction and summarization).

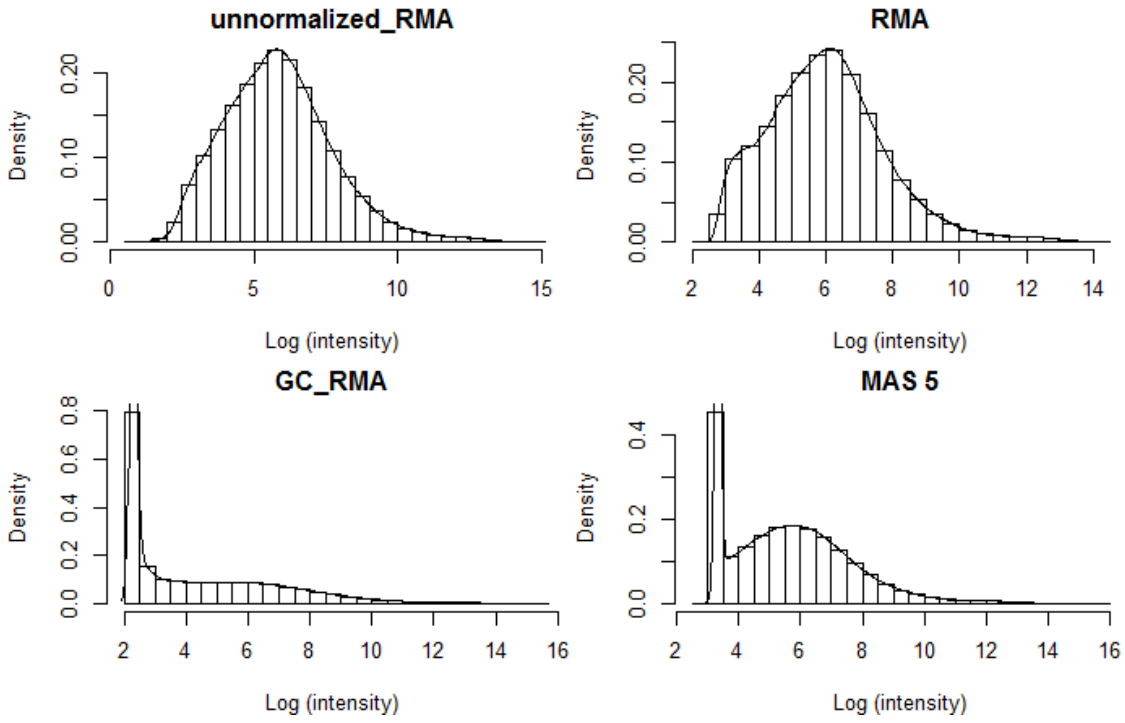


Figure 2 Histograms and density plots of log intensity of all microarrays using different normalization methods (unnormalized_RMA: without normalization but using the same methods as RMA for background correction and summarization)

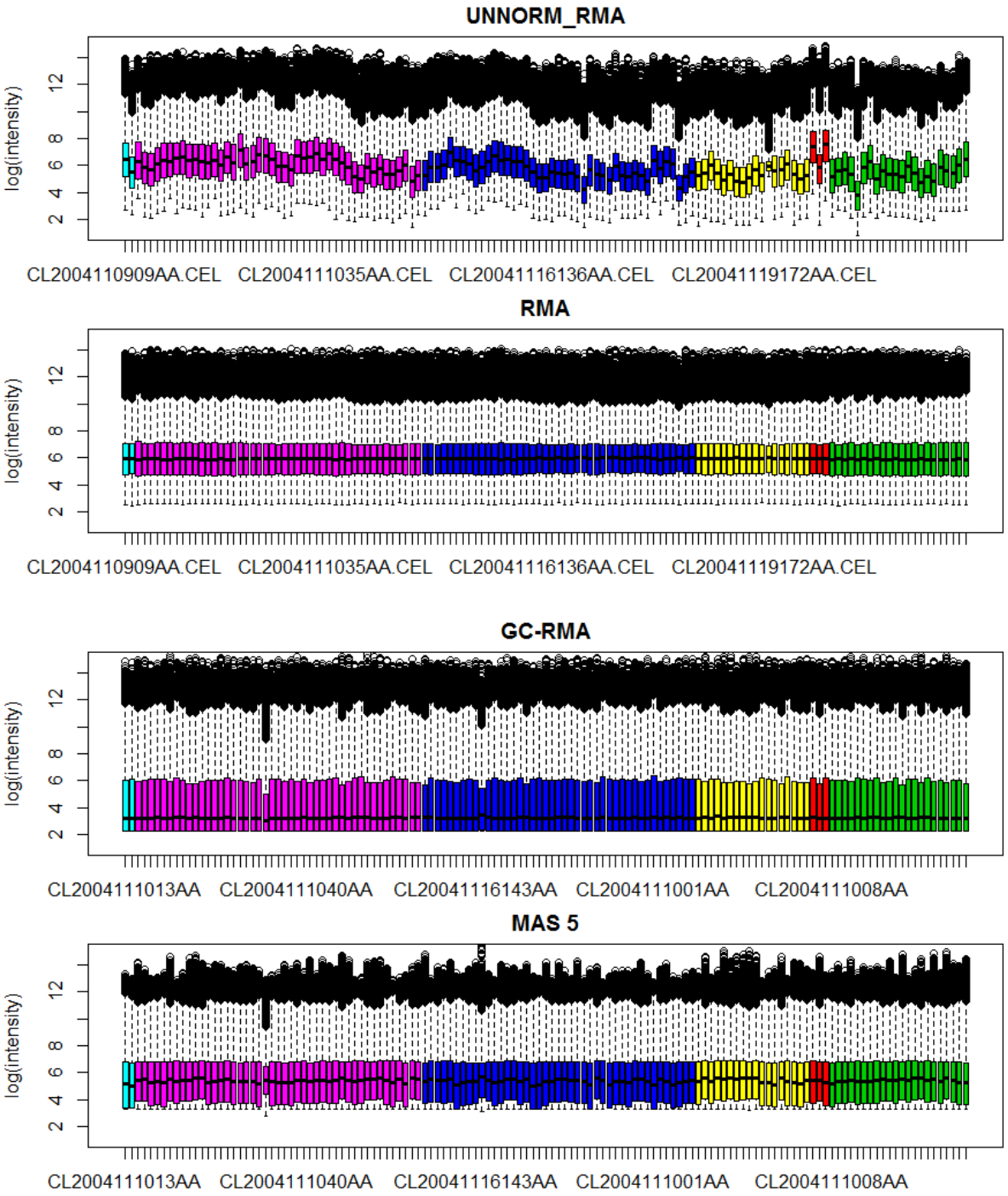


Figure 3 Boxplots of single microarray samples (colored by batch based on microarray gene chip production date). (unnormalized_RMA: without normalization but using the same methods as RMA for background correction and summarization)

2.2 Batch effects

The batch effects represent the systematic technical differences when samples are processed and measured in different batches and which are unrelated to any biological variation during experiments [10]. Here the batch denotes a collection of microarrays (or samples) processed at the same site over a short period of time using the same platform and under approximately identical conditions. During microarray production, only limit numbers of samples can be amplified and hybridized at one time, and replicate samples may be generated several days or month apart. Therefore, non-biological experimental variation or 'batch effects' are commonly observed across multiple batches of microarray experiments, often rendering the task of combining data from these batches difficult. Batch effects can be caused by many factors including the batch of amplification reagent used, the time of day when an assay is done, or even the atmospheric ozone level [11]. Moreover, batch effects are also inevitable when new samples or replicates are incrementally added to an existing array data set or in a meta-analysis of multi studies that pools microarray data across different labs, array types or platforms. The ability to combine microarray data sets is advantageous to increase statistical power to detect biological phenomena from studies. However, normalization could not remove batch effects. In the signal array boxplots shown in Figure 3, each color of the boxplots corresponds to a batch. Although after normalization the gene expressions of all the batches have similar distributions, batch effects still exist, and will be demonstrated in the following analysis.

2.2.1 Batch effects removal methods

There are two main approaches for removing the batch effects: location-scale methods and matrix-factorization methods [10]. The location-scale methods assume a model for location

(mean) and/or scale (variance) of the data within the batches and proceeds to adjust the batches in order to agree with these models. The matrix-factorization technique assumes that the variation in the data corresponding to the batch effects is independent on the variation corresponding to the biological variation of interest. The batch effects can be captured in a small set of factors which can be estimated through some matrix factorization methods, such as singular value decomposition or principal component analysis. For example, the 1st principle component has the highest possible variance, which is associated with batch effects. After remove the factors associate with batch effects, the batch effect adjusted dataset is reconstructed back.

In this study, we only applied a few location-scale methods to remove batch effects, including batch mean centering (BMC), gene standardization (GENENORM), empirical Bayes (also known as COMBAT), and distance weighted discrimination (DWD). The main idea of these methods is to transform the data from each batch to have similar (or equal) mean and / or variance for each gene probeset. BMC is a simplest location-scale method, which only subtracting the mean of each gene probeset over all samples from its observed expression inside of each batch. GENENORM is to standardize the gene expression in each batch by transforming the gene expression of each probeset to have mean zero and standard deviation of 1. Empirical Bayes [12] is a method using estimations for the location-scale parameters (mean and variance). The method incorporates systematic batch bias common across genes in making adjustments, assuming that phenomena resulting in batch effects often affect many genes in similar ways, i.e., increased expression, higher variability, etc. The parameters which represent batch effects are estimated by pooling information across genes in each batch to shrink the batch effect parameter estimates toward the overall mean of batch effect estimates (across genes). These estimates are

then used to adjust data for batch effect, providing more robust adjustment for the batch effect on each gene. The method contains three steps: (1) Standardize the data; (2) Batch effect parameter estimates using parametric empirical priors; and (3) Adjust the data for batch effects. For the DWD method, at a starting point, samples from a single batch are regarded as belonging to a specific class and distance weighted discrimination is used as a classification algorithm by finding the optimal hyperplane $w \times x + b = 0$, where w the normal vector of the hyperplane. The hyperplane separates samples from the different classes (batches). Then the samples in each batch are projected in the direction of the normal vector to the hyperplane by calculating the mean distance from all samples in each batch to the hyperplane and followed by subtracting the normal vector to this plane multiplied by the corresponding mean distance.

2.2.2 Analysis and Results

The 133 microarray gene expressions were firstly normalized by RMA method before remove batch effects. Then the microarray samples were split into 6 batches based on the micro array production dates, as shown in Table 1. The Batch name in the Table 1 contains the information of the month and the date of gene chip production, e.g., '1109' means the microarray chips were produced on November 9. The microarray samples of the 6^{th} batch were produced about 6 to 7 month later than the other 5 batches. Although the simple sizes are small in some batches, especially batch 1 and 5, which are only contain 2 or 3 samples, we still performed all the batch effects removal methods mentioned above. We also used most common and straightforward visualization tools to evaluate the effectiveness of batch effect removal methods.

Figure 4 is the boxplot of all gene probeset expressions by using different batch effects remove methods, while Figure 5 shows the histograms and density plots. For comparison

purpose, we also include the distribution of the microarray dataset without using any batch removal method after RMA normalization. It can be seen that gene expression distribution of both COMBAT and DWD methods are similar as the one without batch effect removal, and the dataset generated by DWD method has slightly larger variation than the COMBAT method. The BMC and GENENORM are simple centering or standardized methods, so the means (or medians) of the gene expression values move to close to zero and their distributions are quite narrow and more symmetric compare to others.

Table 1 Batches of microarray samples

Batch ID	1	2	3	4	5	6
Batch name	1109	1110	1116	1119	1130	0603
number of arrays	2	45	43	18	3	22

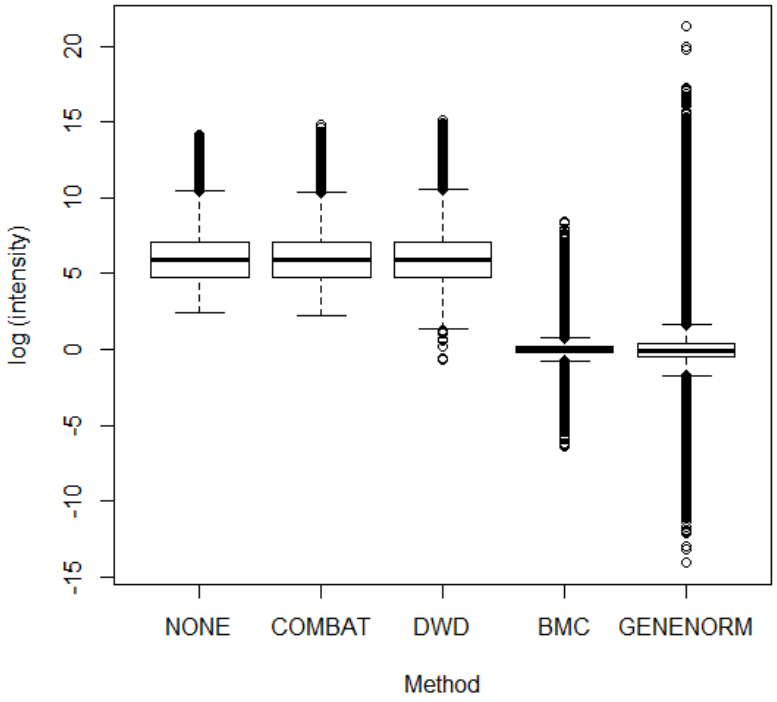


Figure 4 Boxplot of gene expression using different batch effects removal methods

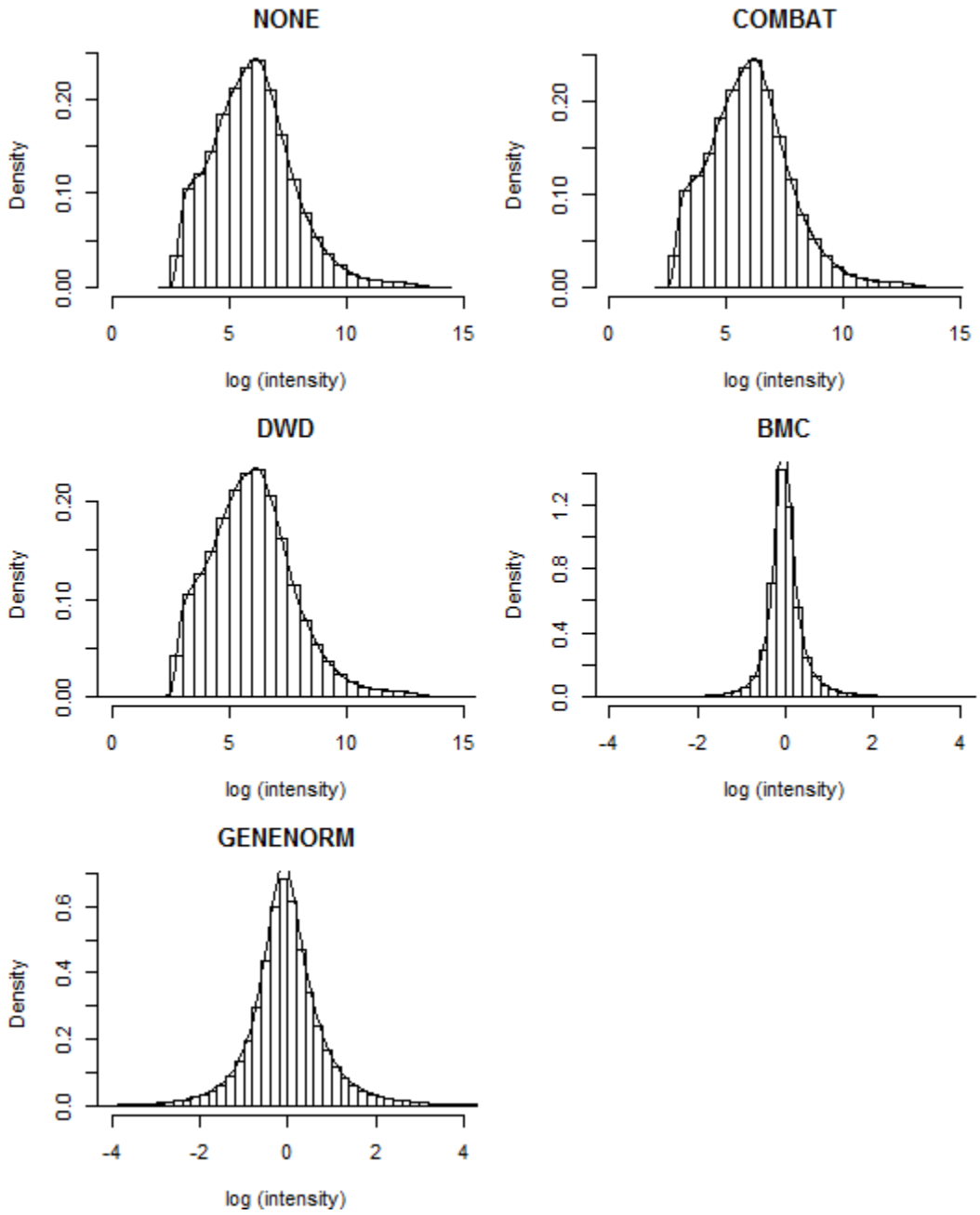


Figure 5 Histograms and density plots of gene expressions using different batch effects removal methods

The visualization tools can provide a crude approximation of the efficiency of batch effects removal. However, if more rigorous evaluation is required, quantitive measures have to be computed to accurately assess the quality of the batch effect removal process. In this work, we only used some visualization tools to evaluate the effectiveness of batch effects removal methods. Some global visualization tools, such as Dendrograms, heatmap, plots of the principal components or relative log expression plot, can provide a 'big picture' of the presence of the batch effect. If the samples group by batch, then it indicates the presents of batch effects [10].

Firstly we used heatmaps, which present hierarchial clustering analysis results, to illustrate the batch effects. Hierarchial clustering is to group similar objects into 'cluster'. In the beginning, each row and column is considered a cluster. In hierarchial clustering, the two most similar clusters are combines and continue to combine until all objects are in the same cluster. Hierarchial clustering produces a tree (dendrogram) that shows the hierarchy of the clusters. This allows for exploratory analysis to see how the microarrays group together based on similar features. Figure 6 shows the heatmap of selected 50 gene probesets with highest variance before and after applying batch effects removal methods. The columns of the heatmap are the 133 microarray samples and the rows of the heatmap are the selected 50 gene probesets. Because we are interested in the batch effects across the microarray samples, only the clustering along the the column direction is considered. Between the dendrogram and the heatmap matrix, there is a narrow colorful strip, in which each color corresponds to a batch. The batch effects could be clearly observed from these strips. Without adjusting batch effects, the samples from the same batch tend to cluster together (e.g., the samples from green batch, and some samples from pink and blue batches). Then we can conclude that there exist batch effects in the dataset. After adjusting batch effects, the samples in the same clusters are from different batch, i.e., the clusters

are formed due to biological features of the sample instead of due to the samples from the same batch. It can be seen that almost no large batch effects exists after applying the batch effect methods.

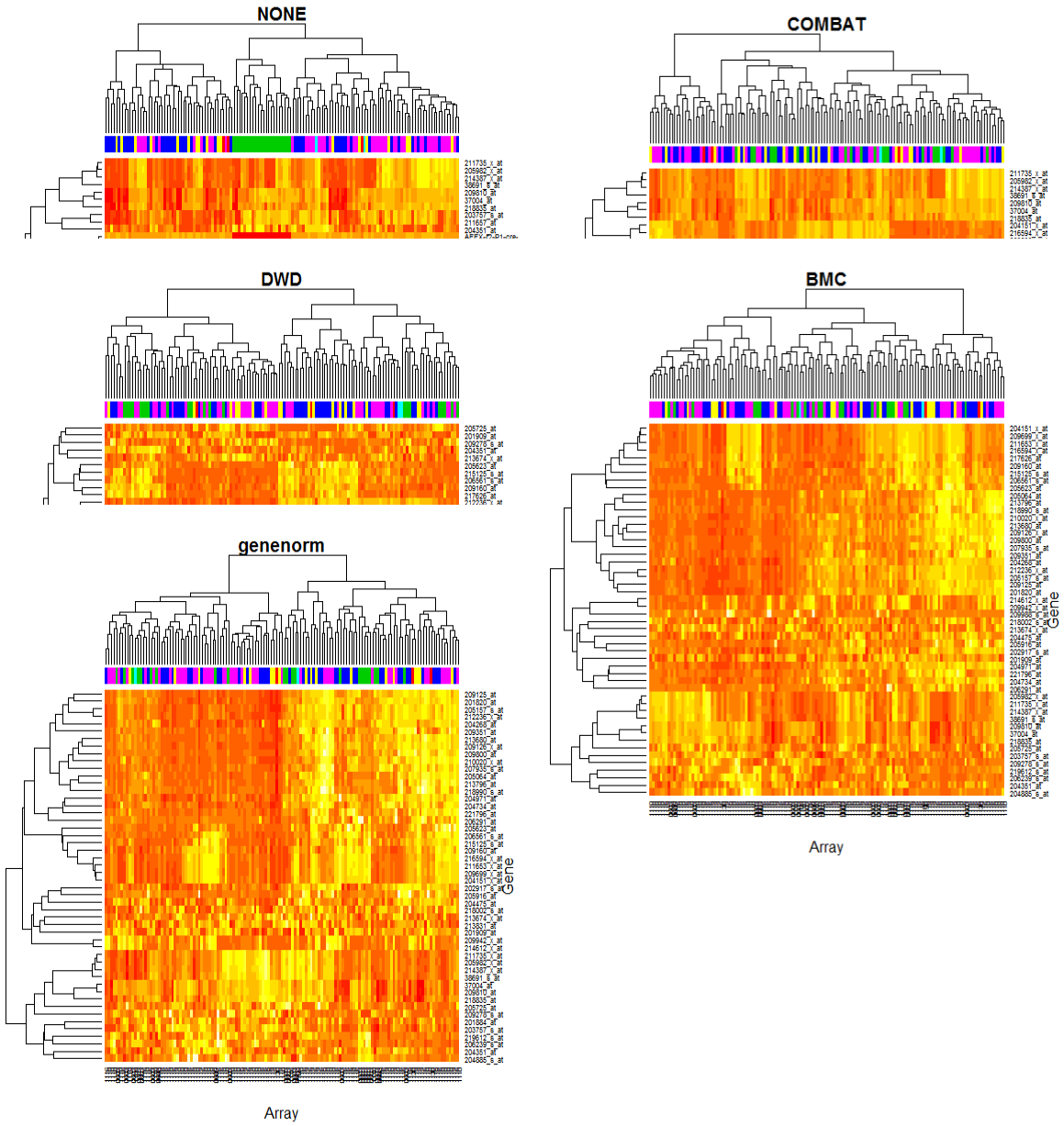


Figure 6 Heatmaps of selected 50 gene probesets with highest variation for batch effects

The principal components plots can also illustrate the batch effects. Principal component analysis [13] makes use of an orthogonal transformation to convert a set of observation of possibly correlated variables into a set of values of linearly uncorrelated variables called 'principal component'. The transformation is defined in such a way that the first principal component has the largest possible variance, and it is orthogonal on the other principal components. The rest of the principal components are ordered according to the amount of variance captured. Based on the observation of a number of various studies, it had been concluded that the greatest source of differentially expression is nearly always across batches rather than across biological groups [14]. Therefore, the plots of the first two principal components are commonly used to visualize the batch effects. According to these plots, batch effects present if the samples from different batches are separated.

Figure 7 illustrates the batch effects by plotting the first two principal components. Before adjusting batch effect, it can be seen that the samples from the same batch have the tendency of clustering. Particularly the samples from batch '6', the samples aggregate at the top-left corner of the plot and separate from other samples. It is consistent with our defined batches based on the microarray chips production time shown in Table 1, i.e., the samples in batch '6' were produced about 6 months later than other samples. Although the samples from other batches do not complete separate with each other, it is clearly seen that the samples from the same batch still group together. After adjusting batch effects, the samples from different batches mixed up and spread on the plots randomly, indicating the batch effects are removed or reduced effectively.

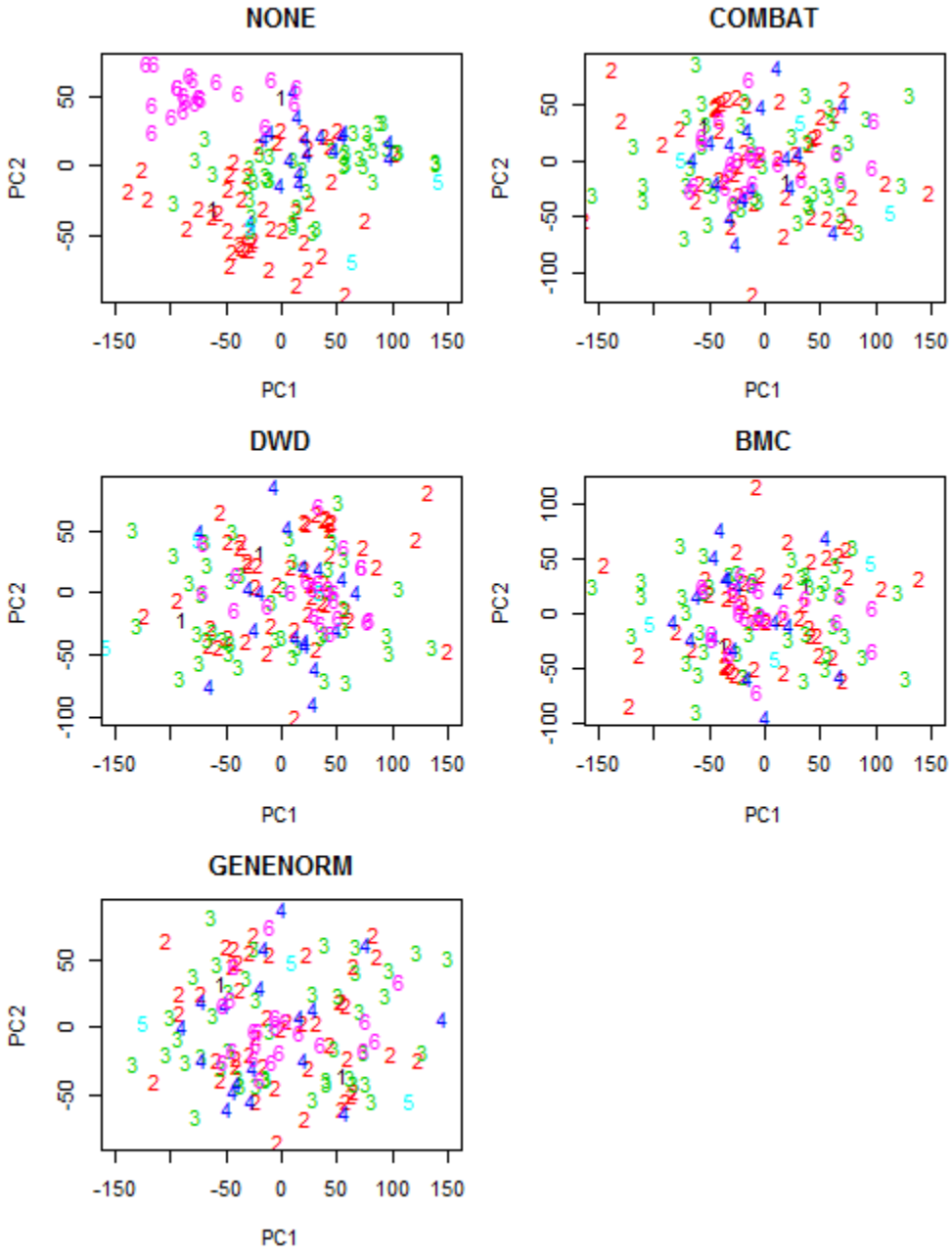


Figure 7 Principal components plots for batch effects. Plots of the first two principal components (The symbol numbers 1 to 6 in the plots corresponding to batch IDs 1 to 6)

2.3 Summary

For the studies involved in multiple microarrays, the raw data are always highly noise and have variations across microarray samples. These variations are not due to the biology of the samples but the technical reasons during array production or combing different studies together. In order to explore the biological difference between arrays, appropriate normalization and batch effect removal methods have to be used to remove those variations. In this work, raw data (*.cel file) of 133 microarray samples from BR. 10 clinical trial were preprocessed by normalization and batch effects removal. Three normalization algorisms, including RMA, GC-RMA and MAS 5 were implemented along with quantile normalization method. The gene expressions showed consistent distribution across array samples after normalization. A few location-scale batch effects removal methods, including COMBAT, DWD, BMC and GENENORM were performed on the data set after RMA normalization. Both the dendrograms from heatmaps, and the principal components plots illustrate that the batch effects exist in the microarray data set, and they can be effectively removed or reduced by applying these batch effects removal methods. For the main analyses hereafter we use the data abstracted from using RMA normalization and DWD batch effect adjustment.

Chapter 3

Predictive Gene Signature Selection

In order to identify the subset of patients that can potentially benefit from a treatment, so that to develop personalized medicine strategies, it is important to study the treatment and covariate interaction in the clinical trials. In this work, we have adjuvant chemotherapy treatment and a large number of microarray gene expressions as covariates for each observation. The purpose is to model their relationship with survival outcomes, and select a group of genes that have significant interactions with the treatment to predict the treatment effects. However, it is very difficult to detect the interactions between the treatment and high dimensional covariates via direct multivariate regression modeling. Firstly, appropriate variable selection methods have to be used to reduce the number of covariates having interaction with the treatment. Secondly, the main effects of each covariate have to be included in the model. The presence of main effect not only increases the difficulty to detect the treatment covariate interactions, but also increases the number of covariates and further compounds the difficulties in covariates dimension reduction since a subset of variables need to be selected for modeling the main effects as well. In this work, we used Tian et al. [15] proposed approach to estimate the covariates and treatment interaction without need for modeling main effects. Elastic net regularization and variable selection method was used to reduce the number of covariates. Multiple gene probesets were selected by fitting Cox's regression models with modified covariates without main effects, and

performing variable selection by elastic net regularization. Bootstrap method was also applied to achieve a more robust predictive gene signature selection. Gene signature was chosen by combining most often selected probesets. Using the first principal component (PC1) to synthesize the information across the selected gene probesets, patients predictive scores via cross validation were generated by the sum of the treatment estimate and the multiplication of the PC1 and the estimated coefficient of the interaction effect between treatment and PC1 from Cox's model with treatment, PC1 and their interaction terms. Then the patients were classified into low, middle and high predictive score groups. The gene signature is selected as the group of probesets that best separate the low predictive score group patients by treatment arm.

3.1 Patients and samples

BR.10 is a randomized controlled trial of adjuvant vinorelbine / cisplatin versus observation alone, and the snap frozen tumor samples were prospectively collected. There are total 482 randomly assigned patients in the trial, of which frozen tissue was collected from 169 patients for future laboratory study. Gene expression profiling was completed in 133 of these samples, using the U133 oligonucleotide microarrays (Affymetrix, Santa Clara, CA) [4]. Of these 133 patients with microarray profiles, 62 were in observation group, while 71 received adjuvant chemotherapy. A previous study [4] had compared demographic features (including treatment received, age, sex, performance status, stage of disease, surgery, pathologic type, and RAS mutation status) of the 133 patients with microarray profiling to 349 without profiling. All factors are similarly distributed, except for stage, where more stage IB patients are present in the microarray-profiled cohort (55% vs. 42%; p=0.01). There is no significant difference in the

overall survival of patients with and without microarray profiling, and a similar beneficial effect of adjuvant chemotherapy was observed.

Microarray data are available at National Center for Biotechnology Information Gene Expression Omnibus (GSE14824) [9]. Raw microarray data (*.cel files) were preprocessed using RMA method for normalization and distance-weighted discrimination method to adjust batch effects by means of R language based Bioconductor microarray analysis packages, as descripted in the Chapter 2.

3.2 Statistical methods

3.2.1 Treatment and covariates interactions

Tian et al. [15] proposed a simple method and general approach to model predictive interaction terms between treatment and covariate by using modified variables without the main effects. The modified variables are derived as follows. The treatment variable is coded as ±1, and the covariates are standardized to have mean 0 and variance of 1. The modified covariates, i.e., the products of the treatment variable with the standardized covariates, are used in the regression model without main effects for multi-variable model selection. The modified covariates framework can be generalized to different types of outcome, i.e., continuous, binary and survival outcome variables. The following shows the details of the method.

Let T be the binary treatment indicator, where the patients received treatment $T = 1$ and control $T = -1$, and Y be the potential outcome. Only Y, T and Z (a p dimensional baseline covariate vector) are observed. It is assume that there are N independent and identical distributed copies of (Y, T, Z), i.e., $\{(Y_i, T_i, Z_i), i = 1, \ldots, N\}$. Furthermore, Let $W(Z)$ be a p dimensional

functions of baseline covariates Z. Here the dimension of W could be larger than the sample size N.

A simple multivariate linear regression model for charactering the interaction between treatment and covariate is

$$Y = \alpha_0 + \gamma_0' W(Z).\frac{T}{2} + \epsilon$$

where ε is the mean zero random error. Based on this model, the modified covariate estimator $\hat{\gamma}$ is proposed as the minimizer of

$$\frac{1}{N}\sum_{i=1}^{N}(Y_i - \gamma' \frac{W(Z_i)\cdot T_i}{2})^2$$

Then, each component of $W(Z_i)$ can be simply multiplied by one-half the treatment assignment indicator (= ±1) and perform a regular linear regression.

The modified covariate approach can be easily generalized to other more complicate model. In general, the following modified covariate approach is proposed

1. Modified covariate: $Z_i \to W_i = W(Z_i) \to W_i^* = W_i \cdot \frac{T_i}{2}$
2. Perform appropriate regression: $Y \sim \gamma_0' W^*$

   Based on the modified observations $(W^*,\ Y_i) = \left\{\frac{W_i \cdot T_i}{2}, Y_i\right\}, i = 1,2, \ldots N.$
3. $\hat{\gamma}' W(z)$ can be used to stratify patients for individualized treatment selection

When the outcome variable is survival time, we can fit a Cox regression model using the above modified covariate, i.e.,

$$h(t|Z,T) = h_0(t)e^{\gamma_0' W^*}$$

where $h(t|Z,T)$ is the hazard function and $h_0(t)$ the baseline hazard function free of Z and T.

3.2.2 Regularization for high dimensional data

For a multiple linear regression model, ordinary least square estimates can be obtained by minimizing the sum of squares of residuals. The least square estimate has low (almost zero) bias but high variance, so the mean square error is high, which leads to poor prediction accuracy, especially when p (the number of predictor) is large. Moreover, least square method also does poorly in interpretation because of the variable selection issue. Scientists prefer a simple model to interpret the relationship between response and covariates. Although some model selection methods are available, such as AIC, BIC, etc., those methods are more suitable for relatively small p, but the models are not stable for high dimensional dataset. For the case of $p >> n$ (n is the number of observations), the least square method could not be performed. Therefore, appropriate variable selection procedure has to be applied to deal statistical analysis for high dimensional dataset such as microarray, which is a typical large p, small n problem.

Penalized regularization techniques had been proposed to improve ordinary lease square methods. For example, ridge regression [16] minimizes the sum of square of residual subject to a bound on an L_2-norm of the coefficients. Ridge regression achieves better prediction performance through a bias-variance trade-off as a continuous shrinking method. However, ridge regression cannot perform variable selection. It only shrinks the coefficient of estimates towards zero but never reach to zero, so it always keeps all the predictors in the model. Tibshirani [17] proposed a promising technique, called the lasso, for 'least absolute shrinkage and selection operator'. It imposes an L_1 penalty to the least square method, to allow both continuous shrinkage and automatic variable selection simultaneously, because some coefficients of the estimates are shrunk all the way to zero. Although the lasso has shown success in many situations, it still has some limitations. (a) In the $p > n$ cases, the lasso select at most n variables.

In microarray analysis, the number of selected gene probesets is bounded by the number of sample. (b) The lasso fails to do group selection. If there is a group of variables among which the pairwise correlations are very high, then the lasso only select one variable. However, in microarray analysis, some genes share the same biological 'pathway' and the correlation between them are very high. We want to select the whole group of genes.

Zou and Trevor [18] proposed an elastic net regularization method using both L_1-norm and L_2-norm penalizations to improve the lasso regularization. The model can be written as

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

where $\lambda_2\|\beta\|^2$ is the L_2-norm penalty term and $\lambda_1\|\beta\|_1$ the L_1-norm penalty term. Similar as the lasso model, the L_1 part of the penalty generates a sparse model, while the quadratic part of the penalty removes the limitation on the number of selected variables, encourages grouping selection, as well as stabilizes the L_1 regularization path. The model can also be written as

$$\min_{\beta} \|\mathbf{y}-\mathbf{X}\beta\|^2 \text{ s.t. } J(\beta) \leq t$$

where the penalty term of elastic net

$$J(\beta) = (1-\alpha)\|\beta\|^2 + \alpha\|\beta\|_1 \qquad (\text{with } \alpha = \frac{\lambda_1}{\lambda_1+\lambda_2}.)$$

Here λ_1 and λ_2 are tuning parameters. Similar as the lasso or ridge model, the tuning parameter(s) controls the strength of the penalty, i.e., the strength of the shrinkage. The larger values of the tuning parameter, the stronger shrinkage will be, and the less number of variables are selected (for λ_1). When $\lambda_1 = 0$ the model becomes ridge regression, while when $\lambda_2 = 0$ the model becomes the lasso model. When both λ_1 and λ_2 equal to zero, the model converts to the simple linear regression model. In the elastic model, λ_2 and the fraction of L_1-norm or L_2-norm (i.e., α or $(1-\alpha)$) can also be the tuning parameters. The advantage of using (λ_2, α) is that α is

always valued within [0,1]. The tuning parameters can be determined by cross validation. Generally there are two rules to select the tuning parameters [17], minimum cross validation rule and one standard error rule. For the minimum cross validation rule, the tuning parameter can be selected at the values that the cross validation error is minimum. Alternately, in order to get simpler model, the tuning parameters can also be the point that the value of minimum cross validation error plus one standard error. An illustration of tuning parameter selection by cross validation is shown in Figure 8. There are two tuning parameters in the elastic net, so we need to cross-validate on a two dimensional surface.

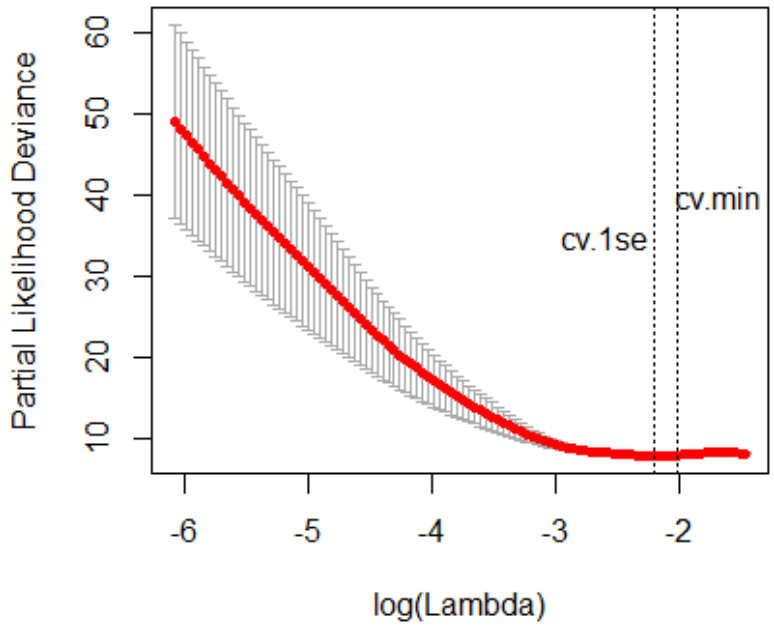


Figure 8 Illustration of tuning parameter selection by cross validation. cv.min represents minimum cross validation rule, while cv.1se represents one standard error rule. (The plot is tuning λ_2 in the elastic net model at $\alpha = 0.1$.)

3.3 Analysis procedures

3.3.1 Preselection of gene probesets

The microarray dataset and BR.10 clinical data were linked by microarray sample names and the patients clinical IDs. The dataset contains 22215 gene probesets and 133 observations. The overall survival was used as survival outcome in this work. We firstly filtered out about 1/3 of the gene probesets with the variance of the expression values less than 0.1 and mean less than 4, which contain relative less information on data variation and those with intensity similar to background noises. The left gene probesets were standardized with mean zero and variance 1 across all 133 microarray samples. Then we performed univariate analysis to fit Cox's regression models for each gene probeset using a modified covariate, as described in section 3.2.1. The modified covariates only contain the interaction terms of gene expression and treatment without main effects, i.e., each Cox's regression model, corresponding to each gene probeset, is fitted by overall survival vs. the modified covariate, the product of one half of the treatment and the standardized gene expression values. Here the treatments are coded as -1 or +1, where the patients who received chemotherapy are coded as +1 and the patients in observation groups are coded as -1. Then the gene probesets were ranked by Cox's regression model p-values, and the gene probesets that have predictive potential i.e., probesets that have significant interaction with treatment were preselected based on Cox's regression model with p-value < 0.05.

3.3.2 Selection of predictive gene signature

The patients were randomly divided into 2 parts with similar survival experience by the treatment arms. In order to obtain similar survival outcomes and similar distributions of some major clinical features in the two subsets, we firstly separated patients into small subsets by

stratifying the treatment arm, disease stage, histology, and survival outcome. Then we performed random sampling without replacement in each small subset. Finally, the selected patients in each subset were combined together to form the training dataset, and the rest patients are the testing datasets.

Only the data in the training set were used to select predictive gene signature. The training microarray dataset containing the gene probesets preselected in the section 3.3.1 were standardized to mean zero and variance 1, denoted as *z*. Similar as the univariate analysis, modified covariates were created here, which are the linear combination of the standardized gene expressions *z* and the coded treatment values (+1 or -1) over two, and were fitted to the Cox' regression model. Because of large number of covariates, the elastic net regularization method was used for variable selection. A sparse estimate coefficients matrix was given by the elastic net, and the probesets with non-zero coefficients are selected as predictive gene probesets. As we mentioned before, elastic net method has two tuning parameters. Here we used λ_2 and α as the tuning parameters. At a given α, the λ_2 was tuned by leave one out cross validation using the minimum cross validation rule. We initially also tuned the α in that range of (0, 1) using the cross validation, and we got the α values very close to 0, which are 0.01 and 0.03 based on minimum cross validation rule and one standard error rule, respectively. These small α values imply that our model is quite close to ridge regression model, such that a large number of gene probesets will be selected as predictive gene signature, then our model may be over fitted. Therefore, we tuned the α with 0.025 increment in the range of (0, 1) based on the selected gene probesets which can best identify the subset of the patients who are beneficial to the chemotherapy treatment in the testing dataset in the following procedure instead of using cross validation error rule.

Because the variation of the dataset, the selected gene probesets only based on one time model fitting is not robust. Here we use bootstrap method to resample the patients in the training dataset, i.e., resample the same number of patients with replacement, and then performed elastic net gene signature selection based on the new datasets created by the bootstrap. In order to maintain similar disease features between the bootstrap resampled datasets and the original training dataset, the disease stage was stratified before resampling. The patients with stage IB and stage II were firstly resampled separately, then were combined together to form the new bootstrap datasets. We implemented 1000 times bootstrap resampling and performed Cox's regression and elastic net variable selection for each bootstrap dataset, and output the frequency for each probeset that had been selected in the model. Then we ranked the gene probesets based on the frequencies of the probesets appeared, and the first 150 highest frequency gene probesets will be used for further predictive gene signature selection.

Based on the original microarray data in the training dataset, principal component analysis of the gene probesets expressions was conducted for the first i (i = 1 to 150) ranked probesets in the previous step, respectively, and the 1^{st} principal components were used for the further analysis to derive the predictive score for each patient. To get more robust results, we use 10 fold cross validation. Firstly we randomly divided the training set into 10 subsets stratified by disease stage and survival event status. Then each Cox's regression model was fitted with study treatment, the 1^{st} principal component (PC1), and their interaction term based on the 9 folds data. Based on the estimate of treatment effect and the interaction effect, the predictive scores of data left out were developed using the formula of $\beta_1 + \beta_3$*PC1, where the β_1 and β_3 are the estimate of treatment effect and the estimate of interaction effect.

The patients were divided into 3 groups using 1/3 and 2/3 quantiles of predictive scores as cut-off points. The log-rank test p-value for comparison between two treatment arms in the low score group will be used to choose the group of probesets for the final model. At this point, we have the p-values from the 150 models with the number of gene probesets 1 to 150 as the covariates. Then we chose the number of probesets with the minimum p-value as the group selected predict gene probesets (predictive gene signature) for our final model. For example, if the minimum log-rank test p-value for comparing the treatment group in the low predictive score patients group was achieved when the first *n* of the 150 gene probesets were in the model, then these first *n* gene probesets are our selected gene probesets.

3.3.3 Prediction of treatment effect

The selected predictive gene probesets were used to predict treatment effect and identify the subset of the patients who are beneficial to the chemotherapy. Considering the correlations among the selected gene probesets, principal component analysis was also implemented for the training dataset, and the PC1 was fitted to a Cox's regression model along with the treatment and their interaction term. Here the PC1 of the training dataset represents about 20% data variation. Then the predictive score of each patient in the training set is defined as the product of estimate of the interaction effect based on the Cox's regression model and the PC1, i.e., β_3*PC1, where β_3 is the estimate of interaction effect of treatment by PC1. Using 1/3 and 2/3 quantiles of the predictive scores as cut-off points, the patients were classified into low predictive score, middle predictive score and high predictive score groups. Then the log-rank test p-value and hazard ratio based on Cox's model between the treatment arms were computed in the each predictive score group of the training dataset.

Furthermore, predictive score of each patient in the testing set was also constructed using the similar formula in the training set, i.e., β_3*PC1_testing. The PC1_testing is attained by using the gene expression matrix times the 1st column of the loading matrix of the principal component analysis from the training dataset. The loading matrix is the transformation matrix from the original dataset to the principal components dataset. Then we used the cut-off points developed in the training dataset to divide the patients in the testing dataset into low, middle and high predictive score groups, and applied log-rank test and Cox's regression to check the treatment effects between the two treatment arms, and identify treatment benefit subset of the patients.

Besides of three gene groups, two predictive score groups, i.e., low and high predictive score groups were also constructed using median predictive score of the training dataset based on the selected gene signature, and similar statistical analysis was carried out to predict the treatment effect.

3.3.4 Internal validation

We performed internal validation to test the selected predictive gene signatures and the proposed method. All the 133 patients were treated as training dataset and the cut-off points (for both 3 gene groups and 2 gene groups) were computed. The validation datasets were generated by the bootstrap resampling, i.e., randomly resampling patients with replacement from the all data by stratifying the treatment arm, such that the resampled datasets contain similar number of patients in each treatment arm. The cut-off points developed based on the all data was used to divide the patients in the validation datasets into 3 or 2 gene groups and the treatment effects were checked in each groups based on the log-rank test p-value and the estimate of hazard ratio

of the treatment effect. 200 times bootstrap resampling were performed for the internal validations.

3.4 Results and discussion

3.4.1 Predictive gene signature selection

After filtering the gene probesets with small variance across the microarrray samples and univariate analysis, 664 gene probesets with significant interaction with the treatment were preselected for further analysis.

Of all 133 patients, 66 patients were randomly selected as training dataset while the rest 67 patients are in the testing dataset. The baseline factors in the all dataset, training dataset and testing dataset are shown in Table 2. It can be seen that the patients are almost evenly distributed in the training and the testing datasets. In the training dataset, 31 patients are in the observation group and 35 patients received adjuvant chemotherapy, while in the testing dataset, the numbers of patients in the two corresponding arms are 31 and 36, respectively.

Kaplan-Meier survival estimates, log-rank test and Cox's regression with treatment as covariate were used to estimate and test difference in survival distributions between treatment for all dataset, training dataset and testing dataset, respectively, shown in Figure 9. For the original dataset containing all 133 patients, beneficial effect of adjuvant chemotherapy can be observed based on the Kaplan-Meier estimates and Cox' regression model (HR (hazard ratio), 0.80; 95% CI, 0.48 to 1.32), but log-rank test shows the difference is not statistically significant (p-value=0.38). The training dataset and the testing dataset show similar survival experience, with HR of 0.79 (95% CI: 0.38 to 1.59), and log-rank test p-value of 0.49 for the training dataset,

while HR of 0.82 (95% CI: 0.40 to 1.69), and log-rank test p-value of 0.59 for the testing dataset. Moreover, the treatment effects are similar between the training and testing sets (p-value=0.87).

Table 2 Baseline demographics of patients in training, testing and all datasets

		Training	Testing	Total
Age, years	Minimum	38	35	35
	Median	62	60	62
	Maximum	81	77	81
Sex	Female	17 (26%)	25 (37%)	42 (32%)
	Male	49 (74%)	42 (63%)	91 (68%)
Treatment	Observation	31 (47%)	31 (46%)	62 (47%)
	Chemo	35 (53%)	36 (54%)	71 (53%)
Histology	Adenocarcinoma	36 (55%)	35 (52%)	71 (53%)
	Squamous cell carcinoma	25 (37%)	27 (40%)	52 (39%)
	Undeferentiated	5 (8%)	5 (8%)	10 (8%)
Stage	IB	36 (55%)	37 (55%)	73 (55%)
	II	30 (45%)	30 (45%)	60 (45%)
Survival event	Censored	36 (55%)	37 (55%)	73 (55%)
	Died	30 (45%)	30 (45%)	60 (45%)
Total		66	67	133

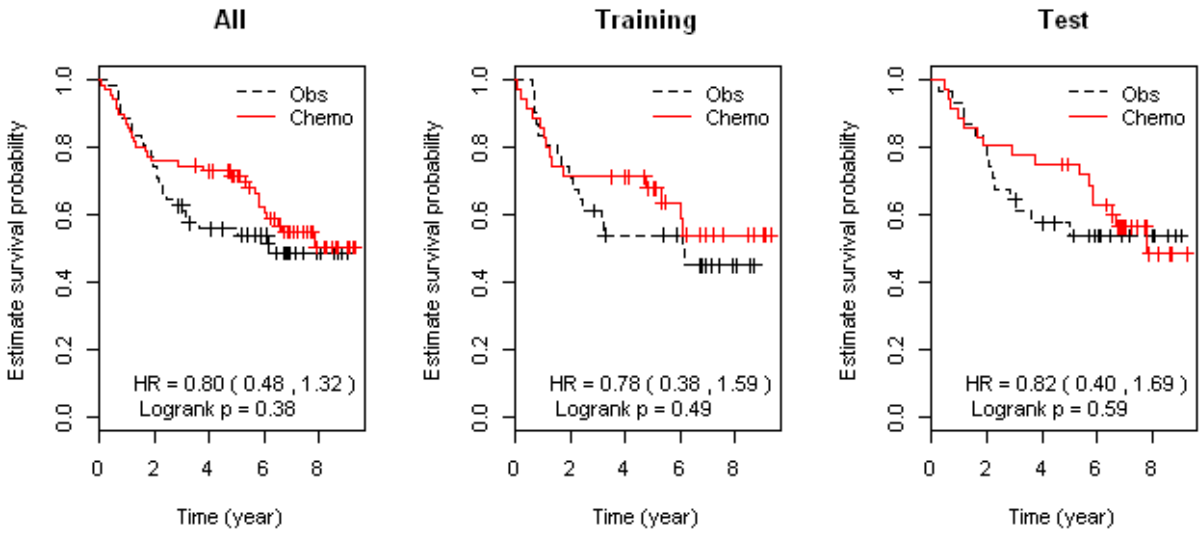


Figure 9 Survival plots by treatment arm for all, training and testing datasets

After applying Bootstrap – Cox' regression, elastic net variable selection and principal component analysis, 34 gene probesets were selected as a group of predictive gene signatures, as shown in Table 3. The corresponding gene symbols were annotated based on Affymetrix HG-U133A annotation file [19] from the website http://www.affymetrix.com/support/technical/annotationfilesmain.affx.

Table 3 Genes and probesets that constitute the 34-gene signature

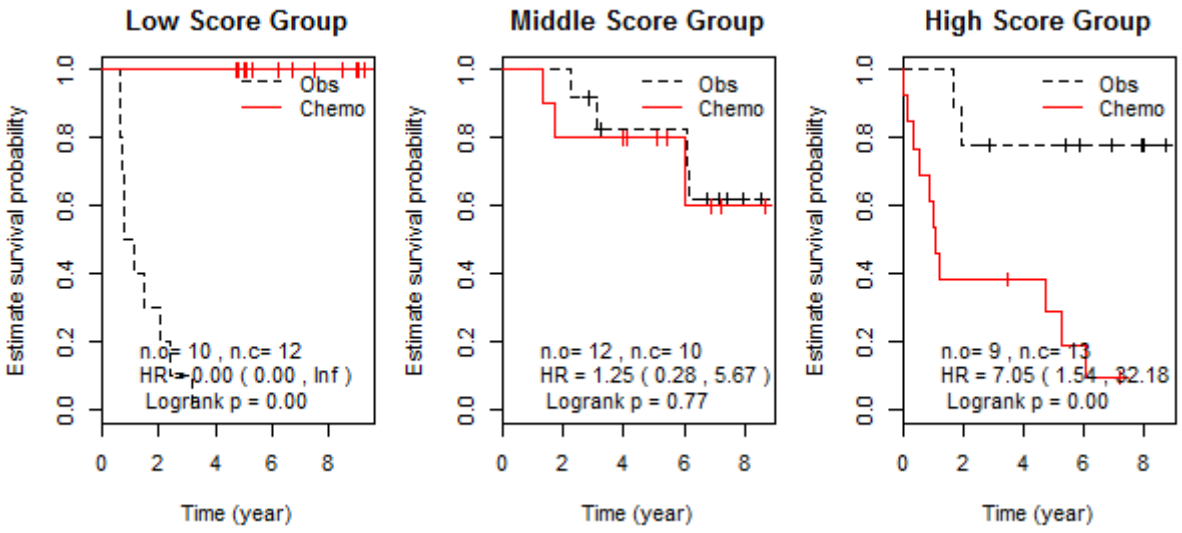
Probeset	Gene Symbol	Gene Title
205241_at	SCO2	SCO cytochrome oxidase deficient homolog 2 (yeast)
219202_at	RHBDF2	rhomboid 5 homolog 2 (Drosophila)
203147_s_at	TRIM14	tripartite motif-containing 14
201016_at	EIF1AX	eukaryotic translation initiation factor 1A, X-linked
221609_s_at	WNT6	wingless-type MMTV integration site family, member 6
220442_at	GALNT4	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 4 (GalNAc-T4)
218640_s_at	PLEKHF2	pleckstrin homology domain containing, family F (with FYVE domain) member 2
203764_at	DLG7	discs, large homolog 7 (Drosophila)
204068_at	STK3	serine/threonine kinase 3 (STE20 homolog, yeast)
207663_x_at	GAGE3	G antigen 3
213032_at	NFIB	nuclear factor I/B
217730_at	TMBIM1	transmembrane BAX inhibitor motif containing 1
219429_at	FA2H	fatty acid 2-hydroxylase
219140_s_at	RBP4	retinol binding protein 4, plasma
205874_at	ITPKA	inositol 1,4,5-trisphosphate 3-kinase A
217995_at	SQRDL	sulfide quinone reductase-like (yeast)
219148_at	PBK	PDZ binding kinase
203929_s_at	MAPT	microtubule-associated protein tau
214106_s_at	GMDS	GDP-mannose 4,6-dehydratase
205076_s_at	MTMR11	myotubularin related protein 11
205552_s_at	OAS1	2',5'-oligoadenylate synthetase 1, 40/46kDa
202036_s_at	SFRP1	secreted frizzled-related protein 1
204437_s_at	FOLR1	folate receptor 1 (adult)
204867_at	GCHFR	GTP cyclohydrolase I feedback regulator
202869_at	OAS1	2',5'-oligoadenylate synthetase 1, 40/46kDa
214056_at	MCL1	Myeloid cell leukemia sequence 1 (BCL2-related)

206134_at	ADAMDEC1	ADAM-like, decysin 1
219670_at	C1orf165	chromosome 1 open reading frame 165
204580_at	MMP12	matrix metallopeptidase 12 (macrophage elastase)
207713_s_at	RBCK1	RanBP-type and C3HC4-type zinc finger containing 1
201017_at	EIF1AX	eukaryotic translation initiation factor 1A, X-linked
200661_at	PPGB	protective protein for beta-galactosidase (galactosialidosis)
203449_s_at	TERF1	telomeric repeat binding factor (NIMA-interacting) 1
217794_at	PRR13	proline rich 13

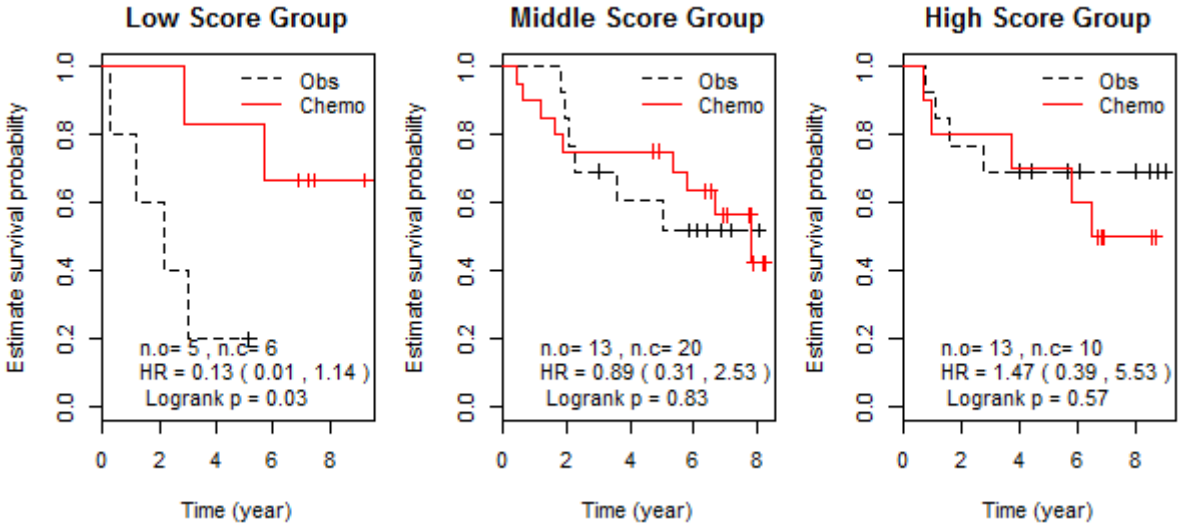
3.4.2 Treatment effect prediction

The Kaplan-Meier estimates of overall survival for the patients in the low, middle and high gene groups of the training and testing data sets based on 34-gene signature are plotted in the Figure 10. In the training dataset, each predictive score group has 22 patients. Significant treatment effects were found in the low and high predictive score groups. In the low predictive score groups, patients are beneficial to the adjuvant chemotherapy, i.e., the patients received the chemotherapy have significant longer survival time (HR < 0.01, and log-rank test p-value < 0.01). However, in the high predictive score group, detrimental effect of the chemotherapy was observed (HR, 7.05; 95% CI, 1.54 to 32.2; log-rank p-value < 0.01). In the middle score group, the survival outcomes were similar between the two treatment arms (HR 1.25; 95% CI, 0.28 to 5.67; log-rank test p-value of 0.77). In the testing dataset, the 34-gene signature classifies the 11 patients (5 of them in observation arm and 6 in treatment arm) into low predictive score group, 33 patients in the middle predictive score group, and 23 patients in the high score group. Similar as the training set, patients who are classified to the low predictive score group are beneficial to the chemotherapy. The patients received chemotherapy has significant longer overall survival than those in the observation arm (HR 0.13; 95% CI, 0.01 to 1.14; p = 0.03 by log-rank test). In the high risk group, no benefit from treatment on the survival outcome was observed, with an estimated HR of 1.47 (95% CI 0.39 to 5.53) and log-rank test p-value of 0.57. In the middle

score group, the survival outcomes of the two treatment arms have no significant difference (HR, 0.89; 95% CI, 0.31 to 2.53; log-rank p-value=0.83).



(a) Training dataset



(b) Testing dataset

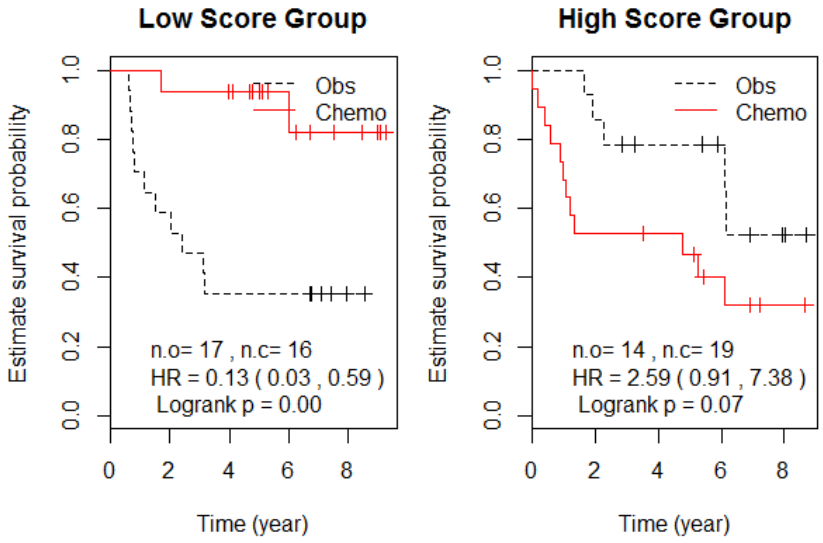
Figure 10 Overall survival in three predictive score groups based on the 34-gene signature. (a) Training dataset; (b) Testing dataset. n.o and n.c in the plots represent the number of patients in the observation arm and chemotherapy arm, respectively.

When the patients were classified into two groups based on the 34-gene signature using the median predictive score of the training set as the cut-off point, the treatment beneficial group still can be identified, shown in Figure 11. In the low score group 26 patients have significant longer overall survival in the chemo treatment arm than those in the observation arm (HR, 0.20; 95% CI, 0.06 to 0.70; log-rank test p-value = 0.01). In the high score group, the patients with adjuvant chemotherapy also show slightly worse survival outcomes but the difference is not statistically different (HR, 1.77; 95% CI, 0.65 to 4.79; log-rank test p-value=0.26). These results indicate that the selected 34 gene signature can identify the patients who are beneficial to the adjuvant chemotherapy.

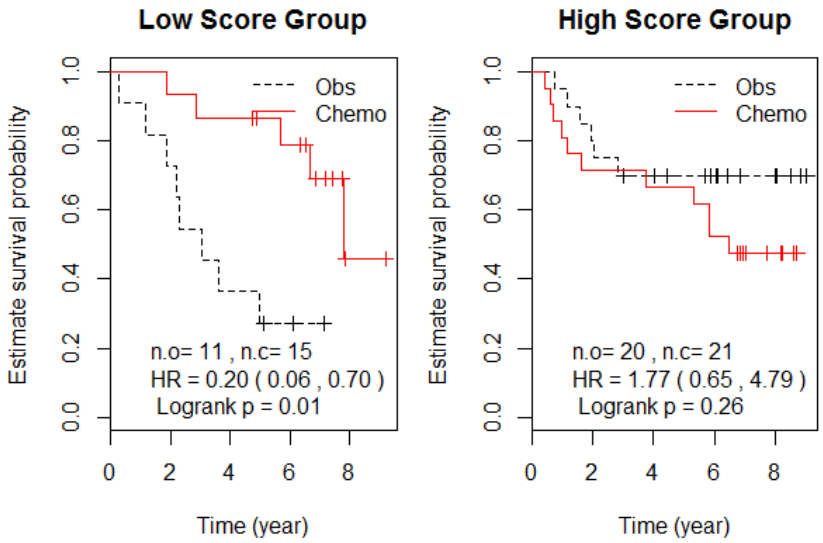
If only consider the patients in the observation arm, from Figures 10 and 11 it can be seen that the patients in the low predictive score group have worse survival than those in the high score group. Therefore, the patients in the prognostic high risk group will be potentially beneficial to the chemotherapy, but the patients in low risk group are lack of benefit of the chemo treatment. This is consistent with the results from other studies about adjuvant chemotherapy treatment effect on early stage NSCLC patients [3, 4].

The predictive ability of the 34-gene signature for the chemotherapy benefits is further validated by bootstrap resampling the 67 samples of the testing dataset. 200 times validations were conducted. When the patients were classified into low, middle, and high predictive score groups, about ¾ of the validations have the log-rank test p-value < 0.05, and the HR < 0.4 for 195 (97.5%) validations. Averagely 11 patients were classified into the chemotherapy benefit group. When the patients were classified into two groups, 65% validations have the log-rank test p-value < 0.05, and about 80% of validations have the hazard ratio < 0.4, while 28 patients were

classified into low predictive score group averagely. Therefore, the 34-gene signature works effectively to identify the patients who are benefit from the chemotherapy.



(a) Training dataset



(a) Testing dataset

Figure 11 Overall survival in two predictive score groups based on the 34-gene signature. (a) Training dataset; (b) Testing dataset. n.o and n.c in the plots represent the number of patients in the observation arm and chemotherapy arm, respectively.

Figure 12 is a flow chart of the predictive gene selection and the treatment effect prediction procedures.

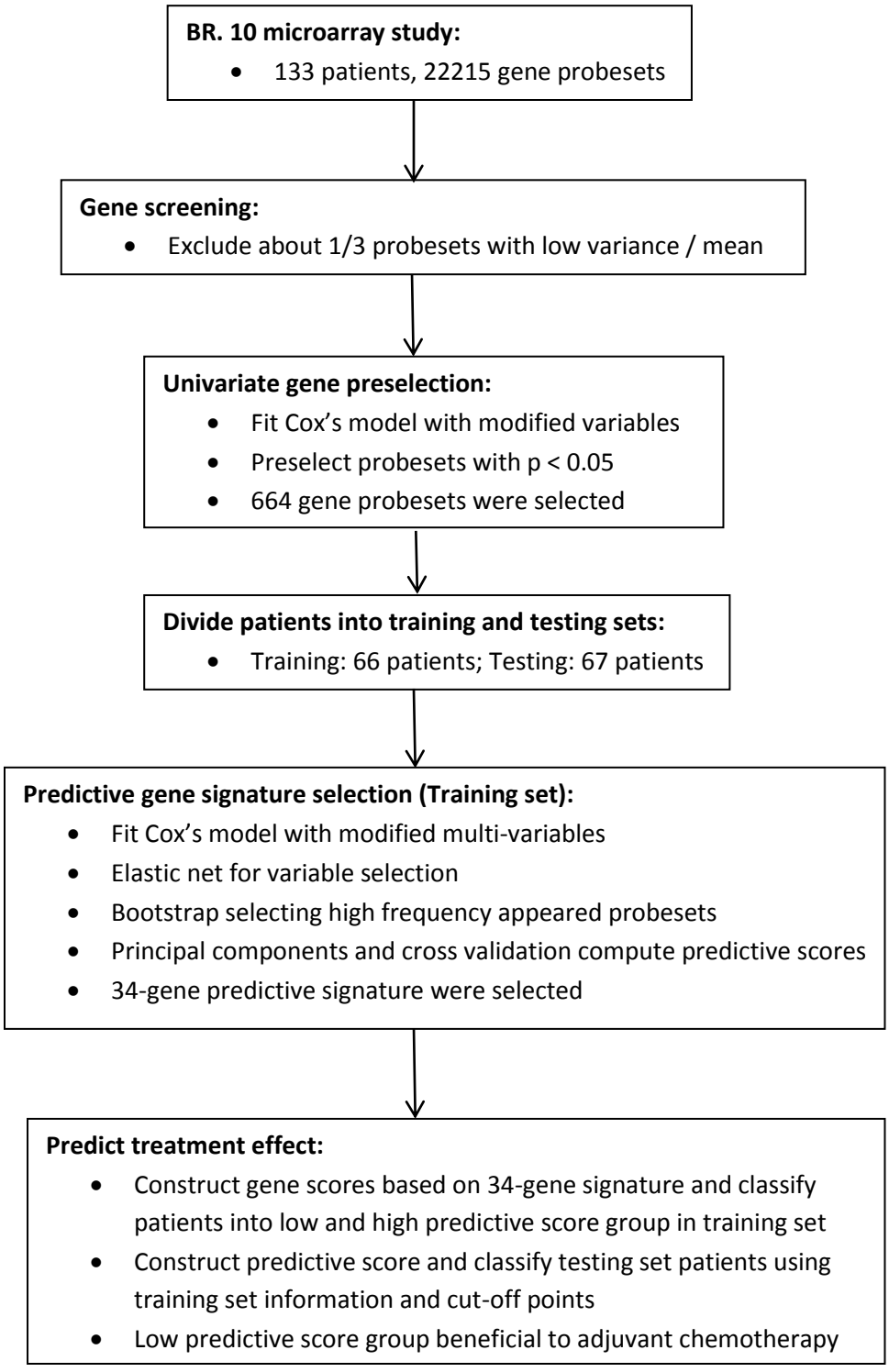
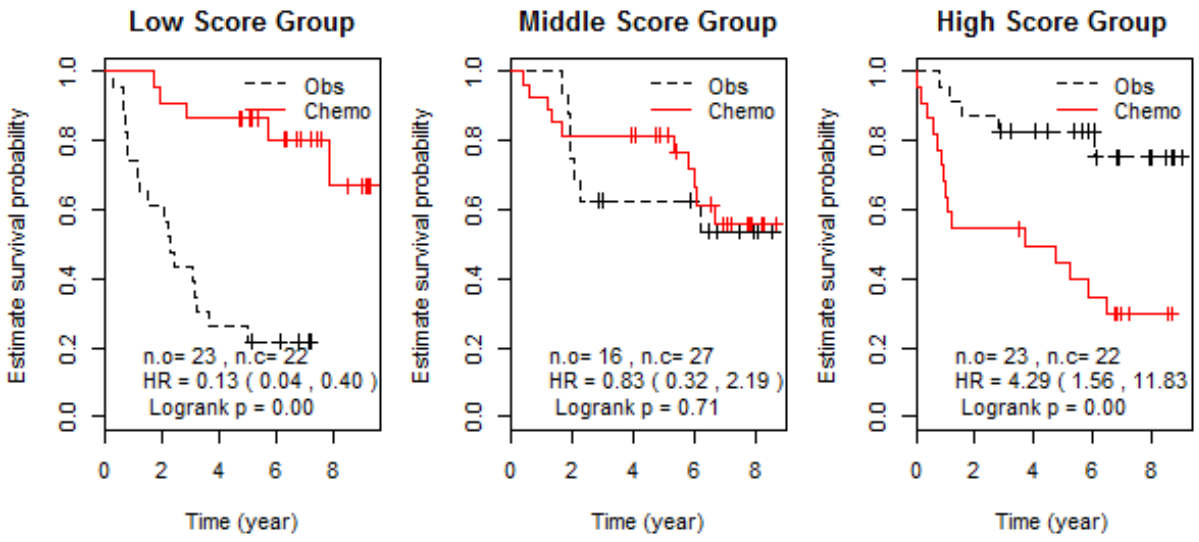
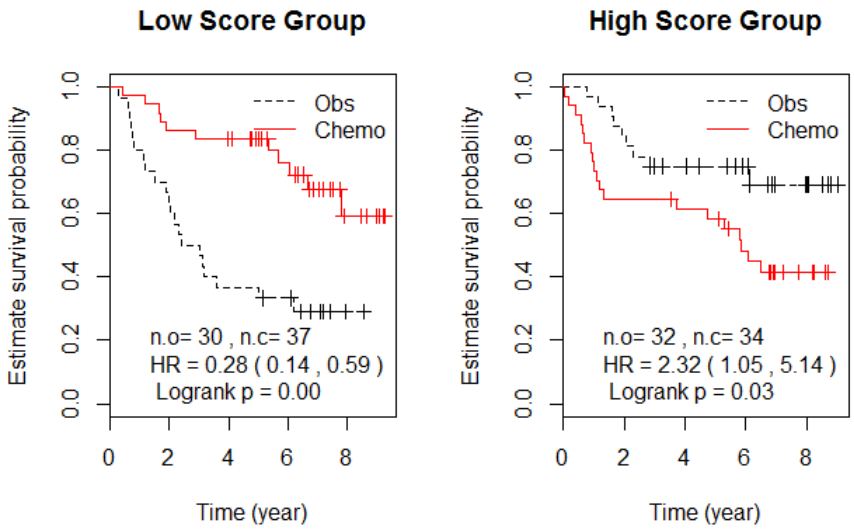


Figure 12 Flow chart of predictive gene signature selection and treatment effect prediction

Based on the 34-gene signature, all the 133 samples of BR.10 are classified into low, middle, and high predictive score groups by using 1/3 and 2/3 of the predictive scores, or low and high predictive score groups by the median of the scores. The Kaplan-Meier survival plots are shown in Figure 13.



(a)



(b)

Figure 13 Overall survival of 133 patients in predictive score groups based on 34-gene signature. Classified patients into (a) 3 predictive score groups; (b) 2 predictive score groups.

It can be seen from Figure 13 that when the patients are classified into 3 predictive score groups, 45 patients are in the low risk groups (23 observation, and 22 chemo) with HR of 0.13 (95% CI 0.04 to 0.40), and log-rank test p-value < 0.01. When classify patients into two score groups, 67 patients (30 observation, and 37 chemo) are in the low score group and with HR of 0.28 (95% CI 0.14 to 0.59) and log-rank test p-value < 0.01. This also demonstrated that the patients in the low score groups are benefit from the chemotherapy. The baseline factors of the patients in the low predictive score groups are shown in the Table 4.

Table 4 Baseline demographics of patients classified into low predictive score groups

		3 score group	2 score group
Age, years	Minimum	43	43
	Median	62	62
	Maximum	75	76
Sex	Female	10 (22%)	17 (25%)
	Male	35 (78%)	50 (75%)
Treatment	Observation	23 (51%)	30 (45%)
	Chemo	22 (49%)	37 (55%)
Histology	Adenocarcinoma	17 (38%)	24 (36%)
	Squamous cell carcinoma	24 (53%)	36 (54%)
	Undeferentiated	4 (9%)	7 (10%)
Stage	IB	21 (47%)	34 (51%)
	II	24 (53%)	33 (49%)

The information used in the algorithm for deriving the 1st principle component, the predictive scores and the cut-off points for benefit group based on all 133 patients are shown in Table 5.

Table 5 Coefficients of individual genes of the 34-gene signature in the 1st principal component (1st column of the loading coefficient)

Probeset	PC1 loading coefficient	Probeset	PC1 loading coefficient
205241_at	0.135	203929_s_at	-0.066
219202_at	0.153	214106_s_at	-0.083
203147_s_at	0.236	205076_s_at	0.197
201016_at	-0.185	205552_s_at	0.262
221609_s_at	-0.080	202036_s_at	-0.169
220442_at	0.120	204437_s_at	0.185
218640_s_at	-0.071	204867_at	0.206
203764_at	-0.199	202869_at	0.254
204068_at	-0.145	214056_at	0.132
207663_x_at	-0.091	206134_at	-0.034
213032_at	-0.075	219670_at	-0.131
217730_at	0.235	204580_at	-0.072
219429_at	0.148	207713_s_at	0.159
219140_s_at	0.108	201017_at	-0.208
205874_at	0.171	200661_at	0.264
217995_at	0.250	203449_s_at	-0.212
219148_at	-0.215	217794_at	0.170

Predictive score = 0.816 * PC1

The cut-off points of patients classified as low, middle or high predictive score groups are -0.734, and 0.810 using 1/3 and 2/3 quantiles of the predictive scores. The cut-off point of patients classified as low and high predictive score groups using median of the predictive score is -0.0132.

3.4.3 Internal validation

71 patients were resampled 200 times from the all 133 patients with replacement to test the predictive ability of the selected 34-gene signature. The log-rank test p-value and hazard ratio

of the two treatment arms (estimate of treatment effect) as well as the number of patients in the low predictive score group when the patients were divided into three and two groups are shown in Figures 14 and 15, respectively. The histograms for the estimated treatment effects (hazard ratios) of the bootstrapped samples are plotted in Figure 16.

From the Figure 14 and 16(a), in which the patients were classified into low, middle, and high gene groups, of the 200 times validations, there are 16 validations (8%) that the log-rank test p-value > 0.05, while 5 validations (2.5%) with the p-value > 0.10. The hazard ratios are all less than 0.4 except one validation. The average number of patients in the low score group is 23, which is slightly lower than 1/3 of the total number of patients in the validation. When the patients were classified into two gene groups, as shown in Figure 15 and 16(b), 42 validations have the log-rank test p-value > 0.05, which is about 20% of the total number of validation, and 18 validations (9%) have the p-value greater than 0.1. 34 validations (17%) have the hazard ratio greater than 0.4, and averagely about half patients were classified into the low predictive score group.

The internal validation results further demonstrate that the selected 34-gene signature can identify the subset of patients who are beneficial to the adjuvant chemotherapy.

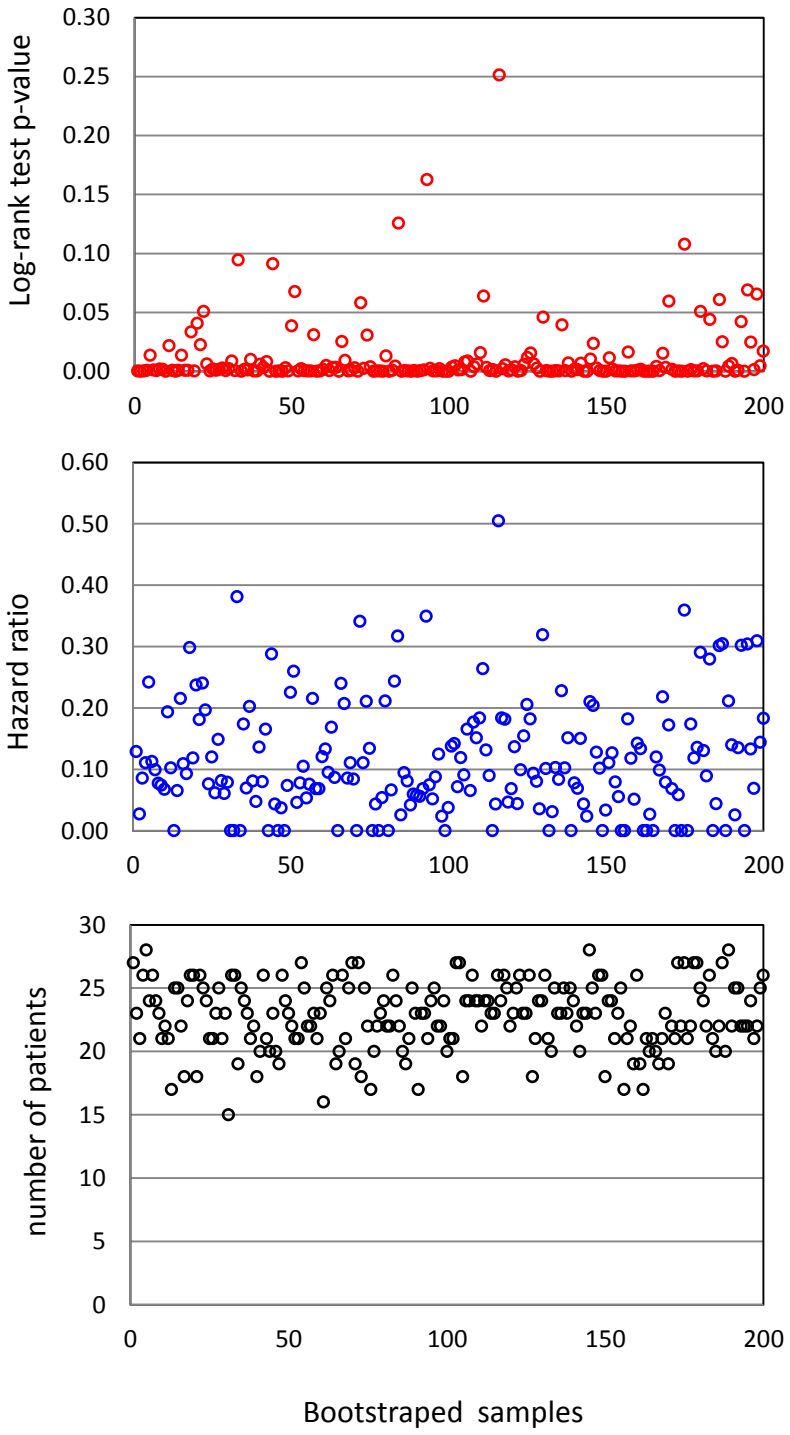


Figure 14 Internal validation results of patients in low predictive score group when patients are classified into three predictive score groups. (a) Log-rank test; (b) Hazard ratio; (3) number of patients in the low score group.

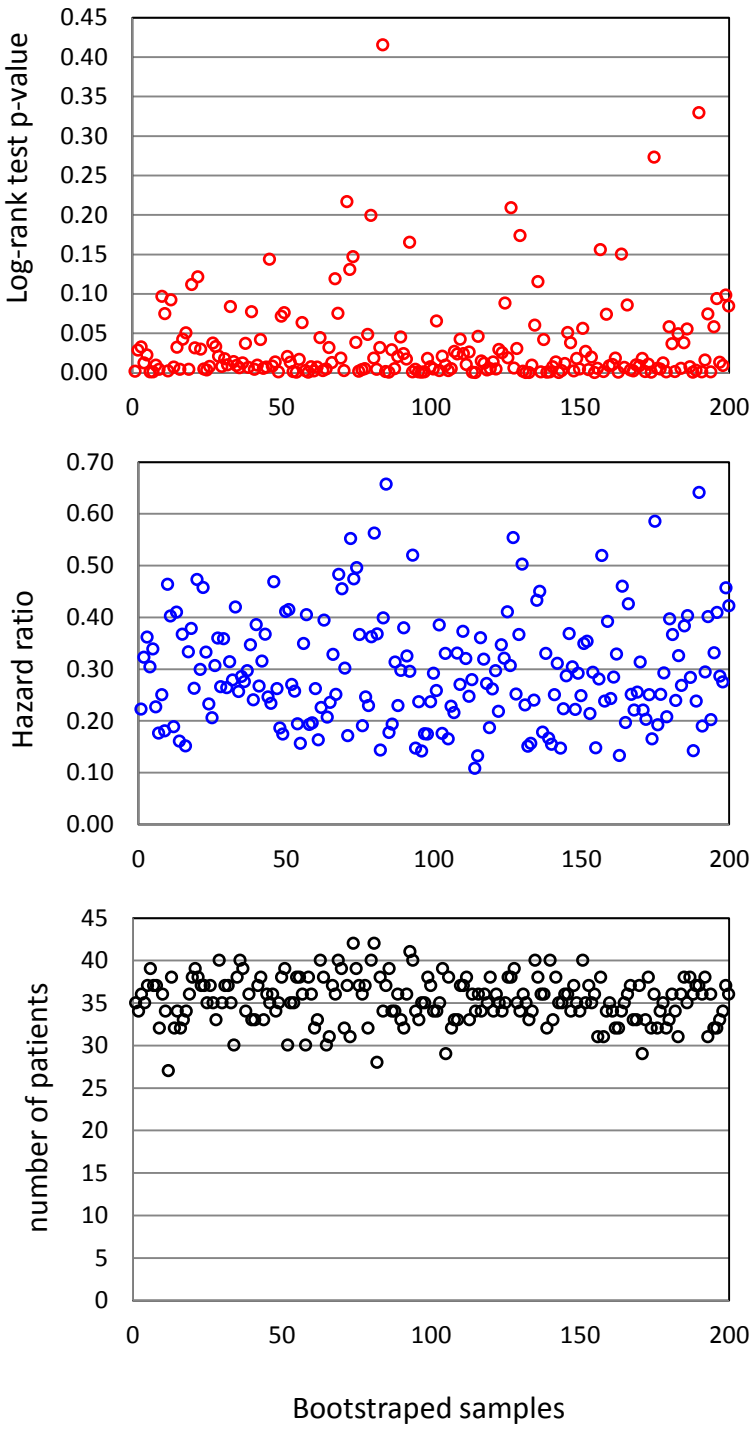
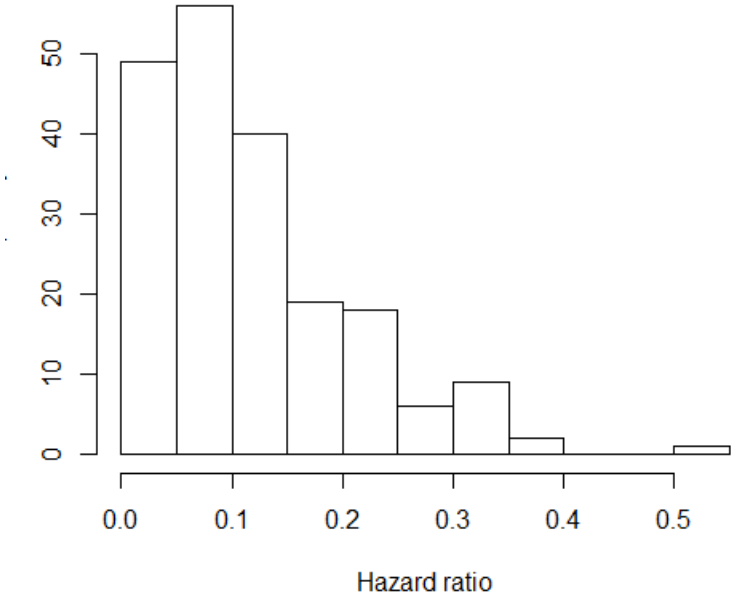
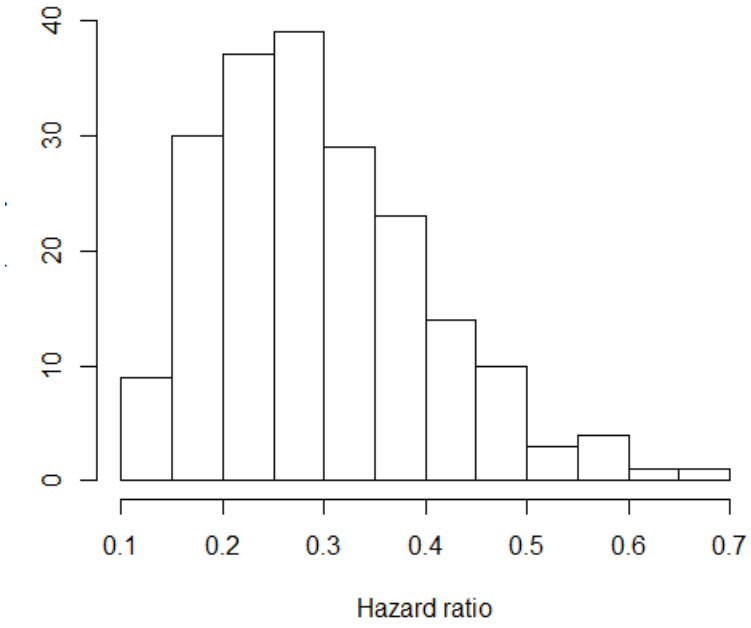


Figure 15 Internal validation results of patients in low predictive score group when patients are classified into two predictive score groups. (a) Log-rank test; (b) Hazard ratio; (3) number of patients in the low score group.



(a)

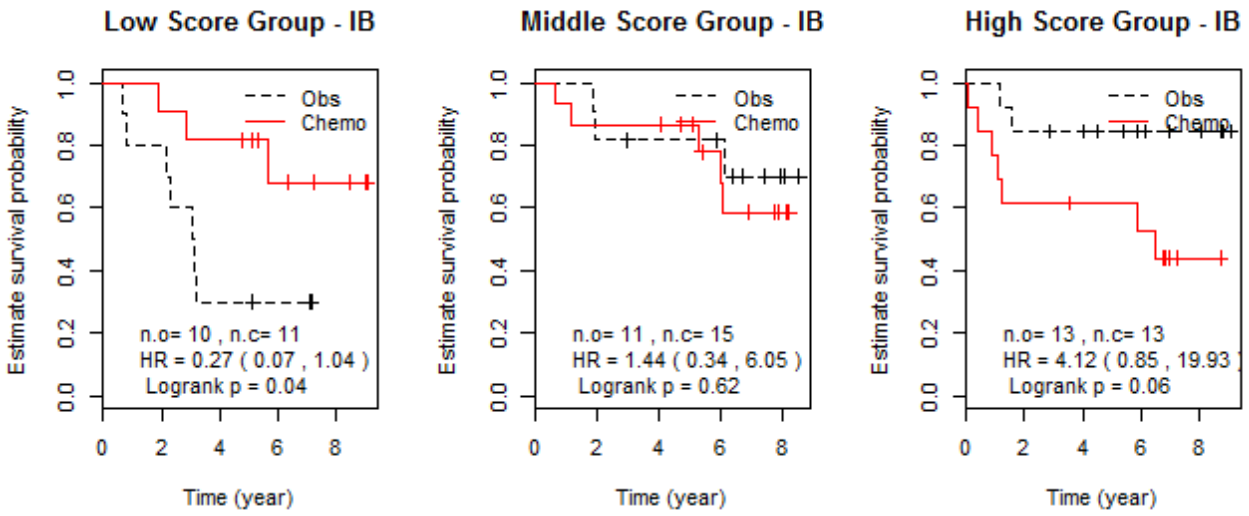


(b)

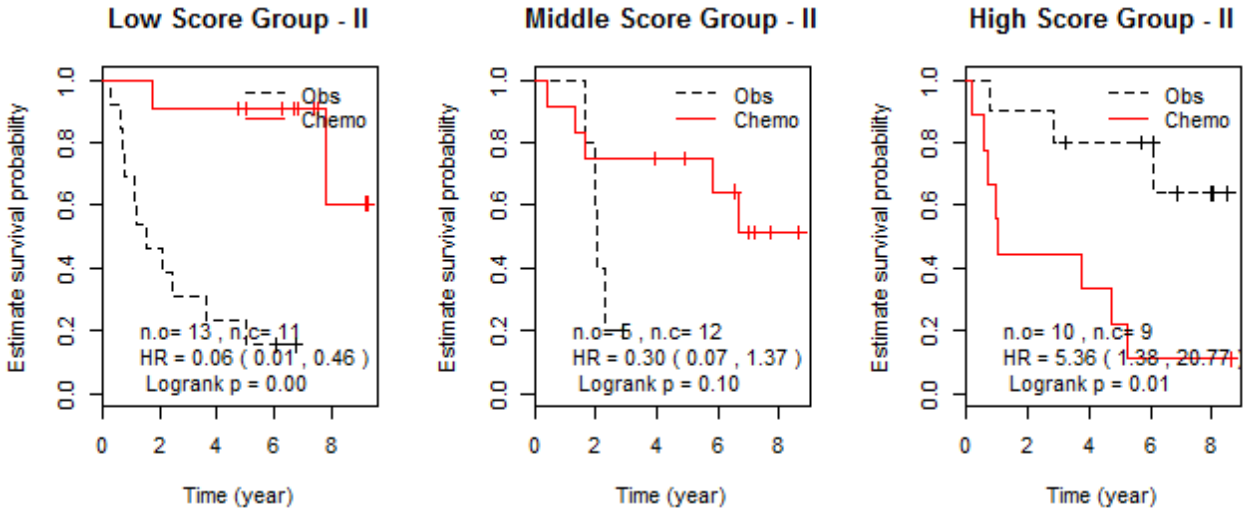
Figure 16 Histograms of the estimated treatment effects (HR) in low predictive score group by internal validations. (a) Three predictive score groups; (b) Two predictive score groups.

3.4.4 Stratified by disease stage

The BR.10 patients have disease stages IB and stage II. In order to investigate the effects of disease stages on the treatment benefit, the all 133 patients are stratified by their disease stage. The overall Kaplan-Meier survival plots in the three predictive score groups using 1/3 and 2/3 quantiles of the predictive scores as cut-off points are shown in Figure 17. It can be seen that 22 patients with disease stage IB in low score group are significant benefit from the chemotherapy (HR 0.27; 95% CI, 0.07 to 1.04, and log-rank test p-value of 0.04). The patients with IB disease stage in the middle score group have similar survival outcomes across the two treatment arms (HR 1.44; 95% CI 0.34 to 6.05, and p-value 0.62), while worse chemo therapy effects were observed for the stage IB patients in the high score group (HR 4.12; 95% CI 0.85 to 19.9; p-value of 0.06). However, for the stage II patients, chemotherapy benefit was observed in both low predictive score and middle score groups (HR 0.06, 95% CI 0.01 to 0.46; and p-value < 0.01 for low score group, and HR 0.30; 95% CI 0.07 to 1.37; and p-value of 0.10 for middle risk group, respectively). Therefore, the chemo therapy benefit is stage dependent. 1/3 quantile of the predictive score seems appropriate to classify the benefit group of the patients with IB disease stage, while a larger value of the cut-off point should be used to classify benefit group for the stage II patients, so that more stage II patients who are benefit from the chemotherapy can be identified. Therefore, 2/3 quantile of the predictive score was used to classify stage II patients into low and high score groups, shown in Figure 18. In this case, 41 patients who are in the low score group are benefit from the chemotherapy with HR of 0.16 (95% CI 0.06 to 0.44) and log-rank test p-value < 0.01, while 19 patients are in the high score group, who are detrimental to the chemotherapy with HR of 5.36 (95% CI of 1.38 to 20.77) and log-rank test p-value of 0.01.



(a)



(b)

Figure 17 Overall survival of 133 patients in three predictive score groups based on 34-gene signature stratified by disease stages. (a) stage IB; (b) stage II.

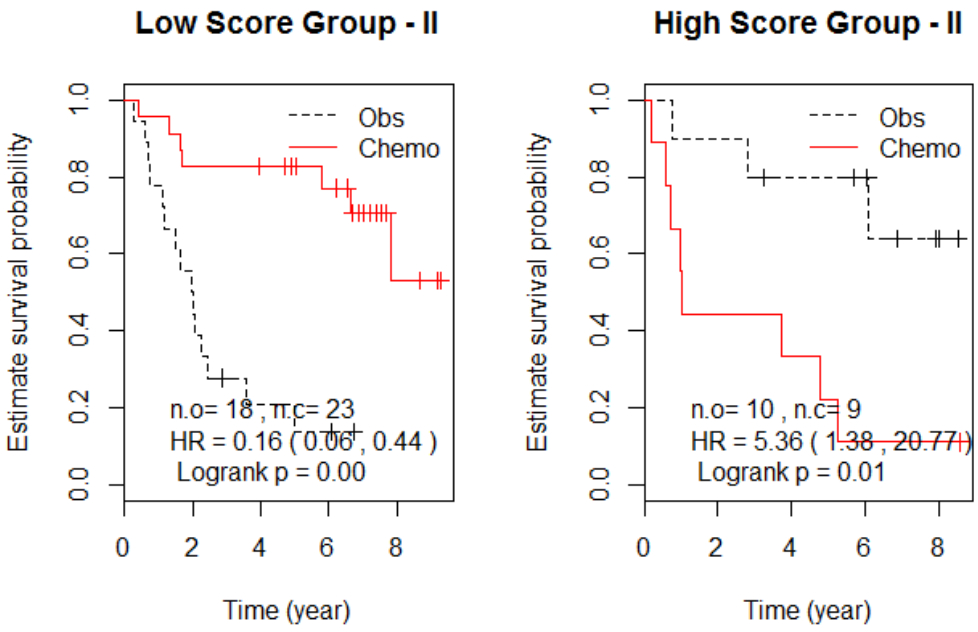


Figure 18 Overall survival of stage II patients in low and high score group based on 34-gene signature using 2/3 quantile of predictive scores as cut-off point.

Due to the stage dependence of the chemotherapy benefit, different cut-off points to classify the patients with stage IB and stage II diseases are used here. Figure 19 shows the survival estimates of patients in low and high predictive score groups when 1/3 quantile of the predictive scores used as the cut-off point for stage IB patients while 2/3 quantile of the predictive scores is as the cut-off point for stage II patients. In this way, 62 patients who are benefit from the chemotherapy are well identified in the low predictive score group with HR of 0.19 (95% CI, 0.09 to 0.42) and log-rank test p-value of < 0.01. In the high predictive score group the patients are lack of benefit from the chemotherapy (HR, 2.87; 95% CI, 1.26 to 6.51; and log-rank test p-value = 0.01).

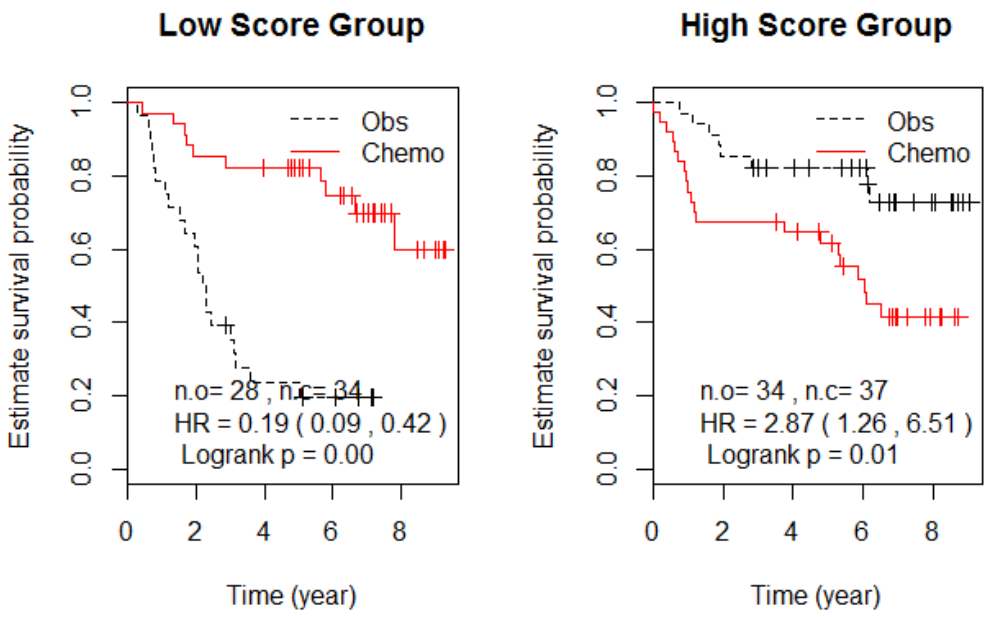


Figure 19 Overall survival of 133 patients in low and high predictive score groups based on 34-gene signature using 1/3 quantile of predictive scores as cut-off point of stage IB patients and 2/3 quantile of predictive scores as cut-off point of stage II patients.

3.5 Summary

In BR. 10 randomized controlled clinical trial, 133 patients with Affymatrix microarray gene expression profiling were used to select predictive gene signature to identify the subset of patients who are truly benefit from the adjuvant chemotherapy. The gene probesets were initially screened by excluding about 1/3 probesets with low variance across the samples. Then univaraite analysis by fitting Cox's regression model using the modified covariate with only interaction term between gene expression level and treatment without main effect was carried out, and 664 gene probesets were preselected. The dataset then was randomly divided into training and testing sets with similar survival experience by treatment arms. The bootstrap resampled training

datasets were fitted to the Cox's regression model using the modified covariates with the preselected gene probesets. Elastic net regularization method was used for variable selection. The predictive scores used for predictive gene selection were computed by the cross validation. 34-gene signature was selected based on the frequency of gene probesets appeared during the model selection and predictive ability of survival benefit in low predictive score group. The predictive scores of patients in the testing set were constructed based on the information in the training set, and patients were classified into low, (middle) and high predictive score groups. The 34-gene signature successfully identified the patients in the low score group are beneficial to adjuvant chemotherapy (HR, 0.13; 95% CI, 0.01 to 1.14; log-rank test p-value = 0.03). Internal validation was implemented in the datasets that randomly resample 71 patients from the all dataset with replacement. Of all 200 times validations, the log-rank test p-values are less than 0.05 in more than 90% validations and most of hazard ratios are less than 0.4. Therefore, the 34-gene signature can well predict the survival benefits from the adjuvant chemotherapy in NSCLC patients. Moreover, stage dependent chemotherapy benefit was also observed, and the chemotherapy benefit group can be better identified using different cut-off points for the stage IB and stage II patients.

Chapter 4

General conclusions and Future Work

4.1 General conclusions

BR.10 microarray data analysis was conducted to select predictive gene signature to identify a subset of early stage resected non-small cell lung cancer patients who are potentially beneficial to the adjuvant chemotherapy. The study involves in 133 patients with prospectively collected tumor samples and Affymetrix U133A microarray gene expression profiling available. The raw microarray data were preprocessed by several normalization and batch effect removal methods.

After normalized by RMA method and adjusting batch effect using the distance-weighted discrimination method, the microarray data were pre-screened to remove the gene probesets with low variance across samples, and then 664 gene probesets were preselected by the univariate analysis using the modified covariate without main effect to fit the Cox's regression model. A group of the gene probesets as the predictive gene signature were extracted based on the bootstrap resampled training datasets and multivariate Cox's regression model with modified covariates, i.e., the interaction term of chemotherapy treatment and the standardized preselected gene probesets expression values, using elastic net regularization for variable selection, and principal component analysis and cross validation for developing predictive scores. A subset of patients who are benefit from the adjuvant chemotherapy was identified based on a 34-gene signature in the low predictive score group of the test dataset. The hazard ratio of 0.13 (95% CI,

0.01 to 1.14) and log-rank test p-value 0.03 were achieved for the patients received chemotherapy vs. the patients in the observation group. Multiple internal validation results further indicated that the 34-gene signature can predict survival benefits from the adjuvant chemotherapy successfully for the early stage resected NSCLC patients.

4.2 Future work

1. In this work we only performed internal validation because lack of other published randomized clinical trial data for comparison of early stage resected non-small cell lung cancer patients who received chemotherapy and observation. External validation using independent datasets will further test the predictive ability of the 34-gene signature and the generalizability of the purposed method.
2. Because disease stage dependence of chemotherapy benefit was observed, i.e., there may exist different cut-off points for the patients with disease stage IB and stage II to identify the chemotherapy benefit group, further analysis to select gene signature and predict treatment effect considering disease stages are suggested.
3. The batch effect removal methods were qualitatively evaluated through visualization techniques, which is only a crude approximation of the efficiency of batch effect removal. However, for a more rigorous evaluation, quantitative measures should also be computed to accurate assess the quality of the batch effect removal process.
4. Moreover, the predict gene signature was selected by statistical methods and mathematical algorism. The selected genes could be linked with their corresponding biological features to better understand the relationship between the gene expression and the disease outcomes.

References

[1] Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2014. Toronto, ON: Canadian Cancer Society (2014).

[2] Raponi M, Zhang Y, Yu J, Chen G, Lee G, Taylor JM, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res. 2006; 66: 7466–72.

[3] Tang H, Xiao G, Behrens C, et al. A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. Clin. Cancer Res., 2013; 19: 6, 1577-86.

[4] Zhu CQ, Ding K, Strumpf D, Weir BA, Meyerson M, Pennell N, et al. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. J. Clin. Oncol. 2010;28: 4417–24.

[5] Simon, R., Korn E., McShane L, et al. Design and analysis of DNA microarray investigations. Series: Statistics for Biology and Health. Springer, 2004

[6] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of affymetrix genechip probe level data. Nucleic Acids Res. 2003; 31:e15.

[7] Gautier,L. et al. Affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004; 20, 307–15.

[8] Lim WK1, Wang K, Lefebvre C, Califano A. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics. 2007 ; 23(13):282-8.

[9] Gene Expression Ominbus (GEO) database – NCBI, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed

[10] Lazar C, Meganck S, Taminau J, Steenhoff D et al., Batch effect removal methods for microarray gene expression data integration: a survey. Brief Bioinform. 2013; 14(4):469-90.

[11] Fare TL., Coffey EM., Dai, H., et al. Effects of atmospheric ozone on microarray data quality. Analytical Chemistry. 2003; 75, 4672–5.

[12] Johnson WE, Li C. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8, 1, 118–127

[13] Jolliffe IT. Principal Component Analysis. 2nd edn. Springer, 2002.

[14] Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010;11(10):733–9.

[15] Tian L, Alizadeh A, Gentles J, Tibshiran R. A Simple method for detecting interactions between a treatment and a large number of covariates. Submitted to JASA, Mar 2014, in press.

[16] Hoerl, A. and Kennard, R. Ridge regression. In Encyclopedia of Statistical Sciences. 1988; 8 : 129–136. New York: Wiley.

[17] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 1996; 58: 267–288.

[18] Zou H., and Trevor H. Regularization and variable selection via the elastic net. J. R. Statist. Soc. B, 2005; 67 : 301-320

[19] HT_HG-U133A Annotations, Affymetrix. http://www.affymetrix.com/support/technical/annotationfilesmain.affx