國立臺灣大學理學院數學系
碩士論文

Department of Mathematics

College of Science

National Taiwan University

Master Thesis

生物晶片資料分析與肺腺癌存活之預測

Microarray Data Analysis and Prognosis of Lung Adenocarcinomas

何昊
Hao Ho

指導教授: 李克昭 教授
Advisor: Professor Ker-Chau Li

中華民國九十九年二月
February 2010

誌謝

七年前，第一次走進台大數學系。一轉眼七年過去了，也將正式離開這充滿回憶的地方。在這裡，我度過了大半的青春時光，失去了很多，得到了更多。所有好的、不好的，開心、不開心的，都將化成我最珍貴的回憶。

首先要感謝的是養我育我、對我萬般包容的家人。感謝爸爸、媽媽無論何時何地都給予我最大的支持。感謝爺爺、奶奶在我小的時候對我的照顧及教育。感謝所有無論住在一起或沒住在一起，一直關心我的家人。請原諒我的不開朗以及任性。我愛你們，謝謝你們給了我一切最好的。

感謝李克昭老師的指導，寬廣了我的視野，更見識到問問題的藝術。感謝老師您給我如此大的空間獨立地思考、學習、研究，又總在關鍵時候指導我什麼才是真正最重要的，讓我不會迷失在自已絮亂的想法之中。感謝陳宏老師，因為您的啟蒙，讓我決定走上統計這條路。感謝江金倉老師亦師亦友、直言不諱的給予當頭棒喝以及鼓勵。感謝陳素雲老師不厭其煩的提點，無論在統計以及英文寫作上都讓我獲益良多。感謝蕭朱杏老師在我的研究過程中給予的所有幫助。感謝袁新盛老師耐心的陪我討論我那多且煩雜的問題。

感謝所有曾在我的世界中留下不同軌跡的朋友們。感謝你們給了我另一份歸屬感。感謝數學系的大伙兒、天外天的兄弟們、ＭＩＢ的同伴們、替代役的長官及夥伴們、歷任資Ａ資Ｂ的家庭成員們，還有所有關心我以及曾關心我的知己們，謝謝你們帶給我生命中不同的美麗風景。感謝我最愛的妳，陪我走過這十年的求學生涯。感謝妳的好，讓我成長。一路上要感謝的人實在太多，然而，誌謝可用的篇幅又太少，請原諒我無法將心中所有的感謝在此一一訴盡。

最後，僅將所有的感謝化成兩個字，再一次的向你們說：謝謝！

中文摘要

近年來肺癌高居國內及全球癌症相關死因首位，其死亡率至今仍然居高不下。非小細胞肺癌乃發生率最高之肺癌，其中又以肺腺癌最為普遍。研究指出，肺癌病患的治療方式不只取決於腫瘤類型，而不同肺癌分期亦應適當選擇給予不同治療方式。因此，為幫助病患選取最有利之治療方式，建立更準確之新肺癌診斷方法，有其重要性及迫切性。其中，因前期肺癌病患仍有數種治療方式可選擇，故對於前期肺癌病患之診斷最為重要。

生物晶片技術的發展使得研究人員得以同步測量數以萬計之基因表現量，並提供了新的研究平台。一項大型的肺腺癌研究建立了豐富的肺腺癌病患之基因表現量資料和臨床資料，用以建立及驗證數種以基因表現量導出之肺癌診斷方法。然而，對於前期病患的存活預測，尚未能找到一種基因訊號對於所有的驗證資料皆能達到百分之五之顯著水準。我們的研究目的即是重新研究這份資料，以期建立一個新的基因訊號對於所有的驗證資料皆能有顯著的存活預測力。

我們引用一個兩階段之維度縮減方法佐以部份的修正來導出肺癌診斷之基因訊號。在第一階段中，我們以基因和存活時間的相關性與配對基因和存活時間的流動關聯性來選出重要的候選基因。第二階段，我們則使用了改良的切片逆迴歸分析方法導出最後的肺癌診斷基因訊號。

分析結果指出，以相同的驗證流程檢驗我們導出的基因訊號，在所有的驗證資料都能達到百分之五之顯著水準的存活預測能力，更進一步的，在另一個包含肺腺癌及鱗狀細胞癌的非小細胞肺癌資料上，我們導出的基因訊號也同樣達到百分之五之顯著水準的存活預測能力。因此，我們認為以TMEM66，CSRP1，BECN1，FOSL2，ERO1L，SRP54及PAWR七個基因所導出的基因訊號對於非小細胞肺癌病患有好的存活預測能力。

Abstract

Purpose Recently, several new gene expressions based signatures were proposed to predictive the survival of Non Small Cell Lung Cancer (NSCLC) patients. However, for stage I patients, the task is more difficult and no signatures had been found from a large study of lung adenocarcinoma. We reanalyzed this large sample data and tried to construct a gene signature, which had significant prediction power for all stage and early stage patients in all the validation sets. We also used an external independent cohort data set containing both adenocarcinomas and squamous cell carcinomas to test if our gene signature still had significant prediction power for all stage and early stage NSCLC patients.

Materials A total of 442 lung adenocarcinoma gene expression profiles from Shedden *et al.* (2008) containing four independent data sets were reanalyzed in our study. Two of the data sets were combined as a training data set to derive our gene signature. The other two data sets were used for validation. An external NSCLC data from Duke lung caner cohort was used for additional validation.

Methods We modified a two-steps dimension reduction method proposed by Wu *et al.* (2008) to derive our gene signature. In the first step, both correlation and liquid association methods were used to select the candidate genes. In the second step, we applied the modified sliced inverse regression proposed by Li *et al.* (1998) to derive a gene signature from the candidate genes.

Results Five genes TMEM66, CSRP1, BECN1, FOSL2 and ERO1L were selected by correlation methods. SRP54 and PAWR (as a LA pair) were selected by liquid association method. The final signature gave significant

prediction power for samples with all stage patients and for samples with stage I patients only in all the validation sets.

Conclusion The gene signature derived from the seven genes (TMEM66, CSRP1, BECN1, FOSL2, ERO1L, SRP54 and PAWR) had good prediction power for all stage and early stage NSCLC patients.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the latest years, lung cancer is the leading cause of caner death in many Western countries and also in Taiwan. The most common cell type of lung cancer is Non Small Cell Lung Cancer (NSCLC), especially the lung adenocarcinomas [1] [2]. The treatment selection of different patients is dependent on various factors including tumor cell types and cancer stages. Recently, the adjuvant chemotherapy has been proved for significantly improving the survival of stage IB and stage II patients [3]. However, for some early stage patients with good prognoses, the benefit is not significant. To help early stage patients select the most beneficial treatment, a new diagnostic method is urgently needed.

The development of microarray technology has helped researchers to measure more than ten thousands of gene expressions simultaneously. From the information contained in these data, several new gene expression based signatures were introduced to predict to survival of NSCLC patients [4]-[9]. However, to reproduce and to validate these signatures in general are not easy. A large sample lung adenocarcinomas data containing four independent cohort data sets was generated to compare the performance of several latest expression based diagnostic signatures [10]. Nevertheless, the study did not found a signature having significant prediction power for samples with stage I patients only in all the validation data sets.

We reanalyzed this large adenocarcinomas data starting from data preprocessing to filter out some non-informative genes. A two-steps dimension reduction method proposed by Wu *et al.* (2008) [11] was then applied with some modifications for deriving diagnostic signature on the training set. In the first step, both correlation and liquid association methods [13] [14] were used to select the candidate genes. In the second step, we applied the modified sliced inverse regression proposed by Li *et al.* (1998) [16] to derive the signature from the candidate genes. Based on the same validation procedure for all the testing data as in Shedden *et al.* (2008), our signature gave significant prediction power for samples with all stage patients and samples with stage I patients only. Furthermore, we also used an external independent cohort data set containing both adenocarcinomas and squamous cell carcinomas to test if our gene signature still had significant prediction power for all stage and early stage patients. Our gene signature still gave significant prediction power in this validation data set.

Here we summarize the main contents of this thesis. We introduce all the cohort data sets used in our analysis and our analysis procedure in Chapter 2. Three different criterions for gene filtering are introduced in Chapter 3. In Chapter 4, we introduce how we select the candidate genes by correlation and liquid association (LA) methods. A modified imputation method for censored data is also given. In Chapter 5, we discuss how we derived the gene signature from the selected candidate genes. A brief introduction for modified sliced inverse regression and its implementation is also given. The validation results and cross platform adjustment for testing sets are presented in Chapter 6. In Chapter 7, we discuss the genes we used in our signature, the relationship between our permutation procedure and Benjamini-Hochberg procedure under independent assumption in our data setting, and the effect, which might be caused by different preprocessing methods.

Chapter 2

Materials and analysis procedure

2.1 Materials

In this section, we introduce several independent cohort data sets we used to construct our gene signature or to test the prediction power of our gene signature. Each data set contains gene expression profiles measured by Affymetrix microarray and survival outcome data of Non Small Cell Lung Cancer (NSCLC) patients. Some relevant clinical and pathological data, such as sex, age, TNM tumor stage and tumor cell type are also available. The summary statistics of clinical variables in each data set are given as table 2.1.

442 lung adenocarcinoma data from Shedden *et al.* (2008)

In Shedden *et al.* (2008), a large lung adenocarcinomas microarray data was generated from four institutions, Moffitt Cancer Center (HLM), University of Michigan Cancer Center (UM), Dana-Farber Cancer Institute (CAN/DF)

Table 2.1: Summary statistics of the clinical and survival data

	HLM	UM	CAN/DF	MSK	Duke
Sample size	79	177	82	104	111
Age (mean)	67	64	61	65	65
Cell type (% adenocarcinomas)	100%	100%	100%	100%	52%
Sex (% male)	51%	56%	56%	36%	57%
Stage IA	11%	39%	13%	26%	36%
Stage IB	43%	27%	55%	35%	24%
Stage II	27%	16%	32%	19%	16%
Stage III	19%	18%	0%	20%	20%
Stage IV	0%	0%	0%	0%	4%
Median follow up (months)	39	54	51	43	31
Number of deaths	60	102	35	39	58

and the Memorial Sloan-Kettering Cancer Center (MSK), using a common platform, Affymetrix U133A. After excluding samples with poor quality of microarray data or incomplete clinical data, a total of 442 samples from four independent data sets were analyzed in their study. They analyzed the data for validating the following four hypotheses: (1) can gene expression predict the survival of all stage patients? (2) Can gene expression predict the survival of stage I patients? (3) Can gene expression with clinical covariates (stage, sex and age) predict the survival of all stage patients? (4) Can gene expression with clinical covariates predict the survival of stage I patients? They combined the first two data sets, HLM and UM, as a training data set to construct the gene signature, and used the rest two independent data sets, CAN/DF and MSK, to test the prediction power of the derived signatures. In their reports, with the clinical covariates, several methods gave significant prediction power for all stage and stage I patients in the rest two validation sets, CAN/DF and MSK. However, without clinical covariates, there were no signatures found with significant prediction power for stage I patients in both validation sets. Therefore, in our analysis, we focused on the first and the second hypotheses. Our purpose was to derive a gene signature from the training data set which had significant prediction power for all stage and stage I patients in both validation data sets. The raw data were downloaded from https://caarraydb.nci.nih.gov/caarray/publicExperimentDetailAction.do?expId=1015945236141280.

111 NSCLC samples from Duke lung caner cohort

Another microarray data set from Duke lung caner cohort containing 111 NSCLC samples was used to test our gene signature. This data set was a more challenging external validation data set. First we noted that it contained patients of two different tumor cell types, adenocarcinomas and squamous cell carcinomas. The prediction power of our gene signature for patients with different tumor cell types can be tested in this validation set. Second, since the gene expression profiles were measured by a different microarray platform Affymetrix HU133plus2, a cross platform adjustment was needed. The detail of our adjustment is discussed in Chapter 6.

2.2 Analysis procedure

In this section, we briefly introduce our analysis procedure and the main methods we used to construct our gene signature. The theoretical and implemented details of these methods are given in later chapters.

The major challenge in microarray data analysis is the large dimensionality. The number of genes (G) is in the range of ten to fifty thousands but the sample size (n) is only about hundreds. To reduce the effect of microarray noise, we used three criterions to filter out non-informative genes. We excluded the genes with low expressions, small variation expressions or with inconsistent expressions between three preprocessing methods, the MAS 5.0 Statistical algorithm from Affymatrix (2001) [18], the dChip algorithm from Li and Wong (2001) [19] and the Robust Multi-chip Average (RMA) from Irizarry *et al.* (2003) [20]. The detail of this part is discussed in Chapter 3.

After gene filtering, we reduced the number of genes (G) to a related smaller number (G^*), but the true effective dimensions to the survival time might be much smaller. Our strategy was to implement a two-steps dimension reduction method proposed by Wu *et al.* (2008) with some modifications. This approach contained gene selection and signature construction two steps. In the gene selection part, we used both correlation and liquid association methods to select the important candidate genes related to survival time. Pearson's correlation coefficient was introduced to measure the strength of linear dependency between two variables. However, the association between gene expressions and patients survival might not be linear and might be more complicated. The liquid association (LA) method was implemented here to explore the interaction of two genes related to the survival time. Due to the data censoring issue, both correlation and liquid association could not be implemented directly. Therefore, we modified a nonparametric imputation method [11] to impute the censored data, then the correlation coefficients could be evaluated by plugging the imputed survival probability. After that, we calculated and ranked the correlation coefficients between gene expressions and the imputed survival probability by the absolute values. Genes in the first few places were selected as candidate genes in this part. We also proposed a permutation procedure to decide how many genes should be selected. For implementing the liquid association (LA) method, our strategy was to select the genes, which recurrently appeared in the first few extreme LA gene pairs. These genes were called the LA hub genes. We selected the LA hub gene and also its paired genes as candidate genes in this part.

We note that the gene expression profiles and the imputed survival probability were normalized by normal quantile transformed in the first two parts, gene filter and the gene selection. The normal quantile transformation is necessary for the liquid association method and makes the procedure robust against the outliers. Both the correlation and liquid association can be computed in the website http://kiefer.stat2.sinica.edu.tw/LAP3/index.php. The details of the imputation methods and liquid association method are given in Chapter 4.

In the signature construction part, the candidate genes selected from the previous step were used to derive a gene signature for survival prediction. First, we applied the modified sliced inverse regression to estimate the effective dimension reduction (e.d.r.) directions and projected the selected gene expression profiles on the e.d.r. space. If there is only one SIR direction, the estimated e.d.r. direction, found significantly by a large sample chi-squared test, we projected the expression profiles on it as our final gene signature. Otherwise, we could use the projected directions to fit other survival model, for example the multivariate Cox proportional hazard model, and derive the final gene signature. We note that the normal quantile transformation was not used in this part. The regressors (**X**) in the dimension reduction model were the candidate gene expression profiles transformed by log-2 transformation and centered toward sample mean in training data set. The theoretical derivation and practical implementation of modified sliced inverse regression are given in Chapter 5.

The prediction power of our gene signature was tested in the independent validation data sets. In each validation data set, we used the linear combination coefficients estimated from the training data set to combine the selected gene expressions into a gene signature. We used two ways to present the prediction power of our signature. First we used median of our signatures in each validation set to separate the samples into two groups, high risk and low risk groups, as a categorical classifier. For this categorical classifier, the log rank test was used to test the difference of the survival distribution of two groups. Second, we used our derived gene signature as a continuous risk score to fit the Cox proportional hazard model. We estimated the hazard ratios with corresponding p-value and the concordance probabilities (CPE) [23] for both categorical classifier and the continuous risk score. The CPE estimated the probability that survival outcome agreed with the risk score or categorical classifier under the Cox proportional hazard model. To compare the derived gene signature and the TNM tumor stage the results of multivariate Cox proportional hazard model were also presented. A flow chart of our procedure is given as figure 2.1.

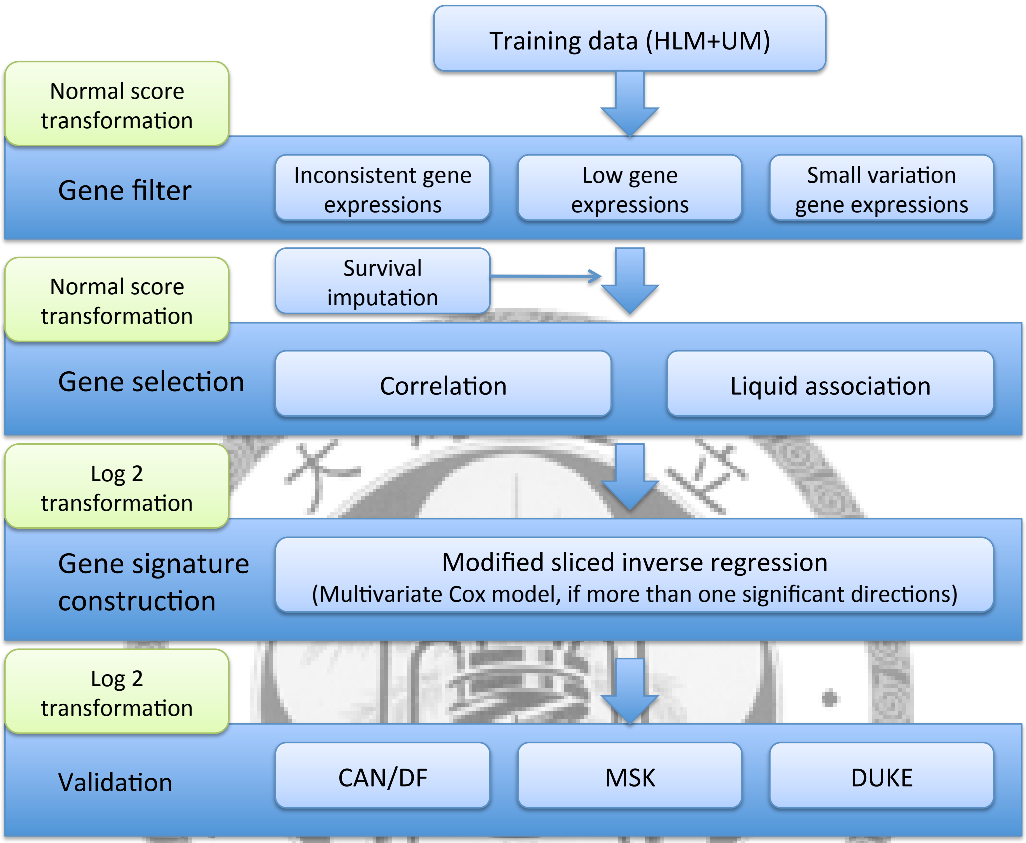

Training data (HLM+UM)
Normal score transformation
Gene filter
Inconsistent gene expressions
Low gene expressions
Small variation gene expressions
Normal score transformation
Survival imputation
Gene selection
Correlation
Liquid association
Log 2 transformation
Gene signature construction
Modified sliced inverse regression
(Multivariate Cox model, if more than one significant directions)
Log 2 transformation
Validation
CAN/DF
MSK
DUKE

Figure 2.1: The flow chart of analysis procedures.

Chapter 3

Gene filter

3.1 Inconsistent gene expressions

At the beginning of microarray analysis, choosing data preprocessing method is still an open issue. MAS 5.0 Statistical algorithm, dChip algorithm and the Robust Multi-chip Average (RMA) method are three widely used data preprocessing methods. However, different methods may lead to different results. In Shedden *et al.* (2008), they preprocessed the expression profiles by running dChip algorithm on all four data sets together. Nevertheless, there was an issue they remarked: running the dChip algorithm on the entire data sets may have removed some of the inter-site differences but is somewhat unrealistic. Figure 3.1 showed a dramatic shift of the data indicating that gene expressions preprocessed separately or together as a group using dChip algorithm are not comparably scaled. This inter-site difference may impact the validation results a lot.

In our data analysis, we chose the MAS 5.0 Statistical algorithm for data preprocessing. The MAS 5.0 algorithm allowed us to preprocess the microar-

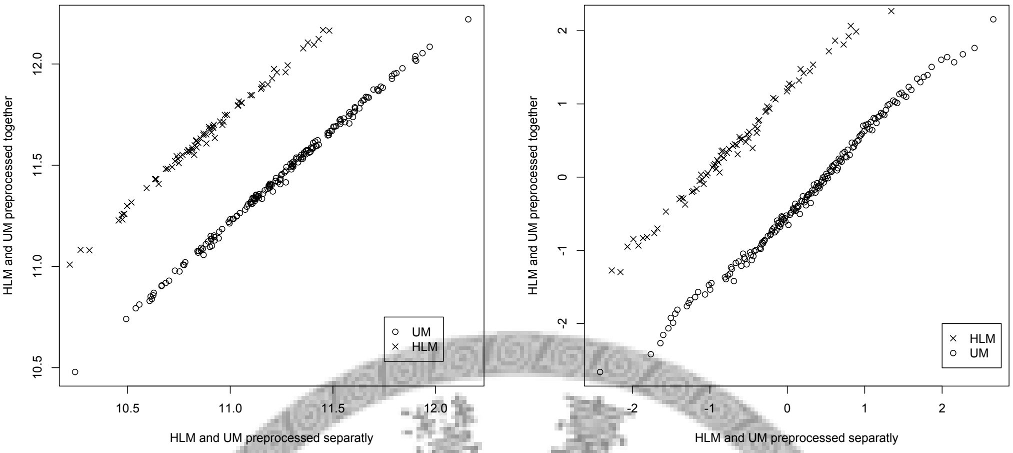


Figure 3.1: Scatter plots of gene expression profiles of DDR1 in HLM and UM two data sets preprocessed together versus preprocessed separately by dChip algorithm: The left panel is log-2 transformed gene expression profiles and the right panel is normal quantile transformed gene expression profiles.

ray data entirely or chip by chip with the same results. However, we thought that the genes with inconsistent expression profiles between three preprocessing methods were unconvinced. Therefore, we filtered out the genes that had inconsistent expression levels between three preprocessing methods. Correlation coefficient is a measure also used to measure the similarity of two variables. Here we used it to measure the similarity of the expression levels preprocessed by each two of the three preprocessing methods for each gene. Nevertheless, since the RMA preprocessed data is in log-2 scale, we may transform it before we calculated the correlation coefficients. Furthermore, the normal quantile transformed correlation coefficients between gene expressions and imputed survival probability is an important selecting criterion in the gene selection part. Thus, we also used the normal quantile transformed correlation coefficient in this part.

Before continuing the introduction of the gene filter, here we give a definition of the normal quantile transformation and note some properties of it.

Definition 3.1.1. For any n observations $\mathbf{x} = (x_1, ..., x_n)'$ of variable X, The normal quantile transformation is define by

$$N(\mathbf{x}) = \left(\Phi^{-1}(\frac{1}{n+1}R_1), \Phi^{-1}(\frac{1}{n+1}R_2), ..., \Phi^{-1}(\frac{1}{n+1}R_n)\right)',$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution and R_i is the rank of x_i in $\mathbf{x}$ for $i = 1, 2, ..., n$.

Since Pearson's correlation coefficient with normal quantile transformation only depends on the rank of observations, it can be viewed as a kind of rank correlation coefficient. It is more robust against outliers than the original Pearson's correlation coefficient. Furthermore, in elementary statistics, the correlation coefficient between N($\mathbf{x}$) and $\mathbf{x}$ is used to test for the null hypothesis that X is normally distributed. The correlation coefficient between N($\mathbf{x}$) and $\mathbf{x}$ is closed to 1 under null hypothesis. If X_1 and X_2 are both normally distributed, the correlation coefficient between N($\mathbf{x_1}$) and N($\mathbf{x_2}$) is closed to the correlation coefficient between $\mathbf{x_1}$ and $\mathbf{x_2}$. Then some properties of the original Pearson's correlation coefficient carried over. We also note that the normal quantile transformation is necessary for the LA calculation. Therefore, we used the normal quantile transformed correlation coefficient for all the correlations between two variables in our analysis procedure.

For our real data analysis, first we preprocessed the expression profiles by using all the three preprocessing methods separately in HLM, UM, CAN/DF and MSK four data sets for three versions of gene expression profiles. Since the CAN/DF and MSK data sets were used for validation, the gene filter was only implemented in the training data sets, HLM and UM. There were 22,215 probe sets on Affymetrix U133A microarray. For each gene in the two training data sets, we evaluated the normal quantile transformed correlation coefficients between each two of the three different preprocessed expression profiles. Thus, there were six correlation coefficients evaluated for each gene. The genes that had as least one of the six correlation coefficients smaller than 0.78 were excluded.

In each data set, the ranks of the mean expressions of the remaining genes with a total of 22,215 probe sets were recorded. The histogram is given in figure 3.2. In the histogram, the proportion of the remaining genes with high rank is found larger than the proportion of the remaining genes with low rank.

3.2 Low gene expressions

Although microarray can be used to detect more than ten thousands of gene expression profiles simultaneously, the proportion of truly expressed genes might be no greater than half. In this gene filtering part, we evaluated the sample mean expression of each gene in the training set and filtered out the genes with small sample mean expressions. In practice, the genes with sample mean expressions smaller than 300 in the training data set were excluded.

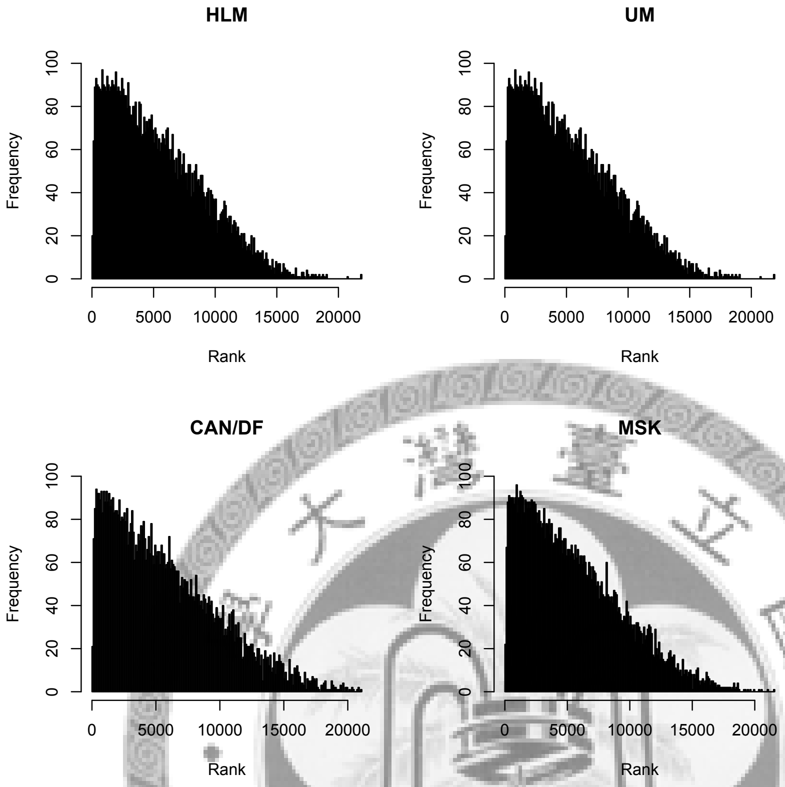


Figure 3.2: Histogram of the ranks of mean expressions of 22,215 probe sets in training data set after excluded inconsistent genes.

3.3 Small variation gene expressions

Another gene filtering criterion we used was the variation of each gene expression profile. Some ”housekeeping” genes expressed constantly high or low for basic reactions. The variations of these gene expression profiles were small and may not relate to the survival time of patients. However, selecting genes by normal quantile transformed correlation coefficients compared the correlations between gene expressions and survival time of patients at the same sale of variation. Some non-informated genes might be selected.

Therefore, we excluded the genes with small sample variation of expressions. In practice, the genes that had standard deviation smaller than 135 in the training data set were excluded.

After three-steps gene filtering, there were 6,252 genes remained in our analysis. The histogram of the ranks of mean expression profiles of the remaining genes with a total of 22,215 probe sets in training data set is given as figure 3.3.

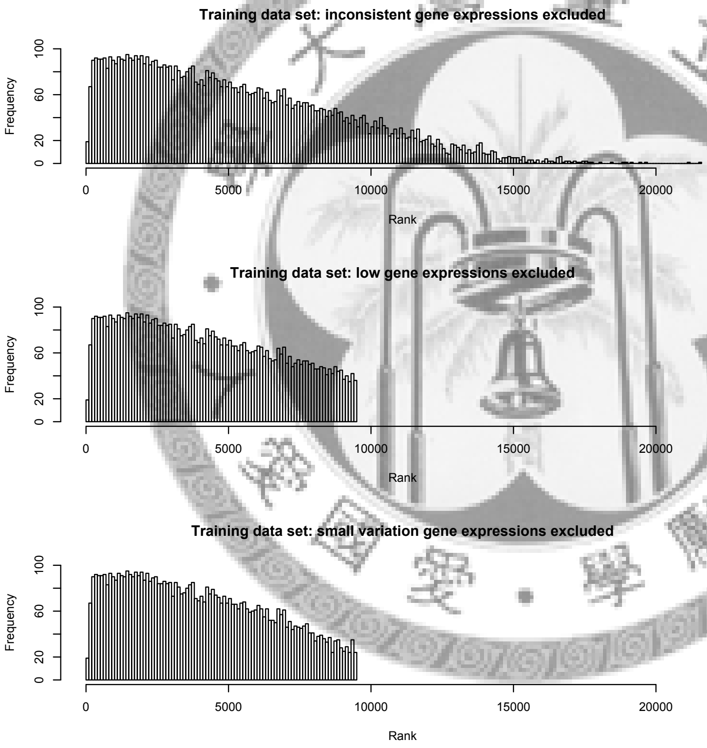


Figure 3.3: Histogram of the ranks of mean expression profiles of the remaining genes with a total of 22,215 probe sets in training data set.

Chapter 4

Gene selection

4.1 Correlation

In the previous chapter, we reduced the number of genes from 22,215 (G) to 6,252 (G^*) through the three-steps gene filter. However, the number is still much larger than the sample size 256 (n). The implementation of sliced inverse regression (SIR) requires the covariance matrix $\Sigma_{\mathbf{x}}$ to be non-singular, which can not hold in such a case. To solve this issue, we used both correlation and liquid association (LA) methods to select the important candidate genes. The genes with greatest correlations to the survival time Y° and gene pairs that had interaction related to the survival time were selected as candidate genes in this part. Due to the censoring of survival data, both correlation and liquid association methods could not be applied directly for the observed time $Y = \min\{Y^{\circ}, C\}$, where C is the censored time. A modified imputation method was proposed to impute the survival probability for the censored time, and then both correlation and liquid association methods could be implemented by plugging the imputed survival probability.

4.1.1 Imputation of survival time with right censoring

Here we used the normal quantile transformed correlation coefficient described in the previous chapter to measure the correlation between survival time $\mathbf{y}^\circ$ and each gene expression profile $\mathbf{x}_g$, where $g = 1, 2, ..., G^*$. Due to data censoring, we could not use the correlation coefficient between observed time $\mathbf{y}$ and gene expression $\mathbf{x}_g$. Let $\delta = \mathbf{1}_{\{Y^\circ \leq C\}}(Y^\circ, C)$ be the indicator that indicated the status of each patient. To reduce the effect caused by right censoring, an imputation $\hat{Y}^\circ$ for $\delta = 0$ was needed.

Suppose that Y° was the true survival time with survival function $S^\circ(y^\circ) = P(Y^\circ > y^\circ)$ and density function $f(y^\circ)$. In elementary statistics, we knew that the conditional mean, $E(Y^\circ \mid Y^\circ > y)$, minimized the 2-norm imputation error loss, $l_2(\hat{Y}^\circ) = E[(\hat{Y}^\circ - Y^\circ)^2 \mid Y^\circ > y]$. However, Wu *et al.* (2008) pointed out the limitation of estimating the conditional mean given that $Y^\circ > y$ by Kaplan-Meier estimate when the last observation was censored in practice. Therefore, we did not adopt the conditional mean. If the 1-norm imputation error loss $l_1(\hat{Y}^\circ) = E[|\hat{Y}^\circ - Y^\circ| \mid Y^\circ > y]$ was used, we could impute the censored data by the conditional median given that $Y^\circ > y$ and evaluate the normal quantile transformed correlation coefficient, $corr(N(\mathbf{x}_g), N(\hat{\mathbf{y}}^\circ))$, to estimate the correlation coefficient between survival time Y° and each gene expression profile X_g.

First we noted that the normal quantile transformation $N(\cdot)$ only depended of the ranks of variables. Therefore, it was invariant under any monotone transformation, that is $N((h(x_1), h(x_2), ..., h(x_n))') = N((x_1, x_2, ..., x_n)')$ for any monotone function h. Second, the conditional median given that

$Y^\circ > y$ satisfied that

$$S^\circ\big(median(Y^\circ \mid Y^\circ > y)\big) = \frac{1}{2}S(y).$$

Furthermore, the distribution function $F^\circ(\cdot) = 1 - S^\circ(\cdot)$ is a monotone function, so that we have

$$\begin{aligned} N(\hat{\mathbf{y}}^\circ) &= N\big((\hat{y}_1^\circ, \hat{y}_2^\circ, ..., \hat{y}_n^\circ)'\big) \\ &= N\Big(\big(F^\circ(\hat{y}_1^\circ), F^\circ(\hat{y}_2^\circ), ..., F^\circ(\hat{y}_n^\circ)\big)'\Big) \\ &= N\Big(\big(1 - S^\circ(\hat{y}_1^\circ), 1 - S^\circ(\hat{y}_2^\circ), ..., 1 - S^\circ(\hat{y}_n^\circ)\big)'\Big) \end{aligned}$$

and

$$S^\circ(\hat{y}_i^\circ) = \begin{cases} S^\circ(y_i), & \text{if } \delta_i = 1 \\ \frac{1}{2}S^\circ(y_i), & \text{if } \delta_i = 0. \end{cases}$$

Therefore, instead of estimating the conditional median, we could evaluate the $N(\hat{\mathbf{y}}^\circ)$ by plugging the estimated survival function $\hat{S}^\circ(\cdot)$. Wu *et al.* (2008) proposed a nonparametric imputation procedure by using the Kaplan-Meier estimation for the survival function. The procedure is summarized as the following steps:

Imputation - Kaplan-Meier based

1. Calculate $\hat{S}_i^\circ$ the Kaplan-Meier estimate of the survival probability;

2. Impute the survival probability by the predicted conditional median

$$\tilde{S}_i^\circ = \begin{cases} \hat{S}_i^\circ, & \text{if } \delta_i = 1 \\ \frac{1}{2}\hat{S}_i^\circ, & \text{if } \delta_i = 0; \end{cases}$$

3. Calculate the percentile $\hat{p}_i = 1 - \tilde{S}_i{}^\circ$;

4. Calculate the imputed $N(\hat{\mathbf{y}}^\circ)$ by performing the normal quantile transformation on $\hat{p}_i$.

The implementation of this nonparametric imputation procedure is easy since we only have to calculate the Kaplan-Meier estimate of the survival probability. However, an issue is how we improve the imputation if we have extra information. The TNM tumor stage is strongly related to the survival time of NSCLC patients. It is not suitable to impute the same survival time for different stage patients at the same censored time. It motivated us to modify the imputation procedure with this extra information.

Modified imputation - Cox proportional hazard model based

Previous studies indicated that the survival of NSCLC patients in different TNM tumor stage were significantly different and it motivated us to modify the imputation procedure. Since the original imputation procedure directly followed by the Kaplan-Meier estimate of survival probability, one nature idea was to modify the survival probability estimation by incorporating the TNM tumor stage Z. We assumed that the conditional survival function given $Z = z$ satisfied Cox proportional hazard model.

Cox proportional hazard model is one of the well-known regression survival models. It modeled that the hazard function given $Z = z$ is proportional to a baseline hazard function and the logarithm of the ratio is linearly dependent on the regressors,

$$\lambda^{\circ}(y^{\circ} \mid Z = z) = \lambda_0^{\circ}(y^{\circ})e^{\gamma z}.$$

Here we let Z be a four levels factor which indicated that the patient's TNM tumor stage is IA, IB, II or III/IV. Then, the relationship between conditional survival function given $Z = z$ and the baseline survival function can be

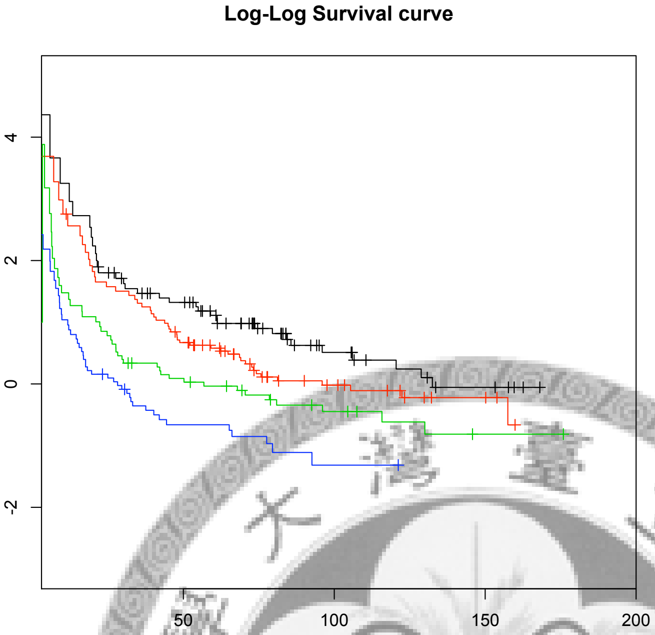


Figure 4.1: The log-log Kaplan-Meier curves of different stages for the training data set.

expressed as

$$S^{\circ}(Y^{\circ} \mid Z = z) = [S_0^{\circ}(Y^{\circ})]^{e^{\gamma z}}. \tag{4.1}$$

To check the assumption of Cox proportional hazard model, first we drew the log-log Kaplan-Meier curves of different stages for the training data set. From the equation (4.1) above, the log-log Kaplan-Meier curves should be parallel if the assumption held. Figure 4.1.1 showed that there was no strong evidence of non-parallelism for our data. Second, we drew the observed versus expected plot for the training data set and it also showed that there was no strong evidence to reject the assumption.

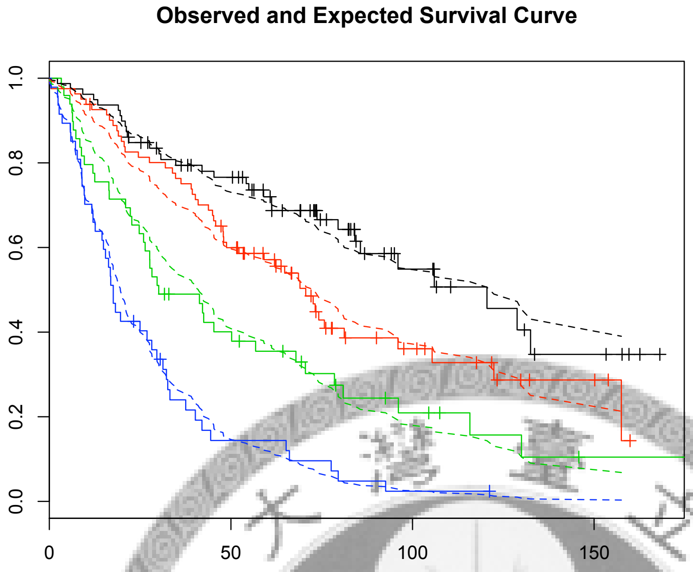


Figure 4.2: Observed Kaplan-Meier plot and Cox proportional hazard model. Cox coefficients of (Stage IB, II, III/IV)= (0.50, 1.05, 1.81). The hazard ratio of (Stage IB, II, III/IV)= (1.64, 2.85, 6.12).

Therefore, we assumed that $Y^\circ \mid Z = z$ satisfied the Cox proportional hazard model and imputed the censored time by $\hat{y}^\circ = median(Y^\circ \mid Y^\circ > y, Z = z)$. Then we had

$$\begin{aligned} N(\hat{\mathbf{y}}^\circ) &= N\big((\hat{y}_1^\circ, \hat{y}_2^\circ, ..., \hat{y}_n^\circ)'\big) \\ &= N\Big(\big(1 - S_0^\circ(\hat{y}_1^\circ), 1 - S_0^\circ(\hat{y}_2^\circ), ..., 1 - S_0^\circ(\hat{y}_n^\circ)\big)'\Big), \end{aligned}$$

where $S_0^\circ(\cdot)$ was the baseline survival function and it was also a monotone function. Furthermore, we had $S_0^\circ(\hat{y}_i^\circ) = S_0^\circ(y_i)$ if $\delta_i = 1$ and

$$\begin{aligned} S_0^\circ(\hat{y}_i^\circ) &= \Big(S^\circ(\hat{y}_i^\circ|Z)\Big)^{\frac{1}{\exp(\gamma Z)}} = \Big(\frac{1}{2}S^\circ(y_i|Z)\Big)^{\frac{1}{\exp(\gamma Z)}} \\ &= \Big(\frac{1}{2}S_0^\circ(y_i)^{\exp(\gamma Z)}\Big)^{\frac{1}{exp(\gamma Z)}} = (\frac{1}{2})^{\frac{1}{\exp(\gamma Z)}} S_0^\circ(y_i), \text{ if } \delta_i = 0. \end{aligned}$$

From the equations above, we observed that the modified imputation method gave different weights for the censored survival probability of the patients in different stages. In practice, the Cox coefficients γ could be estimated by finding the γ that maximized the partial likelihood. The baseline survival function could be estimated by the Nelson-Allen estimate or Breslow estimate. Then the original procedure could be implemented by replacing the survival probability estimation and the imputed weights. The modified imputation procedure is summarized as the following steps:

1. Estimate the Cox coefficients $\gamma' s$ for each TNM tumor stage and the baseline survival probability $\hat{S}^{\circ}_{0i}$;

2. Impute the survival probability by the predicted conditional median

$$\tilde{S}^{\circ}_{0i} = \begin{cases} \hat{S}^{\circ}_{0i}, & \text{if } \delta_i = 1 \\ (\frac{1}{2})^{\frac{1}{\exp \gamma z_i}} \hat{S}^{\circ}_{0i}, & \text{if } \delta_i = 0; \end{cases}$$

3. Calculate the percentile $\tilde{p}_i = 1 - \tilde{S}^{\circ}_{0i}$;

4. Calculate the imputed $N(\hat{\mathbf{y}}^{\circ})$ by performing the normal quantile transformation on $\tilde{p}_i$.

4.1.2 A simulation comparison between two imputation methods

To present the improvement of our modification, we did a simulation study. First we randomly generated 256 survival time samples from Cox proportional hazard model with a four levels factor regressor. The levels of the factor regressor were uniformly random generated. The baseline survival function was exponential distribution with rate parameter set to be 1 and the Cox coefficients γ's for each level were set to be $(0, 2, 4, 8)$. Another 256 censored time samples were randomly generated from exponential distribution with rate parameter 3. We set the minimum of the survival time samples and the censoring time samples to be the observed time samples. The average censoring rate was 0.5099.

To assess the performances of the imputation methods, the normal quantile transformed correlation coefficient between true survival time samples and the imputed values, $corr(N(y^{\circ}), N(\hat{y}^{\circ}))$, was used to measure the closeness. For 1,000 simulation runs, we implemented both imputation methods and recorded the correlation coefficients in each run. The average of the correlation coefficients of the Kaplan-Meier based imputation was 0.8684 and the average correlation coefficients of Modified imputation was 0.9096. Moreover, there were only three times that the Kaplan-Meier based imputation had correlation coefficient greater than the Modified imputation. We concluded that the modified imputation method had better performance when the Cox proportional hazard model assumption held. The results of our simulation were given in the following table.

Table 4.1: Simulation comparison between two imputation methods

Estimated Cox model coefficients	Average	S.D.
Cox model coefficient $\hat{\gamma}_2$ ($\gamma_2 = 2$)	2.1096	0.6260
Cox model coefficient $\hat{\gamma}_3$ ($\gamma_3 = 4$)	4.1355	0.6264
Cox model coefficient $\hat{\gamma}_4$ ($\gamma_4 = 8$)	8.2389	0.7948
Normal quantile transformed correlation coefficient between true survival time and imputed value	Average	S.D.
No imputation (Observed time)	0.7786	0.0312
KM based imputation	0.8683	0.0205
Modified imputation	0.9096	0.0146
Censor rate	0.5099	0.0308

4.1.3 Gene selection in training data by correlation

In this subsection, we presented the results of candidate genes selection by correlation method in the training data set (HLM+UM). For all the samples in the training data set, first we implemented the modified imputation procedure described in the previous subsection to impute the survival probability and performed the normal quantile transformation on the imputed survival probability and each gene expression profile. For each gene, we calculated the correlation coefficient r_g between normalized gene expression $N(\mathbf{x}_g)$ and the imputed $N(\hat{\mathbf{y}}^{\circ})$ and ranked the 6,252 correlation coefficients by the absolute values. The top five genes with the greatest absolute values of correlation coefficients were selected as our candidate genes. These candidate genes were the only five genes which had the absolute value of correlation coefficients greater than 0.25. The cutoff was determined based on controlling the ratio of the expected number of all true null hypotheses to the real observed number of correlation coefficients greater than the cutoff

$$c^* = min\Big\{c \mid \big[E[\sum_{g=1}^{G^*}\mathbf{1}(R_g > c)]/[\sum_{g=1}^{G^*}\mathbf{1}(r_g > c)]\big] \leq \alpha\Big\},$$

where $R_g = corr\big(N(\mathbf{x}_g), N(\hat{\mathbf{y}}^\circ)\big)$ under $X_g \perp Y^\circ$ for all g. A permutation procedure was proposed to select the cutoff. For any given cutoff, we estimated the expected number of correlation coefficients greater than the cutoff if survival time was irrelative to all the 6,252 genes by the average of 1,000 runs permutation. The ratio of the expected number to the real observed number of correlation coefficients greater than the cutoff was used to choose the cutoff. The procedure could be summarized as the following steps:

1. Calculate and rank the true absolute value of correlation coefficients $(r_{(1)}, r_{(2)}, ..., r_{(G^*)})'$ between the imputed value $N(\hat{\mathbf{y}}^\circ)$ and each gene expression $N(\mathbf{x}_g)$, where $g = 1, 2, ..., G^*$ and $G^* = 6,252$;
2. Permute the imputed $N(\hat{\mathbf{y}}^\circ)$ randomly as $N(\hat{\mathbf{y}}^\circ)^*$;
3. Calculate and the absolute value of correlation coefficients $(r_1^*, r_2^*, ..., r_{G^*}^*)'$ between $N(\hat{\mathbf{y}}^\circ)^*$ and each gene expression $N(\mathbf{x}_g)$, where $g = 1, 2, ..., G^*$;
4. Calculate the number of permuted values r_g^* greater than the each ranked true values $\mathbf{m} = (m_1, m_2, ..., m_{G^*})$, where $m_i = \sum_{g=1}^{G^*} \mathbf{1}_{\{r_g^* \geq r_i\}}(r_g^*)$, for $i = 1, 2, ..., G^*$;
5. Repeat step 2 to step 4 for 1,000 times and record the $\mathbf{m}_k$ for the k-th time, where $k = 1, 2, ..., 1000$;
6. Estimate the expected number of correlation coefficients greater each ranked true values by chance by the average $\hat{\mathbf{e}} = \bar{\mathbf{m}} = (\frac{1}{1000}\sum_{k=1}^{1000}(\mathbf{m}_k)_1, \frac{1}{1000}\sum_{k=1}^{1000}(\mathbf{m}_k)_2, ..., \frac{1}{1000}\sum_{k=1}^{1000}(\mathbf{m}_k)_{G^*})'$;
7. Calculate the ratios of expected number to the observed number and determin the cutoff to be the j-th place true absolute value of correlation

coefficients $r_{(j)}$, where $j = \max\{i \mid \hat{\mathbf{e}}_i/i \leq \alpha\}$, where α is specified by user.

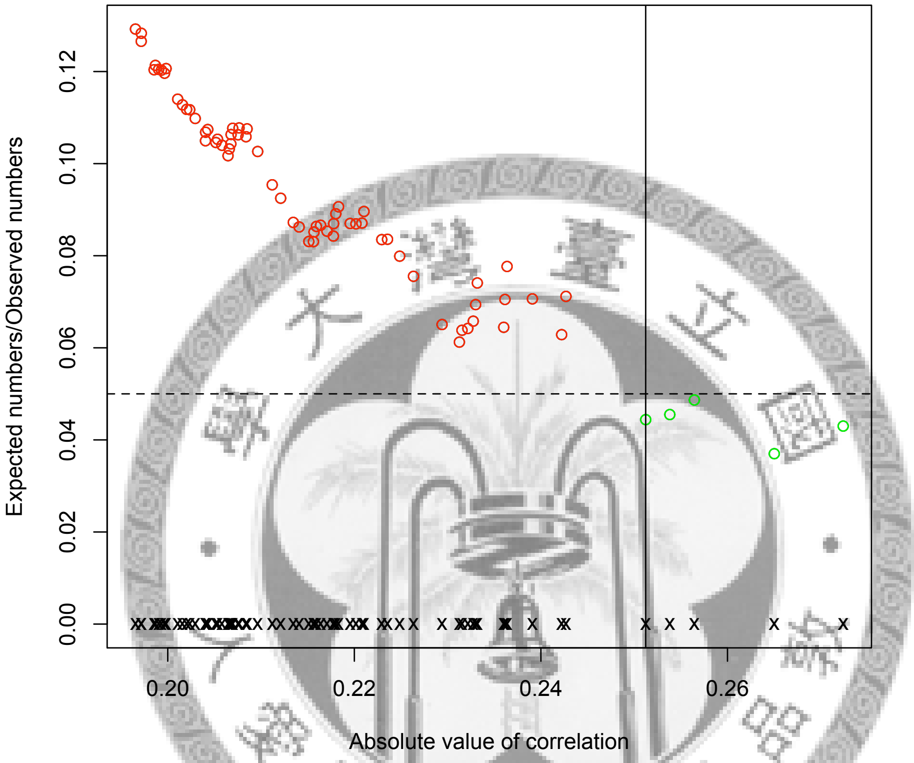


Figure 4.3: The scatter plot of ratios of expected number to the observed number versus cutoffs of the absolute value of correlation coefficient in the training data.

For a conservative criterion, we set α to be 0.05 and implemented this permutation procedure in the training data set to determine the cutoff to be 0.251. A scatter plot of the top 70 places of true absolute value of correlation coefficients and the corresponding ratios of expected number to the observed number is given as figure 4.3. Figure 4.3 showed that the first five ratios of

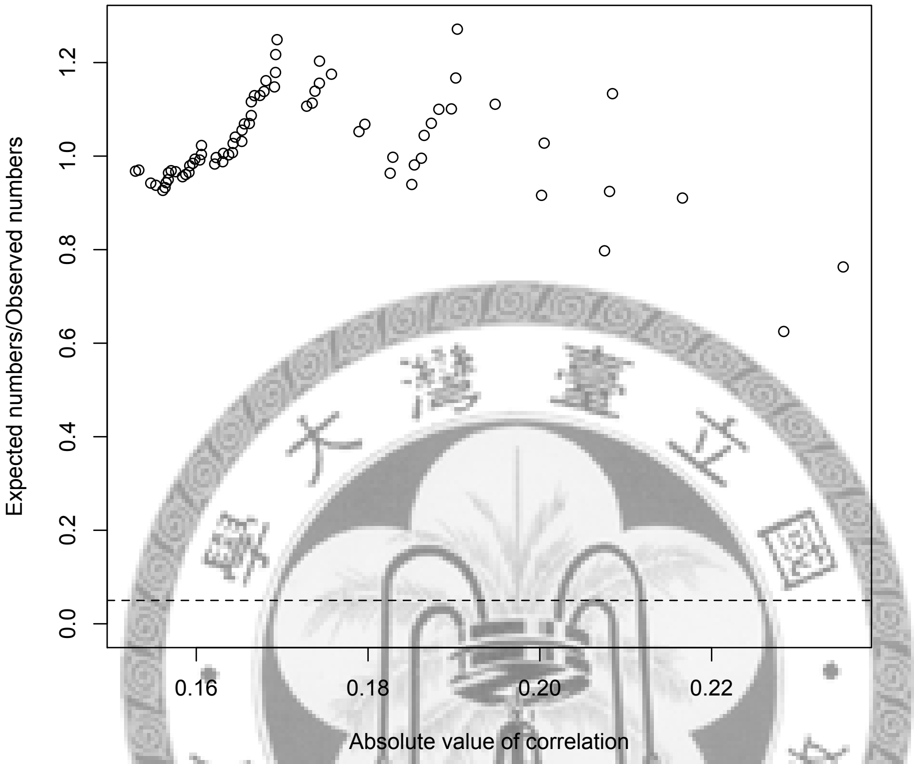


Figure 4.4: The scatter plot of ratios of expected number to the observed number versus cutoffs of the absolute value of correlation coefficient in one of the 1,000 permutations.

expected number to the observed number were smaller than the others. Furthermore, we could see that the ratio increased when the cutoff decreased. This tendency was expected, since we expected the rest correlation coefficients distributed similarly as randomly permuted correlation coefficients. To illustrate the difference between the real data and the permuted data, a scatter plot of ratios versus cutoffs for one of the 1,000 permutations is given as figure 4.4. In figure 4.4, we could see that the first ratio was quite large and the rest ratios oscillated around 1. Thus, we suggested that no genes should be selected in this situation. The candidate genes and the normal quantile transformed correlation coefficients with the imputed survival value were given in table 4.2.

Table 4.2: The candidate genes and the correlation coefficients of the candidate genes and the imputed survival time

Symbols	Full names	Correlation coefficients
TMEM66	transmembrane protein 66	0.272
CSRP1	cysteine and glycine-rich protein 1	0.265
BECN1	beclin 1, autophagy related	0.256
FOSL2	FOS-like antigen 2	-0.254
ERO1L	ERO1-like	-0.251

4.2 Liquid association

In the previous section, we selected the genes correlated with the imputed survival time as the candidate genes for subsequent analysis. Correlation could be viewed as a measure of linear dependency of two variables X and Y. However, some genes which had nonlinear association with survival time might not be detected by our first step selection. To reveal these relations, we applied the liquid association (LA) method from Li (2002) for the second step selection. The details of methodology and the implementation of liquid association method are given in this section.

4.2.1 Methodology of Liquid Association

Liquid association was originally proposed for studying the functionally-associated gene pairs [13]. The correlation between a functionally-associated gene pair (X, Y) is usually not found significantly, because the functional association may be varied from different cellular states, which are usually unknown. However, if there is an expression profile of a third gene Z with variation associated with the cellular state change, then the expression profile of gene Z can be used to reveal the patterns of functionally-associated in the gene pair (X, Y). If the gene Z is known, we may draw the scatter plot of profiles X and Y colored by profile Z to reveal the patterns by eyes. Since the gene Z is usually unknown, to screen more than ten thousands of scatter plots for whole genes searching is impractical. Therefore, LA score, a scoring system for the average rate of change of correlation between a pair (X, Y) with respect to profile Z, was introduced to searching the latent gene Z.

We assume that the correlation of a gene pair (X, Y) depends on the cellular states, for example, the correlation is positive at state 1 and negative at state 2. If there is a latent gene expression profile$Z$ highly expresses at state 1 and lowly expresses at state 2, we can expect that the increase of expression profile Z is associated with the increase of the correlation between X and Y. Then the pair (X, Y) is called a positive LA pair of Z. If the monotone relation holds, the average rate of change of correlation between pairs (X, Y) with respect to profile Z should be significantly greater than zero. Similarly, a pair (X, Y) is called a negative LA pair of Z, if the increase of expression Z is associated with the decrease of the correlation between X and Y. If the monotonic relation holds, the average rate of change of correlation between pairs (X, Y) with respect to profile Z should be significantly smaller than zero. Thus, the LA score is defined as the follow.

Definition 4.2.1. Suppose X, Y and Z are random variables with mean 0 and variance 1. The LA score of X and Y with respect to Z, denoted by $LA(X, Y \mid Z)$, is

$$LA(X, Y \mid Z) = E_Z[g'(Z)]$$

,where

$$g(Z) = E_{XY}[XY \mid Z].$$

If Z follows the normal distribution, the LA score can be evaluated by Stein's Lemma, such that

$$\begin{aligned} LA(X,Y|Z) &= E_Z[g'(Z)] \\ &= E_Z[Zg(Z)] \\ &= E_Z[ZE_{X,Y}[XY|Z]] \\ &= E_Z[E_{X,Y}[XYZ|Z]] \\ &= E_{X,Y,Z}[XYZ]. \end{aligned}$$

In practice, we can use the moment estimate for the sample version of LA score, where

$$\hat{LA}(X,Y \mid Z) = \hat{E}_{X,Y,Z}[XYZ] = \frac{1}{n}\sum_{i=1}^{n} X_iY_iZ_i.$$

The three-tuples with extreme absolute value of LA scores are expected to have liquid associations. Since the Stein's Lemma holds only if the variable Z follows the normal distribution. The normal quantile transformation is necessary for the implementation of liquid association method. We note that each variable is normal quantile transformed separately. Transforming the variables into multivariate normal distribution will cause the violence to the underlying pattern. There will be no significantly extreme LA scores found if the variables are multivariate normal distributed, since the correlations between each two variables are constants.

4.2.2 Implementation of Liquid Association

Wu *et al.* (2008) proposed a strategy for implementing LA method to find the candidate gene pairs associated with the survival time. They took the imputed survival time as the third variable to find gene pairs whose functionally-associated pattern may vary as the imputed survival changes.

The gene pairs with greatest absolute value of LA scores with respect to the imputed survival were selected as candidate genes. However, they also pointed out that due to the large number, $\frac{1}{2}G^*(G^*+1)$, of comparison of LA scores, the signals might be difficult to be detected by examining the individual LA pairs. They suggested an alternative strategy; to select the recurrent genes from a subset of the gene pairs with extreme LA scores with respect to the imputed survival time. The recurrent genes were called LA hub genes and selected as the candidate genes. Their LA hub genes selecting procedure could be summarized as the following steps:

1. Perform the normal quantile transformation for both gene expression profiles and imputed survival value;
2. Calculate and rank the LA scores, $LA(N(\mathbf{x}_i), N(\mathbf{x}_j) \mid N(\hat{\mathbf{y}}^\circ))$, of all possible gene pairs with respect to the imputed survival value;
3. Select the genes appeared at least k times in the top M positive and negative places, where the cutoff k and M need to be specified by user.

In their examples, they took M to be 50 and then selected the cutoff k to be 3 by their permutation result. They permuted the imputed survival time for 1,000 runs and calculated the average number of genes appeared at least k times in the first top M places. Then they used the average number to compare with the observed number. The selection of k depended on M only, which means that the recurrence was defined only by the ranks of all LA scores. However, we found that due to correlation structure of genes, the recurrence of the genes in the first M places was easy to pop out in the randomly permuted cases but the LA scores were related small. Therefore, we suggested selecting recurrent genes with significantly extreme LA score rather than selecting the recurrent gene in the first few places. We suggested a two-steps permutation procedure to replace the third step in the original procedure to select the LA hub genes. Our procedure could be summarized as the following steps:

1. Normal quantile transform both the gene expression profiles and imputed survival time;
2. Calculate the LA scores, $LA(N(X_i), N(X_j) \mid N(\hat{Y}^{\circ}))$, of all possible gene pairs with respect to the imputed survival time;
3. Applied the permutation procedure described in section 4.1.3 with a soft criterion α to decide the cutoff of LA scores;
4. Compare the number of k-times hub genes with the average number of genes appeared at least k times in LA pairs with LA scores greater than the cutoff in 1,000 permutation runs.

4.2.3 Gene selection in training data by LA

We implemented the LA hub genes selection procedure in the training data set. First we implemented the modified imputation. In the second steps, a total of 19,546,878 LA scores of all possible gene pairs with respected to the imputed value were computed. The scatter plot of ratios versus cutoffs in step 3 is given as figure 4.5. Based on the scatter plot of ratios versus cutoffs we decided to choose the cutoff to be 0.326. Then, we noticed that the gene SRP54 appeared 4 times in a total of 11 pairs with LA score greater than the cutoff. Then we permuted the imputed survival time for 1,000 runs and computed the average number of genes appeared in gene pairs with LA score greater than 0.326 at least 4 times. The average number was only 0.032

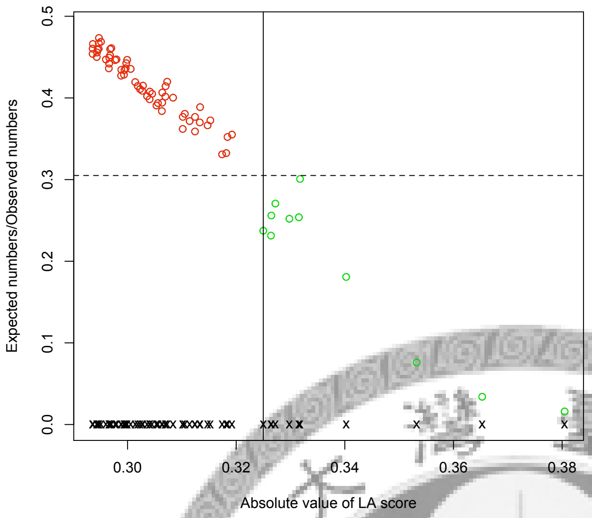


Figure 4.5: The scatter plot of ratios of expected number to the observed number versus cutoffs of the absolute value of LA scores in the training data.

in 1,000 permutation runs. Therefore, we selected these LA pairs as the candidate genes for the subsequent analysis. The LA hub gene SRP54 with the 4 paired genes and the LA scores with respect to the imputed survival value are given in table 4.3.

Table 4.3: The LA hub gene SRP54 and its LA paired genes

Symbols	Full names	LA scores
SRP54	signal recognition particle 54kDa	-
SART3	squamous cell carcinoma antigen recognized by T cells 3	0.3810
NR2C1	nuclear receptor subfamily 2, group C, member 1	0.3659
CROP	cisplatin resistance-associated overexpressed protein	0.3269
PAWR	PRKC, apoptosis, WT1, regulator	0.3268

Chapter 5

Signature construction

5.1 Methodology of modified sliced inverse regression for censored data

In the gene signature construction part, first we applied the modified sliced inverse regression for censored data by Li *et al.* (1999) to reduce dimensions. After finding out the effective dimension reduction (e.d.r.) space, we could project the gene expression profiles on it and fit further survival model if necessary.

Sliced inverse regression by Li (1991) [15] was originally introduced for dimension reduction. Assuming the p-dimension regressor $\mathbf{X}$ and the response Y° satisfied the dimension reduction model

$$Y^\circ = g(\beta_1'\mathbf{X}, \beta_2'\mathbf{X}, ..., \beta_k'\mathbf{X}, \epsilon) \tag{5.1}$$

and the linear design condition; for any b in $\mathbf{R}^p$

$$E(b'\mathbf{X}|\beta_1'\mathbf{X}, \beta_2'\mathbf{X}, ..., \beta_k'\mathbf{X}) = c_0 + c_1\beta_1'\mathbf{X}, ..., c_k\beta_k'\mathbf{X}, \tag{5.2}$$

for some constants $c_1, c_2, ..., c_k$, the effective dimension reduction (e.d.r.) space, $\mathcal{B} = span(\beta_1, \beta_2, ..., \beta_k)$, could be estimated by the eigenvalue decomposition of $\Sigma_{E[\mathbf{X}|Y^\circ]}$ with respect to $\Sigma_\mathbf{X}$, where $\Sigma_{E[\mathbf{X}|Y^\circ]} = cov(E[\mathbf{X} \mid Y^\circ])$ and $\Sigma_\mathbf{X} = cov(\mathbf{X})$. The function g and the distribution of ϵ were not need to be specified for estimating the e.d.r. space. The key observation described as Theorem 3.1 in Li (1991) [15] was that under conditions (5.1) and (5.2), the centered inverse regression curve $E[\mathbf{X} \mid Y^\circ] - E[\mathbf{X}]$ was contained in the linear subspace spanned by $\beta_1\Sigma_\mathbf{X}, \beta_2\Sigma_\mathbf{X}, ..., \beta_k\Sigma_\mathbf{X}$. Therefore, we could estimate the e.d.r. space by estimating the inverse regression curve. To determine how many e.d.r. directions should be select, Li (1991) also proposed a large sample chi-squared test for testing the significance of the estimated e.d.r. directions which is called the SIR directions. In practice, the implementation of sliced inverse regression method is summarized as the following steps:

1. Sort the paired data $(\mathbf{x}, y^\circ)$ by y° and divide it into H slices (with similar proportions);
2. Compute the sample mean within each slice $\bar{\mathbf{x}}_h = \frac{1}{n\hat{p}_h}\sum_{y_i \in I_h} \mathbf{x}_i$, where $\hat{p}_h = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{I_h}(y_i^\circ)$;
3. Compute the sample covariance matrix $\hat{\Sigma}_X = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ and the between slices sample covariance matrix $\hat{\Sigma}_{E[X|Y^\circ]} = \sum_{h=1}^{H} \hat{p}_h(\mathbf{x}_h - \bar{\mathbf{x}})(\mathbf{x}_h - \bar{\mathbf{x}})'$;
4. Conduct a eigenvalue decomposition of $\hat{\Sigma}_{E[X|Y^\circ]}$ with respect to $\hat{\Sigma}_X$;
5. Applied the large sample chi-squared test to select the significant leading eigenvectors to be the SIR directions.

In survival data analysis, due to the data censoring, applying the original sliced inverse regression by directly slicing observed time Y may cause the estimation bias. Li *et al.* (1999) [16] studied the effects to the original sliced inverse regression caused by two data censoring conditions. Under the independent censoring condition

$$C \text{ is independent of } \mathbf{X} \text{ and } Y^\circ,$$

we can show that independent censoring did not affect the sliced inverse regression by directly slicing observed time Y. Under the conditional independent censoring condition, a more general condition,

$$\text{Conditional on } \mathbf{X}, \text{ C is independent of } Y^\circ,$$

directly slicing observed time Y does cause the estimation bias. Li *et al.* (1999) modified the original sliced inverse regression for this situation. As described above, an important step in sliced inverse regression is estimating the inverse regression curve $E[X \mid Y^\circ]$. For $Y^\circ \in [y_i^\circ, y_{i+1}^\circ)$ the inverse regression curve can be expressed by

$$\begin{aligned} E[\mathbf{X} \mid Y^\circ \in [y_i^\circ, y_{i+1}^\circ)] &= \frac{E[\mathbf{X}\mathbf{1}_{[y_i^\circ, y_{i+1}^\circ)}(Y^\circ)]}{E[\mathbf{1}_{[y_i^\circ, y_{i+1}^\circ)}(Y^\circ)]} \\ &= \frac{E[\mathbf{X}\mathbf{1}(Y^\circ \geq y_i^\circ)] - E[\mathbf{X}\mathbf{1}_(Y^\circ \geq y_{i+1}^\circ)]}{E[\mathbf{1}(Y^\circ \geq y_i^\circ)] - E[\mathbf{1}(Y^\circ \geq y_{i+1}^\circ)]}, \end{aligned}$$

where $0 = y_1^\circ < y_2^\circ < ... < y_H^\circ < y_{H+1}^\circ = \infty$ is a partition of true survival time. One can observe that

$$\begin{aligned} & E[\mathbf{X}\mathbf{1}(Y^\circ \geq y_i^\circ)] \\ = \;& E[\mathbf{X}\mathbf{1}(Y \geq y_i^\circ)] + E[\mathbf{X}\mathbf{1}(Y < y_i^\circ, \delta = 0)\mathbf{1}(Y^\circ \geq y_i^\circ)] \\ = \;& E[\mathbf{X}\mathbf{1}(Y \geq y_i^\circ)] + E[\mathbf{X}\mathbf{1}(Y < y_i^\circ, \delta = 0)E[\mathbf{1}(Y^\circ \geq y_i^\circ) \mid Y, \delta, \mathbf{X}]] \\ = \;& E[\mathbf{X}\mathbf{1}(Y \geq y_i^\circ)] + E[\mathbf{X}\mathbf{1}(Y < y_i^\circ, \delta = 0)E[\mathbf{1}(Y^\circ \geq y_i^\circ) \mid C, Y^\circ > C, \mathbf{X}]] \\ = \;& E[\mathbf{X}\mathbf{1}(Y \geq y_i^\circ)] + E\Big[\mathbf{X}\mathbf{1}(Y < y_i^\circ, \delta = 0)\frac{E[\mathbf{1}(Y^\circ \geq y_i^\circ) \mid \mathbf{X}]}{E[\mathbf{1}(Y^\circ \geq Y) \mid \mathbf{X}]}\Big], \end{aligned}$$

where the last equation holds under the condition (5.4). A weighted function is defined by $\omega(t, t', \mathbf{X}) = \frac{E[\mathbf{1}(Y^\circ \geq t)|\mathbf{X}]}{E[\mathbf{1}(Y^\circ \geq t')|\mathbf{X}]} = \frac{S^\circ(t|\mathbf{X})}{S^\circ(t'|\mathbf{X})}$. Under the conditional independent assumption, we can show that $\frac{S^\circ(t|\mathbf{X})}{S^\circ(t'|\mathbf{X})} = \frac{S(t|\mathbf{X})}{S(t'|\mathbf{X})}$ where S° is the survival function of failure time and S is the survival function of observed time. Then we have,

$$E[\mathbf{X}\mathbf{1}(Y^\circ \geq y_i^\circ)] = E[\mathbf{X}\mathbf{1}(Y \geq y_i^\circ)] + E\Big[\mathbf{X}\mathbf{1}(Y < y_i^\circ, \delta = 0)\omega(y^\circ, Y, \mathbf{X})\Big],$$

and

$$E[\mathbf{1}(Y^\circ \geq y_i^\circ)] = E[\mathbf{1}(Y \geq y_i^\circ)] + E\Big[\mathbf{1}(Y < y_i^\circ, \delta = 0)\omega(y^\circ, Y, \mathbf{X})\Big]$$

with a similar argument. Therefore, the inverse regression curve can be estimated by replacing the expectations by the first sample moments and plugging the weighted function $\omega(\cdot, \cdot, \cdot)$ by its kernel estimation $\hat{\omega}(\cdot, \cdot, \cdot)$. The proof of the consistency and the root-n rate convergence under some regularity conditions were given in Li *et al.* (1999) [16].

Since the kernel estimation only performed well in the low-dimension case, they also proposed an initial dimension reduction step called *double-slice* before applying the modified sliced inverse regression for censored data. We assumed that the censor time C also satisfied the dimension reduction model

$$C = h(\theta_1'\mathbf{X}, \theta_2'\mathbf{X}, ..., \theta_l'\mathbf{X}, \epsilon').$$

By applying the original sliced inverse regression, the space spanned by β's and θ's, called the joint e.d.r. space, can be estimated by slicing the observed time Y for $\delta = 1$ and 0 separately. Then we can replace $\mathbf{X}$ by its projection in the estimated joint e.d.r. space for a low-dimension kernel estimation of weight function $\hat{\omega}(\cdot, \cdot, \cdot)$. The modified sliced inverse regression procedure can be summarized as the following steps:

1. Double-slice the survival time and censoring time and apply the original sliced inverse regression;
2. Applied the large sample chi-squared test to select the first few significant joint SIR directions;
3. Project the regressors into the space spanned by the joint SIR directions to estimate the conditional survival function and the weight function by kernel estimation;
4. Compute the estimated conditional expectation in each slice by plugging the estimated weight function and the first sample moments;
5. Compute the estimated between slices covariant matrix $\hat{\Sigma}_{E[\mathbf{X}|Y^{\circ}]}$ and the sample covariant matrix $\hat{\Sigma}_{\mathbf{X}}$;
6. Conduct a eigenvalue decomposition of $\hat{\Sigma}_{E[\mathbf{X}|Y^{\circ}]}$ with respect to $\hat{\Sigma}_{\mathbf{X}}$;
7. Applied the large sample chi-squared test to select the significant leading eigenvectors be the survival time SIR directions.

5.2 Signature construction in the training data set

After the gene filter and the gene selection two steps, 10 candidate genes were selected to construct the gene signature, which included 5 genes X_1, X_2, ..., X_5 selected by correlation method, 1 hub gene X_6 and 4 genes $X_7, X_8, ...,$ X_{10} paired with the hub gene selected by liquid association method. Wu *et al.* (2008) suggested applying the modified sliced inverse regression on the genes selected by correlation method and the LA hub genes. However, we found that in their data example, the four genes with the greatest weights (absolute value) were selected by correlation method and the weights of the LA hub genes were relative small. It might not be suitable to use only the LA hub genes without their paired genes, since the change of survival time was related to the functionally-associated pattern based on the LA methodology. Therefore, we thought that the genes paired with the LA hub gene were not negligible for survival prediction.

Since correlation coefficient measured the linear dependency of two variables, the dimension reduction model assumption that survival time Y° depended on the gene expression profiles $(X_1, X_2, ..., X_5)'$ only trough its linear combinations was reasonable. Nevertheless, this assumption might not be suitable when there were both correlation genes and LA pair genes, due to the nonlinear conception of liquid association. Thus, to incorporate the LA pairs into the dimension reduction model, we added the interaction terms of LA pairs as regressors. However, we did not suggest adding all the interaction pairs. Although the significance of LA hub gene was showed in the previous chapter, the genes paired with it might be selected by chance, due to the correlated structure of thousands of genes. Here we presented how this happened with a simple simulation. First we generated variables $(X_1, X_2, ..., X_5)'$ from multivariate normal distribution with mean $\mathbf{0}$, variance 1 and equal correlation 0.2 for each two variables. We independently generated another cluster of genes $(Z_1, Z_2, ..., Z_{20})'$ from multivariate normal distribution with mean $\mathbf{0}$, variance 1 and equal correlation 0.7 for each two variables. The response variable Y° was generated by $Y^\circ = \exp(0.5X_1 + 0.5X_2 + 0.5X_3 + 0.5X_4 + 0.5X_5Z_1 + (0.5)^2\epsilon)$, where ϵ was generated from standard normal distribution independent to X's and Z's. 200 independent variable $W_1, W_2, ..., W_{200}$ were generated from multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $(\Sigma_W)_{ij} = 0.9^{|i-j|}$. After 1,000 simulation runs, the average number of X_5 appeared in the first 10 LA pairs was 6.52 and there were 873 times X_5 appeared more than one time in the first 10 LA pairs. It showed that several paired genes might be found by chance even there was only one true paired gene. Therefore, to be conservative, we did not incorporate all the LA pairs into the dimension reduction model.

Since our final goal was to derive a gene signature to predict the survival for all stage patients and stage I patients, especially the stage I patients. Our strategy was to incorporate the LA pair that improved the prediction power of derived signature most. For the derived gene signature by modified sliced inverse regression, the prediction power of it was presented in two ways. First we used the gene signature as a continuous risk score $r = \beta'\mathbf{x}$, where β was the significant SIR direction, to fit the Cox proportional hazard model, $\lambda^\circ(y^\circ \mid r) = \lambda_0^\circ(y^\circ)e^{\gamma r}$, and calculated the p-value for testing the null hypothesis that hazard ratio e^γ was equal to 1. To present the prediction power, we calculated the concordance probability estimate (CPE) for the probability that survival outcome agreed with the signature $P(Y_1^\circ > Y_2^\circ \mid \gamma r_1 \leq \gamma r_2)$. Gönen and Heller (2005) proposed that under Cox proportional hazard model, the concordance probability can be expressed by

$$\begin{aligned} P(Y_1^\circ > Y_2^\circ \mid \gamma r_1 \leq \gamma r_2) &= \frac{P(Y_1^\circ > Y_2^\circ, \gamma r_1 \leq \gamma r_2)}{P(\gamma r_1 \leq \gamma r_2)} \\ &= \frac{\int\int_{\gamma r_1 \leq \gamma r_2} P\big(Y^\circ(\gamma r_1) > Y^\circ(\gamma r_2)\big) dF(\gamma r_1) dF(\gamma r_2)}{\int\int_{\gamma r_1 \leq \gamma r_2} dF(\gamma r_1) dF(\gamma r_2)} \\ &= \frac{\int\int_{\gamma r_1 \leq \gamma r_2} [1 + \exp(\gamma r_2 - \gamma r_1)]^{-1} dF(\gamma r_1) dF(\gamma r_2)}{\int\int_{\gamma r_1 \leq \gamma r_2} dF(\gamma r_1) dF(\gamma r_2)}, \end{aligned}$$

where the last equation was followed by the proportional hazard assumption.

Then the concordance probability could be estimated by

$$CPE(\hat{\gamma}) = \frac{2}{n(n-1)} \sum_{i<j} \Big[\frac{\mathbf{1}_{\{\hat{\gamma} r_i \le \hat{\gamma} r_j\}}(r_i, r_j)}{1 + \exp(\hat{\gamma} r_j - \gamma r_i)} + \frac{\mathbf{1}_{\{\hat{\gamma} r_j \le \hat{\gamma} r_i\}}(r_i, r_j)}{1 + \exp(\hat{\gamma} r_i - \hat{\gamma} r_j)} \Big].$$

We noted that the concordance probability estimate was in the range from 0.5 to 1. A CPE close to 1 indicated the good prediction power of the signature, and a CPE close to 0.5 indicated the poor prediction power of the signature. Second we used the signature to separate the patients into two groups; high risk and low risk, by cutting at the median of the signature. Then we used the category classifier to fit the Cox proportional hazard model and evaluated the hazard ratio, the corresponding p-value and the concordance probability estimate. For the survival prediction of stage I patients, we used the same genes and the same coefficients β estimated from all stage samples to construct the signature for stage I patients. The two different risk groups were separated by cutting at the median of signature in samples of stage I patients only.

In practice, first we started from only the five genes selected by correlation method. We noted that we did not perform normal quantile transformation on any variables in this part, since it is somewhat unrealistic in the test set. The expression profiles preprocessed by the MAS 5.0 Statistical algorithm were used as our raw data. Then we took log-2 transformation and centered each gene expression profile at its sample mean. For the five genes selected by correlation method, we implemented the modified sliced inverse regression method and selected the only significant (p-value < 0.05) SIR direction by the large sample chi-squared test. We projected the expression profiles on the only SIR direction as our final gene signature. The Kaplan-Meier survival functions for two different risk groups separated by our signature, the corresponding p-values and the concordance probability estimate are given

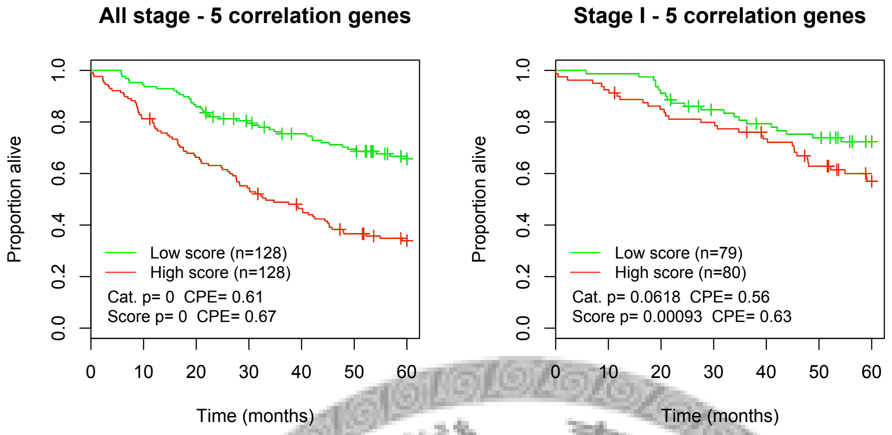


Figure 5.1: Kaplan-Meier survival curves for all stage and stage I samples in training data set separated by gene signature constructed by only five correlation genes.

as figure 5.1.

Figure 5.1 showed that the signature constructed by only five correlation genes had significant prediction power for all stage patients but not for the stage I patients only. However, since survival prediction for early stage patients is a more important issue, we wanted to incorporate LA gene pair for improving the prediction power for sample of stage I patients only. Each LA pair was incorporated to the dimension reduction model by adding their interaction term and the main effect terms. Specifically, we used $\mathbf{X} = (X_1, X_2, ..., X_5, X_6, X_i, X_6X_i)'$ as the regressors in the dimension reduction model, where $i = 7, ..., 10$. The interaction term was added for the nonlinear association of the LA pair with respect to the survival time, and the main effect terms were added for adjusting the miss centered issue for the interaction term. We applied the modified sliced inverse regression for each

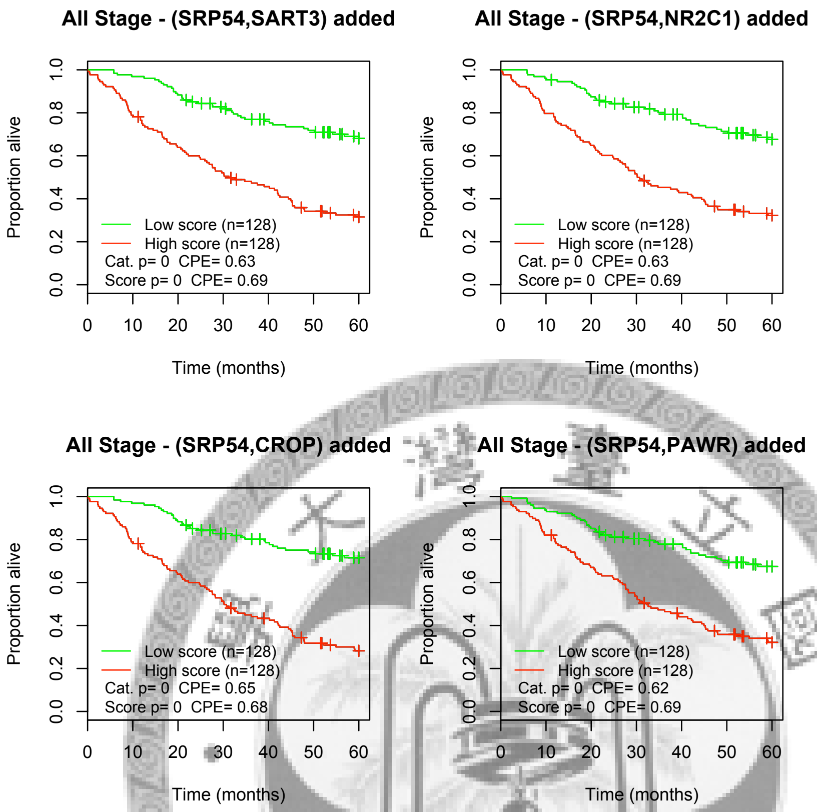
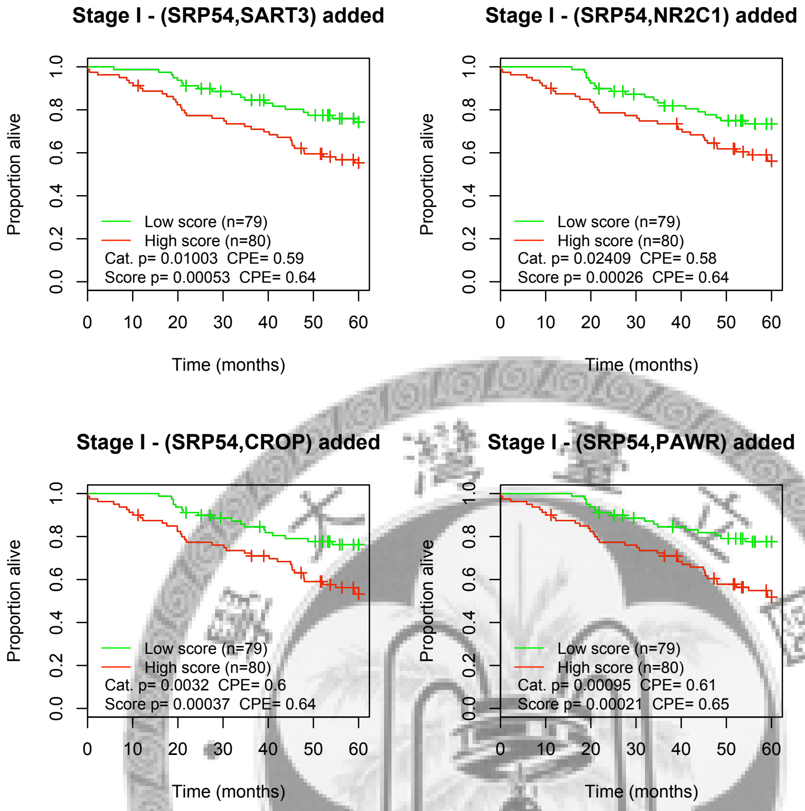


Figure 5.2: Kaplan-Meier survival curves for all stage samples in training data separated by gene signature constructed by five correlation genes and one LA pair.

regressor **X** to find the SIR direction. For each case, there was exact one significant SIR direction selected by the large sample chi-suared test. The results were given as figure 5.2 and 5.3.

Figure 5.3 showed that incorporating the LA pair did improve the prediction power for stage I patient only. We chose the best-performing gene signature constructed by five correlation genes and the LA pair (SRP54, PAWR) to be our final gene signature. To combine our signature and the


Stage I - (SRP54,SART3) added

Stage I - (SRP54,NR2C1) added

Stage I - (SRP54,CROP) added

Stage I - (SRP54,PAWR) added

Figure 5.3: Kaplan-Meier survival curves for stage I samples in training data separated by gene signature constructed by five correlation genes and one LA pair.

clinical covariates (TNM stage, sex ,age) for survival prediction, we fitted them with the multivariate Cox proportional hazard model, where the TNM tumor stage was coded as a four levels factor as in Chapter 3. After a stepwise selection, our gene signature, age and TNM stage III were still significant in the multivariate Cox proportional hazard model. The estimated sliced inverse regression direction was given in table 5.1, and the details of hazard ratio for univariate and multivariate Cox proportional hazard model were given in table 5.2.

The genes with negative coefficients are called protect genes, because the increase of its expression is associated with the decrease of hazard ratio. On the other hand, the genes with positive coefficients are called risk genes because the increase of its expression is associated with the increase of hazard ratio. Table 5.1 showed that these coefficients agreed with our results in gene selection part. All the protect genes had positive correlation with the imputed survival time and all the risk genes had the negative correlation with the imputed survival time. The coefficient of the interaction term of the LA pair was negative, which also agreed with its positive LA score. The absolute values of the coefficients and the standard deviations also presented the strength of the genes affecting our signature. Table 5.1, showed that the products of the coefficient and standard deviation of the regressors were closed except the main effect terms of the LA pair. Thus, all these genes gave important effects for our gene signature. Table 5.2 showed the significant of our gene signature in the Cox proportional hazard model. Furthermore, the p-values of the multivariate Cox proportional hazard model showed that our gene signature was still significant even we incorporated the TNM tumor stage and age.

Table 5.1: The estimated coefficients of SIR direction

Variable	SIR dir. coefficient	S.D.	SIR dir. coefficient*S.D.
TMEM66	-0.6457 (Protect)	0.3862	-0.2494
CSRP1	-0.4606 (Protect)	0.4738	-0.2182
BECN1	-1.1296 (Protect)	0.3671	-0.4147
FOSL2	0.2288 (Risk)	0.8292	0.1897
ERO1L	0.3333 (Risk)	0.9534	0.3178
(SRP54)	0.0253 (-)	0.5447	0.0138
(PAWR)	0.1045 (-)	0.6669	0.0697
SRP54*PAWR	-0.7517 (Protect)	0.4647	-0.3493
p-value	0.032		

Table 5.2: Hazard ratio with the corresponding 95% confident interval, p-value and the CPE of our gene signature

UM+MICH - All stage	Hazard ratio	95% C.I.	p-value	CPE
Risk score	2.22	(1.77, 2.79)	1.50e-13	0.688
Categorical	2.86	(1.96, 4.18)	1.20e-08	0.621
UM+MICH - Stage I	Hazard ratio	95% C.I.	p-value	CPE
Risk score	1.83	(1.31, 2.57)	0.0002	0.648
Categorical	2.55	(1.43, 4.52)	0.0009	0.610

Table 5.3: Hazard ratios with the corresponding 95% confident intervals and p-values of our gene signature and clinical covariates

Multivariate	Hazard ratio	95% C.I.	p-value
Risk score	1.85	(1.48,2.32)	6.98e-08
age	1.02	(1.01,1.04)	9.66e-03
Stage IB	1.28	(0.73,2.25)	3.84e-01
Stage II	2.62	(1.47,4.69)	1.14e-03
Stage III	4.72	(2.68,8.34)	8.41e-08
Multivariate	Hazard ratio	95% C.I.	p-value
Categorical	2.24	(1.51, 3.31)	5.60e-05
age	1.03	(1.01, 1.05)	7.38e-03
Stage IB	1.37	(0.78, 2.41)	2.74e-01
Stage II	2.79	(1.55, 5.02)	5.94e-03
Stage III	5.33	(3.04, 9.36)	5.57e-09

Chapter 6

Signature validation

6.1 Validation procedure

To test the predication power of our gene signature derived from the training data, we reconstructed our gene signature in two independent testing data sets, CAN/DF and MSK. First we applied the MAS 5.0 Statistical algorithm to get the raw data in the test sets. We chose the same probe sets selected from the training data and took log-2 transformation as in the training data. Second, we centered the testing data set at the sample means in the training data. Then, we used the same coefficients derived from the training data to combine the expression profiles into one gene signature as a risk score. We also separated the patients into two groups; high risk and low risk by cutting at the median risk score in the testing set. To present the prediction power, both the continuous risk score and the categorical classifier were used to fit the Cox model. The hazard ratios with the corresponding p-values and the CPE were evaluated. For the samples of stage I patients only, we applied the same procedure with the same probe sets, linear combination coefficients to get the same risk score. The median of the risk score of the stage I patients in each testing set was used to be the cutoff as a categorical classifier.

6.2 Cross platform adjustment

Another independent cohort data from Duke University was used as an external validation data set. There were two challenges for applying our gene signature in this data set. First, there were different tumor cell types, squamous cell carcinomas and adenocarcinomas, in this data set. Second, the platform, HU133plus2 array, used in this data set was different from the training data. Therefore, to apply our signature in this data set, a cross platform adjustment was needed. The HU133plus2 array contained all the probe sets in U133A array but the additional 33,429 probe sets might cause the scale change in the preprocessing procedure. To avoid this systematic bias, first we applied the MAS 5.0 Statistical algorithm to get the raw data. Then we selected the probe set contained both in U133A and HU133plus2 and adjusted the expression profiles of these matched probe sets by implementing the trim mean step again. After this adjustment, we implemented the same validation procedure in this data set as in the other two validation sets.

6.3 Validation results

All the testing results were summarized in table 6.1. To compare with our signature, the testing results of using TNM tumor stage as a classifier and the best-performing method (method A) from Shedden *et al.* (2008) were also included in table 6.1. Here the TNM tumor stage classifier was simply separating the all stage patients into two groups; stage I and late stage. The stage I patients was separated by stage IA or IB. Method A gave a continuous risk score constructed by using the average expression profiles of 100 clusters to fit ridged Cox proportional hazard model. We noted that the results of method A were analyzed from the expression profiles preprocessed by implementing dChip algorithm for entire training and testing data set. In CAN/DF and MSK two testing sets, all of these four methods performed good for the all stage patients prediction. Only our gene signature, both continuous and categorical types, had all hazard ratios significantly greater than 1 for all stage and stage I patients in both data sets. The hazard ratio of method A was not significantly greater than 1 in CAN/DF data. The hazard ratios of TNM tumor stage IA and IB were not significantly greater than 1 in both testing sets. Furthermore, the hazard ratio of TNM tumor stage IA and IB was smaller than 1 in CAN/DF data set. This result did not suggest using the IA and IB as a classifier for stage I patients. In the external validation cohort data from Duke University, the hazard ratios of our signature were also significantly great than 1 for the patients of both the all stage and early stage samples. We noticed that our signature performed better for stage I than all the patients, in this data set. We found that there were five stage IV and fifteen stage IIIB patients in this data set. However, there were only 11 stage IIIB patients and no stage IV patients in training data set. Furthermore, there were five of these late stage patients died within half years. This might be a reason that all patients prediction performed not good as the TNM tumor stage.

The Kaplan-Meier curves were given to illustrate the difference of survival functions between high risk and low risk groups. The Kaplan-Meier curves for classifying patients by TNM tumor stages were also given. Due to the small sample size, highly censored rate and the relatively homogeneous samples, classifying CAN/DF into different risk groups was much harder than other data set. The Kaplan-Meier curve showed that our signature had reasonable good prediction power in such a data set. The significant p-value for the Duke validation set showed that our gene signature, derived from adenocarcinomas patients only, had potential to predict patients with different tumor types. We concluded that our gene signature had good prediction power for all stage or early stage non-small cell lung cancer patients.

Table 6.1: Validation results in CAN/DF, MSK and Duke data sets

CAN/DF All stage	Hazard ratio	95% C.I.	p-value	CPE
Risk score	1.65	(1.17, 2.31)	0.002	0.662
Categorical	3.96	(1.68, 9.34)	0.001	0.651
TNM stage	3.25	(1.54, 6.84)	0.002	0.616
Method A	0.57	(1.20, 2.60)	0.003	0.623
CAN/DF stage I	Hazard ratio	95% C.I.	p-value	CPE
Risk score	1.59	(1.01, 2.51)	0.036	0.666
Categorical	3.78	(1.04,13.74)	0.027	0.648
TNM stage	0.55	(0.17, 1.80)	0.347	0.546
Method A	1.29	(0.84, 1.98)	0.243	0.574
MSK All stage	Hazard ratio	95% C.I.	p-value	CPE
Risk score	1.68	(1.13, 2.51)	0.012	0.614
Categorical	2.65	(1.29, 5.45)	0.006	0.614
TNM stage	3.87	(1.91, 7.85)	0.000	0.642
Method A	1.83	(1.24, 2.70)	0.002	0.627
MSK stage I	Hazard ratio	95% C.I.	p-value	CPE
Risk score	2.23	(1.14, 4.35)	0.023	0.654
Categorical	11.89	(1.53, 92.16)	0.001	0.715
TNM stage	2.60	(0.70, 9.63)	0.127	0.611
Method A	2.10	(1.15, 3.84)	0.014	0.656
Duke All stage	Hazard ratio	95% C.I.	p-value	CPE
Risk score	1.22	(1.02, 1.47)	0.032	0.580
Categorical	1.71	(1.01, 2.87)	0.043	0.566
TNM stage	2.17	(1.29, 3.63)	0.004	0.589
Duke stage I	Hazard ratio	95% C.I.	p-value	CPE
Risk score	1.44	(1.10, 1.87)	0.007	0.635
Categorical	2.93	(1.36, 6.34)	0.005	0.625
TNM stage	1.97	(0.95, 4.10)	0.070	0.580

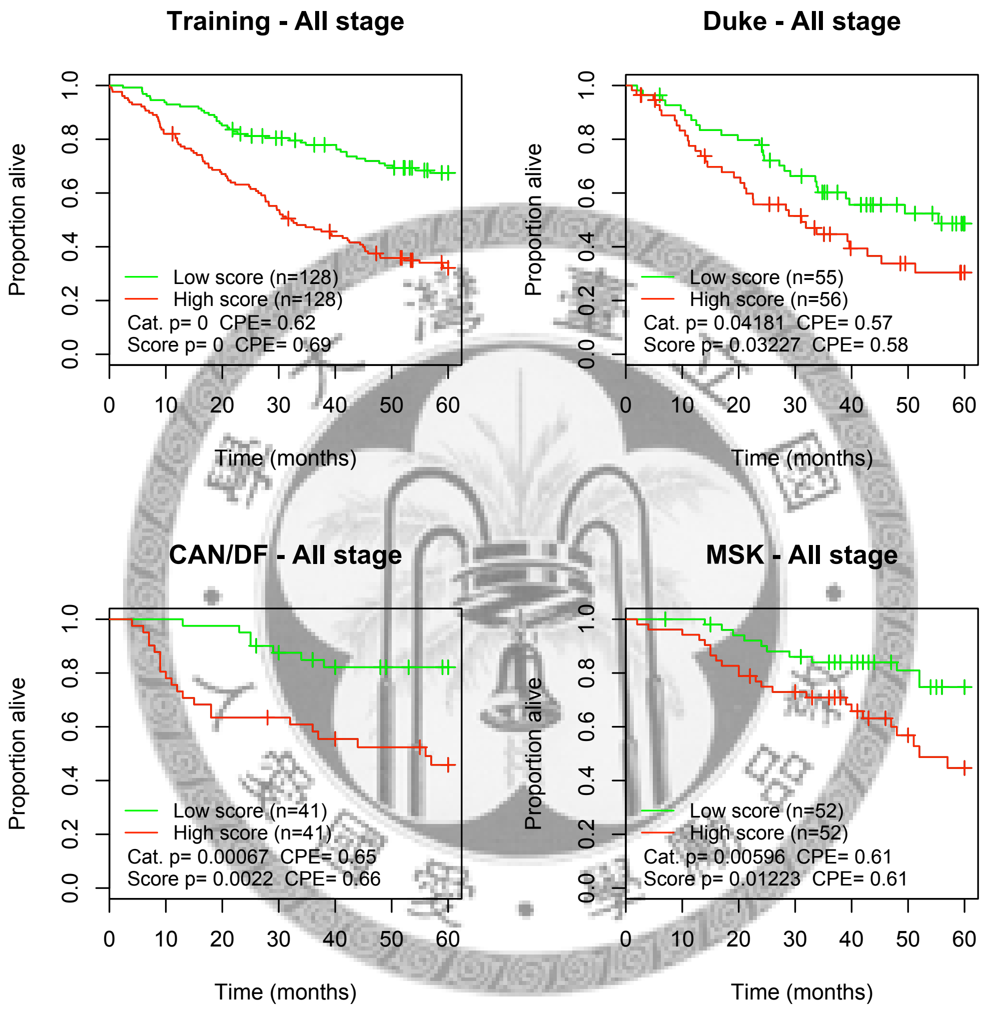


Figure 6.1: Kaplan-Meier curves for all stage samples separated by gene signature

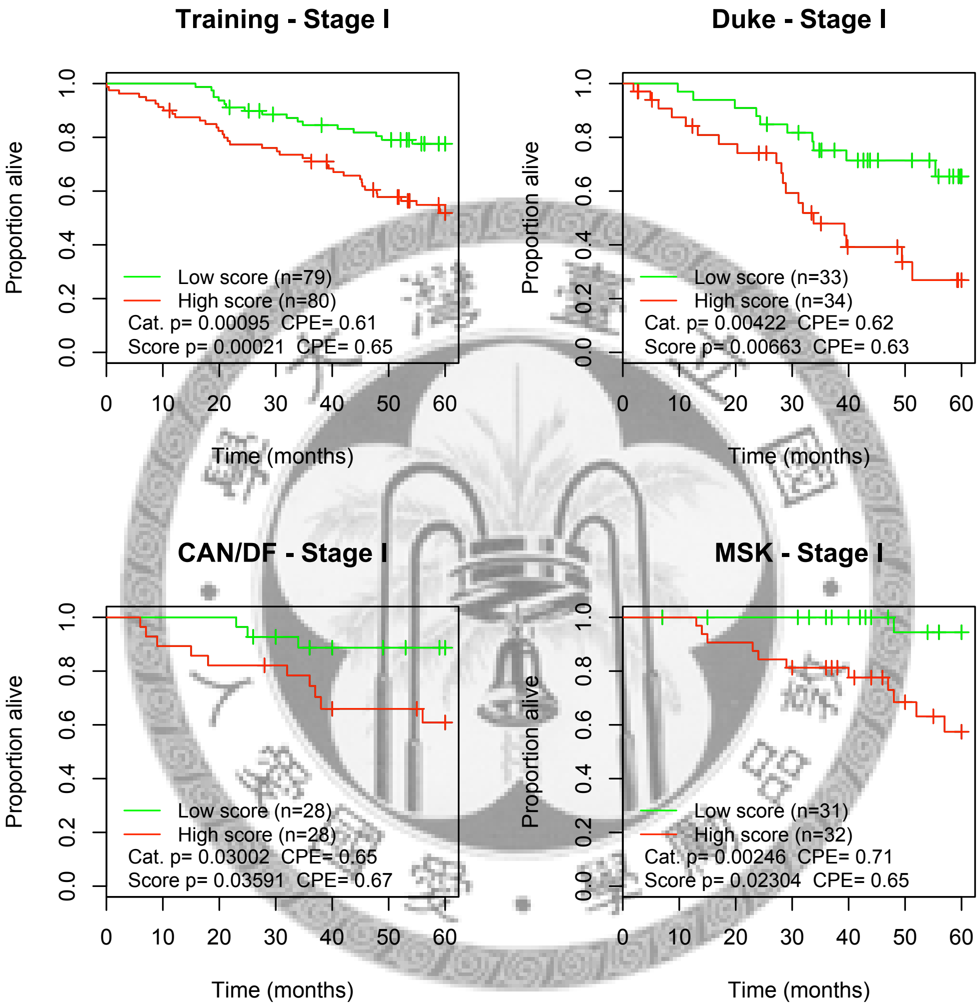


Figure 6.2: Kaplan-Meier curves for stage I samples separated by gene signature

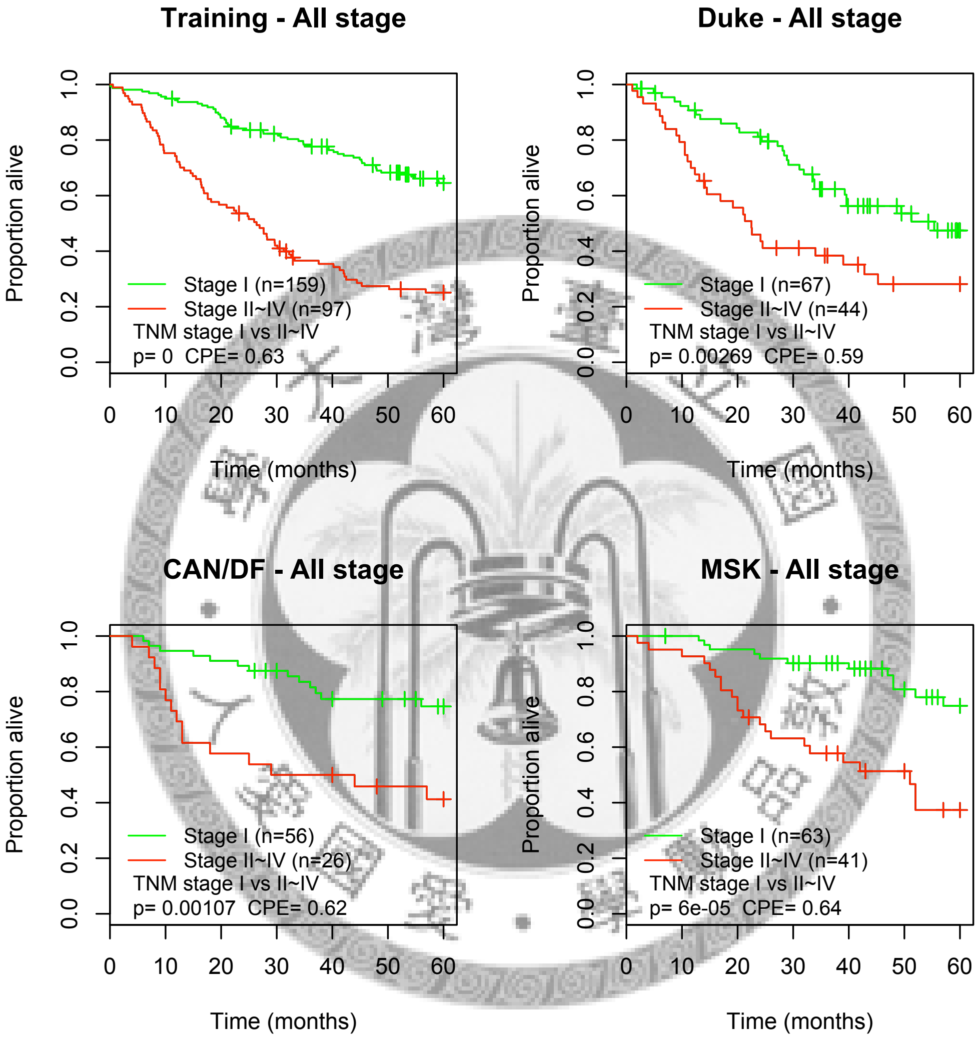


Figure 6.3: Kaplan-Meier curves for all stage samples separated by TNM stage

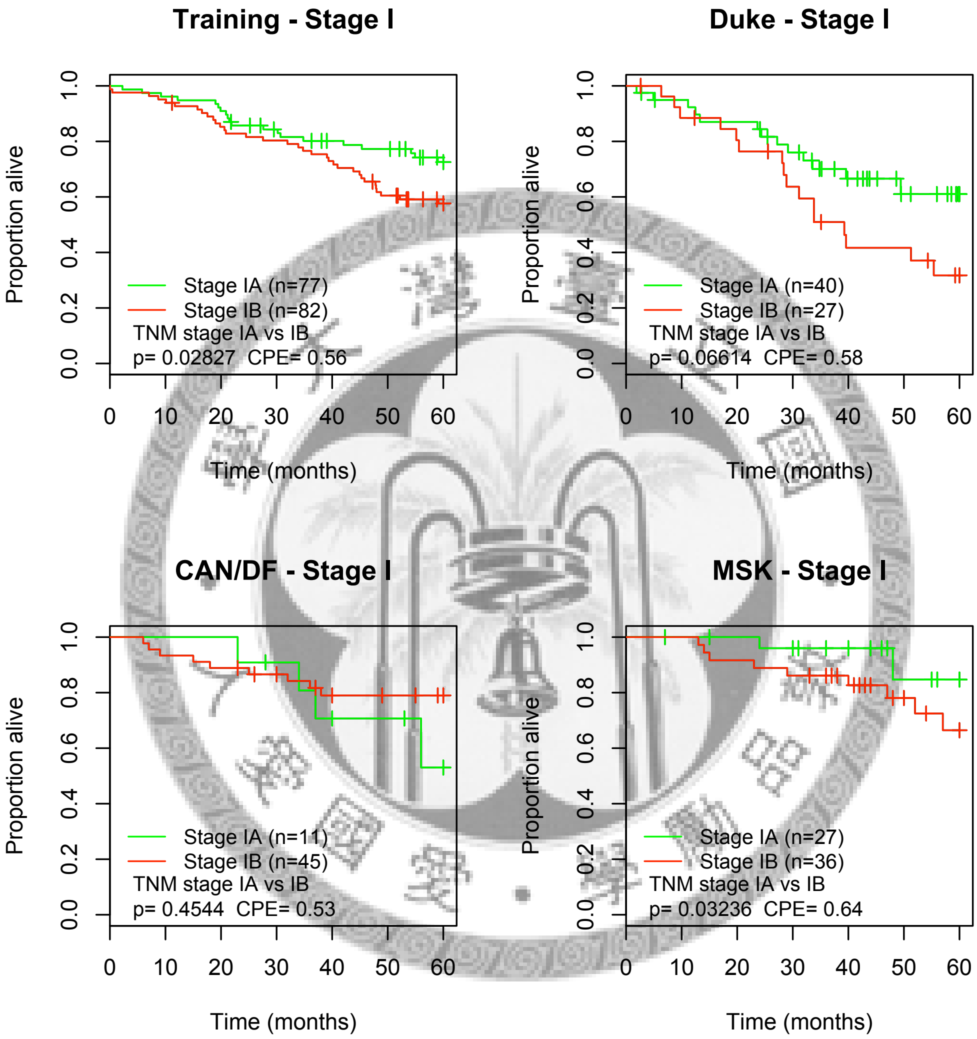


Figure 6.4: Kaplan-Meier curves for stage I samples separated by TNM stage

Chapter 7

Summary and discussion

7.1 Summary

We reanalyzed a large adenocarcinomas data from Shedden *et al.* (2008) and derived a gene signature from seven gene expression profiles. Tested in two independent validation data sets, our gene signature had significant prediction power for the survival of samples of all stage patients or stage I patients only. Furthermore, our signature also had significant prediction power in an external NSCLC data set that contained two different tumor cell types.

Most of the available analysis procedures contain three conceptions; initial data filter, feature selection and signature construction. Our analysis procedure also contains three steps; gene filter, gene selection and signature construction. Compared with other methods, our analysis procedure has several advantages. First, we proposed a new criterion to filter out some inconsistent genes. Second, the gene selection and the signature construction are both supervised by the patient survivals. Third, some non-linear interactive structures are considered in our procedure. Fourth, the model assumption for the gene selection and the signature construction part in our procedure is the least. Furthermore, the interpretation of our signature is easy and clear. However, there are still some unsolved issues. Although our LA hub gene selecting procedure shows the significancy of the selected LA hub gene, but the significancy of its paired genes is not showed. A method to screen the selected paired genes is need. However, in practice, we may rely on some Biological knowledge to choose the paired genes. In our analysis procedure, there is only one LA pair used in the signature construction part. However, if there are more LA pairs selected, a better model to incorporate the LA pairs in the dimension reduction model is needed.

7.2 Discussion

Discussion of the selected genes

Our gene signature was derived from the expression profiles of seven genes, where five genes were selected by correlation method and other two genes were selected by liquid association method. According to the sign of the coefficient of our signature, we called one selected gene protect or risk genes. The negative coefficient indicated the increase of expression associated with good prediction and the positive coefficient indicated the increase of expression associated with pool prediction. Both protect and risk genes were contained in these 7 genes.

The 3 protect genes are TMEM66, CSRP1 and BECN1. BECN1, also known as autophagy-related gene 6, played a key role in autophagy. It has been reported to be involved in various cancers. It can inhibit the growth of colorectal cancer cells. The expression of beclin 1 is associated with favorable prognosis in stage IIIB colon cancers. CSRP1 is related to gene regulation, cell growth, and differentiation. And it is hypothesized to be a colorectal cancer related tumor suppressor gene. TMEM66 is a novel gene known as transmembrane protein 66. The 2 risk genes are FOSL2 and ERO1L. FOSL2 is contained in the Fos gene family which encodes leucine zipper proteins that can dimerize with proteins of the JUN family, thereby forming the transcription factor complex AP-1. It was implicated as regulators of cell proliferation, differentiation, and transformation. The over expression of FOSL2 was also indicated that may play a major role in CCR4 expression and oncogenesis in ATLis associated with a more aggressive tumor phenotype and is probably involved in breast cancer progression in vivo. ERO1L is essential oxidoreductase; a source of oxidative stres. However, it was indicated as a protect gene in another lung adenocarcinoma prognostic study [30]. Another study [31] suggested that ERO1L plays a key role for inhibiting tumor growth via inhibiting VEGF-driven angiogenesis. These results disagreed with our finding. Therefore, to check this issue, we used the gene ERO1L to fit the univariate Cox proportional hazard model. In the training data set, the hazard ratio is 1.35 and it is significantly greater than 1 with 95% confidence interval from 1.19 to 1.54 (p-value $< 10^{-5}$). In other validation data sets, the hazard ratios are all greater than 1, although they are not significant. Thus, in this large sample data set, the increase of expression profile ERO1L was related poor prediction.

The LA hub gene SRP54 is known as signal recognition particle 54kDa. It binds to the signal sequence of presecretory protein when they emerge from the ribosomes and transfers them to TRAM. A gene chromosomes study suggest that SRP54, BAZ1A, NFKBIA, MBIP, HNF3A, and two uncharacterized expressed sequence tags are candidate targets of the amplification mechanism and therefore may be associated, together or separately, with development and progression of esophageal squamous cell carcinoma (ESC). Its paired gene PAWR is a human gene coding for a tumor-suppressor protein that induces apoptosis in cancer cells, but not in normal cells and specifically upregulated during apoptosis of prostate cells.

Relationship between our permutation procedure and Benjamini-Hochberg procedure under independent assumption

In gene selection part, we used a permutation procedure to select the cutoff of the correlation coefficients and LA scores. The cutoff was decided by the ratio of the expected number and the observed number of correlation coefficients greater than the cutoff. However, this issue can be viewed as a multiple testing issue. A total of 6,252 null hypotheses that X_g and Y° was uncorrelated were tested by the test statistics $corr(N(\mathbf{x}_g), N(\mathbf{y}^{\circ}))$, for $g = 1, 2, ..., G^*$. Since the normal quantile transformation only depended on the ranks of the variables, $N(\mathbf{x}_g)$'s were identically distributed. Thus, we can apply the permutation test to construct the reference distribution of $corr(N(\mathbf{x}), N(\mathbf{y}^{\circ}))$ and get the significant level for each observed $r_g = corr(N(\mathbf{x}_g), N(\mathbf{y}^{\circ}))$. The two-sided p-value for a corresponding cutoff r can be evaluated by

$$p(r) = \frac{1}{256!} \sum_{N(\mathbf{x})^* \in \mathcal{N}} \mathbf{1}_{\{|corr(N(\mathbf{x}), N(\mathbf{x})^*)| \geq r\}}(N(\mathbf{x}), N(\mathbf{x})^*), \tag{7.1}$$

where $\mathcal{N}$ was the collection of all the permutations of $N(\mathbf{x})^*$'s. Since the number of elements in $\mathcal{N}$ was too large (256!), in practice, the p-value can be estimated by

$$\hat{p}(r) = \frac{1}{\#(\mathcal{M})} \sum_{N(\mathbf{x})^* \in \mathcal{M}} \mathbf{1}_{\{|corr(N(\mathbf{x}), N(\mathbf{x})^*)| \geq r\}}(N(\mathbf{x}), N(\mathbf{x})^*), \tag{7.2}$$

where $\mathcal{M}$ was a large subset of $\mathcal{N}$. With the p-value given from the permutation test, we might implement the Benjamini-Hochberg procedure [17] to solve the multiple testing issue. The Benjamini-Hochberg procedure can be summarized as the follow. Let $p_{(i)}$ be the ordered p-value, for $i = 1, 2, ..., n$. The null hypotheses $H^0_{(1)}, H^0_{(2)}, ..., H^0_{(k)}$ were rejected for $k = \max\{i \mid p_{(i)} \leq \frac{i}{m}q\}$. Under the independent or positively correlated assumption of the tests, this procedure controlled the false discovery rate (FDR) at level $\frac{m_0}{m}q \leq q$, where m was the total number of hypotheses (G^*) and m_0 was the number of true null hypotheses. Furthermore, in Benjamini and Yekutieli (2001) [17], they concluded that if we replaced q by $q/\sum_{i=1}^{m}\frac{1}{i}$, then the Benjamini-Hochberg procedure still controlled the false discovery rate (FDR) at level $\frac{m_0}{m}q \leq q$. Nevertheless, the modified criterion became more conservative than the Bonferroni method for the first few ordered p-value. In our data, there were no genes could be selected with this modified procedure.

The complex correlation structure of thousands of genes is not clear. The dependency of these tests is also not clear. The modified procedure might be too conservative. Here we discuss the relationship between our permutation procedure and the Benjamini-Hochberg procedure under the independent assumption. If we assume that the expressions $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{G^*}$ were independent, then our procedure constructed a reference distribution with a total number 1,000×6,252. The p-value of the ordered absolute value of correlation coefficients $\hat{r}_{(g)}$ can be estimated by $\hat{e}_g/G^*$ from equation (7.2) with $\#(\mathcal{M}) = 1000G^*$. Then our selecting criterion $j = \max\{i \mid e_i/i \leq \alpha\}$ is the same as the original criterion in the Benjamini-Hochberg procedure by taking $\alpha = q$ and dividing G^* both side. Thus, under the independent assumption, our selecting procedure is the same as the Benjamini-Hochberg procedure with p-value estimated by the permutation test.

Gene signature performance in different data preprocess methods

The expression profiles we analyzed in were preprocessed by MAS 5.0 Statistical algorithm from Affymatrix (2001). One reason we used the MAS 5.0 preprocessing method is that the MAS 5.0 algorithm allows us to implement on each chip separately. To preprocess the training and testing data together is somewhat unrealistic. This reason also motivated us to filter out the genes with inconsistent expression profiles by different preprocessing method. However, we used the expression profiles preprocessed by two different preprocessing methods dChip and RMA of our selected genes to derive the signature by applying modified SIR in training data. All the preprocess methods were implemented separately for each data set. After implementing the modified slice inverse regression method, the p-values of the large sample chi-spuared test were all greater than 0.05 for these two preprocess method. We did not suggest using any leading eigenvector as the gene signature. However, as a comparison, we still validated the signatures with the same procedure on CAN/DF and MSK data sets. The estimated coefficients of SIR direction for expression profiles preprocessed by dChip and RMA were given in table 7.1

We noted that our cross platforms adjustment could not be applied for these preprocessing methods. Therefore, we did not test the signatures on the Duke data set. The estimated hazard ratios with the corresponding 95% confident intervals, p-values and the CPE were given in table 7.2. The results showed that the continuous risk were also significant in all testing sets and

Table 7.1: The estimated coefficients of SIR direction for expression profiles preprocessed by dChip and RMA

dChip	SIR dir. coefficient		RMA	SIR dir. coefficient	
TMEM66	-0.8949	(Protect)	TMEM66	-0.6386	(Protect)
CSRP1	-0.4813	(Protect)	CSRP1	-0.5222	(Protect)
BECN1	-1.7374	(Protect)	BECN1	-1.0540	(Protect)
FOSL2	0.7738	(Risk)	FOSL2	0.4945	(Risk)
ERO1L	0.6041	(Risk)	ERO1L	0.5166	(Risk)
(SRP54)	-0.5563	(-)	(SRP54)	-0.5033	(-)
(PAWR)	0.0448	(-)	(PAWR)	0.0600	(-)
SRP54*PAWR	-0.3652	(Protect)	SRP54*PAWR	-0.7517	(Protect)
p-value		0.10	p-value		0.10

the categorical classifier were not significant for stage I patients in CAN/DF data. The Kaplan-Meier curves also illustrated that these signature performed worse than the original signature for the stage I patients in CAN/DF data preprocessed by MAS 5.0 algorithm. One possible reason was the mean shift caused by preprocessing data set separately. Our dimension reduction model was nonlinear because we added the interaction term. The mean shift issue may affect our nonlinear structure. Therefore, these results suggest that the gene signature had potential for lung adenocarcinomas diagnostic, but for implementing in practice, the improvement of array technique and the preprocessing is still needed.

Table 7.2: Validation results in CAN/DF and MSK data preprocessed by dChip and RMA

CAN/DF All stage	Hazard ratio	95% C.I.	p-value	CPE
Risk score - dChip	2.76	(1.78, 4.28)	0.000	0.710
Risk score - RMA	2.04	(1.47, 2.82)	0.000	0.659
Categorical - dChip	4.35	(1.84, 10.28)	0.000	0.659
Categorical - RMA	3.44	(1.51, 7.84)	0.002	0.639
CAN/DF stage I	Hazard ratio	95% C.I.	p-value	CPE
Risk score - dChip	2.12	(1.13, 3.97)	0.024	0.653
Risk score - RMA	1.70	(1.06, 2.73)	0.049	0.617
Categorical - dChip	2.11	(0.69, 6.50)	0.182	0.591
Categorical - RMA	1.48	(0.49, 4.42)	0.484	0.549
MSK All stage	Hazard ratio	95% C.I.	p-value	CPE
Risk score - dChip	1.79	(1.20, 2.67)	0.004	0.632
Risk score - RMA	1.64	(1.09, 2.48)	0.019	0.607
Categorical - dChip	2.88	(1.38, 6.03)	0.003	0.622
Categorical - RMA	3.04	(1.45, 6.35)	0.002	0.627
MSK stage I	Hazard ratio	95% C.I.	p-value	CPE
Risk score - dChip	3.08	(1.41, 6.71)	0.003	0.715
Risk score - RMA	2.46	(1.18, 5.15)	0.017	0.672
Categorical - dChip	9.97	(1.28, 77.5)	0.003	0.708
Categorical - RMA	9.97	(1.28, 77.5)	0.003	0.708

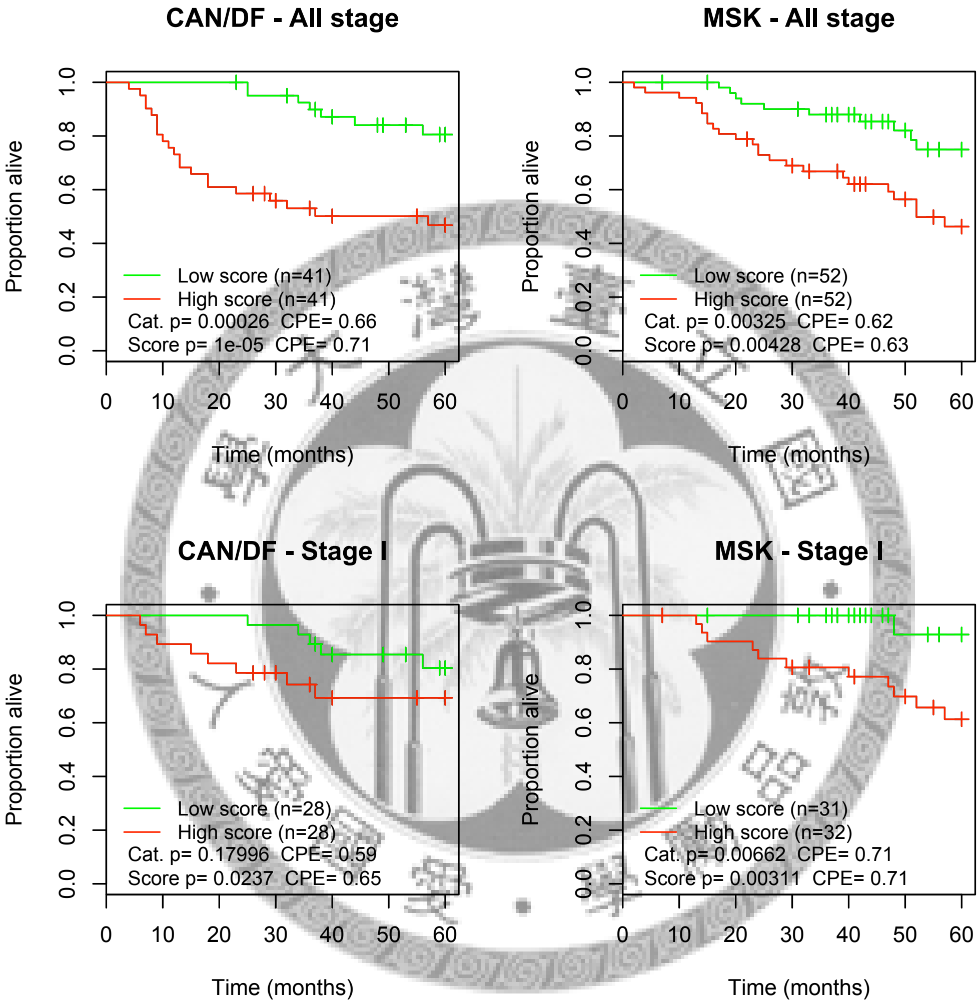


Figure 7.1: Kaplan-Meier curves for all stage and stage I samples separated by gene signature preprocessed by dChip in CAN/DF and MSK data

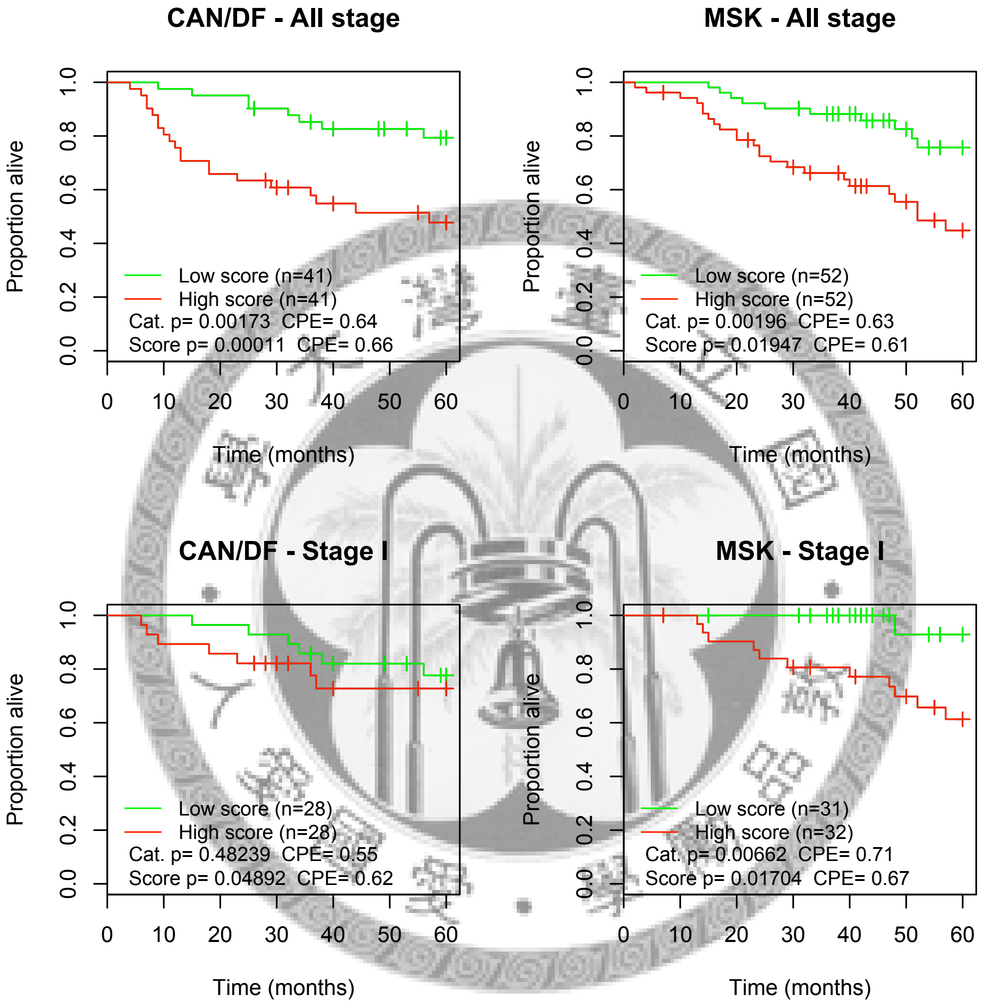


Figure 7.2: Kaplan-Meier curves for all stage and stage I samples separated by gene signature preprocessed by RMA in CAN/DF and MSK data

Bibliography

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu and M. J. Thun, **Cancer Statistics, 2009**, *CA Cancer J. Clin.*, **59** (2009), 225-249.

[2] D. M. Parkin, F. Bray, J. Ferlay and P. Pisani, **Global Cancer Statistics, 2002**, *CA Cancer J. Clin.*, **55** (2005), 74-108.

[3] C. M. Booth and F. A. Shepherd, **Adjuvant chemotherapy for resected non-small cell lung cancer**, *J. Thorac. Oncol.*, **1** (2006), 180-187.

[4] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker and M. Meyerson, **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**, *Proc. Natl. Acad. Sci.*, **98** (2001), 13790-13795.

[5] M. E. Garber, O. G. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. van de Rijn, G. D. Rosen, C. M. Perou, R. I. Whyte, R. B. Altman, P. O. Brown, D. Botstein and I. Petersen, **Diversity of gene expression in adenocarcinoma of the lung**, *Proc. Natl. Acad. Sci.*, **98** (2001), 13784-13789.

[6] D. A. Wigle, I. Jurisica, N. Radulovich, M. Pintilie, J. Rossant, N. Liu, C. Lu, J. Woodgett, I. Seiden, M. Johnston, S. Keshavjee, G. Darling, T. Winton, B. J. Breitkreutz, P. Jorgenson, M. Tyers, F. A. Shepherd and M. S. Tsao, **Molecular profiling of non-small cell lung cancer and correlation with disease-free survival**, *Cancer Res.*, **62** (2002), 3005-3008.

[7] A. Potti, S. Mukherjee, R. Petersen, H. K. Dressman, A. Bild, J. Koontz, R. Kratzke, M. A. Watson, M. Kelley, G. S. Ginsburg, M. West, D. H. Harpole Jr. and J. R. Nevins, **A genomic strategy to refine prognosis in early-stage non–small-cell lung cancer**, *N. Engl. J. Med.*, **355** (2006), 570-580.

[8] H. Y. Chen, S. L. Yu, C. H. Chen, G. C. Chang, C. Y. Chen, A. Yuan, C. L. Cheng, C. H. Wang, H. J. Terng, S. F. Kao, W. K. Chan, H. N. Li, C. C. Liu, S. Singh, W. J. Chen, J. J. W. Chen and P. C. Yang, **A five-gene signature and clinical outcome in non-small-cell lung cancer**, *N. Engl. J. Med.*, **356** (2007), 11-20.

[9] Y. Lu, Y. Lemon, P. Y. Liu, Y. Yi, C. Morrison, P. Yang, Z. Sun, J. Szoke, W. L. Gerald, M. Watson, R. Govindan and M. You, **A gene expression signature predicts survival of subjects with stage I nonsmall cell lung cancer**, *PLoS Med.*, **12** (2006), e467.

[10] K. Shedden, J. M. G. Taylor, S. A. Enkemann, M. S. Tsao, T. J. Yeatman, W. L. Gerald, S. Eschrich, I. Jurisica, T. J. Giordano, D. E. Misek, A. C. Chang, C. Q. Zhu, D. Strumpf, S. Hanash, F. A. Shepherd, K. Ding, L. Seymour, K. Naoki, N. Pennell, B. Weir, R. Verhaak, C. Ladd-Acosta, T. Golub, M. Gruidl, A. Sharma, J. Szoke, M. Zakowski, V.

Rusch, M. Kris, A. Viale, N. Motoi, W. Travis, B. Conley, V. E. Seshan, M. Meyerson, R. Kuick, K. K. Dobbin, T. Lively, J. W. Jacobson and D. G. Beer, **Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study**, *Nat. Med.*, **14** (2008), 822-827.

[11] T. Wu, W. Sun, S. Yuan, C. H. Chen and K. C. Li, **A method for analyzing censored survival phenotype with gene expression data**, *BMC Bioinformatics*, **9** (2008), 417.

[12] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M. B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson Jr., J. R. Marks, H. K. Dressman, M. West and J. R. Nevins, **Oncogenic pathway signatures in human cancers as a guide to targeted therapies**, *Nature*, **439** (2006), 353-357.

[13] K. C. Li, **Genome-wide co-expression dynamics: theory and application**, *Proc. Natl. Acad. Sci.*, **99** (2002), 16875-16880.

[14] K. C. Li, A. Palotie, S. Yuan, D. Bronnikov, D. Chen, X. Wei and O. W. Choi, J. Saarela and L. Peltonen, **Finding disease candidate genes by liquid association**, *Genome Biol.*, **8** (2007), R205.

[15] K. C. Li, **Sliced inverse regression for dimension reduction (with discussion)**, *J. Amer. Statist. Assoc.*, **86** (1991), 316-327.

[16] K. C. Li, J. L. Wang and C. H. Chen, **Dimension reduction for censored regression data**, *The Annals of Statistics*, **27** (1999), 1-23.

[17] Y. Benjamini and D. Yekutieli, **The control of the false discovery rate in multiple testing under dependency**, *The Annals of Statistics*, **29** (2001), 1165–1188.

[18] Affymetrix, *Statistical algorithms reference guide* Technical report, Affymetrix Inc.; (2001).

[19] C. Li, and W. H. Wong, **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error applications**, *Genome Biol.*, **2** (2001), 1-11.

[20] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed, **Exploration, normalization and summaries of high density oligonucleotide array probe level data**, *Biostatistics*, **4** (2003), 249-264.

[21] D. R. Cox and D. Oakes, *Analysis of survival data*, London; New York: Chapman and Hall Ltd.; (1984).

[22] J. P. Klein and M. L. Moeschberger, *Survival Analysis - Techniques for Censored and Truncated Data second edition*, New York: Springer; (2003).

[23] M. Gönen and G. Heller, **Concordance probability and discriminatory power in proportional hazards regression**, *Biometrika*, **92** (2005), 965-970.

[24] L. Pirtoli, G. Cevenini, P. Tini, M. Vannini, G. Oliveri, S. Marsili, V. Mourmouras, G. Rubino and C. Miracco, **The prognostic role of Beclin 1 protein expression in high-grade gliomas**, *Autophagy*, **5** (2009), 930-936.

[25] K. Koneri, T. Goi, Y. Hirono, K. Katayama and A. Yamaguchi, **Beclin 1 gene inhibits tumor growth in colon cancer cell lines**, *Anticancer Res.*, **27** (2007), 1453-1458.

[26] B. X. Li, C. Y. Li, R. Q. Peng, X. J. Wu, H. Y. Wang, D. S. Wan, X. F. Zhu and X. S. Zhang, **The expression of beclin 1 is associated with favorable prognosis in stage IIIB colon cancers**, *Autophagy*, **5** (2009), 303-306.

[27] C. Z. Zhou, G. Q. Qiu, X. L. Wang, J. W. Fan, H. M. Tang, Y. H. Sun, Q. Wang, F. Huang, D. W. Yan, D. W. Li and Z. H. Peng, **Screening of tumor suppressor genes on 1q31.1-32.1 in Chinese patients with sporadic colorectal cancer**, *Chin. Med. j. (Engl.)*, **121** (2008), 2479-2486.

[28] T. Nakayama, K. Hieshima, T. Arao, Z. Jin, D. Nagakubo, A. K. Shirakawa, Y. Yamada, M. Fujii, N. Oiso, A. Kawada, K. Nishio and O. Yoshie, **Aberrant expression of Fra-2 promotes CCR4 expression and cell proliferation in adult T-cell leukemia**, *Oncogene*, **27** (2008), 3221-3232.

[29] K. Milde-Langosch, S. Janke, I. Wagner, C. Schröder, T. Streichert, A. M. Bamberger, F. Jänicke and T. Löning, **Role of Fra-2 in breast cancer: influence on tumor cell invasion and motility**, *Breast Cancer Res. Treat.*, **107** (2008), 337-347.

[30] H. Endoh, S. Tomida, Y. Yatabe, H. Konishi, H. Osada, K. Tajima, H. Kuwano, T. Takahashi and T. Mitsudomi, **Prognostic model of pulmonary adenocarcinoma by expression profiling of eight genes as determined by quantitative real-time reverse transcriptase polymerase chain reaction**, *J. Clin. Oncol.*, **22** (2004), 881-889.

[31] D. May, A. Itin, O. Gal, H. Kalinski, E. Feinstein and E. Keshet, **Ero1-L alpha plays a key role in a HIF-1-mediated pathway to improve**

disulfide bond formation and VEGF secretion under hypoxia: implication for cancer, *Oncogene*, **24** (2005), 1011-1020.

[32] K. Yasui, I. Imoto, Y. Fukuda, A. Pimkhaokham, Z. Q. Yang, T. Naruto, Y. Shimada, Y. Nakamura and J. Inazawa, **Identification of target genes within an amplicon at 14q12-q13 in esophageal squamous cell carcinoma**, *Genes, Chromosomes and Cancer*, **32** (2001), 112-118.

[33] ”Entrez Gene” http://www.ncbi.nlm.nih.gov/gene/.