2011

Computational Hybrid Systems for Identifying Prognostic Gene Markers of Lung Cancer

Ying-Wooi Wan
West Virginia University

Computational Hybrid Systems for Identifying Prognostic Gene Markers of Lung Cancer

Ying-Wooi Wan

Dissertation submitted to the
College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Computer and Information Science

Lan Guo, Ph.D., Chair
Bojan Cukic, Ph.D.
Tim Menzies, Ph.D.
Arun Ross, Ph.D.
Mark Culp, Ph.D.
James Denvir, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2011

Keywords: Gene Signature, Molecular Prognosis, Hybrid System, Implication Network, Biomarker Discovery, Survival Prediction

Abstract

Computational Hybrid Systems for Identifying Prognostic Gene Markers of Lung Cancer

Ying-Wooi Wan

Lung cancer is the most fatal cancer around the world. Current lung cancer prognosis and treatment is based on tumor stage population statistics and could not reliably assess the risk for developing recurrence in individual patients. Biomarkers enable treatment options to be tailored to individual patients based on their tumor molecular characteristics. To date, there is no clinically applied molecular prognostic model for lung cancer. Statistics and feature selection methods identify gene candidates by ranking the association between gene expression and disease outcome, but do not account for the interactions among genes. Computational network methods could model interactions, but have not been used for gene selection due to computational inefficiency. Moreover, the curse of dimensionality in human genome data imposes more computational challenges to these methods.

We proposed two hybrid systems for the identification of prognostic gene signatures for lung cancer using gene expressions measured with DNA microarray. The first hybrid system combined *t*-tests, Statistical Analysis of Microarray (SAM), and *Relief* feature selections in multiple gene filtering layers. This combinatorial system identified a 12-gene signature with better prognostic performance than published signatures in treatment selection for stage I and II patients (log-rank $P<0.04$, Kaplan-Meier analyses). The 12-gene signature is a more significant prognostic factor (hazard ratio=4.19, 95% CI: [2.08, 8.46], $P<0.00006$) than other clinical covariates. The signature genes were found to be involved in tumorigenesis in functional pathway analyses.

The second proposed system employed a novel computational network model, i.e., implication networks based on prediction logic. This network-based system utilizes gene coexpression networks and concurrent coregulation with signaling pathways for biomarker identification. The first application of the system modeled disease-mediated genome-wide coexpression networks. The entire genomic space were extensively explored and 21 gene signatures were discovered with better prognostic performance than all published signatures in stage I patients not receiving chemotherapy (hazard ratio>1, CPE>0.5, $P < 0.05$). These signatures could potentially be used for selecting patients for adjuvant chemotherapy. The second application of the system modeled the smoking-mediated coexpression networks and identified a smoking-associated 7-gene signature. The 7-gene signature generated significant prognostication specific to smoking lung cancer patients (log-rank $P<0.05$, Kaplan-Meier analyses), with implications in diagnostic screening of lung cancer risk in smokers (overall accuracy=74%, $P<0.006$). The coexpression patterns derived from the implication networks in both applications were successfully validated with molecular interactions reported in the literature ($FDR<0.1$).

Our studies demonstrated that hybrid systems with multiple gene selection layers outperform traditional methods. Moreover, implication networks could efficiently model genome-scale disease-mediated coexpression networks and crosstalk with signaling pathways, leading to the identification of clinically important gene signatures.

Acknowledgements

I would like to offer my sincerest gratitude to my advisor, Dr. Nancy L. Guo for all the guidance, advice, and encouragements she had provided me over the years. I would like to thank her for the invaluable training and experience she had given me. I also thank her for her continual patience and helpfulness. Without her, this project would not be able to come about.

I would also like to thank my other committee members: Dr. Bojan Cukic, Dr. Arun Ross, Dr. Tim Menzies, and Dr. Mark Culp for their help in my study and research. Special thanks are also due to Dr. Jim Denvir. He has made available his assistance and insightful ideas for my projects in a number of ways.

It is a pleasure to thank my lab colleagues Swetha Bose, Joseph Putila, Jason Young, Kursad Tosun, Shruti Rathnagiriswaran, Rama kanth Mettu, ChangChang Xiao, JiaJia Wang, DaJie Luo, Naveen Bondalapati, Ebrahim Sabbagh, Rebecca Raese, Maricica Pacurari, and Julian Dymacek for their meaningful discussions, time, and friendships throughout the years.

This project is supported by NIH R01LM009500 and NCRR P20RR16440 (Guo). Software license and training for Pathway Studio is supported by NIH/NCRR P2016477. I would like to thank the sources for these funding.

Last but not least, I would like to show my deepest gratefulness to my family for their endless love and support. Without their understanding and encouragement, I would not have gone this far.

Contents

Chapter 5

Network-based Identification of Smoking-associated Gene Signature for Lung Cancer 95

Chapter 6

Evaluation with Boolean Implication Networks and Bayesian Networks 110

Chapter 7

Contributions and Future Work 123

List of Figures

List of Tables

Chapter 1

Introduction

For the past decades, cancer has been the major health problem to industrialized countries around the world. Among all types of cancers, lung cancer is the leading cause of cancer-related deaths [3]. Treatment failure will lead to death in lung cancer. Currently, surgery is the foremost treatment option for patients with stage I non-small cell lung cancer (NSCLC). However, 35–50% of stage I NSCLC patients will relapse within 5 years [4, 5]. It remains a critical challenge to determine the risk for recurrence in early-stage lung cancer patients. Patients at high risk for recurrence might benefit from adjuvant chemotherapy, whereas those with a low risk for tumor recurrence might be spared from the side effects of chemotherapy. Following this, another critical issue in clinics is to determine an individual patient's predisposition to a specific anticancer drug. The emerging use of biomarkers may enable physicians to make treatment decisions based on the specific characteristics of individual patients and their tumor, instead merely of on population statistics [6].

Microarray technologies present a convenient platform for scientists and clinical investigators to gain new insights into biology and ultimately for developing clinical applications [7]. The advancements in microarray technologies lead to promising achievements in the molecular prediction of individual clinical outcome. Two successful examples include the commercial gene tests for breast cancer, Oncotype DX [8] and MammaPrint [9, 10]. There have been a few studies on lung cancer signatures and molecular prognosis by transcriptional profiling [2, 11-17]. To date, there is no fully-validated and clinically applied model for predicting lung cancer recurrence [18].

On the other hand, microarray technologies pose a few challenges in computational techniques for molecular prognosis and diagnosis researches [19]. The first challenge is the high dimensionality of the data. A typical microarray experiment would be able to profile up to tens of thousands of genes. This nature of microarray data has complicated major diagnostic and prognostic breakthroughs [20] and puts a premium on innovative feature selection and data mining methods. Feature selection methods allow us to select a subset of predictive genes as biomarkers. With the discovered biomarkers, we could construct a faster, cost-effective prognostic or diagnostic classifier with improved performance [21]. The main objective of feature selection is to remove irrelevant features and retain only the informative features. Nevertheless, the search of the optimal subset of informative features in the space of all features is NP-hard. This hard problem becomes more difficult when the microarray data is also small in sample size and clouded with noisy biological confounding effects [22].

The most intuitive approach to identify candidate marker genes is to rank genes according to their association with the clinical outcome and select the top ranked genes. However, studies had shown that individual genes showing strong association with the outcome are not necessarily good classifiers [23-25]. Moreover, instead of functioning alone, genes and proteins interact with one another to form modular machines [26]. Ranking-based approaches that evaluate each gene individually could not model interactions among genes. Therefore, with the completion of the Human Genome Project, understanding the networks of interactions among genes had become increasingly important to reveal the molecular basis of disease for biomarker identification [27].

Currently, various techniques had been applied to microarray studies to identify biomarkers and construct molecular classifier. To discovery predictive gene signatures, statistical methods and feature selection methods is simple and efficient but would not account for the complex interactive machinery among genes. On the other hand, network-based approaches overcome the limitations of statistical and feature selection methods by providing a closer modeling of genetic interactive nature. However, they might suffer from computational complexities. Once the set of signature genes had been identified, construction of the molecular classifier poses another set of challenges. These challenges include assessment of the classifiers' robustness [19] and the true biological validity of the findings [28]. Combining the limitations of various methods with challenges originated from microarray data discussed above, it remains an open problem in this research domain to develop a methodology to efficiently identify a set of predictive genes as the biomarker for molecular prognosis or diagnosis.

In this dissertation, we proposed two computational hybrid systems as the robust platform to identify prognostic gene signatures for lung cancer molecular prognosis. The first hybrid system combined multiple traditional statistics and feature selection methods in different stages for gene filtering. The second hybrid model integrated a novel network model, i.e. the implication networks based on prediction logic. The integration of the network models in the second system incorporates the information of gene interactions with major signaling hallmarks in the identification of prognostic gene signatures. To examine the proposed hybrid system methodologies, three studies were carried out. The first study examined the first combinatorial framework with traditional statistics and feature selection methods. It demonstrated that the combinatorial scheme of using different traditional methods to filter genes in multiple stages identify better gene signatures when applying these methods alone. The second hybrid system that is built upon the innovative implication networks was investigated in the second and the third study. The second study employed the network-based system to explore the prognostic signatures discovery in the whole genomic scale. Extensive gene signatures for lung cancer with better prognostication than all published signatures were identified in this study. Instead of the entire genome, the third study applied the network-based model in a smaller scope: genes significantly associated with lung cancer survival and smoking status. This leads to the identification of prognostic gene signatures specific to the smoking lung cancer patients. The implication networks efficiently and accurately model gene coexpression patterns perturbed by the disease outcome or other factor, such as smoking status. Furthermore, it leads to identification of prognostic genes tightly involved in signaling pathway.

The remainder of the proposal is organized as follows. Chapter 2 presents the related work with focuses on various gene selection methods. These methods would be described and their strengths and weakness would be summarized and discussed. Chapter 3 describes the first study of hybrid models with statistical and feature selection methods. Chapter 4 includes the second study with the integration of implication networks in the hybrid model in genomic scale. Chapter 5 presents the study of using the implication networks in the identification of a smoking-associated signature. Chapter 6 compares the implication networks employed in our studies with two network models, i.e. Boolean implication networks and Bayesian networks. Finally, the last chapter, Chapter 7 discusses the contributions of our studies and the future works for the methodology.

Chapter 2

Related Work

In recent two decades, with advancements in high-throughput biotechnologies and knowledge in genomic profiling such as microarray technologies, researches are able to investigate prognostic factors of cancer using genomic data such as gene expression values. These studies involved identification of gene signatures as biomarkers, construction of molecular prognostic models using the identified biomarkers, and validation of the findings for clinical applications. This chapter provides a review of the methods and works related to our studies. Since public data was use in our studies, a brief description of the few data sets used is included at the end of the chapter. The first three sections will discuss the methods for genomic signatures identification in three categories: ranking-based gene selection methods (2.1), network-based methods (2.2), and regularized linear models (2.3). A thorough discussions on the few major methods reviewed in the first three sections is given in Section 2.4. Statistical methods and bioinformatics tools used to validate the survival and biological aspect of the computational findings will be discussed in Section 2.5. Section 2.6 describes of the few public data sets used in our studies. Finally, we summarize the chapter with Section 2.7.

2.1 Ranking-based Gene Selection Methods

2.1.1 Introduction

In several microarray studies, genes are ranked according to their association with the clinical outcome, and the top ranked genes are identified as the gene signature and included in the prognostic classifier [8, 11, 12, 15, 29, 30]. The top rank genes could be either a fixed number of top tanked genes (such as top 10% of the ranked list). Alternatively, a threshold can be set on the ranking criterion and the genes whose criterion exceeds the threshold are selected as signature genes. Methods to study and rank the association of genes could be grouped into two major categories: statistical methods and traditional feature selection methods. This section will briefly describe various methods found in these two categories.

2.1.2 Statistical Methods

The traditional practice to study the gene association to the clinical outcome from microarray data is to identified genes that differentially expressed between two clinical states, for example between the disease state and normal state. One approach to identified differentially expressed genes is to compute the gene expression fold change between the two states for each gene and assess the observed fold change with statistical significance test. The commonly used statistical significance test is the conventional *t*-tests, which provides the probability (*P*) that the computed changes in expression occurred by chance [31]. Genes that pass certain predetermined threshold of fold change and statistical significance constitute the list of prognostic genes [32].

Microarray data is usually small in sample size with large number of genes. This poses a challenge in determining significance level using *P*-value from conventional *t*-test. For example, the traditional statistical significant level of $P = 0.05$ will lead to discovery of 1,000 false positive genes by chance in a microarray experiment with 20,000 genes. On the other hand, a more stringent threshold such as $P = 0.001$ will decrease false positives but result in high false negatives which will lead to failure in discovering a lot predictive genes [33]. It is a multiple hypothesis testing problem to determine if a gene has significantly different changes amongst large number of other genes. Multiple testing correction methods such as Bonferroni correction is too conservative to be applied in microarray experiment. For example, it would require a gene with $P < 0.05/20{,}000$ in order to be declared significant, which is so small that hardly any genes could achieve that threshold. Therefore, statistical methods used to control the false discovery rate (*FDR*) are used instead. Such a method commonly used in microarray analysis is significant analysis of microarrays (SAM) [34].

SAM is used to identify genes with statistically significant changes in two different biological states. It accounts for the multiple hypothesis testing problems in microarray analysis by estimating the *FDR* for the set of significant genes based on permutation test. In SAM, a modified *t*-test, or known as gene-specific *t*-test is used. For each gene, it computes the ratio change in gene expression relative to standard deviation, known as the "relative difference":

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0} \tag{1}$$

where $\bar{x}_I(i)$ and $\bar{x}_U(i)$ are defined as the mean expression for gene (*i*) in states *I* and *U*, respectively. $s(i)$ is the standard deviation in expression for gene (*i*) and s_0 is a positive constant used to ensure the relative difference $d(i)$ is independent of the gene expression.

The procedure carried out in SAM is depicted in Fig. 2.1, the observed relative differences for all the *n* genes are ranked into ascending order. Null distribution of the relative difference of each gene is generated by random permutations of samples' class labels for π times. In each iteration *p*, null relative difference $d_p(i)$ for each genes are ranked into ascending order as well. From the null distribution of relative differences, the expected relative difference $d_E(i)$ for gene (*i*) is computed by averaging the null relative differences from all iterations:

$$d_E(i) = \frac{\sum_{p=1}^{\pi} d_p(i)}{\pi} \tag{2}$$

The expected relative differences are then ranked into ascending order and plotted against the observed relative differences. From the scatter plot, genes with differences between the observed and expected relative differences greater than the threshold *delta* (Δ) are identified as significant genes (or known as "called significant" genes), which could be defined as the set *T:*

$$T = \{i : |d(i) - d_E(i)| > \Delta\} \tag{3}$$

In each iteration, falsely discovered genes are those whose null relative difference $d_p(i)$ exceeds the horizon cutoffs (du and dl) of observed relative difference for genes called significant found from the scatter plot.

The estimated FDR relative to the set of significant genes T is defined as the ratio of the average number of genes falsely discovered from all π permutations over the total number of genes called significant. Mathematically, this could be represented as:

$$FDR = \frac{\frac{\sum_{p=1}^{\pi} \left| \{ i : d_p(i) > du \ or \ d_p(i) < dl \} \right|}{\pi}}{|T|} \tag{4}$$

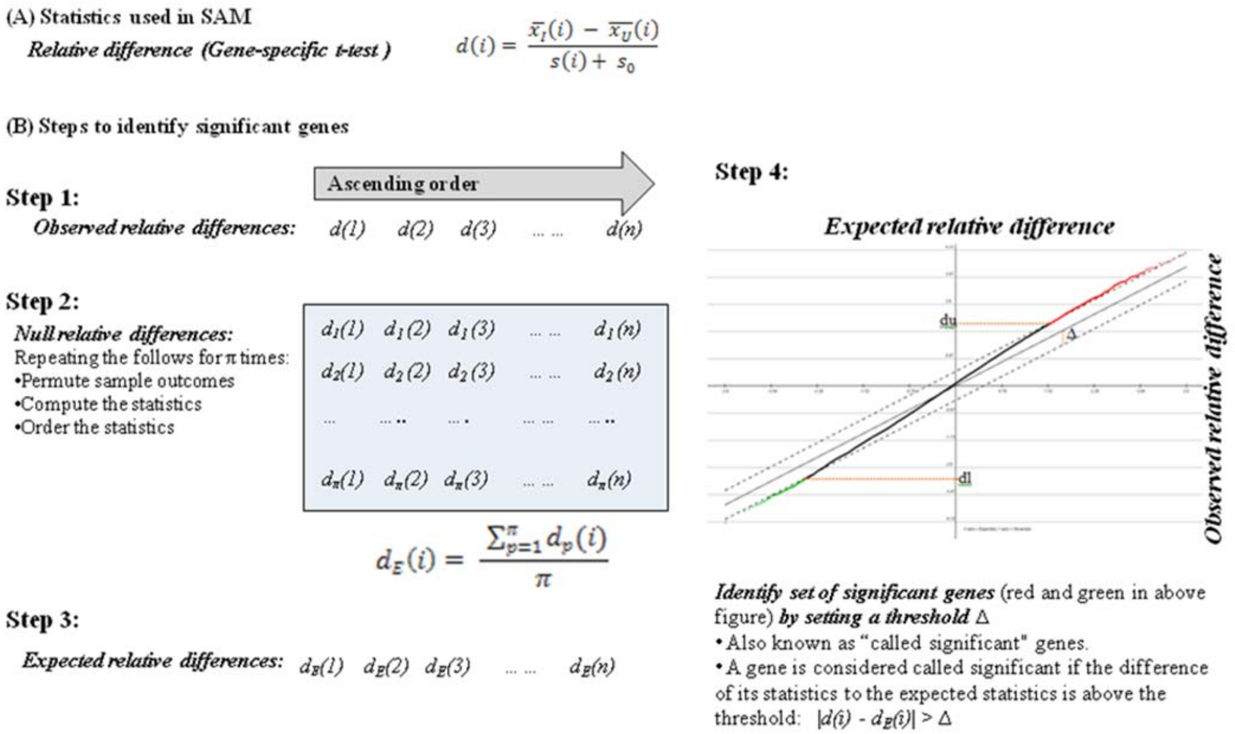

(A) Statistics used in SAM
Relative difference (Gene-specific t-test)
$d(i) = \frac{\overline{x_I}(i) - \overline{x_U}(i)}{s(i) + s_0}$
(B) Steps to identify significant genes
Step 1:
Observed relative differences: d(1) d(2) d(3) … … d(n)
Ascending order
Step 2:
Null relative differences:
Repeating the follows for π times:
•Permute sample outcomes
•Compute the statistics
•Order the statistics
$d_1(1)$ $d_1(2)$ $d_1(3)$ … … $d_1(n)$
$d_2(1)$ $d_2(2)$ $d_2(3)$ … … $d_2(n)$
$d_\pi(1)$ $d_\pi(2)$ $d_\pi(3)$ … … $d_\pi(n)$
$d_E(i) = \frac{\sum_{p=1}^{\pi} d_p(i)}{\pi}$
Step 3:
Expected relative differences: $d_E(1)$ $d_E(2)$ $d_E(3)$ … … $d_E(n)$
Step 4:
Expected relative difference
Observed relative difference
du
dl
Δ
Identify set of significant genes (red and green in above figure) by setting a threshold Δ
•Also known as "called significant" genes.
•A gene is considered called significant if the difference of its statistics to the expected statistics is above the threshold: $|d(i) - d_E(i)| > \Delta$

Figure 2.1. Procedures of SAM

In data analysis involved survival outcome, such as survival status after surgery, censored cases often occur often due to failure in follow up. Using SAM or other discrete-class learning methods, they need to be removed from the analysis as the exact survival outcome for these observations were not known. In microarray data analyses where observations are limited, it's important to include as many available observations as possible to strengthen the statistical power of the study. Therefore, methods that could include all observations yet accounting for censored cases would be preferred from methods that redefine the problem as binary class problem by removing censored cases. Statistical survival analysis methods such as Cox proportional hazard model would account for censored cases. Univariate Cox proportional hazard model analyzes how each gene changes relative to survival status. Gene that passed certain predefined significance threshold or ranked tops according to the statistical significance would be identified as interesting genes that are related to survival outcome [11, 15].

Cox proportional hazard model, or usually known as Cox model, is a regression model proposed by D.R. Cox [35]. It's commonly used in survival analysis to study the relationships between predictors (or known as covariates) and the survival outcome. In survival analysis, the hazard at time *t*, is the probability of an event (such as death) at time *t*, given survival up to time *t* [36], which can be defined as:

$$h(t) = \frac{\text{number of subjects experiencing the event at time } t}{\text{number of subjects at risk starting time } t} \tag{5}$$

In univariate Cox model, the hazard definition is extended to be proportional hazard, which is the probability of an event at time *t*, given survival up to time *t*, and for a specific value of a predictor, *x*:

$$h(t \mid x) = h_0(t) \times \exp(\beta \cdot x) \tag{6}$$

where $h_0(t)$ is known as baseline hazard function. It is the probability that subjects will experience the event when the predictor is zero. Hazards for observations of two survival states could be defined as:

$$h(t \mid x = x_1) = h_0(t) \times \exp(\beta \cdot x_1) \tag{7}$$

$$h(t \mid x = x_2) = h_0(t) \times \exp(\beta \cdot x_2) \tag{8}$$

Thus, the ratio of the two hazards is obtained by:

$$\begin{aligned} \frac{h(t \mid x = x_1)}{h(t \mid x = x_2)} &= \frac{h_0(t) \times \exp(\beta \cdot x_1)}{h_0(t) \times \exp(\beta \cdot x_2)} \\ &= \exp(\beta \cdot (x_1 - x_2)) \end{aligned} \tag{9}$$

Assume the change of the predictors is one unit; it gives us the estimated degree of effect of the predictor on survival, known as hazard ratio:

$$HR = \exp(\beta) \tag{10}$$

The statistical significance of the estimated hazard ratio is assessed by Wald test, under the hypothesis that the coefficient (β) is zero [37], which is analogous to the fact that the predictors has no effect on survival giving a *HR* of 1. Nevertheless, the number of genes is much larger than the number of samples available. The value indicates the significance of the genes ranked by univariate Cox model need to be corrected for multiple hypothesis testing.

2.1.3 Feature Selection Methods

Traditional feature selection methods used in machine learning applications are not commonly employed in genomic studies. Random forest is one of the feature selection methods used to select predictive genes in genomic studies [38-41]. Random forest uses both bagging and random variable selections in the algorithm to construct the ensemble of classification trees. Specifically, each of the classification trees is built using a bootstrap sample of the data, and each split of the tree is based on a random subset of the variables [42]. Random forest could be used for variable selection because in addition to classifications, random forest assesses the importance of each variable in the algorithm. The decrease in a tree splitting criterion, the Gini index and the decrease of permutation accuracy are implemented in random forest as measures to evaluate the importance of variables with respect to the outcome [43, 44]. The out-of-bag (OOB) error rate from the classification could be used as a criterion to select the final set of variables through an iteration random forests [38]. Since the tree split is based on random subset of variables on a bootstrapped sample, it enables random forest to work efficiently in microarray data where number of variables (genes) is much larger than the number of observations [38]. Random forest could select a smaller set of variables which could also achieve comparable prediction performance than other classifiers with larger set of variables [39, 41, 44], especially in data with large amount of noise. These properties of random forest are preferred in genomic studies with noisy high-throughput microarray data. One issue with random forest is the stability of the results given [38]. Multiple sets of selected variables that are equally good in classification performance would be produced by random forest. The lack of uniqueness and overlapping genes in the resulting selected genes will lead to questions on the biological interpretability of the results [45].

Principle Component Analysis (PCA) is another common method used to reduce the feature space dimensionality by transforming the data to a new coordinate system, or feature space. Each coordinate (called the principle component) is a linear combination of the original features [46]. The first few principle components yield the greatest variance present in all the original features and hence usually selected as the new feature for classification. One disadvantage of the projection method is that all of the original input features need to be retained [47]. It was proposed to perform gene selection through a variable selection strategy based on PCA [48]. A variation of PCA, such as the generalized and nonlinear kernel PCA (KPCA) was also proposed to reduce dimension of the microarray gene expression data prior to classification [49].

Relief is another method that could be used for feature selection because it would assess the importance of each variable in differentiating samples between two classes and provide the ranking accordingly. The first *Relief* algorithm was proposed by Kira and Rendell [50]. An extended version with more reliable probabilities estimation was later proposed by Kononenko et al. [51], known as *Relief-F*. In the extended version, instead of calculating the weight of features based on the nearest hit and miss of the randomly selected sample, *k*-nearest hits and *k*-nearest misses of the randomly selected sample are used. As depicted in Fig. 2.2, *Relief* evaluates the importance of a variable by repeatedly sampling an instance and checking the value of the given variable for the *k*-nearest instances from the same and different classes. The values of the variables of the nearest neighbors are compared to the sampled instance and used to update the relevance weights for each variable.

```
set all weights W[A] := 0.0;
for  i := 1 to n do
begin
    randomly select an instance R;
    find nearest hit H and nearest miss M;
    for A:= 1 to #all_attributes do
        $W[A] := W[A] - \mathit{diff}(A, R, H)/n + \mathit{diff}(A,R, M)/n$;
end;
```

Figure 2.2. The Relief algorithm [51].

An approximation of the weight of attribute A computed by *Relief* could be written as:

$$\begin{aligned} W[A] = {} & P(\textit{different value of } A \mid k\textit{ - nearest miss}) \\ & \text{- } P(\textit{different value of } A \mid k\textit{ - nearest hit}) \end{aligned} \tag{11}$$

When the algorithm stops, *Relief* assigns more weight to those variables that have the same value for instances from the same class and differentiate between instances from different classes [52, 53].

2.1.4 Discussion

Among the few ranking-based gene selection methods, *t*-test and fold change is the simplest and most intuitive technique in selecting genes differentially expressed with respect to disease outcome. However, it is sensitive to the gene-specific variances and the estimates suffer more in small sample [54, 55]. These gene-specific variances were adjusted in SAM through a new statistics measurement, the gene-specific *t*-test, or known as the relative difference [34]. Moreover, with *t*-test, multiple testing problem arises from the large number of genes in microarray data and poses high risk in false positive. To address to this problem, repeated measurements of statistics from permutations are used in SAM to control for amount of false positives. Nevertheless, the permutations cause SAM more computationally expensive than *t*-test. Instead of evaluating the differentiation of expression between two disease states, univariate Cox model evaluates the discriminative power of genes with respect to survival outcome, which is over a series of time points. This is the strength of univariate Cox model over *t*-test and SAM in studies involved survival data. Theoretically, SAM could be generalized to implement survival analysis method by defining the gene-specific *t*-test in a different way. There is no available tool for such implementation.

Small sample size is one of the key challenges in microarray data studies. To obtain less biased estimates from small sample, machine learning methods such as random forest and *Relief* employed randomization and repeated measurements. When sample size is small, bagging with bootstrapped aggregating was shown to be able to improve the performance of unstable estimators, such as the classification and regression tree (CART) [56]. Random forest employs bagging in the algorithm and have shown to provide good performance in microarray data with small samples and large variables [38]. Random forest is more flexible as it could be used to rank the genes by providing a variable importance measure for each gene and also could provide a subset of predictive genes through classifications. This presents a convenient framework for removing redundant genes. On the hand, *Relief* does not remove redundant features. In term of computational time, *Relief* is faster than random forest and other feature selection methods as it avoids any exhaustive or heuristic combinatorial search. The *Relief* algorithm is in linear time to the number of features and number of samples selected. Although *Relief* has been applied in a wide aspect of applications, ranging from feature selection before model constructions to provide feature importance guide in other algorithm [57], it is not commonly used in microarray studies.

2.2 Network-based Methods

2.2.1 Introduction

Fundamental mechanisms of molecular functions are based on interactions among genes and proteins [26, 58]. Therefore, it's important to understand the genes interaction networks in order to gain further insights to the relationships between genes and diseases. Molecular network analysis using computational network models has led to promising applications in identifying new disease genes [59], discovering disease-related sub-networks [60], and classifying diseases [61]. Computational network models that have been developed for molecular network analysis can be roughly categorized into three classes: logical model to demonstrate the state of entities (genes/proteins) at anytime as a discrete level; continuous models to represent real-valued molecular network processes and activities over continuous timescale; and single-molecule models to simulate small regulatory networks and mechanisms [62]. Since our studies involved implementing novel computational network models in biomarkers discovery using microarray data, a few logical network models commonly used for molecular network studies will be briefly discussed in this section.

2.2.2 Artificial Neural Networks

Artificial neural networks (ANNs) are computational models consist of set of highly interconnected nodes. The structure and functions of ANNs are modeled with the motivation from the biological neural systems, where each node in the ANNs portray the biological neurons [63]. ANNs are typically organized in layers of nodes, where each node interconnected with one another with a connected line. The lines represent the relationships between the nodes and each connection has an associated weight to describe the strength of the relationship. ANNs are generally described in three layers: input layer, hidden layer, and output layer (Fig 2.3). The input layer contains nodes to which the input is presented. The output layer is where the final predictions/ answers are retrieved. The hidden layer, which could contain more than one layer, is where the processing is done on the incoming data and feed the output to the next layer.

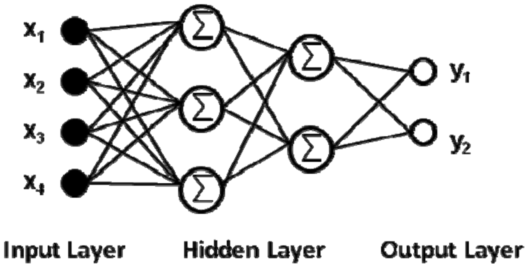


Figure 2.3. Structural diagram of a general artificial neural network.

To construct an ANN model representing the data, back-propagation algorithm is the most common algorithm used to learn the weights. Specifically, the algorithm starts with a random weight, then iterates through the training data set and updates the weights to reduce the error on each observation. The algorithm stops when weights converge, or when the error rate passes a certain threshold [63, 64]. Once the weights are learned, the modeled ANN could be used to obtain prediction on new input. An example of this process in software high-risk program detection is given in Fig. 2.4. The training phase of modeling an ANN is relatively time consuming, but the prediction phase is typically straightforward and fast [63]. The learning rate parameter in the back propagation algorithm could be tuned to control the speed of the training process. However, it would affect the generalization of the constructed ANN, leading to over-fitting the data or too generalized with low precision.

Artificial neural networks were first used in artificial intelligence applications to interpret complex real-world problems, such as speech synthesis, facial recognition, and handwriting recognition [63]. Nowadays, ANNs could be found in a wide range of applications, including applications in biomedical fields and microarray studies. A few examples include modeling classifier for diagnostic or prognostic prediction [65-68], learning and modeling interactions among genes from expression data [69] .

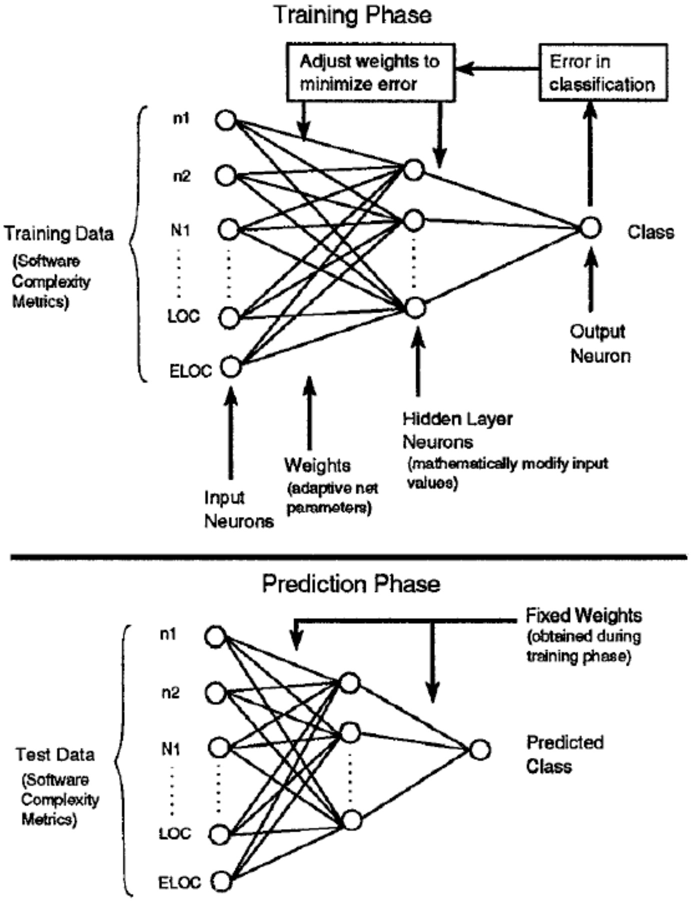

Training Phase
Adjust weights to minimize error
Error in classification
Training Data (Software Complexity Metrics)
n1
n2
N1
LOC
ELOC
Class
Output Neuron
Hidden Layer Neurons (mathematically modify input values)
Weights (adaptive net parameters)
Input Neurons
Prediction Phase
Fixed Weights (obtained during training phase)
Test Data (Software Complexity Metrics)
n1
n2
N1
LOC
ELOC
Predicted Class

Figure 2.4. Construction of artificial neural networks for high-risk programs detection. [64]

2.2.3 Bayesian Networks

Bayesian networks, which also known as Bayesian belief networks or belief networks, are graphical models used to represent the joint probability distributions of a set of random variables and the conditional dependence relationships among the variables based on Bayesian probability. Bayesian networks have been a popular framework for encoding uncertain knowledge in expert systems [70]. Bayesian networks had been applied and shown to be useful in various applications, ranging from manufacturing control, price forecasting, diagnosis, automated vision, to bioinformatics [71-74] .

Bayesian network is a directed acyclic graph (DAG) where the nodes representing random variables and edges representing direct relationships between the connected variables. In the Bayesian network, each node has an associated conditional probability table (CPT) denoting the conditional probabilities of the node given all possible combinations of its parents. For nodes without parent, prior probability of the node is specified [75]. Markov assumptions on conditional independence among variables are hold in Bayesian networks. Markov assumptions state that variable X_i is considered independent of its non-descendants, given its parents and joint distribution could be decomposed into the product form [76]. For a given set of random variables $X = \{X_1, \ldots, X_n\}$, the joint distributions represented by the Bayesian networks could thus be defined as:

$$P(X_1,\ldots,X_n\} = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}^G(X_i)) \tag{12}$$

where $\mathrm{Pa}^G(X_i)$ is the set of parents of X_i in the Bayesian network. Fig. 2.5 gives an example of the Bayesian network with five variables.

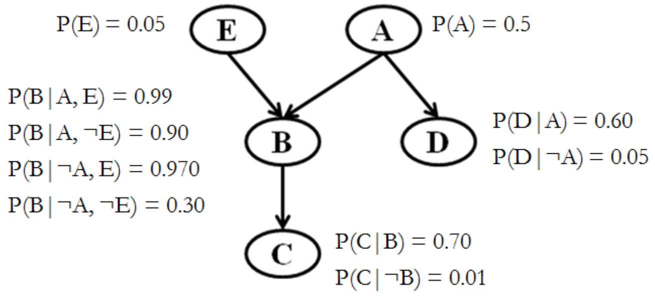


Figure 2.5. An example of Bayesian network structure.

The Markov independencies among the variable in the Bayesian network in Fig 2.5 are $I(A;E)$, $I(B;D \mid A,E)$, $I(C; A,D,E|B)$, $I(D;B,C,E|A)$, and $I(E;A,D)$. The joint distribution of the five variables specified by the Bayesian network is:

$$P(A,B,C,D,E) = P(A) * P(B \mid A,E) * P(C \mid D) * P(D \mid A) * P(E) \tag{13}$$

A scoring function is used to evaluate how well a built Bayesian network matches the data. The score computed could be used as the criterion in the Bayesian network that best represent the probability distributions of the attributes in the data. Given a sufficiently large number of instances, the learning procedure will converge and lead to the exact network structure up to the correct equivalence class [77]. This process of searching for the optimal network in the space of directed acyclic graphs is a NP-hard problem. Multiple search algorithms such as hill-climbing, beam search, or simulated annealing could be used to search for the optimal network. Although these methods provide only suboptimal solution, in which only the local maximal Bayesian network is obtained, it had been shown to give good performance in practice. Another approach for more efficient learning process is sparse candidate algorithm [77], in which a subset of variables was chosen as the set of candidate parents and the search was restricted to networks in which the candidate parents of a variable can be its parents.

Bayesian networks could be used to interpret causal relationships among variables by imposing more stringent interpretation of the edges: the parents of a variable are its immediate causes. In the causal interpretation, variable is considered independent of its earlier causes, given the values of its parents. This is known as the causal Markov assumptions. This causal interpretation of Bayesian networks is a natural interpretation for biological models. For example, in genetic pedigree: once we know the genetic makeup of the individual's parents, the genetic makeup of her ancestors is not informative about her own genetic makeup. [76].

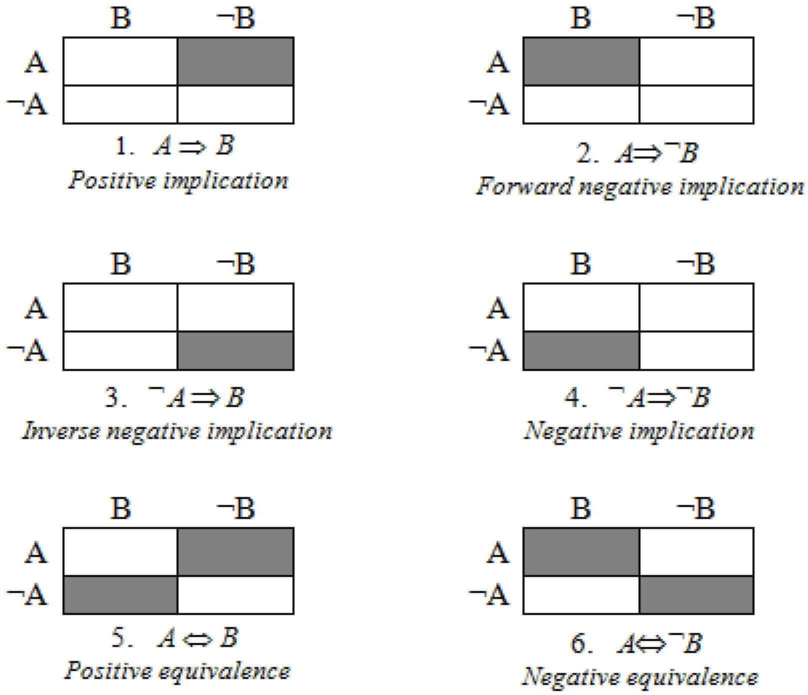
2.2.4 Implication Networks

Similar as Bayesian networks described in the last section, implication networks are also probabilistic graphical models representing the relationships among the variables. In the implication network, each node represents a variable and the edge between pair of nodes represents the type of implications existing between the pair of variables. Instead of acyclic as in

Bayesian networks, implication networks allow cyclic relation, which is an important property over Bayesian networks for biological networks studies.

The first formalism of implication networks was proposed by Liu and Desmarais [78], which is based on binomial distribution. This formalism had not been applied to any aspect of biological studies. Another formalism of implication networks based on prediction logic was proposed by Guo et al. [1], where prediction logic based on formal logic rules was used to derive successful implication relations.

There exist six implication relations between any pair of dichotomous variables (Fig. 2.6).



1. $A \Rightarrow B$
Positive implication

2. $A \Rightarrow \neg B$
Forward negative implication

3. $\neg A \Rightarrow B$
Inverse negative implication

4. $\neg A \Rightarrow \neg B$
Negative implication

5. $A \Leftrightarrow B$
Positive equivalence

6. $A \Leftrightarrow \neg B$
Negative equivalence

Figure 2.6. Six most important implication rules relating two dichotomous variables.

Each table in Fig 2.6 is a contingency table (Table 2.1) where each cell represents the number of co-occurrences. For example, cell $N_{A \wedge B}$ indicates the number of samples where both variables A and variable B are true. The shaded cells of the contingency tables in Fig. 2.6 represent the errors for the corresponding implication rule. For example, $A \wedge \neg B$ is the error cell for the implication rule $A \Rightarrow B$, $N_{A \wedge \neg B}$ represents the number of error occurrences. Cell $A \wedge \neg B$ is erroneous for the rule $A \Rightarrow B$ because in an ideal case, if the implication $A \Rightarrow B$ is the true relationships between *A* and *B*, then we would never expect to find the contradiction case where *A* is true but not *B*.

Table 2.1. Contingency table of two variables for *N* empirical samples.

	B	¬B
A	$N_{A \wedge B}$	$N_{A \wedge \neg B}$
¬A	$N_{\neg A \wedge B}$	$N_{\neg A \wedge \neg B}$

To derive the implication relation between each pair of variables in the dataset, a modified *U*-Optimality method [79] was used in the implication induction algorithm (Fig. 2.7).

```
The Implication Induction Algorithm by Guo et al. [1]
Begin
  Set a significant level $\nabla_{min}$ and a minimal $U_{min}$
  For $node_i$, $i \in [0, v_{max} - 1]$ and $node_j$, $j \in [i+1, v_{max}]$
    (Note: $v_{max}$ is the total number of nodes)
   For all empirical case samples $N$
    Compute a contingency table as in Fig. 2.6
          $M_{ij} = \boxed{\begin{matrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{matrix}}$
   For each relation type $k$ out of the six cases, find the solution
                         $Max\ U_p$
     Subject to          $Max\ U_p \geq U_{min}$
                         $\nabla p \geq \nabla_{min}$
                  $\nabla_{\text{error cells}} > \nabla_{\text{non-error cells}}$
   If the solution exists, then return a type $k$ relation
End
```

Figure 2.7. Implication induction algorithm based on prediction logic.

In the contingency table M_{ij} of the induction algorithm (Fig. 2.7), N_{11} indicates number of samples where both *i* and *j* occur to be true, N_{12} is when *i* is true but not *j*, N_{21} is when *j* is true but not *i*, and N_{22} is when both *i* and *j* are not true.

In the induction algorithm, U_p is the scope of the implication rule, representing the portion of the data covered by the implication relation, and ∇p is the precision of the implication rule, representing the prediction success of the corresponding implication relation. For a single error cell, where N_{ij} is the number of error occurrences, scope U_p, and precision ∇_p are defined as:

$$Up = Uij = \frac{Ni.*N.j}{N^2} \tag{14}$$

$$\nabla_p = \nabla_{ij} = 1 - \frac{N_{ij}}{N * U_p} \tag{15}$$

For the rule types where there are multiple error cells, they are defined as:

$$U_p = \sum_i \sum_j \omega_{ij} * U_{ij} \tag{16}$$

$$\nabla_p = \sum_i \sum_j \left(\frac{\omega_{ij} * U_{ij}}{U_p} \right) \nabla_{ij} \tag{17}$$

where $\omega_{ij} = 1$ for error cells; otherwise, $\omega_{ij} = 0$.

Based on the contingency table for variable A and B (M_{AB}) (Table 2.1), the scope and precision for each of the six implication rules in Fig. 2.6 are defined as follows.

For positive implication, $A \Rightarrow B$,

$$U_{A \Rightarrow B} = U_{A \wedge \neg B} = \frac{N_A * N_{\neg B}}{N^2} \tag{18}$$

$$\nabla_{A \Rightarrow B} = \nabla_{A \wedge \neg B} = 1 - \frac{N_{A \wedge \neg B}}{N * U_{A \Rightarrow B}} \tag{19}$$

Similarly, for forward negative implication, $A \Rightarrow \neg B$,

$$U_{A \Rightarrow \neg B} = U_{A \wedge B} = \frac{N_A * N_B}{N^2} \tag{20}$$

$$\nabla_{A \Rightarrow \neg B} = \nabla_{A \wedge B} = 1 - \frac{N_{A \wedge B}}{N * U_{A \Rightarrow \neg B}} \tag{21}$$

For inverse negative implication, $\neg A \Rightarrow B$,

$$U_{\neg A \Rightarrow B} = U_{\neg A \wedge \neg B} = \frac{N_{\neg A} * N_{\neg B}}{N^2} \tag{22}$$

$$\nabla_{\neg A \Rightarrow B} = \nabla_{\neg A \wedge \neg B} = 1 - \frac{N_{\neg A \wedge \neg B}}{N * U_{\neg A \Rightarrow B}} \tag{23}$$

For negative implication, $\neg A \Rightarrow \neg B$,

$$U_{\neg A \Rightarrow \neg B} = U_{\neg A \wedge B} = \frac{N_{\neg A} * N_B}{N^2} \tag{24}$$

$$\nabla_{\neg A \Rightarrow \neg B} = \nabla_{\neg A \wedge B} = 1 - \frac{N_{\neg A \wedge B}}{N * U_{\neg A \Rightarrow \neg B}} \tag{25}$$

For positive equivalence, $A \Leftrightarrow B$,

$$U_{A \Leftrightarrow B} = U_{A \wedge \neg B} + U_{\neg A \wedge B} = \frac{N_A * N_{\neg B} + N_{\neg A} * N_B}{N^2} \tag{26}$$

$$\nabla_{A \Leftrightarrow B} = 1 - \frac{N_{A \wedge \neg B} + N_{\neg A \wedge B}}{N_A * N_{\neg B} + N_{\neg A} * N_B} * N \tag{27}$$

And for negative equivalence, $A \Leftrightarrow \neg B$,

$$U_{A \Leftrightarrow \neg B} = U_{A \wedge B} + U_{\neg A \wedge \neg B} = \frac{N_A * N_B + N_{\neg A} * N_{\neg B}}{N^2} \tag{28}$$

$$\nabla_{A \Leftrightarrow \neg B} = 1 - \frac{N_{A \wedge B} + N_{\neg A \wedge \neg B}}{N_A * N_B + N_{\neg A} * N_{\neg B}} * N \tag{29}$$

In the implication induction algorithm, the minimum requirement for the scope (U_{min}) and precision (∇_{min}) must be positive values for an implication rule. They are the parameters used to control the significance level for an implication rule. In our studies, we defined the minimum requirement for these two parameters to be at least 95% significant ($P < 0.05$) from one-tail Z-test based on the sample size. In this induction algorithm (Fig. 2.7), the minimum requirements for deriving an implication rule are set for both scope and precision, which is different from the original *U*-Optimality [79] method, where the minimum requirement is set for precision alone.

An implication rule has high precision when the number of error occurrences is a small portion of the data covered by the implication rule. An implication rule is successfully derived from the algorithm if it has the maximum scope, U_p and it satisfies the constraint that its scope ($U_{p)}$ and precision (∇p) are greater than the required minimum values, U_{min} and ∇_{min} , respectively. To simplify the computations of the maximization problem, the precision ∇_{ij} value of every error cell must be greater than that of the non-error cells for the corresponding implication rule [1].

The complexity of the induction algorithm is $O(Nv^2)$, where *N* is the sample size and *v* is the number of variables in the dataset (i.e. nodes in the implication networks) [1].

To represent the strength of the implication relation for the connecting pair of variables, a weight is estimated based on conditional probability. Since each implication rule has a logically equivalent rule, another weight for the corresponding logical equivalence should be estimated. Both weights could be derived at the same time by the induction algorithm [1]. Let W_I be the weight associated with the implication rule, W_I' is the weight associated with its logical equivalence. For example, for implication rule $A \Rightarrow B$, its logical equivalence is $\neg B \Rightarrow \neg A$. Their respective weights are defined as:

$$W_I = \frac{N_{A \wedge B}}{N_{A \wedge B} + N_{A \wedge \neg B}} \tag{30}$$

$$W_I' = \frac{N_{\neg A \wedge \neg B}}{N_{A \wedge \neg B} + N_{\neg A \wedge \neg B}} \tag{31}$$

Given a quintuple representing the implication rule *I*:

$$I \in \ddot{I}, I =< R, N_{ant}, N_{con}, W_I, W_I' > \tag{32}$$

where $\ddot{I}$ represents the set of all possible implication rules, *R* represents the implication rule type, W_I and W_I' are the weight functions mapping the antecedent node N_{ant} and consequent node N_{con} of the implication rule and their negations to the real number between 0 and 1, which are defined as:

$$W_I : N_{ant} \times N_{con} \to [0,1] \tag{33}$$

$$W_I^{'} : \neg N_{ant} \times \neg N_{con} \to [0,1] \tag{34}$$

The formalism of implication networks based on prediction logic integrates formal logic theory and statistics, which provides conceptual value of prediction analysis in constructing and evaluating useful statements, particularly in complex multinomial problems with moderate sample sizes [1]. This feature is essential for clinical applications, in which many clinical parameters are multinomial and the patient sample size is small. The implication induction algorithm in Fig. 2.7 is general for discrete datasets. With the expansion of the contingency table M_{ij}, implication rules can be induced for multinomial datasets, where error cells are those with the highest precision (∇_{ij} values) and satisfying all the constraints. The proposition can then be induced according to the error set.

2.2.5 Boolean Implication Networks

Recently, another formalism of implication networks, Boolean implication networks were constructed to model gene interactions networks in a meta-analysis of microarray data for multiple species [80]. The implication relations in the Boolean implication networks were induced based on scatter plots of expression between two genes. On the scatter plots of gene expressions, a threshold was automatically determined using StepMiner algorithm [81] to discretize the gene expression level as ‘high’ or ‘low’. Based on the discretized levels, the scatter plot is partitioned into four quadrants and the implication relation between the two genes is derived based on the number of data points (occurrences) in the quadrants. The partitioned scatter plot with four quadrants is analogous to the contingency tables in Fig. 2.6 and Table 2.1, where the ‘low’ and ‘high’ expression of gene A corresponds to $\neg A$ and A respectively. In order to derive a successful implication rule between the pair of genes for the Boolean implication networks, two statistics were tested. The first statistic tests if the observed number of occurrences in the sparse quadrant (error cell) is significantly less than the expected number of occurrences under an independent model, given the relative distribution of low and high values of both genes. The second statistic estimates the maximum likelihood of the error rate for the number of occurrences in the error cell. For example, if the error cell for genes *A* and *B* is where both *A* and *B* is low, the observed and expected number of occurrences in the error cell is:

$$obsreved = N_{\neg A \wedge \neg B} \quad (35)$$

$$expected = (\frac{N_{\neg A}}{N} * \frac{N_{\neg B}}{N}) * N = \frac{N_{\neg A} * N_{\neg B}}{N} \quad (36)$$

The first statistics and the error rate are thus defined as:

$$statistic = \frac{expected\text{ - }observed}{\sqrt{expected}} \quad (37)$$

$$error\ rate = 1/2 * (\frac{observed}{N_{\neg A}} + \frac{observed}{N_{\neg B}}) \quad (38)$$

An implication rule representing the pair of genes is successfully derived if the statistic in equation (37) is greater than 3 and the error rate is less than 0.1.

2.2.6 Discussion

The four network models reviewed above falls into the first class of computational models for regulatory network analysis [62], which are logical models. Logical models are suitable for modeling the interactions among genes from microarray data as they require the least amount of data compared with other network models, such as single-molecule network models. Although the logical models are abstract and could only provide qualitative insight to the interactions, they are simpler to be studied. Other benefits shared by logical models reviewed above are that they provide good performance in learning from noisy data and provide a framework for inferring predictions [63, 82, 83].

There are a few limitations with ANNs. The first argument about ANNs is that the modeled networks are overly complex and difficult to be interpreted. It is like a “black box” where the weights learned are hard to be understood by humans compared with other rule-based classifiers [63, 84]. Although it is computationally difficult, the knowledge about the weights learned could be retrieved in the form of Causal Indices (CI) [84]. Another shortcoming about ANNs is that they will over-fit the training data and generalize to new data poorly when the sample size is small. Furthermore, as discussed earlier, it is time consuming in training the weights of the networks. This makes modeling genome-wide gene interactions with ANNs computationally challenging. To our knowledge, there are no applications for the complete modeling for gene-gene interactions for the whole human genome. On the other hand, unlike implication networks where the gene expression values would need to be discretized into binary scale, ANNs could model the interactions of genes at continuous form.

Bayesian network is more commonly known and preferred in molecular network analysis [58, 73, 74]. Bayesian networks are preferred because it could provide causal relationships between pair of genes. More importantly, the noise inherent to biological data could be accommodated by the probabilistic nature of the formalism of Bayesian networks [74]. A causal interpretation for Bayesian networks had been utilized to predict genome-wide protein-protein interactions [73] and model cellular networks [72]. However, it is not viable to evaluate all possible networks as the number of possible networks grows exponentially in the number of genes under consideration. Furthermore, owing to the Markov assumption hold in Bayesian networks, it is not always possible to determine the causal relationships between nodes, i.e., the direction of the edges [85]. More importantly, the acyclic Bayesian network structure was unable to model feedback loops, which are essential in signaling pathways [74] and genetic networks [86-88]. To overcome the acyclic limitation, a more complex scheme, dynamic Bayesian networks, was explored for modeling temporal microarray data [89, 90]. On the other hand, implication networks could model cyclic relations. Therefore, the cyclic implication network is more suitable for studying relationships and interactions of biological networks than Bayesian networks.

Implication networks and Bayesian networks are both belief networks formalized based on statistics derived from data. Implication networks are not as commonly known in the research domain as Bayesian networks. The latest applications of implication networks in the research domain are the Boolean implication networks, which are proposed as a computational platform for genomic evolution of genes interactions and discovery of novel biological relations among genes [80]. It was shown that implication networks is computationally efficient and feasible to be applied to construct genome-wide networks [80]. Moreover, implications networks are suitable for genes networks representation because both the symmetric and asymmetric relationships between pair of genes could be represented with the six implication rules [80], where asymmetric relationships could be represented by the first four implication rules ($A \Rightarrow B$, $A \Rightarrow \neg B$, $\neg A \Rightarrow B$, and $\neg A \Rightarrow \neg B$); symmetric relationships could be represented by positive equivalence ($A \Leftrightarrow B$) and negative equivalence ($A \Leftrightarrow \neg B$).

2.3 Regularized Linear Models

2.3.1 Introduction

In general context, linear models are used to study the effects of multiple factors on the response variable or used to construct a prediction model. In microarray studies, linear models such as ANOVA or ordinary least square (OLS) linear regression models were used to analyze gene expression changes or to construct classification models [91, 92]. In this section, we will briefly review the general properties of linear models and their shortcomings in microarray analyses through descriptions of two specific regularized linear regression models.

A few properties were desired in linear models for genomic studies. The first property is to have a good fit in the modeling data but also accurate prediction in new observed data. In fitting the regression model, when the number of predictors is relatively large, the fitted models will tempt to overfit the data available but predict poorly in new observed data. The curse of dimensionality phenomenon with the large p (number of predictors) small n (number of samples) found in microarray data not only posts a challenge in fitting the models but also makes the over fitting problem worse [93]. Shrinkage methods were recommended to avoid the overfitting problem in fitting the regression model in small data sets [94]. The second property desired is to select the whole group of genes sharing the same biological pathway instead of individual genes [95]. Furthermore, since most genomic studies involve construction of patient classification model using the set of genes selected, it is desired for linear models to have a property where the gene selection method is built into the classification procedure. In this section, lasso and elastic net, the two regression-based methods with these properties, will be discussed.

2.3.2 Lasso

Lasso (least absolute shrinkage and selection operator) is a linear regression method proposed by Tibshirani [96] as a regularization for OLS linear regression. Linear regression model is a model used to obtain a predicted response $\hat{y}$ with liner combinations of p predictors $x_1, \ldots , x_p$, which could be formulated as:

$$\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + ... + x_p\hat{\beta}_p \tag{39}$$

The model fitting procedures produce the vector of estimated coefficients $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$.

Suppose that we have a data of n samples: (x^i, y_i), $i = 1, 2, \ldots, n$, where $x^i = (x_{i1}, \ldots, x_{ip})^T$ are the predictors and y_i are the responses. OLS linear regression estimates the coefficients by minimizing the residual squared error:

$$\hat{\beta}^{OLS} = \arg\min_{\beta}\left[\sum_{i=1}^{n}(y_i - \sum_j \beta_j x_{ij})^2\right] \tag{40}$$

OLS often fits the given data well but performs poorly in future data. Thus, alternative methods for the coefficient estimations were desired. Lasso was proposed as an alternative estimation method through regularizing the OLS regression model by adding a L$_1$-norm penalty. In other words, lasso estimates the coefficients by minimizing the residuals sum of squares subject to a bound on the L$_1$-norm of the coefficients:

$$\hat{\beta}^{lasso} = \arg\min_{\beta}\left[\sum_{i=1}^{n}(y_i - \sum_j \beta_j x_{ij})^2\right], \text{subject to } \sum_j |\beta_j| \le t \ , (t \ge 0) \tag{41}$$

which is equivalent to solving the following problem:

$$\min_{\beta}\left[\sum_{i=1}^{n}(y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|\right] \tag{42}$$

Due to the L$_1$-norm constraint, lasso will assign a zero coefficient to some of the variables, causing these variables being “dropped out” from the regression model automatically. Thus, lasso could also be treated as a method for variable selection.

The criterion $\sum_{i=1}^{n}(y_i - \sum_j \beta_j x_{ij})^2$ in equation (41) is equivalent to the quadratic equation $(\beta - \hat{\beta}^{OLS})^T X^T X(\beta - \hat{\beta}^{OLS})$ that forms elliptical contours centered around $\hat{\beta}^{OLS}$. Take an example with two features as shown in Fig. 2.8A. The L1-norm constraint $|\beta_1|+|\beta_2|$ forms the rotated square. The solution to the above equation is the first point where the contours meet the rotated square. When it happens at the corner, the corresponding coefficient will be zero (β_2 in this example), which is analogous to dropping the corresponding variable out of the model, providing a mechanism for automatic variable selection. In practice, one can tune the parameter *t* in order for the contours to meet the constraint. As a comparison, the L2-norm penalty employed in the ridge regression could not provide variable selection mechanism because the constraint ${\beta_1}^2+{\beta_2}^2$ forms a circle instead of a square (Fig. 2.8B), in which there is no corner for the contours to hit and hence zero coefficients will rarely occur [96].

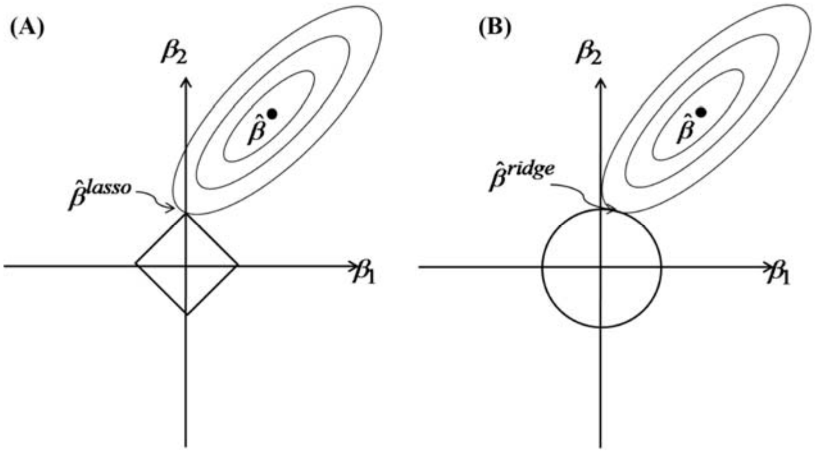


Figure 2.8. Geometry of the coefficient estimation for (A) lasso and (B) ridge regression.

2.3.3 Elastic Net

Elastic net proposed by Zou and Hastie [97] is another regularized regression model that could be used for variable selection by taking the advantage of both lasso and ridge regression. In elastic net, regularization is done through the combination of L1-norm and L2-norm. The coefficients are estimated by minimizing the residuals sum of squares subject to a bound on a function with L1- and L2-norm of the coefficients:

$$\hat{\beta}^{naiveEN}=\arg\min_{\beta}\left[\sum_{i=1}^{n}(y_i-\sum_{j}\beta_j x_{ij})^2+\lambda_1\sum_{j}|\beta_j|+\lambda_2\sum_{\mathrm{j}}{\beta_{\mathrm{j}}}^2\right] \tag{43}$$

which is equivalent to:

$$\hat{\beta}^{naiveEN} = \underset{\beta}{\arg\min} \left[\sum_{i=1}^{n} (y_i - \sum_{j} \beta_j x_{ij})^2 \right],$$
$$subject\ to\ (1 - \alpha) \sum_{j} \left| \beta_j \right| + \alpha \sum_{j} \beta_j^{\ 2} \le t,\ \ with\ \ \alpha = \frac{\lambda_2}{\lambda_2 + \lambda_1} \tag{44}$$

The combined constraints for elastic net (at α = 0.5) forms a diamond circle as compared with the rotating square by lasso and circle by ridge regression in the two-dimensional example (Fig. 2.9).

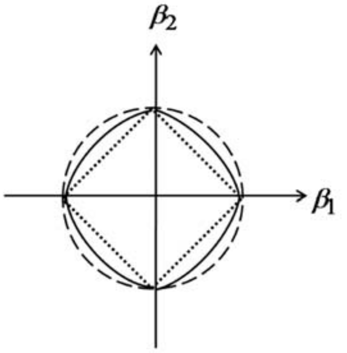


Figure 2.9. Two-dimensional contour plot of elastic net penalty at α=0.5 (----) as compared with lasso penalty (…..) and ridge penalty(-- -- --).

The addition of the L$_2$-norm penalty provides grouping effects in addition to the variable selection feature provided by the L$_1$-norm penalty. In lasso, if multiple variables are closely important and highly correlated to one another, only one of them will be selected and the other will be dropped out from the model. This is inappropriate for genomic studies as the groups of genes sharing the same biological pathway are desired to be identified together as they all function as a whole. However, the problem of multicollinearity phenomenon in regression exists as the groups of genes sharing same biological pathway are highly correlated to one another. With the L$_2$-norm penalty, the group of highly related variables will all be retained in the model with a more stable estimation; while the unimportant variables will be dropped out from the model with the L$_1$-norm penalty. With the combined advantages from both penalties, elastic net is preferred than lasso in genetic studies for variable selections, such as identification of multiple genetic variants from the whole genome [98].

Empirical studies showed that the estimation done in obtaining the $\hat{\beta}^{naiveEN}$ from equation (44) incurs double shrinkage [97]. Double shrinkage causes the estimation to perform well only when the problem is close to either lasso or ridge regression. Thus, it's known as the naïve elastic net. In order to correct the double shrinkage problem, a more stable estimation for elastic net is obtained by rescaling the naïve elastic net coefficients:

$$\hat{\beta}^{EN} = (1+\lambda_2)\hat{\beta}^{naiveEN} \tag{45}$$

2.3.4 Discussion

Between the two regularized linear models reviewed, lasso is simpler to be interpreted compared to elastic net and computationally lighter. However, lasso performs poorly in data with high collinearity [99] and select only one out of the group of genes sharing the same biological pathway. Although elastic net is more complex, is has all the three properties desired for linear models in genomic studies introduced at the beginning of the section. Elastic net performs better than lasso and provide groups mechanism leading to selection of genes sharing the same pathways. The grouping effects of elastic net could sometime turn into drawback of the method as it would lead to selection of highly redundant genes and incapable of providing small subset of predictive genes.

2.4 Critiques of Gene Selection Methods and Our Proposed Frameworks

Among the various gene selection methods discussed in the previous three sections, ranking-based gene selection methods are the easiest to be implemented. Despite the efficient computation time and scalability to the large dimension, ranking-based methods evaluate genes in isolation, without considering the effect of interactions among genes. Regularized linear model could capture the interaction effects among genes. However, the interactions limited to linear relationships. Since genes function through a series of complex interactions, the non-linear network models would be more appropriate to model the biological networks, such as gene regulations and genetic pathways. Compared with network-based methods, regularized linear models provide better computational complexity; but more computationally expensive than ranking-based methods.

Implication networks are computationally efficient and easier to be interpreted than ANNs. Most importantly, implication networks allow cyclic relations. This property makes implication networks more appropriate than Bayesian networks in studying genomic networks where feedback loop is common. Nonetheless, implication networks are more computationally heavier than ranking-based methods.

In summary, ranking-based methods are simple, scalable but do not account for true interconnecting nature among genes. On the other hand, network-based methods could provide a closer representation of the true phenomenon among genes but are much more computationally intensive. Therefore, we proposed combinatorial framework with different gene selection methods in multiple stages in order to systematically exploit the benefits while avoiding the drawbacks of various methods for novel gene signatures identification and better molecular prognosis. The first proposed combinatorial framework employed a combination of traditional statistics and feature selection methods. These methods include *t*-test, SAM, and *Relief* feature selection. The second proposed framework was built upon novel computational network models, i.e., implication networks. Implication networks efficiently model genome-wide coexpression networks and allow us to utilize signaling pathways for identifying prognostic genes signatures.

Survival prediction classifiers were constructed using the identified biomarkers to predict clinical outcome in individual patients. Prognostic performance of the classifiers is evaluated using statistical methods such as Kaplan-Meier (KM) analysis and concordance probability of estimate (CPE). Topological structure of the coexpression networks derived from implication networks were evaluated and confirmed with reported molecular interactions in literature. Bioinformatics tools were used to validate the clinical and biological aspects of the computational findings. These methods and tools will be introduced in the next section.

2.5 Validation Methods and Tools

2.5.1 Introduction

A common objective shared by our works and studies is to apply the findings for clinical use in the future, such as for disease prognosis. In order to apply the findings for clinical use, we should confirm if the computational findings agree to the true biological phenomenon. Due to the high cost and risk involves in in-vivo studies, the computational findings would be first evaluated using statistical methods, bioinformatics tools, or information obtained from genomic databases before the in-vivo studies. In this section, we will discuss the methods used to evaluate the performance of the prognostic model, the interactions revealed by computational network models, and the biological relevance of the gene sets identified.

2.5.2 Prognostic Evaluation

To evaluate the clinical value of the gene signature identified and the survival prediction model construct, statistical methods such as Kaplan-Meier (KM) survival analysis and concordance probability estimate (CPE) would be used to validate if the prediction obtained agreed to the true survival outcome.

Kaplan-Meier (KM) survival analysis is a non-parametric statistical method used to estimate a survival function from lifetime data [100]. In KM analysis, the survival function is estimated based on the life-table of the data, where survival time of patients could be of different length [101]. In collecting clinical outcomes from patients, it is hard to have stringent control on patients over time and therefore some patients who are valid at the beginning of the study would become invalid from one particular time onwards. For example, contacts with certain patients were lost before their death. These patient samples would be considered censored cases. Therefore, analysis methods that could consider censoring samples over the time series are important in survival analysis. Table 2.2 gives an example of the life- table.

Table 2.2. An example of life-table.

No. of months in study	No. of patients at risk	No. of patients who died	No. of patients censored
0-5	10	0	0
5-12	10	1	1
12-15	8	1	2
16-20	5	2	0
21-30	3	1	0

KM analysis is also known as the product limit estimator. Based on the life-table, the survival probability of each interval is estimated as the ratio of number of survival patients over number of patients at risk. If n_i is the number of patients at risk just prior to time t_i and d_i is the number of patients died at time t_i, the KM estimate of survival at time t is the non-parametric maximum likelihood estimate $S(t)$, which is the product of survival probability of intervals prior to time t:

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i} \tag{46}$$

The plot of the estimated survival function is a series of declining horizontal steps. An example of KM curve of the estimated based on life table in Table 2.2 is shown in Fig 2.10.

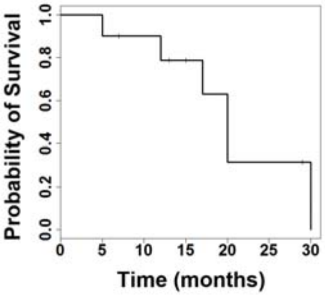


Figure 2.10. Example of Kaplan-Meier curve estimated based on life-table in Table 2.2.

When comparing survival functions of multiple groups estimated from KM analysis, the Mantel-Cox log-rank test could be used to evaluate the statistical significance between the survival curves for different groups [101]. For example, if the prognostic classifier predicts patients into two groups, a KM analysis could be used to estimate the survival function of each predicted group and log-rank test is used to test if the two groups are significantly different in terms of survival. Fig 2.11 give an example of comparing two KM curves.

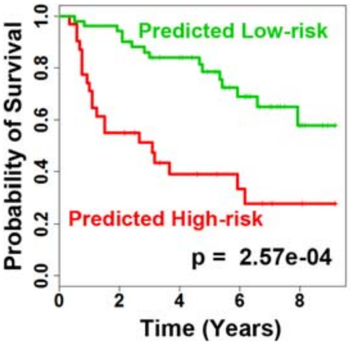


Figure 2.11. An example of comparing two Kaplan-Meier curves.

In general, concordance probability (CPE) is used to evaluate how the predicted outcomes of a nonlinear statistical model agreed with the actual outcomes. Concordance probability of a pair of bivariate observations (X_1, T_1) and (X_2, T_2) is thus defined as:

$$K_{X,T} = P(T_2 > T_1 \mid X_2 \geq X_1) \quad (47)$$

In our studies, we used Cox proportional hazard model to estimate the risk scores of each subject. Therefore, we would like to evaluate how the risk scores obtained from our model agreed with the actual survival outcomes of patient samples. In order to evaluate how concordant the risk scores estimated is to the actual survival outcomes, the CPE proposed by Gonen and Heller could be used [102]. This estimation is focused on Cox model and is defined as:

$$K(\beta) = P(T_2 > T_1 \mid \beta^T x_1 \geq \beta^T x_2) \quad (48)$$

where T is the response variable (the actual survival outcomes of patient samples) and $\beta^T x$ corresponds to risk scores obtained from the Cox model.

In their estimation, partial likelihood estimator $\hat{\beta}$ is used to substitute β and the empirical distribution of $\beta^T x$ is used to represent the distribution of risk scores. To resolve the asymptotic nature of the Cox partial likelihood estimator, a kernel function is used for smoothing. The final estimator used in obtaining the concordance probability of the model obtained would be purely based on the regression coefficients and covariates from Cox model, without the patients' survival time and outcomes. Therefore, this estimation is not sensitive to the censoring cases in the patient samples.

If the CPE obtained is close to 0.5, it indicates that model has poor predictive performance on the actual survival outcome (it's as good as the random chance of tossing the coin). The model showed better predictive performance when the CPE is approaching closer to 1.

2.5.3 Gene Coexpression Networks Assessment

The gene coexpression networks derived from the implication networks are evaluated on precision, false discovery rate (*FDR*), and stability. Precision and false discovery rates evaluate the biological relevance of the derived coexpression networks. Stability examines if the derived coexpression relations were stable or unpredictable.

Five gene set collections (positional, curated, motif, computational, and Gene Oncology) and canonical pathway databases from the MSigDB[1] were used to evaluate the precision and *FDR* of the derived coexpression networks. A coexpression relation was considered a true positive (TP) if the pair of genes belongs to the same gene set or pathway in any investigated database. If a pair of genes does not share any gene set or pathway, the coexpression relation was considered a false positive (FP). A coexpression relation was labeled as non-discriminatory (ND) if at least one gene in the pair is not annotated in a database [103]. Coexpression relations labeled as ND were excluded in the evaluation as they were not confirmed.

Precision and *q*-value of the coexpression networks are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{49}$$

$$q - value = \frac{FP}{TP + FP} \tag{50}$$

To generate the null distributions of precisions and *q*-values, class labels of patient samples in the test data were randomly permuted for 1,000 iterations and the coexpression networks were derived based on the permuted data. From the null statistics, the statistical

[1] http://www.broadinstitute.org/gsea/msigdb/collections.jsp

significance (*P*) of the precision is indicated by the chance of getting higher precision from the null distribution. The *FDR* of the disease-mediated coexpression networks is the average of *q*-value from the null distribution.

The stability of the computationally derived coexpression networks was evaluated with different subsets of patient samples from the training set in 100 iterations. The stability is defined as the portion of the coexpression relations obtained from the original data that are retrieved by using only a random subset of the training data and the full test data.

2.5.4 Topological Validation

To confirm if the interactions obtained from the computational network models truly exists in biological context, bioinformatics tools and public genes/protein interactions databases would be used.

Ingenuity Pathway Analysis (IPA) software (Ingenuity Systems, Redwood City, CA) is a proprietary web-based curated database which provides contents of gene and protein interactions reported in the literature. The databases and software toolsets weigh and integrate information from numerous sources, including experimental repositories and text collections from published literature. Therefore, IPA allows researchers to derive curated molecular interactions, including both physical and functional interactions, and pathway relevance. In studies related to our work, IPA enables us to delineate molecular networks of genes interacting with the set of gene interested and identify the most significant biological processes and functions from the networks delineated from core analysis. Pathway Studio[2] is another bioinformatics application like IPA to allow user to carry out pathway analysis based on curated data from literature. Literature available in Pathway Studio is extracted from PubMed by MedScan application. STRING 8 (Search Tool for the Retrieval of Interacting Genes/Proteins) is a similar tool as IPA that retrieves protein/gene interactions reported in the literature [104]. Compared with STRING 8, IPA is more commonly used in industrial sectors, such as pharmaceutical firms because interactions included in the database are manually curated by scientist from reported literature; while interactions found in the STRING 8 database include predicted interactions and interactions resulted from automatic literature-mining searches. On the other hand, STRING 8 is

[2] http://www.ariadnegenomics.com/products/pathway-studio/

freely available and also includes a URL-based programming interface that allows researchers to query STRING from their applications.

The interactions derived from the computational network model could be confirmed against databases of known interactions among genes. A few recognized gene interaction databases include Kyoto Encyclopedia of Genes and Genomes (KEGG) [105, 106], NCI pathway interactions database (PID)[3] , and PubMed. The former two databases present gene pathways maps and molecular networks in diagrams. The latter one, PubMed, is primarily a web-based portal developed by National Centre for Biotechnology Information (NCBI) at National Library of Medicine (NLM) of U.S. National Institutes of Health (NIH). It comprises millions of citations for literature from MEDLINE, life science journals, biomedical journals, and books. From PubMed, users could view the related literatures on the genes interested and their respective interactions. From PubMed, users could also be redirected to specific information of the interactions retrieved from the Gene database under NCBI.

2.5.5 Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) allows assessment of gene sets in the genome-wide expression profiles [107]. Based on the genome-wide gene expression of a set of samples and their respective phenotype, GSEA would determine how well the members in the gene set correlated to the phenotypes. Specifically, according to the differential expression between the two, GSEA maintained a ranked list of genes (L). By going through the ranked list L, a measurement called enrichment score (ES) would be computed for each gene set using running-sum statistics with weighted correlation of the genes with the phenotype. ES reflects the degree to which a gene set is overrepresented to both ends of L. The statistical significance of the computed ES is indicated by a nominal P-value estimated by randomly permuting the samples phenotypes. If a gene set is significantly overrepresented with respect to the phenotypes (either one or both), then it would have extreme ES at both ends of the ranked list L, as shown in Fig. 2.12.

[3] http://pid.nci.nih.gov/

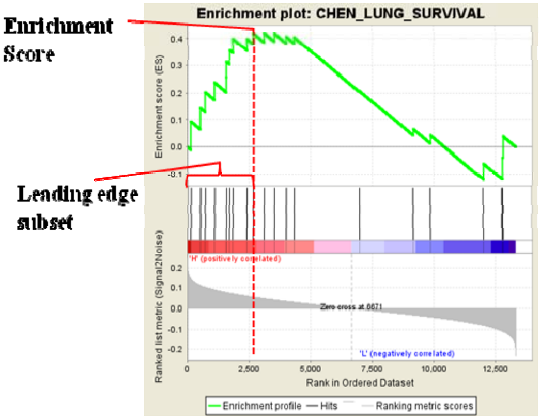


Figure 2.12. An example of enrichment plot for a gene set.

GSEA also allows evaluation of multiple gene sets at once, which is comparing the enrichment of multiple gene sets against one another in the input genome-wide expression profiles. This multiple gene sets comparison is actually a multiple comparison problem. In order to correct the measurements (*ES*) according to multiple hypotheses testing, the phenotype labels were randomly permuted. A normalized enrichment score (*NES*) for each gene set is generated by averaging enrichment scores from all permutations. Statistical significance of the *NES* corresponding to each gene set is indicated by false discovery rate (*FDR*) in the permutation analysis by permuting the phenotypes [107]. The gene set that gets high absolute *NES* and low *FDR* compared with other signatures implies that it is more significantly enriched than others in the provided gene expression profiles.

2.6 Data Used in Experiments

Four sets of published microarray gene expression profiles of lung cancer patients were used throughout our studies. All four sets of gene expression profiles were quantified with Affymetrix GeneChip® human genome expression arrays.

The first set is the largest lung cancer microarray data publicly available till date. It contains gene expression profiles quantified with Affymetrix HG-U133A on 442 lung adenocarcinoma patient samples obtained from a multi-center microarray study of lung cancer published by Shedden et al, which also known as the Director's Challenge Study[2]. This study cohort is composed of four data sets (University of Michigan, H. Lee Moffitt Cancer Center, Memorial Sloan-Kettering Cancer Center, and Dana-Farber Cancer Institute) contributed by six institutions. The raw microarray data are available from caArray website[4]. The second set contains 130 adenocarcinoma and squamous cell lung cancer samples published by Raponi et al. [17]. The third set contains 111 non-small cell lung carcinoma samples published by Bild et al. [13]. Table 2.3 provides a summary on the data and clinical characteristics of each cohort. The fourth set contains expressions quantified with Affymetrix HG-U133A on 164 airway epithelial cells from current and former smokers published by Spira et al. [108]. This cohort is composed of lung cancer patients and smokers without lung cancer. It was particular used in our study of a smoking-associate signature and separated into independent training and test sets. Table 2.4 gives a summary of patient characteristics in each set.

Table 2.3. Characteristics summary of three lung cancer patients cohorts.

	Director's Challenge Study[2] (n=442)	**Raponi et al. [17] (n=130)**	**Bild et al. [13] (n=111)**
Affymetrix GeneChip®	HG-U133A	HG-U133A	HG-U133 Plus 2
Histology			
Adenocarcinoma	100%		52%
Squamous cell		100%	48%
Median follow-up (months)	47	34	31
Age (mean, s.d.)	64 (10)	67 (10)	65 (10)
Sex (% male)	50%	63%	57%
Tumor Stage			
Stage I	62%	56%	60%
Stage II	22%	26%	16%
Stage III	15%	18%	22%
Stage IV	-	-	2%
Unknown	1%	-	-

[4] https://array.nci.nih.gov/caarray/project/details.action?project.id=182

Table 2.4. Patient characteristics from Spira et al [108].

	Training (*n*=77)	Test 1 (*n*=52)	Test 2 (*n*=35)
Age (mean, s.d.)	57 (14)	55 (16)	64 (11)
Sex (% male)	78%	75%	69%
Lung Cancer Histology			
Small Cell	15% (6/40)	25% (5/20)	17% (3/18)
Non-small Cell	83% (33/40)	75% (15/20)	78% (14/18)
Unknown	2% (1/40)	-	5% (1/18)
Small Cell Tumor Stage			
Limited	3 (3/6)	4	2
Extensive	3 (3/6)	1	1
NSCLC Tumor Stage			
Stage I	30% (10/33)	7% (1/15)	14% (2/14)
Stage II	-	13% (2/15)	-
Stage III	30% (10/33)	47% (7/15)	36% (5/14)
Stage IV	39% (13/33)	33% (5/15)	29% (4/14)
Unknown	-	-	21% (3/14)

2.7 Summary

In this chapter, we had reviewed current methods and tools used in gene signature identification and prognostic prediction using microarray data. Having studied current approaches and problems in these methods, we proposed two methodologies to efficiently discover prognostic gene signatures for lung cancer molecular prognosis.

We first proposed a hybrid system comprised of statistical and machine learning feature selection methods to identify gene signatures for lung cancer prognosis and chemoresponse prediction, which will be presented in Chapter 3. From the review of the numerous state-of-the-art ranking-based gene selection methods used in microarray studies, each method had their strengths while other methods missing. Therefore, we hypothesized that through a framework of a systematic multiple-stage gene filtering approach, we could exploit the strengths from various methods and lead to identification of predictive biomarkers.

In order to incorporate the interactive machinery of gene functions and signaling pathway information in biomarker discovery, we proposed the second methodology based on gene coexpression networks modeled with implication networks. Implication networks were chosen over Bayesian networks and ANNs because it could be efficiently constructed and unlike Bayesian networks, it allows formation of cyclic relations. In Chapter 4, the network-based approach was studied at the genome-wide scale. Extensive study was carried out to examine the performance of the network-based approach when it was applied alone or in combination with feature selection methods. In Chapter 5, the network-based approach was studied on a smaller pool of genes: the genes that are associated with smoking and lung cancer survival. Chapter 6 evaluates the performance of the implication networks employed in our studies in comparison with the Boolean implication networks.

Chapter 3

Hybrid Models Identified Gene Signatures for Lung Cancer Prognosis and Chemoresponse Prediction

As discussed in Chapter 2, ranking-based methods are simple to be implemented for gene selection. Statistics methods such as *t*-test and SAM scale efficiently to large number genes and control for false positive genes. However, it usually leads to fairly large gene sets (~10^2). Feature selection methods such as *Relief* incorporate the classification in the evaluation of gene predictive performance. Nonetheless, it's computationally heavier than statistics methods and thus does not scale well to the whole genome.

In this chapter, we present our first proposed hybrid system for efficient prognostic gene signature identification. The proposed hybrid system combined traditional statistics and feature selection methods in multiple stages to identify predictive gene signatures for lung cancer prognosis. This system has a few appealing characteristics. The first appealing characteristic is that it exploits of the strengths of each gene selection methods while avoiding the drawbacks in the systematic integration. The second characteristic of the system is that it leads to identification of small set of genes with high prognostic performance. Smaller size of gene signatures will not only reduce the time and cost of further validation but also make the clinical application more feasible.

The proposed hybrid system identified a 12-gene and 15-gen lung cancer prognostic signatures. These two signatures are more accurate compared with previously published signatures in the largest lung adenocarcinoma data samples (*n* = 442) [2]. Moreover, the 12-gene signature could identify stage I and stage II patients who might benefit from adjuvant chemotherapy and who could be spared from it. This implies that the 12-gene signature could be used to select treatment for stage I and II patients. Quantitative RT-PCR analyses of independent NSCLC tissue samples confirmed the gene expression patterns of these two signatures. Functional pathway analysis revealed that the signature genes had interactions with well established cancer hallmarks, indicating the important roles of the signature genes in tumor initiation and progression. The 12-gene signature also accurately predicted chemoresistance and chemosensitivity to Cisplatin, Carboplatin, Paclitaxel (Taxol), Etoposide, Gefitinib and Erlotinib in a panel of 60 cancer cell lines (NCI-60).

The remainder of this chapter is organized as follows. Section 3.1 illustrates the methodology of the proposed hybrid system. The experiment design is presented in Section 3.2. Section 3.3 describes the identification of various gene signatures using the proposed system. Survival prediction performance of the identified signatures is presented in Section 3.4 to 3.6. Section 3.7 presents the survival prediction of early stage patients. Section 3.8 presents implications of the 12-gene signature in treatment selection for stage I and II NSCLC. Section 3.9 compares the identified gene signatures with clinical and demographical parameters. Our gene signatures were compared with published lung cancer signatures in Section 3.10. Section 3.11 confirms the expression patterns of the identified genes. Section 3.12 shows the chemoresponse prediction capability provided by the 12-gene signature. Functional pathway of the 12-gene signature can be found in Section 3.13. The last section, Section 3.14 discusses the study and concludes the chapter.

3.1 Methodology

We developed a hybrid system with combination of traditional statistics and feature selection methods for the identification of gene signatures and lung cancer prognosis. As depicted in Fig. 3.1, the proposed system comprised the following steps: 1) Selection a pool of candidate genes from the whole genome using statistical methods. 2) Ranked the pool of candidate genes with *Relief* feature selection. 3) From the top ranked gene, one gene was added at each step to the gene set, until the classification accuracy could not be improved by adding one more gene, the gene set is the prognostic gene signature identified. Specifically, this could be interpreted as a system with two phases, with statistical methods in the first phase and step-wise forward selection with feature selection in the second phase.

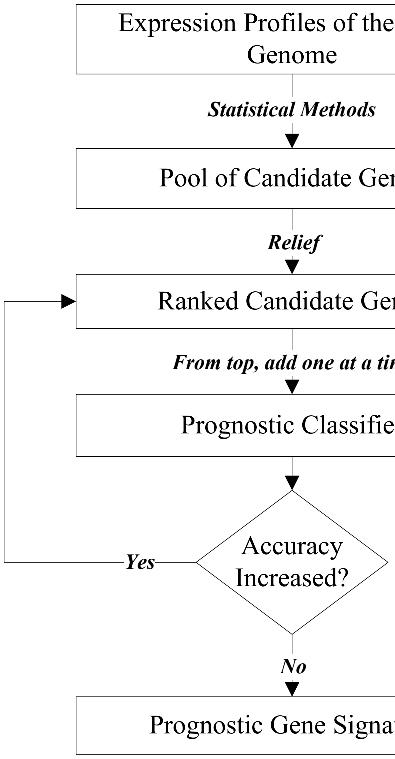


Figure 3.1. Hybrid system with traditional raking-based gene selection methods.

This hybrid system utilize the statistical methods in the first phase because statistical methods is more computational efficient in large scale. Step-wise forward feature selection with in the second phase allows us to obtain the smallest set of genes with the optimized prognostic performance.

3.2 Prognostic Model System

In order to take advantage of different algorithms of gene selection, hybrid models of different methods in different stages are needed for biomarker discovery and good disease classification. In this study, we combined statistical methods and machine learning algorithms to identify prognostic biomarkers of lung adenocarcinoma. The 442 lung adenocarcinoma patient samples from the Director's Challenge Study [2] were used in this study. The UM & HLM cohorts from the sample formed the training set (*n* = 256), whereas the samples from MSK (*n* = 104) and the DFCI (*n* = 82) formed two independent test sets.

In general, the hybrid systems examined in this study included three phases (Fig 3.2): 1) identification of a small set of signature genes by combining statistical methods and feature selection methods from genome-scale transcriptional profiles of the training cohort, 2) construction of a classifier to predict overall survival in lung cancer patients, and 3) validation of the gene expression-based prognostic model in two independent patient cohorts. The model validation and evaluation of the identified gene signature were also compared with over previously published lung cancer prognostic signatures on the two independent test sets. Specifically, in the first phase, two combinatorial schemes were studied by joining statistical methods and feature selection algorithms. The first scheme was combination of pooled-variance t-test and Relief algorithm. The second scheme was combination of Significance Analysis of Microarrays (SAM) [34], different-variance t-test, and *Relief* algorithm. A functional pathway analysis after the previous two schemes was carried out to explore the biological functions shared by the gene sets. Since signatures identified from various approaches performed differently in different classifiers, Cox model and Naïve Bayes classifiers were used to model the prognostic model to predict overall survival in lung cancers.

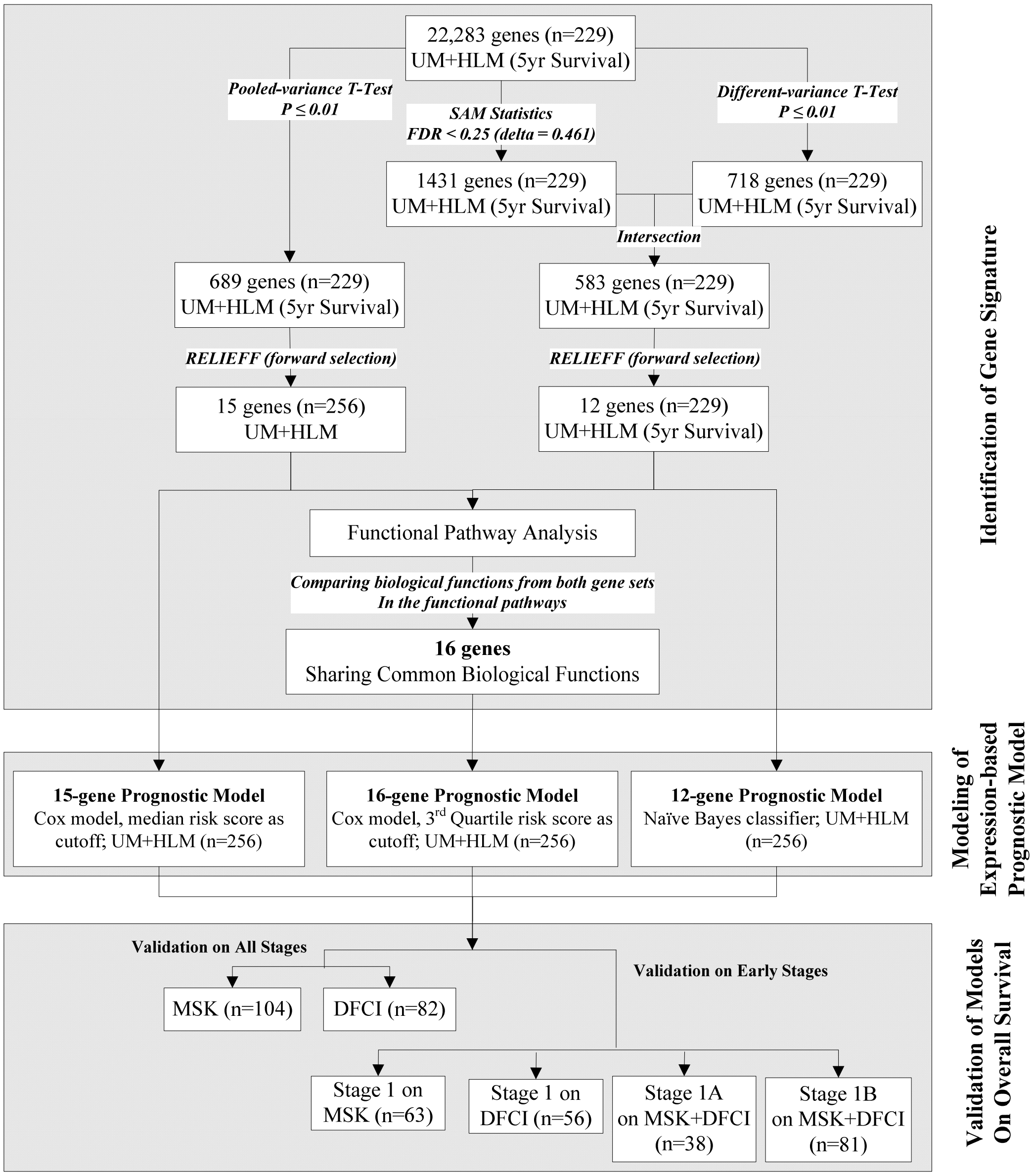

22,283 genes (n=229)
UM+HLM (5yr Survival)
Pooled-variance T-Test
P ≤ 0.01
SAM Statistics
FDR < 0.25 (delta = 0.461)
Different-variance T-Test
P ≤ 0.01
1431 genes (n=229)
UM+HLM (5yr Survival)
718 genes (n=229)
UM+HLM (5yr Survival)
Intersection
689 genes (n=229)
UM+HLM (5yr Survival)
583 genes (n=229)
UM+HLM (5yr Survival)
RELIEFF (forward selection)
RELIEFF (forward selection)
15 genes (n=256)
UM+HLM
12 genes (n=229)
UM+HLM (5yr Survival)
Functional Pathway Analysis
Comparing biological functions from both gene sets
In the functional pathways
16 genes
Sharing Common Biological Functions
Identification of Gene Signature
15-gene Prognostic Model
Cox model, median risk score as cutoff; UM+HLM (n=256)
16-gene Prognostic Model
Cox model, 3rd Quartile risk score as cutoff; UM+HLM (n=256)
12-gene Prognostic Model
Naïve Bayes classifier; UM+HLM (n=256)
Modeling of Expression-based Prognostic Model
Validation on All Stages
MSK (n=104)
DFCI (n=82)
Validation on Early Stages
Stage 1 on MSK (n=63)
Stage 1 on DFCI (n=56)
Stage 1A on MSK+DFCI (n=38)
Stage 1B on MSK+DFCI (n=81)
Validation of Models On Overall Survival

Figure 3.2. Overview of the hybrid model to molecular prognosis.

3.3 Identification of 15-, 12-, and 16-gene Signature

Three combinatorial schemes with multiple gene selection methods were adopted to examine the hybrid system for prognostic signatures identification. In the first scheme, *t*-test was used to select candidate genes from 22,283 probes quantified on the training cohort (n = 256) in the first phase. The pooled-variance *t*-test selected 689 genes with significant differential expression (P <

0.01) between the low-risk groups (patient who survived longer than 5 years) and high-risk (those who died within 5 years following surgery) groups. Twenty-seven censored cases with follow-up time less than 5 years were removed from this analysis due to the uncertainty of patient post-operative status. In order to refine the gene set into a more feasible size for clinical application, *Relief* algorithm implemented in WEKA 3.4 was used to rank each of these 689 genes in terms of the power to separate low-risk and high-risk groups. On the ranked list, step-wise forward selection was used to identify a gene subset with the highest prognostication accuracy. Specifically, starting from the top ranked gene, one gene was added at each step to the gene set, until the classification accuracy could not be improved by adding one more gene. At each step, the gene set was used to classify good-prognosis and poor-prognosis groups with Cox model, with median risk score of the training set as the cutoff for stratification. On the ranked list of the 689 genes, the process stopped when the addition of a new gene did not increase the Cox model stratification after the top 15-gene set. As a result, a 15-gene signature (Table 3.1) was identified.

In the second scheme, a combination of *t*-test and SAM was then used to select candidate prognostic with a predefined false discovery rate. Specifically, a different-variance *t*-test selected 718 genes with significant differential expression ($P < 0.01$) between the two prognosis groups. With false discovery rate (*FDR*) of 25% (*delta* = 0.46), SAM selected 1,431 genes that significantly differentiated the two prognostic groups. There were 583 genes selected by both *t*-tests and SAM, and these were considered the set of candidate prognostic genes for the next stage of the analysis. In the next step where gene set was further refined, similar approach as adopted in the first scheme discussed above was employed. *Relief* algorithm implemented in WEKA 3.4 was used to rank these 583 genes and forward selection was used to select the most signature genes starting from the top ranked gene. The gene set was used to classify good-prognosis and poor-prognosis groups with Naïve Bayes algorithm. The forward selection process stopped when the addition of a new gene did not increase the classification accuracy as evaluated in a 10-fold cross validation. As a result, a 12-gene signature (Table 3.2) was identified as the most accurate prognostic genes from the set of candidate genes for overall survival prediction.

The third approach combined all the steps adopted in the first two approaches with a biological functional pathway analysis. Specifically, functional pathway analysis was done on the 15-gene and 12-gene signatures using IPA. By comparing the biological functions of 15- and 12-gene signatures, there were 16 genes sharing the same functions (Table 3.4). As a result, the 16 genes with common functions were selected as the signature gene (Table 3.3).

Table 3.1. List of 15-gene signature.

Probe Set ID	**Gene**	**Functions**	**Classification**
204854_at	GPR162 /// LEPREL2	Collagen biosynthesis, folding, and assembly	Metabolism
206150_at	CD27	B-cell activation and immunoglobulin synthesis; signaling transduction	Oncogene
205171_at	PTPN4	Cell growth, differentiation, mitotic cycle, and oncogenic transformation	Oncogene
201107_s_at	THBS1	Cell-to-cell and cell-to-matrix interactions.	Oncogene
210762_s_at	DLC1	A candidate tumor suppressor gene	Oncogene
218340_s_at	UBA6	Ubiquitin-activating protein	Protein Degradation
211327_x_at	HFE	Iron absorption	Signaling Transduction
208772_at	ANKHD1	Unknown	Structure
211603_s_at	ETV4	Cellular movement	Transcription
207296_at	ZNF343	Unknown	Transcription
214717_at	DKFZp434H1419	Unknown	N/A
213779_at	EMID1	Unknown	N/A
215598_at	TTC12	Binding	N/A
201581_at	TXNDC13	Cell redox homeostasis, electron transport chain	N/A
205308_at	FAM164A	Unknown	N/A

Table 3.2. List of 12-gene signature.

Probe Set ID	Gene	Protein Functions	Classification
212041_at	ATP6V0D1	ATPase	Metabolism
222078_at	PKLR	Pyruvate kinase	Metabolism
219808_at	SCLY	Catalyzes the decomposition of L-selenocysteine to L-alanine and elemental selenium	Metabolism
209420_s_at	SMPD1	Converts sphingomyelin to ceramide	Metabolism
210762_s_at	DLC1	A candidate tumor suppressor gene	Oncogene
204524_at	PDPK1	Cell signal protein	Oncogene
218833_at	ZAK	Cell signal protein	Oncogene
208855_s_at	STK24	Protein kinase	Signaling Transduction
208775_at	XPO1	Mediates nuclear export of cellular proteins	Signaling Transduction
46142_at	LMF1	Maturation of specific proteins in the endoplasmic reticulum	Structure
205308_at	FAM164A	Unknown	N/A
221685_s_at	CCDC99	Cell cycle	Signaling Transduction

Table 3.3. List of 16-gene signature.

Probe Set ID	Gene	Functions	Classification
206150_at	CD27	B-cell activation and immunoglobulin synthesis; signaling transduction	Oncogene
205171_at	PTPN4	Cell growth, differentiation, mitotic cycle, and oncogenic transformation	Oncogene
201107_s_at	THBS1	Cell-to-cell and cell-to-matrix interactions.	Oncogene
211327_x_at	HFE	Iron absorption	Signaling Transduction
211603_s_at	ETV4	Cellular movement	Transcription
201581_at	TXNDC13	Cell redox homeostasis, electron transport chain	N/A
212041_at	ATP6V0D1	Atpase	Metabolism
222078_at	PKLR	Pyruvate kinase	Metabolism
219808_at	SCLY	Catalyzes the decomposition of L-selenocysteine to L-alanine and elemental selenium	Metabolism
209420_s_at	SMPD1	Converts sphingomyelin to ceramide	Metabolism
210762_s_at	DLC1	A candidate tumor suppressor gene	Oncogene
204524_at	PDPK1	Cell signal protein	Oncogene
218833_at	ZAK	Cell signal protein	Oncogene
208855_s_at	STK24	Protein kinase	Signaling Transduction
208775_at	XPO1	Nuclear protein transport	Signaling Transduction
46142_at	LMF1	Maturation of specific proteins in the endoplasmic reticulum	Structure

Table 3.4. Comparison of biological functions between the 12- and 15-gene signatures with curated database.

Category	Category	12-gene	15-gene	Common
Diseases and Disorders	Cancer			✓
	Cardiovascular Disease		✓	
	Connective Tissue Disorders		✓	
	Dermatological Diseases and Conditions		✓	
	Genetic Disorder			✓
	Hematological Disease			✓
	Hepatic System Disease			✓
	Immunological Disease			✓
	Infection Mechanism	✓		
	Inflammatory Disease		✓	
	Inflammatory Response		✓	
	Metabolic Disease			✓
	Neurological Disease			✓
	Reproductive System Disease			✓
	Respiratory Disease			✓
	Skeletal and Muscular Disorders		✓	
Molecular and Cellular Functions	Amino Acid Metabolism			✓
	Antigen Presentation		✓	
	Carbohydrate Metabolism		✓	
	Cell Cycle		✓	
	Cell Death			✓
	Cell Morphology		✓	
	Cell Signaling			✓
	Cell-To-Cell Signaling and Interaction		✓	
	Cellular Assembly and Organization			✓
	Cellular Compromise		✓	
	Cellular Development			✓
	Cellular Function and Maintenance			✓
	Cellular Growth and Proliferation			✓
	Cellular Movement			✓
	DNA Replication, Recombination, and Repair	✓		
	Drug Metabolism		✓	
	Gene Expression	✓		
	Lipid Metabolism			✓
	Molecular Transport			✓
	Nucleic Acid Metabolism		✓	
	Post-Translational Modification			✓

	Protein Synthesis		✓	
	Protein Trafficking		✓	
	RNA Trafficking	✓		
	Small Molecule Biochemistry			✓
Physiological System Development and Function	Cardiovascular System Development and Function		✓	
	Cell-mediated Immune Response		✓	
	Hematological System Development and Function		✓	
	Immune Cell Trafficking		✓	
	Nervous System Development and Function		✓	
	Organ Development		✓	
	Skeletal and Muscular System Development and Function			✓
	Tissue Development		✓	
	Tumor Morphology		✓	
	Visual System Development and Function		✓	

3.4 Survival Prediction Using 15-gene Prognostic Model

Using expression profiles of the 15 genes as predictors, a prognostic classifier was constructed to stratify patients into low- and high-risk of failure in survival (i.e. death) using a multivariate Cox proportional hazard model. The Cox model of overall survival was constructed based on the 15-gene signature, with each gene variable as a covariate. In the UM & HLM training samples (n = 256), a survival risk score was generated for every patient, with a higher risk score representing a greater probability of death. From the gene expression-defined risk scores in the training cohort, median of the risk score (value of -1.79) was identified as the cut-off to stratify patients into low- and high-risk groups. The constructed training model and the cut-off value were then applied to the two validation sets. In all three patient cohorts, the 15-gene defined model stratified patients into prognostic groups with distinct overall survival (log-rank $P < 0.03$; Fig. 3.3)

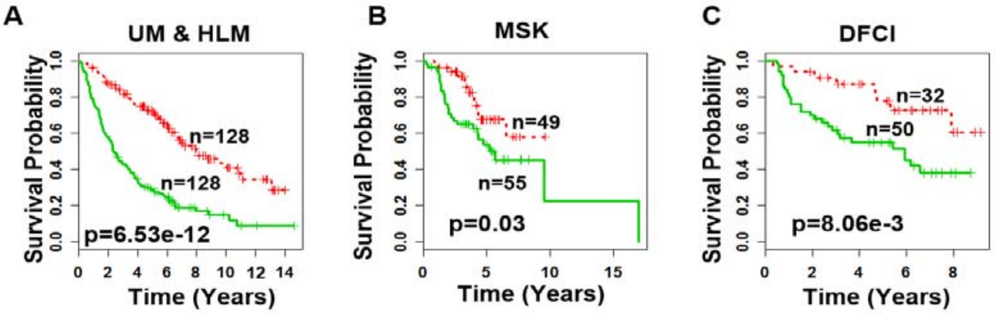


Figure 3.3. Kaplan-Meier analysis of the 15-gene signature on patients on all stages.

3.5 Survival Prediction Using 12-gene Prognostic Model

To predict overall survival using the 12-gene signatures, expression profiles of the identified 12 genes were used as predictors in a prognostic classifier to stratify patients into low-risk (5-year survival) and high-risk (non-5-year survival) groups. The Naïve Bayes classifier implemented in WEKA 3.4 was used in the classification on UM & HLM training samples (low-risk n = 104; high-risk n = 125). Twenty-seven censored cases without sufficient follow-up information were removed in the model construction. Priors estimated by the model are 0.45 for low-risk class and 0.55 for high-risk class. Other parameters of the trained Naïve Bayes model, including the mean and standard deviation for each of the 12 genes in both low- and high-risk groups, are listed in Table 3.5.

Table 3.5. Parameters estimated in the 12-gene Naive Bayes classifier.

Gene (attribute)	Low-risk mean (μ_{Li})	Low-risk standard deviation (σ_{Li})	High-risk mean (μ_{Hi})	High-risk standard deviation (σ_{Hi})
LMF1	101.6708	31.6461	88.6869	29.5986
DLC1	868.5886	578.3862	648.4284	530.6969
PKLR	14.3474	6.872	11.002	5.5501
ATP6V0D1	1388.054	398.6874	1209.6369	325.7233
CCDC99	277.1923	56.2284	300.0086	60.678
SCLY	58.3824	13.2988	63.6222	13.7703
PDPK1	297.6373	117.3514	253.7384	103.0455
FAM164A	264.8707	106.5128	223.8295	96.6066
SMPD1	278.5686	84.5316	239.3571	65.4393
XPO1	1674.3741	344.9824	1824.6274	400.4278
ZAK	132.694	67.7063	159.0546	79.1456
STK24	2248.6647	529.6098	2457.9982	576.496

The Naïve Bayes classifier computes the posterior probability of death within 5 years after surgery in each patient. This posterior probability represents the risk for tumor recurrence in patients, since recurrence is the major cause of treatment failure (i.e. death) in lung cancer. Based on the posterior probability, a patient is classified into the high-risk group if the value is greater than 0.5; or into the low-risk group otherwise. The training model was evaluated in a 10-fold cross validation. Without parameter re-estimation, this model was then used to predict posterior probability representing the risk for tumor recurrence in each patient in two test sets (MSK and DFCI), as well as the censored cases left out of the model construction. The distribution of the posterior probability of 442 patients in this study was illustrated in Fig. 3.4A. After obtaining the predicted outcomes, Kaplan-Meier (KM) analysis was carried out to estimate the average survival probability at the 5-year mark following surgery. Results show that high-risk posteriors from the prognostic model are strongly associated with the 5-year survival probabilities (Fig. 3.4B). Patients with a high probability of tumor recurrence tend to be more likely to have treatment failure after surgery. This indicates that the high-risk posterior probability computed by the model is a good prognostic factor of lung cancer survival. The wide 95% confidence interval at posteriors ranging from 0.35 to 0.6 (Fig. 3.4B) might be due to the small sample size in this distribution (Fig. 3.4A). Furthermore, a posterior of 0.5 means that the chance of tumor recurrence is random, which also leads to a looser confidence interval.

Using the prognostic categorization scheme described above, the 12-gene signature separated patients into high- and low-risk groups with significantly distinct (log-rank P = 6.96e-7) post-operative survival on the training cohort in Kaplan-Meier analysis (Fig. 3.5A). This scheme generated significant patient stratification on independent validation sets MSK (log-rank P = 9.88e-4; Fig. 3.5B) and DFCI (log-rank P = 2.57e-4; Fig. 3.5C).

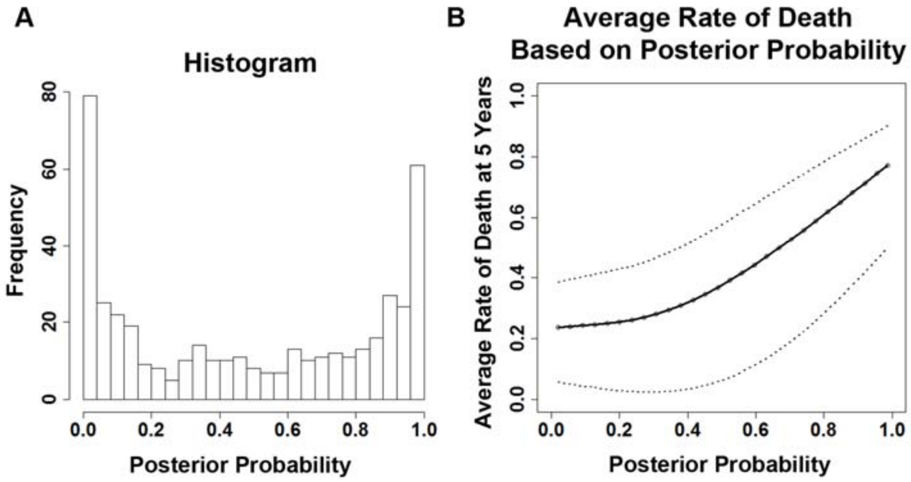


Figure 3.4. Association of the 12-gene risk score algorithm and lung cancer survival. (A) Histogram showing the distribution of the risk scores (posterior probabilities of high-risk) in the whole studied cohort. (B) Average rate of death at five years after surgery corresponding to 12-gene risk score (posterior probability). The dotted lines represent 95% confidence interval.

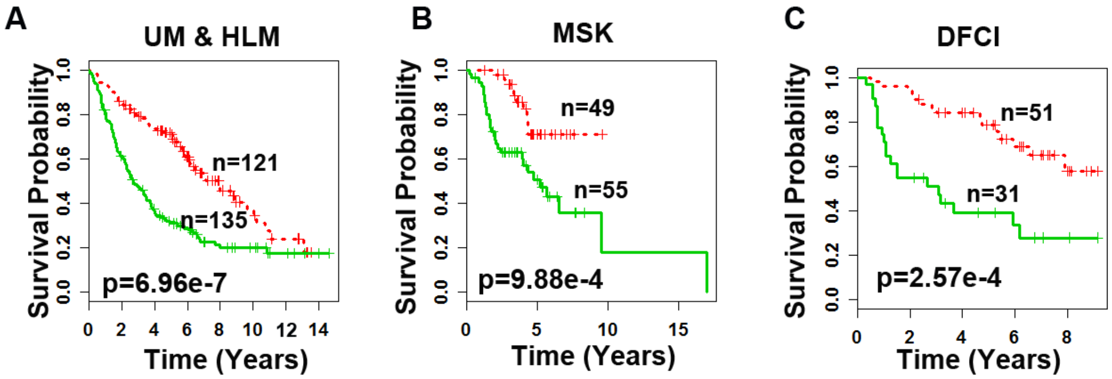


Figure 3.5. Kaplan-Meier analysis of the 12-gene prognostic classification in lung cancer patients.

3.6 Survival Prediction Using 16-gene Prognostic Model

A prognostic classifier was constructed using the expressions of the 16-gene signature to stratify patients into low- and high-risk of death using a multivariate Cox proportional hazard model with a similar approach adopted for 15-gene prognostic model. With the 16 genes as a covariates, a survival risk score was generated for patients in the UM & HLM training samples (*n* = 256). From the distribution of gene expression-defined risk scores in the training cohort, the 3rd quartile (value of -1.5724) was identified as the cut-off to stratify patients into low- and high-risk groups. Then, the constructed training model and the cut-off value were applied to the two validation sets. In all three patient cohorts, the 16-gene prognostic model stratified patients into prognostic groups with distinct overall survival (log-rank $P < 0.03$; Fig. 3.6)

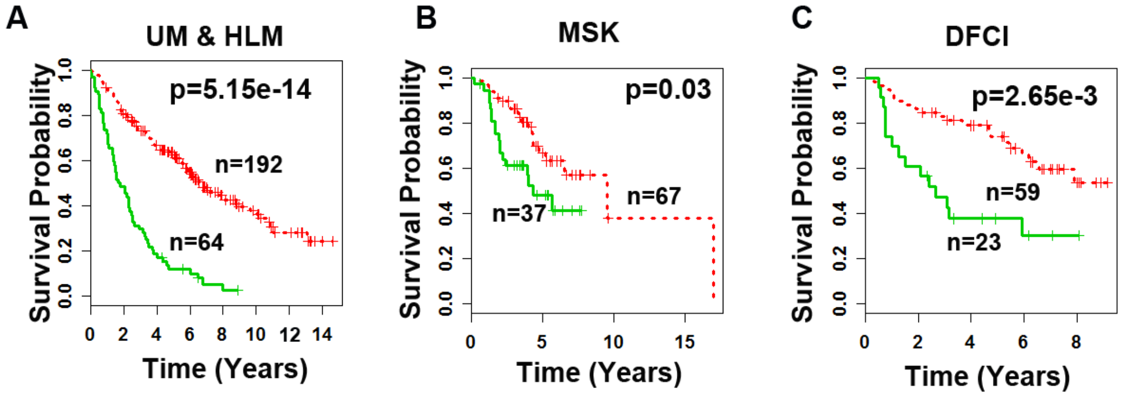


Figure 3.6. Prognostic performance of the 16-gene signature in patients on all stages.

3.7 Survival Prediction for Stage I NSCLC Patients

In current practice, treatment for patients diagnosed with NSCLC is based on AJCC tumor stage. Surgical resection to remove the tumor is the major treatment option for stage I NSCLC patients. However, about 35-50% of stage I NSCLC patients will develop and die from tumor recurrence within the five years following surgery [4, 5]. On the other hand, stage IB patients who received surgical resection followed by adjuvant chemotherapy showed improved survival rate [30]. Thus, we sought to explore whether the 15-, 12-, and 16-gene expression-defined prognostic classifier could identify specific high-risk patients with stage I tumors for the aggressive adjuvant chemotherapy.

Results show that the 15-gene prognostic signature could identify high-risk patients with stage I tumors on training cohort (results not shown) and DFCI test cohort (log-rank $P = 0.02$; Fig 3.7B) but not on the MSK Stage I patients (log-rank $P = 0.12$; Fig 3.7A) and the stage IA patients in the combined cohort of MSK and DFCI (results not shown). The 15-gene prognostic model could also separate high- and low-risk groups (log-rank *P* = 0.008) within stage IB patients in the combined test sets (Fig. 3.7C).

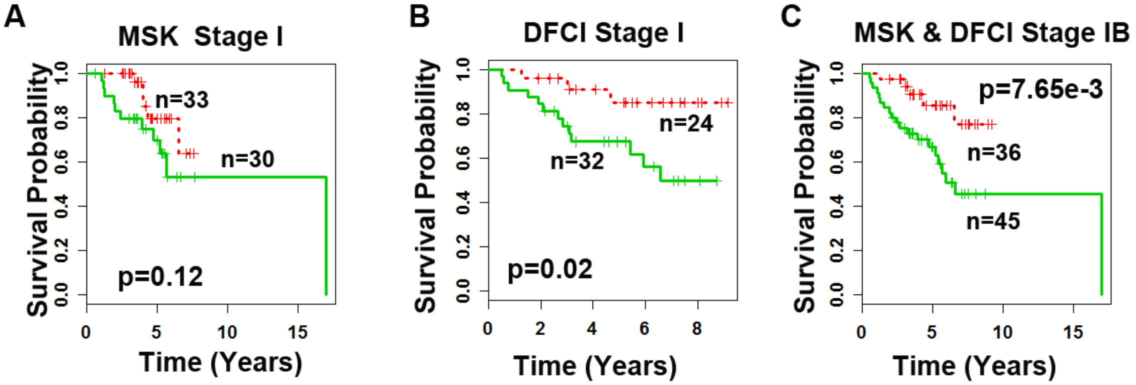


Figure 3.7. Prognostic performance of the 15-gene signature in stage I patients.

The 16-gene prognostic signature performed similarly as the 15-gene model in stage I patients. The 16-gene prognostic model generated significant stratifications in patients with stage I tumors on training cohort (results not shown) and DFCI test cohort (log-rank *P* = 0.01; Fig 3.8B), but not on the MSK test cohort (log-rank *P* = 0.34; Fig 3.8A) and the stage IA patients in the combined test cohort (result not shown). The 16-gene prognostic model also separated high- and low-risk groups (log-rank *P* = 0.02) within stage IB patients in the combined test sets (Fig. 3.8C).

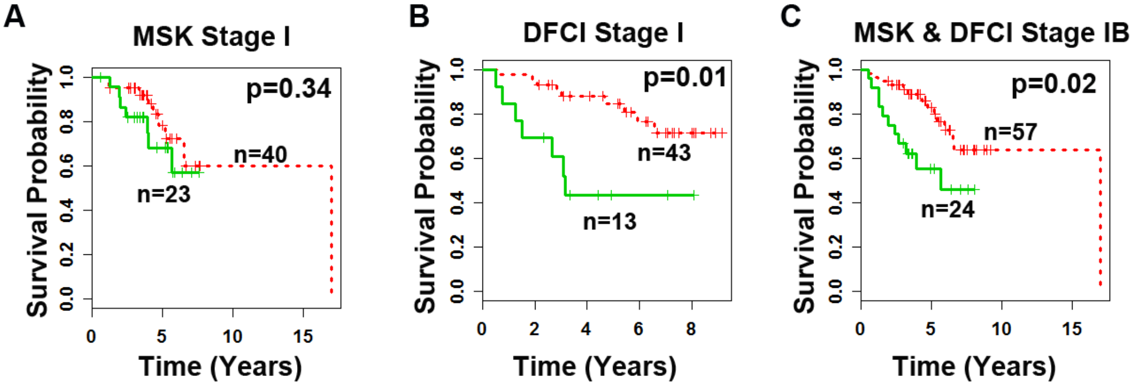


Figure 3.8. Prognostic performance of the 16-gene signature in stage I patients.

The 12-gene prognostic signature could reliably identify high-risk patients with stage I tumors on both the training cohort (results not shown) and two independent test cohorts (log-rank *P* = 0.04; Fig. 3.9A, Fig. 3.9B). The prognostic model also separated high- and low-risk groups (log-rank *P* = 4.73e-3) within stage IB patients in the combined test sets (Fig. 3.9C).

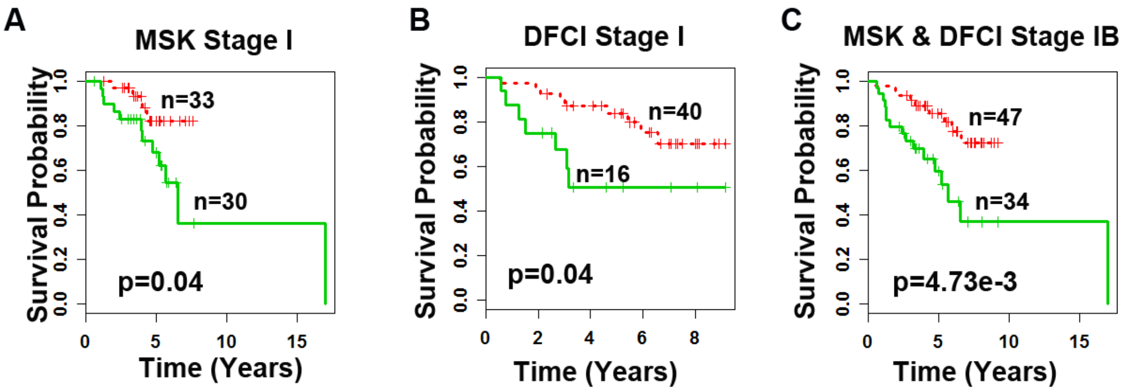


Figure 3.9. Prognostic performance of the 12-gene signature in stage I patients.

These results demonstrate that the identified 12-gene signature is independent of the current AJCC staging system. Result from the KM analyses that the 12-gene signature could stratify Stage I patients into two significantly distinct survival groups demonstrate that the 12-gene signature provides more precise prognosis than the current AJCC staging system. Using the 12-gene model, stage I NSCLC patients could be advised to receive adjuvant chemotherapy according to the expression profiles of the 12 signature genes.

3.8 Treatment Selection for Stage I and II NSCLC Patients with the 12-gene Signature

Among the three prognostic models constructed, 12-gene prognostic model was the only model that generated significant stratification on the stage I patients in both the training cohort (results not shown) as well as two test cohorts (Fig 3.10). Therefore, we further assessed whether the 12-gene signature could be used for treatment selection for stage I and II NSCLC patients. Patients who did not receive chemotherapy were selected for this analysis. Results from the KM analysis show that the prognostic model separated high- and low-risk stage I patients without chemotherapy in the training (UM & HLM; log-rank *P* = 0.04; Fig. 3.10A) and test cohorts (MSK & DFCI; log-rank *P* = 0.02; Fig. 3.10B). Similarly, the model differentiated high- and low-risk stage II patients without chemotherapy in the training (log-rank *P* = 0.06; Fig. 3.10C) and test cohorts (log-rank *P* = 0.03; Fig. 3.10D) in KM analyses. These results indicate that the 12-gene expression-defined prognostic model could reliably select patients with early stage NSCLC for adjuvant chemotherapy. Meanwhile, it could also spare some low-risk stage I and II NSCLC patients from chemotherapy based on the expression patterns of the identified gene markers in the tumors.

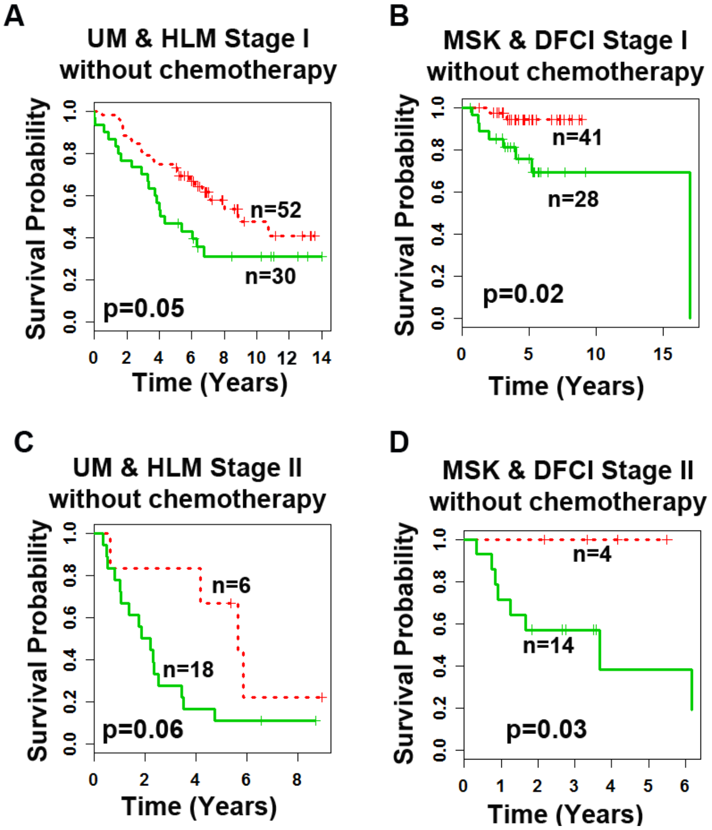


Figure 3.10. Evaluation of the 12-gene signature in treatment selection.

3.9 Prognosis Evaluation of the Identified Signature with Clinical Covariates

To confirm the prognostic power of the identified signatures, the expression-defined prognostic model was evaluated with commonly used prognostic factors of lung cancer, including gender, age, and tumor stage on the combined testing cohorts (DFCI and MSK).

The posterior probability of high-risk estimated by the 12-gene Naïve Bayes classifier, termed as 12-gene risk score, was used as a covariate in the multivariate Cox analysis. Risk scores estimated by the 15-gene and 16-gene fitted Cox model was used as a covariate in the analysis (Table 3.6). Results showed that without the 12-, 15-, or the 16-gene risk score, tumor stage was the only factor significantly ($P < 0.00006$) associated with risk of lung cancer death. When the 12-gene risk score was added to the multivariate Cox model, the 12-gene risk score demonstrated a strong association with the lung cancer survival (hazard ratio = 3.94, 95% CI: [2.07, 7.52]), and tumor stage remained significant (Table 3.6). When the 15-gene risk score was added to the multivariate Cox model, tumor stage remained significant and the 15-gene risk score also showed significantly association with the lung cancer survival (hazard ratio = 1.99, 95% CI: [1.37,2.89]; Table 3.6). In the analysis with the 16-gene risk score, the 16-gene risk score also appeared to be a significant factor associated with the lung cancer survival (hazard ratio = 2.50, 95% CI: [1.33,3.59]) and tumor stage remained significant (Table 3.6).

A comprehensive evaluation was carried out with all available clinical covariates and demographic factors in the dataset, including smoking history, race, and tumor differentiation (Table 4). In this comprehensive evaluation, the 12-gene risk score remained as a highly significant prognostic factor with a hazard ratio of 4.19 (95% CI: [2.08, 8.46]; Table 3.7). The risk scores of the 15- and 16-gene demonstrated to be significant factors in this analysis while comparing with all clinical and demographic factors, with a hazard ratio of 1.81 (95% CI: [1.23, 2.65]) and 2.45 (95% CI: [1.72, 3.50]) respectively (Table 3.7).

In both multivariate analyses, the hazard ratios of the 12-gene risk score algorithm were higher than other clinical covariates except tumor stage (III vs. I), while there is no significant difference between the hazard ratio of the 12-gene signature and tumor stage. Hazard ratio of the 15-gene risk score was comparatively good as the hazard ratio of the tumor stage (III vs. I) in the first analysis with major clinical covariates. However, in the comprehensive analysis, the hazard ratio of tumor stage (III vs. I) was significantly higher than the hazard ratio of 15-gene risk score. The hazard ratio of the 16-gene risk score was comparatively good as the hazard ratio of tumor stage with stage II vs. I but not significantly higher than the hazard ratio of tumor stage III vs. I in both multivariate analyses. These results demonstrate that the 12-gene signature is a more accurate prognostic factor than some commonly used clinical parameters.

Table 3.6. Multivariate Cox proportional analysis of the 12-, 15-, and 16-gene risk score and major clinical covariate including gender, age, and tumor stage on testing cohorts (MSK and DFCI).

Variable*	***P*-value**	**Hazard Ratio (95% CI)** $^{\Psi}$	
Analysis without gene signature risk score			
Gender (Male)	0.22	1.34	(0.84-2.16)
Age at diagnosis (>60)	0.08	1.61	(0.95-2.74)
Cancer Stage			
Stage II	6.25E-05	2.91	(1.72-4.91)
Stage III	1.09E-05	4.16	(2.20-7.85)
Analysis with 12-gene risk score			
Gender (Male)	0.17	1.40	(0.87-2.26)
Age at diagnosis (> 60)	0.29	1.34	(0.78-2.31)
Cancer Stage			
Stage II	3.47E-04	2.61	(1.54-4.43)
Stage III	7.40E-06	4.31	(2.28-8.16)
12-gene risk score	**3.10E-05**	**3.94**	**(2.07-7.52)**
Analysis with 15-gene risk score			
Gender (Male)	0.20	1.36	(0.85-2.18)
Age at diagnosis (> 60)	0.08	1.60	(0.94-2.74)
Cancer Stage			
Stage II	1.32E-04	2.80	(1.65-4.74)
Stage III	4.82E-05	3.73	(1.98-7.05)
15-gene risk score	**2.84E-04**	**1.99**	**(1.37-2.89)**
Analysis with 16-gene risk score			
Gender (Male)	0.11	1.49	(0.92-2.41)
Age at diagnosis (> 60)	0.18	1.44	(0.84-2.48)
Cancer Stage			
Stage II	5.36E-05	2.97	(1.75-5.03)
Stage III	7.52E-07	5.19	(2.70-9.96)
16-gene risk score	**6.24E-07**	**2.50**	**(1.33-3.59)**

*Gender was binary variable (0 for female and 1 for male); age at diagnosis was a binary variable (0 for < 60 years old and 1 otherwise); tumor stage was categorical variable of 3 categories (Stage I [as the reference group], Stage II, and Stage III). Risk score was continuous variable; where hazard ratio describes the relative risk between the mean risk scores of high-risk and low-risk groups. $^{\Psi}$ denotes confidence interval.

Table 3.7. Multivariate Cox proportional analysis of all available clinical covariates and 12-, 15-, and 16-gene risk score on testing cohorts (DFCI and MSK).

Variable*	***P*-value**	**Hazard Ratio (95% CI)$^{\Psi}$**	
Analysis without 12-gene risk score			
Gender (Male)	0.43	1.22	(0.74-1.99)
Age at diagnosis (>60)	0.05	1.70	(0.99-2.92)
Race			
Others/Unknown	0.28	0.43	(0.09-1.97)
White	0.10	0.28	(0.06-1.28)
Tumor differentiation			
Moderately differentiated	0.14	0.53	(0.23-1.24)
Poorly differentiated	0.70	1.17	(0.53-2.61)
Smoking history			
Smokers	0.62	0.84	(0.43-1.66)
Unknown	0.91	0.89	(0.11-7.10)
Cancer Stage	3.31E-04	2.72	(1.57-4.69)
Stage II	2.38E-05	4.93	(2.35-10.33)
Stage III	0.43	1.22	(0.74-1.99)
Analysis with 12-gene risk score			
Gender (Male)	0.38	1.25	(0.76-2.08)
Age at diagnosis (>60)	0.12	1.56	(0.89-2.72)
Race			
Others/ Unknown	0.52	0.60	(0.13-2.77)
White	0.11	0.29	(0.07-1.32)
Tumor differentiation			
Moderately differentiated	0.17	0.56	(0.24-1.29)
Poorly differentiated	0.83	0.91	(0.41-2.06)
Smoking history			
Smokers	0.61	0.84	(0.43-1.64)
Unknown	0.79	0.75	(0.09-5.98)
Cancer Stage			
Stage II	1.37E-03	2.44	(1.41-4.22)
Stage III	5.12E-06	5.88	(2.75-12.58)
12-gene risk score	**6.34E-05**	**4.19**	**(2.08-8.46)**
Analysis with 15-gene risk score			
Gender (Male)	0.36	1.26	(0.77-2.06)
Age at diagnosis (>60)	0.04	1.75	(1.02-3.01)
Race			
Others/ Unknown	0.38	0.50	(0.11-2.31)
White	0.14	0.32	(0.07-1.45)
Tumor differentiation			

Moderately differentiated	0.16	0.55	(0.24-1.27)
Poorly differentiated	0.99	0.99	(0.44-2.23)
Smoking history			
Smokers	0.93	0.97	(0.49-1.91)
Unknown	0.85	1.22	(0.15-9.89)
Cancer Stage			
Stage II	2.61E-04	2.76	(1.60-4.77)
Stage III	5.19E-05	4.66	(2.21-9.82)
15-gene risk score	**2.47E-03**	**1.81**	**(1.23-2.65)**
Analysis with 16-gene risk score			
Gender (Male)	0.17	1.42	(0.86-2.35)
Age at diagnosis (>60)	0.09	1.63	(0.93-2.85)
Race			
Others/ Unknown	0.22	0.38	(0.08-1.77)
White	0.05	0.22	(0.05-1.00)
Tumor differentiation			
Moderately differentiated	0.16	0.55	(0.23-1.28)
Poorly differentiated	0.96	1.02	(0.45-2.30)
Smoking history			
Smokers	0.53	0.81	(0.41-1.59)
Unknown	0.96	0.94	(0.12-7.54)
Cancer Stage			
Stage II	2.37E-04	2.79	(1.62-4.83)
Stage III	2.09E-06	6.34	(2.96-13.58)
16-gene risk score	**7.49E-07**	**2.45**	**(1.72-3.50)**

* Gender was binary variable (0 for female and 1 for male); age at diagnosis was a binary variable (0 for < 60 years old and 1 otherwise); race was a categorical variable of 3 categories (African American [as the reference group], White, and Others [composed of Asian (5) , Hawaiian or Pacific Islander (1), and unknown]); tumor grade was categorical variable of 3 categories (Well [as the reference group], Moderately, and Poorly differentiate); Smoking history was a categorical variable of 3 categories (Non-smokers, Smokers, and Unknown); tumor stage was categorical variable of 3 categories (Stage I [as the reference group], Stage II, and Stage III). Risk score was continuous variable; where hazard ratio describes the relative risk between the mean risk scores of high-risk and low-risk groups.

$^{\Psi}$ denotes confidence interval.

3.10 Comparison with other Lung Cancer Gene Signatures

In the Director's Challenge study [2], prognostic classifiers were constructed with gene expression signatures alone or gene expression signatures combined with clinical covariates. Among twelve gene signatures analyzed in their study (Table A.1), the best signature was reported as “method A” (referred to as “Shedden A” in this study), which contains about 9,591 genes/probes. In order to compare the predictive performance of our prognostic models with their best model, the estimated hazard ratio and the concordance probability estimate (CPE) of the models were evaluated. Hazard ratios greater than 1 indicate that patients with high predicted risk scores have poor clinical outcome. CPE value close to 1 indicates that the model has strong predictive; the model has poor predictive power (comparable to random prediction) when the CPE value close to 0.5.

Results show that the proposed 12-gene signature has the highest hazard ratio and CPE in both test sets when compared to the gene signatures from Director’s Challenge Study [2] (Fig. 3.11A, 3.11B). Although the hazard ratios of the 15-gene signature in both test sets were slightly lower than the 12-gene, there was no significant difference between the two signatures. Comparatively, the 16-gene signature didn’t perform as well as the 12- and 15-gene signature because the hazard ratio was not significant in MSK test cohort (Fig 3.11A). In patient cohorts with stage I tumors only, the three identified signatures had comparative performance as the “method A” because each signature was able to generate significant hazard ratio in only one of the two test cohort (Fig. 3.11C).

Among the three signatures presented thus far, the 12-gene signature gave the best performance. Therefore, we further compared the 12-gene signature with other published lung cancer signatures. To evaluate the 12-gene signature with previously published 14 lung cancer signatures [2, 11, 12, 15-17, 29, 30, 109, 110] (Table A.2), Gene Set Enrichment Analysis (GSEA[5]) was used to assess the enrichment of these signatures on 5-year survival. The normalized enrichment score (*NES*) and its corresponding false discovery rate (*FDR*) associated with each gene signature were evaluated on all 442 samples used in this study. In general, a gene set with high *NES* and low *FDR* is desired, as it indicates that the gene set expresses diversely with respect to the clinical outcome and the finding is unlikely to be by chance. In comparison to 14 other published gene signatures, the 12-gene signature exhibits high enrichment in groups of patients survived 5 year or longer with significantly low *FDR* (absolute *NES* = 1.5; *FDR* < 0.10) (Fig. 3.12). In this analysis, the most enriched signature with the lowest *FDR* was SHEDDEN_MH of 244 genes (absolute *NES* = 2.00; *FDR* < 0.002). Overall, among the 15 gene

[5] http://broad.harvard.edu/gsea/

sets studied, the 12-gene signature is one of the best lung cancer signatures evaluated with GSEA.

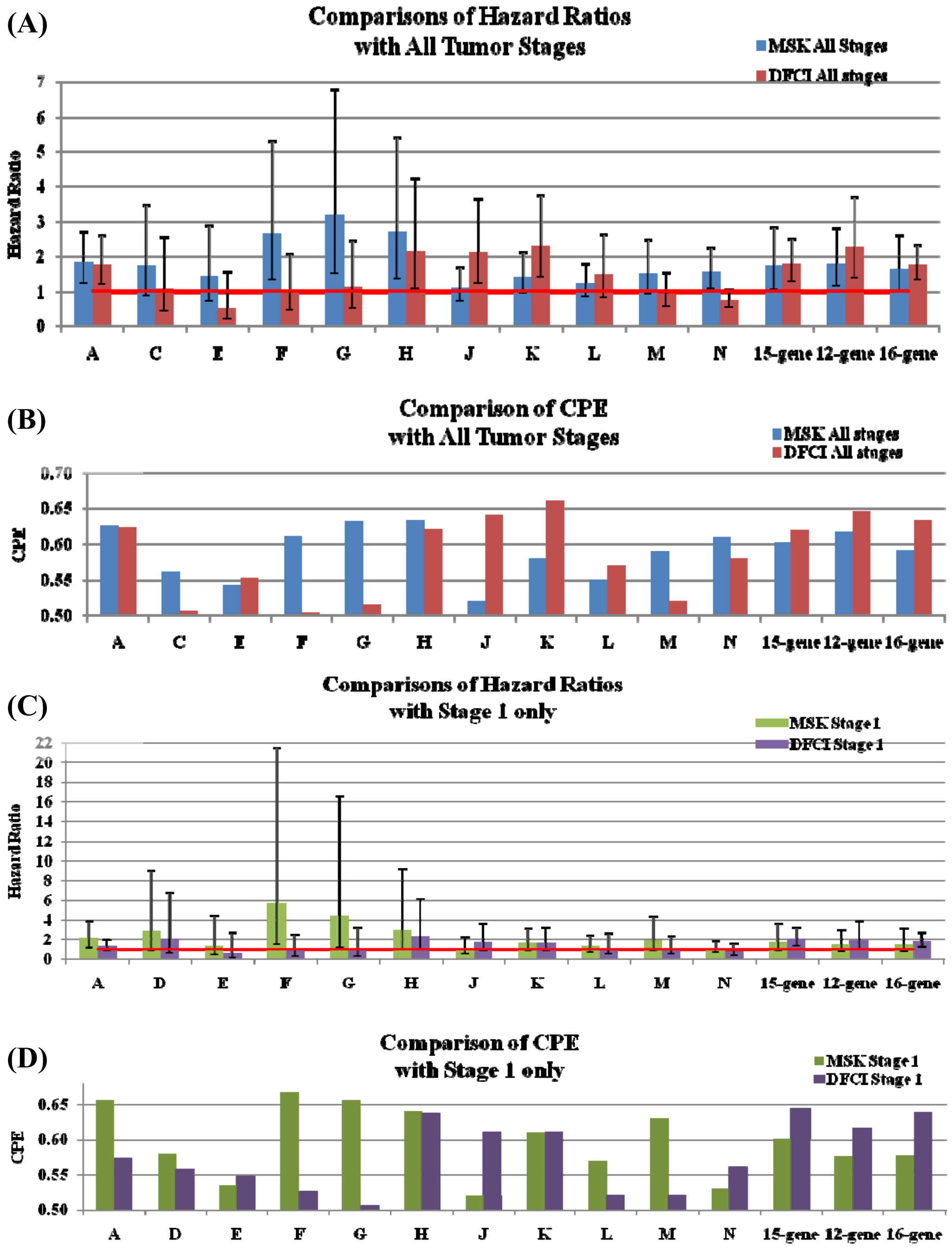


Figure 3.11. Evaluation of the 15-, 12-, and 16-gene prognostic models with molecular prognostic models presented by Shedden et al. [2]. Hazard ratio (A, C) and concordance probability estimate (CPE) (B, D) were compared on patients in all stages (A, B) and stage I (C, D) of lung cancer. Error bars in (A) and (C) represent 95% confidence interval of hazard ratio.

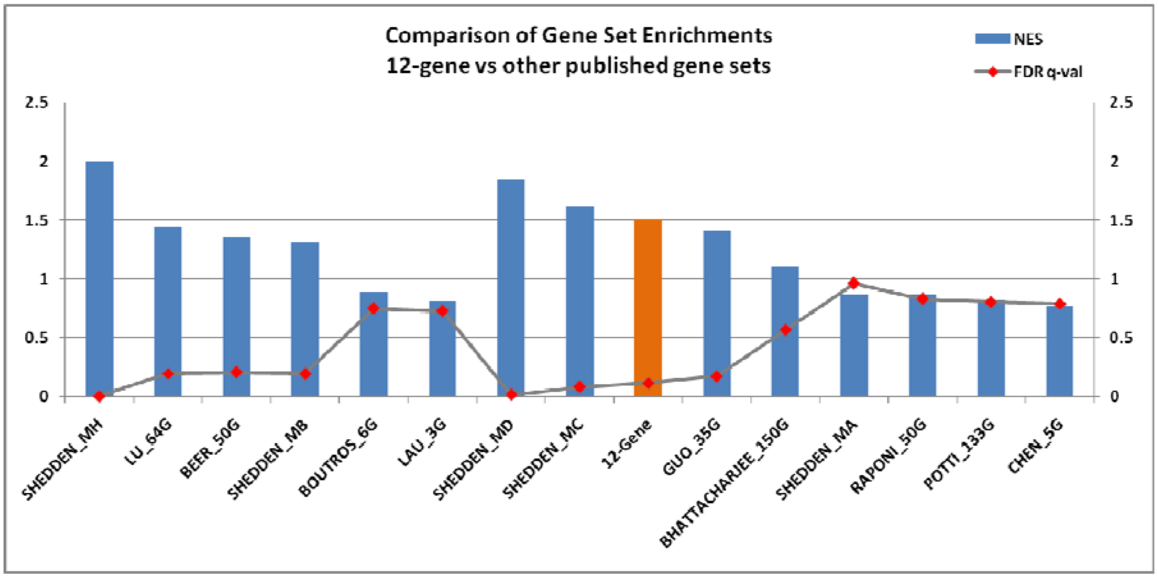
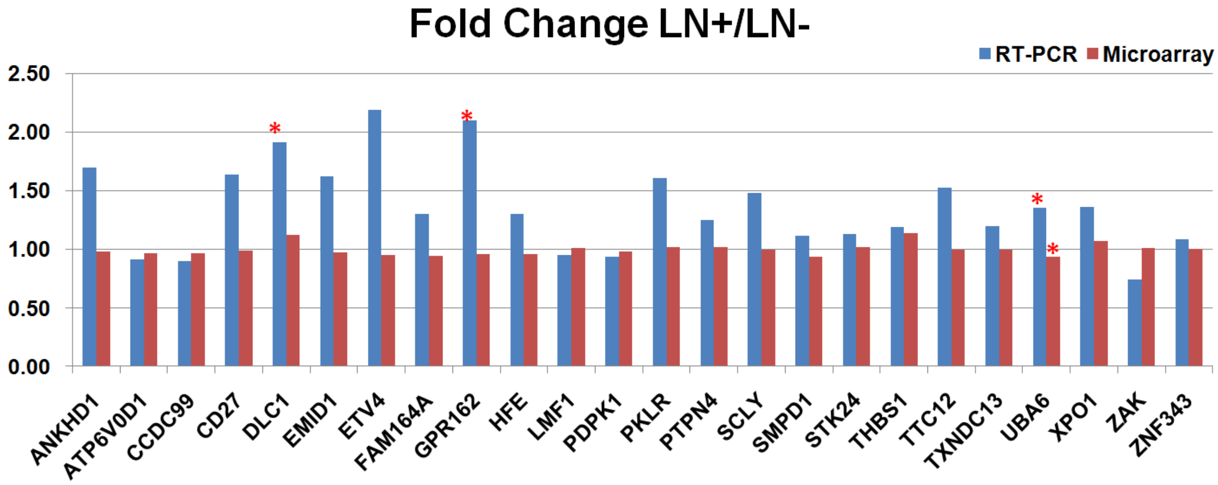


Figure 3.12. Gene set enrichment analysis of the 12-gene signature along with 14 published gene signatures for NSCLC.

3.11 RT-PCR Validation of Gene Expression Patterns

In order to further confirm the expression patterns of the 25 genes from the three signatures identified, RT-PCR microfluidic low density arrays were used to analyze independent NSCLC tumor samples. 91 NSCLC specimens obtained from West Virginia University Tissue Bank and the Cooperative Human Tissue Network (CHTN) (Ohio State University Tissue Bank, Columbus, OH) were analyzed.

First, the gene expression patterns for the 25 genes obtained from both microarray and RT-PCR were compared in terms of lymph node metastasis (Fig. 3.13A). On the RT-PCR data normalized with *POLR2A*, gene expression fold changes of the 25 genes in lymph node positive (LN+) versus lymph node negative (LN-) samples were compared with those in microarray data from Director's Challenge study [2]. The results show that the expression patterns of the 25 genes measured in both platforms are concordant in terms of lymph node metastasis.



Sample Size	LN- (N0)	N1	N2	LN+	Total
RT-PCR LDA	43	26	5	31	74
Microarray	299	87	53	140	439

Figure 3.13. Comparison of gene expression patterns of the 25 signature genes measured with DNA microarray and RT-PCR microfluidic low density arrays (LDA). Gene expression fold change in lymph node positive (LN+) patients vs. lymph node negative (LN-) patients was compared (A). Samples included in the fold change comparison are summarized in (B).

3.12 Prediction of Chemoresponse in NCI-60 Cell Lines

After demonstrating the promising performance of the 12-gene signature in predicting lung adenocarcinoma overall survival, we sought to explore whether the signature can predict chemoresponse to anti-lung cancer agents, including Cisplatin, Carboplatin, Paclitaxel, Etoposide, Erlotinib, and Gefitinib.

In this analysis, transcriptional gene expression profiles and activity profiles of various drugs used in chemotherapy in the NCI-60 cell lines [111] were used. The transcriptional gene expression profiles in all the 60 cell lines included in the study retrieved with CellMiner[6]. The data retrieved were generated on Affymetrix U133A and normalized using the *GCRMA* method

[6] http://discover.nci.nih.gov/cellminer

[112]. The drug activity profiles, measure in $\log_{10}$ (GI_{50}), were retrieved from Developmental Therapeutic Program at NCI/NIH through DTP Data Search[7]. The latest screening results for each studied drug were used in the analysis. The drug activity data was further processed to define drug resistance and sensitivity. Specifically, for each drug, $\log_{10}(GI_{50})$ values were first normalized across the 60 cell lines. Cell lines with $\log_{10}(GI_{50})$ at least 0.5 standard deviations (SDs) above the mean were defined as resistant to the drug. Those with $\log_{10}(GI_{50})$ at least 0.5 SDs below the mean were defined as sensitive to the drug. The remaining cell lines with $\log_{10}(GI_{50})$ within 0.5 SDs were defined as intermediate [113, 114].

For each drug, cancer cell lines that are either sensitive or resistant to the drug were included to build a chemoresponse classifier based on the 12-gene expression profiles in the cell lines. The performance of the classifiers was evaluated with leave-one-out cross validation (Table 3.8). Statistical significance of the classification was evaluated by comparing the overall accuracy of the 12-gene signature with that of 1000 random signatures of the same size using the same algorithm. Result show that the overall prediction accuracy of chemoresponse was 81% ($P < 0.004$) for Paclitaxel (Taxol), 78% ($P < 0.001$) for Carboplatin, 80% ($P < 0.005$) for Cisplatin, 73% ($P < 0.017$) for Etoposide, 79% ($P < 0.001$) for Erlotinib, and 94% ($P < 0.001$) for Gefitinib. These results demonstrate that the 12-gene signature accurately predicted sensitivity and resistance to common lung cancer chemotherapeutic agents in cancer cell lines.

Table 3.8. Prediction accuracy of chemoresponse in NCI-60 cell lines using the 12-gene signature.

Drug	Sensitivity (chemoresistance)	Specificity (chemosensitivity)	Overall accuracy	*P*-value*
Carboplatin	76% (19/25)	80% (16/20)	78% (35/45)	< 0.001
Paclitaxel	72% (8/11)	87% (13/15)	81% (21/26)	0.004
Cisplatin	85% (22/26)	74% (14/19)	80% (36/45)	0.005
Etoposide	80% (16/20)	67% (14/21)	73% (30/41)	0.017
Erlotinib	79% (11/14)	80% (16/20)	79% (27/34)	0.001
Gefitinib	92% (11/12)	95% (20/21)	94% (31/33)	< 0.001

* A *P*-value < 0.05 represents that the overall accuracy of the 12-gene signature is significantly higher than that of random gene signatures with the same size using the same classifier in 1000 tests

[7] http://dtp.nci.nih.gov/dtpstandard/dwindex/index.jsp

The differential expression in sensitive and resistant lung cancer cell lines was also analyzed for each signature gene. The drug responses of the lung cancer cell lines in the NCI-60 panel were provided in Table 3.9.

Among the signature genes, the over-expression of *STK24* was linked to chemoresistance to all the studied drugs except Gefitinib in the lung cancer cell lines; whereas the over-expression of *FAM14A* was associated with chemosensitivity to all the studied drugs except Gefitinib in lung cancer cell lines. The under-expression of *STK24* was associated with resistance to Gefitinib (*P* < 0.05). The under-expression of *CCDC99* was observed in resistance to Paclitaxel (*P* < 0.05). The over-expression of *DLC1* was associated with chemoresistance to Erlotinib (*P* < 0.05), Paclitaxel, and Cisplatin; whereas its under-expression was associated with chemoresistance to Etoposide and Carboplatin (not statistically significant) (Fig. 3.14).

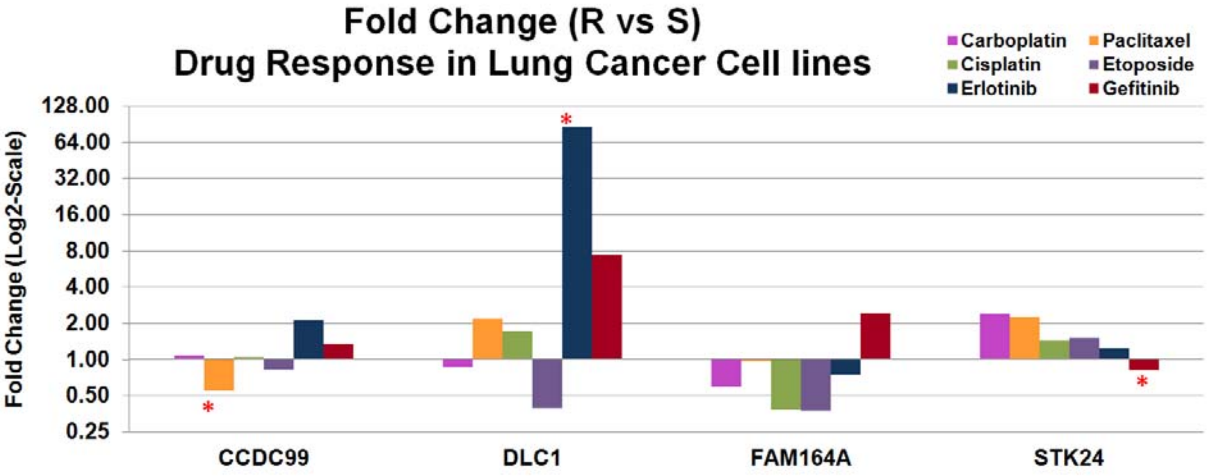


Figure 3.14. Genes with at least 1.5-fold expression fold change in resistant vs. sensitive lung cancer cell lines to six anticancer drugs. In the graph, differential expression with statistical significance (P < 0.05, t-tests) is marked by a red asterisk.

Table 3.9. Machine learning algorithm and genes used in chemoresponse prediction using 12-gene signature.

Anti-cancer Agent	Machine learning algorithm	Genes Selected	Resistant lung cancer cell lines	Sensitive lung cancer cell lines
Carboplatin	RBF Network (seed = 2)	ATP6V0D1 CCDC99 FAM164A LMF1 PDPK1 PKLR SCLY SMPD1 STK24 XPO1	LC:EKVX LC:NCI_H322M	LC:NCI_H460 LC:NCI_H522 (LC:NCI_H23 not included due to missing values)
Paclitaxel	IBK (k=3)	CCDC99 DLC1 LMF1 PKLR SMPD1 XPO1 ZAK	LC:HOP_92 LC_EKVX	LC:NCI_H460 LC:NCI_H522
Cisplatin	Decorate (PART as base learner)	ATP6V0D1 CCDC99 FAM164A LMF1	LC:NCI_H226 LC:EKVX LC:NCI_H322M	LC:HOP_62 LC:NCI_H460 (LC:NCI_H23 not included due to missing values)
Etoposide	AdaBoostM1 (seed = 2, Random Tree as base learner)	CCDC99 LMF1 SCLY STK24 XPO1	LC:EKVX LC:NCI_H322M	LC:HOP_62 LC:NIC_H460
Erlotinib	RBF Network	DLC1 LMF1 XPO1 SMPD1 STK24 PDPK1 ZAK PKLR CCDC99	LC:NCI_H226 (LC:NCI_H23 not included due to missing values)	LC:EKVX LC:NCI_H322M LC:NCI_H522

Gefitinib	Multilayer Perceptron (seed=2, learning rate=0.4)	ATP6V0D1 SMPD1 XPO1 PKLR STK24 SCLY	LC:A549 LC:HOP_62 LC:HOP_92 LC:NCI_H226 (LC:NCI_H23 not included due to missing values)	LC:EKVX LC:NCI_H322M

3.13 Functional Pathway Analysis of 12-gene Signature

Having established the clinical relevance of the 12-gene prognostic signature, we sought to explore the functional involvement of this gene set in lung tumorigenesis and tumor progression. Two functional pathway analysis tools, Ingenuity Pathway Analysis (IPA) and Pathway Studio 7.0, were used to obtain molecular interaction related to the 12 genes reported in literature. Results from IPA show that the signature genes interact with major cancer signaling pathways, such as *TNF* and *AKT* (Fig. 3.15A). In the study with Pathway Studio 7.0, interactions among the 12 genes and 13 major lung cancer hallmarks (*EGF, EGFR, KRAS, MET, RB1, TP53, E2F1, E2F2, E2F3, E2F4, E2F5, AKT1,* and *TNF*) reported in the literature were explored. Results from Pathway Studio revealed various types of interactions ranging from regulation to protein modification among the 12 genes and eight out of 13 cancer hallmarks (Fig. 3.15B). Results from both functional pathway analyses suggest that the 12 signature genes are involved in lung cancer oncogenesis and tumor progression.

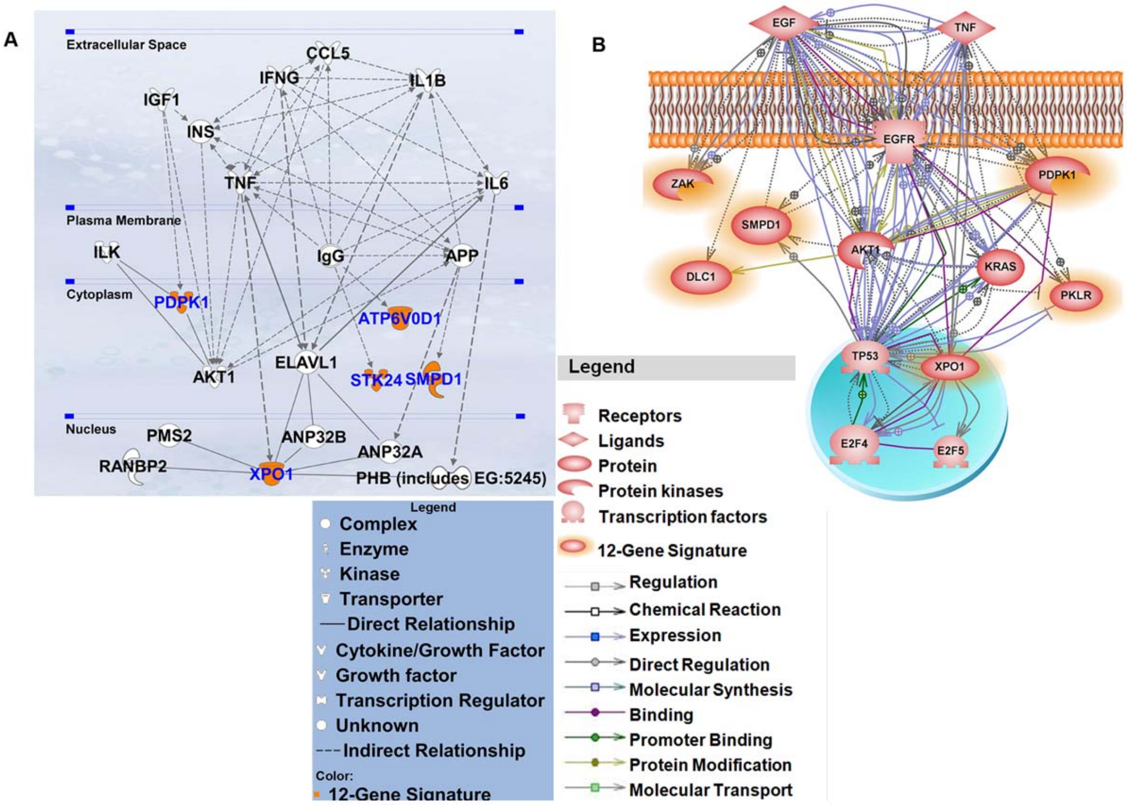


Figure 3.15. Functional pathway analysis of the 12 signature genes. (A) Using core analysis from Ingenuity Pathway Analysis (IPA), curated interactions were revealed among the identified signature genes and major lung cancer signaling pathways. (B) Six of the 12 genes also exhibited various curated interactions with eight prominent lung cancer hallmarks with Pathway Studio 7.0.

3.14 Conclusions and Discussions

Gene signatures are essential for the development of personalized medicine for precise lung cancer prognosis. With the availability of genome-wide profiles in the post-genomic era, innovative computational models are needed to identify clinically important gene markers. Given the current scale of high throughput data with thousands of genes, traditional methods for gene selections would not be adequate. Instead, a hybrid system with combinatorial gene selection scheme of different gene filtering methods at different stages is needed. This study presents a hybrid model system for the identification of gene signatures for lung cancer prognosis. The hybrid model systems identified three signatures: a 15-gene, a 12-gene, and a 16-gene signature.

In the hybrid model, SAM and *t*-tests was used to identify candidate genes showing differential expression between two prognostic groups in the training set. SAM controls for multiple testing problem and is very similar to *t*-test. We used *t*-tests ($P < 0.01$) to select genes with certain level of differential expression between two prognostic groups, and used SAM to control for false discovery rate ($FDR < 25\%$). The results from SAM and *t*-test are not exactly the same, because the SAM method adds a constant (s) in the denominator to ensure that genes with a very small variance in the samples and a small differential expression are not selected as significant markers. When s=0, SAM is exactly the same as *t*-test [34]. This hybrid system was able to identify a small set of genes that are more accurate than previously published lung cancer gene signatures on the same datasets. We have experimented to stringent the threshold in SAM statistics. As a result, there were 87 genes with a FDR <10% and no genes were selected with a FDR < 1% from the training set. The 87 genes were not able to generate significant stratification in all three patient cohorts. These results indicate that using SAM method alone is not sufficient to identify the most accurate prognostic gene signature.

Among the 12- and 15-gene signatures identified using *t*-tests, SAM, and *Relief*, 16 genes share common biological functions (Table 3.3). The performance of these gene signatures is comparable to one another in term of Kaplan-Meier analyses, hazard ratio of the prognostic model, and multivariate analyses with clinical covariate and demographic factors. When 3-year survival was used to define high- and low-risk groups (high-risk: death within 3-y; low-risk: alive after 3-y), the 12-gene risk algorithm achieved a sensitivity (correctly predicted high-risk patients) of 73.65% in the training set, 86.96% in MSK, and 68.18% in DFCI, and a specificity (correctly predicted low-risk patients) of 59.21% in the training set, 57.75% in MSK, and 76.36% in DFCI (Table 3.10). The sensitivity and specificity of the 15-gene signature in predicting 3-year survival was also similar as the 12-gene signature, with sensitivity of 76.84%, 82.61%, 86.36% and specificity of 64.47%, 50.70%, 47.27% in training, MSK, and DFCI respectively. Compared to 12- and 15-gene signature, the 16-gene signature gave lower sensitivity but higher specificity (Table 3.10).

Table 3.10. Sensitivity and specificity of the 12-, 15- and 16-gene prognostic models.

	Sensitivity (% of correctly predicted high-risk patients)				**Specificity (% of correctly predicted low-risk patients)**			
	n	**12-gene**	**15-gene**	**16-gene**	***n***	**12-gene**	**15-gene**	**16-gene**
3-year survival as the cutoff (high-risk: death within 3-y; low-risk: alive after 3-y)								
UM & HLM	95	73.65	76.84	47.37	152	59.21	64.47	87.50
MSK	23	86.96	82.61	60.87	71	57.75	50.70	70.42
DFCI	22	68.18	86.36	54.55	55	76.36	47.27	81.82
5-year survival as the cutoff (high-risk: death within 5-y; low-risk: alive after 5-y)								
UM & HLM	125	72.80	72.80	44.80	104	66.35	69.23	93.27
MSK	34	70.59	67.65	50.00	31	48.39	41.94	67.74
DFCI	28	64.29	78.57	50.00	36	77.78	47.22	86.11
2.5-year and 5-year survival as the high- and low-risk cutoffs (high-risk: death within 2.5-y; low-risk: alive after 5-y)								
UM & HLM	84	75.00	77.38	48.81	104	66.35	69.23	93.27
MSK	21	95.24	85.71	66.67	31	48.39	41.94	67.74
DFCI	20	70.00	85.00	55.00	36	77.78	47.22	86.11

According to the hazard ratio of the prognostic model for the three signatures in both test cohorts, the 12-gene signature exhibited highest potential for lung cancer prognosis as it was the only signature generated significant hazard ratio in stage I patients of both test cohorts. In addition, the 12-gene signature accurately quantifies survival in patients in all stages, stage I only, stage IB only, and patients in stage I or II who did not receive chemotherapy. The 12-gene expression-defined risk score is a more accurate prognostic factor than commonly used clinical parameters. Due to the high prognostication performance, chemoresponse prediction was further studied using the 12-gene signature. Results show that the signature also predicts chemoresistance and chemosensitivity to several major anti-lung cancer drugs in NCI-60 cancer cell lines. Together, the results indicate that the 12-gene signature could be used to select early stage lung adenocarcinoma patients at high risk for tumor recurrence for adjuvant chemotherapy. Meanwhile, it may spare stage I and II low-risk patients from unnecessary chemotherapy. Furthermore, the 12-gene signature has the potential to be used to inform physicians which anticancer drugs should be used in treating a particular patient. The expression patterns of the 12-gene signature were confirmed in RT-PCR. Curated interactions between the signature genes and major cancer signaling hallmarks revealed in the functional pathway analysis provides further evidence that the 12-gene signature might be involved in lung cancer oncogenesis and tumor progression.

Overall, the combinatorial gene selection scheme presented in this study identified 25 prognostic genes. This study demonstrates that combination of different stages of gene filtering identified gene signatures with higher prognostic performance than traditional gene selection approach. Feature selection algorithms included in the system is crucial not only to reduce the size of the identified signatures but also to provide a set of genes with strong prognostic classification. The choice to use a different feature selection technique depends on an evaluation with an independent classifier. If the classification performance cannot be further improved with the current algorithm, a different feature selection algorithm should be used. In conclusion, hybrid models with combination of statistics and feature selection methods are efficient, robust, and could identify prognostic gene signatures feasible for clinical utility.

Chapter 4

Network-based Models for Lung Cancer Prognostic Signatures Identification

With the completion of the Human Genome Project, cataloging the “parts list” of disease genes is no longer the focus of biomarker identification. Understanding the networks of interactions that take place among the genes has become the new emphasis to identify marker genes because the gene networks provide insights to unravel the molecular basis of disease [27]. Molecular network analysis had been shown to be useful in disease classification [61] and identification of novel therapeutic targets [115]. Nonetheless, the development of efficient methods for constructing genome-wide coexpression networks and the identification of a particular set of markers, from among the enormous number of potential markers, that has the highest predictive ability for disease outcome remains the challenges for this research domain [7].

We had demonstrated that the combinatorial framework with multiple gene filtering layers identify better prognostic genes signatures than traditional methods when being applied alone. In this chapter, we will present another hybrid system that is built upon a computational network model for the identification of lung cancer prognostic signatures. The network model incorporated in the hybrid system is the implication networks induced from prediction logic [1]. With this network-based system, users could specify the signaling pathways and identify signature genes that are tightly related to the set of signaling proteins in that particular pathway. This presents an efficient framework for scientists to retrieve prognostic genes from the disease-mediated coexpression networks linked to signaling pathways. By combining additional layers of gene selection methods after retrieving prognostic genes from the modeled networks, sets of gene signatures with strong prognostic classification performance were identified.

The remainder of the chapter is organized as follows. Section 4.1 illustrates the methodology of the proposed system and the identification of extensive prognostic gene signatures for lung cancer. Section 4.2 presents the prognostic performance evaluation of the identified signatures. Comparison of the identified signatures with all published lung cancer gene signatures is discussed in Section 4.3. Section 4.4 describes the construction of molecular prognostic classifiers and the performance using a particular signature identified, i.e. the 10-gene signature. Functional pathway analysis was carried out to study the biological aspect of the identified 10 genes to lung cancer oncogenesis and will be presented in Section 4.5. Section 4.6 presents the evaluation of the disease-mediated coexpression networks derived using the implication network algorithm. The last section, Section 4.7 concludes the chapter.

4.1 Methodology

The methodology is based on the genome-wide coexpression networks modeled with the implication networks. The implication induction algorithm (Fig. 2.7) was used to construct pair-wise genome-scale coexpression networks for predicting risk from developing recurrence in lung cancer. The methodology was motivated by the hypothesis that the combined analysis of disease-mediated genome-wide coexpression networks, signaling pathways, and clinical approaches would lead to prognostic biomarker for more informed clinical use.

Patient samples from the largest public lung cancer microarray data published by Shedden et al. [2] were used in this study. Training set was formed with patient samples from UM and HLM (n = 256), whereas samples from MSK (n = 104) and DFCI (n = 82) constituted two independent test sets. Data preprocessing were done before the analysis. First, whenever a gene has missing measurements in at least half of the samples, the gene were removed from the analysis. Then, for genes measured using multiple probes, the average expression of the duplicates was used to represent the expression profile of the unique gene. This gave a final set of 12,566 unique genes for the implication network analysis.

To construct implication networks, the mean expression of each gene in a patient cohort was used as a cut-off to partition the expression profiles. If the expression of a gene in a patient sample was greater than the mean in the cohort, this gene was denoted as *up-regulated* in this tumor sample; otherwise, it was denoted as *down-regulated* in the tumor sample. In the training set, patients who died within 5 years were labeled as poor-prognosis (n = 125), and those who survived 5 years after surgery were labeled as good-prognosis (n = 104). Censored cases (those with follow-up of less than 5 years) were removed from the analysis (n = 27). For each patient group in the training set, a genome-wide coexpression network was constructed using the implication induction algorithm. Between each pair of genes, possible significant ($P < 0.05$; one-sided z-tests) coexpression relations were derived in each patient group, constituting disease-mediated gene coexpression networks. By comparing the implication rules connecting each pair of nodes between the two networks, disease-specific differential network components were identified. These differential components contain the coexpression relations that were either present in the poor-prognosis group but missing in the good-prognosis group, or conversely, those present in the good-prognosis group but missing in the poor-prognosis group (Fig. 4.1).

Next, candidate genes were obtained by retrieving genes displaying a direct significant ($P < 0.05$, z-tests) co-regulation relation with major NSCLC signal proteins from the differential components associated with each prognosis group. From the human NSCLC signaling pathways delineated by the KEGG pathway database[8], 11 signaling proteins (*TP53*, *MET*, *RB1*, *EGF*, *EGFR*, *KRAS*, *E2F1, E2F2, E2F3, E2F4,* and *E2F5*) were included in this study. To analyze the performance of methodology, candidate genes with significant coexpression relations with any combination of 6 or 7 signaling proteins were included for further analysis (Fig. 4.1).

Three approaches were taken to identify gene signatures from the pool of candidate genes. In the first approach, probes with significant association with survival ($P < 0.05$, univariate Cox model) were identified as signature genes. In the second approach, random forests were used to obtain a refined set of signature genes from the significant probes ($P < 0.05$; univariate Cox model). In the third approach, *Relief* algorithm was used to rank the significant probes ($P < 0.05$; univariate Cox model), and a step-wise forward selection was used to identified the final gene signatures. Specifically, starting from the top ranked gene, one gene

[8] http://www.genome.jp/kegg/pathway/hsa/hsa05223.html

was added at each step to the gene set, until the prognostic accuracy could not be improved by adding more genes. The final gene set was identified as the gene signature. Fig. 4.1 gives an overview of the methodology.

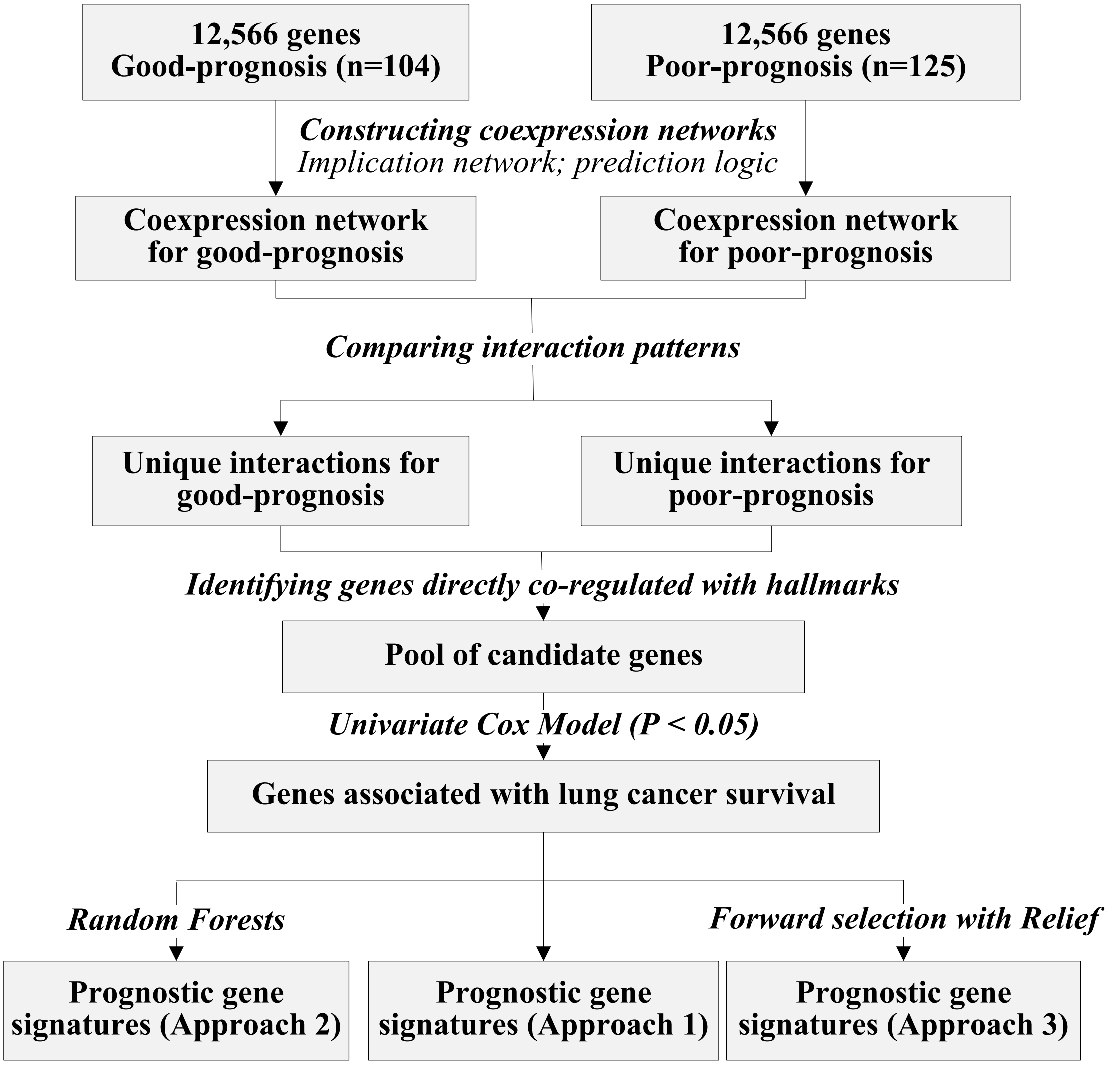

12,566 genes
Good-prognosis (n=104)
12,566 genes
Poor-prognosis (n=125)
Constructing coexpression networks
Implication network; prediction logic
Coexpression network for good-prognosis
Coexpression network for poor-prognosis
Comparing interaction patterns
Unique interactions for good-prognosis
Unique interactions for poor-prognosis
Identifying genes directly co-regulated with hallmarks
Pool of candidate genes
Univariate Cox Model (P < 0.05)
Genes associated with lung cancer survival
Random Forests
Forward selection with Relief
Prognostic gene signatures (Approach 2)
Prognostic gene signatures (Approach 1)
Prognostic gene signatures (Approach 3)

Figure 4.1. Overview of the study design for identifying prognostic gene signatures with implication networks and feature selection methods.

4.2 Evaluation of Identified Prognostic Gene Signatures

To evaluate if the identified signatures could provide accurate prognostic prediction for lung adenocarcinoma, multivariate Cox proportional hazard model was used to construct prognostic classifiers to stratify patients. On training samples, gene expressions of the identified prognostic signature genes were fitted to the Cox proportional hazard models as covariates. Coefficients obtained for each covariate in the constructed model were used to represent the training model. Using the training model, a survival risk score was generated for each patient. From the training risk scores, a cut-off value was identified to stratify patients into high- or low-risk groups. The model and cutoff values defined using the training set were applied to the independent test sets without re-estimating parameters. The prognostic performance of each identified gene signature was evaluated according to the following criteria: log-rank tests in Kaplan-Meier analyses and hazard ratio of death from lung cancer for all cancer stages, for stage I only and for stage I without receiving chemotherapy in training and test cohorts. The prognostic performance of patients from all tumor stages was evaluated on the two independent test sets individually. Due to small sample size, the two independent test sets were combined while evaluating the prognostic performance for stage I and stage I without receiving chemotherapy.

In the first approach, among the 462 sets of candidate genes that co-regulated with 6 signaling proteins, 9 gene signatures generated significant stratification (log-rank P <0.05) with significant hazard ratios ($P < 0.05$) in all three patient cohorts (Table 4.1). Among these 9 gene signatures, 5 of them also had significant hazard ratios (P <0.05) on stage I patients in all three cohorts. Among the 5 gene signatures that could give accurate prognostic categorization in all stages and stage I tumors, 4 gene signatures (referred to as S1-S4; Table B.1) generated significant stratifications (log-rank P <0.05 in Kaplan-Meier analysis, with hazard ratio significantly greater than 1) for stage I patients without receiving chemotherapy (Table 4.1). Similarly, among the 330 sets of candidate genes co-regulated with 7 signaling proteins in the first approach, 4 gene signatures generated accurate prognostic stratification (log-rank P <0.05 in Kaplan-Meier analysis, with hazard ratio significantly greater than 1) in all three patient cohorts, and one of them also generated accurate prognostic prediction in stage I patients in all three datasets (Table 4.1). In the second approach, only 1 gene signature that co-regulated with 7 signaling proteins (referred to as S5; Table B.2) provided significant stratifications in patients with all tumor stages, stage I only, and stage I without receiving chemotherapy (Table 4.1). The third approach identified 16 such gene signatures (referred to as S6 to S21; Table B.3) from the candidate genes co-regulated with 6 signaling proteins.

In summary, a total of 21 gene signatures were identified using the three approaches in this study, which, in turn, generated significant prognostic categorizations in lung adenocarcinoma patients with all cancer stages, stage I only, and stage I without chemotherapy (Table 4.1). These results demonstrate that the methodology provides a platform to efficiently identify prognostic gene signatures which also co-regulate with major signaling proteins for lung adenocarcinomas. Most importantly, the size of these gene signatures (4 ~ 33 genes) is feasible to be further validated with biology experiments and used for clinical application.

Table 4.1. Summary of prognostic signature discovered using the methodology in the extensive study.

Gene Selection Approach	**Number of signaling hallmarks**	**No. of signatures giving significant stratifications (log-rank *P*< 0.05) in all dataset**		
		with significant hazard ratio in all stages	**& with significant hazard ratio in stage I ***	**& with significant hazard ratio in Stage I without chemotherapy #**
Network-based (Approach 1)	**7**	4	1	0
	6	9	5	4
	Average signature size	21 genes	21 genes	24 genes
Network + Random Forests (Approach 2)	**7**	4	4	1
	6	3	2	0
	Average signature size	12 genes	10 genes	5 genes
Network + Relief (Approach 3)	**7**	7	4	0
	6	47	26	16
	Average signature size	14 genes	12 genes	14 genes
Summary	**Total number of signatures**	**74**	**42**	**21**

* Gene signatures in this column also had significant hazard ratio in all cancer stages in all three patient cohorts.
# Gene signatures in this column also have significant hazard ratio in all cancer stages in all three patient cohorts and stage I in training and combined test cohorts.

4.3 Comparison with other Lung Cancer Gene Signatures

To further investigated the prognostic performance of the 21 prognostic signatures identified from the proposed methodology, we compared the signatures with gene expression-based lung cancer signatures reported to date. Eleven lung cancer gene signatures were evaluated in the Director's Challenge Study [2], among which five of them were identified from previous studies on lung cancer molecular prognosis [15, 16]. Among the 11 gene signatures evaluated, the best signature reported was "method A" (referred to as "A" in Fig. 4), which contains about 9,591 genes/probes. The prognostic performance of our gene signatures was compared with the best lung cancer gene signatures reported to date in terms of the estimated hazard ratio and the concordance probability estimate (CPE) in two test sets (Fig. 4.2A and 4.2B).

Results from the comparison show that the 21 gene signatures (S1-S21) identified in this study perform better than all other previously identified lung cancer gene signatures (Fig. 4.2). Among the 11 previously identified gene signatures, "method A" is the only model with hazard ratio significantly ($P < 0.05$) greater than 1 in all three patient cohorts (Fig. 4.2A). On the other hand, all 21 gene signatures identified from the proposed system generated hazard ratio significantly ($P < 0.05$) greater than 1 in all three patient cohorts (Fig. 4.2A). Moreover, all 21 gene signatures had a significant hazard ratio and a CPE significantly greater than 0.5 ($P<0.05$) in stage I patients (Fig. 4.2C-4.2D). Most significantly, the 21 identified signatures also had significant hazard ratio and CPE in stage I patients without receiving chemotherapy (Fig. 4.2E-4.2F), which is a prognostic capacity which has not been reported in the previous studies [2, 15, 16].

These results demonstrate that the gene signatures discovered with the network-based methodology are clinically important in identifying specific high-risk patients diagnosed with early stage lung adenocarcinoma for adjuvant chemotherapy.

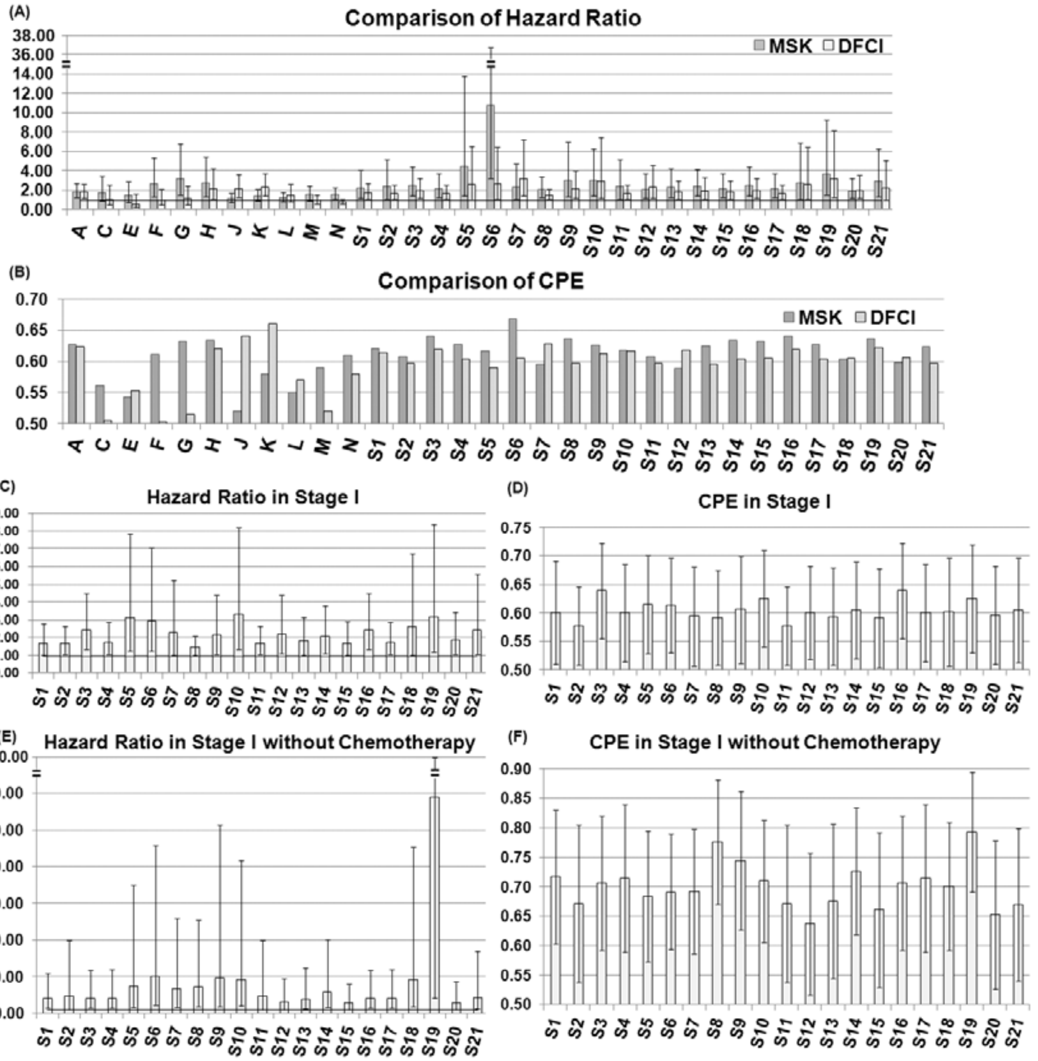
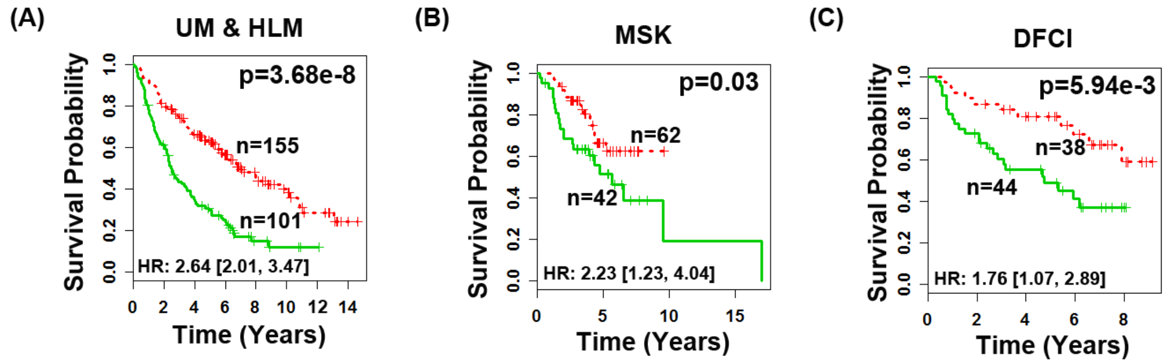


Figure 4.2. **Comparison of 21 identified gene signatures with other lung cancer gene signatures.** The 21 prognostic gene signatures were compared with 11 gene signatures evaluated in the Director's Challenge Study [2] in two test sets in terms of hazard ratio (A) and concordance probability estimate [CPE] (B). The prognostic performance of the 21 gene signatures was evaluated for stage I patients by hazard ratio (C) and CPE (D), as well as for stage I patients without receiving chemotherapy in the combined test cohorts (E, F). The error bar in the charts represents 95% confidence interval of the measurement.

4.4 Survival Prediction Using the Identified 10-gene Prognostic Signature

To confirm the prognostic performance of the identified signatures, we further evaluated one of the 21 identified signatures. A 10-gene signature identified using the third approach (S13; was Table B.3) was selected for further evaluation. From the disease-mediated prognosis groups, 154 candidate prognostic genes showed direct coexpression with signaling proteins *EGF*, *KRAS*, *TP53*, *RB1*, *E2F1*, and *E2F2*; in which 57 were identified from the good-prognosis group and 106 were identified from the poor-prognosis group (with 9 genes common in both groups). From the training set of the original continuous microarray data, 26 probes out of these 154 genes were significantly associated with overall survival ($P < 0.05$, univariate Cox mode). Based on the forward selection and ranking with *Relief* [116], the top 10 genes were identified as the final signature (S13; Table B.3).

Multivariate Cox proportional hazard model was fitted with the 10 genes as covariates on bootstrapped training samples for 1,000 times. The average of the 1,000 coefficients obtained for each covariate was used to represent the final coefficients in the training model. Using the training model, a survival risk score was generated for each patient. A risk score of -12.04 was identified as a cut-off value for patient stratification in the training set (Fig. 4.3A). This training model and cut-off value was then applied to the two validation sets to generate prognostic categorization without re-estimating parameters (Fig. 4.3B and 4.3C). In all three patient cohorts, this scheme stratified patients into two prognostic groups with significantly distinct survival outcome (log-rank $P < 0.03$, Kaplan-Meier analyses). When the high-risk group is defined as a group of patients who survived 5 years or less, and the low-risk group with patients who survived 5 years or longer, this model accurately classify 64% of the patients on training, 57% on MSK and 66% on DFCI. The model also achieved sensitivity (correctly predicted high-risk patients) of 55.20% on the training set, 52.94% on MSK, and 75% on DFCI. The specificity (correctly predicted low-risk patients) was 75% on the training set, 61.29 % on MSK, and 58.33% on DFCI (Fig. 4.3D).



(D)

5-year Survival Prediction	UM & HLM	MSK	DFCI
Accuracy (%)	64.19	56.92	65.63
Sensitivity (%)	55.20	52.94	75.00
Specificity (%)	75.00	61.29	58.33
CPE [95% CI]	0.65 [0.61, 0.69]	0.62 [0.54, 0.70]	0.60 [0.52, 0.68]

Figure 4.3. Prognostication of disease-specific survival using the 10-gene signature in lung adenocarcinoma patients. The model stratified patients into two prognostic groups with significantly different (P < 0.03) survival outcome in the training set UM&HLM (A) and both test sets MSK (B) and DFCI (C) in Kaplan-Meier analyses. Log-rank tests were used to assess the difference in survival probability between the two prognostic groups. Performance of 5-year survival prediction on training and two test sets (D).

Furthermore, the 10-gene prognostic signature could identify high-risk patients with stage I cancers on both the training set and combined test sets (log-rank $P \leq 0.007$; Fig. 4.4A-4.4B). The prognostic model also successfully separated high- and low-risk groups within stage IB patients in the training and combined test sets (log-rank $P \leq 0.04$; Fig. 4.4C-4.4D). In stage I patients who did not receive chemotherapy, the prognostic model stratified high- and low-risk groups with distinct survival outcome in both training and test sets (log-rank $P \leq 0.04$; Fig. 4.4E-4.4F). These results demonstrate that the 10-gene signature provides a more refined prognosis than the current AJCC staging system. Using this model, patients with stage I NSCLC could be advised to either receive or be spared from chemotherapy according to the expression profiles of the 10 prognostic genes.

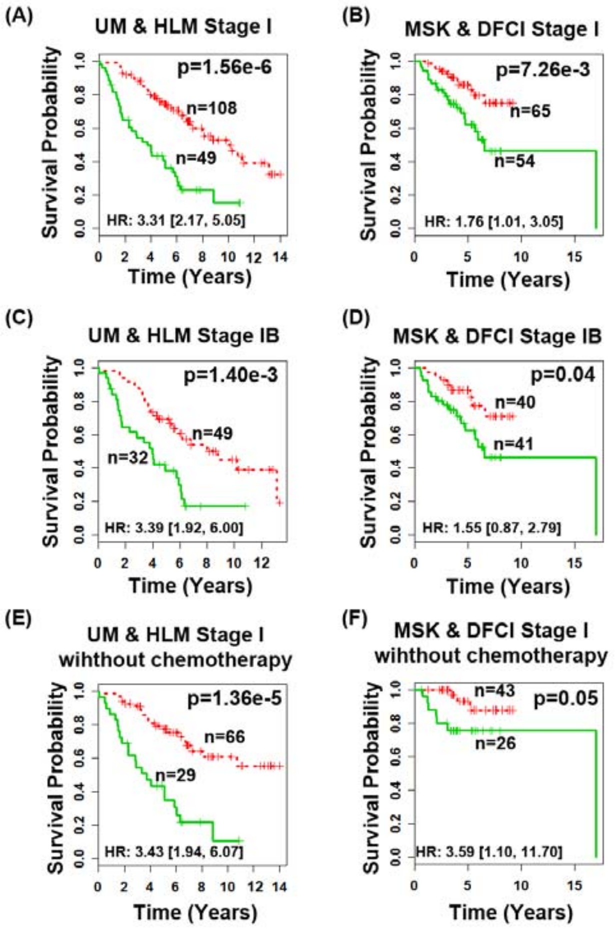


Figure 4.4. Prognostic performance of the 10-gene signature in stage I lung adenocarcinoma. The model generated significant prognostic categorization for stage I patients in both training set UM&HLM (A) and combined test sets MSK&DFCI (B), for stage IB patients in training (C) and combined test sets (D), as well as for stage I patients without receiving chemotherapy in both training (E) and combined test sets (F). Statistical significance of the difference in survival probability between the two prognostic groups was assessed with log-rank tests in Kaplan-Meier analyses.

4.5 Prognostic Evaluation with Clinical Covariates

To further validate the prognostic power of the model, the constructed 10-gene prognostic model was evaluated with common lung cancer prognostic factors using multivariate Cox analysis on the combined testing cohorts (MSK and DFCI). The constructed 10-gene risk score algorithm was evaluated using clinical factors, including gender, age, cancer stage, smoking history, race, and tumor differentiation. In the analysis without the 10-gene risk score, among major clinical

cal covariates age, gender and cancer stage, cancer stage was the only significant predictor of death from lung cancer (Table 4.2). After the 10-gene risk score was included, the gene risk score became a highly significant prognostic factor with a hazard ratio of 3.63 (95% CI: [1.70, 7.77]). The hazard ratio of the gene risk score was higher than other clinical covariates, except cancer stage (III vs. I; with no significant difference). Similar results were obtained in the more comprehensive analysis with all the clinical covariates (Table 4.3). These results demonstrate that the 10-gene signature is a more accurate prognostic factor than most commonly used clinical factors.

Table 4.2. Multivariate Cox proportional hazard analysis of the 10-gene risk score and major clinical covariates including gender, age, and tumor stage on the combined testing cohorts (MSK and DFCI).

Variable*	***P*-value**	**Hazard Ratio (95% CI)** $^{\Psi}$	
Analysis without 10-gene risk score			
Gender (Male)	0.22	1.34	(0.84,2.16)
Age at diagnosis (>60)	0.08	1.61	(0.95,2.74)
Cancer Stage			
Stage II	6.25E-05	2.91	(1.72,4.91)
Stage III	1.09E-05	4.16	(2.20,7.85)
Analysis with 10-gene risk score			
Gender (Male)	0.28	1.30	(0.81, 2.09)
Age at diagnosis (> 60)	0.09	1.59	(0.93, 2.70)
Cancer Stage			
Stage II	1.62E-04	2.74	(1.62, 4.63)
Stage III	4.58E-06	4.45	(2.35, 8.43)
10-gene risk score	**8.61E-04**	**3.63**	**(1.70, 7.77)**

* Gender was a binary variable (0 for female and 1 for male); age at diagnosis was a binary variable (0 for < 60 years old and 1 otherwise); cancer stage was a categorical variable with 3 categories (Stage I [as the reference group], Stage II, and Stage III).

$^{\Psi}$ denotes confidence interval.

Table 4.3. Multivariate Cox proportional analysis of all available clinical covariates and 10-gene risk score in the combined test cohorts (MSK and DFCI).

Variable*	*P*-value	Hazard Ratio (95% CI)$^{\Psi}$
Analysis without 10-gene risk score		
Gender (Male)	0.43	1.22 (0.74,1.99)
Age at diagnosis (>60)	0.05	1.70 (0.99,2.92)
Race		
Others/Unknown	0.28	0.43 (0.09,1.97)
White	0.10	0.28 (0.06,1.28)
Smoking history		(0.00,0.00)
Smokers	0.62	0.84 (0.43,1.66)
Unknown	0.91	0.89 (0.11,7.10)
Tumor differentiation		
Moderately differentiated	0.14	0.53 (0.23,1.24)
Poorly differentiated	0.70	1.17 (0.53,2.61)
Cancer Stage		
Stage II	3.31E-04	2.72 (1.57,4.69)
Stage III	2.38E-05	4.93 (2.35,10.33)
Analysis with 10-gene risk score		
Gender (Male)	0.37	1.25 (0.76, 2.04)
Age at diagnosis (>60)	0.05	1.69 (0.99, 2.89)
Race		
Others/ Unknown	0.20	0.37 (0.08, 1.67)
White	0.10	0.28 (0.06, 1.25)
Smoking history		
Smokers	0.81	0.92 (0.47, 1.80)
Unknown	0.87	1.18 (0.15, 9.64)
Tumor differentiation		
Moderately differentiated	0.13	0.52 (0.23, 1.21)
Poorly differentiated	0.81	1.10 (0.50, 2.41)
Cancer Stage		
Stage II	4.19E-04	2.66 (1.54, 4.58)
Stage III	3.47E-05	4.79 (2.28, 10.05)
10-gene risk score	**3.31E-03**	**3.23 (1.48, 7.06)**

* Gender was a binary variable (0 for female and 1 for male); age at diagnosis was a binary variable (0 for < 60 years old and 1 otherwise); race was a categorical variable of 3 categories (African American [as the reference group], White, and Others [composed of Asian (5) , Hawaiian or Pacific Islander (1), and unknown]); tumor grade was categorical variable of 3 categories (Well [as the reference group], Moderately, and Poorly differentiated); Smoking history was a categorical variable of 3 categories (Non-smokers, Smokers, and Unknown); cancer stage was a categorical variable with 3 categories (Stage I [as the reference group], Stage II, and Stage III).

$^{\Psi}$ denotes confidence interval.

4.6 Functional Pathway Analysis

Having established the prognostic performance of the 10 prognostic genes identified, we sought to explore the functional involvement of this gene set in lung tumorigenesis and tumor progression. Curated molecular interactions between the major NSCLC signaling pathways and the identified 10-gene signature were retrieved using functional pathway analysis tools, Ingenuity Pathway Analysis (IPA, Ingenuity® Systems). The IPA functional pathway analysis demonstrated that nine canonical pathways were significantly (*P*<0.05; adjusted with BH tests) associated with the 10 prognostic genes. These pathways include methane metabolism and phenylalanine metabolism related to cell cycle, eicosanoid signaling that mediates inflammation and immunity, and MAPK signaling related to cell death, tissue morphology and inflammatory response (Fig. 4.5A). The pathway analysis also showed that cancer is among the top 5 most significant disease and disorders (*P*<0.05; adjusted with BH tests) in the network related to the 10 prognostic genes (Fig. 4.5B). Furthermore, 4 of the 10 prognostic genes were involved in interactions with major lung cancer signaling proteins, including *TP53*, *KRAS*, *EGF*, *E2F1*, and *RB1* as reported in the literature (Fig. 4.5C). These results suggest that the identified 10 genes are involved in lung cancer oncogenesis and tumor progression.

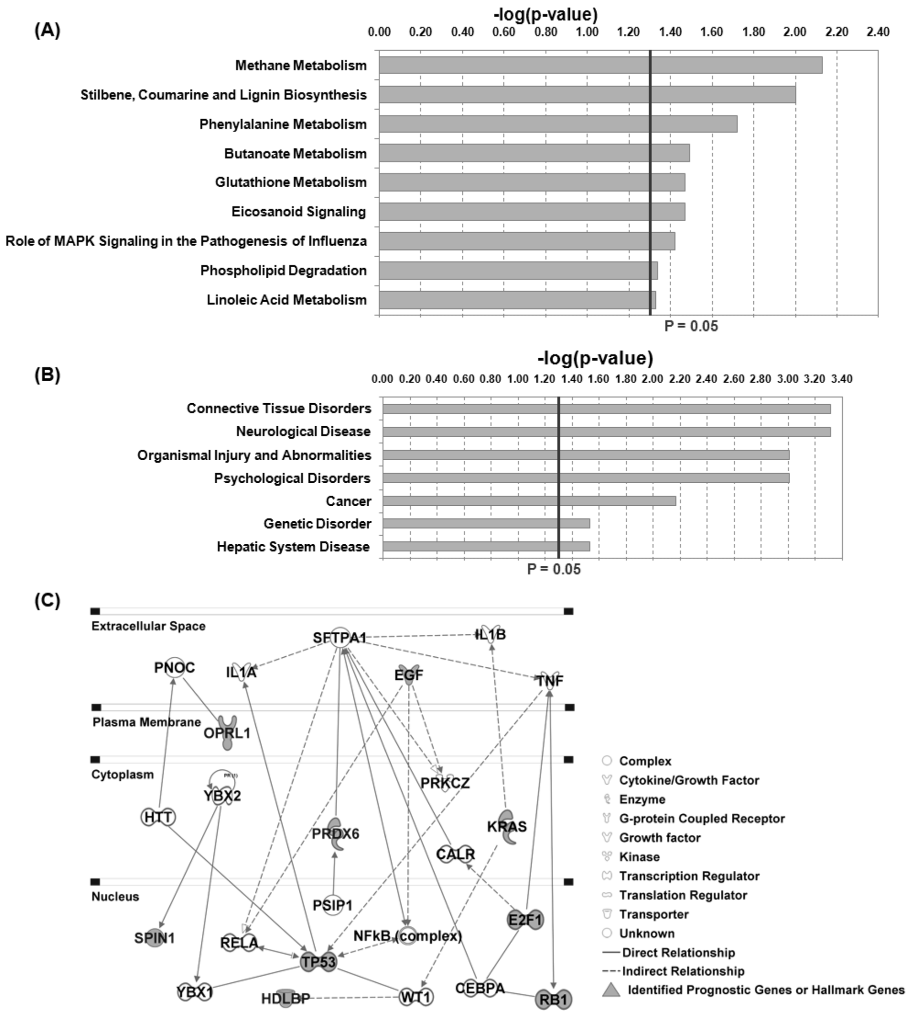


Figure 4.5. Functional pathway analysis of the 10 prognostic genes. Core analysis was performed with Ingenuity Pathway Analysis (IPA). Significant canonical pathways retrieved from IPA (A). Cancer was a significant biological function in the disease and disorders category (B). Curated interactions related to the 10 signature genes were also revealed from the literature (C).

4.7 Evaluation of Disease-mediated Gene Coexpression Networks

We further examined the disease-mediated coexpression networks derived from the system. The coexpression relations among the 10 signature genes and the 6 signaling proteins specific to each prognostic group were retrieved. Those commonly present in both training and test sets were considered robust for further study and biological evaluation. There were 4 common coexpression relations specific to good-prognosis group (Fig 4.6A) and 2 specific to poor-prognosis group (Fig 4.6B) in both training and test sets. These 6 coexpression relations represent the gene coexpression patterns specifically associated with metastasis in lung cancer patients. Among these 6 coexpression relations, the interaction between *CPEB1* and *TP53* was confirmed in a reported study [117] (Fig. 4.6A). Based on the five gene collections from MSigDB[9], the disease-mediated coexpression networks were also assessed in term of precision and false discovery rate (*FDR*). In 1,000 permutations, the precision of disease-mediated coexpression networks is 1 (P <0.001) and the *FDR* is 0. These results indicate that implication networks can reveal biologically relevant gene associations. Moreover, results from the stability test showed that more than 60% of the coexpression relations confirmed in the test set could be derived by using as few as 70% of the training samples, indicating the implication network algorithm is stable (Fig. 4.6D).

In addition to the 10-gene signatures, biological robustness of the other 20 identified signatures was also evaluated with the known molecular relations found in MSigDB. For each gene signature, the coexpression relations among the signature genes and their co-regulated signaling proteins were generated for each prognosis groups and those commonly found in training and two independent test cohorts were retrieved for the assessment (Fig. C.1 – C.22). Results show that two signatures (S2, S7) generated coexpression networks with the most coexpression relations derived incorrectly with *FDR* of 0.1 (-log(*FDR*) = 1, Fig. 4.7)). One the other hand, the disease-mediated coexpression networks for seven of the 21 signatures (including the 10-gene signature) has *FDR* < 0.001 (-log(*FDR*) ~ 3; Fig. 4.7). These results demonstrate that the coexpression relations derived from the implication relation induction algorithm are successfully validated with molecular interactions reported in the literature.

[9] http://www.broadinstitute.org/gsea/msigdb/collections.jsp

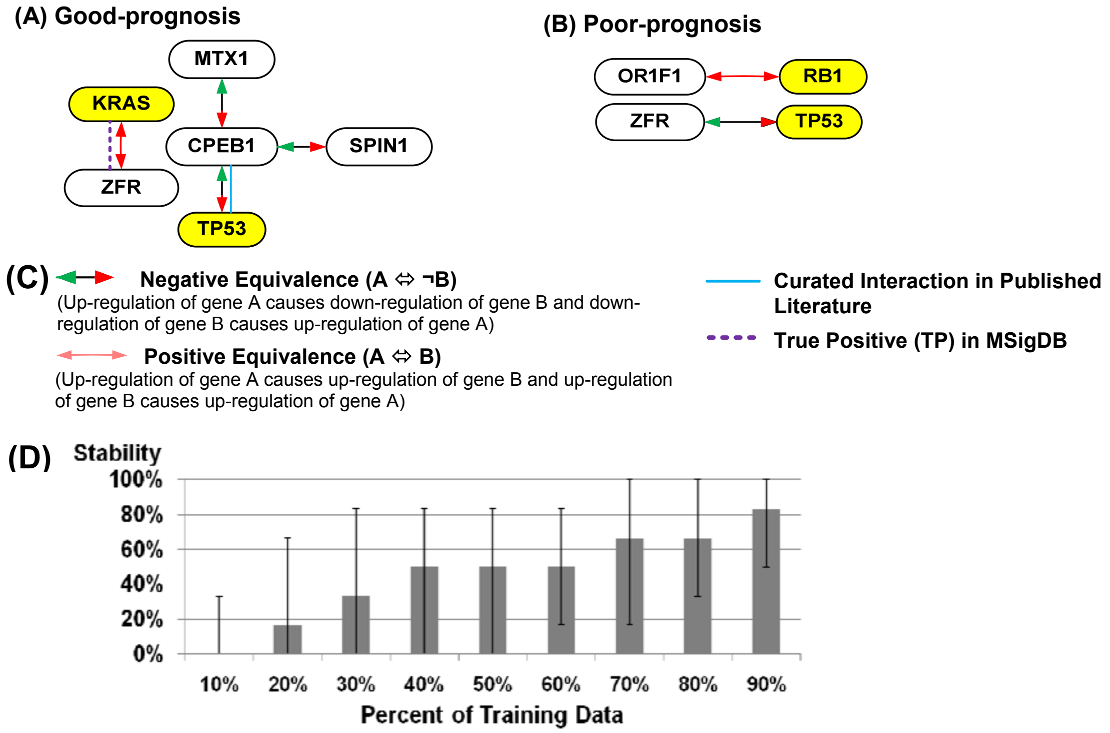

(A) Good-prognosis
MTX1
KRAS
CPEB1
SPIN1
ZFR
TP53
(B) Poor-prognosis
OR1F1
RB1
ZFR
TP53
(C)
Negative Equivalence (A ⇔ ¬B)
(Up-regulation of gene A causes down-regulation of gene B and down-regulation of gene B causes up-regulation of gene A)
Positive Equivalence (A ⇔ B)
(Up-regulation of gene A causes up-regulation of gene B and up-regulation of gene B causes up-regulation of gene A)
Curated Interaction in Published Literature
True Positive (TP) in MSigDB
(D)
Stability
100%
80%
60%
40%
20%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
Percent of Training Data

Figure 4.6. Disease-specific coexpression relations among the 10 prognostic signature genes and the 6 lung cancer signaling proteins. The disease-specific expression patterns for the good-prognosis group (A) and the poor-prognosis group (B) that were commonly present in both training and test cohorts were illustrated. The interpretation of the coexpression patterns is provided in (C). The stability of the networks in (A) and (B) was evaluated by using random subsets of the training samples in 100 iterations (D).

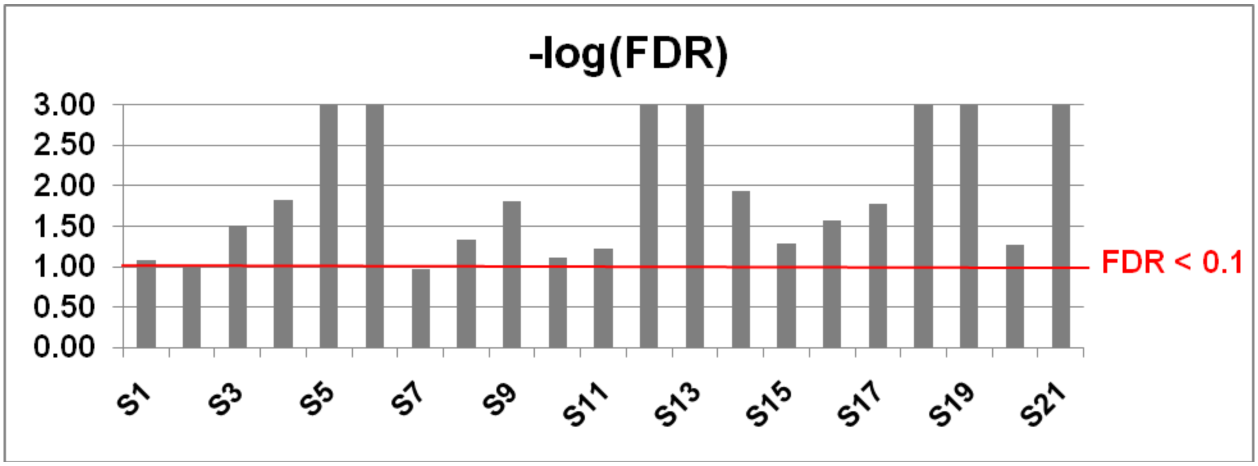


Figure 4.7. False discovery rate of the disease-mediated coexpression networks for the identified 21 prognostic signatures. The false discovery rate of the disease-specific coexpression relations among the signature genes and co-regulated hallmarks found in all three studied cohorts validated with MSigDB in 1,000 permutations.

4.8 Conclusions

This study presents a novel network-based methodology for modeling gene coexpression networks with major NSCLC signaling hallmarks for biomarker identification. The network model is flexible; it could be used alone, or in conjunction with other gene selection algorithms, such as random forests or *Relief*, in signature identification. This study demonstrates that the implication network methodology based on prediction logic is suitable for constructing genome-wide coexpression networks for analyzing perturbed gene/protein expression patterns in different disease states. The disease-mediated differential network components may contain important information for the discovery of biomarkers and pathways with implications for prognostic prediction. The implication network methodology provides a convenient and more predictive structure of gene regulation than the networks constructed based on correlation coefficients.

Our previous study identified a 12-gene signature using hybrid models combining *t*-test, significant analysis of microarray (SAM), and *Relief* algorithm [118]. The hazard ratio of the 12-gene signature was significant for all cancer stages in three patient cohorts, but not significant in any test sets for stage I only in the Director's Challenge Study. The network-based methodology presented in this chapter demonstrates the extensive identification of lung cancer prognostic gene signatures with strong prognostication performance in all tumor stages, stage I only, and stage I patients without receiving chemotherapy. All 21 gene signatures identified in this study outperformed other lung cancer signatures reported in the literature on the same patient cohorts. Most importantly, the identified signatures were all in feasible size to be further validated with biology experiments. These results indicate that modeling disease-mediated coexpression networks and crosstalk with NSCLC signaling hallmarks is crucial to identifying clinically important biomarkers for lung cancer. The identified gene signatures could potentially be used to advise patient selection for adjuvant chemotherapy in personalized lung cancer treatment.

The discovered gene signatures may also reveal essential molecular mechanisms of the disease and enhance our understanding of why patients with certain molecular tumor characteristics have a poor clinical outcome and how their outcome could be improved. Functional pathway studies with IPA confirmed the interactions between the major NSCLC signaling pathways and the identified gene signatures.

In addition to the 10-gene signature discussed in the chapter, a 14-gene and 13-gene prognostic signature was identified using the third approach of the implication network-based methodology [119, 120]. The 14-gene prognostic signature was identified from genes having direct coexpression relation with *TP53*, *KRAS, EGF, EGFR, E2F3,* and *E2F4* [120]; where the 13-gene prognostic signatures were directly co-regulated with *MET*, *EGF*, *KRAS*, *TP53*, *E2F2*, and *E2F4* in the disease-mediated differential network components [119]. Both these signatures generated significant patient stratification on the training set and two validation sets, with all tumor stage, and stage IB. However, they could not generate significant stratifications on both test cohorts of stage I patients and stage I patients without receiving chemotherapy. In patients with all tumor stages, the prognostic performance of the 14- and 13-gene signatures is comparable to the 10-gene signature presented in this chapter. (Table 4.4).

Table 4.2. Sensitivity and specificity of the 13- and 14-, and 10-gene gene prognostic models.

	Sensitivity (% of correctly predicted high-risk patients)				**Specificity (% of correctly predicted low-risk patients)**			
	n	**13-gene**	**14-gene**	**10-gene**	***n***	**13-gene**	**14-gene**	**10-gene**
3-year survival as the cutoff (high-risk: death within 3-y; low-risk: alive after 3-y)								
UM & HLM	95	56.84	52.63	58.95	152	75.66	67.76	73.03
MSK	23	73.91	78.26	65.22	71	54.93	45.07	63.68
DFCI	22	68.18	90.91	77.27	55	56.36	34.55	58.18
5-year survival as the cutoff (high-risk: death within 5-y; low-risk: alive after 5-y)								
UM & HLM	125	52.00	53.60	55.20	104	77.88	73.08	75.00
MSK	34	67.65	79.41	52.94	31	51.61	51.61	61.29
DFCI	28	67.86	89.29	75.00	36	61.11	30.56	58.33
2.5-year and 5-year survival as the high- and low-risk cutoffs (high-risk: death within 2.5-y; low-risk: alive after 5-y)								
UM & HLM	84	59.52	51.19	59.52	104	77.88	73.08	75.00
MSK	21	76.19	76.19	61.90	31	51.61	51.61	61.29
DFCI	20	70.00	90.00	75.00	36	61.11	30.56	58.33

These results conclude that the presented implication network-based methodology accurately model the disease relevant gene coexpression patterns for the discovery of clinically important prognostic gene signatures. Most importantly, gene signatures identified with this novel network-based methodology provide strong prognostic performance and in viable size for biology validation and clinical application.

Chapter 5

Network-based Identification of Smoking-associated Gene Signature for Lung Cancer

Studies have demonstrated that smoking contributes to about 90% of all lung cancer cases and it appears to be a strong risk factor in the development of lung cancer [108, 121, 122]. However, smoking is not an established determinant in lung cancer prognosis as its effect in lung cancer progression remains unclear. In this study, we sought to identify a smoking-associated gene signature with implications in lung cancer diagnosis and prognosis using genome-wide transcriptional profiles from lung cancer patients.

In the previous chapter, implication networks were employed to model disease-mediated genome-wide coexpression networks for the identification of prognostic gene signatures. In this study, implication networks were used to infer the relevance to signaling pathways in a set of selected genes associated with smoking and lung cancer survival.

This chapter is organized into ten sections. The first section presents the methodology. Section 5.2 describes the identification of the smoking-associated signature using the proposed methodology. Prognostic evaluation of the identified signature will be presented in Section 5.3. Section 5.4 provides the results on association study between the signature and smoking. The prognostic evaluation of the signature with clinical covariates is presented in Section 5.5. Section 5.6 validates the prognostic performance of the signature on different subtypes of NSCLC. Results on potential usage of the signature for early detection of lung cancer will be presented in Section 5.7. The assessment of interactions retrieved using implication networks will be discussed in Section 5.8. Biology eexperiment validation result is presented in Section 5.9. The conclusions of the study will be discussed in Section 5.10.

5.1 Methodology

The methodology studied in this chapter is similar to the implication network-based system illustrated in Chapter 4. Major difference between the two methodologies is that the gene coexpression networks were not modeled for the whole genome in this study. Instead of the whole genome, the implication networks were used to model coexpression patterns of a smaller pool of genes: genes associated with smoking and also prognostic for lung cancer. This application also demonstrates the use of implication networks in modeling gene coexpression patterns mediated with the smoking practice, instead of disease outcome as studied in Chapter 4.

Specifically, the methodology contains the following steps: 1) identifying genes significantly associated with lung cancer survival, 2) from the survival genes, selecting genes which are differentially expressed in smoker versus non-smoker groups, 3) from these candidate genes, constructing gene co-expression networks based on prediction logic for smokers and non-smokers, 4) identifying smoking-mediated differential components, i.e., the unique gene co-expression patterns specific to smoker group or non-smoker group, and 5) from the differential components, identifying genes directly co-expressed with major lung cancer hallmarks as the smoking-associated gene signature for lung cancer (Fig. 5.1).

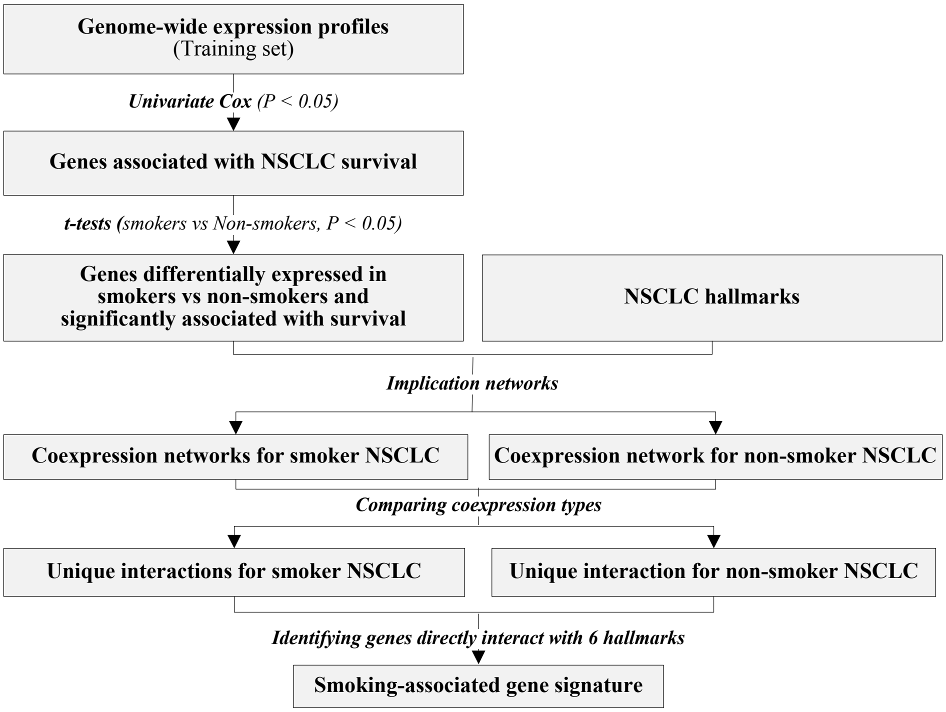

Genome-wide expression profiles
(Training set)
Univariate Cox (P < 0.05)
Genes associated with NSCLC survival
t-tests (smokers vs Non-smokers, P < 0.05)
Genes differentially expressed in smokers vs non-smokers and significantly associated with survival
NSCLC hallmarks
Implication networks
Coexpression networks for smoker NSCLC
Coexpression network for non-smoker NSCLC
Comparing coexpression types
Unique interactions for smoker NSCLC
Unique interaction for non-smoker NSCLC
Identifying genes directly interact with 6 hallmarks
Smoking-associated gene signature

Figure 5.1. Methodology for network-based identification of smoking-associated signatures.

5.2 Identification of a Smoking-associated 7-gene Signature

In this study, 442 lung adenocarcinoma patient samples obtained from the Director's Challenge Study [2] were used. In this study, the UM and HLM cohorts from the Director's Challenge Study [2] formed the training set (*n*=256), whereas MSK and DFCI cohorts formed the test set (*n*=186). Before the analysis, genes with missing values in at least half of the samples were removed, which left 19,866 genes for the analysis.

Survival genes were first selected from the whole genome. A total of 2,310 genes were significantly associated with overall survival ($P < 0.05$, univariate Cox modeling) in the training data. Next, from the set of 2,310 survival genes, 217 genes showed significant differential expression ($P < 0.05$, *t*-tests) in smokers versus non-smokers in the training data were further extracted. These 217 survival and smoking-associated genes as well as six major signaling proteins, including *EGF*, *EGFR*, *MET*, *KRAS*, *E2F3*, and *E2F5,* were included in the network analysis. Although these six hallmarks were not significantly associated with survival nor differentially expressed in smokers, they were major signaling proteins included in human non-small cell lung cancer disease mechanisms delineated by the KEGG Pathway Database[10].

To construct implication networks, expression profiles in each patient were partitioned into binary values using the mean expression profile of each gene as the cutoff. If the expression of a gene in a patient sample was greater than the mean in the cohort, this gene was denoted as *up-regulated* in this tumor sample; otherwise, it was denoted as *down-regulated* in the tumor sample. Patient samples in the training set were separated into two groups: smokers (patients who smoked in the past or who are currently smoking) and non-smokers (patients who never smoked). For each patient group, coexpression network among the 217 genes and six signally hallmarks was constructed using the implication induction algorithm. Between each pair of the 223 genes, possible significant ($P < 0.05$; z-tests) coexpression relations (interactions) were derived in the smoker group and the non-smoker group separately, constituting smoking-mediated gene co-expression networks for lung cancer. By comparing the implication rules between each pair of nodes in the two smoking-mediated networks, differential network components were identified. These differential components are interactions that were present in the smoker group but missing in the non-smoker group, or conversely, those present in the non-smoker group but absent in the smoker group.

From the differential components associated with smoker group and non-smoker group, genes having direct interactions with the six lung cancer hallmarks were identified. As a result, six genes were identified from the smoker group and one gene was identified from the non-smoker group. This constituted the smoking-associated 7-gene signature for lung cancer prognosis (Table 5.1). Fig. 5.2 gives an overview of the whole methodology.

[10] http://www.genome.jp/kegg/pathway/hsa/hsa05223.html

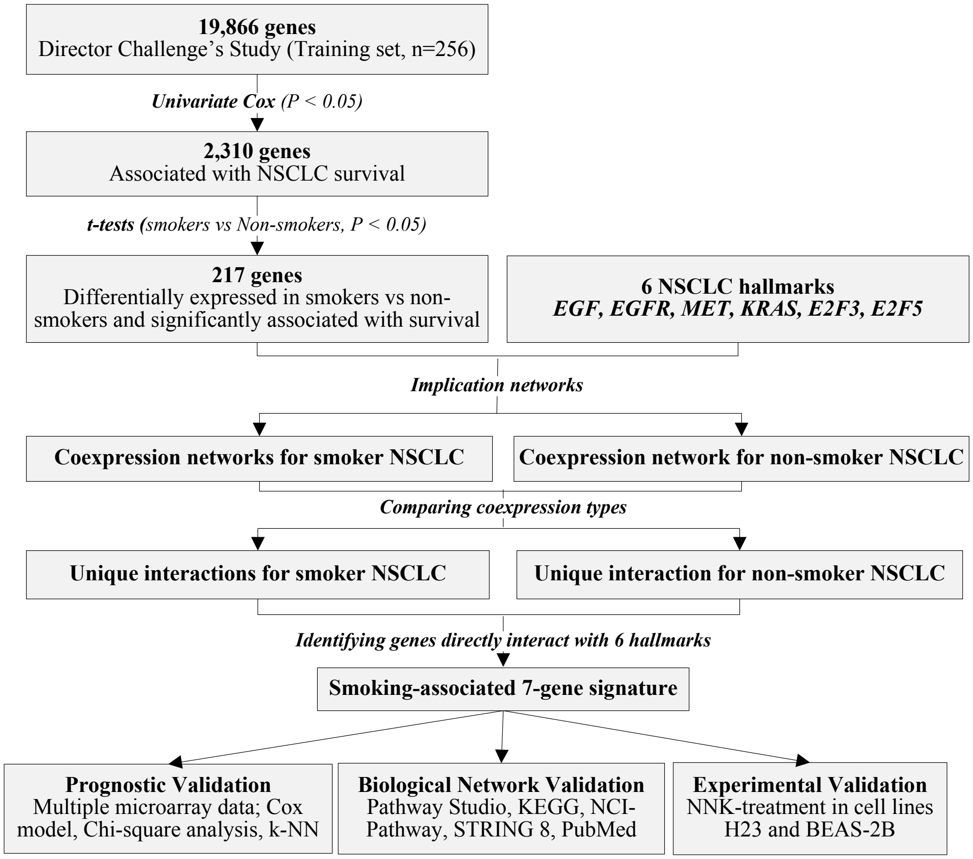

19,866 genes
Director Challenge's Study (Training set, n=256)
Univariate Cox (P < 0.05)
2,310 genes
Associated with NSCLC survival
t-tests (smokers vs Non-smokers, P < 0.05)
217 genes
Differentially expressed in smokers vs non-smokers and significantly associated with survival
6 NSCLC hallmarks
EGF, EGFR, MET, KRAS, E2F3, E2F5
Implication networks
Coexpression networks for smoker NSCLC
Coexpression network for non-smoker NSCLC
Comparing coexpression types
Unique interactions for smoker NSCLC
Unique interaction for non-smoker NSCLC
Identifying genes directly interact with 6 hallmarks
Smoking-associated 7-gene signature
Prognostic Validation
Multiple microarray data; Cox model, Chi-square analysis, k-NN
Biological Network Validation
Pathway Studio, KEGG, NCI-Pathway, STRING 8, PubMed
Experimental Validation
NNK-treatment in cell lines H23 and BEAS-2B

Figure 5.2. Identification of 7-gene smoking-associated signature.

Table 5.1. The identified 7-gene smoking associated signature.

Gene Symbol	Gene Title	Molecular Function (Gene Ontology)
ABCA3	ATP-binding cassette, sub-family A (ABC1), member 3	ATP, nucleotide binding; ATPase, transporter activity
CRTAC1	Cartilage acidic protein 1	Calcium ion binding
CYP3A4	Cytochrome P450, family 3, subfamily A, polypeptide 4	Monooxygenase, electron carrier, oxidoreductase activity; heme, metal ion, and steroid binding
GPRC5C	G protein-coupled receptor, family C, group 5, member C	Receptor activity; protein binding
LTF	Lactotransferrin	Ferric iron, heparin, metal ion, protein binding; peptidase, serine-type endopeptidase activity
PIGN	Phosphatidylinositol glycan anchor biosynthesis, class N	Phosphotransferase, transferase activity
SEMA3C	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	Receptor activity; semaphorin receptor binding

5.3 Prognostic Evaluation of the Signature

We sought to study if the gene signature identified could provide accurate prognostic prediction of survival for lung cancer patients. The six hallmarks were not fitted in the model as they were not significantly associated with survival. On the training cohort, the original continuous expression profiles of the seven probes were fitted into a Cox proportional hazard model as covariates. A survival risk score was generated for each patient in the training set. To identify the best patient stratification scheme, various cutoff values of the risk scores from the training set were evaluated. The cutoff value that gave the shortest distance to the point of perfect prediction, i.e. point [0,1] of the 3-year ROC curve (Fig. 5.3A), produced the best patient stratification in the training set (Fig. 5.3B). Therefore, the training model and cutoff value were applied to the test set (Fig. 5.3C). In both training and test set, this classification scheme generated significant patient stratifications (log-rank $P < 0.007$, Kaplan-Meier analysis).

To evaluate the statistical significance of the signature identified from the proposed network analysis, a set of seven genes from the 217 survival and smoking-associated genes were randomly selected and constructed as a classifier using the same approach with the Cox proportional hazard model. Results showed that the signature identified gave significantly ($P <$ 0.04) better lung cancer prognosis compared with 1000 random signatures.

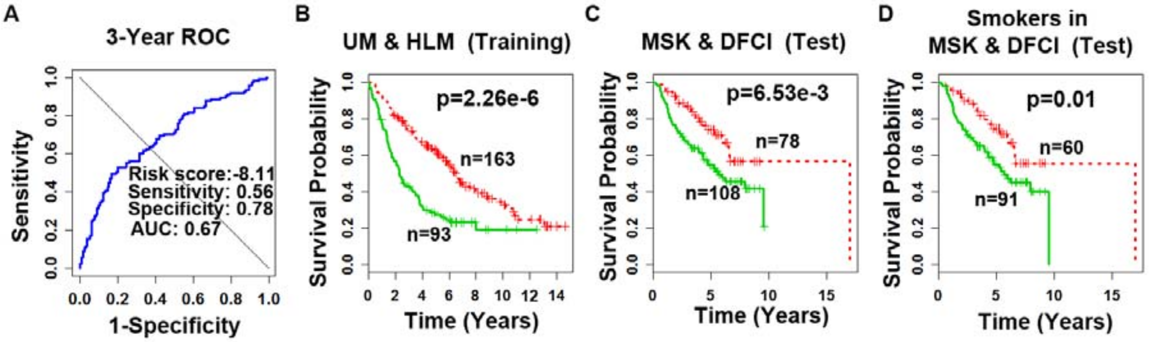


Figure 5.3. Prognostic prediction of patients survival by smoking-associated gene signature. On the cohorts from Shedden et al. [2], the risk score giving the best prediction on the 3-year ROC curve was identified as the cutoff for patient stratification (A). This cutoff value generated significant patient stratification on the training set (B), test set (C), and smokers of test set (D) in Kaplan-Meier analyses. Log-rank tests were used to assess the statistical significance in survival probability between the two prognostic groups.

5.4 Smoking Association and Smoking Cessation

To evaluate the smoking association of the identified gene signature, we evaluated the performance of the prognostic signature on smokers in the studied cohorts. Results showed that the signature gave accurate prognostic prediction in smokers in the test cohort (log-rank $P < 0.01$, Kaplan-Meier analysis) (Fig. 5.3D) but not in non-smokers (log-rank $P < 0.12$, Kaplan-Meier analysis, results not shown). In addition, gene expression-defined high and low-risk groups showed significant association with smoking ($P < 0.02$, Chi-square tests) and smoking cessation ($P < 0.00001$, Chi-square tests) (Table 5.2). Specifically, smokers were significantly associated with high-risk group compared with non-smokers, and current smokers showed a stronger association with the high-risk group compared with former smokers.

Table 5.2. Associations between smoking status and the classifier's prediction.

	Low-risk	High-risk	Chi-square Test
Smoker	143	157	**Smoking association**
Non-smoker	33	16	$\chi^2 = 5.76$ ($P = 0.02$)
Current Smoker	3	29	**Smoking cessation**
Former Smoker	140	128	$\chi^2 = 19.37$ ($P = 1.08\text{e-}5$)

5.5 Prognostic Evaluation with Clinical Covariates

To validate the prognostic power of the identified 7-gene signature, the constructed expression-defined prognostic model was evaluated with common lung cancer prognostic factors, including gender, age, tumor stage, and tumor differentiation on smokers in the test cohort. The predicted 7-gene risk score was used as the covariate in the multivariate Cox analysis.

Results from the multivariate Cox proportional analysis showed that tumor stage was the only factor significantly ($P < 0.002$) associated with elevated risk of lung cancer death when the model was fitted without the 6-gene prognostic prediction (Table 3). When the 7-gene risk score was added to the multivariate Cox model, the 7-gene risk score demonstrated a significantly strong association with the risk of lung cancer death (hazard ratio = 1.89, 95% CI: [1.06, 3.38]), and tumor stage remained significant (Table 3). The hazard ratio of the 7-gene risk score was higher than other cancer prognostic factors except tumor stage, while there is no significant difference between the hazard ratio of the 7-gene risk score and tumor stage (II vs. I). The results demonstrate that the 7-gene risk score could provide more accurate prognosis than some commonly used clinical parameters.

Table 3. Multivariate Cox proportional analysis of the 7-gene risk score and major clinical covariates in smoking lung cancer patients of the test cohort.

Variable*	***P*-value**	**Hazard Ratio (95% CI)**$^{\Psi}$	
Analysis without 7-gene risk score			
Gender (Male)	0.55	1.17	(0.70, 1.95)
Age at diagnosis (>60)	0.35	1.31	(0.74, 2.29)
Tumor differentiation			
Moderately differentiated	0.30	0.63	(0.26, 1.51)
Poorly differentiated	0.89	1.06	(0.47, 2.38)
Cancer Stage			
Stage II	1.54E-03	2.60	(1.44, 4.71)
Stage III	5.53E-05	4.48	(2.16, 9.29)
Analysis with7-gene risk score			
Gender (Male)	0.51	1.19	(0.71, 1.99)
Age at diagnosis (>60)	0.49	1.22	(0.69, 2.16)
Tumor differentiation			
Moderately differentiated	0.33	0.65	(0.27, 1.55)
Poorly differentiated	0.93	0.96	(0.43, 2.16)
Cancer Stage			
Stage II	1.64E-03	2.61	(1.44, 4.74)
Stage III	3.29E-05	4.79	(2.29, 10.04)
7-gene risk score	**0.03**	**1.89**	**(1.06, 3.38)**

* Gender was a binary variable (0 for female and 1 for male); age at diagnosis was a binary variable (0 for < 60 years old and 1 otherwise); tumor grade was categorical variable of 3 categories (Well [as the reference group], Moderately, and Poorly differentiated); tumor stage was categorical variable of 3 categories (Stage I [as the reference group], Stage II, and Stage III).

$^{\Psi}$ denotes confidence interval.

5.6 Prognostic Validation on other Histology Subtypes of NSCLC

The prognostic performance of the 7-gene signature was further evaluated on Raponi [17] and Bild [13] cohorts including squamous cell carcinoma. Due to small sample size, patient samples in the studied cohort were randomly partitioned into separate training and test sets. Then, a prognostic classifier was constructed on training set using the Cox proportional hazard model and validated on the test set without re-estimation of parameters. On Raponi's cohort with squamous cell carcinoma patients, the 7-gene signature stratified patients into two distinct survival groups (log-rank $P < 0.005$, Kaplan-Meier analysis) in the training set but border line in the test set (Fig. 5.4A, 5.4B). The border line performance in the test cohort could be due to the reason that 8 percent of the patients in the cohort were non-smokers and the 7-gene prognostic signature is specific to smokers. On the Bild's cohort with lung adenocarcinoma or squamous cell carcinoma patients, the 7-gene signature stratified patients into two distinct survival groups in both training and test set (log-rank $P < 0.04$, Kaplan-Meier analysis) (Fig. 5.4C, 5.4D).

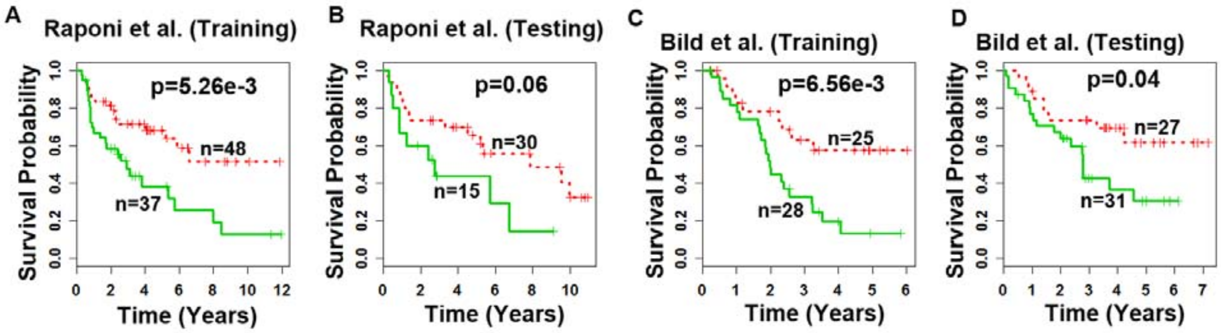


Figure 5.4. Prognostic prediction of patient survival by the smoking-associated gene signature on two cohorts with different histology. In Kaplan-Meier analyses, significant patient stratifications were also obtained in the training and test sets on cohorts from Raponi et al. [17] (A, B) and Bild et al. [13] (C, D). Log-rank tests were used to assess the statistical significance in survival probability between the two prognostic groups.

5.7 Early Detection of Lung Cancer

We further evaluated whether the 7-gene signature could be used for the diagnosis of lung cancer in smokers. The smoking cohort from Spira et al. [108] was separated into a training set (*n*=77) and two independent test sets (*n*=52 and *n*=35). With the nearest neighbor algorithm implemented in WEKA [116], the classifier could accurately identify lung cancer patients from normal patients with overall accuracy of 65% in training and 73% or higher in test sets (Table 5.3). The sensitivity in identifying lung cancer patients is at least 72% (Table 5.3). The odds ratio of predicted lung cancer risk was highly significant in all three sets (*OR* = 3.85, 95% CI: [1.45, 10.20], *P* < 0.007 in training; *OR* = 7.35, 95% CI: [2.16, 25.04], *P* < 0.001 in Test set 1; *OR* = 8.45, 95% CI: [1.84, 38.75], *P* < 0.006 in Test set 2; Table 5.3). Furthermore, the classifier's performance was significantly (*P* < 0.002) better than that of random signatures with the same size using the same classifier in 1000 tests, on the same training and test sets.

Table 5.3. Prediction of lung cancer risk in smokers.

	Sensitivity (lung cancer)	**Specificity (normal)**	**Overall Accuracy**	**Odds Ratio [95% CI]**	***P*-value**
Training (10-fold CV)	74% (26/35)	57% (24/42)	65% (50/77)	3.85 [1.45, 10.20]	0.007
Test 1	72% (18/25)	74% (20/27)	73% (38/52)	7.35 [2.16, 25.04]	0.001
Test 2	72% (13/18)	76% (13/17)	74% (26/35)	8.45 [1.84, 38.75]	0.006

5.8 Assessment of Smoking-mediated Gene Coexpression Networks

To assess the smoking-mediated coexpression relations derived by the implication network, differential network components among the signature genes and the six signaling hallmarks present in both training and test sets were retrieved as they were consider robust for further evaluation. There were 17 common interactions specifically associated with smokers (Fig. 5.5A) and one interaction specifically associated with non-smokers (Fig. 5.5B).

The biological relevance of the derived coexpression relations was validated by retrieving curated interactions related to these genes using bioinformatics tools including Pathway Studio (Fig. 5.5D) and other curated signal pathway databases. Among 18 coexpression relations derived from the implication networks, 11 interactions specific to smokers were confirmed (Fig. 5.5A and 5.5B). The *FDR* of the smoking-mediated coexpression networks derived is 0.01. These results indicate that implication networks can reveal biologically relevant gene associations.

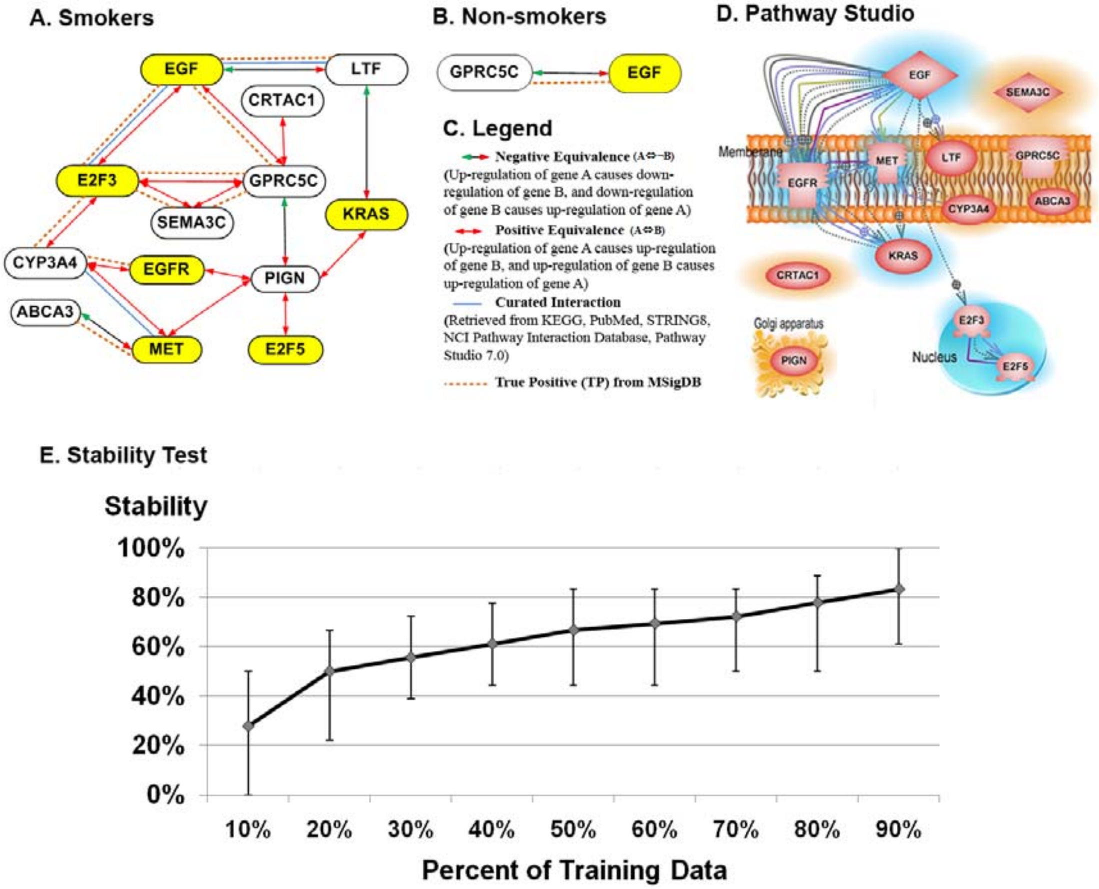

A. Smokers
EGF
LTF
CRTAC1
E2F3
GPRC5C
SEMA3C
KRAS
CYP3A4
EGFR
PIGN
ABCA3
MET
E2F5
B. Non-smokers
GPRC5C
EGF
C. Legend
Negative Equivalence (A⇔¬B)
(Up-regulation of gene A causes down-regulation of gene B, and down-regulation of gene B causes up-regulation of gene A)
Positive Equivalence (A⇔B)
(Up-regulation of gene A causes up-regulation of gene B, and up-regulation of gene B causes up-regulation of gene A)
Curated Interaction
(Retrieved from KEGG, PubMed, STRING8, NCI Pathway Interaction Database, Pathway Studio 7.0)
True Positive (TP) from MSigDB
D. Pathway Studio
EGF
SEMA3C
Membrane
MET
LTF
GPRC5C
EGFR
CYP3A4
ABCA3
KRAS
CRTAC1
Golgi apparatus
PIGN
E2F3
Nucleus
E2F5
E. Stability Test
Stability
100%
80%
60%
40%
20%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
Percent of Training Data

Figure 5.5. Smoking-mediated coexpression relations among the signature genes and lung cancer hallmarks. Gene coexpression patterns specific to smokers (A) and non-smokers (B) derived by the implication network algorithm ($P < 0.05$) in both training and test sets ($FDR = 0.01$). The biological interpretation of the implication relations are described in (C). Interactions reported in literature retrieved from Pathway Studio (D). The stability of smoking-mediated networks as evaluated with random subsets of patients from the training cohort in 100 iterations (E).

Results from the stability test of the smoking-mediated coexpression networks (Fig. 5.5A and 5.5B) show that the implication network algorithm is stable as most of the coexpression relations (about 70%) could be derived using as few as half of the training samples (Fig. 5.5E)

5.9 Experiment Validation

We further confirm the biological aspect of smoking-mediated gene coexpression relations derived from the implication networks and the perturbation of signaling pathway mechanisms in smokers among the identified signature genes and hallmark genes using the expression quantified from cell lines. H23 and BEAS-2B cells were exposed to NNK for 15 minutes, one hour, and 16 hours. Then, qRT-PCR low-density arrays were used to analyze the gene expressions in the NNK-treated cells. On the qRT-PCR data normalized with *POLR2A*, gene expression fold changes of the genes in treated cell lines versus control were computed.

Based on the fold changes computed for the signature genes and the hallmarks observed in the *NNK*-treated H23 cell lines, coexpression relations among the 7 signature genes and the six hallmarks were derived. These represented the observed perturbations among the signature genes and signaling pathway mechanism specific for smokers. Comparing the observed perturbations with coexpression relations unique for smokers (Fig. 5.5A), results showed that the coexpression relations derived with the implication networks in smokers were confirmed by the coexpression relations observed in NNK-treated H23, at different time points (Fig. 7).

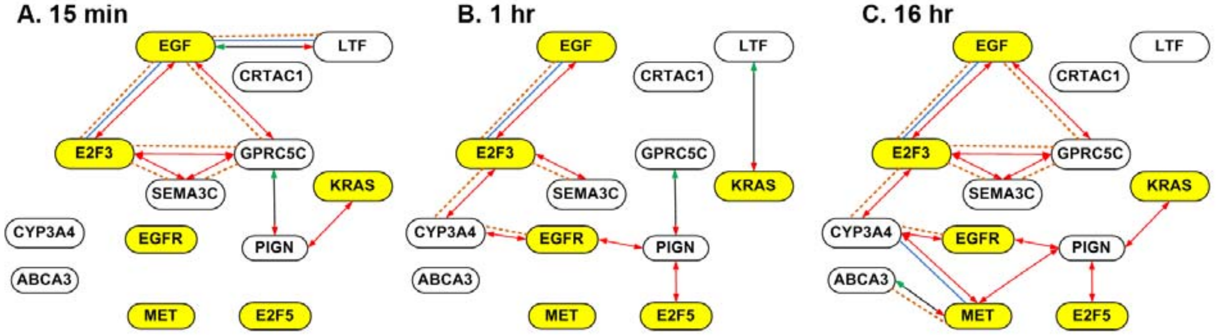


Figure 5.6. Coexpression relations observed in the NNK-treated H23 cell lines for 15 minutes (A), 1 hours (B), and 16 hours (C).

5.10 Conclusions

This study examined the implication network-based model discussed in Chapter 4 in a different scenario. Instead of modeling disease-mediated coexpression networks at the genome-wide scale, smoking-mediated coexpression networks were constructed on a subset of genes associated with smoking and lung cancer survival. From the smoking-mediated coexpression networks derived, a smoking-associated 7-gene prognostic signature that co-regulated with major lung cancer signaling proteins were identified. The identified 7-gene signature showed strong implications in providing accurate estimate for lung cancer survival and risk of diagnose with lung cancer in smokers. The 7-gene defined prognostication also showed strong association with smoking and smoking cessation. Furthermore, the 7-gene prognostic model also appeared to be a more accurate prognostic factor than commonly used clinical factors for lung cancer.

Using the same methodology, a 6-gene and an 8-gene smoking-associated prognostic signature were also identified from having direct coexpression relations with different major lung cancer signaling hallmarks (Table 5.4). The prognostic performance of the 6-gene and 8-gene signatures was comparable with the 7-gene signature (Fig. 5.7).

Table 5.4. 6-gene and 8-gene smoking-associated signatures identified using the implication network-based methodology.

Signature	Hallmarks	Identified Signature Genes
6-gene signature	MET, EGF, KRAS, TP53, E2F1, E2F4	HERC3, NUPR1, SEMA3C, EEF1B2, SOSTDC1, TFAP2A
8-gene signature	MET, EGF, EGFR, KRAS, E2F2, E2F5	LUC7L3, CRTAC1, CYP3A4, GPRC5C, HOMER1, PIGN, SEMA3C, EEF1B2

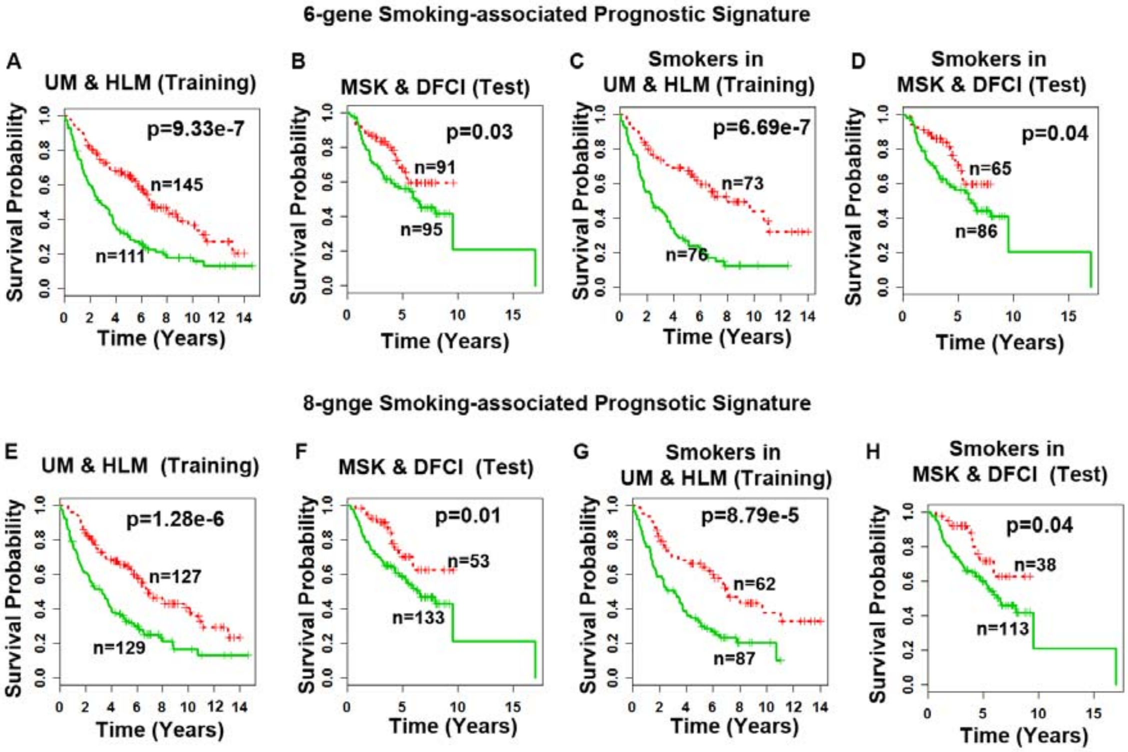


Figure 5.7. Prognostic performance of the 6-gene and 8-gene signature identified with the network-based methodology on patient cohorts from the Director's Challenge Study [2]. Using multivariate Cox proportional hazard model fitted with the 6 genes, the risk score giving the best predicted on the 3-year ROC curve (value of -15.36) generated significant patient stratification in both the training and test cohort (A,B) and smokers in training and test cohorts (C,D). Similarly, on the 8-gene fitted Cox model, the mean of risk scores from training samples (value of -5.31) generated patients into two risk groups with significantly distinct survival outcome in the training and test cohorts with all patients (E,F) and smoker patients only (G,H).

In comparison with the clinical covariates, the 6-gene and 7-gne prognostic power was comparable to one another (Table 5.5) but the 8-gene signature did not give better prognostic power than clinical factors.

Table 5.5. Multivariate Cox analyses of the 6-gene expression-defined prognostication and major clinical covariates in smoking lung cancer patients in the test cohort.

Variable*	*P*-value	Hazard Ratio (95% CI)$^{\Psi}$	
Analysis without 6-gene prognostic prediction			
Gender (Male)	0.55	1.17	(0.70, 1.95)
Age at diagnosis (>60)	0.35	1.31	(0.74, 2.29)
Tumor differentiation			
Moderately differentiated	0.30	0.63	(0.26, 1.51)
Poorly differentiated	0.89	1.06	(0.47, 2.38)
Cancer Stage			
Stage II	1.54E-03	2.60	(1.44, 4.71)
Stage III	5.53E-05	4.48	(2.16, 9.29)
Analysis with 6-gene prognostic prediction			
Gender (Male)	0.42	1.24	(0.74, 2.08)
Age at diagnosis (>60)	0.52	1.20	(0.68, 2.13)
Tumor differentiation			
Moderately differentiated	0.39	0.68	(0.28, 1.64)
Poorly differentiated	0.89	0.94	(0.42, 2.15)
Cancer Stage			
Stage II	7.30E-04	2.83	(1.55, 5.19)
Stage III	1.51E-05	5.36	(2.50, 11.46)
6-gene prognostic prediction	**0.04**	**1.89**	**(1.04, 3.43)**

* Gender was a binary variable (0 for female and 1 for male); age at diagnosis was a binary variable (0 for < 60 years old and 1 otherwise); tumor grade was categorical variable of 3 categories (Well [as the reference group], Moderately, and Poorly differentiated); cancer stage was categorical variable of 3 categories (Stage I [as the reference group], Stage II, and Stage III).

$^{\Psi}$ denotes confidence interval.

Results from this study showed that the 7-gene smoking-associated signature is highly potential to be used to develop clinical gene test to screen smokers for risk of developing lung cancer and provide a precise prognostic test for smoking lung cancer patients. This would be beneficial to a large population of the lung cancer patients.

The application of the implication network-based methodology in this study again demonstrated that the methodology correctly modeled the biologically perturbed coexpression patterns. In this study, the coexpression patterns perturbed by smoking practices were correctly modeled and validated with coexpression relations observed from the experiments with NNK-treated cell lines. Moreover, this study once more demonstrated that the integration of biologically perturbed coexpression patterns with signaling pathway mechanisms lead to identification of strong prognostic gene signatures.

Chapter 6

Evaluation with Boolean Implication Networks and Bayesian Networks

Results presented in Chapter 4 and Chapter 5 demonstrated that the implication networks based on prediction logic could efficiently model the disease-mediated gene coexpression networks for signature genes identification. Furthermore, the coexpression patterns derived were successfully validated with molecular interactions reported in the literature. Another similar formalism of implication networks, i.e., Boolean implication networks were used in a meta-analysis to discover relationships among genes for different species [80]. In contrast to small patient cohorts we had studied, large sample of microarray data were used in their meta-analysis: 4,787 human, 2,154 mice, and 450 *Drosophila*. Moreover, the Boolean implication networks were used as the framework to study the genomic evolution, where our framework was used for gene selection. While implication networks could efficiently model the gene regulatory networks, it is not as commonly known as the Bayesian networks in this research domain.

In this chapter, we will discuss the performance and characteristics of the employed implication networks in comparison with the Boolean implication networks and Bayesian networks. The comparison of the two implication networks in term of the size of the networks will be presented in the first section, Section 6.1. Section 6.2 discusses the comparativeness of both implication networks after fine-tuning the implication networks based on prediction logic. Section 6.3 examines the biological robustness of networks derived from both implication networks. To compare the employed implication networks with the Bayesian networks, we evaluate the biological strength of the derived disease-mediated gene coexpression networks.

The results are presented in Section 6.4. The last section, Section 6.5 provides a brief conclusion on the comparison studies.

6.1 Comparison with Boolean Implication Networks

In the framework to derive Boolean implication networks, StepMiner algorithm [81] was first used to automatically assign a threshold (*t*) for each gene. Based on the assigned threshold *t*, the gene expression level was defined as up-regulated if the expression value is above *t* + 0.5; down-regulated if the expression value is below *t* - 0.5. If the expression value is between *t* – 0.5 and *t* + 0.5, the expression level is defined as intermediate and will be ignored during the implication relations derivation. The choice of the interval width (±0.5) is based on the standard deviations of genes over all arrays, and the 5th percentile from the bottom is selected. In the data used in their study, the standard deviation is a little less than 0.26. The interval is defined as two standard deviations from the threshold, thus *t* ± 0.5. We adopted similar approach in deciding the width for the interval width.

In our data, the standard deviations of all the genes (12,566 genes) ranged from 0.13 to 40.24. There were some outlier genes with very large standard deviation; this could be due to noise (Fig 6.1A). After removing the outliers, the distribution of the standard deviations is a little skewed toward the range of 0 to 0.5 (Fig 6.1B, 6.1C). The 5th percentile from the bottom (value 0) of the distribution is 0.23. Therefore, after obtaining the threshold (*t*) from StepMiner, we defined the gene as up-regulated if the expression value is *t* + 0.46, down-regulated if the expression value is *t* – 0.46, or intermediate if the expression value lies between *t* – 0.46 and *t* + 0.46.

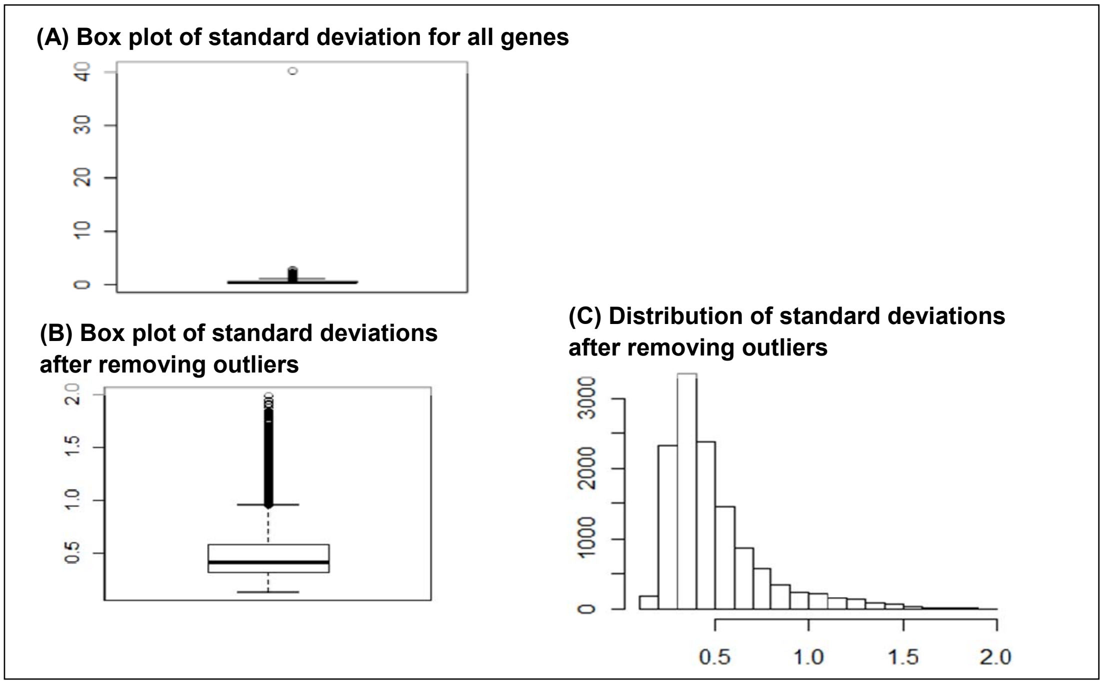


Figure 6.1. Distribution of standard deviations for all the genes (A) and after removing the outlier genes (B,C).

After defining the gene expression levels into up- or down-regulated by *t* + 0.46 and *t* - 46, implication relations between the gene pairs were derived using the Boolean implication networks proposed by Sahoo et al. [80] and the implication networks based on prediction logic (Fig. 2.7). Implication relations were derived for good-prognosis group (patients who survived 5 years or longer after surgery) and poor-prognosis group (patients who died within 5 years after surgery). By comparing the implication rule types in both prognosis group, implication relations specific to each prognosis group were also obtained.

Results show that Boolean implication networks derived less coexpression relations than those derived from implication networks based on prediction logic (Fig. 6.2), except for the implication relations specific to poor-prognosis group. After removing the samples with expression falls within the intermediate range (*t*±0.46), the average sample size used for deriving the implication relations is 23 (out of 125) for the poor prognosis group and 20 (out of 140) for the good prognosis group. Since the number of samples in deriving the successful implication relation between the pair of genes was small, these results are not reasonable to be further investigated.

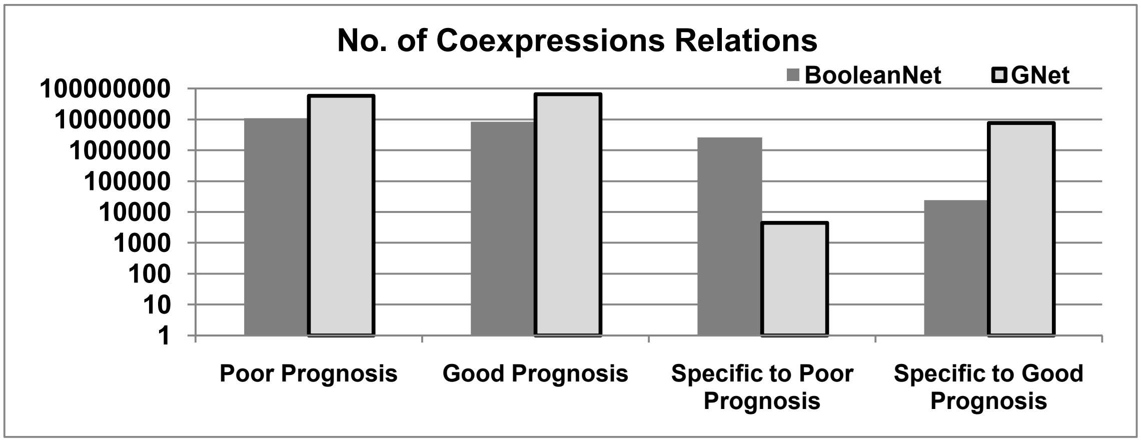


Figure 6.2. Comparison between Boolean implication networks (BooleanNet) and the implication algorithms based on prediction logic (GNet) in term of the number of gene coexpression relations derived.

6.2 Comparison after Parameters Tuning

For a reasonable comparison of both methods with larger and more representative sample, we used two different approaches to define the expression level of genes as up- or down-regulated in order to increase the sample size for deriving the implication relations. In the first approach, the expression level of each gene was defined as up-regulated if the expression value is greater than or equal to the mean of the gene in the cohort, and down-regulated otherwise. In the second approach, the expression level of a gene was defined as up-regulated if the value is half the standard deviation above the mean, down-regulated if it is half the standard deviation below the mean, and intermediate if the expression value lies within half the standard deviation below or above the mean.

In both approaches, the number of implication relations derived using the implication algorithm based on prediction logic is larger than those derived from Boolean implication networks. However, after tuning the minimum precision (∇_{min}) in the induction algorithm based on prediction logic, the number of interactions was reduced to the comparable scale as the Boolean implication networks. (Fig. 6.3).

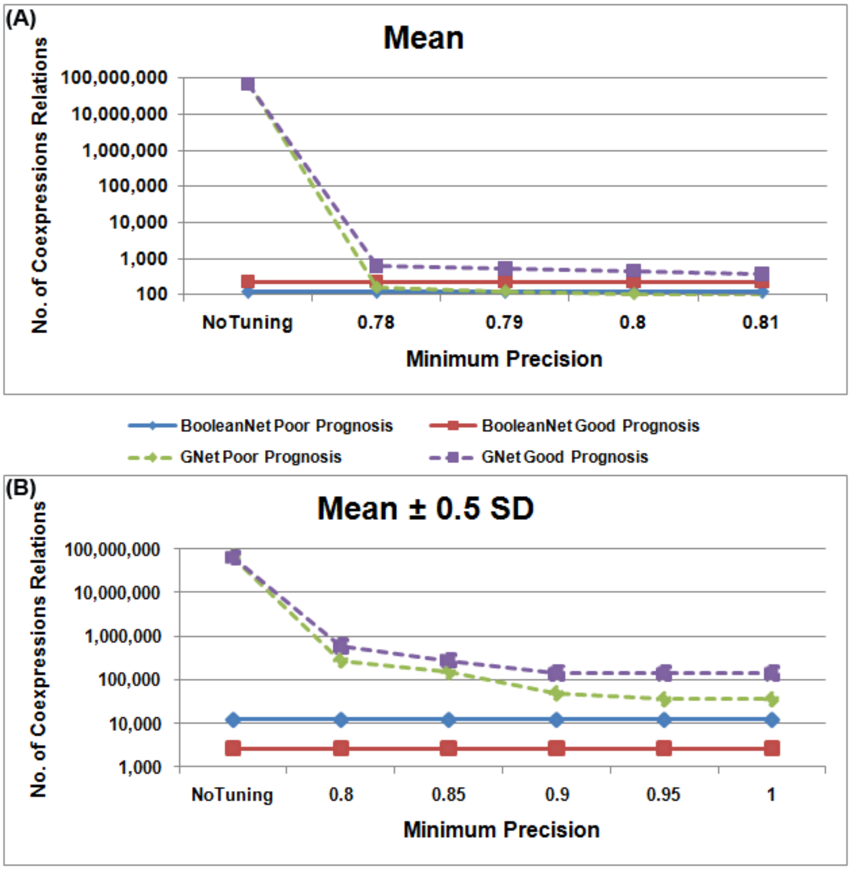


Figure 6.3. Number of implication relations derived from Boolean implication network (BooleanNet) and implication algorithm based on prediction logic (GNet) after tuning the minimum precision parameter in data partitioned by mean only (A) and data partitioned by mean and half the standard deviation (B).

The ∇_{min} was tuned because precision represents the prediction success of the implication rule, which is comparative to the error rate parameter used to decide a successful implication relation in the Boolean implication networks. Results show that the networks derived from implication networks based on prediction logic are comparable to those from Boolean implication networks after tuning ∇_{min}.

6.3 Assessment of Implication Networks with Biological Databases

Having looked at the size of the derived coexpression networks, the biological aspect of the derived coexpression networks was examined with the five gene collections from MSigDB. The precision and false discovery rate of the derived coexpression networks for each prognosis group was measured on the training data. Genes were partitioned into up- or down-regulated with the mean expression of each gene in the cohort. The two measurements from the Boolean implication networks and the implication networks based on prediction logic tuned at ∇_{min} were examined.

Results show that the precision for all derived networks were greater than 95%. However, only precision of the implication networks with $\nabla_{min} = 0.78$ was statistically significant ($P < 0.04$). The precision of the implication networks with $\nabla_{min} = 0.75$ was borderline significant ($P < 0.06$) and the Boolean implication networks was not significant ($P < 0.21$) (Fig. 6.4A). On the other hand, the false discovery rate of the derived networks was all less than 5% (Fig. 6.4B). These results demonstrate that tuning the parameter∇_{min} not only reduces the size of derived implication networks but also affects the biological robustness of the networks.

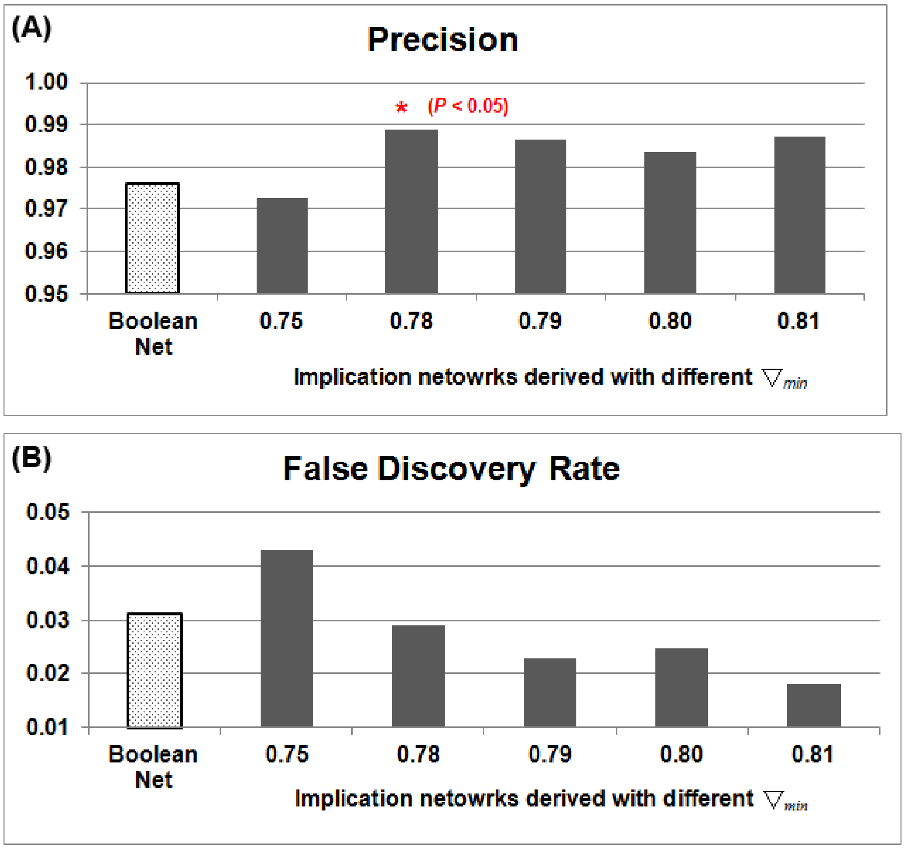


Figure 6.4. Precisions (A) and false discovery rate (B) of the derived implication networks. An asterisk (*) above the bar indicates that the precision is significantly (P < 0.05) higher than the null precisions in 1,000 permutations.

6.4 Comparison with Bayesian Networks

As discussed in Section 2.2.5, Bayesian networks are the most common computational network model for modeling biological networks. It was preferred in biomedical research studies over other network models due to its characteristics in probabilistic structure and tolerance to noise in biological data. Nevertheless, its shortcoming over implication networks in gene coexpression networks is that it could not model feedback loop. In this section, we will compare the biological strength of the disease-specific coexpression networks derived from the implication networks based on prediction logic and the Bayesian networks. Specifically, the precisions and *FDR* of the disease-specific coexpression networks derived from both methodologies for the 21 prognostic signatures identified using the network-based approach (Table B.1-B.3) were compared.

On the patient cohorts from the Director's Challenge Study [2], good-prognosis group is defined as group of patients who survived 5 years or longer after surgery, whereas poor-prognosis group is defined as group of patients who died within 5 years after surgery. For each prognosis group, Bayesian networks were derived using the TETRAD IV[11]. As part of the products from the TETRAD Project by the Carnegie Mellon University, TETRAD IV is the state of the art application for causal models that implements Bayesian networks. TETRAD IV is freely available in various versions, including the user-friendly GUI version and the command line based executable JAR file. In our study, we employed the executable JAR file (version 4.3.10-3) as it allows us to program in scripts and makes the analysis for all 21 signatures more efficient.

Results show that the precision of the disease-specific coexpression networks derived on the training cohort using both methods are comparably high (Fig. 6.5A). Among the 21 signatures, five signatures have precision of 1 for both methods. For the remaining 16 signatures, the networks derived from Bayesian networks had precision comparable to those derived from the implication networks. The precisions of networks from the two methods are significantly different ($P < 0.05$, two-proportion z-tests) on two signatures (S8, S13). For these two signatures, the networks derived from the implication networks have higher proportion of true relations among genes compared with those derived from the Bayesian networks.

Most of the disease-specific coexpression networks derived from both methods have low *FDR* (*FDR* < 0.1; Fig. 6.5B). While only one network (signature S7) derived with the implication networks algorithm has *FDR* higher than 10%; two networks (signatures S2 and S11) derived with the Bayesian network algorithm has *FDR* above 10%. Similar results are obtained in the two testing cohorts (MSK and DFCI; Fig. 6.6). The precision and *FDR* of the disease-specific coexpression networks derived using both methods are comparable to one another when evaluated on each cohort independently.

[11] http://www.phil.cmu.edu/projects/tetrad/

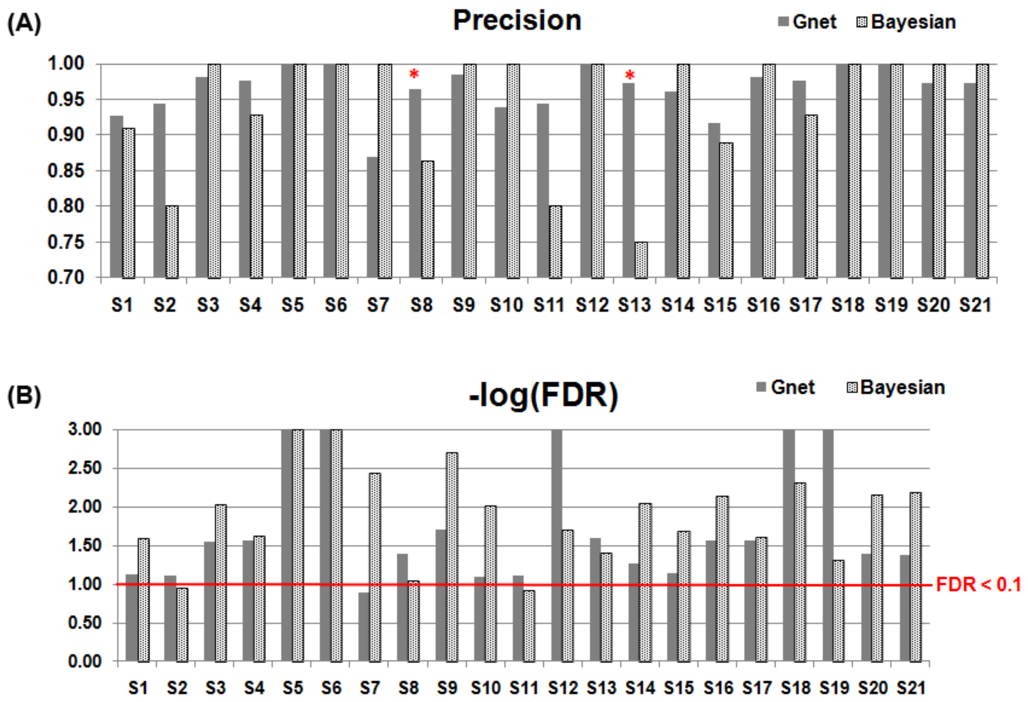


Figure 6.5. Comparison of the disease-specific coexpression networks derived using implication networks and Bayesian networks on the training cohorts from the Director's Challenge Study [2] in terms of precision (A) and false discovery rate (B) for the 21 prognostic signatures. An asterisk (*) above the bar in (A) indicates that the precision of the two derived coexpression networks is statistically significant ($P < 0.05$).

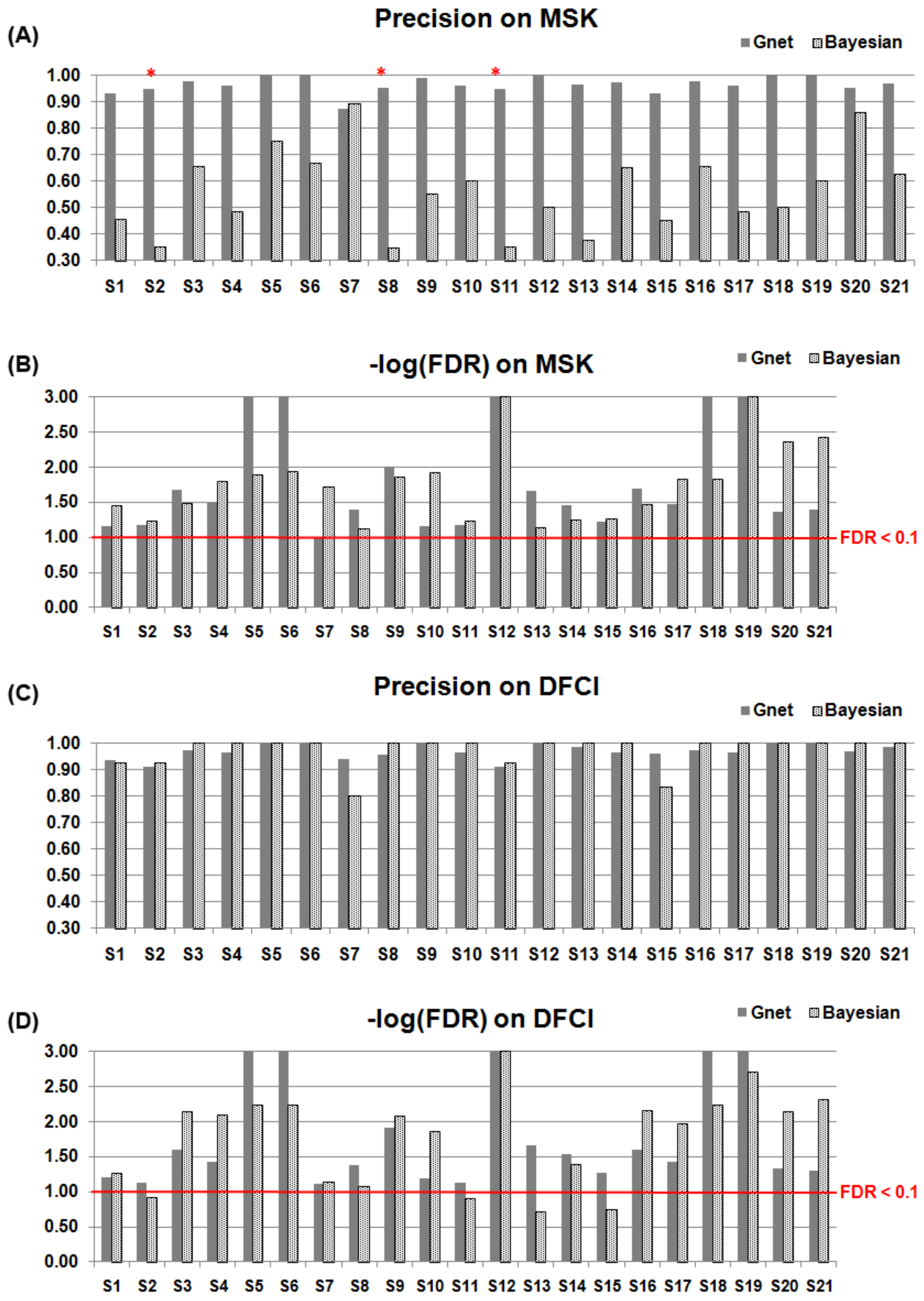


Figure 6.6. Comparison precision and *FDR* of the disease-specific coexpression networks derived using implication networks and Bayesian networks on two independent test cohorts. An asterisk (*) above the bar in (A) indicates that the precision of the two derived coexpression networks is statistically significant ($P < 0.05$).

The more robust approach to evaluate the biological strength of the derived coexpression relations in the networks would be based on the set of coexpression relations commonly present in the networks derived on the training cohort and the two test cohorts. However, for all the 21 signatures, there was no relation commonly found in the disease-specific coexpression networks derived in the training and test cohorts using Bayesian networks. On the other hand, the relations derived from training cohort using the implication network algorithms could be successfully reproduced in both independent test cohorts. The precision of the disease-mediate coexpression networks common in three cohorts is 1 with statistical significance (P < 0.05) as evaluated in 1,000 random permutations for 18 of the 21 signatures (Fig. 6.7A). Furthermore, these networks have *FDR* lower than 10% for all 21 signatures (Fig. 6.7B). Since there was no common relation between the coexpression networks derived from both the training and two test cohorts using Bayesian networks, the precisions are represented with zeros and false discovery rates are represented with NAs (not applicable) in Fig. 6.7.

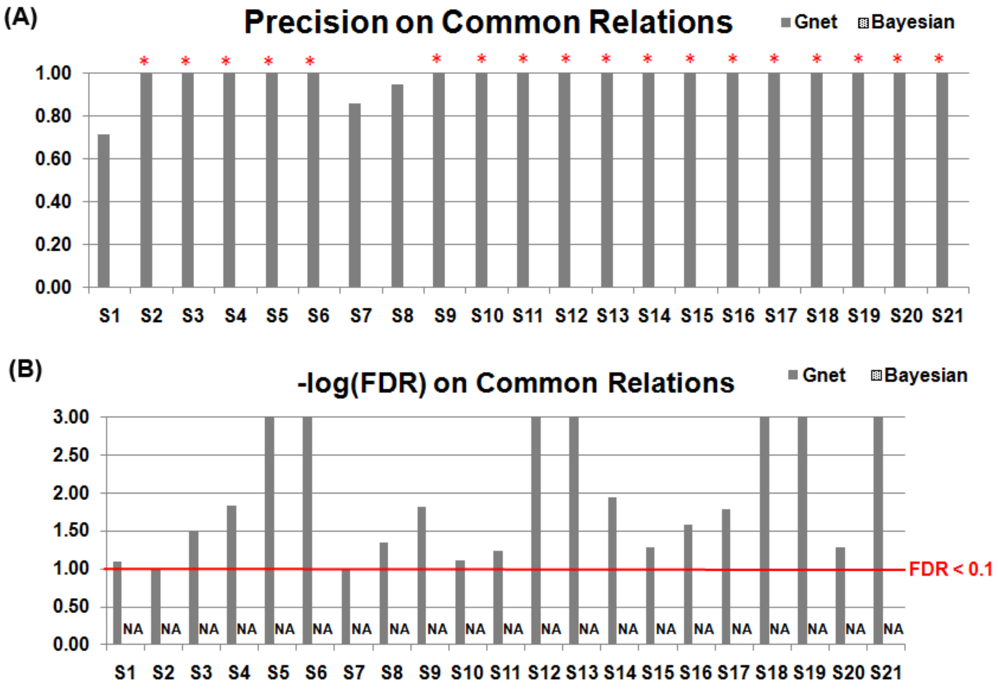


Figure 6.7. Comparison of the implication networks with the Bayesian networks on the disease-specific coexpression relations commonly found on the three studied cohorts. The precision is zero for the Bayesian networks for all 21 signatures in (A) and the *FDR* is NA in (B) as no relation was commonly derived in all three cohorts. The asterisk (*) above the bar in (A) indicates that the precision is significantly (P < 0.05) greater than null precisions in 1,000 permutations.

For the 7-gene smoking-associated signature identified in Chapter 5, seven relations were obtained on the smoking-mediated coexpression networks on the training cohort and 10 relations were obtained on the test cohort [123]. Among these relations, only one smoker-specific coexpression relation was commonly found in both the training and test cohort (Fig. 6.8C). In comparison, with the implication networks algorithm employed in our studies, 18 (17 smoker-specific, 1 non-smoker-specific) coexpressions were commonly derived in both the training and test cohorts (Fig. 6.8A, 6.8B). The precision of the smoking-mediated coexpression networks derived from both methods is 1 and 0.91 for Bayesian and implication networks respectively, with no statistical difference ($P < 0.75$, two-proportion z-test).

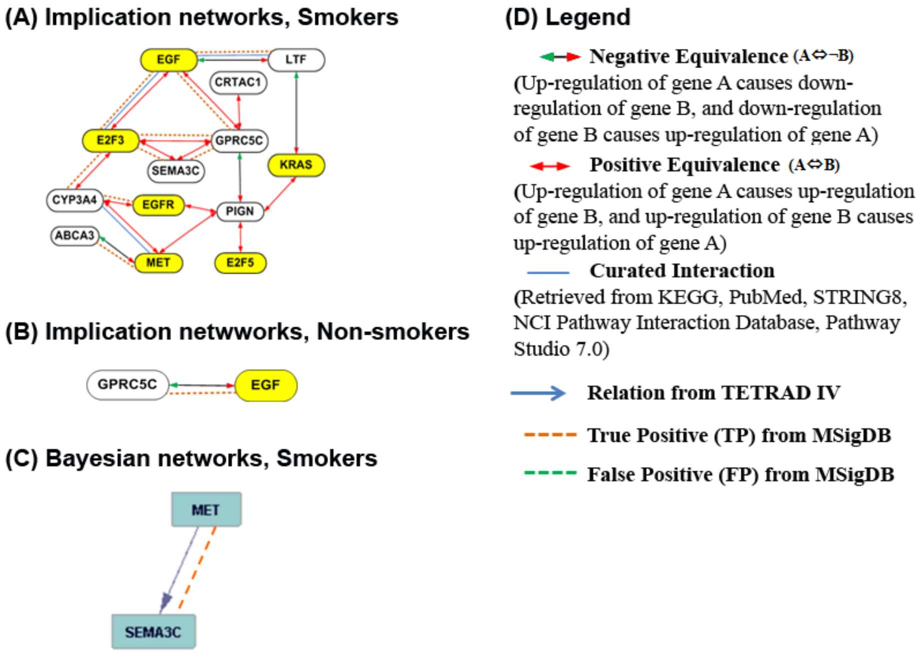

(A) Implication networks, Smokers
EGF
LTF
CRTAC1
E2F3
GPRC5C
SEMA3C
KRAS
CYP3A4
EGFR
PIGN
ABCA3
MET
E2F5
(B) Implication netwworks, Non-smokers
GPRC5C
EGF
(C) Bayesian networks, Smokers
MET
SEMA3C
(D) Legend
Negative Equivalence (A⇔¬B)
(Up-regulation of gene A causes down-regulation of gene B, and down-regulation of gene B causes up-regulation of gene A)
Positive Equivalence (A⇔B)
(Up-regulation of gene A causes up-regulation of gene B, and up-regulation of gene B causes up-regulation of gene A)
Curated Interaction
(Retrieved from KEGG, PubMed, STRING8, NCI Pathway Interaction Database, Pathway Studio 7.0)
Relation from TETRAD IV
True Positive (TP) from MSigDB
False Positive (FP) from MSigDB

Figure 6.8. Comparison of the smoking-mediated coexpression networks for the 7-gene smoking-associated signature from the implication networks and the Bayesian networks. 18 gene coexpression relations specific to smokers (A) and non-smokers (B) derived by the implication networks algorithms commonly present in both training and test cohort. Only one smoker-specific coexpression relations was commonly derived by Bayesian networks in both training and test cohort (C). (D) lists the interpretation of various relations in the networks.

6.5 Conclusions

As discussed at the beginning of the chapter, Boolean implication networks were applied for large scale evolutionary conservation study of genes interactions. Its objective was to explore the biological meaningful relationships among genes for a species or between different species [80]. On the other hand, our applications focused on integrating genes interactions to study how genes change in healthy and disease states and used for gene signatures discovery. Despite the differences of the applications and objectives, the theoretical constructions of both implication networks are similar and worth to be compared.

In the application of the Boolean implication networks, no further information was given on fine tuning the parameters for deriving a successful implication relation. The two statistics used to derive an implication relations are tested against a constant value (3 for the first statistics and 0.1 for the error rate). On the other hand, the minimum requirements for scope (U_p) and precision (∇_{min}) in the implication algorithm adopted in our studies could be adjusted according to the sample size or users' specification. This makes the algorithm we had adopted more flexible. In addition, results from the three comparison studies also demonstrate that the implication networks implemented in our studies is comparable to the Boolean implication networks. The size and biological robustness of the derived networks is comparable after tuning the ∇_{min} parameter of the algorithm. Most importantly, in the lung cancer patient cohorts used in the study, the coexpression relations in the derived Boolean implication networks do not involved with most of the major lung cancer hallmarks. This makes the selection of marker genes with crosstalk to signaling pathways unfeasible.

In the comparison to Bayesian networks, implication networks employed in our studies could reveal more true biological relations than the Bayesian networks. Although the precision and *FDR* of the disease-specific coexpression networks are comparable in both methods when evaluated on the training and test cohorts separately, the relations derived with Bayesian networks in the training cohort could not be reproduced on the test cohort. On the other hand, the gene coexpression relations derived with implication networks in the training cohort could be reproduced in the testing cohorts with low false discovery rate.

Chapter 7

Contributions and Future Work

7.1 Contributions

Lung cancer remains the leading cause of death worldwide. In search for a cure for this fatal disease, it is important to identify clinically relevant prognostic biomarkers in order to develop personalized medicine. More importantly, the discovered biomarkers may reveal fundamental molecular mechanisms of this fatal disease, and improve our knowledge of why patients with certain tumor molecular characteristics have a poor clinical outcome and how to improve their survival time. Studies of biomarker discoveries had been enhanced proficiently by the emergence and development of microarray technologies. Our studies provided a few contributions to this research area.

The first contribution of our studies is the development of a hybrid system for the identification of prognostic genes for lung cancer with traditional statistics and feature selection methods. Results demonstrated that the systematic combinatorial framework of multiple traditional methods in the hybrid model provided improved performance over traditional methods when being applied alone. The 12-gene signature identified from this hybrid model showed better prognostic performance than published signatures. The 12-gene expression-defined prognostic model precisely identified risk for stage I and II patients for different treatments, and accurately predicted chemotherapy drug response. These results implied the clinical utility of the 12-gene signature in the development of personalized therapy.

The second contribution of our works is the extensive identification of prognostic signatures for lung cancer using a network-based hybrid system. The novel network model employed in the second system is the implication networks based on prediction logic. This is an innovative application of using implication networks to model the genome-wide disease-mediated coexpression networks. The implication network-based system not only efficiently scales to the entire genome, but also conveniently couples with information from signaling pathways for the identification of prognostic genes. Using this system, we extensively explored the prognostic signatures of the whole genomic space and discovered 21 signatures with better performance than all published signatures in prognostic categorization on the same patient cohorts. The prognostication evaluation of the signatures were carried out on patients with all tumor stages, stage I only, and stage I without receiving chemotherapy, which was the prognostic capacity never been reported till date. Results also implied that the 21 identified gene signatures could potentially provide a more precise patient selection scheme in stage I patients for adjuvant chemotherapy in personalized lung cancer treatment. Furthermore, the coexpression patterns derived from the implication networks were also successfully validated with molecular interactions reported in the literature.

The third contribution of our studies included the discovery of a 7-gene smoking-associated signature using the implication network-based system. Smoking has been known to be highly associated with lung cancer but yet is not an established clinical factor used in lung cancer prognosis. Our discovery contributed more information to the genes with association to both smoking and lung cancer survival. From our study, the identified 7-gene smoking-associated signature showed strong prognostication for smoking lung cancer patients and accurately identified high-risk patients from a cohort of smokers. The smoking-mediated coexpression networks derived from the implication networks were being validated in experiment by our collaborators. These results implied that the 7-gene signature could potentially be used to develop gene test for more precise prognosis for smoking lung cancer patients, which occupies 90% of all lung cancer patients. It could be used to screen for risk of developing lung cancer for smokers, which could raise cautions to smokers and help advocating them on quitting smoking to reduce health risks.

7.2 Future Work

Results from the comparison studies with the Boolean implication networks presented in Chapter 6 lead us to the future analysis of the methodology through examination of the characteristics of the two implication networks as affected by different parameters, such as the minimum precision, minimum scope, as well as the weights associated with the implication rule and its logical equivalence. In addition, we could also study the comparativeness of the parameters in both algorithms. Through these examinations, we hope to acquire the set of parameters for the derivation of a smaller gene coexpression networks with strong biological robustness (high precision with statistical significance and low false discovery rate) for this research domain to identify marker genes for complex diseases.

Results from Chapter 6 had demonstrated that the number of samples used for deriving the implication networks affects the implication networks derived. Another direction to study the methodology in the future is to study the use of other approach to define genes as up- or down-regulated and the effects on the implication networks derived for prognostic genes identifications. Instead of using mean alone as the threshold to define gene as up- or down-regulated, alternative approach such as the more stringent threshold with standard deviations could be used. However, this would lead to the removal of patient samples that do not pass the threshold, which would lead to smaller sample size.

Appendix A

Published Lung Cancer Molecular Classifiers and Gene Signatures

Table A.1: Summary of gene selection and classification methods of molecular classifiers reported in (Shedden et al, 2008).

Molecular Classifier*	Number of signature genes	Gene selection method(s)	Classification method(s)
Shedden A	~ 9591 Genes	Clustering analysis	Ridged Cox proportional hazard model
Shedden C	23 Genes	SAM, Maximizing Chi-Square analysis (MCA, univariate Cox model and k-mean clustering)	Binary Tree-Structured Vector Quantization (BTSVQ)
Shedden D	37 Genes	SAM, Maximizing Chi-Square analysis (MCA, univariate Cox model and k-mean clustering)	Binary Tree-Structured Vector Quantization (BTSVQ)
Shedden E	1 Gene	Gene Expression Fold Change	Post-hoc split of expression of one gene
Shedden F	42 Genes	Univariate Cox Model	Principal Components and Cox Model
Shedden G	38 Genes	Univariate Cox Model	Principal Components and Cox Model
Shedden H	252 Genes	Scoring and filtering on set of mitosis genes	Majority vote
Shedden J	5 Genes	Univariate Cox model (Chen et al, NEJM 07)	Ridged Cox proportional hazard model
Shedden K	16 Genes	Univariate Cox model (Chen et	Ridged Cox

		al, NEJM 07)	proportional hazard model
Shedden L	9 Genes (from 80 Genes)	Principal Components (Potti et al, NEJM 06)	Ridged Cox proportional hazard model
Shedden M	45 Genes (from 80 Genes)	Principal Components (Potti et al, NEJM 06)	Ridged Cox proportional hazard model
Shedden N	80 Genes	Principal Components (Potti et al, NEJM 06)	Ridged Cox proportional hazard model

*Gene signatures A-H were identified in (Shedden et al, 2008). Gene signatures J and K were identified in (Chen et al, 2007). Gene signatures L, M, and N were identified in (Potti et al, 2006).

Table A.2: 14 published lung cancer gene signatures evaluated in GSEA in Chapter 3.

Signature Name (GSEA)	**First Author**	**Publication PubMed ID**	**No. of Signature Genes/Probes**	**No. of Genes matched in GSEA (By gene symbol)**
Beer_50g	Beer, DG	PMID:12118244	50	45
Bhattacharjee_150g	Bhattacharjee, A	PMID:11707567	150	130
Boutros_6g	Boutros, PC	PMID:19196983	6	6
Chen_5g	Chen, HY	PMID:17202451	5	5
Guo_35g	Guo, L	PMID:16740756	35	34
Lau_3g	Lau, SK	PMID:18065728	3	3
Lu_64g	Lu, Y	PMID:17194181	64	62
Potti_133g	Potti, A	PMID:16899777	133	129
Raponi_50g	Raponi, M	PMID:16885343	50	44
Shedden_MA	Shedden, K	PMID:18641660	13830	8319
Shedden_MB	Shedden, K	PMID:18641660	52	50
Shedden_MC	Shedden, K	PMID:18641660	26	23
Shedden_MD	Shedden, K	PMID:18641660	42	34
Shedden_MH	Shedden, K	PMID:18641660	313	244

Appendix B

Significant Prognostic Signatures Identified Using Network-based Models

Table B.1: Prognostic signatures identified with Approach 1 that generated significant stratifications in patients with all stages, stage I only, and stage I without receiving chemotherapy.

No.	Size of Signature	Hallmarks	Signature genes
S1	21	MET, EGF, KRAS, RB1, E2F1, E2F5	HSPA9, PRDX6, SUPT7L, LEPROT, MPI, QPCT, SLC39A8, ADH1B, MTX1, RAD17, HIPK1, ZFR, CLIC2, TFPI, HEXA, LYST, DYNLRB1, GCC1, CPEB1, ATP1A1, ABHD11
S2	19	EGF, EGFR, KRAS, TP53, E2F3, E2F4	TOMM34, RPS6KA1, ADD2, MPPED1, DNAJC4, IL12RB2, ICA1, THY1, LOC399491, FHL1, WDR43, LRRC23, MRPL13, ZC3H7A, GRHL2, APOA2, CPEB1, LOC100294391, ATP1A1
S3	24	EGF, KRAS, TP53, E2F1, E2F2, E2F4	EEF1B2, TOMM70A, TOMM34, IRF3, DDT, RPS6KA1, SC65, SMAD3, PPM1E, MOCS3, DNAJC4, DNAJA2, GRK6, ZNF592, THY1, FHL1, ACTA2, GRM8, GRHL2, APOA2, CPEB1, FBXO31, PDCD1LG2, HDLBP
S4	32	EGF, KRAS, TP53, E2F1, E2F2, E2F5	PRDX6, ANXA6, TOMM70A, TOMM34, IRF3, RPS6KA1, KATNA1, MPHOSPH9, CCDC9, ZNF141, SCNN1G, DNAJA2, ABCF2, HBS1L, APLP1, ITCH, MTX1, GRK6, NUP214, ANXA9, ELN, ZFR, ZNF592, ACTA2, GRM8, NRN1, APOA2, CPEB1, PDCD1LG2, MUM1, HDLBP, RING1

Table B.2: Prognostic signatures identified with Approach 2 that generated significant stratifications in patients with all stages, stage I only, and stage I without receiving chemotherapy.

No.	Size of Signature	Hallmarks	Signature genes
S5	5	MET, EGFR, E2F2, KRAS, TP53, E2F1, E2F3	CD86, LHX2, GBX1, HEMK1, CPEB1

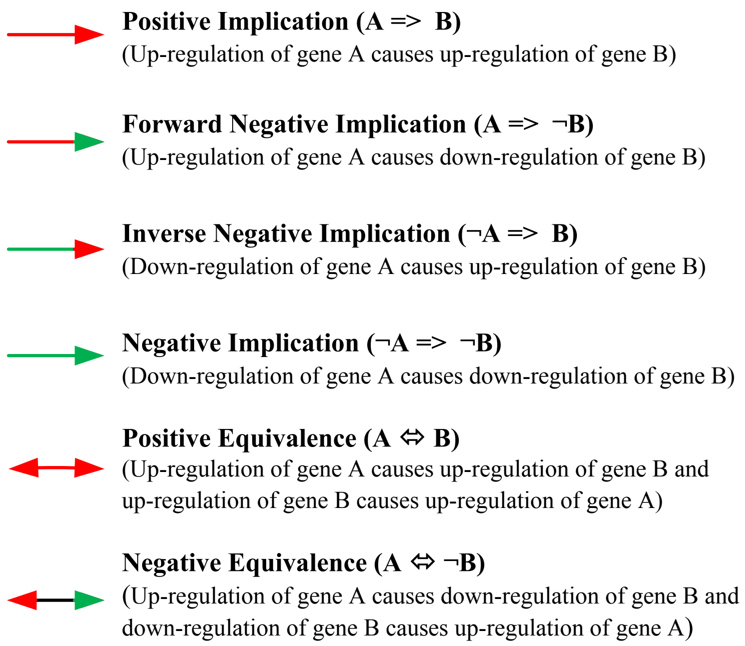
Table B.3: Prognostic signatures identified with Approach 3 that generated significant stratifications in patients with all stages, stage I patients only, and stage I patients without receiving chemotherapy.

No.	Size of Signature	Coexpressed Signaling Hallmarks	Signature genes
S6	4	MET, EGF, EGFR, KRAS, TP53, E2F3	CD86, ICA1, RPAP3, CPEB1
S7	7	MET, EGF, EGFR, KRAS, E2F2, E2F3	ANXA6, SLC17A7, CD86, GAS7, TAF4, ARNT, CPEB1
S8	33	MET, EGF, KRAS, TP53, E2F1, E2F2	EEF1B2, SNRPD2, PRDX6, ANXA6, TOMM70A, NIPSNAP1, IL13RA1, IRF3, DDT, ABCC4, RPS6KA1, SMAD3, CD86, CCDC9, OPRL1, CLDN6, DNAJA2, CCL19, MTX1, MAPK9, ANXA9, ZFR, THY1, SFRS2B, IVD , MKRN2, GRHL2, CPEB1, FBXO31, PDCD1LG2, C20orf30, MUM1, OR1F1
S9	23	MET, EGF, KRAS, E2F1, E2F3, E2F5	HSPA9, ANXA6, MPI, ACTL6A, RPS6KA1, RTCD1, SLC12A2, CCDC9, NDUFAF3, FLT3LG, ANXA9, ZFR, CLIC2, SOSTDC1, TRMU, TCF3, DYNLRB1, CPEB1, C20orf46, LOC100294391, ATP1A1, MUM1, ABHD11
S10	7	EGF, EGFR, KRAS, TP53, RB1, E2F2	RPL18, VIPR2, MOCS3, DNAJC4, ADAMTSL3, WDR12, HDLBP
S11	19	EGF, EGFR, KRAS, TP53, E2F3, E2F4	TOMM34, RPS6KA1, ADD2, MPPED1, DNAJC4, IL12RB2, ICA1, THY1, LOC399491, FHL1, WDR43, LRRC23, MRPL13, ZC3H7A, GRHL2, APOA2, CPEB1, LOC100294391, ATP1A1
S12	7	EGF, EGFR, TP53, RB1, E2F1, E2F2	MOCS3, DNAJC4, CCBP2, THY1, SFRS2B, PUM2, HDLBP
S13	10	EGF, KRAS, TP53, RB1, E2F1, E2F2	PRDX6, MOCS3, OPRL1, HBS1L, MTX1, ZFR, SPIN1, CPEB1, OR1F1, HDLBP
S14	15	EGF, KRAS, TP53, RB1, E2F1, E2F4	DDT, MOCS3, MPPED1, DNAJC4, RGL1, CEP57, THY1, TFPI, LRRC23, MRPL13, CPEB1, FBXO31, ATP1A1, HDLBP, SFTPB
S15	21	EGF, KRAS, TP53, RB1, E2F1, E2F5	RPL30, PRDX6, SNX2, LEPROT, MPI, KATNA1, SLC39A8, HBS1L, MTX1, ELN, ZFR, ANGEL1, TFPI, LRRC23, NRN1, SLC35F2, HMBOX1, CPEB1, ATP1A1, GINS2, HDLBP

S16	24	EGF, KRAS, TP53, E2F1, E2F2, E2F4	EEF1B2, TOMM70A, TOMM34, IRF3, DDT, RPS6KA1, SC65, SMAD3, PPM1E, MOCS3, DNAJC4, DNAJA2, GRK6, ZNF592, THY1, FHL1, ACTA2, GRM8, GRHL2, APOA2, CPEB1, FBXO31, PDCD1LG2, HDLBP
S17	32	EGF, KRAS, TP53, E2F1, E2F2, E2F5	PRDX6, ANXA6, TOMM70A, TOMM34, IRF3, RPS6KA1, KATNA1, MPHOSPH9, CCDC9, ZNF141, SCNN1G, DNAJA2, ABCF2, HBS1L, APLP1, ITCH, MTX1, GRK6, NUP214, ANXA9, ELN, ZFR, ZNF592, ACTA2, GRM8, NRN1, APOA2, CPEB1, PDCD1LG2, MUM1, HDLBP, RING1
S18	6	EGF, KRAS, TP53, E2F2, E2F3, E2F5	KIAA0040, KCNS3, KCNA4, COL14A1, CPEB1, RING1
S19	3	EGF, KRAS, RB1, E2F1, E2F3, E2F5	HSPA9, ABHD11, C9orf156
S20	9	EGFR, KRAS, RB1, TP53, E2F1, E2F2	TRAP1, PRMT2, MOCS3, DNAJC4, CCL8, TFCP2L1, LOH3CR2A, HDLBP, PKNOX2
S21	9	EGFR, KRAS, RB1, E2F5, TP53, E2F2,	TRAP1, VIPR2, TCP10, TBX1, CCL8, LDLR, WDR12, PRR15L, HDLBP

Appendix C

Disease-specific Coexpression Networks for the Prognostic Signatures Identified Using Network-based Models



Positive Implication (A => B)
(Up-regulation of gene A causes up-regulation of gene B)

Forward Negative Implication (A => ¬B)
(Up-regulation of gene A causes down-regulation of gene B)

Inverse Negative Implication (¬A => B)
(Down-regulation of gene A causes up-regulation of gene B)

Negative Implication (¬A => ¬B)
(Down-regulation of gene A causes down-regulation of gene B)

Positive Equivalence (A ⇔ B)
(Up-regulation of gene A causes up-regulation of gene B and up-regulation of gene B causes up-regulation of gene A)

Negative Equivalence (A ⇔ ¬B)
(Up-regulation of gene A causes down-regulation of gene B and down-regulation of gene B causes up-regulation of gene A)

Figure C.1. Legend of expression relations of the disease-specific coexpression networks represented in the six implication rules.

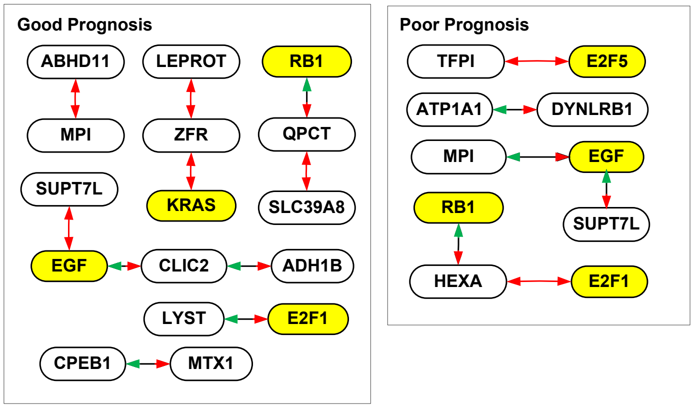


Figure C.2. Disease-specific coexpression networks for signature S1 (precision = 0.71, FDR = 0.08)

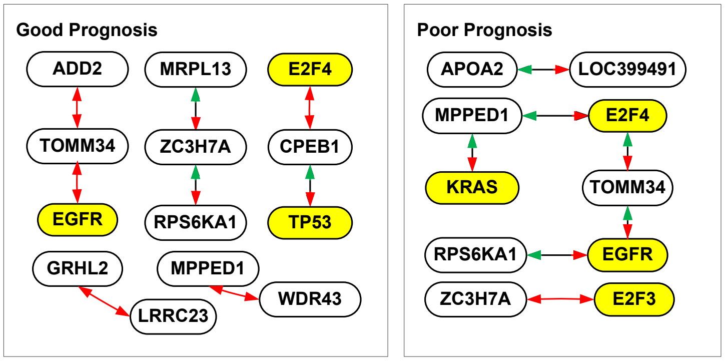


Figure C.3. Disease-specific coexpression networks for signature S2 (precision = 1, *FDR* = 0.10)

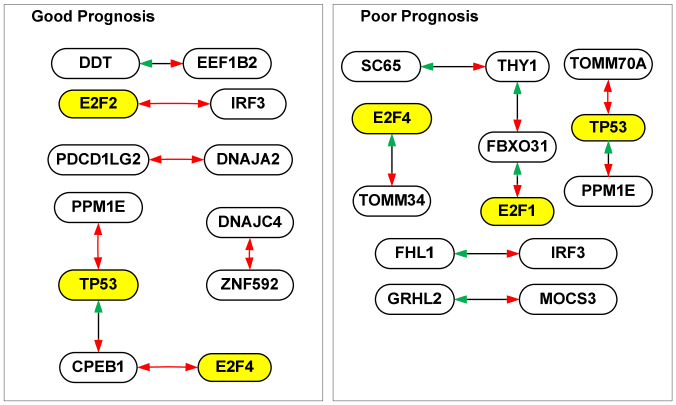


Figure C.4. Disease-specific coexpression networks for signature S3 (precision = 1, *FDR* = 0.03)

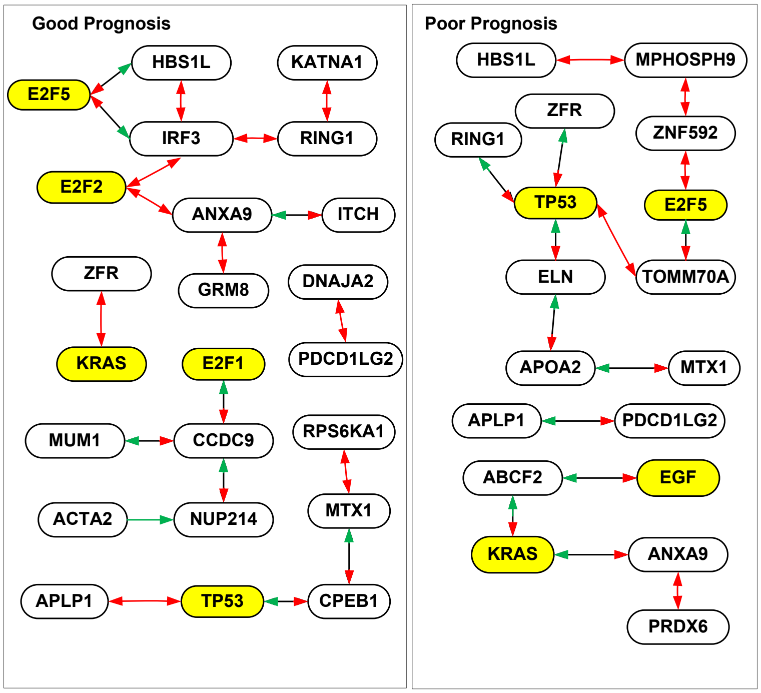


Figure C.5. Disease-specific coexpression networks for signature S4 (precision = 1, *FDR* = 0.01)

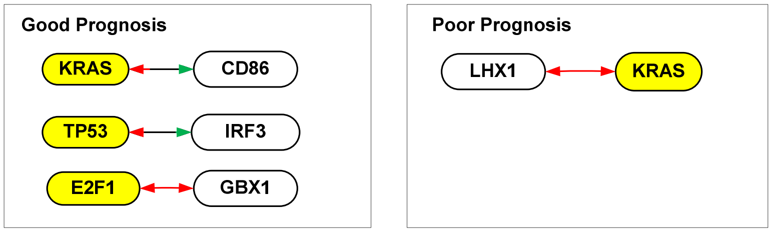


Figure C.6. Disease-specific coexpression networks for signature S5 (precision = 1, *FDR* = 0)

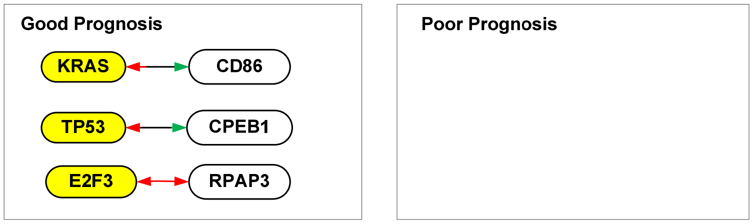


Figure C.7. Disease-specific coexpression networks for signature S6 (precision = 1, *FDR* = 0)

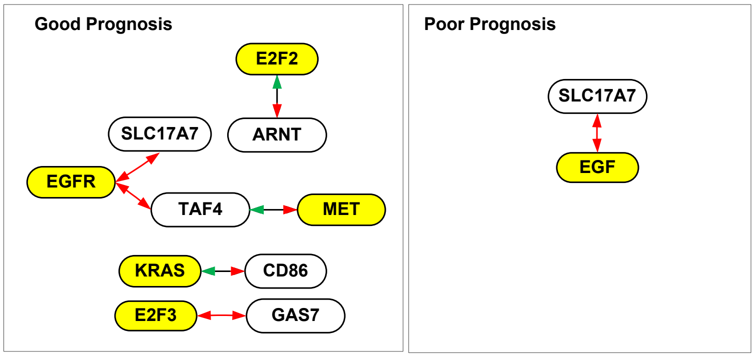


Figure C.8. Disease-specific coexpression networks for signature S7 (precision = 0.86, *FDR* = 0.10)

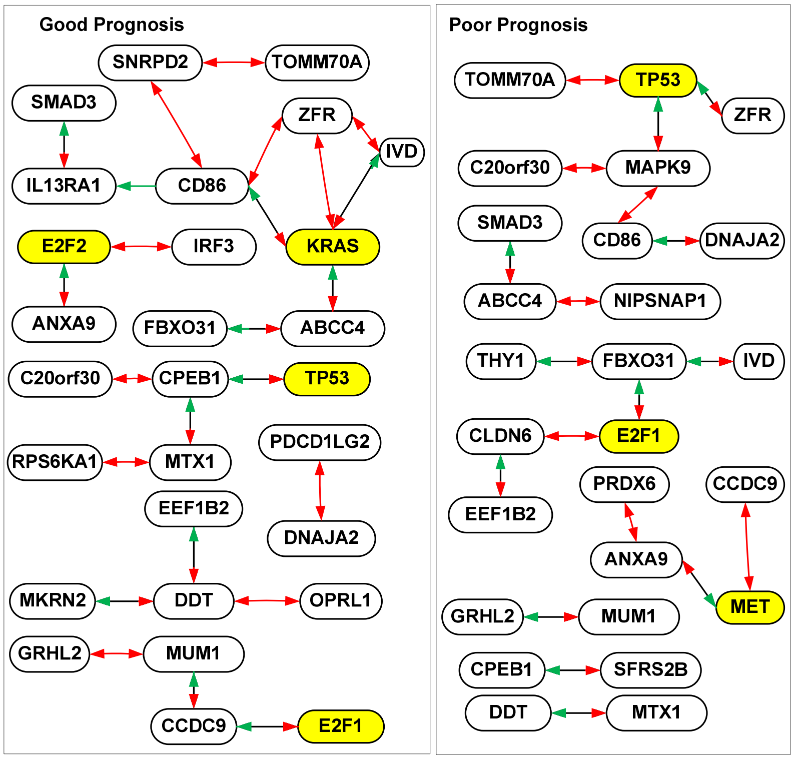


Figure C.9. Disease-specific coexpression networks for signature S8 (precision = 0.95, *FDR* = 0.05)

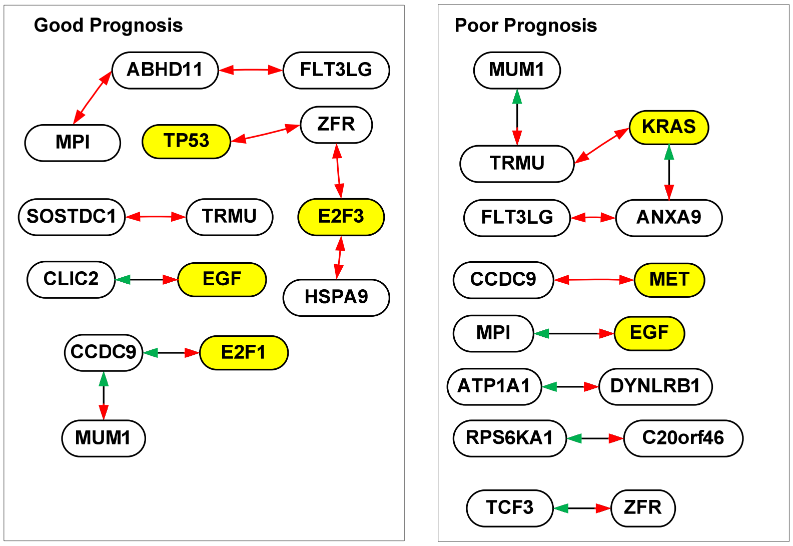


Figure C.10. Disease-specific coexpression networks for signature S9 (precision = 1, *FDR* = 0.02)

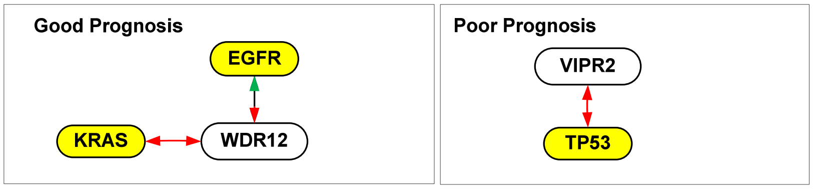


Figure C.11. Disease-specific coexpression networks for signature S10 (precision = 1, *FDR* = 0.08)

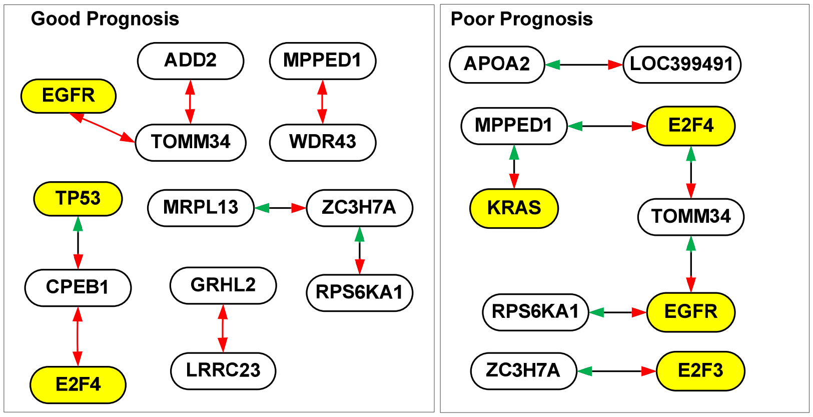


Figure C.12. Disease-specific coexpression networks for signature S11 (precision = 1, *FDR* = 0.06)

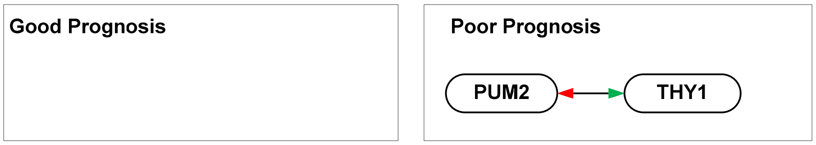


Figure C.13. Disease-specific coexpression networks for signature S12 (precision = 1, *FDR* = 0)

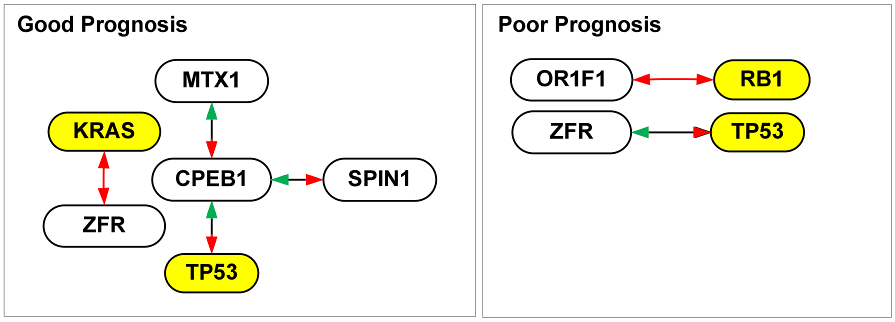


Figure C.14. Disease-specific coexpression networks for signature S13 (precision = 1, *FDR* = 0)

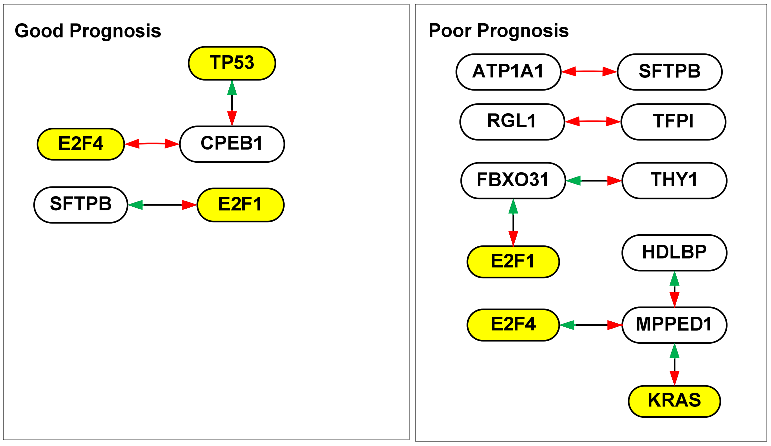


Figure C.15. Disease-specific coexpression networks for signature S14 (precision = 1, *FDR* = 0.01)

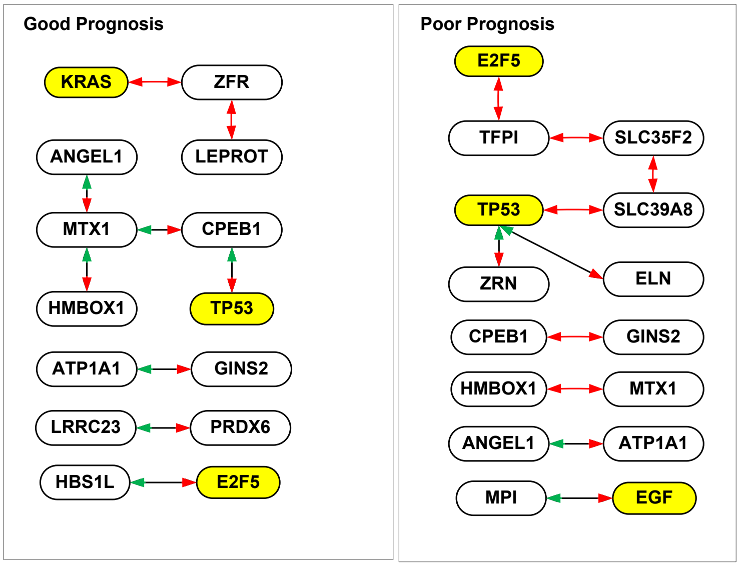


Figure C.16. Disease-specific coexpression networks for signature S15 (precision = 1, *FDR* = 0.05)

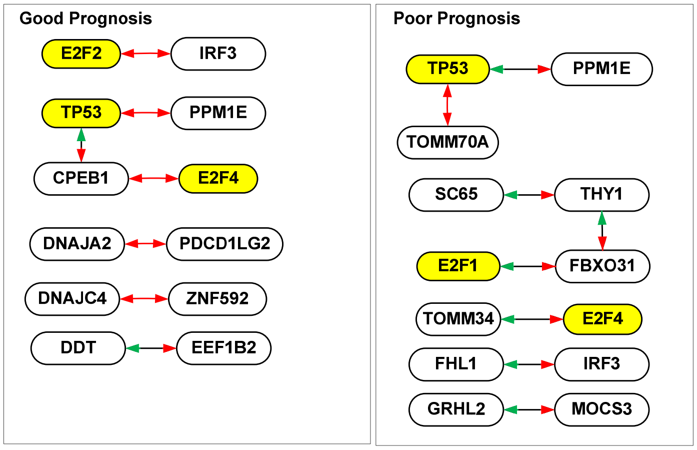


Figure C.17. Disease-specific coexpression networks for signature S16 (precision = 1, *FDR* = 0.03)

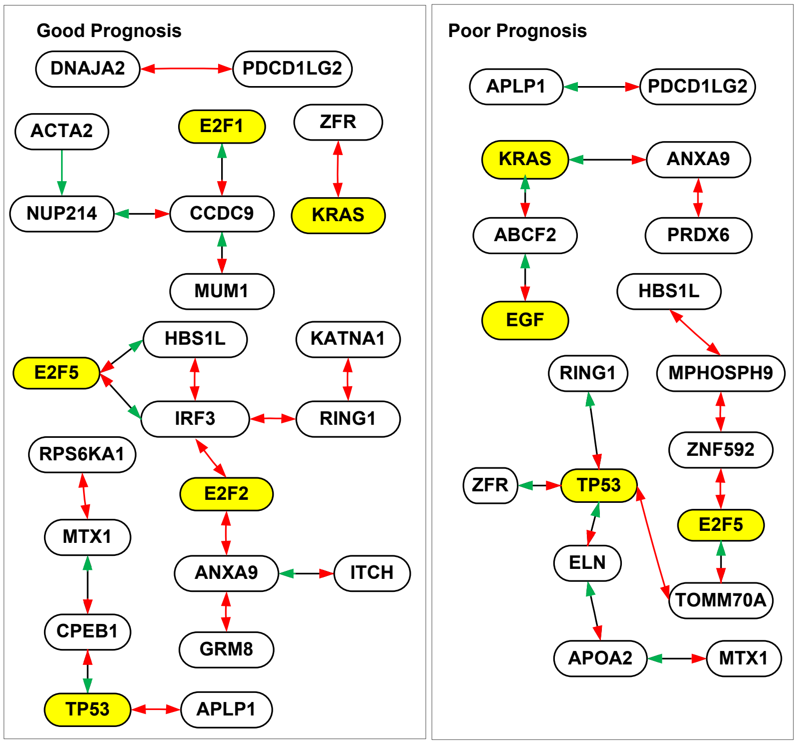


Figure C.18. Disease-specific coexpression networks for signature S17 (precision = 1, *FDR* = 0.02)

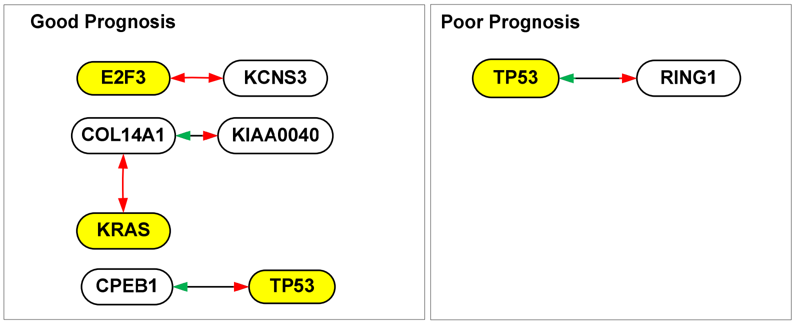


Figure C.19. Disease-specific coexpression networks for signature S18 (precision = 1, *FDR* = 0)

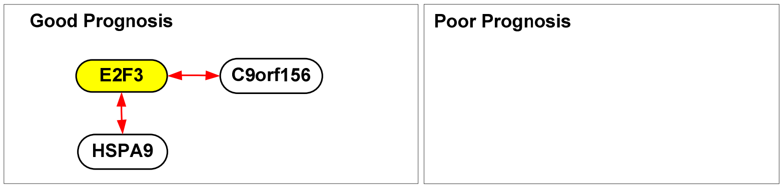


Figure C.20. Disease-specific coexpression networks for signature S19 (precision = 1, *FDR* = 0)

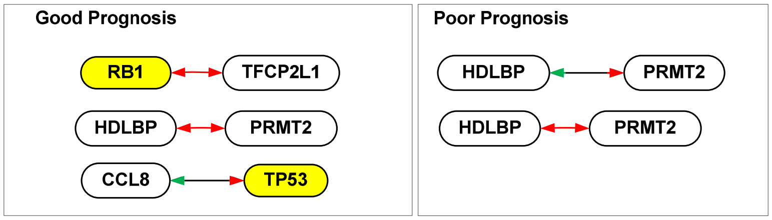


Figure. C.21. Disease-specific coexpression networks for signature S20 (precision = 1, *FDR* = 0.05)

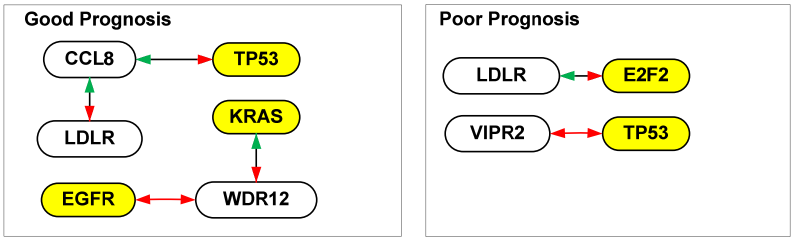


Figure C.22. Disease-specific coexpression networks for signature S21 (precision = 1, *FDR*=0.001)

Bibliography

[1] L. Guo, B. Cukic, and H. Singh, "Predicting Fault Prone Modules by the Dempster-Shafer Belief Networks," *18th IEEE International Conference on Automated Software Engineering (ASE'03)*. pp.249-252, 2003.

[2] K. Shedden, J. M. Taylor, S. A. Enkemann et al., "Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study," *Nat.Med,* vol. 14, no. 8. pp.822-827, Aug., 2008.

[3] American Cancer Society, "Cancer Facts and Figures 2005.," *Atlanta, Ga: American Cancer Society*, 2005.

[4] T. Naruke, T. Goya, R. Tsuchiya et al., "Prognosis and survival in resected lung carcinoma based on the new international staging system," *J Thorac.Cardiovasc.Surg.,* vol. 96, no. 3. pp.440-447, Sept., 1988.

[5] P. C. Hoffman, A. M. Mauer, and E. E. Vokes, "Lung cancer," *Lancet,* vol. 355, no. 9202. pp.479-485, Feb., 2000.

[6] W. S. Dalton and S. H. Friend, "Cancer biomarkers--an invitation to the table," *Science,* vol. 312, no. 5777. pp.1165-1168, May, 2006.

[7] C. Sotiriou and M. J. Piccart, "Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?," *Nat.Rev.Cancer,* vol. 7, no. 7. pp.545-553, July, 2007.

[8] S. Paik, S. Shak, G. Tang et al., "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer," *N.Engl.J.Med.,* vol. 351, no. 27. pp.2817-2826, Dec., 2004.

[9] L. J. van 't Veer, H. Dai, M. J. van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature,* vol. 415, no. 6871. pp.530-536, Jan., 2002.

[10] M. J. van de Vijver, Y. D. He, L. J. van 't Veer et al., "A gene-expression signature as a predictor of survival in breast cancer," *N.Engl.J.Med.,* vol. 347, no. 25. pp.1999-2009, Dec., 2002.

[11] D. G. Beer, S. L. Kardia, C. C. Huang et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nat.Med.,* vol. 8, no. 8. pp.816-824, Aug., 2002.

[12] A. Bhattacharjee, W. G. Richards, J. Staunton et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc.Natl.Acad.Sci.U.S.A,* vol. 98, no. 24. pp.13790-13795, Nov., 2001.

[13] A. H. Bild, G. Yao, J. T. Chang et al., "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature,* vol. 439, no. 7074. pp.353-357, Jan., 2006.

[14] A. C. Borczuk, H. K. Kim, H. A. Yegen et al., "Lung adenocarcinoma global profiling identifies type II transforming growth factor-beta receptor as a repressor of invasiveness," *Am.J.Respir.Crit Care Med.,* vol. 172, no. 6. pp.729-737, Sept., 2005.

[15] H. Y. Chen, S. L. Yu, C. H. Chen et al., "A five-gene signature and clinical outcome in non-small-cell lung cancer," *N.Engl.J.Med.,* vol. 356, no. 1. pp.11-20, Jan., 2007.

[16] A. Potti, S. Mukherjee, R. Petersen et al., "A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer," *N.Engl.J.Med.,* vol. 355, no. 6. pp.570-580, Aug., 2006.

[17] M. Raponi, Y. Zhang, J. Yu et al., "Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung," *Cancer Res.,* vol. 66, no. 15. pp.7466-7472, Aug., 2006.

[18] J. Subramanian and R. Simon, "Gene expression-based prognostic signatures in lung cancer: ready for clinical use?," *J.Natl.Cancer Inst.,* vol. 102, no. 7. pp.464-474, Apr., 2010.

[19] R. L. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics,* vol. 19, no. 12. pp.1484-1491, Aug., 2003.

[20] L. Hood, J. R. Heath, M. E. Phelps et al., "Systems biology and new technologies enable predictive and preventative medicine," *Science,* vol. 306, no. 5696. pp.640-643, Oct., 2004.

[21] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics,* vol. 23, no. 19. pp.2507-2517, Oct., 2007.

[22] J. K. Lee, P. D. Williams, and S. Cheon, "Data mining in genomics," *Clinics in Laboratory Medicine,* vol. 28, no. 1. pp.145-+, Mar., 2008.

[23] S. G. Baker, B. S. Kramer, and S. Srivastava, "Markers for early detection of cancer: statistical guidelines for nested case-control studies," *BMC.Med.Res.Methodol.,* vol. 2. pp.4, 2002.

[24] B. Emir, S. Wieand, J. Q. Su et al., "Analysis of repeated markers used to predict progression of cancer," *Stat.Med.,* vol. 17, no. 22. pp.2563-2578, Nov., 1998.

[25] M. S. Pepe, H. Janes, G. Longton et al., "Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker," *Am.J.Epidemiol.,* vol. 159, no. 9. pp.882-890, May, 2004.

[26] L. H. Hartwell, J. J. Hopfield, S. Leibler et al., "From molecular to modular cell biology," *Nature,* vol. 402, no. 6761 Suppl. pp.C47-C52, Dec., 1999.

[27] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Res.,* vol. 18, no. 4. pp.644-652, Apr., 2008.

[28] J. K. Lee, P. D. Williams, and S. Cheon, "Data mining in genomics," *Clinics in Laboratory Medicine,* vol. 28, no. 1. pp.145-+, Mar., 2008.

[29] P. C. Boutros, S. K. Lau, M. Pintilie et al., "Prognostic gene signatures for non-small-cell lung cancer," *Proc.Natl.Acad.Sci.U.S.A,* vol. 106, no. 8. pp.2824-2828, Feb., 2009.

[30] Y. Lu, W. Lemon, P. Y. Liu et al., "A gene expression signature predicts survival of patients with stage I non-small cell lung cancer," *PLoS.Med.,* vol. 3, no. 12. pp.e467, Dec., 2006.

[31] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biol.,* vol. 4, no. 4. pp.210, 2003.

[32] K. Kadota, Y. Nakai, and K. Shimizu, "Ranking differentially expressed genes from Affymetrix gene expression data: methods with reproducibility, sensitivity, and specificity," *Algorithms.Mol.Biol.,* vol. 4. pp.7, 2009.

[33] J. K. Lee, P. D. Williams, and S. Cheon, "Data mining in genomics," *Clin.Lab Med.,* vol. 28, no. 1. pp.145-66, viii, Mar., 2008.

[34] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc.Natl.Acad.Sci.U.S.A,* vol. 98, no. 9. pp.5116-5121, Apr., 2001.

[35] D. Cox, "Regression models and life-tables (with discussion).," *Journal of the Royal Statistical Society, Series B, Methodological,* vol. 34. pp.187-220, 1972.

[36] F. E. Harrell, Jr., K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Stat.Med.,* vol. 15, no. 4. pp.361-387, Feb., 1996.

[37] J. Fox, "Cox Proportional-Hazards Regression for Survival Data," *Appendix to An R and S-PLUS Companion to Applied Regression*, 2002.

[38] R. Diaz-Uriarte and d. A. Alvarez, "Gene selection and classification of microarray data using random forest," *BMC.Bioinformatics.,* vol. 7. pp.3, 2006.

[39] R. Jiang, W. Tang, X. Wu et al., "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC.Bioinformatics,* vol. 10 Suppl 1. pp.S65, 2009.

[40] K. Y. Kim, D. H. Ki, H. C. Jeung et al., "Improving the prediction accuracy in classification using the combined data sets by ranks of gene expressions," *BMC.Bioinformatics,* vol. 9. pp.283, 2008.

[41] Y. A. Meng, Y. Yu, L. A. Cupples et al., "Performance of random forest when SNPs are in linkage disequilibrium," *BMC.Bioinformatics,* vol. 10. pp.78, 2009.

[42] L. Breiman, "Random Forests," *Machine Learning,* vol. 45. pp.5-32 , 2001.

[43] B. H. Menze, B. M. Kelm, R. Masuch et al., "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinformatics,* vol. 10. pp.213, 2009.

[44] C. Strobl, A. L. Boulesteix, A. Zeileis et al., "Bias in random forest variable importance measures: illustrations, sources and a solution," *BMC.Bioinformatics.,* vol. 8. pp.25, 2007.

[45] R. L. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics,* vol. 19, no. 12. pp.1484-1491, Aug., 2003.

[46] I. T. Jolliffe, *Principle Component Analysis,* 2nd Edition: Springer, 2002.

[47] I. Guyon, J. Weston, S. Barnhill et al., "Gene selection for cancer classification using support vector machines," *Machine Learning,* vol. 46, no. 1-3. pp.389-422, 2002.

[48] A. Wang and E. A. Gehan, "Gene selection for microarray data analysis using principal component analysis," *Stat.Med.,* vol. 24, no. 13. pp.2069-2087, July, 2005.

[49] Z. Liu, D. Chen, and H. Bensmail, "Gene expression data classification with Kernel principal component analysis," *J.Biomed.Biotechnol.,* vol. 2005, no. 2. pp.155-159, June, 2005.

[50] K. Kira and L. Rendell, "A Practical Approach to Feature Selection," *Proceedings of the Ninth International Workshop on Machine Learning (Aberdeen, Scotland, UK)*. pp.249-256, 1992.

[51] I. Kononenko, E. Simec, and M. Robnik-Sikonja, "Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF," *Applied Intelligence,* vol. 7, no. 1. pp.39-55, 1997.

[52] M. A. Hall and G. Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining," *IEEE Transactions on Knowledge and Data Engineering,* vol. 15, no. 3. pp.1437-1447, 2003.

[53] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition)* : Morgan Kaufmann, 2005.

[54] R. Breitling, P. Armengaud, A. Amtmann et al., "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments," *Febs Letters,* vol. 573, no. 1-3. pp.83-92, Aug., 2004.

[55] J. T. Dudley, R. Tibshirani, T. Deshpande et al., "Disease signatures are robust across tissues and experiments," *Mol.Syst.Biol.,* vol. 5. pp.307, 2009.

[56] L. Breiman, "Bagging predictors," *Machine Learning,* vol. 24, no. 2. pp.123-140, Aug., 1996.

[57] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning,* vol. 53, no. 1-2. pp.23-69, Oct., 2003.

[58] J. Zhu, B. Zhang, E. N. Smith et al., "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks," *Nat Genet,* vol. 40, no. 7. pp.854-861, July, 2008.

[59] V. Emilsson, G. Thorleifsson, B. Zhang et al., "Genetics of gene expression and its effect on disease," *Nature,* vol. 452, no. 7186. pp.423-428, Mar., 2008.

[60] S. E. Calvano, W. Xiao, D. R. Richards et al., "A network-based analysis of systemic inflammation in humans," *Nature,* vol. 437, no. 7061. pp.1032-1037, Oct., 2005.

[61] H. Y. Chuang, E. Lee, Y. T. Liu et al., "Network-based classification of breast cancer metastasis," *Mol.Syst.Biol.,* vol. 3. pp.140, 2007.

[62] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nat.Rev.Mol.Cell Biol.,* vol. 9, no. 10. pp.770-780, Oct., 2008.

[63] T. M. Mitchell, *Machine Learning*, pp. 1-413: McGraw Hill, 1997.

[64] T. M. Khoshgoftaar, D. L. Lanning, and A. S. Pandya, "A neural network modeling methodology for the detection of high-risk programs." *Proceedings of 1993 IEEE International Symposium on Software Reliability Engineering* , pp. 302-309. Nov. 93 A.D. 11-3-0093.
Ref Type: Conference Proceeding

[65] J. Khan, J. S. Wei, M. Ringner et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nat.Med.,* vol. 7, no. 6. pp.673-679, June, 2001.

[66] J. S. Wei, B. T. Greer, F. Westermann et al., "Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma," *Cancer Res.,* vol. 64, no. 19. pp.6883-6891, Oct., 2004.

[67] M. C. O'Neill and L. Song, "Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect," *BMC Bioinformatics,* vol. 4. pp.13, Apr., 2003.

[68] Y. Xu, F. M. Selaru, J. Yin et al., "Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer," *Cancer Res.,* vol. 62, no. 12. pp.3493-3497, June, 2002.

[69] E. Keedwell and A. Narayanan, "Discovering gene networks with a neural-genetic hybrid," *IEEE/ACM Trans.Comput.Biol.Bioinform.,* vol. 2, no. 3. pp.231-242, July, 2005.

[70] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian Networks - the Combination of Knowledge and Statistical-Data," *Machine Learning,* vol. 20, no. 3. pp.197-243, Sept., 1995.

[71] D. Heckerman, A. Mamdani, and M. P. Wellman, "Real-World Applications of Bayesian Networks - Introduction," *Communications of the Acm,* vol. 38, no. 3. pp.24-26, Mar., 1995.

[72] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science,* vol. 303, no. 5659. pp.799-805, Feb., 2004.

[73] R. Jansen, H. Yu, D. Greenbaum et al., "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science,* vol. 302, no. 5644. pp.449-453, Oct., 2003.

[74] K. Sachs, O. Perez, D. Pe'er et al., "Causal protein-signaling networks derived from multiparameter single-cell data," *Science,* vol. 308, no. 5721. pp.523-529, Apr., 2005.

[75] E. Charniak, "Bayesian Networks Without Tears," *Ai Magazine,* vol. 12, no. 4. pp.50-63, 1991.

[76] N. Friedman, M. Linial, I. Nachman et al., "Using Bayesian networks to analyze expression data," *J Comput.Biol.,* vol. 7, no. 3-4. pp.601-620, 2000.

[77] N. Friedman, I. Nachman, and D. Pe'er, "Learning Bayesian network structure from massive datasets: The "sparse candidate" algorithm." *Proc.Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI '99)* , pp. 196-205. 1999.
Ref Type: Conference Proceeding

[78] J. Liu and M. C. Desmarais, "A Method of Learning Implication Networks from Empirical Data: Algorithm and Monte-Carlo Simulation-Based Validation," *IEEE Transactions on Knowledge and Data Engineering,* vol. 9, no. 6. pp.990-1004, 1997.

[79] D. K. Hildebrand, J. D. Laing, and H. Rosenthal, *Prediction Analysis of Cross Classifications*: John Wiley & Sons, 1977.

[80] D. Sahoo, D. L. Dill, A. J. Gentles et al., "Boolean implication networks derived from large scale, whole genome microarray datasets," *Genome Biol.,* vol. 9, no. 10. pp.R157, 2008.

[81] D. Sahoo, D. L. Dill, R. Tibshirani et al., "Extracting binary signals from microarray time-course data," *Nucleic Acids Res.,* vol. 35, no. 11. pp.3705-3712, 2007.

[82] M. C. O'Neill and L. Song, "Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect," *BMC Bioinformatics,* vol. 4. pp.13, Apr., 2003.

[83] L. Guo, *Software Quality and Reliability Prediction Using Dempster-Shafer Theory*: Ph.D. Dissertation, West Virginia University, 2004.

[84] Z. Boger, "Artificial neural networks methods for identification of the most relevant genes from gene expression array data." *Proceedings Of The International Joint Conference On Neural Networks 2003* vol. 4, pp. 3095-3100. July 2003.
Ref Type: Conference Proceeding

[85] J. Zhu, M. C. Wiener, C. Zhang et al., "Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations," *PLoS.Comput.Biol.,* vol. 3, no. 4. pp.e69, Apr., 2007.

[86] R. Milo, S. Shen-Orr, S. Itzkovitz et al., "Network motifs: simple building blocks of complex networks," *Science,* vol. 298, no. 5594. pp.824-827, Oct., 2002.

[87] R. Milo, S. Itzkovitz, N. Kashtan et al., "Superfamilies of evolved and designed networks," *Science,* vol. 303, no. 5663. pp.1538-1542, Mar., 2004.

[88] S. Wuchty, Z. N. Oltvai, and A. L. Barabasi, "Evolutionary conservation of motif constituents in the yeast protein interaction network," *Nat Genet,* vol. 35, no. 2. pp.176-179, Oct., 2003.

[89] S. Y. Kim, S. Imoto, and S. Miyano, "Inferring gene networks from time series microarray data using dynamic Bayesian networks," *Brief.Bioinform.,* vol. 4, no. 3. pp.228-235, Sept., 2003.

[90] D. Pe'er, A. Regev, G. Elidan et al., "Inferring subnetworks from perturbed expression profiles," *Bioinformatics,* vol. 17 Suppl 1. pp.S215-S224, 2001.

[91] M. K. Kerr, M. Martin, and G. A. Churchill, "Analysis of variance for gene expression microarray data," *J.Comput.Biol.,* vol. 7, no. 6. pp.819-837, 2000.

[92] R. D. Wolfinger, G. Gibson, E. D. Wolfinger et al., "Assessing gene significance from cDNA microarray expression data via mixed models," *J.Comput.Biol.,* vol. 8, no. 6. pp.625-637, 2001.

[93] F. E. Harrell, Jr., K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Stat.Med.,* vol. 15, no. 4. pp.361-387, Feb., 1996.

[94] E. W. Steyerberg, M. J. Eijkemans, F. E. Harrell, Jr. et al., "Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets," *Med.Decis.Making,* vol. 21, no. 1. pp.45-56, Jan., 2001.

[95] M. R. Segal, K. D. Dahlquist, and B. R. Conklin, "Regression approaches for microarray data analysis," *Journal of Computational Biology,* vol. 10, no. 6. pp.961-980, 2003.

[96] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society Series B-Methodological,* vol. 58, no. 1. pp.267-288, 1996.

[97] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B-Statistical Methodology,* vol. 67. pp.301-320, 2005.

[98] S. Cho, H. Kim, S. Oh et al., "Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis," *BMC.Proc.,* vol. 3 Suppl 7. pp.S25, 2009.

[99] H. Zou and H. H. Zhang, "ON THE ADAPTIVE ELASTIC-NET WITH A DIVERGING NUMBER OF PARAMETERS," *Ann.Stat.,* vol. 37, no. 4. pp.1733-1751, 2009.

[100] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association,* vol. 53, no. 282. pp.457-481, June, 1958.

[101] G. R. Norman and D. L. Streiner, *Biostatistics: The Bare Essentials (2nd Edition)*: B.C. Decker Inc., 2000.

[102] Mithat Gönen and Glenn Heller, "Concordance probability and discriminatory power in proportional hazards regression," *Biometrika,* vol. 92, no. 4. pp.965-970, 2005.

[103] D. Ucar, I. Neuhaus, P. Ross-MacDonald et al., "Construction of a reference gene association network from multiple profiling data: application to data analysis," *Bioinformatics,* vol. 23, no. 20. pp.2716-2724, Oct., 2007.

[104] L. J. Jensen, M. Kuhn, M. Stark et al., "STRING 8--a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Res,* vol. 37, no. Database issue. pp.D412-D416, Jan., 2009.

[105] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.,* vol. 28, no. 1. pp.27-30, Jan., 2000.

[106] M. Kanehisa, S. Goto, M. Hattori et al., "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Res,* vol. 34, no. Database issue. pp.D354-D357, Jan., 2006.

[107] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, no. 43. pp.15545-15550, Oct., 2005.

[108] A. Spira, J. E. Beane, V. Shah et al., "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer," *Nat Med.,* vol. 13, no. 3. pp.361-366, Mar., 2007.

[109] L. Guo, Y. Ma, R. Ward et al., "Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma," *Clin.Cancer Res.,* vol. 12, no. 11. pp.3344-3354, June, 2006.

[110] S. K. Lau, P. C. Boutros, M. Pintilie et al., "Three-gene prognostic classifier for early-stage non small-cell lung cancer," *J.Clin.Oncol.,* vol. 25, no. 35. pp.5562-5569, Dec., 2007.

[111] U. T. Shankavaram, W. C. Reinhold, S. Nishizuka et al., "Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study," *Mol.Cancer Ther.,* vol. 6, no. 3. pp.820-832, Mar., 2007.

[112] Z. Wu, R. A. Irizarry, R. Gentleman et al., "A Model-Based Background Adjustment for Oligonucleotide Expression Arrays," *Journal of the American Statistical Association,* vol. 99. pp.909, 2004.

[113] Y. Ma, Z. Ding, Y. Qian et al., "An Integrative Genomic and Proteomic Approach to Chemosensitivity Prediction," *Int.J Oncol.*, no. 34. pp.107-115, 2009.

[114] Y. Ma, Z. Ding, Y. Qian et al., "Predicting cancer drug response by proteomic profiling," *Clin.Cancer Res.,* vol. 12, no. 15. pp.4583-4589, Aug., 2006.

[115] P. Csermely, V. Agoston, and S. Pongor, "The efficiency of multi-target drugs: the network approach might help drug design," *Trends Pharmacol.Sci.,* vol. 26, no. 4. pp.178-182, Apr., 2005.

[116] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition)*: Morgan Kaufmann, 2005.

[117] D. M. Burns and J. D. Richter, "CPEB regulation of human cellular senescence, energy metabolism, and p53 mRNA translation
1," *Genes Dev.,* vol. 22, no. 24. pp.3449-3460, Dec., 2008.

[118] Y. W. Wan, E. Sabbagh, R. Raese et al., "Hybrid models identified a 12-gene signature for lung cancer prognosis and chemoresponse prediction," *PLoS ONE,* vol. 5, no. 8, 2010.

[119] N. L. Guo, Y. W. Wan, S. Bose et al., "A novel network model identified a 13-gene lung cancer prognostic signature," *Int.J.Comput.Biol.Drug Des,* vol. 4, no. 1. pp.19-39, 2011.

[120] Y. W. Wan, S. Bose, J. Denvir et al., "A Novel Network Model for Molecular Prognosis," *ACM Intl.Conference on Bioinformatics and Computational Biology (ACM-BCB 2010)*. pp.342-345, 2010.

[121] P. P. Massion, Y. Zou, H. Chen et al., "Smoking-related genomic signatures in non-small cell lung cancer," *Am.J.Respir.Crit Care Med.,* vol. 178, no. 11. pp.1164-1172, Dec., 2008.

[122] M. Woenckhaus, L. Klein-Hitpass, U. Grepmeier et al., "Smoking and cancer-related gene expression in bronchial epithelium and non-small-cell lung cancers," *J.Pathol.,* vol. 210, no. 2. pp.192-204, Oct., 2006.

[123] C. Xiao, "Evaluation of a smoking-associated gene signature for lung cancer,", West Virginia University,pp.44-46, 2010.