

# Pathway-based identification of a smoking associated 6-gene signature predictive of lung cancer risk and survival

Nancy Lan Guo<sup>a, b\*</sup> and Ying-Wooi Wan<sup>b, c</sup>

<sup>a</sup>Department of Community Medicine, <sup>b</sup>Mary Babb Randolph Cancer Center, and <sup>c</sup>Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, West Virginia 26506, United States of America

## \*Corresponding author:

Nancy L. Guo

Tel: +1-304-2936455

Fax: +1-304-2934667

Email: [lguo@hsc.wvu.edu](mailto:lguo@hsc.wvu.edu)

## Correspondence address:

Nancy L. Guo

2816 HSS

Mary Babb Randolph Cancer Center

Morgantown, WV 26506-9300, USA

## Co-author address:

Ying-Wooi Wan

2833 HSS

Mary Babb Randolph Cancer Center

Morgantown, WV 26506-9300, USA

**Key words:** implication networks based on prediction logic, gene coexpression networks based on formal logic, smoking, gene signature, lung cancer diagnosis and prognosis, signaling pathways

## **Abstract**

**Objective:** Smoking is a prominent risk factor for lung cancer. However, it is not an established prognostic factor for lung cancer in clinics. To date, no gene test is available for diagnostic screening of lung cancer risk or prognostication of clinical outcome in smokers. This study sought to identify a smoking associated gene signature in order to provide a more precise diagnosis and prognosis of lung cancer in smokers.

**Methods and materials:** An implication network based methodology was used to identify biomarkers by modeling crosstalk with major lung cancer signaling pathways. Specifically, the methodology contains the following steps: 1) identifying genes significantly associated with lung cancer survival; 2) selecting candidate genes which are differentially expressed in smokers versus non-smokers from the survival genes identified in Step 1; 3) from these candidate genes, constructing gene coexpression networks based on prediction logic for the smoker group and the non-smoker group, respectively; 4) identifying smoking-mediated differential components, i.e., the unique gene coexpression patterns specific to each group; and 5) from the differential components, identifying genes directly co-expressed with major lung cancer signaling hallmarks.

**Results:** A smoking-associated 6-gene signature was identified for prognosis of lung cancer from a training cohort ( $n=256$ ). The 6-gene signature could separate lung cancer patients into two risk groups with distinct post-operative survival (log-rank  $P < 0.04$ , Kaplan-Meier analyses) in three independent cohorts ( $n=427$ ). The expression-defined prognostic prediction is strongly related to smoking association and smoking cessation ( $P < 0.02$ ; Pearson's Chi-squared tests). The 6-gene signature is an accurate prognostic factor (hazard ratio = 1.89, 95% CI: [1.04, 3.43]) compared to common clinical covariates in multivariate Cox analysis. The 6-gene signature also provides an accurate diagnosis of lung cancer with an overall accuracy of 73% in a cohort of smokers ( $n=164$ ). The coexpression patterns derived from the implication networks were validated with interactions reported in the literature retrieved with STRING8, Ingenuity Pathway Analysis, and Pathway Studio.

**Conclusions:** The pathway-based approach identified a smoking-associated 6-gene signature that predicts lung cancer risk and survival. This gene signature has potential clinical implications in the diagnosis and prognosis of lung cancer in smokers.

## 1. Introduction

Lung cancer remains the leading cause of cancer deaths in the United States [1]. Non-small cell lung cancer (NSCLC) accounts for about 80% of lung cancer cases. Two major subtypes of NSCLC are lung adenocarcinoma and squamous cell lung cancer. Smoking is a strong risk factor in lung cancer development and is responsible for about 90% of lung cancer cases [2-4]. Our previous study showed that smoking intensity at the time of diagnosis is a significant and independent prognostic factor of lung cancer[5]. Nevertheless, smoking is not an established lung cancer prognostic determinant in clinical practice, and its mechanistic effect on 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 by analyzing genome-wide transcriptional profiles of lung cancer patient samples.

Traditional approaches to biomarker discovery rank genes based on their association with the clinical outcome and select the top-ranked genes as signature genes [6-8]. However, these approaches do not account for the interactions among genes. It is known that genes function through a series of interactions with one another, and disease is one possible result of these interactions. Recent studies indicate that molecular network analyses could be used to improve disease classification [9-11], identify disease genes [12], discover novel therapeutic targets [13,14], and reveal disease related sub-networks [15].

Boolean networks have been used to gain insights into gene regulation functions [16-19]. The Boolean implication networks presented by Sahoo et al. [20,21] used scatter plots of expression between two genes to derive the implication relations. Their study did not use Boolean implication networks as a gene selection system. We developed an induction algorithm based on prediction logic [22] to derive implication relations. In our previous studies, implication networks were employed to model disease-mediated genome-wide coexpression networks for the identification of prognostic gene signatures [23,24]. In this study, implication networks were used to infer the relevance of signaling pathways in a set of selected genes associated with smoking and lung cancer survival.

Genes implicated in cancer initiation and progression show dysregulated interactions with their molecular partners [25], and cancer genes are more likely to actively interact with signaling proteins [26]. We hypothesized that an analysis of genes associated with smoking and major lung cancer signaling pathways could lead to the identification of a gene signature that provides a

more accurate diagnosis and prognosis of lung cancer. The following steps were carried out to test the hypothesis: 1) Genes that were significantly associated with lung cancer survival were identified from genome-wide expression profiles using the training set ( $n=256$ ). 2) Genes with differential expression in smokers versus non-smokers were then selected for further analysis. 3) The implication network algorithm was employed to construct smoking mediated gene coexpression networks. 4) By comparing the coexpression patterns from smoking mediated gene coexpression networks, the unique coexpression patterns that are specific to each group are identified as the smoking-mediated differential components. 5) From the differential components, genes that had common coexpression relations with *MET*, *EGF*, *KRAS*, *TP53*, *E2F1*, and *E2F4* were pinpointed. The identified signature was then validated for prognostic ( $n=427$ ) and diagnostic ( $n=164$ ) prediction in 5 independent patient cohorts. The prognostic performance of the identified gene signature was also evaluated by comparing it with clinical covariates. Furthermore, the smoking-mediated gene coexpression patterns were confirmed with curated interactions published in the literature.

## **2. Materials and Methods**

### **2.1. Implication induction algorithm for pair-wise coexpression network construction**

An implication network is a directed graph with variables as nodes, and adjacent nodes are connected with arch representing implications. The first induction algorithm for an implication network was proposed by Liu et al. [27,28] based on binomial distribution, which is suitable for binary datasets. An alternative network induction algorithm was proposed by Guo et al. [22] based on prediction logic [29], which is applicable for more general applications, including multinomial datasets and multi-classification problems. Prediction logic reveals the implication relationships among variables in a dataset and evaluates propositions in formal logic by integrating formal logic theory and statistics. The most important aspect of prediction logic is the conceptual value of prediction analysis in constructing and evaluating useful statements, particularly in complex multinomial problems with moderate sample sizes. This feature is vital for clinical applications, in which many clinical parameters are multinomial and the patient sample size is small.

We used prediction logic based on formal logic rules relating two dichotomous variables to induce the implication network. The six most important implication rules relating two dichotomous variables are shown in Fig. 1, where each table is a contingency table and the shaded cells 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. In the biological context,  $A \Rightarrow B$ : upregulation of gene  $A$  causes upregulation of gene  $B$ ;  $A \Rightarrow \neg B$ : upregulation of gene  $A$  causes downregulation of gene  $B$ ;  $\neg A \Rightarrow B$ : down-regulation of gene  $A$  causes upregulation of gene  $B$ ;  $\neg A \Rightarrow \neg B$ : down-regulation of gene  $A$  causes down-regulation of gene  $B$ ;  $A \Leftrightarrow B$ : upregulation of gene  $A$  causes upregulation of gene  $B$ ; and upregulation of gene  $B$  causes upregulation of gene  $A$ ;  $A \Leftrightarrow \neg B$ : upregulation of gene  $A$  causes down-regulation of gene  $B$ ; and down-regulation of gene  $B$  causes upregulation of gene  $A$ .

A modified  $U$ -Optimality method [29] (Fig. 2) was used to derive the implication relation between each pair of variables in the dataset. In the algorithm,  $U_p$  is the scope of the implication rule, representing the portion of the data covered by the implication relation, and  $\nabla_p$  is the precision of the implication rule, representing the prediction success of the corresponding implication relation. An implication rule has high precision when the number of error occurrences is a small portion of the data covered by the implication rule. The minimum scope and precision required by the implication rule are indicated respectively by  $U_{min}$  and  $\nabla_{min}$ , which must be positive for a valid implication relation. The induction algorithm derives an implication rule if it has the maximum scope,  $U_p$  and it satisfies the constraint that its scope,  $U_p$  and precision,  $\nabla_p$  are greater than the required minimum values,  $U_{min}$  and  $\nabla_{min}$ . To simplify the computations of the maximization problem, the  $\nabla_{ij}$  value of every error cell must be greater than that of the non-error cells for the corresponding implication rule [22].

For a single error cell, where  $N_{ij}$  is the number of error occurrences, the scope,  $U_p$  and precision,  $\nabla_p$  are defined as:

$$U_p = U_{ij} = \frac{N_{i.} * N_{.j}}{N^2}, \quad \nabla_p = \nabla_{ij} = 1 - \frac{N_{ij}}{N * U_p}.$$

For multiple error cells, they are defined as:

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

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

This implication induction algorithm is general for discrete datasets. With the expansion of the contingency table  $M_{ij}$  (Fig. 2), implication rules can be induced for multinomial datasets, where error cells are those with top precision ( $\nabla_{ij}$ ) values and satisfying all the constraints. The proposition can then be induced according to the error set.

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) [22]. The difference between this algorithm and that of Hildebrand et al. [29] is that minimum requirements for deriving an implication rule were set for both scope ( $U_p$ ) and precision ( $\nabla_p$ ), instead of for precision alone.

## 2.2. Microarray profiles and patient samples

Four sets of published microarray gene expression profiles were used in this study. The first set contains 442 lung adenocarcinoma patient samples in the Director's Challenge Study [30]. The second set contains 130 adenocarcinoma and squamous cell lung cancer samples published by Raponi et al. [8]. The third set contains 111 NSCLC samples published by Bild et al. [31]. The fourth set contains 164 airway epithelial cells from current and former smokers published by Spira et al. [2]. Gene expression profiles from these studies were quantified with Affymetrix HG-U133A, except for the set from Bild et al. [31], which was quantified with Affymetrix HG-U133 Plus 2. The data used in the analyses was quantile-normalized and  $\log_2$  transformed with dChip [32].

## 3. Results and Discussion

### 3.1. Identification of a smoking-associated gene signature for prognosis in lung adenocarcinoma

In this study, the UM and HLM cohorts from the Director's Challenge Study [30] formed the training set ( $n = 256$ ), whereas MSK and DFCI cohorts formed the test set ( $n = 186$ ). Genes with missing values in at least half of the samples were removed, which left 19,866 genes for the analysis.

Genes associated with lung cancer survival were first selected from the entire genome. A total of 2,310 genes were significantly associated with overall survival ( $P < 0.05$ , univariate Cox

model) in the training data. Second, from this set of 2,310 genes, 217 genes showed significant differential expression ( $P < 0.05$ , unpaired  $t$ -tests) in smokers versus non-smokers in the training data. These 217 survival and smoking-associated genes as well as 6 major signaling proteins, including *EGF*, *TP53*, *MET*, *KRAS*, *E2F1*, and *E2F4*, were included in the network analysis. These signaling pathways are included in human NSCLC disease mechanisms delineated by the KEGG Pathway Database<sup>1</sup>. *KRAS* is involved in many signaling transduction pathways and its mutation is related to many human cancer types. *TP53* regulates cell cycle and functions as a tumor suppressor gene. *EGF* is a growth factor and regulates cell growth, proliferation and differentiation by binding to its receptor *EGFR*. *MET* is an oncogene and plays an important role in embryonic development and wound healing. *E2F1* and *E2F4* are members of the *E2F* family of transcription factors. The *E2F* family is essential for the control of cell cycle and action of tumor suppressor proteins. These signaling proteins are selected based on their reported clinical relevance in non-small cell lung cancer. Because tumors utilize different signaling pathways, we hypothesize that including a diverse set of pathways would perform more uniformly across heterogeneous tumor sets. These 6 hallmarks were not significantly associated with survival nor differentially expressed in smokers.

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, a coexpression network among the 223 genes was constructed using the implication induction algorithm. Between each pair of the 223 genes, possible significant ( $P < 0.05$ ;  $z$ -tests) coexpression relations were derived in the smoker group and the non-smoker group, respectively, constituting smoking-mediated gene coexpression networks for lung cancer. By comparing the implication rules between each pair of nodes in the two networks, differential network components were identified. These differential components are implication

---

<sup>1</sup> <http://www.genome.jp/kegg/pathway/hsa/hsa05223.html>

relations (co-expressions) 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 the smoker group and the non-smoker group, genes having direct co-expressions with the 6 lung cancer hallmarks were identified (detailed gene list provided in Supplementary File). From the non-smoker group, certain genes had direct coexpression with some of the 6 hallmarks but no gene had direct coexpression with all the 6 lung cancer hallmarks. From the smoker group, 6 genes were identified having direct coexpression with all the 6 lung cancer hallmarks. This constituted the smoking-associated 6-gene signature for lung cancer prognosis (Table 1).

### **3.2. Prognostic evaluation of the 6-gene signature in lung adenocarcinoma**

We sought to investigate if the gene signature identified could provide accurate prognostication of survival in NSCLC patients. The 6 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 6 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. 3A), produced the best patient stratification in the training set (Fig. 3B). Therefore, the training model and this cutoff value were applied to the test set without re-estimating the parameters (Fig. 3C). In both training and test sets, this classification scheme generated significant patient stratifications (log-rank  $P < 0.03$ , Kaplan-Meier analyses).

To evaluate the statistical significance of the 6-gene signature identified from the proposed network analysis, a random set of 6 genes from the 217 survival and smoking-associated genes were selected and constructed as a classifier using the same approach with the Cox proportional hazard model. Results showed that the identified signature gave significantly ( $P < 0.05$ ) better lung cancer prognosis compared with 1000 random signatures.

### **3.3. 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 generated significant prognostic stratifications in smokers from both training and test cohorts (log-rank  $P < 0.04$ , Kaplan-Meier analysis) (Fig. 4), but not in non-smokers (log-rank  $P < 0.83$ , Kaplan-Meier analyses, results not shown). In addition, gene expression-defined high- and low-risk groups showed significant association with smoking ( $P < 0.00008$ , Chi-square tests) and smoking cessation ( $P < 0.02$ , Chi-square tests) (Table 2). Specifically, smokers were significantly associated with the high-risk group compared with non-smokers, and current smokers had a stronger association with the high-risk group compared with former smokers (Table 2).

### **3.4. Prognostic validation on squamous cell lung cancer**

The prognostic performance of the 6-gene signature was further evaluated on the Raponi [8] and Bild [31] cohorts which include squamous cell lung cancer. For a rigorous evaluation, patient samples in the studied cohorts were randomly partitioned into separate training and test sets. A prognostic classifier was constructed on the training set using a Cox proportional hazard model and validated on the test set without re-estimating parameters. In the training set from both cohorts, the cutoff value that gave the shortest distance to the point of perfect prediction of the 3-year ROC curve produced the best patient stratification. In both training and test sets, the 6-gene signature stratified patients into two prognostic groups with distinct survival (log-rank  $P < 0.04$ ; Fig. 5). These results indicate that the identified smoking-associated 6-gene signature could be used for prognosis for NSCLC patients.

### **3.5. Prognosis evaluation with clinical covariates**

To validate the prognostic power of the identified 6-gene signature, the constructed expression-defined prognostic model was evaluated with common lung cancer prognostic factors, including gender, age, cancer stage, and tumor differentiation in smokers in the test cohort. The prognostic outcome predicted by the gene expression model was used as a covariate in the multivariate Cox analysis.

Results from the multivariate Cox proportional analysis showed that cancer 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 model (Table 3). When the 6-gene prognostic model was added to the multivariate Cox model, the gene model demonstrated a strong association with the risk of lung cancer death (hazard ratio = 1.89, 95% CI: [1.04, 3.43]), and cancer stage remained significant (Table 3). The hazard ratio of the 6-gene prognostic model was higher than other cancer prognostic factors except for cancer stage, with no significant difference between cancer stage and the gene model. The results demonstrate that the 6-gene prognostic model is a more significant prognostic factor than some commonly used clinical parameters.

### **3.6. Early detection of lung cancer**

We further evaluated whether the 6-gene signature could be used for lung cancer diagnosis in smokers. The smoking cohort from Spira et al. [2] was randomly separated into a training set ( $n = 77$ ) and two independent test sets ( $n = 52$  and  $n = 35$ ). With the Naïve Bayes classification algorithm implemented in software package WEKA [33], the classifier could accurately identify lung cancer patients from normal patients with an overall accuracy of 73% and 69% in the two test sets (Table 4). Furthermore, the classifier's performance was significantly ( $P < 0.005$ ) better than that of random signatures with the same size using the same classifier in 1000 tests, on the same training and test sets. These results indicate that the 6-gene signature could be potentially used in diagnostic screening of lung cancer risk in smokers.

### **3.7. Confirmation of smoking-mediated gene coexpression relations**

The coexpression relations derived by the implication network were also evaluated. Differential network components among the signature genes and the 6 signaling hallmarks present in both training and test sets were retrieved to represent smoking-mediated gene coexpression patterns in lung cancer patients. There were 9 common coexpression relations specifically associated with smokers (Fig. 6A), and 3 specifically associated with non-smokers (Fig. 6B) in both training and testing cohorts.

The biological relevance of the derived coexpression relations was confirmed by retrieving curated interactions related to these genes using bioinformatics tools in Ingenuity Pathway Analysis<sup>2</sup> (IPA, Ingenuity Systems<sup>®</sup>), Pathway Studio, and the signaling pathway

---

<sup>2</sup> <http://www.ingenuity.com/>

database STRING8. Among 12 coexpression relations derived from the implication networks, 1 interaction specific to smokers and 1 interaction specific to non-smokers were confirmed (Fig. 6).

The stability of the smoking mediated coexpression networks (Fig. 6A and 6B) was evaluated with different subsets of patient samples from the training set in 100 iterations (Fig. 6D). The stability is defined as the portion of smoking-mediated 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. Results show that the implication network algorithm is stable as most of the coexpression relations (about 60%) could be derived using as few as half of the training samples (Fig. 6D).

In addition, we also evaluated the precision and false discovery rate (FDR) of the coexpression relations derived in the smoking-mediated coexpression networks (Fig 6A and 6B). Five gene set collections (positional, curated, motif, computational, and Gene Oncology) and canonical pathway databases from the MSigDB<sup>3</sup> were used to evaluate the biological relevance of computationally derived coexpression relations. 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 [34]. Coexpression relations labeled as ND were excluded in the evaluation as they were not confirmed. With precision defined as  $TP/(TP + FP)$ , the precision of the smoking-mediated coexpression networks (Fig. 6A and 6B) was 100% (7 relations were labeled as TP and no relation was labeled as FP; Fig. 6C). Null distribution was generated in 1,000 random permutations of the class labels in the test cohort. The precision of the smoking-mediated coexpression networks is significant at  $P < 0.001$ , with no TP generated in the random tests. With FDR defined as the average of  $FP/(TP+FP)$  in 1,000 permutations, the FDR of the smoking-mediated coexpression networks is 0.0099. These results indicate that implication networks can reveal biologically relevant gene associations.

### 3.8 Comparison with gene association networks based on Pearson's correlation

---

<sup>3</sup> <http://www.broadinstitute.org/gsea/msigdb/collections.jsp>

Large-scale gene coexpression networks have been used in biomarker discovery and disease classification, based on the observation that functionally related genes are frequently coexpressed across multiple datasets and different organisms [35-37]. These studies construct pair-wise gene coexpression networks by using correlation coefficients computed from gene expression profiles. Such networks indicate the *distance* or *similarity* between each pair of gene expression profiles but do not provide the *direction* or *causal* relations in the gene regulatory patterns. A new algorithm is needed to efficiently construct genome-scale coexpression networks and provide a convenient predictive structure of gene regulation. Prediction logic provides a convenient and more predictive structure association than correlation coefficients [29]. Boolean implications networks constructed with a similar algorithm have been used to infer gene regulations [20,21].

For comparison with implication networks, we used Pearson's correlation coefficient to construct gene association networks for smoker and non-smoker groups in both training and testing sets using the same methodology (Figure 7). The implications networks derived more gene association rules than the networks based on Pearson's correlation coefficients. We then evaluated the precision and FDR of the interactions specific to smoker and non-smoker groups that were present in both training and testing sets (Figure 7C). Both networks have the same precision of 0.96 and FDR of 0.04 in the evaluation with MSigDB. These results indicate that implication networks could retrieve more biologically relevant gene associations without any loss of precision and increase of FDR when compared with gene association networks based on Pearson's correlation coefficients. Furthermore, we examined the smoking-specific and non-smoking-specific gene association networks based on Pearson's correlation coefficients in the training set. No gene was coexpressed with all 6 lung cancer hallmarks based on the Pearson's correlation. In other words, using gene association networks based on Pearson's correlation coefficients, we would not be able to identify any gene with concurrent coexpression with the 6 signaling pathways using the proposed methodology. Together, these results demonstrate the advantage of implication networks based on prediction logic in biomarker discovery.

#### **4. Conclusions**

This study presents an implication network-based approach to the identification of a smoking-associated 6-gene signature that was co-expressed with major NSCLC signaling pathways. The identified 6-gene signature could accurately estimate disease-specific survival in NSCLC patients and could potentially be used for screening of lung cancer risk in smokers. The gene expression-defined prognostication also showed strong association with smoking and smoking cessation. This gene signature is a more accurate prognostic factor than some commonly used clinical parameters such as age, gender, and tumor differentiation, and is comparable with cancer stage in terms of hazard ratio. Some of the computationally derived coexpression patterns have been experimentally verified in previous studies.

Our previous studies have demonstrated that implication network-based methodology is efficient in modeling disease-mediated genome-scale coexpression networks for biomarker identification [23,24]. In this study, the methodology was applied to a more focused set of genes related to smoking and lung cancer survival. The results from this study demonstrate that combined analysis of smoking mediated coexpression networks and crosstalk with lung cancer signaling pathways could identify important biomarkers and elucidate mechanistic and possibly synergistic processes underlying oncogenesis and metastasis in lung cancer. The gene coexpression relations derived with implication networks have been validated with biological experiments (results not shown).

### **Acknowledgements**

We are grateful for Rebecca Raese for editing the manuscript. We thank Changchang Xiao for processing the microarray data. This project is supported by NIH R01LM009500 (PI: Guo) and NIH/NCRR P20RR16440 and Supplement (PD: Guo). Software license and training for Ingenuity Pathway Analysis and Pathway Studio is supported by NIH/NCRR P2016477.

### **References**

- [1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M.J. Thun, Cancer statistics, 2009, *CA Cancer J. Clin.*, 59 (2009) pp. 225-249.
- [2] A. Spira, J.E. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri et al, Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer, *Nat Med.*, 13 (2007) pp. 361-366.

- [3] P.P. Massion, Y. Zou, H. Chen, A. Jiang, P. Coulson, C.I. Amos et al, Smoking-related genomic signatures in non-small cell lung cancer, *Am. J. Respir. Crit Care Med.*, 178 (2008) pp. 1164-1172.
- [4] M. Woenckhaus, L. Klein-Hitpass, U. Grepmeier, J. Merk, M. Pfeifer, P. Wild et al, Smoking and cancer-related gene expression in bronchial epithelium and non-small-cell lung cancers, *J. Pathol.*, 210 (2006) pp. 192-204.
- [5] N.L. Guo, K. Tosun, and K. Horn, Impact and interactions between smoking and traditional prognostic factors in lung cancer progression, *Lung Cancer*, 66 (2009) pp. 386-392.
- [6] D.G. Beer, S.L. Kardia, C.C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek et al, Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nat. Med.*, 8 (2002) pp. 816-824.
- [7] H.Y. Chen, S.L. Yu, C.H. Chen, G.C. Chang, C.Y. Chen, A. Yuan et al, A five-gene signature and clinical outcome in non-small-cell lung cancer, *N. Engl. J. Med.*, 356 (2007) pp. 11-20.
- [8] M. Raponi, Y. Zhang, J. Yu, G. Chen, G. Lee, J.M. Taylor et al, Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung, *Cancer Res.*, 66 (2006) pp. 7466-7472.
- [9] H.Y. Chuang, E. Lee, Y.T. Liu, D. Lee, and T. Ideker, Network-based classification of breast cancer metastasis, *Mol. Syst. Biol.*, 3 (2007) pp. 140.
- [10] F.J. Muller, L.C. Laurent, D. Kostka, I. Ulitsky, R. Williams, C. Lu et al, Regulatory networks define phenotypic classes of human stem cell lines, *Nature*, 455 (2008) pp. 401-405.
- [11] I.W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria et al, Dynamic modularity in protein interaction networks predicts breast cancer outcome, *Nat Biotechnol.*, 27 (2009) pp. 199-204.
- [12] V. Emilsson, G. Thorleifsson, B. Zhang, A.S. Leonardson, F. Zink, J. Zhu et al, Genetics of gene expression and its effect on disease, *Nature*, 452 (2008) pp. 423-428.
- [13] P. Csermely, V. Agoston, and S. Pongor, The efficiency of multi-target drugs: the network approach might help drug design, *Trends Pharmacol. Sci.*, 26 (2005) pp. 178-182.
- [14] M.A. Yildirim, K.I. Goh, M.E. Cusick, A.L. Barabasi, and M. Vidal, Drug-target network, *Nat. Biotechnol.*, 25 (2007) pp. 1119-1126.
- [15] S.E. Calvano, W. Xiao, D.R. Richards, R.M. Felciano, H.V. Baker, R.J. Cho et al, A network-based analysis of systemic inflammation in humans, *Nature*, 437 (2005) pp. 1032-1037.
- [16] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung et al, A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, 302 (2003) pp. 449-453.
- [17] D. Sahoo, D.L. Dill, A.J. Gentles, R. Tibshirani, and S.K. Plevritis, Boolean implication networks derived from large scale, whole genome microarray datasets, *Genome Biol.*, 9 (2008) pp. R157.
- [18] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan, Causal protein-signaling networks derived from multiparameter single-cell data, *Science*, 308 (2005) pp. 523-529.
- [19] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Network motifs: simple building blocks of complex networks, *Science*, 298 (2002) pp. 824-827.

- [20] D. Sahoo, D.L. Dill, A.J. Gentles, R. Tibshirani, and S.K. Plevritis, Boolean implication networks derived from large scale, whole genome microarray datasets, *Genome Biol.*, 9 (2008) pp. R157.
- [21] D. Sahoo, J. Seita, D. Bhattacharya, M.A. Inlay, I.L. Weissman, S.K. Plevritis et al, MiDReG: a method of mining developmentally regulated genes using Boolean implications, *Proc. Natl. Acad. Sci. U. S. A.*, 107 (2010) pp. 5732-5737.
- [22] L. Guo, B. Cukic, and H. Singh, Predicting Fault Prone Modules by the Dempster-Shafer Belief Networks, *Proceedings of 18th IEEE International Conference on Automated Software Engineering (ASE'03)*, (2003) pp. 249-252.
- [23] N.L. Guo, Y.W. Wan, S. Bose, J. Denvir, M.L. Kashon, and M.E. Andrew, A novel network model identified a 13-gene lung cancer prognostic signature, *Int. J. Comput. Biol. Drug Des.*, 4 (2011) pp. 19-39.
- [24] Y.W. Wan, S. Bose, J. Denvir, and N.L. Guo, A Novel Network Model for Molecular Prognosis, *Proc. ACM International Conference on Bioinformatics and Computational Biology*, (2010) pp. 342-345.
- [25] K.M. Mani, C. Lefebvre, K. Wang, W.K. Lim, K. Basso, R. la-Favera et al, A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas, *Mol. Syst. Biol.*, 4 (2008) pp. 169.
- [26] Q. Cui, Y. Ma, M. Jaramillo, H. Bari, A. Awan, S. Yang et al, A map of human cancer signaling, *Mol. Syst. Biol.*, 3 (2007) pp. 152.
- [27] 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*, 9 (1997) pp. 990-1004.
- [28] J. Liu, D. Maluf, and M.C. Desmarais, A New Uncertainty Measure for Belief Networks with Applications to Optimal Evidential Inferencing, *IEEE Transactions on Knowledge and Data Engineering*, 13 (2001) pp. 416-425.
- [29] D.K. Hildebrand, J.D. Laing, and H. Rosenthal, *Prediction Analysis of Cross Classifications*, (John Wiley & Sons, 1977).
- [30] K. Shedden, J.M. Taylor, S.A. Enkemann, M.S. Tsao, T.J. Yeatman, W.L. Gerald et al, Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study, *Nat. Med.*, 14 (2008) pp. 822-827.
- [31] A.H. Bild, G. Yao, J.T. Chang, Q. Wang, A. Potti, D. Chasse et al, Oncogenic pathway signatures in human cancers as a guide to targeted therapies, *Nature*, 439 (2006) pp. 353-357.
- [32] C. Li, Automating dChip: toward reproducible sharing of microarray data analysis, *BMC. Bioinformatics.*, 9 (2008) pp. 231.
- [33] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition)*, (Morgan Kaufmann, 2005).
- [34] D. Ucar, I. Neuhaus, P. Ross-MacDonald, C. Tilford, S. Parthasarathy, N. Siemers et al, Construction of a reference gene association network from multiple profiling data: application to data analysis, *Bioinformatics*, 23 (2007) pp. 2716-2724.
- [35] J.K. Choi, U. Yu, O.J. Yoo, and S. Kim, Differential coexpression analysis using microarray data and its application to human cancer, *Bioinformatics.*, 21 (2005) pp. 4348-4355.

- [36] L.L. Elo, H. Jarvenpaa, M. Oresic, R. Lahesmaa, and T. Aittokallio, Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process, *Bioinformatics.*, (2007).
- [37] C.C. Liu, W.S. Chen, C.C. Lin, H.C. Liu, H.Y. Chen, P.C. Yang et al, Topology-based cancer classification and related pathway mining using microarray data, *Nucleic Acids Res*, 34 (2006) pp. 4069-4080.

## Tables

**Table 1. The identified smoking associated 6-gene signature.**

<b>Probe</b>	<b>Gene symbol</b>	<b>Gene title</b>	<b>Molecular function (GO)</b>
200705_s_at	EEF1B2	Eukaryotic translation elongation factor 1 beta 2	Translation elongation factor activity; protein binding
203788_s_at	SEMA3C	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	Receptor activity; semaphorin receptor binding
206183_s_at	HERC3	HECT domain and RCC1-like domain-containing protein 3	Ligase activity; acid-amino acid ligase activity
209230_s_at	NUPR1	Nuclear protein, transcriptional regulator, 1	N/A
210669_at	TFAP2A	Transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha)	DNA and protein binding; transcription factor activity; transcription coactivator activity; protein dimerization activity
213456_at	SOSTDC1	Sclerostin domain containing 1	Protein binding

**Table 2. Association between smoking status and gene expression-defined prognostic risk groups.**

	<b>Low-risk</b>	<b>High-risk</b>	<b>Chi-square tests</b>
<b>Smoker</b>	138	162	<b>Smoking association</b> $\chi^2 = 15.53$ ( $P = 8.10e-5$ )
<b>Non-smoker</b>	38	11	
<b>Current smoker</b>	8	24	<b>Smoking cessation</b> $\chi^2 = 5.45$ ( $P = 0.02$ )
<b>Former smoker</b>	130	138	

**Table 3. Multivariate Cox analyses of the gene expression-defined prognostication and major clinical covariates in smoking lung cancer patients in the test cohort.**

<b>Variable*</b>	<b>P-value</b>	<b>Hazard ratio (95% CI)<sup>ψ</sup></b>	
<i>Analysis without 6-gene prognostic prediction</i>			
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)
<i>Analysis with 6-gene prognostic prediction</i>			
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)
<b>6-gene prognostic prediction</b>	<b>0.04</b>	<b>1.89</b>	<b>(1.04, 3.43)</b>

\* 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).  
<sup>ψ</sup> denotes confidence interval.

**Table 4. Prediction of lung cancer risk in smokers using the 6-gene signature with the Naïve Bayes algorithm.**

	<b>Sensitivity (lung cancer)</b>	<b>Specificity (normal)</b>	<b>Overall accuracy*</b>
<b>Training (10-fold CV)</b>	71% (25/35)	62% (26/42)	66% (51/77)
<b>Test 1</b>	76% (19/25)	65% (19/27)	73% (38/52)
<b>Test 2</b>	72% (13/18)	65% (11/17)	69% (24/35)

\*The 6-gene signature gave significantly ( $P<0.005$ ) accurate performance in all three data sets when compared with 1000 random sets of 6 genes using the same algorithm.

## Figure captions

**Figure 1. Six implication rules relating two dichotomous variables.**

**Figure 2. Implication induction algorithm for constructing coexpression networks.**

**Figure 3. Prognostication of survival using the 6-gene signature in lung adenocarcinoma.**

On the training cohort from the Director's Challenge Study (Shedden, 2008), 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 ( $P < 0.03$ ) patient stratification in the training (B) and test (C) cohorts in Kaplan-Meier analyses. Log-rank tests were used to assess the significance of difference between survival probability of the two prognostic groups

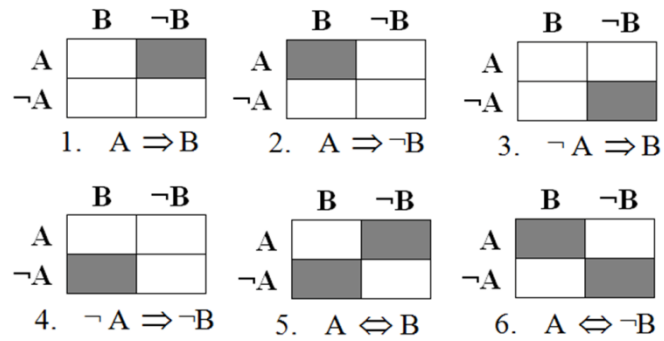
**Figure 4. Survival prediction in smoking lung cancer patients by the 6-gene signature.** The prognostic classifier stratified smoking patients into two prognostic groups with significantly distinct survival ( $P < 0.04$ ) in both the training (A) and test (B) cohorts in the Director's Challenge Study.

**Figure 5. Prognostic performance of the smoking-associated signature on other histological subtypes of NSCLC.** In Kaplan-Meier analyses, significant ( $P < 0.04$ ) stratifications were obtained in the randomly partitioned training and test cohorts of patients with squamous cell lung cancer (A, B) and patients with lung adenocarcinoma or squamous cell lung cancer (C, D).

**Figure 6. Smoking-mediated coexpression networks.** Gene coexpression patterns specific to smokers (A) and non-smokers (B) derived by the implication network model ( $P < 0.05$ ; z-tests) commonly present in both training and test sets. The biological interpretation of the implication relations is described in (C). The stability of smoking-mediated networks as evaluated with random subsets of patients from the training cohort in 100 iterations (D).

**Figure 7. Comparison with gene association networks based on Pearson's correlation coefficients.** Number of gene associations derived with implication networks and Pearson's correlation coefficients on the training set (A), testing set (B) and common gene associations in both training and testing sets (C). Genet: implications networks.

# Figure 1



# Figure 2

**Begin**

Set a significance 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 Figure 1

$$M_{ij} = \begin{array}{|c|c|} \hline N_{11} & N_{12} \\ \hline N_{21} & N_{22} \\ \hline \end{array}$$

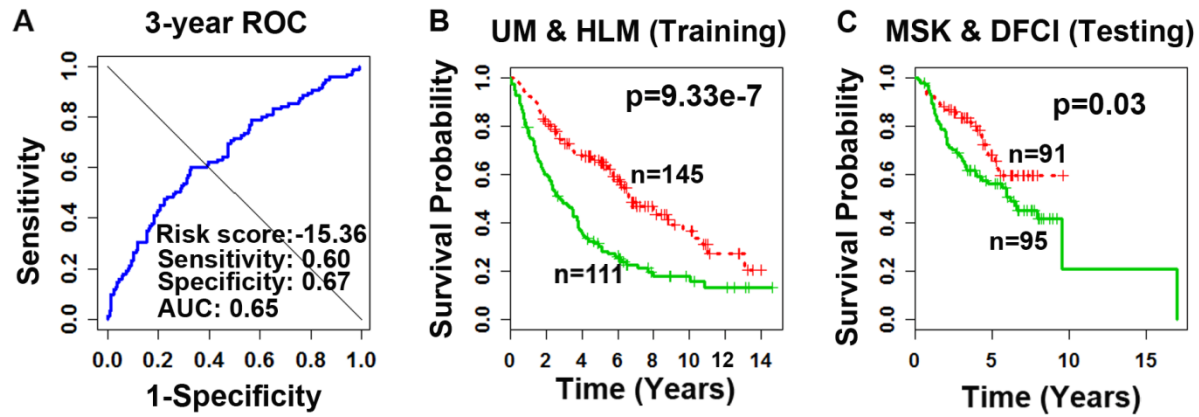
For each relation type  $k$  out of the six cases find the solution

$$\begin{array}{l} \text{Subject to} \\ \text{Max } U_p \\ \text{Max } U_p > U_{min} \\ \nabla_p \geq \nabla_{min} \\ \nabla_{error\ cells} > \nabla_{non-error\ cells} \end{array}$$

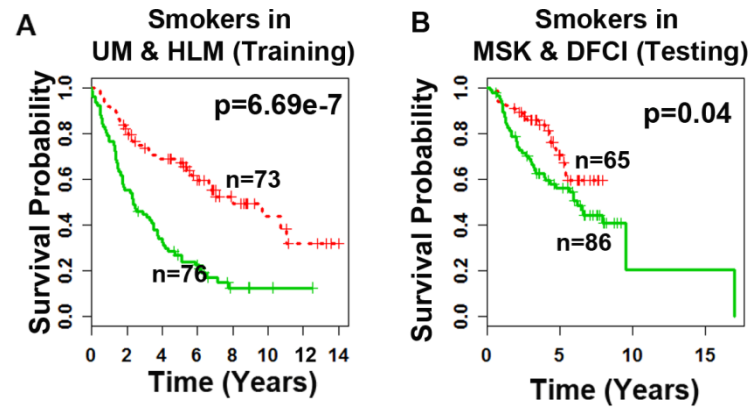
If the solution exists, then return a type  $k$  relation

**End**

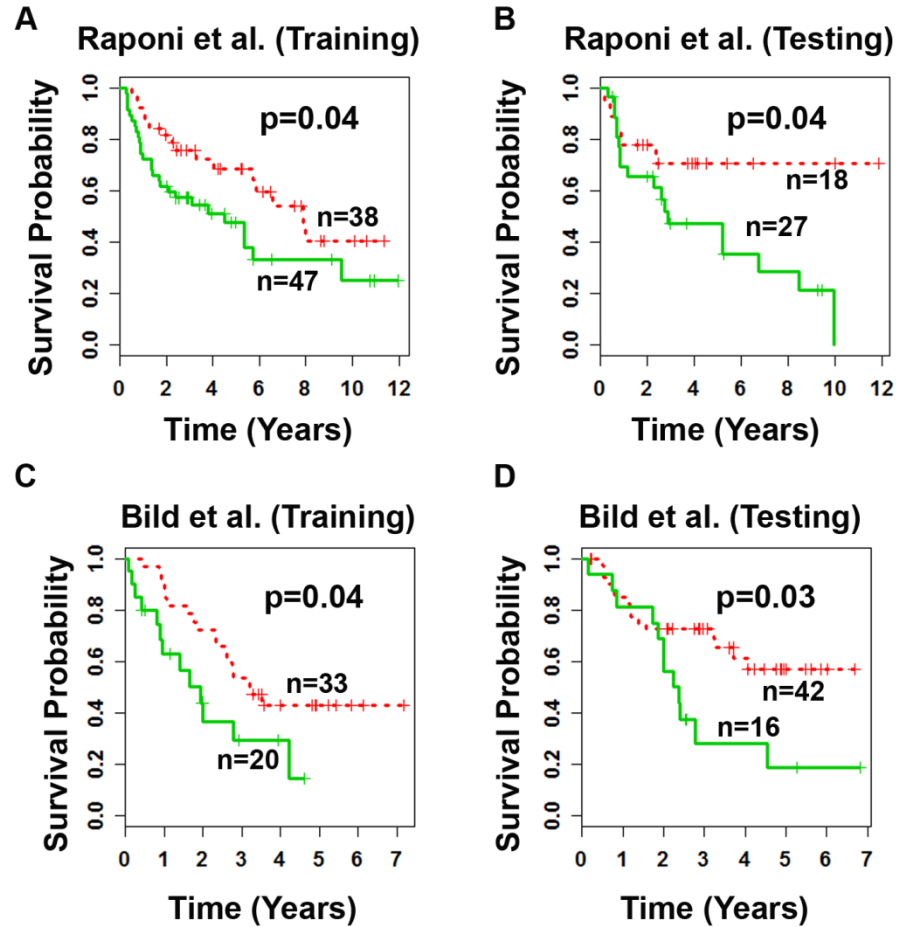
# Figure 3



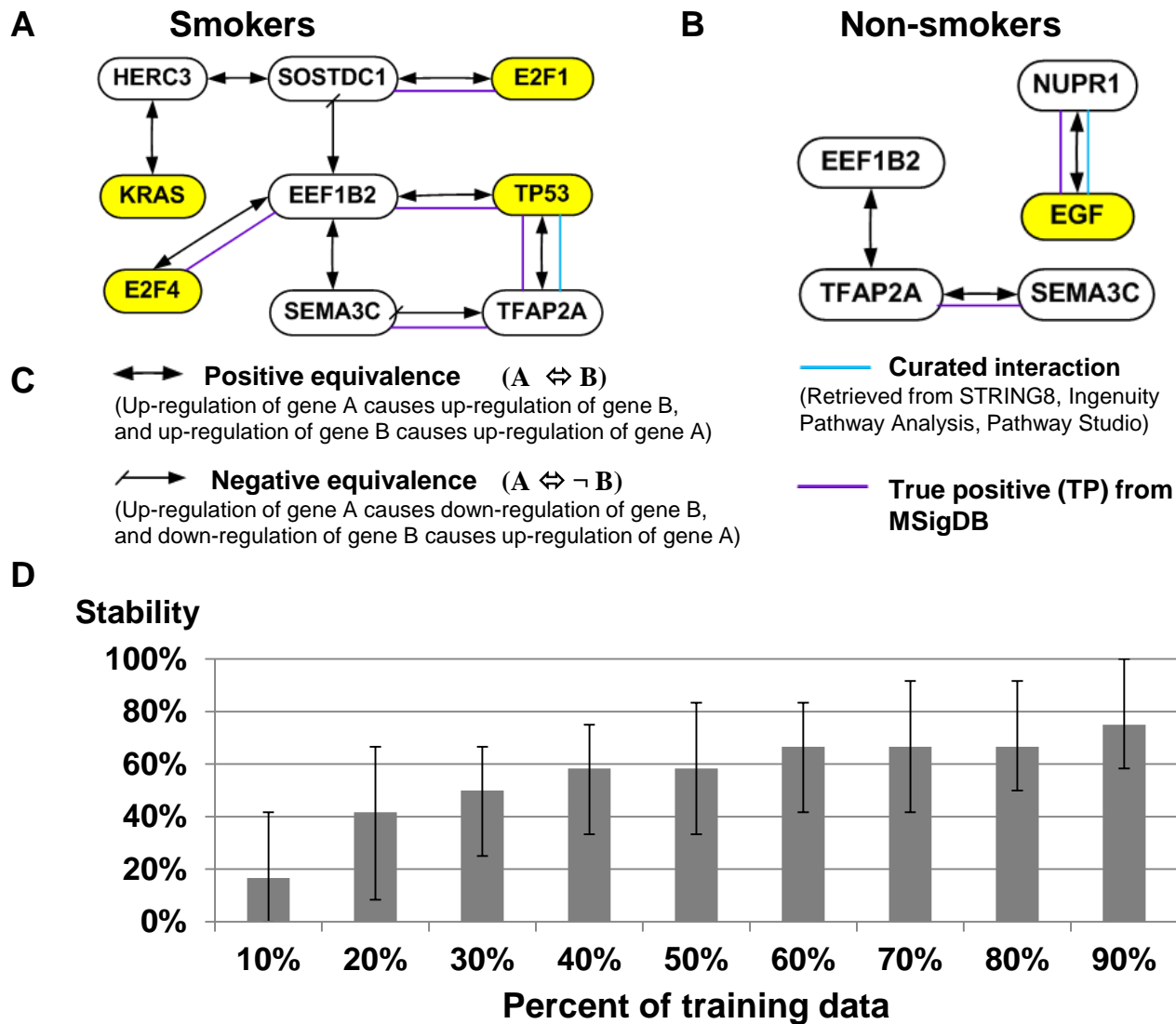
# Figure 4



# Figure 5

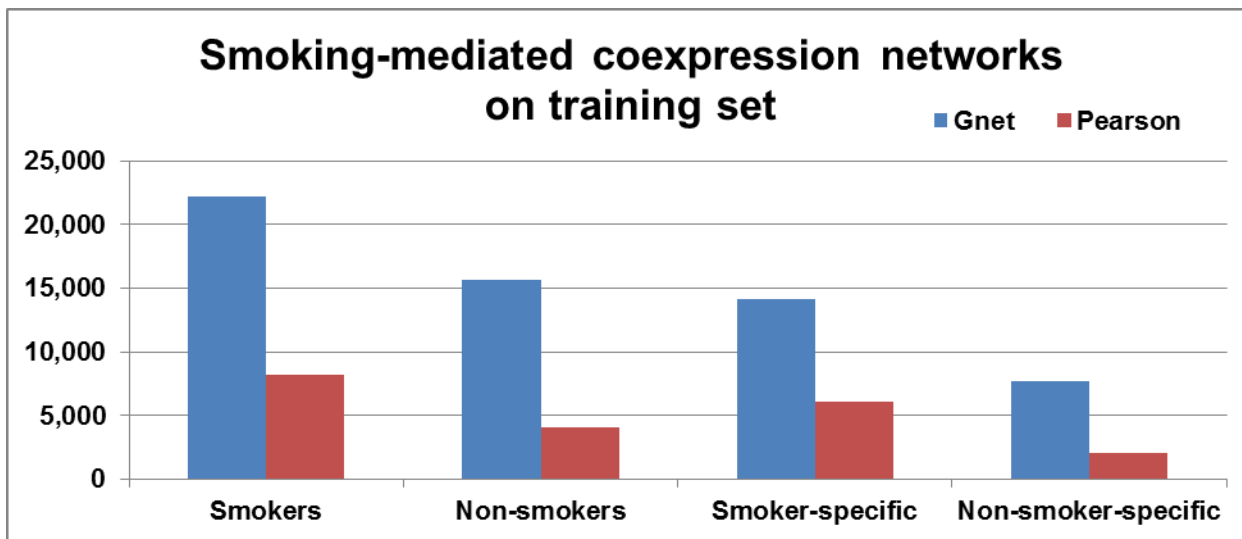


# Figure 6

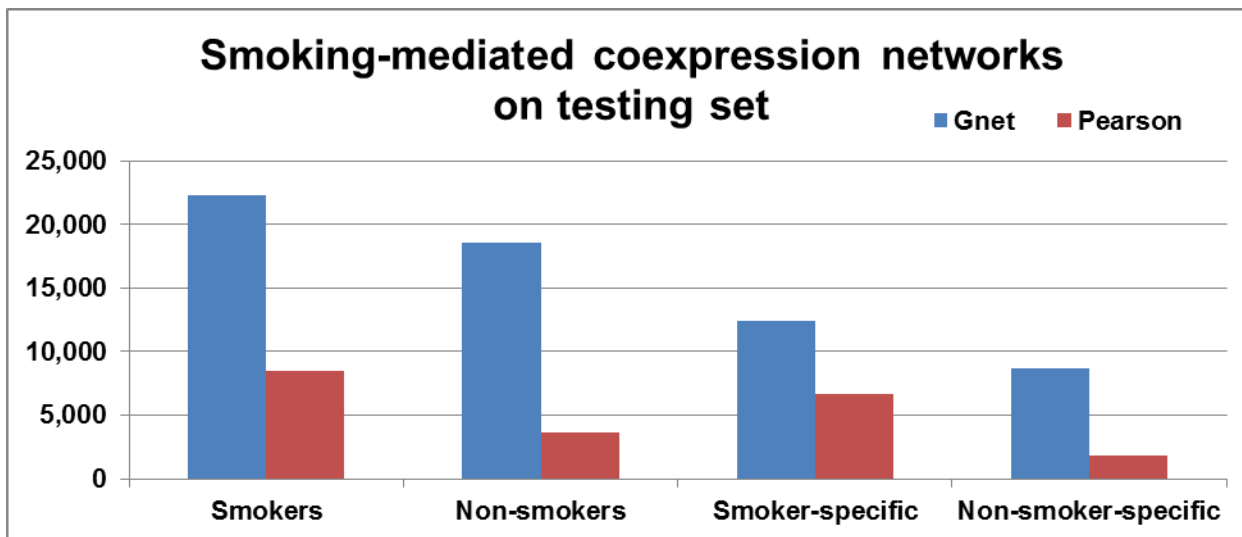


# Figure 7

A



B



# Figure 7

C

