
A novel network model identified a 13-gene lung cancer prognostic signature

Nancy Lan Guo*

Mary Babb Randolph Cancer Centre,
Department of Community Medicine,
West Virginia University,
Morgantown, WV 26506 – 9300, USA
Fax: 304-293-4667
E-mail: lguo@hsc.wvu.edu
Website: <http://www.hsc.wvu.edu/mbrcc/fs/GuoLab/products.asp>
*Corresponding author

Ying-Wooi Wan and Swetha Bose

Lane Department of Computer Science and Electrical Engineering,
West Virginia University,
Morgantown, WV 26506, USA
E-mail: ywan2@mix.wvu.edu
E-mail: swetha_bose@yahoo.com

James Denvir

Department of Statistics,
West Virginia University,
Morgantown, WV 26506, USA
E-mail: Jim.Denvir@mail.wvu.edu

Michael L. Kashon and Michael E. Andrew

National Institute of Occupational Safety and Health,
Biostatistics and Epidemiology,
1095 Willowdale Road, Morgantown, WV 26505, USA
E-mail: mkashon@cdc.gov
E-mail: mta6@cdc.gov

Abstract: This study presents a novel network methodology to identify prognostic gene signatures. Implication networks based on prediction logic are used to construct genome-wide coexpression networks for different disease states. From the differential components associated with specific disease states, candidate genes that are co-expressed with major disease signal hallmarks are selected. From these candidate genes, top genes that are the most predictive of clinical outcome are identified using univariate Cox model and *Relief* algorithm. Using this approach, a 13-gene lung cancer prognosis signature was identified, which generated significant prognostic stratifications (log-rank $P < 0.05$) in Director's Challenge Study ($n = 442$).

Keywords: prognostic gene signature; lung cancer; implication networks; gene co-expression networks; signalling pathways.

Reference to this paper should be made as follows: Guo, N.L., Wan, Y-W., Bose, S., Denvir, J., Kashon, M.L. and Andrew, M.E. (2011) 'A novel network model identified a 13-gene lung cancer prognostic signature', *Int. J. Computational Biology and Drug Design*, Vol. 4, No. 1, pp.19–39.

Biographical notes: Nancy Lan Guo is an Assistant Professor of Community Medicine/Mary Babb Randolph Cancer Centre, and Adjunct Assistant Professor of Computer Science at West Virginia University. She is Program Co-Director of Biomedical Informatics in West Virginia Clinical and Translational Science Institute. She has PhD in Computer and Informatics Science and BS in Biochemistry and Molecular Biology.

Ying-Wooi Wan has MS in Computer Science. She is currently a PhD candidate in Computer and Information Science at West Virginia University.

Swetha Bose has MS in Electrical Engineering. She is currently a Clinical Data Analyst at Cancer Treatment Centres of America (CTCA).

James Denvir has PhD in Mathematics. He is currently a Research Associate at West Virginia University Research Corporation and Marshall University.

Michael L. Kashon has PhD in Neuroscience and MS in Statistics. He is currently a Mathematical Statistician at Biostatistics and Epidemiology Branch of Health Effects Laboratory Division in National Institute of Occupational Safety and Health.

Michael E. Andrew has a PhD in Statistics. He is currently a Mathematical Statistician at Biostatistics and Epidemiology Branch of Health Effects Laboratory Division in National Institute of Occupational Safety and Health.

1 Introduction

Lung cancer is the leading cause of cancer-related deaths in industrialised countries. Non-Small Cell Lung Cancer (NSCLC) accounts for about 80% of lung cancer cases. Currently, surgery is the major treatment option for patients with stage I NSCLC. However, 35–50% of stage I NSCLC patients will develop recurrence and die within five years (Hoffman et al., 2000). It remains an unsolved challenge for physicians to reliably identify patients at high risk for recurrence as candidates for chemotherapy. A few studies have described transcriptional profiling for lung cancer prognosis (Chen et al., 2007; Potti et al., 2006; Shedden et al., 2008). Nevertheless, there is no clinically applied gene test for this deadly disease.

The accurate assessment of disease progression in individual patients is a critical prerequisite in personalised medicine. With the completion of the Human Genome Project, the emphasis of biomarker identification has shifted from cataloguing the 'parts list' of signature genes and proteins to elucidating the networks of interactions that take place among them (Ideker and Sharan, 2008). Molecular network analysis had been shown to be useful in disease classification (Chuang et al., 2007) and identification of

novel therapeutic targets (Csermely et al., 2005). Nevertheless, major challenges have been the development of methods for efficiently constructing genome-scale 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 (Sotiriou and Piccart, 2007). In this study, we hypothesised that the combined analysis of disease-mediated genome-wide coexpression networks, hallmark signalling pathways, and clinical approaches would lead to more informed clinical decision-making. This study will focus on the molecular prognosis of lung cancer relapse and metastasis.

In current genome-wide expression studies, genes are ranked according to their association with the clinical outcome, and the top-ranked genes are included in the classifier. It has been noted that individual biomarkers showing strong association with disease outcome are not necessarily good classifiers (Emir et al., 1998). Genes and proteins do not function in isolation, but rather interact with one another to form modular machines (Hartwell et al., 1999). Molecular network analysis has led to promising applications in identifying new disease genes (Emilsson et al., 2008) and disease-related sub networks (Calvano et al., 2005), and classifying diseases (Chuang et al., 2007).

Boolean networks can provide important biological insights into regulation functions (Albert and Othmer, 2003). Nevertheless, as the number of global states is exponential in the number of entities and the analysis relies on an exhaustive enumeration of all possible trajectories, this method is computationally expensive and only practical for small networks (Karlebach and Shamir, 2008). A recent formalism, causal Bayesian belief networks, have been utilised to model cellular networks (Friedman, 2004). Nevertheless, the number of possible networks is exponential in the number of nodes under consideration, which makes it impossible to evaluate all possible networks. Furthermore, it is not always possible to determine the causal relationships between nodes, i.e., the direction of the edges, owing to a property known as Markov equivalence (Zhu et al., 2008). More importantly, the acyclic Bayesian network structure was unable to model feedback loops, which are essential in signalling pathways (Sachs et al., 2005) and genetic networks (Milo et al., 2002, 2004; Wuchty et al., 2003). To overcome this limitation, a more complex scheme, dynamic Bayesian networks, was explored for modelling temporal microarray data (Kim et al., 2003; Pe'er et al., 2001).

As an alternative to Bayesian networks, an implication network model employs a *Partial Order Knowledge Structure* (POKS) for structural learning and uses the Bayesian theory for inference propagation (Desmarais et al., 1996, 2006). An implication network is a general methodology for reasoning under uncertainty. POKSs are closed under union and intersection of implication relations, and have the formal properties of directed acyclic graphs. The constraints on the partial order can be entirely represented by AND/OR graphs (Desmarais et al., 1996; Falmagne et al., 1990). When the constraints on the partial order are relaxed, the implication networks can represent cyclic relations among the nodes. In this condition, the implication network structure is a directed graph with nodes connected by implication (causal) rules, which can contain cycles such as feedback loops.

Liu and Desmarais (1997) presented the first implication network formalism based on binomial distribution. Boolean implication networks (Sahoo et al., 2008) used scatter plots of expression between two genes to induce the implication relations. In this study, we developed an induction algorithm based on prediction logic (Guo et al., 2003) to derive implication relations. The implication network was employed for efficient construction of disease-mediated genome-wide coexpression networks for the

identification of prognostic gene signatures. The prognostic performance of the identified gene signature was evaluated by comparing with clinical covariates and other gene expression signatures. Furthermore, functional pathway analysis was done to confirm the biological relevance of the identified gene signature.

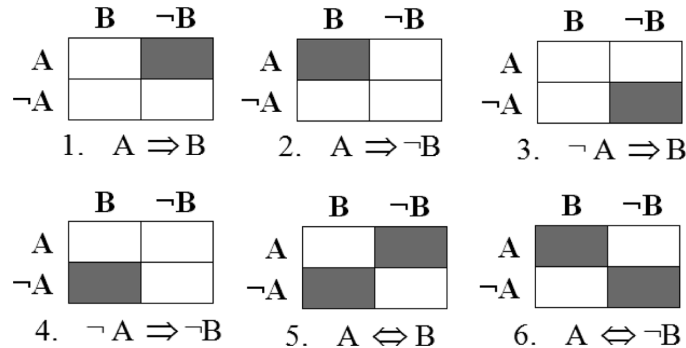
2 Materials and methods

2.1 Implication induction algorithm

The first implication network induction algorithm is based on binomial distribution, which is suitable for binary datasets (Liu and Desmarais, 1997). We proposed an alternative network induction algorithm based on prediction logic (Guo et al., 2003), 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. Prediction logic integrates formal logic theory and statistics to build a convenient predictive structure for a dataset. 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 essential 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 induct the implication network. The six most important implication rules relating two dichotomous variables are shown in Figure 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.

Figure 1 Six important implication rules relating two dichotomous variables. 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



Our algorithm modified U -Optimality method (Hildebrand et al., 1977) (Figure 2), and was used to derive the implication relation between each pair of variables in the dataset. In the implication induction algorithm (Figure 2), 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. 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 ∇_{\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, ∇_p are greater than the required minimum values, U_{\min} and ∇_{\min} , respectively. To simplify the computations of the maximisation problem, the ∇_{ij} value of every error cell must be greater than that of the non-error cells for the corresponding implication rule (Guo et al., 2003).

Figure 2 Implication induction algorithm for building coexpression networks

The Implication Induction Algorithm

Begin

Set a significant level ∇_{\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

$$M_{ij} = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix}$$

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

Subject $Max U_p$
 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

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

$$U_p = U_{ij} = \frac{N_i \times N_j}{N^2}, \quad \nabla_p = \nabla_{ij} = 1 - \frac{N_{ij}}{N \times U_p}.$$

For multiple error cells, they are defined as:

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

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} (Figure 2), implication rules can be induced for multinomial datasets, where error cells are those with top precision (∇_{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) (Guo et al., 2003). The difference between this algorithm and the original U-Optimality (Hildebrand et al., 1977) is that minimum requirements for deriving an implication rule were set for both scope (U_p) and precision (∇_p), instead of for precision alone.

2.2 Relief feature selection algorithm

From the set of prognostic genes identified from implication network methodology, *Relief* was used to rank these genes with software WEKA 3.4 (Witten and Frank, 2005) in order to select the most predictive genes. *Relief* evaluates the importance of a variable by repeatedly sampling an instance and checking the value of the given variable for the nearest instance from the same and different classes. The values of the attributes of the nearest neighbours are compared to the sampled instance and used to update the relevance scores for each attribute. As approximated in following equation, *Relief* computes the weight of attributed as:

$$W[A] = P(\text{different value of } A_{\text{near miss}}) - P(\text{different value of } A_{\text{near hit}}).$$

Relief assigns more weight to those attributes that have the same value for instances from the same class and differentiate from instances in different classes (Hall and Holmes, 2003; Witten and Frank, 2005).

2.3 Functional pathway analysis

A proprietary web-based software Ingenuity Pathway Analysis (IPA, Ingenuity® Systems)¹ were used to derive curated molecular interactions, including both physical and functional interactions, and pathway relevance 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. Core analysis was used to identify significant biological processes and functions from the merged network related to the identified 13-gene signature in human tissues and cell lines.

2.4 Microarray profiles and patient samples

Gene expression profiles quantified with Affymetrix HG-U133A on 442 lung adenocarcinoma samples from a published study (Shedden et al., 2008) were used in this study. This study cohort is composed of four data sets (University of Michigan, H. Lee Moffitt Cancer Centre, Memorial Sloan-Kettering Cancer Centre, and Dana-Farber Cancer Institute) contributed by six institutions. Tumours were collected by surgical resection from patients who have provided consent and protocols were approved by the Institutional Review Boards (IRB-Med) of the respective institutions. None of the

patients received preoperative chemotherapy or radiation and least two years of follow-up information was available. Regions containing a minimum of 60% tumour cellularity were required for macro-dissection, and in most instances tumour cellularity of at least 70–90% was identified for inclusion in the sample for RNA isolation. The raw microarray data are available from caArray website.² The data used in the analysis was quantile-normalised and \log_2 -transformed with dChip (Li, 2008).

3 Results and discussion

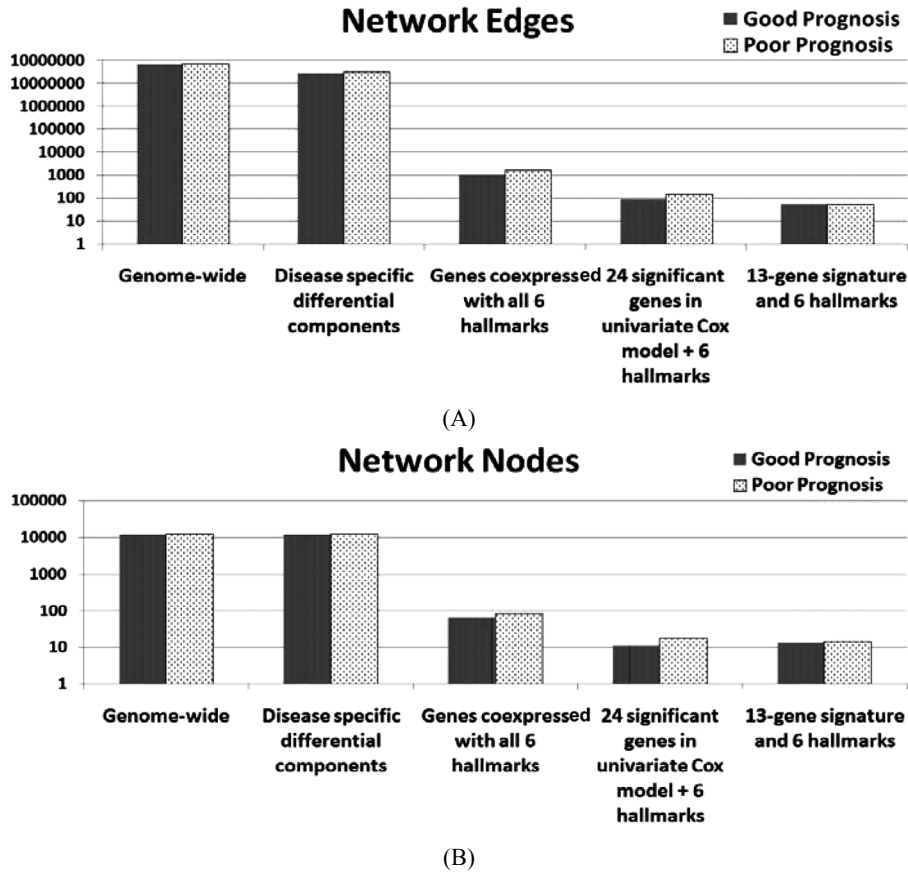
3.1 Identification of prognostic genes using implication networks

In this study, patient samples from UM and HLM formed the training set ($n = 256$), whereas samples from MSK ($n = 104$) and DFCI ($n = 82$) constituted two independent test sets. Genes with missing measurements in at least half of the samples were removed from analysis. Furthermore, for genes measured with multiple probes, the average expression of the duplicates was used to represent the expression profile of a unique gene. This gave 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 tumour sample; otherwise, it was denoted as *down-regulated* in the tumour sample. In the training set, patients who died within five years were labelled as poor-prognosis ($n = 125$), and those who survived five years after surgery were labelled as good-prognosis ($n = 104$). Censored cases (those with follow-up of less than five years) were removed from the analysis ($n = 27$). For each patient group in the training set, a genome-scale coexpression network was constructed using the implication induction algorithm. Between each pair of genes, possible significant ($P < 0.05$; one-sided z -tests of U_{\min} and ∇_{\min}) 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 (interactions) 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.

Next, genes displaying direct co-regulation with major NSCLC signal proteins were identified from the disease-specific differential network components. The signal proteins included in the study were retrieved from the human NSCLC signalling pathways delineated by the KEGG pathway database.³ Genes of a significant ($P < 0.05$; z -tests) coexpression relation with *MET*, *EGF*, *KRAS*, *TP53*, *E2F2*, and *E2F4* were pinpointed from the differential components associated with each prognosis group. As a result, 76 genes were identified from the poor-prognosis group, 58 genes from the good-prognosis group, and 9 genes common in both groups, yielding a set of 125 genes. The number of nodes (genes) and edges (interactions) in the networks associated with specific prognostic groups in each analytical step are shown in Figure 3.

Figure 3 The characteristics of the coexpression networks at each step in the identification of the 13-gene signature: (A) the number of the edges of the coexpression networks in each step, starting from disease-associated genome-wide coexpression networks, disease-specific differential components, genes significantly ($P < 0.05$) coexpressed with 6 cancer signalling hallmarks, 24 genes significant ($P < 0.05$) in univariate Cox model, and the 13-gene signature and 6 hallmarks in good- and poor-prognosis groups and (B) The number of the nodes of the coexpression networks in each step described in (A)



3.2 Survival prediction using the identified prognostic genes

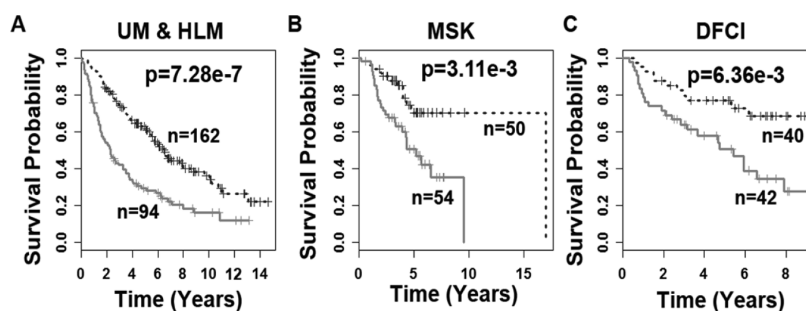
We sought to evaluate whether the genes identified from the proposed implication network analysis could generate accurate prognostic prediction. From the training set of the original continuous microarray data, 24 probes out of the 125 genes selected in the previous steps were significantly associated with overall survival ($P < 0.05$, univariate Cox modelling). These 24 significant probes were ranked by *Relief* (Witten and Frank, 2005) and a step-wise forward selection was used to identify the subset with the highest prognostic accuracy. Specifically, starting from the top ranked gene, one gene was added at each step to the gene set, until the prognostic accuracy could not be improved by adding more genes. As a result, the top 13 genes were identified as the most accurate prognostic gene signature (Table 1).

Table 1 The identified 13-gene lung cancer prognostic signature

| Probe set ID | Gene symbol | Gene title |
|--------------|-------------|--|
| 209157_at | DNAJA2 | Dnaj (Hsp40) homolog, subfamily A, member 2 |
| 200705_s_at | EEF1B2 | Eukaryotic translation elongation factor 1 beta 2 |
| 219785_s_at | FBXO31 | F-box protein 31 /// hypothetical protein LOC100288525 |
| 219388_at | GRHL2 | Grainyhead-like 2 (Drosophila) |
| 210981_s_at | GRK6 | G protein-coupled receptor kinase 6 |
| 219357_at | GTPBP1 | GTP binding protein 1 |
| 202621_at | IRF3 | Interferon regulatory factor 3 |
| 203144_s_at | KIAA0040 | Kiaa0040 |
| 207581_s_at | MAGEB4 | Melanoma antigen family B, 4 |
| 218558_s_at | MRPL39 | Mitochondrial ribosomal protein L39 |
| 203379_at | RPS6KA1 | Ribosomal protein S6 kinase, 90kda, polypeptide 1 |
| 205177_at | TNNI1 | Troponin I type 1 (skeletal, slow) |
| 206505_at | UGT2B4 | UDP glucuronosyltransferase 2 family, polypeptide B4 |

Multivariate Cox proportional hazard model was fitted with the 13 genes as covariates on bootstrapped training samples for 1000 times. The average of the 1000 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 8.87 was identified as a cut-off value for patient stratification in the training set (Figure 4(A)). This training model and cut-off value were then applied to the two validation sets to generate prognostic categorisation without re-estimating parameters (Figure 4(B) and (C)). In all three patient cohorts, this scheme stratified patients into prognostic groups with distinct post-operative overall survival (log-rank $P < 0.006$, Kaplan-Meier analyses).

Figure 4 Prognostic prediction of lung cancer survival after surgery by the 13-gene prognostic model. The model stratified patients into two significantly distinct ($P < 0.006$) prognostic groups in the training set (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



When high-risk group is defined as a group of patients who survived 5 years or less, and low-risk group as a group of patients who survived 5 years or longer, this model achieved sensitivity (correctly predicted high-risk patients) of 52% in the training set,

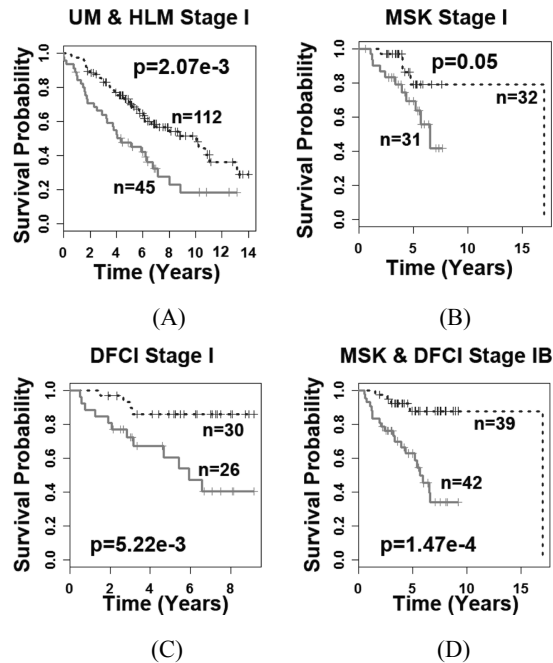
67.65% in MSK, and 67.86% in DFCI; and specificity (correctly predicted low-risk patients) of 77.88% in the training set, 51.61% in MSK, and 61.11% in DFCI.

3.3 Survival prediction on early stage lung cancer

Current treatment options for patients with NSCLC are given based on AJCC tumour stage. Surgical resection is the major treatment option for stage I NSCLC patients. However, about 35–50% of stage I NSCLC patients will develop and die from tumour recurrence within the five years following surgery (Hoffman et al., 2000; Naruke et al., 1988). On the other hand, stage IB patients who received surgical resection followed by adjuvant chemotherapy showed improved survival rate (Lu et al., 2006). This indicates that certain patients with stage I NSCLC are at high risk for developing recurrent diseases. Thus, we sought to explore whether the constructed 13-gene prognostic model could identify specific high-risk patients with stage I tumours.

Results show that the 13-gene prognostic signature could identify high-risk patients with stage I tumours on both the training set and two test sets (log-rank $P \leq 0.05$; Figure 5(A)–(C)). The prognostic model also separated high- and low-risk groups within stage IB patients in the combined test sets (log-rank $P = 1.47e-4$; Figure 5(D)). The 13-gene signature was able to generate significant prognostic stratification on the stage IA patients on the training set but not the test set (results not shown).

Figure 5 Prognostic performance of the 13-gene model in stage I lung cancer. The 13-gene model could further stratify stage I patients into high- and low-risk groups with significantly distinct ($P < 0.05$) survival in all three studied cohorts, including training set UM/HLM (A), test set MSK (B), and test set DFCI (C). The model also generated significant stratifications ($P < 0.0001$) in stage IB patients in MSK/DFCI (D). Statistical significance of the difference in survival probability between the two prognostic groups was assessed with log-rank tests

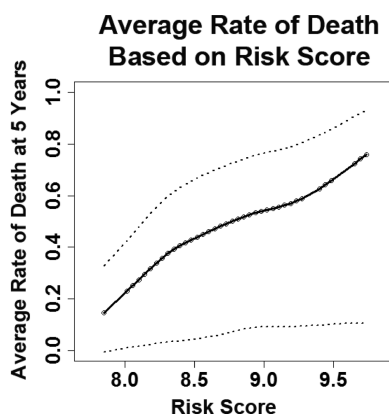


These results demonstrate that the constructed prognostic model provides more refined prognosis than the current AJCC staging system. Using this model, patients with stage I NSCLC could be advised to either receive or spare from chemotherapy according to the expression profiles of the 13 prognostic genes.

3.4 Prognosis evaluation with clinical covariates

The constructed 13-gene prognostic model was evaluated with common lung cancer prognostic factors to further validate the prognostic power of the model. The clinical factors studied include gender, age, tumour stage, smoking history, race, and tumour differentiation. The predicted 13-gene risk score on the combined testing cohorts (MSK and DFCI) was used as a covariate in the analysis. Hazard ratio of the 13-gene risk score represents the likelihood of death from lung cancer for predicted high-risk patients (with estimated probability of death within 5-year after surgery $\geq 50\%$) vs. predicted low-risk patients (with estimated probability of death $< 20\%$), based on the estimated average rate of death associated with gene expression defined-risk scores (Figure S1).

Figure S1 Association of the predicted 13-gene risk score and lung cancer survival. The solid line represents the average rate of death at three years after surgery corresponding to 13-gene risk scores. The dotted lines represent 95% confidence interval



In the first multivariate Cox analysis with major prognostic factors (Table 2), tumour stage was the only factor significantly ($P < 0.00006$) associated with elevated risk of lung cancer death (recurrence) when the model was fitted without the 13-gene risk score. When the 13-gene risk score was added to the multivariate Cox model, the 13-gene risk score demonstrated a strong association with the lung cancer survival (hazard ratio = 2.25, 95% CI: [1.28, 3.98]), and tumour stage remained significant (Table 2). Similarly, a comprehensive evaluation was carried out with all available clinical covariates and demographic data in the dataset, including smoking history, race, and tumour differentiation (Table 3). In this comprehensive evaluation, the 13-gene risk score remained a highly significant prognostic factor with a hazard ratio of 2.28 (95% CI: [1.23, 4.21]). Furthermore, in the comprehensive multivariate analysis, the hazard ratios of the 13-gene risk score algorithm were higher than other clinical covariates except tumour stage (III vs. I). These results demonstrate that the 13-gene signature is a more accurate prognostic factor than some commonly used clinical parameters.

Table 2 Multivariate Cox proportional analysis of 13-gene risk score and major clinical covariates including gender, age, and tumour stage on the combined testing cohorts (MSK and DFCI)

| <i>Variable*</i> | <i>P-value</i> | <i>Hazard ratio</i> | <i>(95% CI)[¶]</i> |
|--|----------------|---------------------|-----------------------------|
| <i>Analysis without 13-gene risk score</i> | | | |
| Gender (Male) | 0.22 | 1.34 | (0.84, 2.16) |
| Age at diagnosis (>60) | 0.08 | 1.61 | (0.95, 2.74) |
| Tumour stage | | | |
| Stage II | 6.25E-05 | 2.91 | (1.72, 4.91) |
| Stage III | 1.09E-05 | 4.16 | (2.20, 7.85) |
| <i>Analysis with 13-gene risk score</i> | | | |
| Gender (Male) | 0.13 | 1.44 | (0.89, 2.32) |
| Age at diagnosis (>60) | 0.10 | 1.57 | (0.92, 2.66) |
| Tumour stage | | | |
| Stage II | 3.46E-04 | 2.62 | (1.55, 4.44) |
| Stage III | 8.99E-06 | 4.24 | (2.24, 8.01) |
| 13-gene risk score | 5.10E-03 | 2.25 | (1.28, 3.98) |

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); tumour stage was categorical variable of 3 categories (Stage I [as the reference group], Stage II, and Stage III).

[¶]Denotes confidence interval.

Table 3 Multivariate Cox proportional analysis of all available clinical covariates and 13-gene risk score in the combined test cohorts (MSK and DFCI)

| <i>Variable*</i> | <i>P-value</i> | <i>Hazard ratio</i> | <i>(95% CI)[¶]</i> |
|--|----------------|---------------------|-----------------------------|
| <i>Analysis without 13-gene risk score</i> | | | |
| 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 | | | |
| Smokers | 0.62 | 0.84 | (0.43, 1.66) |
| Unknown | 0.91 | 0.89 | (0.11, 7.10) |
| Tumour differentiation | | | |
| Moderately differentiated | 0.14 | 0.53 | (0.23, 1.24) |
| Poorly differentiated | 0.70 | 1.17 | (0.53, 2.61) |
| Tumour stage | | | |
| Stage II | 3.31E-04 | 2.72 | (1.57, 4.69) |
| Stage III | 2.38E-05 | 4.93 | (2.35, 10.33) |

Table 3 Multivariate Cox proportional analysis of all available clinical covariates and 13-gene risk score in the combined test cohorts (MSK and DFCI) (continued)

| <i>Variable*</i> | <i>P-value</i> | <i>Hazard ratio</i> | <i>(95% CI)[‡]</i> |
|---|----------------|---------------------|-----------------------------|
| <i>Analysis with 13-gene risk score</i> | | | |
| Gender (Male) | 0.24 | 1.36 | (0.82, 2.24) |
| Age at diagnosis (>60) | 0.05 | 1.71 | (0.99, 2.94) |
| Race | | | |
| Others/ unknown | 0.38 | 0.50 | (0.11, 2.33) |
| White | 0.12 | 0.30 | (0.07, 1.37) |
| Smoking history | | | |
| Smokers | 0.36 | 0.73 | (0.37, 1.44) |
| Unknown | 0.88 | 0.85 | (0.11, 6.80) |
| Tumour differentiation | | | |
| Moderately differentiated | 0.19 | 0.56 | (0.24, 1.32) |
| Poorly differentiated | 0.78 | 1.12 | (0.50, 2.51) |
| Tumour stage | | | |
| Stage II | 8.38E-04 | 2.52 | (1.46, 4.33) |
| Stage III | 2.58E-05 | 5.04 | (2.37, 10.70) |
| 13-gene risk score | 0.01 | 2.28 | (1.23, 4.21) |

*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]); tumour 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); tumour stage was categorical variable of 3 categories (Stage I [as the reference group], Stage II, and Stage III).

[‡]Denotes confidence interval.

3.5 Comparison with other lung cancer gene signatures

Multiple prognostic classifiers were evaluated in the Director’s Challenge Study (Shedden et al., 2008). There were 12 classifiers constructed with gene signatures alone. Five of the 12 analysed signatures were from previous studies on lung cancer molecular prognosis (Chen et al., 2007; Potti et al., 2006). These published lung cancer gene signatures were identified using traditional statistical and machine learning methods (Table S1). Among the 12 gene signatures compared in the Director’s Challenge Study, the best signature was reported as ‘method A’ (referred to as ‘Shedden A’ in this study), which contains about 9591 genes/probes. In order to compare the prognostic performance of our gene signature with the best lung cancer gene signatures reported to date, the estimated hazard ratio and the Concordance Probability Estimate (CPE) of the gene signatures in both test sets were evaluated. Hazard ratios greater than 1 indicate that patients with high predicted risk scores have poor clinical outcome. The model has strong predictive power if the CPE value is close to 1; CPE value close to 0.5 indicates that the model has poor predictive power (comparable to random prediction).

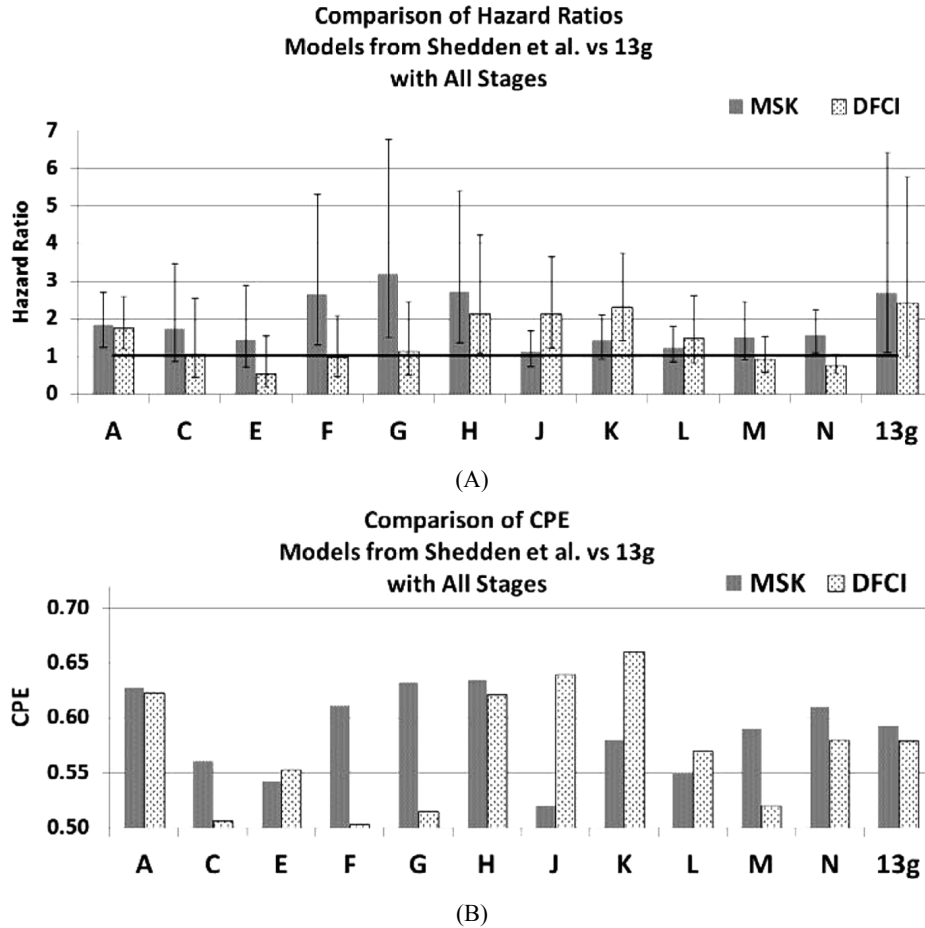
Table S1 Summary of gene selection and classification methods of molecular classifiers compared in Figure 5. Gene signatures A-N were reported in Shedden et al. (2008)

| <i>Molecular classifier*</i> | <i>Number of signature genes</i> | <i>Gene selection method(s)</i> | <i>Classification method(s)</i> |
|------------------------------|----------------------------------|---|--|
| Shedden A | ~ 9591 genes | Clustering analysis | Ridged Cox proportional hazard model |
| Shedden C | 23 genes | SAM, maximising chi-square analysis (MCA, univariate Cox model and k-mean clustering) | Binary Tree-Structured Vector Quantisation (BTSVQ) |
| Shedden D | 37 genes | SAM, maximising chi-square analysis (MCA, univariate Cox model and k-mean clustering) | Binary Tree-Structured Vector Quantisation (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 al., NEJM 07) | Ridged Cox 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 |
| 13-gene | 13 Genes | Implication network, RELIEF | 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).

Results show that the 13-gene prognostic model gives comparative performance as ‘Shedden A’, and is better than all other lung cancer gene signatures. The 13-gene model and ‘Shedden A’ are the only two models with hazard ratio significantly ($P < 0.05$) greater than 1 in patients with all tumour stages (Figure 6(A)). CPE of the 13-gene model is close to 0.6 in patients samples with all stages (Figure 6(C)), which is comparable with ‘Shedden A’. Nevertheless, ‘Shedden A’ is composed of more than 9000 genes, which would be infeasible to be implemented as a clinical prognostic test.

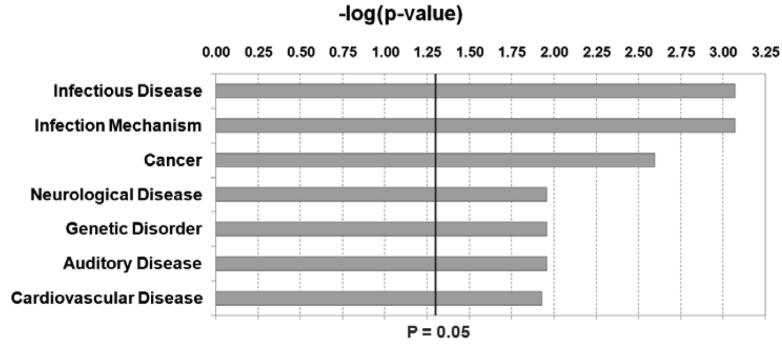
Figure 6 Comparison of 13-gene prognostic model and various gene expression-defined models presented in the director's challenge study (Shedden et al., 2008) in two test sets in term of hazard ratio (A) and concordance probability estimate (B)



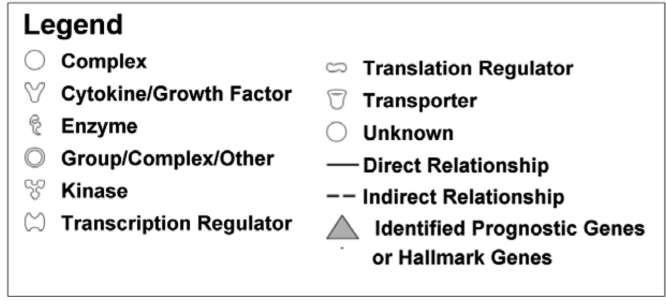
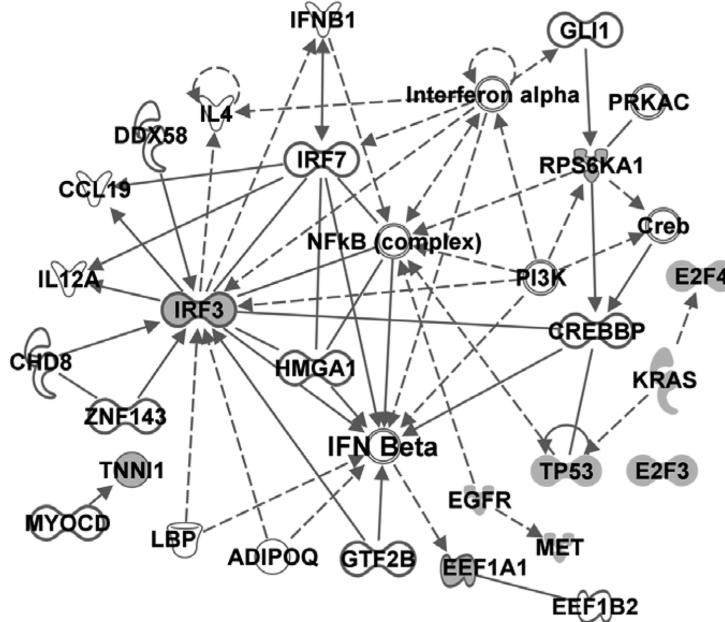
3.6 Functional pathway analysis

Having established the clinical relevance of the 13 prognostic genes identified, we sought to explore the functional involvement of this gene set in lung tumorigenesis and tumour progression. Curated molecular interactions related to the 13 genes were retried using functional pathway analysis tools, Ingenuity Pathway Analysis (IPA, Ingenuity® Systems). IPA results show that cancer is among the top five most significant disease and disorder functions in the network related to the 13 genes (Figure 7(A)). Furthermore, four of the prognostic genes exhibit indirect interactions with major lung cancer signalling pathways, such as *TP53* and *EGFR* (Figure 7(B)). The functional pathway analysis suggests that the 13 genes are involved in lung cancer oncogenesis and tumour progression.

Figure 7 Functional pathways analysis of the 13 prognostic genes. Using core analysis from Ingenuity Pathway Analysis (IPA), cancer was the third most significant biological function in the disease and disorders category (A). Curated interactions related to the 13 signature genes were also revealed (B)



(A)



(B)

4 Discussion

Our previous study identified a 12-gene signature using hybrid models combining differential expression analysis (SAM) and *Relief* algorithm (Wan et al., 2010). Both 12- and 13-gene signatures had significant hazard ratios in prognostic categorisation for patients with all tumour stages in two test sets (MSK and DFCI) of the Director's Challenge Cohorts. The overall accuracies of the 12- and 13-gene prognostic models were not significantly different in 5-year survival prediction (Table S2). For stage I patients, the 13-gene had a significant ($P < 0.05$) hazard ratio of 2.96 in DFCI set but not in MSK, whereas the hazard ratio of 12-gene signature was not significant in any test sets. Taken together, the 13-gene signature identified in this network-based study has better prognostic performance in stage I lung adenocarcinoma patients than our previously identified 12-gene signature using traditional statistical methods.

Table S2 Sensitivity and specificity of the 13- and 12-gene prognostic models on 5-year survival. Patients who survived 5 years or longer were defined as low-risk; patients who died within 5 years were defined as high-risk

| | Sensitivity (% of correctly predicted high-risk patients) | | | Specificity (% of correctly predicted low-risk patients) | | | Overall accuracy (%) | | | |
|--------|---|---------|---------|--|---------|---------|----------------------|---------|---------|-----------------|
| | <i>n</i> | 13-gene | 12-gene | <i>n</i> | 13-gene | 12-gene | <i>n</i> | 13-gene | 12-gene | <i>P</i> -value |
| UM&HLM | 125 | 52.00 | 72.80 | 104 | 77.88 | 66.35 | 229 | 63.75 | 69.87 | 0.08 |
| MSK | 34 | 67.65 | 70.59 | 31 | 51.61 | 48.39 | 65 | 60.00 | 60.00 | 0.50 |
| DFCI | 28 | 67.86 | 64.29 | 36 | 61.11 | 77.78 | 64 | 64.06 | 71.88 | 0.17 |

In this study, a significance level of $P < 0.05$ (one-sided z -tests of U_{\min} and ∇_{\min}) for coexpression relation was used as cut-offs to define possible gene interactions in the network construction. Our approach in the first step of this methodology is not to attempt to produce a list of gene-gene interactions or co-expressions with which we can associate a robust measure of significance. In order to do this, we would need to implement one of two types of mechanism to control for multiple hypothesis testing. The first type of mechanism is to compute a raw P -value for each interaction and then to adjust those P -values using a correction such as Bonferroni or Benjamini-Hochberg test. These corrections tend to be highly conservative when there is a high degree of dependence between the hypotheses, which is clearly the case here when the set of hypotheses is all possible gene-gene interactions, and the consequent loss of statistical power would be too high. The second mechanism is to perform a large number of random permutations of the class labels of the data and to perform the algorithm for each permutation, and then to compare the actual results to the generated null distribution of data. This mechanism is prohibitively computationally expensive in this context.

Instead, our approach is as follows. Rather than determine pairs of genes between which interactions exist, for each pair of genes we determine which type of interaction is most strongly supported by the data.

In other words, our methodology classifies a coexpression relation (if any) between a pair of genes, as represented in one of six common logical rules. Our cutoffs determined by U_{\min} and ∇_{\min} eliminate a relatively small number of pairs of genes for which no logical rule provides strong evidence. We then determine a list of differential components, which are interactions between pairs of genes that classify differently between the different disease states. This list will inevitably contain a number of false positives, for reasons outlined above. Instead of controlling for the number of false positives at this stage of the methodology, we subsequently filter the differential components through steps including selecting significant genes from univariate Cox models and then from genes ranked top in the *Relief* algorithm, as outlined in the previous sections. As a matter of fact, we have experimented using more stringent significance levels for U_{\min} and ∇_{\min} , and the size of the resultant differential components associated with different disease-states was not reduced remarkably (data not shown).

5 Conclusions

This study presents a novel network-based approach to identifying a 13-gene signature for lung cancer prognosis. The identified signature could accurately estimate post-operative survival in lung adenocarcinoma patients. Furthermore, the signature could stratify patients within stage I and specifically, stage IB, into distinct prognostic groups. The gene expression-defined risk score is a more accurate prognostic factor than commonly used clinical covariates. In functional pathway analysis, the 13 genes also exhibit strong associations with cancer signalling hallmarks and oncogenesis.

This study demonstrates that the implication network methodology based on prediction logic is suitable for constructing genome-wide coexpression networks for analysing 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 targeted therapy and prognostic prediction. These results conclude that the presented implication network methodology can retrieve disease relevant gene coexpression patterns and is useful in the discovery of clinically important biomarkers. Furthermore, gene signatures identified with this novel network-based methodology provide better prognostic performance than those identified with traditional statistical and machine learning methods on the same datasets.

Acknowledgements

We thank Dr. David Beer at University of Michigan and Dr. Trey Ideker at University of California in San Diego for thoughtful discussions. This project is supported by NIH R01LM009500 (PI: Guo) and NCRR P20RR16440 and Supplement (PD: Guo). Software license and training for Ingenuity Pathway Analysis is supported by NIH/NCRR P2016477.

References

- Albert, R. and Othmer, H.G. (2003) 'The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*', *J. Theor. Biol.*, Vol. 223, No. 1, pp.1–18.
- Calvano, S.E., Xiao, W., Richards, D.R., Felciano, R.M., Baker, H.V., Cho, R.J., Chen, R.O., Brownstein, B.H., Cobb, J.P., Tschoeke, S.K., Miller-Graziano, C., Moldawer, L.L., Mindrinos, M.N., Davis, R.W., Tompkins, R.G. and Lowry, S.F. (2005) 'A network-based analysis of systemic inflammation in humans', *Nature*, Vol. 437, No. 7061, pp.1032–1037.
- Chen, H.Y., Yu, S.L., Chen, C.H., Chang, G.C., Chen, C.Y., Yuan, A., Cheng, C.L., Wang, C.H., Terng, H.J., Kao, S.F., Chan, W.K., Li, H.N., Liu, C.C., Singh, S., Chen, W.J., Chen, J.J. and Yang, P.C. (2007) 'A five-gene signature and clinical outcome in non-small-cell lung cancer', *N. Engl. J. Med.*, Vol. 356, No. 1, pp.11–20.
- Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. and Ideker, T. (2007) 'Network-based classification of breast cancer metastasis', *Mol. Syst. Biol.*, Vol. 3, p.140.
- Csermely, P., Agoston, V. and Pongor, S. (2005) 'The efficiency of multi-target drugs: the network approach might help drug design', *Trends Pharmacol. Sci.*, Vol. 26, No. 4, pp.178–182.
- Desmarais, M.C., Maluf, A. and Liu, J. (1996) 'User-expertise modeling with empirically derived probabilistic implication networks', *User Modeling and User-Adapted Interaction*, Vol. 5, Nos. 3–4, pp.283–315.
- Desmarais, M.C., Meshkinfam, P. and Gagnon, M. (2006) 'Learned student models with item to item knowledge structures', *User Modeling and User-Adapted Interaction*, Vol. 16, No. 5, pp.403–434.
- Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G.H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadóttir, A., Jonasdottir, A., Jonasdottir, A., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K.P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H.G., Stefansson, T., Leifsson, B.G., Thorsteinsdottir, U., Lamb, J.R., Gulcher, J.R., Reitman, M.L., Kong, A., Schadt, E.E. and Stefansson, K. (2008) 'Genetics of gene expression and its effect on disease', *Nature*, Vol. 452, No. 7186, pp.423–428.
- Emir, B., Wieand, S., Su, J.Q. and Cha, S. (1998) 'Analysis of repeated markers used to predict progression of cancer', *Stat. Med.*, Vol. 17, No. 22, pp.2563–2578.
- Falmagne, J.C., Doignon, J.P., Koppen, M., Villano, M. and Johannesen, L. (1990) 'Introduction to knowledge spaces: how to build, test and search them', *Psychological Review*, Vol. 97, No. 2, pp.201–224.
- Friedman, N. (2004) 'Inferring cellular networks using probabilistic graphical models', *Science*, Vol. 303, No. 5659, pp.799–805.
- Guo, L., Cukic, B. and Singh, H. (2003) 'Predicting fault prone modules by the Dempster-Shafer belief networks', *18th IEEE International Conference on Automated Software Engineering (ASE'03)*, pp.249–252.
- Hall, M.A. and Holmes, G. (2003) 'Benchmarking attribute selection techniques for discrete class data mining', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 3, pp.1437–1447.
- Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) 'From molecular to modular cell biology', *Nature*, Vol. 402, No. 6761, Suppl, p.C47–C52.
- Hildebrand, D.K., Laing, J.D. and Rosenthal, H. (1977) *Prediction Analysis of Cross Classifications*, John Wiley & Sons, Hoffman, PC.
- Hoffman, P.C., Mauer, A.M. and Vokes, E.E. (2000) 'Lung cancer', *Lancet*, Vol. 355, No. 9202, pp.479–485.

- Ideker, T. and Sharan, R. (2008) 'Protein networks in disease', *Genome Res.*, Vol. 18, No. 4, pp.644–652.
- Karlebach, G. and Shamir, R. (2008) 'Modelling and analysis of gene regulatory networks', *Nat. Rev. Mol. Cell Biol.*, Vol. 9, No. 10, pp.770–780.
- Kim, S.Y., Imoto, S., and Miyano, S. (2003) 'Inferring gene networks from time series microarray data using dynamic Bayesian networks', *Brief. Bioinform.*, Vol. 4, No. 3, pp.228–235.
- Li, C. (2008) 'Automating dChip: toward reproducible sharing of microarray data analysis', *BMC Bioinformatics.*, Vol. 9, p.231.
- Liu, J. and Desmarais, M.C. (1997) '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.
- Lu, Y., Lemon, W., Liu, P.Y., Yi, Y., Morrison, C., Yang, P., Sun, Z., Szoke, J., Gerald, W.L., Watson, M., Govindan, R. and You, M. (2006) 'A gene expression signature predicts survival of patients with stage I non-small cell lung cancer', *PLoS. Med.*, Vol. 3, No. 12, p.e467.
- Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. and Alon, U. (2004) 'Superfamilies of evolved and designed networks', *Science*, Vol. 303, No. 5663, pp.1538–1542.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) 'Network motifs: simple building blocks of complex networks', *Science*, Vol. 298, No. 5594, pp.824–827.
- Naruke, T., Goya, T., Tsuchiya, R. and Suemasu, K. (1988) '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.
- Pe'er, D., Regev, A., Elidan, G. and Friedman, N. (2001) 'Inferring subnetworks from perturbed expression profiles', *Bioinformatics.*, Vol. 17, Suppl 1, pp.S215–S224.
- Potti, A., Mukherjee, S., Petersen, R., Dressman, H.K., Bild, A., Koontz, J., Kratzke, R., Watson, M.A., Kelley, M., Ginsburg, G.S., West, M., Harpole Jr., D.H. and Nevins, J.R. (2006) '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.
- Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. and Nolan, G.P. (2005) 'Causal protein-signaling networks derived from multiparameter single-cell data', *Science*, Vol. 308, No. 5721, pp.523–529.
- Sahoo, D., Dill, D.L., Gentles, A.J., Tibshirani, R. and Plevritis, S.K. (2008) 'Boolean implication networks derived from large scale, whole genome microarray datasets', *Genome Biol.*, Vol. 9, No. 10, p.R157.
- Shedden, K., Taylor, J.M., Enkemann, S.A., Tsao, M.S., Yeatman, T.J., Gerald, W.L., Eschrich, S., Jurisica, I., Giordano, T.J., Misek, D.E., Chang, A.C., Zhu, C.Q., Strumpf, D., Hanash, S., Shepherd, F.A., Ding, K., Seymour, L., Naoki, K., Pennell, N., Weir, B., Verhaak, R., Ladd-Acosta, C., Golub, T., Gruidl, M., Sharma, A., Szoke, J., Zakowski, M., Rusch, V., Kris, M., Viale, A., Motoi, N., Travis, W., Conley, B., Seshan, V.E., Meyerson, M., Kuick, R., Dobbin, K.K., Lively, T., Jacobson, J.W. and Beer, D.G. (2008) 'Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study', *Nat. Med.*, Vol. 14, No. 8, pp.822–827.
- Sotiriou, C. and Piccart, M.J. (2007) '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.
- Wan, Y.W., Sabbagh, E., Raese, R., Qian, Y., Luo, D., Denvir, J., Vallyathan, V., Castranova, V. and Guo, N.L. (2010) 'Hybrid models identified a 12-gene signature for lung cancer prognosis and chemoresponse prediction', *PLoS. ONE.*, Vol. 5, No. 8, p.e12222.

A novel network model identified a 13-gene lung cancer prognostic signature 39

- Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann.
- Wuchty, S., Oltvai, Z.N. and Barabasi, A.L. (2003) 'Evolutionary conservation of motif constituents in the yeast protein interaction network', *Nat. Genet.*, Vol. 35, No. 2, pp.176–179.
- Zhu, J., Zhang, B., Smith, E.N., Drees, B., Brem, R.B., Kruglyak, L., Bumgarner, R.E. and Schadt, E.E. (2008) 'Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks', *Nat. Genet.*, Vol. 40, No. 7, pp.854–861.

Notes

¹www.ingenuity.com

²<https://array.nci.nih.gov/caarray/project/details.action?project.id=182>

³<http://www.genome.jp/kegg/pathway/hsa/hsa05223.html>