# Analysis of Pathway Activity in Primary Tumors and NCI60 Cell Lines Using Gene Expression Profiling Data

Xing-Dong Feng[1], Shu-Guang Huang[2], Jian-Yong Shou[3], Bi-Rong Liao[1], Jonathan M. Yingling[3], Xiang Ye[3], Xi Lin[3], Lawrence M. Gelbert[3], Eric W. Su[1], Jude E. Onyia[1], and Shu-Yu Li[1]*

*[1] Integrative Biology, [2] Global Discovery & Development Statistics, [3] Cancer Discovery Research, Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, USA.*

**To determine cancer pathway activities in nine types of primary tumors and NCI60 cell lines, we applied an *in silico* approach by examining gene signatures reflective of consequent pathway activation using gene expression data. Supervised learning approaches predicted that the Ras pathway is active in ∼70% of lung adenocarcinomas but inactive in most squamous cell carcinomas, pulmonary carcinoids, and small cell lung carcinomas. In contrast, the TGF-$\beta$, TNF-$\alpha$, Src, Myc, E2F3, and $\beta$-catenin pathways are inactive in lung adenocarcinomas. We predicted an active Ras, Myc, Src, and/or E2F3 pathway in significant percentages of breast cancer, colorectal carcinoma, and gliomas. Our results also suggest that Ras may be the most prevailing oncogenic pathway. Additionally, many NCI60 cell lines exhibited a gene signature indicative of an active Ras, Myc, and/or Src, but not E2F3, $\beta$-catenin, TNF-$\alpha$, or TGF-$\beta$ pathway. To our knowledge, this is the first comprehensive survey of cancer pathway activities in nine major tumor types and the most widely used NCI60 cell lines. The "gene expression pathway signatures" we have defined could facilitate the understanding of molecular mechanisms in cancer development and provide guidance to the selection of appropriate cell lines for cancer research and pharmaceutical compound screening.**

**Key words: cancer pathways, gene expression profiling, supervised learning, classification**

## Introduction

Cancer is a genetic disease driven by mutations in three types of genes: oncogenes, tumor suppressors, and genome stability genes involved in DNA repair and mitotic processes (*1*). It has been estimated that three to seven mutations are required for the development of cancers (*2*). At the molecular level, these mutations drive the neoplastic process through deregulation of cellular pathways and biological processes that control cell fate, growth, differentiation, and survival. Mutations of oncogenes and tumor suppressors increase tumor cell number by stimulating cell proliferation and inhibiting differentiation and apoptosis pathways (*1*). For example, the activation of Ras proteins by mutations of the ras oncogene recruits the Raf kinase that subsequently activates transcription factors Fos and Jun through MAP kinase signaling pathways. Fos and Jun in turn form AP1 and up-regulate growth promoting genes (*2*). Mutations of different oncogenes or tumor suppressors have been associated with different cancer types, suggesting that specific pathways may be responsible for the development of specific cancers (*2*). Therefore, determining pathway activities in cancers is critical not only for understanding molecular mechanisms in tumor progression but also for designing targeted therapeutic strategies.

Cell lines derived from primary tumor tissues have provided a valuable tool for the understanding of cancer biology at the molecular level. Much of the knowledge that we have today on fundamental processes in cancer cells has largely depended on the use of cell lines (*3*). In addition, since cancer cell lines provide an unlimited source of malignant cells, they are widely used in screening for anti-cancer drugs. However, because cells cultured *in vitro* lack the overall tissue architecture and relevant microenvironment, and cells continuously maintained in culture may lose the attributes of the tumors from which they are derived (*4*), the value of cancer cell lines is limited by the extent to which they represent the primary tumors'

***Corresponding author.**
**E-mail: li_shuyu_dan@lilly.com**

origin and activities. Several approaches have been utilized to characterize cancer cell lines. The ability to form tumors when cell lines were transplanted subcutaneously into nude mice allows a direct comparison of histopathology between tumors formed in nude mice and the human tumors of origin (*5*). Efforts have been made to delineate morphological features of cell lines in comparison with archival tumor tissues that the cell lines are derived from (*6*, *7*). At molecular levels, expression of key proteins such as HER2/neu and p53 in breast and non-small cell lung cancer cell lines as well as their corresponding tumors have been assessed using immunohistochemistry (*6*, *7*). Previously, we carried out a direct comparison between NCI60 cell lines and 9 primary tumor types using gene expression profiling data generated from more than 500 primary tumor samples (*8*). Our computational analysis suggested that 51 of the 59 NCI60 cell lines represent their presumed tumors of origin. These cell lines were also classified into tumor subtypes or different stages in cancer development (*8*). However, it remains unclear that what pathways are activated in each of these cell lines. Therefore, further analysis of pathway activation status in cancer cell lines could provide guidance to the selection of cell lines as appropriate models for studying cancer pathways and for target-based drug screening.

DNA microarray technology has created a new paradigm for understanding cancer biology by simultaneous measurement of tens of thousands of genes in malignant or normal cells. Gene expression profiles have been utilized to identify gene signatures that are associated with tumor progression and alterations in cancer pathways. Recently, gene expression signatures have been identified to reflect the activities of five oncogenic pathways, namely Ras, Myc, Src, E2F3, and $\beta$-catenin (*9*). These signatures derived from primary cell cultures have been validated in transgenic animal models and are correlated with sensitivity to therapeutic agents targeting specific pathways (*9*). Here we exploited the gene expression signatures for these five oncogenic pathways and two receptor-mediated signaling pathways, namely transforming growth factor (TGF)-$\beta$ and tumor necrosis factor (TNF)-$\alpha$, to predict pathways in nine major types of primary cancers and NCI60 cell lines. Supervised learning-based prediction suggested that different pathways are involved in the development of different tumor types. Moreover, our assessment of pathway activation status in NCI60 cells highlights the value of specific cell lines in studying these pathways and their roles in oncogenesis.

# Results

## Developing gene signatures representing active pathways and building supervised models for classification

We used gene expression profiles generated in primary mammary epithelial cell cultures (*9*) to derive signatures for the activated Ras, Myc, Src, E2F3, or $\beta$-catenin pathways. The training dataset includes two groups, cells transfected with adenovirus expressing green fluorescent protein (GFP) or one of the oncogenes. Gene signatures for the activated TGF-$\beta$ or TNF-$\alpha$ pathways were identified using gene expression profiles of TGF-$\beta$ or TNF-$\alpha$ treated by a non-small cell lung cancer cell line Calu6 or of the vehicle control (Yingling and Ye, unpublished results). Two criteria were considered in our selection of signature gene sets for the pathways. First, several candidate signatures were determined, which would give rise to a minimal cross validation error rate. Second, from multiple signature gene sets that satisfy a threshold of cross validation error rates, we selected the one with the smallest number of genes. As a result, there is limited overlap between the gene signatures for different pathways. Unlike the previous study on the five oncogenic pathways where authors built gene classifiers that are overlapping between different pathways (*9*), we believe our approach has generated signature gene sets that are more specific for each pathway and may provide more accurate predictions. Lists of genes selected for subsequent principle component analysis (PCA) and classification are provided in Supporting Online Material (Table S1). Many of these genes are known downstream targets for each of the pathways.

To predict what pathways are active in each of the primary tumor samples and NCI60 cell lines, we used supervised learning approaches (Figure 1). After gene features were selected from the training dataset, supervised predictors were built using a support vector machine (SVM) algorithm. Parameters were adjusted in model building to ensure minimal leave-one-out cross validation (LOOCV) error rates. Table S2 illustrates an example of this process for the Ras pathway. Analysis of variance was carried out to identify genes differentially expressed between the two groups in the training dataset, that is, cells transfected with adenovirus expressing GFP or the activated H-Ras.

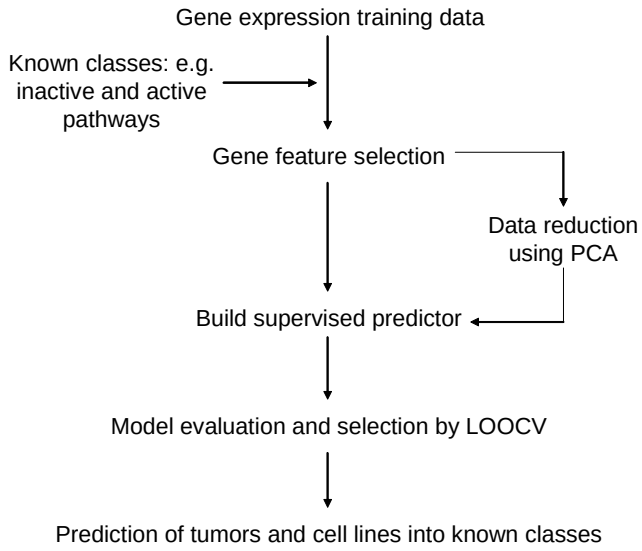



**Fig. 1** Feature classification using supervised learning methods. PCA: principal component analysis; LOOCV: leave-one-out cross validation.

Figure 2A clearly depicts a completely opposite expression pattern of these genes in the control group and in the group with a constitutively active Ras pathway. Then the data reduction using PCA and the subsequent building of classification models were carried out. Multiple models were evaluated using different numbers of principle components, different SVM kernel functions, and different *cost* parameters. Based on the criteria described in Materials and Methods, we chose three principle components as the discriminants, the Sigmoid kernel function, and a *cost* parameter of 8 that gave rise to the optimal error rate in LOOCV. Supervised models for other pathways were also built and tested using the same approach (data not shown).

## Classification of primary cancers

We first attempted to classify lung cancers into an active vs. inactive status for each pathway. The testing gene expression profiling data were previously published using 186 primary lung cancer samples, including 139 adenocarcinomas, 21 squamous cell lung carcinomas, 20 pulmonary carcinoids, and 6 small cell lung cancers (*10*) (Table 1). Our prediction results (Table 2) suggest that the Ras pathway is activated in almost 70% of lung adenocarcinoma patients, but is inactive in most squamous cell carcinomas, pulmonary carcinoids, and small cell lung carcinomas. In contrast, the Src, Myc, E2F3, $\beta$-catenin, TGF-$\beta$, and TNF-$\alpha$ pathways are inactive in almost all of the lung adenocarcinomas. Figure 2B is a graphic illustration of gene expression patterns in lung cancers with an active or inactive Ras pathway. It is noteworthy that differential expression of these signature genes in active vs. inactive primary tumors (Figure 2B) has less magnitude than that observed in the primary cell cultures (Figure 2A), raising the possibility that subtle changes in the pathways may be sufficient to trigger tumorigenesis. An alternative explanation is that tumor biopsy samples often contain a certain percentage of tumor cells and other non-tumor cell types. Therefore, gene expression patterns in tumors are mixed with noise from non-tumor cells. Significant numbers of pulmonary carcinoid and small cell lung cancer samples exhibited a gene signature representing an active E2F3 pathway (Table 2). Our prediction of the activity status of the Ras pathway in lung adenocarcinomas and squamous cell lung carcinomas using the dataset from Bhattacharjee *et al* (*10*) is consistent with the results reported by Bild and colleagues based on a different cohort of patients (*9*).

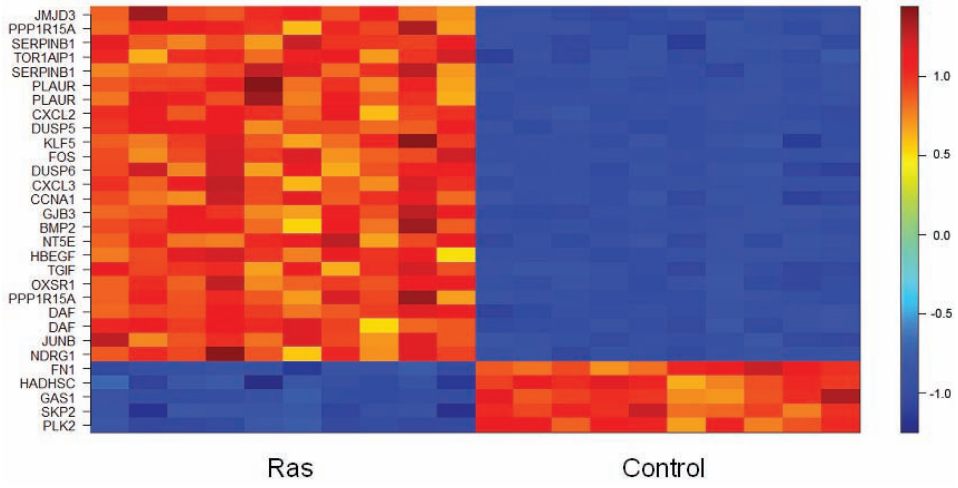
**Table 1 Gene expression profiling datasets on NCI60 cell lines and primary tumors analyzed in this study**

| Cancer type | Sample size | Data format | URL for data downloading | Ref. |
|---|---|---|---|---|
| NCI60 cell lines | – | MAS5 | http://dtp.nci.nih.gov/mtargets/madownload.html | – |
| Lung | 186 | MAS5 | http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi | *10* |
| Prostate | 52 | MAS5 | http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi | *25* |
| Leukemia | 72 | MAS5 | http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi | *26* |
| CNS | 50 | MAS5 | http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi | *27* |
| Melanoma | 29 | MAS5 | http://www.mskcc.org/genomic/ccsmsp/ | *28* |
| Breast | 171 | MAS5 | http://data.cgt.duke.edu/oncogene.php | *9* |
| Ovary | 146 | MAS5 | http://data.cgt.duke.edu/oncogene.php | *9* |
| Colon | 23 | MAS4 | http://www.gnf.org/cancer/epican/ | *29* |
| Kidney | 11 | MAS4 | http://www.gnf.org/cancer/epican/ | *29* |

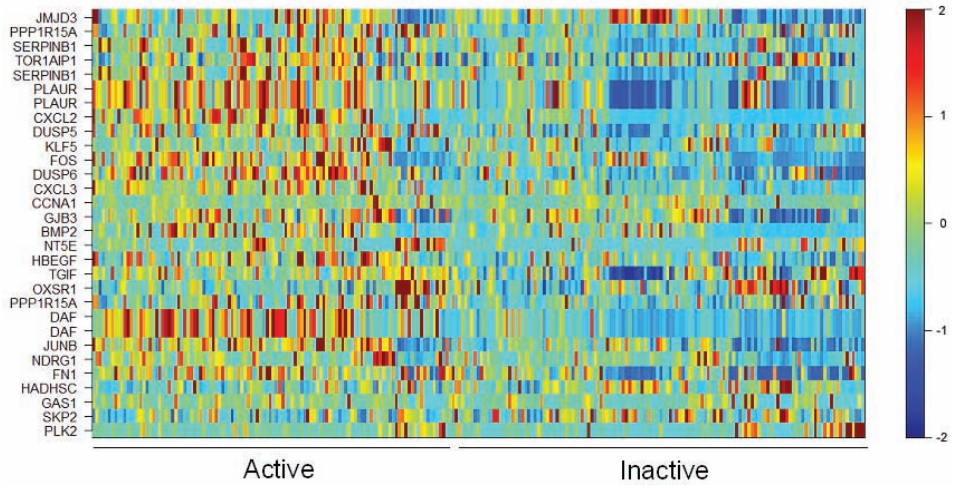
**Table 2 Pathway activity in lung cancers***

| Pathway | Adenocarcinoma | Squamous cell carcinoma | Pulmonary carcinoid | Small cell lung cancer |
|---|---|---|---|---|
| Ras | **0.683** (95/139) | 0.14 (3/21) | 0 (0/20) | 0.17 (1/6) |
| Myc | 0.029 (4/139) | 0 (0/21) | 0 (0/20) | 0.17 (1/6) |
| Src | 0.029 (4/139) | 0 (0/21) | 0 (0/20) | 0 (0/6) |
| E2F3 | 0.022 (3/139) | 0 (0/21) | **0.40** (8/20) | 0.50 (3/6) |
| $\beta$-catenin | 0 (0/139) | 0 (0/21) | 0 (0/20) | 0 (0/6) |
| TGF-$\beta$ | 0 (0/139) | 0 (0/21) | 0 (0/20) | 0 (0/6) |
| TNF-$\alpha$ | 0.065 (9/139) | 0 (0/21) | 0 (0/20) | 0 (0/6) |

*The percentages of patients with predicted active pathways are shown. The numbers in parentheses are the numbers of patients with active pathways vs. the total numbers of patient samples in each subtype of lung cancers. Bolded numbers indicate a significant percentage ($> 20\%$ for sample size $\geq 20$) of samples exhibiting a gene signature of active pathways.




**A**




**B**

**Fig. 2** Classification of primary lung cancers and NCI60 cell lines with respect to active vs. inactive Ras pathways. **A**. A 30-gene signature developed from the training dataset for the Ras pathway. Red and blue represent high and low levels of expression respectively. The y-axis represents the 30 genes and the x-axis represents two groups in the training dataset, that is, cells transfected with adenovirus expressing the activated H-ras or GFP as a control. **B**. Gene expression patterns of the signature genes in 186 lung cancers and 59 NCI60 cell lines with an activated or inactive Ras pathway.

We next carried out classification of other tumor types with publicly available oligonucleotide microarray data (Table 1). Analysis results (Table 3) reveal that Ras may be the most prevailing oncogenic pathway, since gene expressions in significant amount of tumor samples in each cancer type are indicative of an active Ras pathway according to our computational prediction. Different cancer types, however, behave differently with respect to activities of other pathways. For example, while Ras is the only active pathway in lung adenocarcinomas, we predicted an active Ras, Myc, Src, and E2F3 pathway in 73%, 70%, 21%, and 30% of breast cancer patients, respectively. Upon further investigation, 74%, 67%, and 69% of Myc, Src, and E2F3 active samples, respectively, also have an active Ras pathway, suggesting multiple oncogenic pathways may coordinately promote breast cancer progression in these patients. The observations of multiple and overlapping activated pathways in breast tumors reflect the heterogenous nature of cancer. An active status in multiple pathways has also been predicted in brain, colon, kidney, and ovarian cancers. In contrast, Ras is the only active pathway in leukemia, melanoma, and prostate cancers, which is similar to what was observed in lung adenocarcinomas. Except for ovarian cancers, the $\beta$-catenin, TGF-$\beta$, and TNF-$\alpha$ pathways are inactive in almost all of the primary tumors. Collectively, these results substantiate the notion that different pathways may play critical roles in the development of different cancer types.

## Classification of NCI60 cell lines

NCI60 represents the most commonly used cancer cell lines in cancer research and drug screening. In order to evaluate them as models for primary tumors, we estimated pathway activities in NCI60 cell lines using the supervised learning-based classification. Listed in Table 4 are the cell lines with predicted active pathways. Although these results await further experimental validation, they could provide directions to the selection of specific cell lines to study specific pathways in cancer cells. Even though most of the NCI60 cell lines were suggested to represent their corresponding tumor origin (*8*), we postulate that distinct pathways are active in each of these cell lines according to our *in silico* analysis. For example, except for NCI/ADR-RES, all of the breast cell lines in the NCI60 panel have global gene expression profiles more similar to that of primary breast cancers than other tumor types (*8*), but their expression patterns for pathway specific gene signatures are different. BT-549, MDA-MB-231, and HS578 exhibited an active expression signature for the Ras pathway, and MCF7 is the only line that we predicted to have an active Src pathway (Table 4). Interestingly, many cell lines

**Table 3 Pathway activity in other primary cancers***

| Pathway | Breast cancer | CNS cancer | Colon cancer | Kidney cancer | Leukemia | Melanoma | Ovarian cancer | Prostate cancer |
|---|---|---|---|---|---|---|---|---|
| Ras | **0.73** | **0.44** | **0.43** | **0.72** | **0.36** | **0.31** | **0.48** | **0.50** |
| | (125/171) | (22/50) | (10/23) | (8/11) | (26/72) | (9/29) | (70/146) | (26/52) |
| Myc | **0.70** | **0.42** | **0.35** | 0.091 | 0.055 | 0.14 | **0.22** | 0.19 |
| | (120/171) | (21/50) | (8/23) | (1/11) | (4/72) | (4/29) | (32/146) | (10/52) |
| Src | **0.21** | **0.56** | **0.91** | **0.91** | 0.014 | 0.069 | 0 | 0.019 |
| | (36/171) | (28/50) | (21/23) | (10/11) | (1/72) | (2/29) | (0/146) | (1/52) |
| E2F3 | **0.30** | 0.12 | 0.043 | 0 | 0.17 | 0.17 | **0.28** | 0.17 |
| | (51/171) | (6/50) | (1/23) | (0/11) | (12/72) | (5/29) | (41/146) | (9/52) |
| $\beta$-catenin | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | (0/171) | (0/50) | (0/23) | (0/11) | (0/72) | (0/29) | (0/146) | (0/52) |
| TGF-$\beta$ | 0 | 0 | 0 | 0 | 0 | 0 | **0.82** | 0 |
| | (0/171) | (0/50) | (0/23) | (0/11) | (0/72) | (0/29) | (120/146) | (0/52) |
| TNF-$\alpha$ | 0.058 | 0.04 | 0.087 | 0 | 0.055 | 0.10 | 0 | 0.038 |
| | (10/171) | (2/50) | (2/23) | (0/11) | (4/72) | (3/29) | (0/146) | (2/52) |

*The percentages of patients with predicted active pathways are shown. The numbers in parentheses are the numbers of patients with active pathways vs. the total numbers of patient samples in each cancer type. Bolded numbers indicate a significant percentage ($> 20\%$) of samples exhibiting a gene signature of active pathways.

**Table 4 NCI60 cell lines with predicted active pathways***

| Tumor type | Ras | Myc | Src | E2F3 | TNF-$\alpha$ |
|---|---|---|---|---|---|
| Breast | BT-549, MDA-MB-231, HS578T | MDA-MB-435, BT-549, NCI/ADR-RES | MCF7 | – | MDA-MB-231, HS578T |
| CNS | – | SF-268 | – | – | SF-268 |
| Colon | HT-29, COLO205, HCT-15, KM12, HCT-116 | COLO205, KM12, HCT-116, SW-620 | KM12, HCC-2998 | HCT-15 | – |
| Kidney | 786-0 | – | RXF-393, 786-0 | – | – |
| Leukemia | – | RPMI-8226, CCRF-CEM, K-562, MOLT-4, HL-60 | RPMI-8226, SR, K-562, HL-60 | CCRF-CEM, MOLT-4 | – |
| Lung | NCI-H460, NCI-H23, NCI-H522, HOP-92 | NCI-H460, EKVX, NCI-H522 | NCI-H23 | EKVX, NCI-H522 | – |
| Melanoma | LOX IMVI, UACC-257, SK-MEL-28 | LOX IMVI, UACC-62, SK-MEL-2, SK-MEL-5 | LOX IMVI, UACC-62, UACC-257, SK-MEL-5 | – | – |
| Ovary | OVCAR-5 | IGROV1, OVCAR-4, OVCAR-8 | IGROV1, OVCAR-8 | – | – |
| Prostate | PC3, DU-145 | PC3 | PC3 | – | – |

*The $\beta$-catenin and TGF-$\beta$ pathways were predicted to be inactive in all of the cell lines and thus are omitted in the table.

are active in only one or two pathways. This is not unexpected given the homogeneity of the cultured cells due to clonal selection.

## Discussion

Although genome-wide expression profiling has become a mainstay in cancer research, it remains a challenge to extract biological insight from gene expression data. In a typical experiment, individual genes are identified according to their differential expression between the control group and the experimental group, followed by mapping of these genes to biological pathways. However, it has been demonstrated that a biological pathway could play a significant role in physiological processes even though each gene in the pathway only exhibits subtle gene expression changes to external perturbations but collectively they exert significant impact to the cells (*11*). Several algorithms have been proposed to analyze expression data focusing on pathways rather than on individual genes (*12*, *13*). However, before we consider applying these methods to cancer microarray data, two issues need to be addressed. First, it has been a common practice to measure pathway activity by analyzing expression of genes involved in signal transduction. We believe this approach is problematic in studying signaling pathways in cancers. Activation of those pathways often involves post-translational modification of proteins in the signaling but does not depend on an increased expression of genes encoding those proteins. A more sensitive and robust approach would be interrogating downstream genes, that is, gene expression changes that reflect pathway acti-

vation. Second, the frequently used computational methods for pathway analysis compare gene expression patterns between the control group (such as normal tissues) and the experimental group (such as cancerous cells). Given the variability between individuals and limited sample sizes typical of human studies, it could be difficult to distinguish true difference from noise. In this study, we developed a strategy to overcome the above mentioned shortcomings in the current methodology. Gene signatures for the seven pathways were developed from experimental data. Alterations in signature gene expression are associated with and can be used as a direct "readout" of pathway activation. Furthermore, we applied supervised learning methods to predict pathway status in individual samples, which should provide more accurate and sensible results.

Computational analysis requires laboratory experimentation to validate the results. Some of our predictions have already been confirmed by experimental data reported in the literature. It shows that 68% of lung adenocarcinomas exhibited a gene expression signature of active Ras pathways (Table 2). This is consistent with the finding that PCR-based method has detected ras mutations in non-small cell lung cancers at frequencies that may exceed 50% (*14*). We predicted an active Src pathway in the majority (21 of 23, 91%) of colorectal carcinoma samples (Table 3), which is supported by studies that described over expression of c-src and deregulation of the Src pathway in more than 70% of human colon cancers (*15*). Previously, gene amplification has been examined in glioblastomas using an array-based comparative genomic hybridization, and Myc amplification was detected in 42% of the samples (*16*). This again is consistent with our *in silico* pathway analysis indicating 42% of the gliomas have an active Myc pathway. Gene expression patterns in 70% of breast cancer samples represent an active Myc pathway (Table 3). This is not surprising since immunohistochmistry has detected over expression of c-Myc proteins in 45% of 440 primary breast carcinomas (*17*). Even though some oncogenes are not mutated or amplified in certain cancer types, it is still possible that the oncogenic pathways are active in these cancers through other mechanisms. For example, we report here that half of the prostate cancers may have an activated Ras pathway, yet it has been well documented in the literature that ras mutations are rare in prostate cancers (*18–20*). However, in a very recent study Ras downstream MAP kinase activity in prostate cancers was investigated using immunohistochemistry for p44/ERK1 and p42/ERK2, and active MAPK signaling was detected in 51% of the analyzed tumors (*21*), strikingly similarly to our predictions. High and low frequencies of an activated Ras pathway in lung adenocarcinomas and squamous lung cell carcinomas respectively reported by us in this study are in agreement with recent results also based on computational prediction but using gene expression data generated from a different cohort of patients (*9*). Taken together, these evidence strongly supports our approach to examine pathway activity using gene expression profiling data.

We also recognize the limitations in our study. First, gene signatures were developed from an *in vitro* system where the pathways were experimentally activated. The differential expressions of the signature genes are augmented artificially. While the control group and the experimental group in the training dataset can be clearly defined into two classes, there is a significantly greater variability of pathway activity in primary tumors. Therefore, our prediction of a pathway in cancers into either the inactive or the active status is rather arbitrary. Second, gene signatures were derived from data using the primary mammary epithelial cells or non-small cell lung cancer cells Calu6. However, downstream genes regulated by these pathways could be cell type specific. As a result, using the gene signature identified from one cell type to predict pathway activity in other cell types may cause high rate of false negatives. Third, although mutations occur primarily in tumor cells, some pathways play a pivotal role in non-tumor cells to provide a microenvironment for promoting tumor progression and angiogenesis. For example, an activated TGF-$\beta$ pathway creates a favorable microenvironment for tumor growth and invasion (*22*). The effects of TGF-$\beta$ pathway activation is mainly executed in tumor microenvironment but not in tumor cells. Consequently, the importance of TGF-$\beta$ pathway in cancer development should not be undermined even though it is in an inactive state in primary tumor cells. Fourth, one of our main goals is to predict pathway activation status in NCI60 cell lines. An inactive pathway in a cell line, however, only indicates a low baseline activity and does not necessarily exclude the cell line as an ideal model to study the pathway. In fact, TGF-$\beta$ target genes in Calu6 cells are expressed at minimal levels but are robustly up-regulated in response to the TGF-$\beta$ ligand. Accordingly, if a cell line has intact signaling components of a pathway and responds to ligand stimulation, it should be still considered as

a good model system even the basal pathway activity is minimal. Finally, gene expression profiles in cell culture *in vitro* may not reflect gene expressions evaluated when cells are grown *in vivo*, as evidenced by a recent study that although two glioblastoma cell lines (U251 and U87) have disparate gene expression profiles when grown in monolayer cell cultures, they share similar gene expression patterns when grown as intracerebral xenografts in nude mice (*23*). Therefore, the next level approach to evaluate cell lines would be using gene expression profiles of cell lines grown in xenograft models when such data become available. Nevertheless, we believe that with more gene expression profiling studies being carried out, gene signatures for more pathways can be developed in multiple cell types. Our computational approach in predicting pathway activities provides a valuable tool that can be generally applied to studying biological pathways under normal and pathological conditions.

# Materials and Methods

## Data source

The gene expression profiling data on NCI60 cell lines provided by NCI's DTP program (http://dtp.nci.nih.gov/mtargets/madownload.html) are based on Affymetrix U95Av2 oligonucleotide array platforms. While oligonucleotide arrays measure the amount of mRNA in a single sample, gene expression data generated using cDNA array platforms are ratios of expression values in experimental samples over those in a reference sample. The fundamental difference between the two array platforms poses a technical barrier in integrative analysis of gene expression data based on these two different platforms. Therefore, we chose only Affymetrix oligonucleotide array-based data in publicly available gene expression profiling databases on primary tumors (Table 1).

Gene expression data on NCI60 cell lines and primary tumor samples were downloaded from the URL addresses shown in Table 1. Gene expression data of lung, prostate, central nervous system (CNS) cancers, and leukemia were originally generated with Affymetrix MAS4 software. The breast and ovarian cancer datasets were in gcRMA format. We downloaded the .cel files and analyzed them using Affymetrix MAS5 algorithm with trimmed mean values normalized to 500. A trimmed mean is the average value after removing the lowest 2% and the highest 2% of all expression values. The downloaded array data for NCI60 cell lines and melanomas were in MAS5 format and we re-normalized the data by setting the trimmed means to 500. Data were only available for 59 of the NCI60 cell lines. For colon and kidney cancers, we were only able to obtain MAS4 gene expression data and similarly, these data were normalized with trimmed means equal to 500.

We compiled the gene expression data for a total of 799 samples after averaging the expression values over the technical replicates in the lung dataset and in NCI60 cell lines. Expression of each probe set was standardized to a mean of 0 and standard deviation of 1. The standardization procedure is performed for the training dataset and the testing dataset separately.

## Feature selection and classification

All statistical analysis was implemented using SAS and R statistical languages. Pair-wise t-tests were used to identify genes differentially expressed between the control group and the active pathway group in the training dataset. Probe sets were ranked by $p$-values. Multiple $p$-values were tested as thresholds to select gene features for subsequent PCA and classifications. For each pathway, we chose a cutoff $p$-value that gave rise to minimal cross validation error rates in classifications (see below).

SVM was used as the classification method (*24*). We performed PCA after gene features were selected. The number of principal components $p$ and the parameter *cost* that corresponding to constant of the regularization term in the Lagrange formulation was determined based on LOOCV error rate. LOOCV is a procedure in which we trained classifier based on the training dataset after one object is removed and the classifier was tested on the removed one. The procedure was implemented for each object in the dataset and the proportion of errors counted throughout the process is called LOOCV error rate. The value $p$ and the *cost* parameter were chosen to be the smallest one that satisfies two criteria: (1) the LOOCV error rate of the classifier is smaller than 0.05; and (2) the three consecutive classifiers built on the features with $p$, $p$+1, and $p$+2 principal components give consistent predictions. Four most commonly used SVM kernel functions were tested: linear, polynomial with degree 3, radial bases, and neural network. Similarly, we chose a kernel function that minimizes the LOOCV error rate for analysis of each pathway.

## Authors' contributions

SGH, JYS, BRL, and SYL designed the study. XDF and SYL carried out data analysis. JMY, XY, XL, and LMG generated microarray data for the TGF-$\beta$ and TNF-$\alpha$ pathways. XDF, SGH, JTS, JEO, and SYL interpreted the results. XDF and SYL drafted the manuscript. SGH, JYS, BRL, XY, LMG, EWS, and JEO revised the manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors have declared that no competing interests exist.

# References


1. Vogelstein, B. and Kinzler, K.W. 2004. Cancer genes and the pathways they control. *Nat. Med.* 10: 789-799.
2. Vogelstein, B. and Kinzler, K.W. (eds.) 2002. *The Genetic Basis of Human Cancers* (second edition). McGraw-Hill Companies, Inc., New York, USA.
3. Masters, J.R. 2002. HeLa cells 50 years on: the good, the bad and the ugly. *Nat. Rev. Cancer* 2: 315-319.
4. Kamb, A. 2005. What's wrong with our cancer models? *Nat. Rev. Drug Discov.* 4: 161-165.
5. Fogh, J., *et al.* 1977. One hundred and twenty-seven cultured human tumor cell lines producing tumors in nude mice. *J. Natl. Cancer Inst.* 59: 221-226.
6. Wistuba, II, *et al.* 1998. Comparison of features of human breast cancer cell lines and their corresponding tumors. *Clin. Cancer Res.* 4: 2931-2938.
7. Wistuba, II, *et al.* 1999. Comparison of features of human lung cancer cell lines and their corresponding tumors. *Clin. Cancer Res.* 5: 991-1000.
8. Wang, H., *et al.* 2006. Comparative analysis and integrative classification of NCI60 cell lines and primary tumors using gene expression profiling data. *BMC Genomics* 7: 166.
9. Bild, A.H., *et al.* 2006. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439: 353-357.
10. Bhattacharjee, A., *et al.* 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* 98: 13790-13795.
11. Mootha, V.K., *et al.* 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34: 267-273.
12. Pang, H., *et al.* 2006. Pathway analysis using random forests classification and regression. *Bioinformatics* 22: 2028-2036.
13. Subramanian, A., *et al.* 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 102: 15545-15550.
14. Clements, N.C. Jr., *et al.* 1995. Analysis of K-ras gene mutations in malignant and nonmalignant endobronchial tissue obtained by fiberoptic bronchoscopy. *Am. J. Respir. Crit. Care Med.* 152: 1374-1378.
15. Griffiths, G.J., *et al.* 2004. Expression of kinase-defective mutants of c-Src in human metastatic colon cancer cells decreases Bcl-xL and increases oxaliplatin- and Fas-induced apoptosis. *J. Biol. Chem.* 279: 46113-46121.
16. Hui, A.B., *et al.* 2001. Detection of multiple gene amplifications in glioblastoma multiforme using array-based comparative genomic hybridization. *Lab. Invest.* 81: 717-723.
17. Naidu, R., *et al.* 2002. Protein expression and molecular analysis of c-myc gene in primary breast carcinomas using immunohistochemistry and differential polymerase chain reaction. *Int. J. Mol. Med.* 9: 189-196.
18. Carter, B.S., *et al.* 1990. ras gene mutations in human prostate cancer. *Cancer Res.* 50: 6830-6832.
19. Gumerlock, P.H., *et al.* 1991. Activated ras alleles in human carcinoma of the prostate are rare. *Cancer Res.* 51: 1632-1637.
20. Moul, J.W., *et al.* 1992. Infrequent RAS oncogene mutations in human prostate cancer. *Prostate* 20: 327-338.
21. Burger, M., *et al.* 2006. Mitogen-activated protein kinase signaling is activated in prostate tumors but not mediated by B-RAF mutations. *Eur. Urol.* 50: 1102-1109.
22. Yingling, J.M., *et al.* 2004. Development of TGF-beta signalling inhibitors for cancer therapy. *Nat. Rev. Drug Discov.* 3: 1011-1022.
23. Camphausen, K., *et al.* 2005. Influence of *in vivo* growth on human glioma cell line gene expression: convergent profiles under orthotopic conditions. *Proc. Natl. Acad. Sci. USA* 102: 8287-8292.
24. Hastie, T., *et al.* 2001. *The Elements of Statistical Learning.* Springer-Verlag, New York, USA.
25. Singh, D., *et al.* 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1: 203-209.
26. Armstrong, S.A., *et al.* 2002. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* 30: 41-47.
27. Nutt, C.L., *et al.* 2003. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.*

63: 1602-1607.

28. Segal, N.H., *et al.* 2003. Classification of clear-cell sarcoma as a subtype of melanoma by genomic profiling. *J. Clin. Oncol* 21: 1775-1781.

29. Su, A.I., *et al.* 2001. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.* 61: 7388-7393.

**Supporting Online Material**

https://netfiles.uiuc.edu/xfeng2/shared/
Tables S1 and S2