Research Article

Cancer Epidemiology, Biomarkers & Prevention

A P53-Deficiency Gene Signature Predicts Recurrence Risk of Patients with Early-Stage Lung Adenocarcinoma

Yanding Zhao[1], Frederick S. Varn[1], Guoshuai Cai[1], Feifei Xiao[2], Christopher I. Amos[1,3,4], and Chao Cheng[1,3,4]

Abstract

Background: Lung cancer is associated with the highest mortality rate of all cancer types, and the most common histologic subtype of lung cancer is adenocarcinoma. To apply more effective therapeutic treatment, molecular markers that are able to predict the recurrence risk of patients with adenocarcinoma are critically needed. Mutations in *TP53* tumor suppressor gene have been found in approximately 50% of lung adenocarcinoma cases, but the presence of a *TP53* mutation does not always associate with increased mortality.

Methods: The Cancer Genome Atlas RNA sequencing data of lung adenocarcinoma were used to define a novel gene signature for P53 deficiency. This signature was then used to calculate a sample-specific P53 deficiency score based on a patient's transcriptomic profile and tested in four independent lung adenocarcinoma microarray datasets.

Results: In all datasets, P53 deficiency score was a significant predictor for recurrence-free survival where high P53 deficiency score was associated with poor survival. The score was prognostic even after adjusting for several key clinical variables including age, tumor stage, smoking status, and P53 mutation status. Furthermore, the score was able to predict recurrence-free survival in patients with stage I adenocarcinoma and was also associated with smoking status.

Conclusions: The P53 deficiency score was a better predictor of recurrence-free survival compared with P53 mutation status and provided additional prognostic values to established clinical factors.

Impact: The P53 deficiency score can be used to stratify early-stage patients into subgroups based on their risk of recurrence for aiding physicians to decide personalized therapeutic treatment. *Cancer Epidemiol Biomarkers Prev; 27(1); 86–95. ©2017 AACR.*

Introduction

Lung cancer is the leading cause of all cancer-associated deaths in the United States, accounting for approximately a quarter of all cancer mortalities (1). It is associated with poor prognosis, with only 17.4% of lung cancer patients surviving over 5 years after diagnosis (2). Lung adenocarcinoma (LUAD) makes up 40% of lung cancers, making adenocarcinoma the most common histologic subtype of lung cancers (3). Around 10% to 25% of LUAD patients are never-smokers (4). A primary reason for the high mortality of lung cancer is that only 25% of cases are diagnosed at an early stage, when surgery is most effective (5). Indeed, surgical resection of stage I non–small cell lung cancer (NSCLC) results in 5-year survival rates of approximately 70% (6–9). As screening for lung cancer becomes more widely adopted in medical practice, a larger proportion of lung cancer cases will be detected at an early stage (10–12).

LUAD cases are heterogeneous in many aspects, including histopathologic patterns, molecular features, and driver mutations. For example, approximately 30% to 60% of LUAD cases are associated with mutations in *EGFR*, *KRAS*, or *ALK* genes. This heterogeneity likely explains the significant variation in prognosis among LUAD patients, with survival time ranging from a few months to more than 7 years (13–16). Interestingly, about 15.7% of stage I lung cancer patients with complete surgical resection have reported cancer recurrence (17). Thus, it is critical to develop methods for predicting patient-specific prognosis so that the most effective therapeutic and management strategies can be designed for distinct subsets of LUADs. For example, some early-stage lung cancers are highly aggressive at the time of diagnosis and should be treated more aggressively (18–20). Indeed, it has been suggested that some stage I lung cancer patients have developed occult micro-metastases, via which they would develop disease recurrence (21).

Gene expression profiling has been widely applied to investigate the transcriptional regulation underlying lung cancer development and progression. Several gene signatures have been defined to predict the recurrence of patients with early-stage NSCLC (22–27). For example, Beer and colleagues identified 50 genes most of which are highly correlated with survival of 86 lung cancer cases, and found that a risk index calculated based on these 50 genes predicted the risk of patients in stage I cancers

[1]Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire. [2]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, South Carolina. [3]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire. [4]Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Corresponding Author: Chao Cheng, Geisel School of Medicine at Dartmouth, One Medical Center Drive, Lebanon, NH 03756. Phone: 603-653-9032; Fax: 603-650-1188; E-mail: chao.cheng@dartmouth.edu

doi: 10.1158/1055-9965.EPI-17-0478

(26). In another study, Chen and colleagues developed a five-gene signature that was predictive of survival outcome in NSCLC, but the risk strata were highly correlated with stage, and stage was far more associated with survival than the expression profile (28). Importantly, a multisite blinded validation study was performed to examine the performance of several prognostic models based on gene expression alone or in combination with clinical variables (25). This study found that most models achieved better prognosis prediction accuracy when using samples containing all stages compared to using just stage I samples, suggesting their ability to discriminate stages even when stage information was not included in the model. It should be noted that, without including clinical variables, none of the models achieved a significant hazard ratio (HR) in the two validation datasets for stage I lung cancer samples.

The tumor suppressor gene *TP53*, which encodes tumor suppressor protein P53, is the most frequently mutated gene in human cancers, and is associated with more than 50% of tumor cases (29). In lung cancer, somatic mutation frequency of *TP53* varies between different histopathologic subtypes. The highest rates of *TP53* mutation have been observed in small cell lung cancer (>90%; ref. 30) and squamous cell carcinomas (81%; ref. 31), which are the subtypes that are most consistently associated with long-term smoking. In LUAD, 46% of cases have somatic mutations in *TP53* (31). Notably, LUAD patients who are current or former smokers have been shown to have higher *TP53* mutation rates and significantly higher somatic mutation burdens in their tumors compared to patients who are never-smokers (32, 33). In addition, very different *TP53* mutation frequency and mutation types have been observed between current or former smokers and never-smokers in LUAD (33). Despite the high frequency of this mutation, the prognostic value of *TP53* mutation status is still not clear in lung cancer, even though such information is commonly available for lung cancer patients (30). Previous studies have reported inconsistent or even controversial findings on the prognostic value of *TP53* mutation status (34–36).

In this study, we developed a gene signature to quantify P53 pathway activity in LUAD samples by comparing gene expression between P53-mutant and wild-type samples. Our rationale is that P53 pathway activity can more accurately reflect the aggressiveness of cancer cells and is thus a better prognostic indicator than P53 mutation status. Although most nonsynonymous mutations of *TP53* result in aberrant P53 activity in a dominant manner (37), the effects of different mutations vary significantly. Thus, the full impact of mutations cannot be captured by a binary indicator like P53 mutation status. In addition, loss of P53 activity can be caused by other mechanisms, including DNA hypermethylation in the promoter of *TP53* (38, 39), deletion of *TP53* (40), or indirectly, by the dysregulation of P53 regulators (41, 42). Here, we use a *TP53* nonsynonymous mutations-based gene signature to calculate P53 deficiency scores (PDS) in LUAD samples. Our results indicate that PDS can reliably and significantly predict the rate of recurrence for early-stage LUAD patients.

Materials and Methods

LUAD gene expression datasets

The RNA sequencing (RNA-seq) data for LUAD samples generated by The Cancer Genome Atlas (TCGA) project were used to define a gene signature for characterizing P53 pathway activity. The data were downloaded from FireHose (http://gdac.broadinstitute.org/) as Level 3 processed RNA-seq data in November 2016. The data contain gene expression profiles for a total of 515 tumor samples, and provide the Reads Per Kilobase per Million mapped reads (RPKM) for 20,502 genes. In addition, somatic mutation and clinical information associated with these samples were also downloaded.

We used six lung cancer gene expression datasets from microarray experiments to validate the effectiveness of the P53-based gene signatures in predicting survival of patients with LUAD. All these datasets are available from the public Gene Expression Omnibus (GEO) database with the following accession IDs: GSE31210, GSE8894, GSE68465, GSE13213, GSE3141, and GSE42127. The number of adenocarcinoma samples in these datasets are 226, 63, 443, 117, 58, and 133, respectively. Among them, the first four datasets provide recurrence-free survival (RFS) of patients, whereas the other two datasets provide only overall survival (OS). Patient smoking information is available for GSE31210, GSE68465, and GSE13213 datasets. The mutation status of P53 is only available for GSE13213 dataset. A summary of these six datasets is provided in the Supplementary Table S1. These datasets were all generated from one-channel microarray platforms, and were downloaded as a matrix containing the expression levels of all probesets. Probeset-level expression was converted into gene-level expression by choosing the probeset with the maximum average expression to represent the gene with multiple probesets.

Define P53-deficency gene signature based on TCGA LUAD RNA-seq data

The P53-deficency gene signature was defined by comparing the differential expression of genes between P53-mutant and wild-type samples while considering confounding variables using TCGA LUAD RNA-seq data. Samples containing synonymous *TP53* mutations were assigned to the wild-type group. For each gene, a logistic regression model was constructed using patient class as the response variable ($Y = 1$ for P53-mutant samples, and $Y = 0$ for P53 wild-type samples).

$$Y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)}}$$

The predictor variables include expression level of the gene of consideration (X_1), age at the time of diagnosis (X_2), gender (X_3), tumor stage (I, II, III, or IV, labeled as X_4), and smoking status (X_5). Gene expression level was represented as log(RPKM+1) to avoid extreme values. By applying these models to the TCGA LUAD data, we estimated the coefficients (β values) and their statistical significance (p value) for all genes. Second, given (β, p) values for all genes, we defined the P53-deficency gene signature using a pair of weight profiles, w^+ and w^-, that assigned all genes two values in the following way: For gene i, $w_i^+ = -\log(p_i)I(\beta_i > 0)$ and $w_i^- = -\log(p_i)I(\beta_i < 0)$. To avoid extreme values, the weights were trimmed at 10, and then transformed into a value within [0,1], by subtracting the minimum value and then dividing by the range. If a gene i is more significantly upregulated in P53-mutant versus wild-type samples, it will associate with a higher w_i^+ and w_i^- of zero. Conversely, a more significantly downregulated gene will associate with a higher w_i^- and w_i^+ of zero.

Calculate patient-specific PDS based on their expression profiles

Given the expression profiles for a number of LUAD samples, sample-specific PDSs were calculated for all samples based on the P53-deficient gene signature as described above. Specifically, we applied a modified version of a statistical method called BASE (43) with the following steps: First, gene expression data were converted into relative expression of genes by comparing with a calculated reference profile that contained the median expression of genes across all samples. Second, genes were sorted in a descending order based on their relative expression to obtain an expression profile $(e_1, e_2, \ldots, e_g)$, where g is the total number of genes. The biased distribution of upregulated (with large values in w^+) and downregulated (with large values in w^-) genes in P53-mutant samples were examined by comparing two cumulative functions, a foreground $f(i)$ and a background $b(i)$.

$$f(i) = \sum_{k=1}^{i} |e_k w_k| \Big/ \sum_{k=1}^{g} |e_k w_k|, \quad 1 \le i \le g$$

$$b(i) = \sum_{k=1}^{i} |e_k(1 - w_k)| \Big/ \sum_{k=1}^{g} |e_k(1 - w_k)|, \quad 1 \le i \le g$$

If genes with large weights in w (w_i^+ for upregulated genes and w_i^- for downregulated genes in P53-mutant samples) tend to have large values in tumor expression profile e, $f(i)$ will increase in value more rapidly than $b(i)$ as i increases. Third, the maximum deviation between the two functions was calculated and normalized against a null distribution that was estimated by permutation to obtain PDS$^+$ (if $w = w^+$) or PDS$^-$ (if $w = w^-$). Finally, the two scores were combined by taking their difference (PDS$^+$ − PDS$^-$) to obtain the final PDS for this sample. Patients with high PDSs are more likely to have P53 mutation, whereas patients with low PDSs are likely to have wild-type P53. Therefore, high PDS refers to low P53 pathway activity and vice versa. After accomplishing this procedure, the PDSs were calculated for all samples in the lung cancer expression datasets.

Predict patient survival using PDSs

Cox proportional hazard models were constructed to investigate the effectiveness of patient-specific PDSs in predicting patients' survival (RFS or OS). Patient samples were dichotomized into two groups by using an indicator function $I(\text{PDS} \ge t)$, where t is a prespecified threshold. Normally, we set $t = 0$. If this threshold resulted in no or a small number of samples in one group, we set t as the median of PDSs in the sample. A univariate Cox regression model was used to determine the association between dichotomized PDSs and patient survival. Multivariate Cox regression model was used to determine the effect of PDSs on survival after adjusting for potential confounding variables such as age, tumor stage, smoking status, etc. Kaplan–Meier method and log rank test were used to plot survival curve (44). The difference between the survival curves of different groups was compared with significance being estimated by using a log-rank test. The R package "survival" was used to implement statistical analyses. Specifically, the "coxph" function was used to construct cox proportional hazard models; the "survfit" function was used to create Kaplan–Meier survival curves, and the "survdiff" function was used to compare the difference between two survival curves.

P53 mutation types in TCGA

The TCGA LUAD patients were separated into three groups (P53 gain of function, P53 loss of function, and P53 wild-type) based on their P53 mutation types. The P53 gain of function group was determined by containing R248Q, R27H, or R175H mutation in protein sequence (45, 46). The P53 loss of function group was determined by having P53 nonsense mutations or frame shift mutations in the transcript sequence.

P53 target genes

The target genes of P53 were downloaded from the ChIP Enrichment Analysis (CHEA) database (47), which provides P53 targets in four different human cell lines, HCT116, U2OS, IMR90, and HFKS, identified from ChIP-chip or ChIA-PET experiments. None of those four cell lines has P53 mutation. Genes identified in at least two cell lines were selected, resulting in a total of 627 P53 target genes.

Results

Overview of this study

To assess whether a gene signature that reflects P53 activity is a better prognostic marker than P53 mutation status, we performed a series of analyses in LUAD as diagrammed in Fig. 1. We used TCGA RNA-seq data for LUAD to define a P53-deficient gene signature by comparing gene expression between P53-mutant and wild-type samples. In contrast with traditional gene signatures,

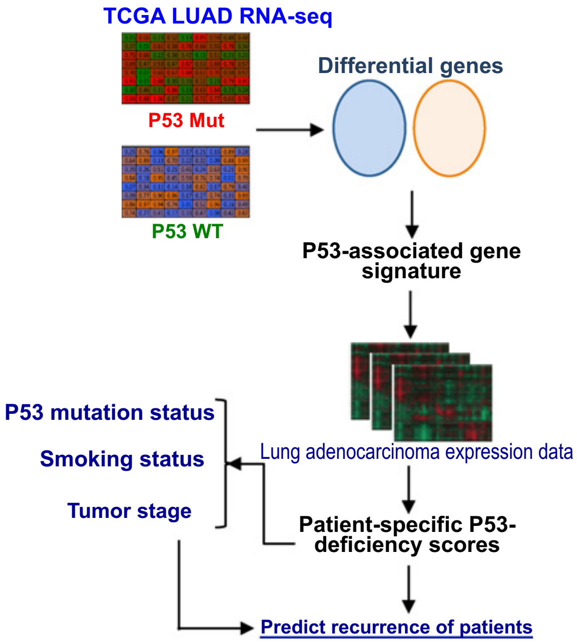
TCGA LUAD RNA-seq
P53 Mut
P53 WT
Differential genes
P53-associated gene signature
Lung adenocarcinoma expression data
P53 mutation status
Smoking status
Tumor stage
Patient-specific P53-deficiency scores
Predict recurrence of patients

Figure 1.
The schematic diagram of this study. TCGA RNA-seq data for LUAD were used to define a P53-deficient gene signature. The signature was applied to multiple adenocarcinoma microarray datasets to calculate a patient-specific PDS for each sample in the datasets. The associations of PDS with P53 mutation status, tumor stage, and smoking status were investigated. The ability of PDS to predict RFS of patients with or without considering clinical variables was examined in all datasets.

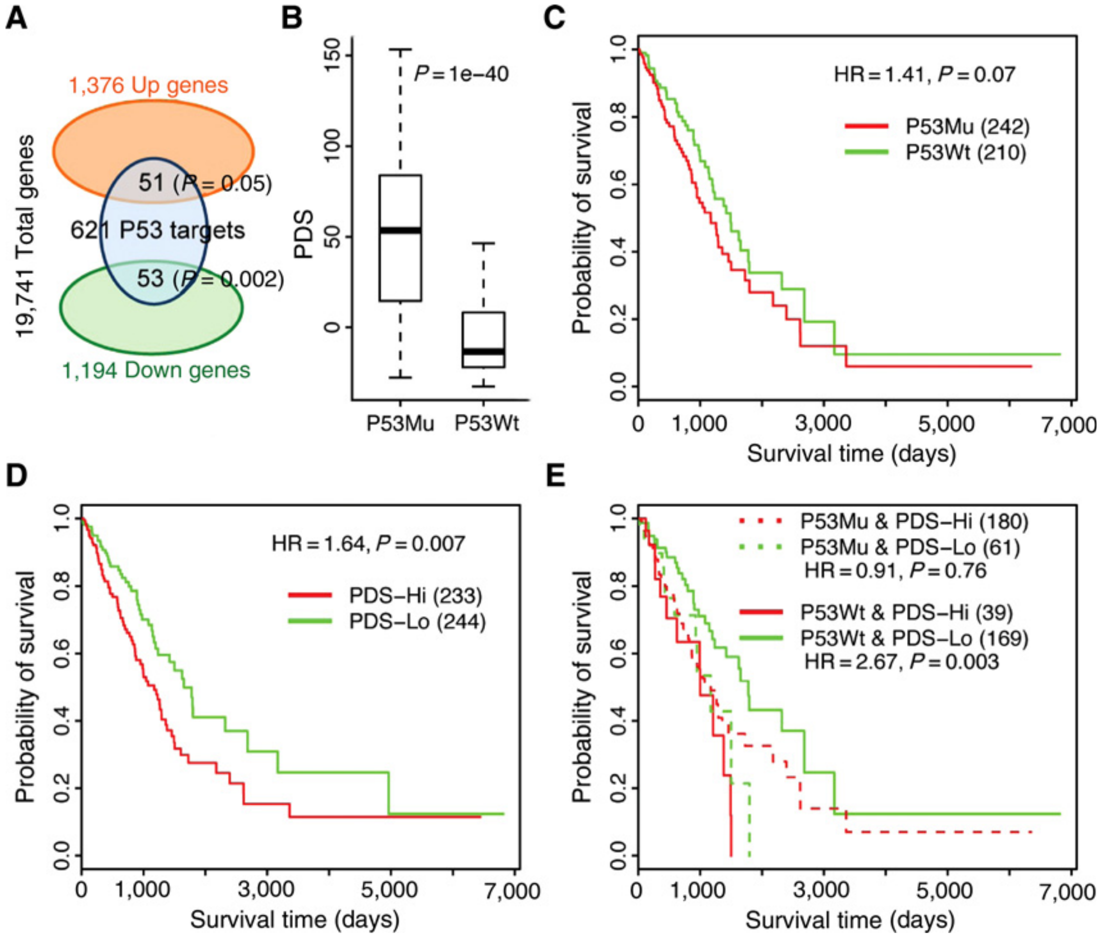

Figure 2.

PDS is predictive of patient OS in TCGA LUAD dataset. **A,** Genes up- and downregulated in P53-mutant LUAD samples are enriched in P53 target genes identified from ChIP-chip or ChIA-PET experiments. **B,** The PDSs of P53-mutant samples are significantly higher than those of P53 wild-type samples. Statistical significance is calculated by Mann–Whitney *U* test. **C,** Mutation status of P53 is not predictive of OS of patients. **D,** PDS is predictive of OS of patients. Patients with high PDSs have significantly shorter survival than those with low PDSs. **E,** Prediction power of PDS in P53-mutant and P53 wild-type sample subsets. The median PDS was used as the cutoff to dichotomize patients into high- (PDS-Hi) and low- (PDS-Lo) groups. Numbers in the parenthesis indicate the number of patients in each group. HR is calculated by using a Cox regression model. *P* values between survival curves were calculated by using the log-rank tests.

our signature consisted of a pair of weighted profiles that indicate the magnitude of up- and downregulation of genes in P53-mutant samples relative to wild-type samples after adjusting for clinical variables such as age at time of diagnosis and tumor stage. This whole transcriptomic signature was applied to infer P53 pathway activity in samples based on their gene expression profiles. Patients having high similarities between their transcriptomic profiles and P53-deficient gene signature would have high PDSs, which leads to low P53 pathway activity, and vice versa. After PDS calculation, we next evaluated the PDS's ability to discriminate samples with or without P53 mutation, and examined the association of PDS with tumor stage and smoking status. Moreover, we examined the effectiveness of PDS in predicting the RFS of patients in four independent LUAD datasets. Particularly, we investigated its predictive power in early stage samples. We demonstrated that PDS is prognostic even after adjusting clinical variables including age, stage, P53 mutation status, and smoking status.

Association of PDS with P53 mutation and OS

Statistical analysis of TCGA LUAD RNA-seq data resulted in the identification of 1,376 and 1,194 genes that were up- and downregulated in P53-mutant versus wild-type samples at a false discovery rate of 0.001 (Supplementary Table S2). Gene set enrichment analysis indicated that upregulated genes were enriched for cell cycle and DNA replication genes, whereas downregulated genes were enriched for ribosome genes (Supplementary Table S3). We investigated the overlap of these genes with the 621 P53 target genes that were available from the LUAD RNA-seq data (Fig. 2A). Out of these target genes, 51 and 53 were up- and downregulated in P53-mutant samples, representing a significant enrichment of 1.3-fold ($P = 0.05$) and 1.5-fold ($P = 0.002$), respectively. This indicates that P53 target genes are more likely to be differentially expressed between P53-mutant and wild-type LUAD samples. Moreover, P53 may primarily function as a transcriptional activator, because its target genes are more significantly enriched in downregulated genes. As shown, P53 targets only represent a small fraction (~4%) of the differentially expressed genes, suggesting that the majority of them are indirectly regulated by the P53 pathway. Additionally, according to our previous study, we provided the importance of using the indirect targets of transcriptional factor to estimate its pathway activity (48, 49). Therefore, even though those P53 target genes were indirectly regulated by P53 pathway, it is reasonable to include them into P53 gene signature to enhance the power of statistical inference on P53 pathway activity for the future analysis.

To further validate PDS is associated with P53 pathway activity, we stratified TCGA LUAD patients into P53 gain of function group, P53 loss of function group and wild-type P53 group and compared their PDSs difference. As shown in the Supplementary Fig. S1, patients with P53 loss of function mutation showed the significantly higher PDSs comparing to the patients with wild-type P53 ($P = 2e{-}7$). Because of the limited sample size in the P53 gain of function group ($N = 4$), we did not observe PDSs showed significant difference between P53 gain of function group and P53 loss of function group. However, PDSs in the P53 loss of function group trended to be higher compared to the PDSs in the P53 gain of function group. These observations further indicate that our PDS is associated with P53 pathway activity in the LUAD.

Next we examined whether PDSs of samples inform their P53 mutation status. As shown in Fig. 2B, we observed significantly higher PDSs in P53-mutant samples compared with wild-type samples ($P = 1e{-}40$). The mutation status of P53 was not predictive of patient OS as shown in Fig. 2C. However, when patients were dichotomized into two groups with either high or low PDS, we observed significantly shorter OS of the high-PDS (PDS-Hi) group compared to the low-PDS (PDS-Lo) group (Fig. 2D). This suggested that P53 deficiency was associated with poor prognosis. We next separated patients into two subsets based on their P53 mutation status and examined the predictive power of PDS in each subset (Fig. 2E). As shown in Fig. 2E, we observed a significant difference between PDS-Hi and PDS-Lo patients in the P53 wild-type subset ($P = 0.003$, solid curves), but not in the P53-mutant subset ($P > 0.1$, dotted curves). For wild-type patients, the PDS-Hi group had overall mortality rates 2.67-fold higher than the PDS-Lo group. These results suggested that in P53 wild-type samples there might exist other mechanisms that result in defective P53 pathway activity. However, most P53-mutant samples have defective P53 pathway activity, and therefore PDS does not provide further prognostic significance in these patients.

Predicting RFS in four LUAD datasets

Having shown the association of PDS with patient OS in TCGA LUAD, we subsequently examined its ability to predict RFS in four independent datasets. These datasets were generated by using different microarray platforms, and varied in the number of samples from 63 to 442. The samples in the GSE68465 dataset were mostly from white American patients, whereas samples in the other three datasets were from East Asian patients. As shown in Fig. 3, results in the four datasets consistently showed that high-PDS is associated with significantly shorter RFS in LUAD.

Although we showed that there was an association between PDS and P53 mutation status in the TCGA LUAD data, the underlying P53-deficient gene signature was defined based on

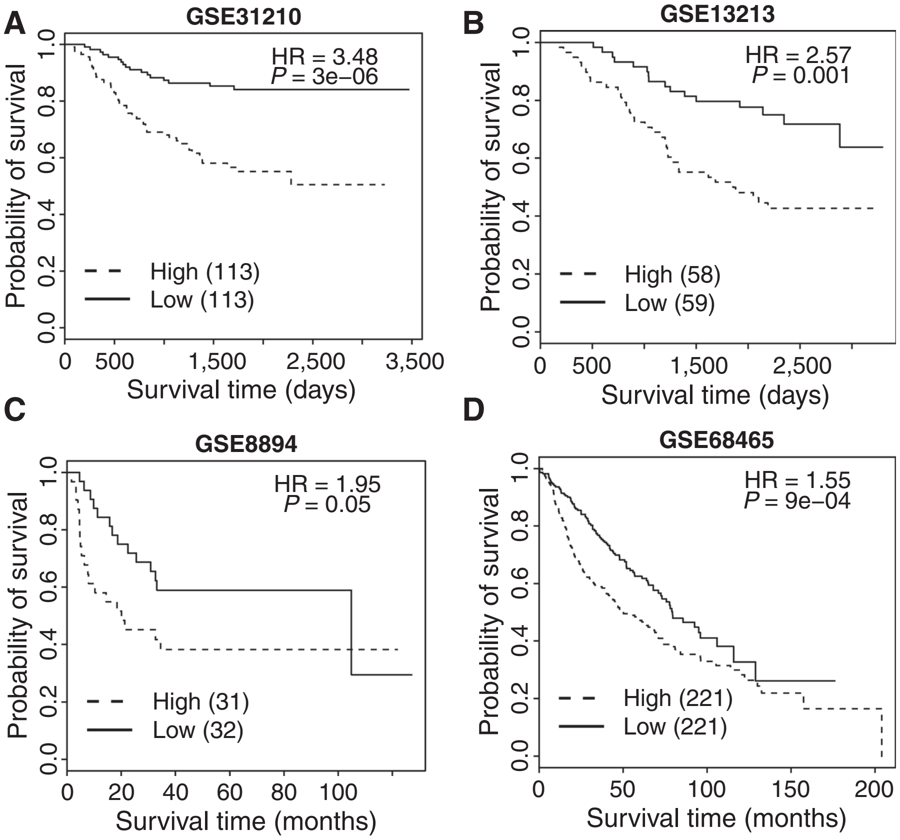

Figure 3.
PDS is predictive of RFS in four microarray LUAD datasets. **A,** GSE31210; **B,** GSE13213; **C,** GSE8894; **D,** GSE68465. For each dataset, the median PDS was used as the cutoff to dichotomize patients into high-PDS and low-PDS groups. Numbers in the parenthesis indicate the number of patients in each group. HR is calculated by using a Cox regression model. *P* values between survival curves were calculated by using the log-rank tests.

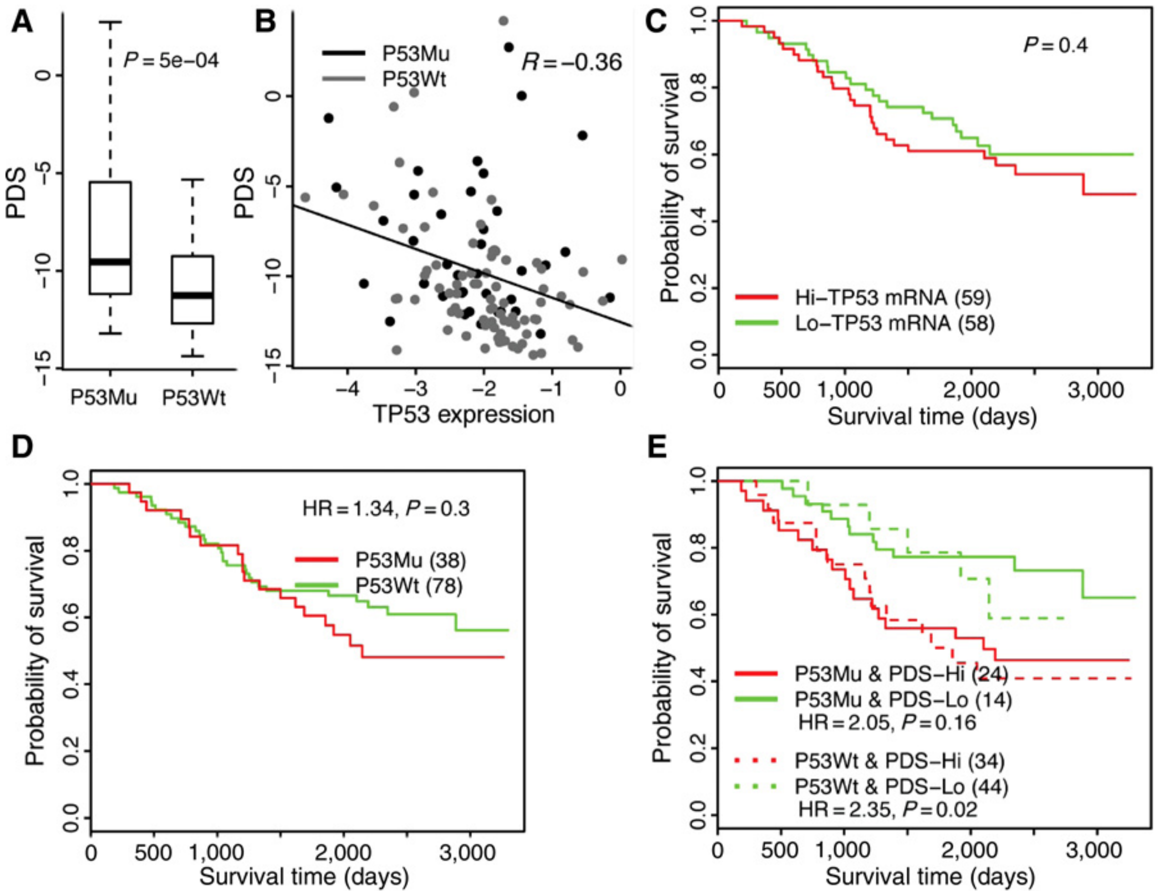

Figure 4.

PDS provides a more effective prognostic marker than *TP53* expression and P53 mutation status. **A,** The PDSs of P53-mutant samples are significantly higher than those of P53 wild-type samples. Statistical significance is calculated by Mann–Whitney *U* test. **B,** PDS is negatively correlated with *TP53* gene expression level. *TP53* expression was log transformed and correlation coefficient was calculated by Spearman correlation. **C,** Expression level of *TP53* is not predictive of RFS of patients. The median of *TP53* expression was used as the cutoff to dichotomize patients. **D,** Mutation status of P53 is not predictive of RFS of patients. **E,** Prediction power of PDS in P53-mutant and P53 wild-type sample subsets. All analyses were based on the GSE13213 dataset. Numbers in the parenthesis indicate the number of patients in each group. HR is calculated by using a Cox regression model. *P* values between survival curves were calculated by using the log-rank tests.

LUAD data and the prognostic analysis was based on OS instead of RFS. Thus, we further validated the association between PDS and P53 mutation status in the GSE13213 dataset. As shown in Fig. 4A, P53-mutant samples had significantly higher PDSs compared to wild-type samples ($P = 5\mathrm{e}{-4}$). In addition, PDSs were negatively correlated with *TP53* expression levels (Fig. 4B), although the correlation was moderate ($R = -0.36$). Neither *TP53* expression level (Fig. 4C) or P53 mutation status (Fig. 4D) was predictive of RFS in this dataset, however, consistent with Fig. 2E, PDS was significantly associated with RFS in P53 wild-type but not in P53-mutant patients (Fig. 4E). Altogether, these results indicate that PDS correctly reflects P53 pathway activity and provides a better prognostic marker than *TP53* expression and P53 mutation status in LUAD.

In addition to these analyses, we also examined the prognostic value of PDS in two other LUAD microarray datasets in which OS but not RFS is available. Our results indicated that high-PDS was associated with poor OS, but less predictive power was observed (Supplementary Fig. S2).

Predicting RFS in early-stage LUAD

We next investigated whether PDS was prognostic in early-stage LUAD after adjusting for clinical variables. To do this, we constructed a multivariate Cox regression model on the GSE31210 dataset that included dichotomized PDSs, gender, age at diagnosis, tumor stage, and smoking status as predictor variables. As shown in the forest plot, PDS remained significant ($P = 0.002$) for predicting RFS even after considering key clinical variables (Fig. 5A). However, stage II samples had significantly higher PDSs than stage I samples (Fig. 5B). Stage-specific survival analysis indicated that PDS was a significant predictor in stage I samples ($P = 0.008$) with HR = 2.33 (Fig. 5C). In stage II samples, high PDS was also associated with poor survival with HR = 2.19, but was not significant ($P = 0.07$), likely due to the small sample size (Fig. 5D).

Association of PDS with smoking status

A large fraction of lung cancer cases are associated with long-term smoking. Therefore, we investigated the association between

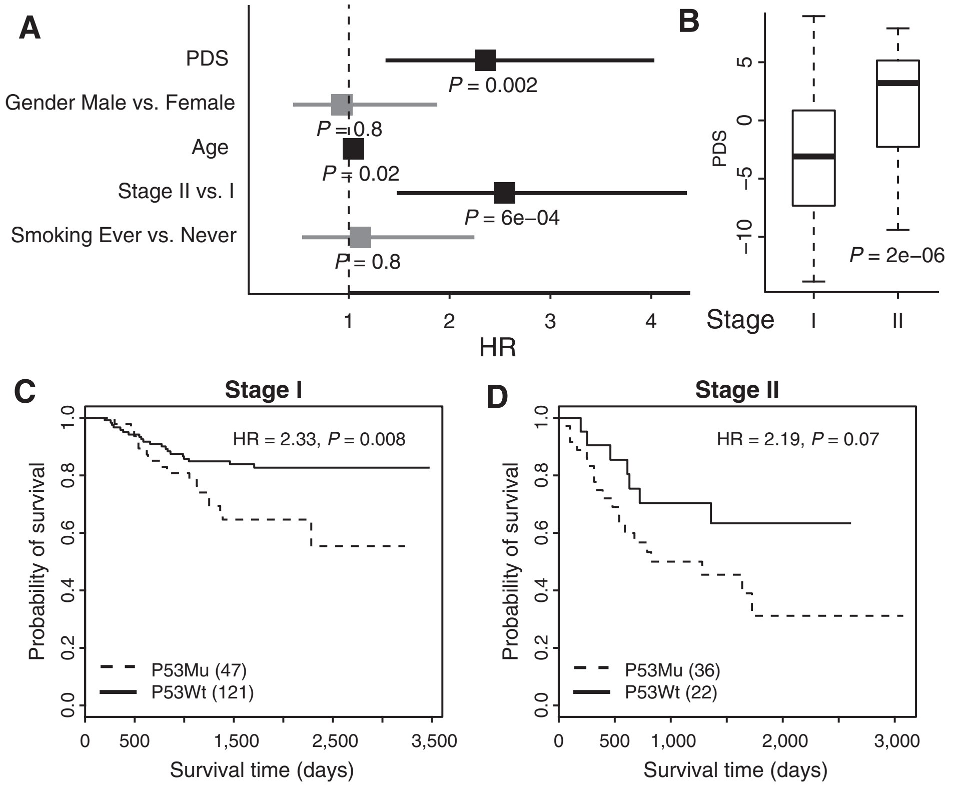

Figure 5.

PDS provides prognostic value in addition to tumor stage and other clinical information. **A,** A forest plot shows the HR and *P* value of PDS and several clinical variables estimated by using the multivariate Cox regression model. Model is adjusted by clinical variables which are gender, age at diagnosis, stage, and smoking status. HRs (95% confidence intervals) are denoted by black boxes. Statistical significance is calculated by Wald test. **B,** Stage II samples have significantly higher PDSs than stage I samples. Statistical significance is calculated by Mann–Whitney *U* test. **C,** PDS is predictive of RFS of patients in a subset of stage I samples. **D,** PDS is predictive of RFS of patients in a subset of stage II samples. All analyses were based on the GSE31210 dataset. Numbers in the parenthesis indicate the number of patients in each group. HR is calculated by using a Cox regression model. *P* values between survival curves were calculated by using the log-rank tests.

PDS and smoking status of patients. As shown, patients that were current or former smokers showed significantly higher PDSs than those who were never-smokers ($P = 7e{-}5$) in the GSE31210 dataset (Fig. 6A). However, smoking status was not a significant predictor for RFS (Fig. 6B). We separated patients into ever- and never-smokers, and performed survival analysis in each subset. We found that high-PDS was significantly associated with patient RFS with a high HR = 4.16 ($P = 7e{-}6$) for ever-smokers, but only moderate significance was observed for never-smokers ($P = 0.09$). The same analysis was performed in the GSE68465 and the GSE13213 datasets, which results in similar observations but less significant association between PDS and survival in both ever- and never-smoker subgroups (Supplementary Fig. S3). These results suggest that P53 might play more critical roles in the progression of LUAD associated with tobacco exposure.

Association of PDS with survival in lung squamous cell carcinoma

We expanded our prognostic analyses in patients with lung squamous cell carcinoma using the TCGA LUAD RNA-seq data derived P53-deficient gene signature. Specifically, we performed this analysis using the squamous cell carcinoma samples available from the TCGA LUSC dataset ($n = 501$, OS), the GSE8894 dataset ($n = 75$, OS), the GSE3141 dataset ($n = 53$, OS), the GSE4573

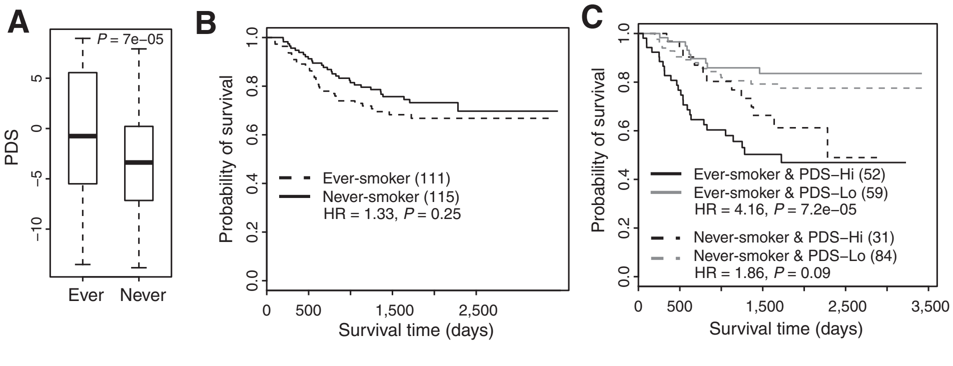

Figure 6.

PDS is associated with smoking status of patients. **A,** Patients in the ever-smoking group have significantly higher PDSs than those in the never-smoking group. Statistical significance is calculated by Mann–Whitney *U* test. **B,** Smoking status is not predictive of RFS of patients. **C,** Prediction power of PDS in ever-smoking and never-smoking subgroups. All analyses were based on the GSE31210 dataset. Numbers in the parenthesis indicate the number of patients in each group. HR is calculated by using a Cox regression model. *P* values between survival curves were calculated by using the log-rank tests.

dataset ($n = 130$, OS), and the GSE14814 dataset ($n = 52$, DSS: disease-specific survival). Interestingly, the inferred PDS was not predictive of OS or DSS of patients in these datasets (Supplementary Fig. S4). This might be explained by the fact that the P53-deficient gene signature was defined based on LUAD data and was not be able to accurately reflect P53 activity in squamous cell carcinoma. However, a squamous cell carcinoma-specific P53-deficient gene signature defined based on TCGA LUSC RNA-seq data did not predict patient survival either (Supplementary Fig. S5). This may be because the *TP53* gene has a very high mutation frequency (>80%) in lung squamous cell carcinoma (31). Thus, it is possible that all cases of this lung cancer subtype are associated with a defective P53 pathway, and therefore the P53 pathway activity is not of prognostic value.

Discussion

Several studies have reported prognostic gene signatures for lung cancer especially in NSCLC. However, none of those gene signatures presented significance in prognosis prediction when clinical variables were taken into consideration. Often the genes involved in those signatures were poorly overlapped, suggesting a lack of reproducibility. Moreover, *TP53* mutations, the most common mutations in human cancers, have been researched as a predictive biomarker for prognosis in lung cancer patients in recent years. Because of the different mutation types of *TP53*, the prognostic value of *TP53* in lung cancer has not been clearly determined.

In this study, we defined a P53-deficient gene signature by comparing gene expression between P53-mutant and wild-type samples in TCGA LUAD RNA-seq data. This signature was used to calculate the PDSs of patients based on their expression measured by microarray platforms. Our results indicated that PDS was associated with P53 mutation status and smoking status, and was predictive of patient RFS with high consistency in multiple LUAD datasets. Furthermore, its predictive ability was independent of clinical variables including tumor stage, suggesting it has the potential to be used for stratifying early-stage patients based on their prognosis to adopt personalized treatment.

We showed that the PDS outperforms *TP53* expression and P53 mutation status in terms of prognostic prediction in LUAD (Fig. 2 and 4). The activity of the P53 protein is determined by the posttranscriptional regulation and posttranslational modification, and as a result, the mRNA level of *TP53* does not accurately reflect its protein activity. As shown, *TP53* expression is only weakly correlated with PDS in the GSE31210 dataset (Fig. 2B), and does not predict patient survival (Fig. 2C). More than 75% of *TP53* mutations result in an abnormal P53 protein that deactivates the P53 pathway via a dominant-negative regulation of wild-type P53 (37); however, the severity of distinct P53 mutations varies substantially. Moreover, the P53 pathway can also be deactivated by other biological mechanisms, including epigenetic regulation and deletion of *TP53* genes, or through alterations of other genes in this pathway. For these reasons, somatic mutation status does not fully capture the P53 pathway activity, and is not a significant predictor for patient survival (Fig. 4D). In contrast, the inferred PDS provides a quantitative measurement of P53 pathway activity, and is predictive of patient prognosis, especially, in P53 wild-type samples (Figs. 2E and 4E).

The P53-deficient gene signature is defined by comparing transcriptome profiles between P53-mutant and wild-type LUAD samples from TCGA. However, smoking status, which is the most important distinction in LUAD classification, might confound the differential gene expression analysis used to define this signature. Nonsmokers tend to have fewer somatic mutations compared to current or former smokers. To ensure that our P53-deficient gene signature solely picked up differences in P53 activity and was not confounded by smoking status and other clinical variables, we adjusted for these factors during the differential expression analysis used to create our signature. We validated that this adjustment was sufficient in regards to smoking status through two follow-up analyses. A multivariate Cox regression model found that PDS was the most significant predictor of patient survival even when adjusting for smoking status in the GSE31210 dataset (Fig. 5A). Furthermore, when patients were stratified based on smoking status, the PDS was predictive of survival with high significance in the ever-smoker group and moderate significance in the never-smoker group in the GSE31210 dataset (Fig. 6C).

In breast cancer, a 32-gene signature has been proposed by Miller and colleagues to predict P53 mutation status and patient prognosis (50). In this study, we showed in LUAD that a P53-deficient gene signature defined based on RNA-seq provides a significant prognostic predictor that is applicable to both RNA-seq and microarray platforms. In addition, we applied a novel method to calculate P53 deficiency in tumor samples, which utilized the whole gene signature instead of selecting a small set of genes. The P53-deficient gene signature consists of all genes and for each gene a weight is assigned on the basis of its ability to discriminate P53-mutant against wild-type samples. This whole-transcriptome strategy is easy to be implemented and achieves a high statistical power.

In summary, we have defined a gene signature that captures P53 pathway activity in LUAD samples and predicts patient prognosis. The computational framework introduced in this study can be applied to define prognostic signatures for any cancer types based on matched gene expression and somatic mutation data.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: C. Cheng
Development of methodology: C. Cheng
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): C.I. Amos, C. Cheng
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): Y. Zhao, F.S. Varn, G. Cai, C.I. Amos, C. Cheng
Writing, review, and/or revision of the manuscript: Y. Zhao, F.S. Varn, G. Cai, F. Xiao, C.I. Amos, C. Cheng
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): F. Xiao, C. Cheng
Study supervision: C. Cheng

Acknowledgments

This work was supported by the National Center for Advancing Translational Sciences of the NIH under award number KL2TR001088, the Center of Biomedical Research Excellence grant under award number GM103534, and the start-up funding package provided to Chao Cheng by the Geisel School of Medicine at Dartmouth College.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received June 1, 2017; revised August 17, 2017; accepted October 23, 2017; published OnlineFirst November 15, 2017.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin 2015;65:5–29.
2. NIH. Cancer stat facts: lung and bronchus cancer. Bethesda, MD: NIH. Available from: http://seer.cancer.gov/statfacts/html/lungb.html.
3. Abe Y, Tanaka N. The Hedgehog signaling networks in lung cancer: the mechanisms and roles in tumor progression and implications for cancer therapy. BioMed Res Int 2016;2016:7969286.
4. Thun MJ, Hannan LM, Adams-Campbell LL, Boffetta P, Buring JE, Feskanich D, et al. Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. PLoS Med 2008;5:e185.
5. Masters GA, Krilov L, Bailey HH, Brose MS, Burstein H, Diller LR, et al. Clinical cancer advances 2015: annual report on progress against cancer from the American Society of Clinical Oncology. J Clin Oncol 2015;33: 786–809.
6. Rusthoven CG, Kavanagh BD, Karam SD. Improved survival with stereotactic ablative radiotherapy (SABR) over lobectomy for early stage non-small cell lung cancer (NSCLC): addressing the fallout of disruptive randomized data. Ann Transl Med 2015;3:149.
7. Choi PJ, Jeong SS, Yoon SS. Prediction and prognostic factors of post-recurrence survival in recurred patients with early-stage NSCLC who underwent complete resection. J Thorac Dis 2016;8:152–60.
8. Pyka T, Bundschuh RA, Andratschke N, Mayer B, Specht HM, Papp L, et al. Textural features in pre-treatment [F18]-FDG-PET/CT are correlated with risk of local recurrence and disease-specific survival in early stage NSCLC patients receiving primary stereotactic radiation therapy. Radiat Oncol 2015;10:100.
9. Naruke T, Goya T, Tsuchiya R, Suemasu K. Prognosis and survival in resected lung carcinoma based on the new international staging system. J Thorac Cardiovasc Surg 1988;96:440–7.
10. Chen C-Y, Chen C-H, Shen T-C, Cheng W-C, Hsu C-NLiao C-H, et al. Lung cancer screening with low-dose computed tomography: experiences from a tertiary hospital in Taiwan. J Formos Med Assoc Taiwan Yi Zhi 2016;115:163–70.
11. Deffebach ME, Humphrey L. Lung cancer screening. Surg Clin North Am 2015;95:967–78.
12. Fontana RS, Sanderson DR, Woolner LB, Taylor WF, Miller WE, Muhm JR. Lung cancer screening: the Mayo program. J Occup Med 1986; 28:746–50.
13. Kadota K, Yeh Y-C, Sima CS, Rusch VW, Moreira AL, Adusumilli PS, et al. The cribriform pattern identifies a subset of acinar predominant tumors with poor prognosis in patients with stage I lung adenocarcinoma: a conceptual proposal to classify cribriform predominant tumors as a distinct histologic subtype. Mod Pathol 2014;27:690–700.
14. Warth A, Muley T, Meister M, Stenzinger A, Thomas M, Schirmacher P, et al. The novel histologic International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society classification system of lung adenocarcinoma is a stage-independent predictor of survival. J Clin Oncol 2012;30:1438–46.
15. Yanagawa N, Shiono S, Abiko M, Ogata S, Sato T, Tamura G. New IASLC/ATS/ERS classification and invasive tumor size are predictive of disease recurrence in stage I lung adenocarcinoma. J Thorac Oncol 2013;8:612–8.
16. Woo T, Okudela K, Mitsui H, Tajiri M, Yamamoto T, Rino Y, et al. Prognostic value of the IASLC/ATS/ERS classification of lung adenocarcinoma in stage I disease of Japanese cases. Pathol Int 2012;62: 785–91.
17. Battafarano RJ, Piccirillo JF, Meyers BF, Hsu H-S, Guthrie TJ, Cooper JD, et al. Impact of comorbidity on survival after surgical resection in patients with stage I non-small cell lung cancer. J Thorac Cardiovasc Surg 2002; 123:280–7.
18. Lou F, Sima CS, Rusch VW, Jones DR, Huang J. Differences in patterns of recurrence in early-stage versus locally advanced non-small cell lung cancer. Ann Thorac Surg 2014;98:1755-1760-1761.
19. Lou F, Huang J, Sima CS, Dycoco J, Rusch V, Bach PB. Patterns of recurrence and second primary lung cancer in early-stage lung cancer survivors followed with routine computed tomography surveillance. J Thorac Cardiovasc Surg 2013;145:75-81-82.
20. Lou F, Sarkaria I, Pietanza C, Travis W, Roh MS, Sica G, et al. Recurrence of pulmonary carcinoid tumors after resection: implications for postoperative surveillance. Ann Thorac Surg 2013;96:1156–62.
21. Kratz JR, Jablons DM. Genomic prognostic models in early-stage lung cancer. Clin Lung Cancer 2009;10:151–7.
22. Okayama H, Kohno T, Ishii Y, Shimada Y, Shiraishi K, Iwakawa R, et al. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Cancer Res 2012;72:100–11.
23. Tomida S, Takeuchi T, Shimada Y, Arima C, Matsuo K, Mitsudomi T, et al. Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis. J Clin Oncol 2009;27:2793–9.
24. Lee E-S, Son D-S, Kim S-H, Lee J, Jo J, Han J, et al. Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. Clin Cancer Res 2008;14:7397–404.
25. Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Shedden K, Taylor JMG, Enkemann SA, Tsao M-S, Yeatman TJ, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 2008;14:822–7.
26. Beer DG, Kardia SLR, Huang C-C, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–24.
27. Kadara H, Behrens C, Yuan P, Solis L, Liu D, Gu X, et al. A five-gene and corresponding protein signature for stage-I lung adenocarcinoma prognosis. Clin Cancer Res 2011;17:1490–501.
28. Chen H-Y, Yu S-L, Chen C-H, Chang G-C, Chen C-Y, Yuan A, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 2007;356:11–20.
29. Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. Science 1991;253:49–53.
30. Campling BG, el-Deiry WS. Clinical implications of p53 mutations in lung cancer. Methods Mol Med 2003;75:53–77.
31. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature 2013; 502:333–9.
32. Gibbons DL, Byers LA, Kurie JM. Smoking, p53 mutation, and lung cancer. Mol Cancer Res 2014;12:3–13.
33. Halvorsen AR, Silwal-Pandit L, Meza-Zepeda LA, Vodak D, Vu P, Sagerup C, et al. TP53 mutation spectrum in smokers and never smoking lung cancer patients. Front Genet 2016;7:85.
34. Ye X-H, Bu Z-B, Feng J, Peng L, Liao X-B, Zhu X-L, et al. Association between the TP53 polymorphisms and lung cancer risk: a meta-analysis. Mol Biol Rep 2014;41:373–85.
35. Steels E, Paesmans M, Berghmans T, Branle F, Lemaitre F, Mascaux C, et al. Role of p53 as a prognostic factor for survival in lung cancer: a systematic review of the literature with a meta-analysis. Eur Respir J 2001;18:705–19.
36. Ahrendt SA, Hu Y, Buta M, McDermott MP, Benoit N, Yang SC, et al. p53 mutations and survival in stage I non-small-cell lung cancer: results of a prospective study. J Natl Cancer Inst 2003;95:961–70.
37. Muller PAJ, Vousden KH. p53 mutations in cancer. Nat Cell Biol 2013; 15:2–8.
38. Saldaña-Meyer R, Recillas-Targa F. Transcriptional and epigenetic regulation of the p53 tumor suppressor gene. Epigenetics 2011;6:1068–77.
39. Pogribny IP, James SJ. Reduction of p53 gene expression in human primary hepatocellular carcinoma is associated with promoter region methylation without coding region mutation. Cancer Lett 2002;176:169–74.
40. Rivlin N, Brosh R, Oren M, Rotter V. Mutations in the p53 tumor suppressor gene: important milestones at the various steps of tumorigenesis. Genes Cancer 2011;2:466–74.
41. Kruse J-P, Gu W. Modes of p53 regulation. Cell 2009;137:609–22.
42. Huang L, Yan Z, Liao X, Li Y, Yang J, Wang Z-G, et al. The p53 inhibitors MDM2/MDMX complex is required for control of p53 activity in vivo. Proc Natl Acad Sci U S A 2011;108:12001–6.
43. Cheng C, Yan X, Sun F, Li LM. Inferring activity changes of transcription factors by binding association with sorted expression profiles. BMC Bioinform 2007;8:452.

44. Cox D. Regression models and life-tables. J R Stat Soc, Ser B; 1972;34: 187–220.
45. Xu J, Reumers J, Couceiro JR, De Smet F, Gallardo R, Rudyak S, et al. Gain of function of mutant p53 by coaggregation with multiple tumor suppressors. Nat Chem Biol 2011;7:285–95.
46. Oren M, Rotter V. Mutant p53 gain-of-function in cancer. Cold Spring Harb Perspect Biol 2010;2:a001107.
47. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinforma Oxf Engl 2010;26:2438–44.
48. Khaleel SS, Andrews EH, Ung M, DiRenzo J, Cheng C. E2F4 regulatory program predicts patient survival prognosis in breast cancer. Breast Cancer Res 2014;16:486.
49. Cheng C, Lou S, Andrews EH, Ung MH, Varn FS. Integrative genomic analyses yield cell-cycle regulatory programs with prognostic value. Mol Cancer Res 2016;14:332–43.
50. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci USA 2005;102:13550–5.