Published OnlineFirst December 1, 2009; DOI: 10.1158/0008-5472.CAN-09-1378



# Gene Expression Profiles in Peripheral Blood Mononuclear Cells Can Distinguish Patients with Non–Small Cell Lung Cancer from Patients with Nonmalignant Lung Disease

Michael K. Showe,[1] Anil Vachani,[2] Andrew V. Kossenkov,[1] Malik Yousef,[1] Calen Nichols,[1] Elena V. Nikonova,[1] Celia Chang,[1] John Kucharczuk,[2] Bao Tran,[2] Elliot Wakeam,[2] Ting An Yie,[3] David Speicher,[1] William N. Rom,[3] Steven Albelda,[2] and Louise C. Showe[1]

[1]The Wistar Institute; [2]Division of Pulmonary, Allergy, and Critical Care Medicine, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania; and [3]Division of Pulmonary and Critical Care Medicine, New York University School of Medicine, New York, New York

## Abstract

**Early diagnosis of lung cancer followed by surgery presently is the most effective treatment for non–small cell lung cancer (NSCLC). An accurate, minimally invasive test that could detect early disease would permit timely intervention and potentially reduce mortality. Recent studies have shown that the peripheral blood can carry information related to the presence of disease, including prognostic information and information on therapeutic response. We have analyzed gene expression in peripheral blood mononuclear cell samples including 137 patients with NSCLC tumors and 91 patient controls with nonmalignant lung conditions, including histologically diagnosed benign nodules. Subjects were primarily smokers and former smokers. We have identified a 29-gene signature that separates these two patient classes with 86% accuracy (91% sensitivity, 80% specificity). Accuracy in an independent validation set, including samples from a new location, was 78% (sensitivity of 76% and specificity of 82%). An analysis of this NSCLC gene signature in 18 NSCLCs taken presurgery, with matched samples from 2 to 5 months postsurgery, showed that in 78% of cases, the signature was reduced postsurgery and disappeared entirely in 33%. Our results show the feasibility of using peripheral blood gene expression signatures to identify early-stage NSCLC in at-risk populations.** [Cancer Res 2009;69(24):9202–10]

## Introduction

Lung cancer is the second most prevalent cancer occurring in both men and women in the United States, accounting for 162,000 deaths in 2008 (1), more than any other cancer. High-risk populations include smokers and former smokers, as well as individuals exposed to second-hand smoke, asbestos, and radon. Presently, there is no easily applied screening protocol for lung cancer similar to those used for breast, prostate, and colon cancers. Screening high-risk patients with low-dose spiral computed tomography (CT; refs. 2–5) identifies small, noncalcified pulmonary nodules in approximately 30% to 70% of high-risk individuals, but only a small proportion (0.4 to 2.7%) of detected nodules ultimately are diagnosed as lung cancers (6–8). Even using the best clinical algorithms, 20% to 55% of patients selected to undergo surgical lung biopsy for indeterminate lung nodules are found to have benign disease (4), and those that do not undergo immediate biopsy or surgery require sequential imaging studies resulting in continued radiation exposure.

Accordingly, efforts are in progress to develop complementary noninvasive diagnostics using techniques such as detection of methylated tumor DNA in sputum (9), serum proteomics (10–12), detection of autoantibodies (13, 14), and gene expression profiling in sputum (15) and airway epithelial brushings (16). Although each of these approaches has its own merits, none has yet passed the exploratory stage. Biomarkers that could be identified from a simple blood test, a routine event associated with regular clinical office visits, would be ideal.

Given previous studies that have analyzed gene expression from peripheral blood mononuclear cells (PBMC) for cancer diagnosis or prognosis (17–21), the goals of this study were to determine whether we could identify a gene expression signature in PBMCs that would accurately distinguish patients with early-stage lung cancer from noncancer controls with similar risk factors (i.e., matched for age, gender, race, and smoking history) and whether such a signature had value in predicting whether lung nodules detected by diagnostic X-ray or CT scans were malignant or benign.

## Materials and Methods

**Study populations.** Study participants (Supplementary Table S1*A–B*) for the initial training sets were recruited from the University of Pennsylvania Medical Center (Penn) during the period 2003 through 2007: 91 subjects with a history of tobacco use without lung cancer, including 41 subjects that had one noncalcified lung nodule diagnosed as benign after biopsy, and 137 patients with newly diagnosed, histopathologically confirmed, non–small cell lung cancer (NSCLC). All participants had blood collection in conjunction with a clinical visit or just before surgery. None of the case subjects had received any cancer therapy before blood collection. Subjects with any prior history of cancer, except nonmelanoma skin cancer, were excluded. Obstructive lung disease was defined as an FEV1/FVC < 70%. We recruited a total of 298 cases and controls from Penn. We excluded 10 NSCLC patients that were diagnosed to have a second cancer, and arrays for 6 samples were removed as technical outliers (see Materials and Methods). The Penn samples were specifically recruited for this study. PBMCs were purified at Penn and RNA extracted at Wistar. The study was approved by the Penn Institutional Review Board (IRB). We also received 90 RNA samples processed at the New York University Medical Center (NYUMC); 27 had acceptable RNA quality based on gel electrophoresis

**Note:** Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

A. Vachani and A.V. Kossenkov contributed equally.

Present address for M. Yousef: Institute for Applied Research, The College of Sakhnin, Israel.

**Requests for reprints:** Louise C. Showe, The Wistar Institute, 3601 Spruce Street, Philadelphia, PA 19104. Phone: 215-898-3901; Fax: 215-898-4521; E-mail: lshowe@wistar.upenn.edu.



doi:10.1158/0008-5472.CAN-09-1378

Published OnlineFirst December 1, 2009; DOI: 10.1158/0008-5472.CAN-09-1378



and Bioanalyzer analysis and only these 27 were further processed for array analysis. Samples from NYUMC were all collected under IRB approval, and are listed in Supplementary Table S1*C*.

**PBMC collection and processing.** Blood samples from Penn were drawn in two "CPT" tubes (BD). PBMCs were isolated within 90 min of blood draw, washed in PBS, transferred into RNAlater (Ambion), and then stored at 4°C overnight before transfer to −80°C. A subset of patient PBMCs was analyzed by flow cytometry, with anti-CD3, CD4, CD8, CD14, CD16, CD19, or CD-56 antibodies or isotype controls (BD Biosciences), and analyzed using FlowJo software. Samples collected at NYUMC were processed within 2 h from collection; PBMC were transferred to Trizol (Invitrogen) and stored at −80°C. Extracted RNA was transferred to the Wistar Institute for further processing.

**Sample processing.** RNA purification of the Penn samples was carried out at Wistar using TriReagent (Molecular Research), as recommended and controlled for quality using the Bioanalyzer. Only samples with 28S/16S ratios of >0.75 were used for further studies. A constant amount (400 ng) of total RNA was amplified, as recommended by Illumina. The NYU samples required DNase treatment before hybridization. Samples were processed as mixed batches of cases and controls and hybridized to the Illumina WG-6v2 human whole genome bead arrays.[4]

**Array quality control and preprocessing.** All arrays were processed in the Wistar Institute Genomics Facility. Arrays were checked for outliers by computing the gene-wise, between-array, median correlation for all the arrays and comparing it with correlation for each array. An array was declared an outlier if the difference between its median correlation with other arrays versus the overall between-array median correlation was greater than eight median absolute deviations. Nonoutlier arrays were quantile normalized and background was subtracted from expression values. Noninformative probes were removed if their intensity was low relative to background in the majority of samples or if maximum ratio between any two samples was not at least 1.2. (see Supplementary Materials and Methods for details).

**Analysis.** Classification was performed using a support vector machine (SVM) with recursive feature elimination (22) using random, 10-fold, cross-validation repeated 10 times. Classification scores for each tested sample were recorded at each reduction step, down to a single gene. Average accuracy for each reduction step was calculated and all the genes at the points of maximal accuracy formed the initial discriminator, which then underwent additional reduction to form the final discriminator (see Supplementary Materials and Methods for details). Pathway analysis was carried out using Ingenuity Pathways Analysis software.[5] Significance of the changes in the SVM score before and after surgery was determined with a one-sided *t* test.

**Validation of the classifier on independent samples.** Each of the genes in the signature from SVM analysis of the microarray data identified in the training set is assigned a coefficient that defines its importance in the classifier. In validating or testing the accuracy of the signature on new samples that are not identified by class association, the analysis is carried out essentially as follows: the signature is applied as an equation of the form:

$$X = a[A] + b[B] + c[C]... + z[Z] + constant$$

where *A, B, C*, etc., are the microarray expression levels of each of the signature genes, and *a, b, c*, etc., are the coefficients by which each expression level is multiplied to give a value for X (the classification score). The expression levels of the 29 genes (*A, B, C…Z*) determined by microarray for a new patient are each multiplied by the appropriate coefficient (*a, b, c…z*) to determine a classification score, "X." If the threshold value of X is set to be zero, then patients with positive scores will be declared to have malignant disease and those with negative scores will be called nonmalignant. The higher the positive score, the greater is the confidence of malignancy, and the more negative the score, the greater is the confidence of no malignancy (Supplementary Fig. S2).

[4] http://www.illumina.com/pages.ilmn?ID=197
[5] http://www.ingenuity.com/

## Results

**Characteristics of the case and control populations.** Clinical and demographic variables for 137 NSCLC cases and 91 controls with nonmalignant lung disease, including those with pathologically diagnosed benign nodules collected at the Penn, are summarized in Table 1 and detailed in Supplementary Table S1*A* and *B*. The case and control groups were similar in terms of age, race, gender, and smoking history. Fifty-five percent of the cancer patients were stage 1, 13% were stage 2, and 32% were stages 3 and 4. Eighty-four percent of the control group and 93% of the NSCLC group were current or previous smokers. Samples used for independent validation included additional 12 cases and 15 controls collected at the NYUMC and 26 additional cases and 2 controls collected at Penn (Supplementary Table S1*C*). These samples were not included in the studies to develop a general classifier.

Flow cytometry was performed on PBMCs from 35 cases and 14 controls collected at Penn. As shown in Supplementary Table S2, there were no significant differences in the percentages of T cells, CD4 cells, B cells, monocytes, or natural killer cells. The tumor group had a slightly lower percentage of CD8 cells (18.9%) than the controls (24.5%), which did reach significance ($P = 0.03$).

**Gene expression in PBMC can identify individuals with NSCLC.** We compared gene expression profiles in PBMC samples

**Table 1.** Demographics of patients

| Category | Cases (*n* = 137) | Controls (*n* = 91) |
|---|---|---|
| Age (y) | | |
| Average | 66 | 63 |
| Median | 68 | 64 |
| Max | 84 | 88 |
| Min | 39 | 38 |
| Gender | | |
| Male | 69 | 55 |
| Female | 68 | 36 |
| Race | | |
| Caucasian | 125 | 78 |
| African-American | 11 | 11 |
| Other | 1 | 1 |
| Tobacco use | | |
| Current | 26 | 8 |
| Former | 102 | 68 |
| Never | 9 | 15 |
| Histology | | |
| Adenocarcinoma | 85 | |
| Squamous cell carcinoma | 42 | NA |
| NSCLC, NOS | 10 | |
| Cancer stage | | |
| Stage I | 75 | |
| Stage II | 18 | |
| Stage III | 39 | NA |
| Stage IV | 5 | |
| Obstructive lung disease | | |
| Yes | 63 | 65 |
| No | 65 | 17 |
| Unknown | 9 | 9 |
| Benign lung nodule | | |
| Yes | NA | 41 |
| No | | 50 |

Abbreviation: NA, not applicable.

Published OnlineFirst December 1, 2009; DOI: 10.1158/0008-5472.CAN-09-1378



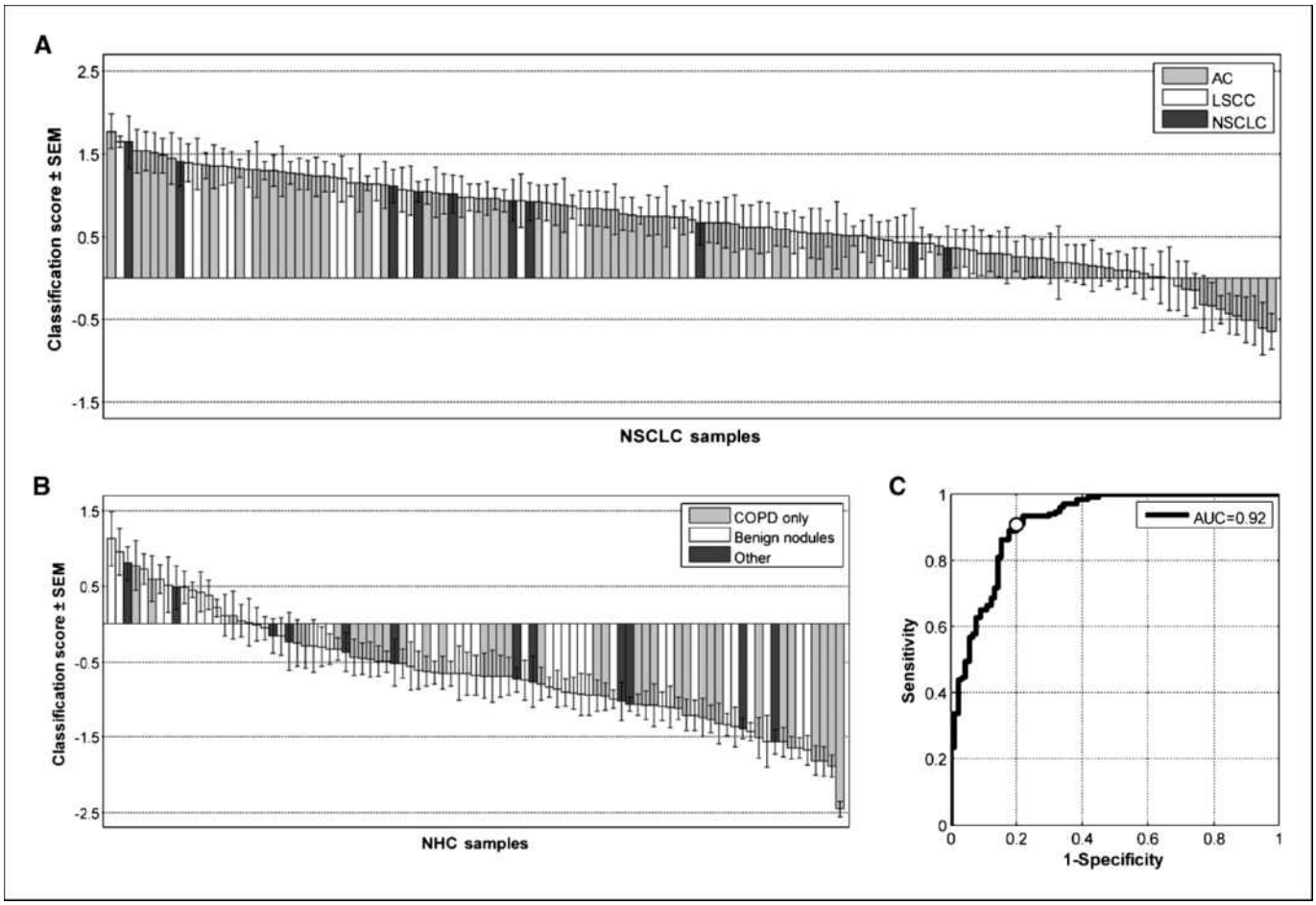


**Figure 1.** Classification scores assigned by the NSCLC classifier to 137 NSCLC patients and 91 patients with nonmalignant lung disease. A positive score indicates classification as a cancer; a negative score as a nonmalignant disease. The column heights are a measure of how well the sample is classified by the SVM algorithm for the 29 genes and the error bars are a measure of the classification variance across the 100 resamplings. *A,* NSCLC patients: *AC,* adenocarcinoma; *LSCC,* lung squamous cell carcinoma; *NSCLC,* samples not further characterized. *B,* NHCs include patients with nonmalignant lung disease: *COPD,* only COPD; *Benign nodules,* determined by biopsy; *other,* various types of lung diseases. *C,* receiver-operator characteristic curve for classification of samples shown in *A* and *B*. *AUC,* area under the curve. *White circle,* sensitivity-specificity value corresponding to classification score threshold of 0.

from the 137 NSCLC cases to 91 controls with nonmalignant lung disease. We applied a SVM with recursive feature elimination and 10-fold cross-validation (22) to the data to find the minimal number of genes that could most accurately distinguish the case and control groups by their PBMC gene expression (see Supplementary Materials and Methods and Supplementary Fig. S1). We identified a 29-gene signature that distinguished the cases from controls with an overall classification accuracy of 86%, a sensitivity of 91%, and a specificity of 80%. The distribution of SVM scores, which measure how well a particular sample is classified, is shown in Fig. 1*A* for each NSCLC patient and in Fig. 1*B* for each control. The numerical classification score of each sample, together with its clinical annotation, is listed in Supplementary Table S3. The 29 genes used for classification are listed in Table 2 ordered by their SVM score, which is a measure of each gene's contribution to the classifier.

Although an SVM score of 0 achieved the greatest degree of accuracy in separating case and control classes, additional clinical utility can be derived from these data by taking advantage of the value of the assigned SVM predictive score in the class assignments. For example, individuals with an SVM score of <–0.65 are classified as controls with 100% specificity. Similarly, an SVM threshold of +0.65 or above would eliminate 12 of 17 false positives and could identify a lung cancer case with 95% sensitivity. The scores have confidence levels that are proportionate to the score itself as shown in Supplementary Fig. S2. The receiver-operator characteristic curve (Fig. 1*C*) shows the full spectrum of performance characteristics for various cutoffs of the SVM scores. The overall area under the curve achieved by the classifier was 0.92.

To address the issue of data overfitting and to test the generality of the classification model, we also performed the analysis using only 80% of the samples for training and set aside 20% of the samples for validation. We repeated that process for five, nonoverlapping, 20% set-asides. Similar average accuracies were found over the five training sets (81.8%) and the five validation sets (81.1%; Supplementary Table S4), demonstrating the ability of the algorithm to classify new samples with the predicted accuracy. The overall accuracy is slightly reduced when using the smaller training sets (81% versus 86%). The average accuracy of the analysis with randomly permuted sample labels was 58% across 10 permutation runs.

**Classification accuracy for tumor subclasses and by smoking status with the NSCLC classifier.** We also determined the accuracy of the NSCLC classifier on histologic subtypes and clinical tumor stages (Supplementary Table S6). The sensitivity

Published OnlineFirst December 1, 2009; DOI: 10.1158/0008-5472.CAN-09-1378



for adenocarcinoma samples was 86%, whereas the squamous cell carcinomas were classified significantly better with 98% sensitivity ($P = 0.04$, $\chi^2$ test). We also determined whether classification sensitivity varied with increasing pathologic stages. As shown in Supplementary Table S6, we find a significant increase in sensitivity from stage 1A (83%) to stages 3 and 4 (100%; $P = 0.005$, $\chi^2$ test), suggesting the PBMC cancer signature becomes more pronounced with disease burden.

The accuracy of the NSCLC classifier varied slightly based on the smoking status of the participants (although there are a limited number of nonsmokers in the study population). The overall accuracy was 79%, 87%, and 88% for current, former, and never smokers, respectively (nonsignificant difference, $P = 0.28$ by Fisher exact test; the accuracy data based on smoking status and case/control status are shown in Supplementary Table S7).

The NSCLC signature was generated with controls from two different at-risk populations. About half (50) were "high risk" based on underlying lung disease and smoking history, whereas an additional 41 had been further diagnosed by CT or chest X-ray with lung nodules and were to undergo surgical evaluation. When we calculated classification accuracy for the two control populations separately, the NSCLC classifier had a specificity of 89%, if only the high-risk controls without lung nodules are considered, whereas the specificity was 71% for the controls with confirmed benign nodules. Although the difference in specificity seems to be large for these two control groups, it does not quite reach statistical significance ($P = 0.051$, Fisher Exact test), limited in part by sample numbers. However, we further explored this difference in accuracy by analyzing patients with confirmed benign nodules separately. We were able to obtain a 24-gene nodule classifier by cross-validation (Supplementary Table S5) using only the 41 benign nodule samples as the control group and data from a randomly selected group of 54 NSCLC case samples. This classifier had a somewhat better apparent specificity of 80% as determined by SVM, but the difference in accuracy between the NSCLC and nodule classifiers did not reach significance ($P = 0.44$, Fisher Exact test). Because of its higher accuracy and potentially broader applicability, the following analyses were carried out with the 29-gene NSCLC classifier.

**Validation of the NSCLC classifier on independent samples.** Although we had used cross-validation to establish our NSCLC classifier, to further validate the utility of the classifier for analyzing new samples, we assessed the classification accuracy using samples not included in the 29-gene selection process. The validation set included 38 NSCLC samples and 17 controls. Twenty-seven of the validation samples (Supplementary Table S1*C*) were collected at the NYU Lung Cancer Biomarker Center, an Early Detection Research Network Clinical and Epidemiologic Validation Center. The data set included 12 stage 1 NSCLC (5 of whom were never

**Table 2.** Twenty-nine genes that distinguish patients with NSCLC from controls with nonmalignant lung disease ordered by their contribution to the final classification score

| # | Accession | Symbol | Description | Fold change |
|---|---|---|---|---|
| 1 | NM_016578 | *RSF1* | Remodeling and spacing factor 1 | 1.27 |
| 2 | NM_003583 | *DYRK2* | Dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2 | −1.34 |
| 3 | NM_003403 | *YY1* | YY1 transcription factor | −1.08 |
| 4 | NM_001031726 | *C19orf12* | Chromosome 19 open reading frame 12 | 1.36 |
| 5 | NM_018473 | *THEM2* | Thioesterase superfamily member 2 | −1.13 |
| 6 | NM_007118 | *TRIO* | Triple functional domain (PTPRF interacting) | −1.16 |
| 7 | NM_001020820 | *MYADM* | Myeloid-associated differentiation marker | −1.34 |
| 8 | NM_017450 | *BAIAP2* | BAI1-associated protein 2 | −1.34 |
| 9 | NM_024589 | *ROGDI* | Rogdi homologue (Drosophila) | −1.18 |
| 10 | NM_024920 | *DNAJB14* | DnaJ (Hsp40) homologue, subfamily B, member 14 | −1.14 |
| 11 | NM_199191 | *BRE* | TNFRSF1A modulator | 1.04 |
| 12 | NM_080652 | *TMEM41A* | Transmembrane protein 41A | 1.15 |
| 13 | NM_032307 | *C9orf64* | Chromosome 9 open reading frame 64 | −1.14 |
| 14 | NM_031424 | *FAM110A* | Family with sequence similarity 110, member A | −1.14 |
| 15 | NM_014801 | *PCNXL2* | Pecanex-like 2 (Drosophila) | 1.21 |
| 16 | NM_005612 | *REST* | RE1-silencing transcription factor | 1.29 |
| 17 | NM_014173 | *C19orf62* | Chromosome 19 open reading frame 62 | 1.10 |
| 18 | NM_138779 | *C13orf27* | Chromosome 13 open reading frame 27 | −1.18 |
| 19 | NM_022091 | *ASCC3* | Activating signal cointegrator 1 complex subunit 3 | 1.83 |
| 20 | NM_005628 | *SLC1A5* | Solute carrier family 1 (neutral amino acid transporter), member 5 | −1.16 |
| 21 | NM_016395 | *PTPLAD1* | Protein tyrosine phosphatase-like A domain containing 1 | −1.22 |
| 22 | NM_005590 | *MRE11A* | MRE11 meiotic recombination 11 homologue A (S. cerevisiae) | −1.18 |
| 23 | NM_033107 | *GTPBP10* | GTP-binding protein 10 (putative; GTPBP10), transcript variant 2 | −1.27 |
| 24 | BX118737 | N/A | BX118737 Soares fetal liver spleen 1NFLS | −1.40 |
| 25 | NM_006217 | *SERPINI2* | Serpin peptidase inhibitor, clade I (pancpin), member 2 | −1.41 |
| 26 | AK126342 | *CREB1* | CAMP responsive element binding protein 1 | −1.45 |
| 27 | NM_016053 | *CCDC53* | Coiled-coil domain containing 53 | −1.07 |
| 28 | NM_032236 | *USP48* | Ubiquitin specific peptidase 48 | −1.17 |
| 29 | NM_001007072 | *ZSCAN2* | Zinc finger and SCAN domain containing 2 | 1.18 |

NOTE: Fold change, average change of NSCLC/NHC.

Published OnlineFirst December 1, 2009; DOI: 10.1158/0008-5472.CAN-09-1378



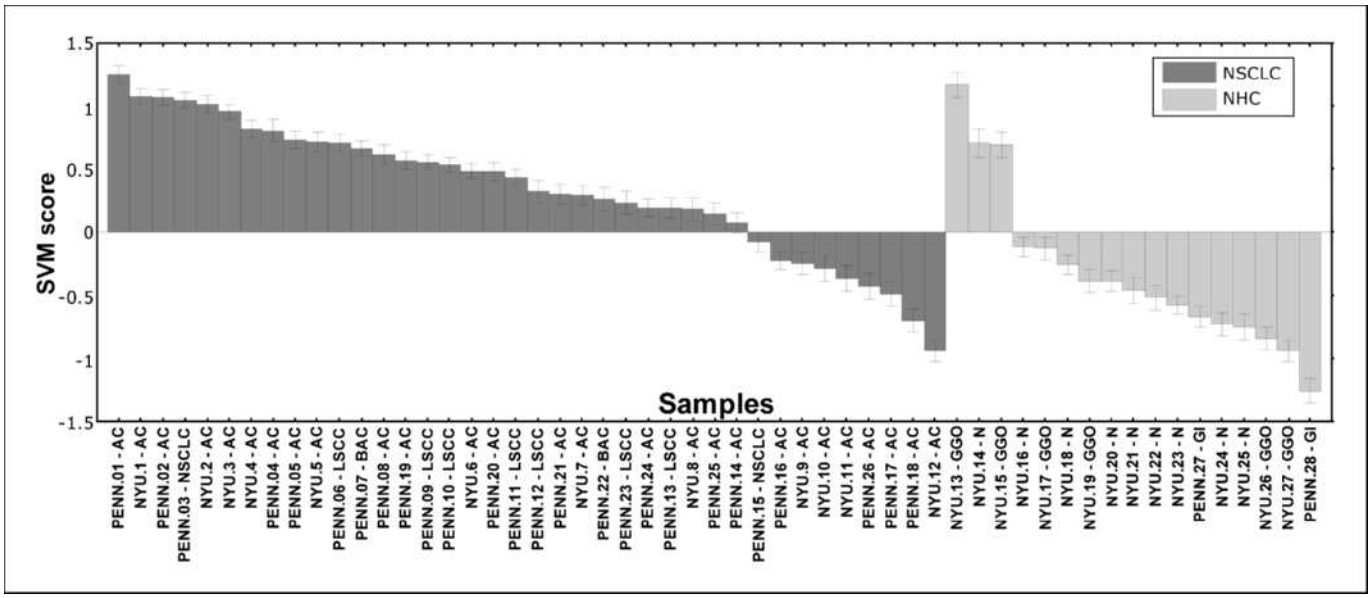


**Figure 2.** Application of the NSCLC classifier to independent validation sets. PBMC-derived RNA of lung cancer patients and controls collected at the NYU Lung Cancer Biomarker Center have labels prefaced by NYU. Lung cancer and control RNAs collected at Penn are prefaced by Penn. *IDs that end in GGO,* ground glass opacities; *GI,* granulomatous inflammation; *AC,* adenocarcinoma; *LSCC,* lung squamous cell carcinoma; *NSCLC,* non–small cell lung cancer; and *NHC,* nonhealthy control.

smokers) and 15 smoker and exsmoker controls. Six of the controls were diagnosed by serial CT scans as having nonmalignant ground glass opacities (23). No ground glass opacities patient samples were included in our original training set. The RNA for these samples was prepared at NYU. An additional of 26 patients and 2 control samples were collected at Penn and had not been analyzed previously. The NSCLC classification algorithm is applied to these samples with no knowledge of whether a sample is a case or control (see Materials and Methods). The classification for the validation set is shown in Fig. 2 and in more detail in Supplementary Table S8. The overall accuracy for the validation set was 78%, with 76% sensitivity and 82% specificity. This small decrease in accuracy and sensitivity (although with an increase in specificity) was not unexpected because the NYU samples were not specifically collected for these studies and, as a result, the sample collection and RNA purification were not standardized for these samples.

**Effect of tumor removal on individual classification scores.** Eighteen of the NSCLC patients in the validation set shown in Fig. 2 also had postresection blood samples that were collected 2 to 5 months after surgery (Supplementary Table S9). To assess how the removal of the tumor affected the NSCLC SVM score, we had determined for the presurgery samples and we also determined the scores for the postresection samples from each pair (Fig. 3). Of the 14 patients that classified as cancer in the validation set (i.e., had positive SVM scores), 13 (93%) showed a decrease in their SVM scores in the postresection samples. Five of these postsurgery samples (4, 5, 6, 10, and 13) had clearly negative SVM scores and would be classified as noncancer samples in the analysis. Of the four misclassified, presurgery patients, one showed a highly decreased score and three showed increases in their scores. Although the time intervals between the first and second samples ranged between 2 and 5 months (Supplementary Table S9), there was no

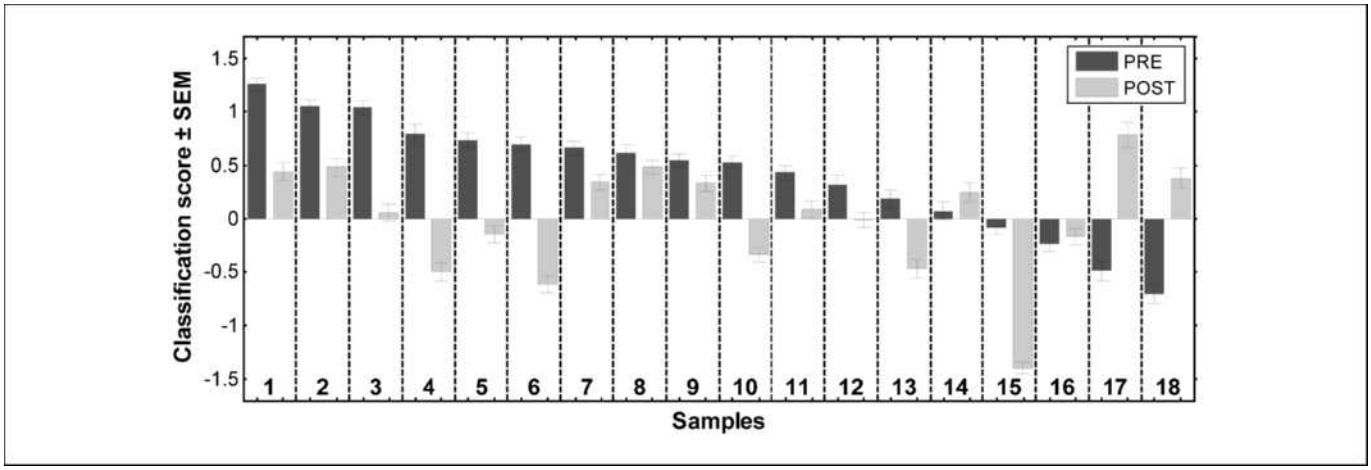


**Figure 3.** Classification scores are altered by tumor removal. The samples are arranged as paired presurgery and postsurgery samples to allow a comparison of the classification scores with the 29-gene diagnostic panel.

Published OnlineFirst December 1, 2009; DOI: 10.1158/0008-5472.CAN-09-1378



obvious relationship between the change in the scores and the time to postresection sample collection. In the large majority of the patients (14 of 18), tumor removal was associated with a decrease in the cancer signature score.

**Effect of tumor presence on expression of genes associated with immune functions.** Although 29 genes were sufficient to distinguish cancer and control classes, many more statistically significant genes were differentially expressed, providing some indication of the nature of the changes we are detecting. We used Ingenuity Core Analysis to determine the functions significantly and preferentially represented after correction for multiple testing in the top 1,000 significant genes from the NSCLC versus nonhealthy control samples (NHC), and NSCLC versus benign nodule comparisons (from a total of 2,386 and 3,276 differentially expressed genes respectively, $P < 0.05$ by $t$ test). We did both analyses to further assess the similarities and differences between the genes identified in the two comparisons. Details are in Supplementary Materials and Methods. A list of statistically significantly enriched pathways is shown in Fig. 4. As expected, pathways associated with specific immune functions are well represented, and highly significant, including pathways for *CD28* and T-cell receptor signaling, calcium-induced T-cell apoptosis, and macrophage and monocytes phagocytosis. The top five pathways by *P* value in the NSCLC/NHC comparison are also found to be significant for the NSCLC versus benign nodule comparison and rank among the top six pathways for that analysis. There were, in addition, three significantly enriched pathways that were unique to the latter comparison, stress-activated protein kinase/*c-Jun*-$NH_2$-kinase signaling, *p38* mitogen-activated protein kinase signaling, and *lymphotoxin β receptor* signaling.

In addition to identifying significant canonical pathways, we looked at genes associated with functional categories. We focused on those functional categories associated with the innate and humoral immune response, in particular, those functions associated with inflammation and infection. The overlap of genes associated with these two processes is significant. Under the functional categories of cell-mediated and humoral immunity, we found that 13 of 13 ($P$ = 9.2E-06) differentially expressed antipathogen response genes and 8 of 9 genes ($P$ = 5.04E-04) associated with the generation of reactive oxidative species, an end product of Toll-like receptor (*TLR*) activation, are downregulated in the NSCLCs compared with controls with benign nodules. In parallel, we found that 7 of 7 antibacterial response genes are downregulated in the NSCLCs compared with all NHC ($P$ = 4.15E-02). Five genes are common to the two comparisons including *TLR5*, the surface receptor for bacterial lipopolysaccahrides. *TLRs 1, 7*, and *8* are down in NSCLCs compared with either control class. We also find that genes associated with activation of the *NFκb* pathway, through which the TLR signals are transmitted (24), are down, whereas pathway inhibitory genes such as *IkB* are up in NSCLC PBMC. Recently, an important role for *TLR* functions in respiratory diseases has emerged, in particular for chronic obstructive pulmonary disease (COPD), a condition affecting the majority of both our case and control subjects (24–26), suggesting that innate response pathways are suppressed in our cancer samples despite the presence of the activating condition of COPD.

## Discussion

We previously suggested that chemokines and cytokines released by malignant cells could impose a tumor-specific signature on normal immune cells of patients with nonhematopoietic cancers (27). Gene expression profiles from PBMC that identify blood signatures associated with a variety of cancers, including metastatic melanoma (18), breast (20), renal (17, 21), and bladder cancers (19) have now been reported. However, most of these studies have focused on later-stage cancers or response to therapy and used healthy control groups for comparison. We now have identified gene expression signatures in PBMC that can distinguish patients with early-stage NSCLC from appropriate at-risk controls with nonmalignant lung diseases common to both patient and control classes.

The observed classification is not likely to be influenced by circulating tumor cells because (*a*) our classifiers do not contain genes characteristic of lung tumors such as *SFTBP* (28) or lung-specific keratins (29); and (*b*) any tumor cells would be diluted to an extraordinary degree by the PBMC without efforts to enrich for such cells. This classifier appears not to be smoking dependent. Lung cancer in individuals who have never smoked has been shown to have several important differences from tobacco-associated lung tumors, and some molecular changes have been suggested to be unique to nonsmokers (30, 31). There were 14 NSCLC patients in our study that had no prior history of smoking. Despite this, 11 of the 14 "never" smokers in our data set were correctly classified as cancer by our NSCLC panel.

The mechanism(s) for the effect we have detected remains to be determined. Interactions between the tumor and immune cells could be direct or mediated by cytokines or other tumor-released factors. The effects are enhanced with tumor progression, as evidenced by the increased accuracy of our gene panel in classifying late-stage NSCLC. Our ability to build a classifier from peripheral immune cells is consistent with recent findings from both mouse models and studies of immune suppression by tumors in humans. For example, Redente and colleagues (32) showed, in a mouse lung cancer model, that soluble factors produced in lung premalignant lesions influenced expression of specific macrophage activation markers in bone marrow macrophages and that the effect on gene expression was enhanced with tumor progression. The ability of tumors to induce myeloid-derived suppressor cells in lymph nodes, spleens, and peripheral blood in mouse models is now well established (33–35). The observation that tumor resection results in disappearance of these myeloid-derived suppressor cells (36) supports our observations that the PBMC tumor signature diminished after tumor removal in the majority of the patients we examined. Similar tumor-induced suppressor cells in the PBMC fraction of blood also have been identified in human cancer patients (37, 38). Evidence from recent studies, comparing gene expression in PBMC and tumor-infiltrating lymphocytes from patients with either liver cirrhosis alone or in conjunction with liver cancer, suggests that the tumor presence can be communicated to the peripheral immune system and that the signal can be detected in the PBMC gene expression patterns (39). These observations support our finding that the NSCLC signature detected in PBMC diminishes in a majority of postsurgery patients.

The five pathways most significantly represented among the top 1,000 differentially expressed genes between cases and controls were significant for both the comparison of NSCLC and all controls and for the comparison of NSCLC and nodule controls. There is significant, but not complete, overlap in the genes associated with these five pathways for the two comparisons. For three of the pathways (1, 2, and 5), <50% of the genes are common to both comparisons. Clearly there are significant similarities as well as some differences in the two comparisons we have carried out to identify

Published OnlineFirst December 1, 2009; DOI: 10.1158/0008-5472.CAN-09-1378



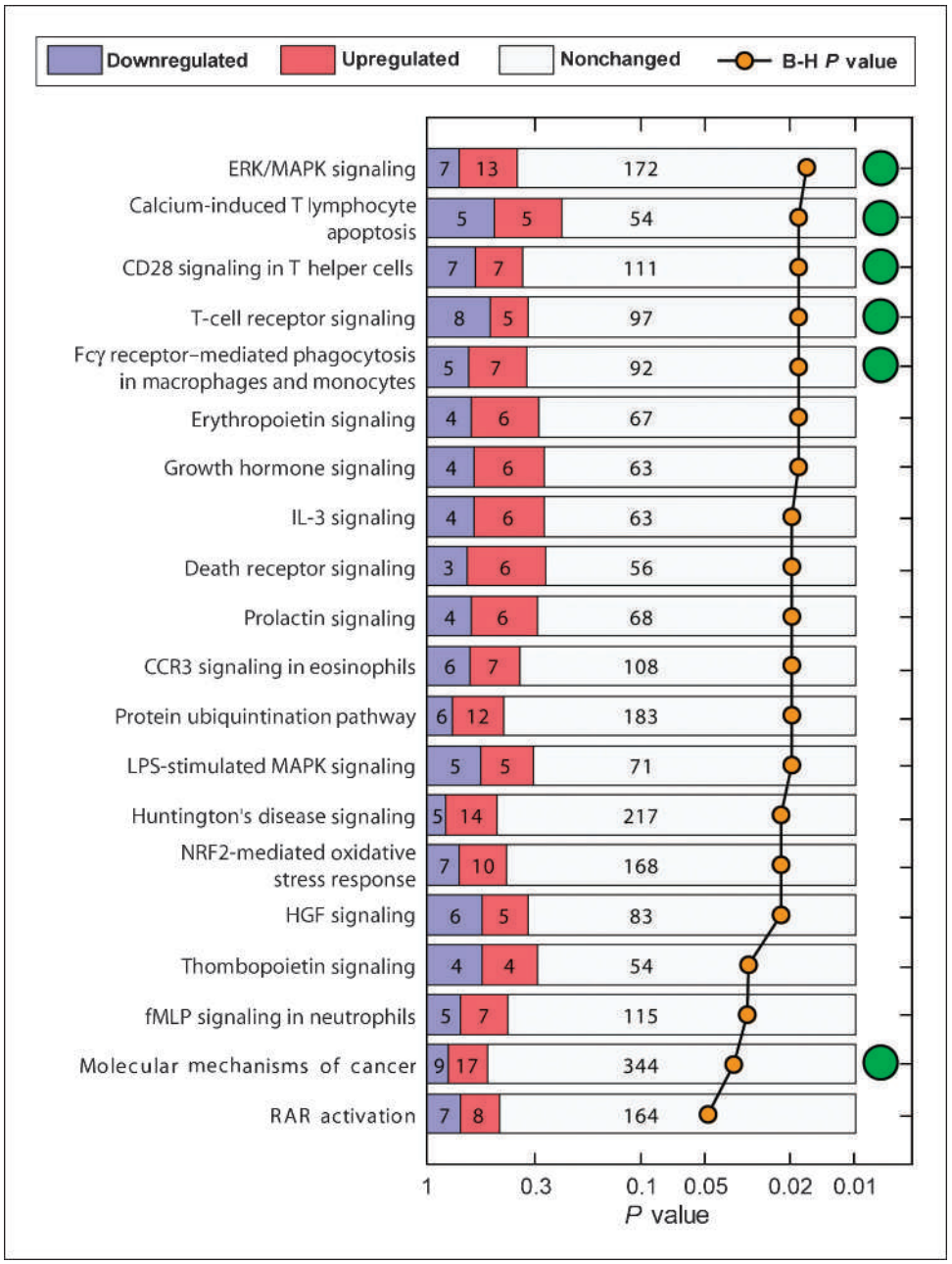


**Figure 4.** Significantly enriched canonical pathways from Ingenuity Pathway Analysis of the genes differentially regulated between NSCLC and NHC samples. *Numbers in the bars,* the number of genes in the pathway significantly higher in cancer (*red*) or lower in cancer (*blue*). *B-H,* Benjamini-Hochberg multiple testing correction. *Green circles,* pathways that were also enriched in NSCLC versus benign nodule comparison.

our NSCLC general classifier. Recent studies have suggested that although diagnostic genes detected in various pathways may vary, the pathways themselves are better classifiers (40, 41).

We also identified some interesting differences between cases and controls in relation to immune response functional categories. The reduction in *TLR* expression in NSCLC was somewhat surprising as a high proportion of our patients and controls have COPD, which would normally be expected to have activated *TLR* pathways (25). *TLR* function has been studied primarily in response to pathogens but a more expansive role in immunoregulation has been emerging for recognition of self-antigens associated with autoimmunity (42–45). In addition, endogenous ligands for TLRs have been identified including *MUC1*, a tumor expressed antigen that has been shown to be a negative regulator of *TLR* signaling (46) and heat shock proteins (47–51).

Our study follows the paradigm for biomarker development described by Pepe and colleagues (52) and adopted by the National Cancer Institute Early Detection Research Network. This paradigm first outlines the use of cross-sectional studies of patients with cancer versus appropriately chosen controls without disease to document initial estimates of sensitivity and specificity. Biomarkers meeting appropriate thresholds are then to be tested in external populations and finally in prospective studies. Following this model, our first analysis showed that a 29-gene panel could differentiate between a lung cancer population and an appropriate at-risk control population. Additional validation studies were then carried out on an external, independent data set. Plans for prospective studies are in progress.

Although the NSCLC signature could be developed as a screening tool for high-risk patients, the initial clinical use of our biomarkers

Published OnlineFirst December 1, 2009; DOI: 10.1158/0008-5472.CAN-09-1378



is more likely to provide additional data to a clinician trying to evaluate a pulmonary nodule diagnosed by CT scan or chest X-ray. Based on prevalence data from a large CT screening study (3), the 29-gene NSCLC classifier has a positive predictive value of 0.06 and a negative predictive value of 1.00 (Supplementary Table S10; ref. 3). This is comparable with the positive predictive and negative predictive values calculated using the same prevalence values for the 80-gene classifier derived from lung epithelial cells obtained from bronchial brushing recently described by Spira and colleagues (16).

Because higher SVM score increases the likelihood of a sample being cancer, the specific SVM value may be useful for clinical decision making in patients with suspected lung cancer or a noncalcified nodule and thus could help determine which patients require immediate interventions such as biopsy or surgical resection. This could potentially decrease the number of patients with benign lung nodules that would otherwise undergo biopsy or surgery (i.e., false positives).

Our results represent an encouraging first step, but several tasks remain to be addressed. Additional external validation sets are required to establish a standard collection protocol and to confirm the gene signatures and their accuracy. A larger prospective cohort study in patients with lung nodules is needed to more fully determine the role of smoking or other potentially confounding effects or diseases and to evaluate the overall clinical feasibility and utility of this approach. In addition, the observed reduction of the NSCLC cancer signature in the postsurgery samples suggests the possibility that postsurgery gene expression profiles might contain information predictive of recurrence. Ongoing follow-up studies are being conducted to determine the applicability of our approach to recurrence and response to therapy.

In summary, we have found gene expression signatures in PBMC that can distinguish individuals with early-stage NSCLC from individuals with nonmalignant lung disease. The changes in PBMC gene expression with tumor removal suggest some specific functional effects of the tumor on the immune system that can be detected in the gene expression profiles. Although we have only examined NSCLC in this study, other types of lung cancer also may be detectable by gene expression in the peripheral immune cells.

Gene expression data are available in the gene expression omnibus. The index code is GSE13255.

## Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

## Acknowledgments

Received 4/13/09; revised 8/25/09; accepted 9/9/09; published OnlineFirst 12/1/09.

**Grant support:** PA DOH Tobacco Settlement grants SAP 4100020718 and 4100038714, the PA DOH Commonwealth Universal Research Enhancement Program, Early Detection Research Network Set-Aside funds, and the Wistar Cancer Center Support Grant P30 CA010815. A. Vachani was supported by NCI K07 CA111952.

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.

We thank WenHwai Horng, Linda Alila, and Shere Billouin for technical assistance and support from the Genomics and Bioinformatics Cores.

## References

1. ACS. Cancer facts and figures 2007. Atlanta: American Cancer Society; 2008.
2. Diederich S, Wormanns D. Impact of low-dose CT on lung cancer screening. Lung Cancer 2004;45 Suppl 2: S13–9.
3. Henschke CI, Yankelevitz DF, Libby DM, Pasmantier MW, Smith JP, Miettinen OS. Survival of patients with stage I lung cancer detected on CT screening. N Engl J Med 2006;355:1763–71.
4. Jett JR. Limitations of screening for lung cancer with low-dose spiral computed tomography. Clin Cancer Res 2005;11:4988–92s.
5. Mulshine JL. Current issues in lung cancer screening. Oncology (Huntingt) 2005;19:1724–30; discussion 30–1.
6. Bach PB, Jett JR, Pastorino U, Tockman MS, Swensen SJ, Begg CB. Computed tomography screening and lung cancer outcomes. JAMA 2007;297:953–61.
7. Deppermann KM. Lung cancer screening-where we are in 2004 (take home messages). Lung Cancer 2004; 45 Suppl 2:S39–42.
8. Ikeda K, Awai K, Mori T, Kawanaka K, Yamashita Y, Nomori H. Differential diagnosis of ground-glass opacity nodules: CT number analysis by three-dimensional computerized quantification. Chest 2007; 132:984–90.
9. Machida EO, Brock MV, Hooker CM, et al. Hypermethylation of ASC/TMS1 is a sputum marker for late-stage lung cancer. Cancer Res 2006;66:6210–8.
10. Gao WM, Kuick R, Orchekowski RP, et al. Distinctive serum protein profiles involving abundant proteins in lung cancer patients based upon antibody microarray analysis. BMC Cancer 2005;5:110.
11. Patz EF, Jr., Campa MJ, Gottlin EB, Kusmartseva I, Guan XR, Herndon JE II. Panel of serum biomarkers for the diagnosis of lung cancer. J Clin Oncol 2007;25: 5578–83.
12. Yanagisawa K, Shyr Y, Xu BJ, et al. Proteomic patterns of tumour subsets in non-small-cell lung cancer. Lancet 2003;362:433–9.
13. Brichory FM, Misek DE, Yim AM, et al. An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. Proc Natl Acad Sci U S A 2001;98:9824–9.
14. Pontes ER, Matos LC, da Silva EA, et al. Autoantibodies in prostate cancer: humoral immune response to antigenic determinants coded by the differentially expressed transcripts FLJ23438 and VAMP3. Prostate 2006;66:1463–73.
15. Belinsky SA, Liechty KC, Gentry FD, et al. Promoter hypermethylation of multiple genes in sputum precedes lung cancer incidence in a high-risk cohort. Cancer Res 2006;66:3338–44.
16. Spira A, Beane JE, Shah V, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nat Med 2007;13:361–6.
17. Burczynski ME, Twine NC, Dukart G, et al. Transcriptional profiles in peripheral blood mononuclear cells prognostic of clinical outcomes in patients with advanced renal cell carcinoma. Clin Cancer Res 2005;11: 1181–9.
18. Critchley-Thorne RJ, Yan N, Nacu S, Weber J, Holmes SP, Lee PP. Down-regulation of the interferon signaling pathway in T lymphocytes from patients with metastatic melanoma. PLoS Med 2007;4:e176.
19. Osman I, Bajorin DF, Sun TT, et al. Novel blood biomarkers of human urinary bladder cancer. Clin Cancer Res 2006;12:3374–80.
20. Sharma P, Sahni NS, Tibshirani R, et al. Early detection of breast cancer based on gene-expression patterns in peripheral blood cells. Breast Cancer Res 2005;7: R634–44.
21. Twine N, Stover J, Marshall B, et al. Disease-associated expression profiles in peripheral blood mononuclear cells from patients with advanced renal cell carcinoma. Cancer Res 2003;6:6069–75.
22. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002;46:389–422.
23. Ohta Y, Shimizu Y, Kobayashi T, et al. Pathologic and biological assessment of lung tumors showing ground-glass opacity. Ann Thorac Surg 2006;81:1194–7.
24. Brody JS, Spira A. State of the art. Chronic obstructive pulmonary disease, inflammation, and lung cancer. Proc Am Thorac Soc 2006;3:535–7.
25. Pan MM, Sun TY, Zhang HS. [Expression of toll-like receptors on CD14+ monocytes from patients with chronic obstructive pulmonary disease and smokers]. Zhonghua Yi Xue Za Zhi 2008;88:2103–7.
26. Sabroe I, Whyte MK. Toll-like receptor (TLR)-based networks regulate neutrophilic inflammation in respiratory disease. Biochem Soc Trans 2007;35:1492–5.
27. Kari L, Loboda A, Nebozhyn M, et al. Classification and prediction of survival in patients with the leukemic phase of cutaneous T cell lymphoma. J Exp Med 2003; 197:1477–88.
28. Vachani A, Nebozhyn M, Singhal S, et al. A 10-gene classifier for distinguishing head and neck squamous cell carcinoma and lung squamous cell carcinoma. Clin Cancer Res 2007;13:2905–15.
29. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 2001;98:13790–5.
30. Subramanian J, Govindan R. Lung cancer in never smokers: a review. J Clin Oncol 2007;25:561–70.
31. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers-a different disease. Nat Rev Cancer 2007;7: 778–90.
32. Redente EF, Orlicky DJ, Bouchard RJ, Malkinson AM. Tumor signaling to the bone marrow changes the phenotype of monocytes and pulmonary macrophages

Published OnlineFirst December 1, 2009; DOI: 10.1158/0008-5472.CAN-09-1378



during urethane-induced primary lung tumorigenesis in A/J mice. Am J Pathol 2007;170:693–708.

**33.** Marigo I, Dolcetti L, Serafini P, Zanovello P, Bronte V. Tumor-induced tolerance and immune suppression by myeloid derived suppressor cells. Immunol Rev 2008; 222:162–79.

**34.** Serafini P, Borrello I, Bronte V. Myeloid suppressor cells in cancer: recruitment, phenotype, properties, and mechanisms of immune suppression. Semin Cancer Biol 2006;16:53–65.

**35.** Sinha P, Clements VK, Bunt SK, Albelda SM, Ostrand-Rosenberg S. Cross-talk between myeloid-derived suppressor cells and macrophages subverts tumor immunity toward a type 2 response. J Immunol 2007;179:977–83.

**36.** Salvadori S, Martinelli G, Zier K. Resection of solid tumors reverses T cell defects and restores protective immunity. J Immunol 2000;164:2214–20.

**37.** Diaz-Montero C, Salem M, Nishimura M, Garrett-Mayer E, Cole D, Montero A. Increased circulating myeloid-derived suppressor cells correlate with clinical cancer stage, metastatic tumor burden, and doxorubicin-cyclophosphamide chemotherapy. Cancer Immunol Immunother 2009;58:49–59.

**38.** Kusmartsev S, Su Z, Heiser A, et al. Reversal of myeloid cell-mediated immunosuppression in patients with metastatic renal cell carcinoma. Clin Cancer Res 2008; 14:8270–8.

**39.** Sakai Y, Honda M, Fujinaga H, et al. Common transcriptional signature of tumor-infiltrating mononuclear inflammatory cells and peripheral blood mononuclear cells in hepatocellular carcinoma patients. Cancer Res 2008;68:10267–79.

**40.** Efroni S, Schaefer CF, Buetow KH. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. PLoS ONE 2007;2:e425.

**41.** Lee E, Chuang H-Y, Kim J-W, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol 2008;4:e1000217.

**42.** Li M, Zhou Y, Feng G, Su SB. The critical role of Toll-like receptor signaling pathways in the induction and progression of autoimmune diseases. Curr Mol Med 2009;9:365–74.

**43.** Fischer M, Ehlers M. Toll-like receptors in autoimmunity. Ann N Y Acad Sci 2008;1143:21–34.

**44.** Krieg AM, Vollmer J. Toll-like receptors 7, 8, and 9: linking innate immunity to autoimmunity. Immunol Rev 2007;220:251–69.

**45.** Ehlers M, Ravetch JV. Opposing effects of Toll-like receptor stimulation induce autoimmunity or tolerance. Trends Immunol 2007;28:74–9.

**46.** Ueno K, Koga T, Kato K, et al. MUC1 mucin is a negative regulator of toll-like receptor signaling. Am J Respir Cell Mol Biol 2008;38:263–8.

**47.** Chen K, Huang J, Gong W, Iribarren P, Dunlop NM, Wang JM. Toll-like receptors in inflammation, infection and cancer. Int Immunopharmacol 2007;7: 1271–85.

**48.** Qazi KR, Oehlmann W, Singh M, Lopez MC, Fernandez C. Microbial heat shock protein 70 stimulatory properties have different TLR requirements. Vaccine 2007;25:1096–103.

**49.** Tsan MF, Gao B. Endogenous ligands of Toll-like receptors. J Leukoc Biol 2004;76:514–9.

**50.** Tsan MF, Gao B. Heat shock proteins and immune system. J Leukoc Biol 2009;85:905–10.

**51.** Vabulas RM, Wagner H, Schild H. Heat shock proteins as ligands of toll-like receptors. Curr Top Microbiol Immunol 2002;270:169–84.

**52.** Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93:1054–61.

Cancer Research
The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

AACR American Association for Cancer Research

# Gene Expression Profiles in Peripheral Blood Mononuclear Cells Can Distinguish Patients with Non–Small Cell Lung Cancer from Patients with Nonmalignant Lung Disease

Michael K. Showe, Anil Vachani, Andrew V. Kossenkov, et al.

*Cancer Res* 2009;69:9202-9210. Published OnlineFirst December 1, 2009.