Published in final edited form as:

Cancer Prev Res (Phila). 2011 October ; 4(10): 1599–1608. doi:10.1158/1940-6207.CAPR-10-0170.

A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma

Melissa Rotunno¹, Nan Hu¹, Hua Su¹, Chaoyu Wang¹, Alisa M. Goldstein¹, Andrew W. Bergen^{1,2}, Dario Consonni³, Angela C Pesatori³, Pier Alberto Bertazzi³, Sholom Wacholder¹, Joanna Shih⁴, Neil E. Caporaso¹, Phil R. Taylor¹, and Maria Teresa Landi¹

¹Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland ²Center for Health Sciences, SRI International, Menlo Park, California ³Unit of Epidemiology, Fondazione IRCCS Ospedale Maggiore Policlinico and Department of Occupational and Environmental Health, Università degli Studi di Milano, Milan, Italy ⁴Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland

Abstract

Affordable early screening in subjects with high risk of lung cancer has great potential to improve survival from this deadly disease. We measured gene expression from lung tissue and peripheral whole blood (PWB) from adenocarcinoma cases and controls to identify dysregulated lung cancer genes that could be tested in blood to improve identification of at-risk patients in the future. Genome-wide mRNA expression analysis was conducted in 153 subjects (73 adenocarcinoma cases, 80 controls) from the Environment And Genetics in Lung cancer Etiology (EAGLE) study using PWB and paired snap-frozen tumor and non-involved lung tissue samples. Analyses were conducted using unpaired t-tests, linear mixed effects and ANOVA models. The area under the receiver operating characteristic curve (AUC) was computed to assess the predictive accuracy of the identified biomarkers. We identified 50 dysregulated genes in stage I adenocarcinoma *versus* control PWB samples (False Discovery Rate ≤ 0.1 , fold change ≥ 1.5 or ≤ 0.66). Among them, eight (*TGFBR3*, *RUNX3*, *TRGC2*, *TRGV9*, *TARP*, *ACPI*, *VCAN*, and *TSTA3*) differentiated paired tumor *versus* non-involved lung tissue samples in stage I cases, suggesting a similar pattern of lung cancer-related changes in PWB and lung tissue. These results were confirmed in two independent gene expression analyses in a blood-based case-control study (n=212) and a tumor-non tumor paired tissue study (n=54). The eight genes discriminated patients with lung cancer from healthy controls with high accuracy (AUC=0.81, 95% CI=0.74–0.87). Our finding suggests the use of gene expression from PWB for the identification of early detection markers of lung cancer in the future.

Keywords

microarray gene expression; peripheral blood; lung cancer; stage I

Corresponding Author: Maria Teresa Landi, MD, PhD, landim@mail.nih.gov, National Cancer Institute, 6120 Executive Blvd., MSC 7114, Bethesda, MD 20892-7248, Telephone: 301-402-9519, Fax: 301-402-4489.

Conflicts of interest: The authors declare no conflicts of interest.

Introduction

Lung cancer causes more deaths than any other cancer in both men and women, with over 160,000 deaths annually in the United States and one million worldwide (1). Unfortunately, the average 5-year survival rate has remained relatively stable at 15% over many decades, due to minimal improvements in early detection and treatment. Non-invasive assays for detection of lung cancer at a curable stage could offer the best therapeutic option for these patients. Although promising, imaging techniques such as low-dose helical computed tomography are expensive and potentially associated to increased risk due to ionizing radiation exposure. Blood-based biomarker assays are a potentially important alternative non-invasive method to screen for lung cancer. Technological advances in methods of blood collection and RNA stabilization have only recently increased our ability to detect transcript levels in gene expression studies of human blood samples. Recent studies of gene expression from blood cells have successfully identified gene signatures for diverse exposures (e.g., tobacco smoking (2) or benzene (3)), and health conditions, including autoimmune disorders (4–6), inflammatory diseases (7), and cancer (8–11). In our study, we first compared gene expression changes in blood between adenocarcinoma cases and non-cancer controls to select the genes whose expression mostly differentiated cases from controls. We then compared this signature in paired adenocarcinoma *vs.* non-involved lung tissue samples to identify the subset of genes differentiating both cases/controls (blood samples) and tumor/non-tumor (tissue samples). These expression changes could be specifically due to early development of cancer. Finally, we validated the overlapping gene expression signature in additional blood-based and tissue-based independent studies. If confirmed in prospective studies, gene expression changes from blood tests can provide a useful tool for the early detection of cancer in at-risk individuals.

Materials and Methods

Overview of Strategy

Our study design included three phases. **(i)** First, we aimed to identify molecular changes in blood due to cancer, by comparing stage I adenocarcinoma cases (n=26) to controls (n=80). We restricted the analysis to stage I cases to focus on early molecular changes not affected by systematic metabolic disruption such as weight loss or other sequelae of advanced disease. We then verified whether these gene changes in PWB were present also in later stages (n=47). Since tobacco smoking is the most important risk factor for lung cancer (12) and has been associated with lung cancer progression (13), we also explored potentially distinct gene signatures by smoking groups. **(ii)** We then compared the blood-related gene expression signature distinguishing stage I cases from controls with the gene expression signature differentiating fresh frozen paired tumor *versus* non-involved tissue samples in a subgroup (n=15) of the same stage I cases. With this comparison we aimed to identify expression changes in PWB due to lung cancer that paralleled changes in the target organ. **(iii)** Finally, we sought to validate the main results using a) qRT-PCR analysis from PWB for all identified genes in additional 82 stage I adenocarcinoma patients and 130 controls from the same population and b) microarray gene expression data for all identified genes in 54 lung adenocarcinoma and non-involved paired tissue samples from a previously published independent study (14).

Population

Individuals with lung adenocarcinoma (n=73 for the microarray experiment; n=82 for qRT-PCR validation) and healthy controls (n=80 for the microarray experiment; n=130 for qRT-PCR validation) were randomly sampled from a large, well-defined population-based case-control study, the Environment And Genetics in Lung cancer Etiology (EAGLE) study (15–

21), including 2,100 consecutive incident lung cancer cases and 2,120 controls (all Caucasians) from Italy. Selected cases had histologically confirmed primary adenocarcinoma of the lung including all stages and controls were matched to cases by age, sex, and smoking status (never, former, and current smoking). For the validation set we focused on current smoker stage I cases and controls. Detailed subjects' characteristics are described in Table 1.

The study was approved by the Institutional Review Board (IRB) of each participating institutions in Italy and by the National Cancer Institute, Bethesda, MD. All participants signed an informed consent.

Blood and tissue collection for RNA extraction

PWB was collected for all EAGLE participants (after lung cancer diagnosis and before treatment for cases, and at enrollment for controls) using the Paxgene® Blood RNA System (PreAnalytiX, Hombrechtikon, CH) containing a proprietary solution that reduces RNA degradation and gene induction (22;23). Fresh lung tissue samples were snap-frozen within 20 minutes of surgical resection.

Microarray gene expression data

Data on microarray gene expression from PWB were obtained using the Affymetrix GeneChip® HG-U133A v2.0. After exclusion of two samples with poor quality profile (see quality assessment in Supplemental Material 1), the remaining 162 samples were processed and normalized with the Robust Multichip Average (RMA) method. Corresponding CEL files and information conform to the MIAME guidelines are publicly available on the GEO database (accession number GSE20189). Nine subjects were excluded after data normalization because of reclassification to non-adenocarcinoma morphology during histologic review. The final analyses were based on 73 adenocarcinoma cases and 80 controls. All 22,277 probe sets based on RMA summary measures were used in the analyses.

The detailed description of the gene expression examination of lung tissues in lung cancer cases in EAGLE (also based on the Affymetrix HG-U133A GeneChip®) and sample inclusion and exclusion criteria have been published previously (24). For the present study we used data from paired tumor and non-involved lung tissue samples from 15 of the same stage I adenocarcinoma cases included in the PWB based study.

The validation lung tissue set consisted of 27 tumor and 27 non-involved paired lung tissue samples from a previously published independent study (14). Details of the specimens, mRNA processing and hybridization, and data access are described in the relative publication (14).

qRT-PCR gene expression data

We followed the procedure described in Hu N *et al.* (25). Briefly, RNA quality and quantity was determined using the RNA 6000 Labchip/Aligent 2100 Bioanalyzer. RNA purification was performed according to the manufacturer's instructions (Qiagen Inc.). After reverse transcription of RNA, all real-time PCR reactions were performed using an ABI Prism 7000 Sequence Detection System with the designed primers and probes for target genes and an internal control gene, glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*). Each sample for each gene was run in triplicate. Quantitative methods require that PCR efficiencies be similar for all genes and $\geq 90\%$. Efficiency was measured using a standard curve generated by serial dilutions of the RNA as described in <http://docs.appliedbiosystems.com/search>.

Statistical analysis

(i) A two sample t-test was conducted to test whether blood RNA expression differed between cases and controls (overall and stratified by stage and by smoking status). Age, sex, and smoking variables were similarly distributed across the groups (Table 1) and were not associated with the expression of the 61 selected gene targeting probes (gene-probes) among controls or cases. Analyses adjusted or unadjusted for these factors provided almost identical results. Unadjusted results are shown throughout the paper. We used the Benjamini-Hochberg procedure (26) to calculate the False Discovery Rate (FDR) to adjust for the ~22,000 comparisons and only further considered results with a maximum $FDR \leq 0.1$ (based on single gene-probe p-value threshold of 0.001). In addition, only gene-probes with a fold change (FC) ≤ 0.66 for down-regulated gene-probes or ≥ 1.5 for up-regulated gene-probes were considered for follow-up in subsequent analyses. (ii) Since significantly fewer hypotheses (61 probes) were tested in the following analyses, less stringent significance criteria were applied (p-value < 0.005). For analyses of tumor *versus* non-involved paired tissues from the same subjects, a linear mixed effects model was used to account for intra-person correlation. Gene-probes with p-value < 0.005 and same FC direction and intensity (i.e., $FC \leq 0.66$ or ≥ 1.5) as in the case/control blood RNA comparison were selected for validation analyses. (iii) To validate the significant results, we analyzed: a) the qRT-PCR gene expression PWB-based data using the $2^{-\Delta\Delta C_t}$ method (27) to compare cases to controls and b) the microarray gene expression tissue-based data using linear mixed effects model to compare tumor to non-involved paired lung tissue samples. In addition, receiver operating characteristic (ROC) analysis was performed on the PWB-based validation data and the area under the curve (AUC) was estimated to assess the accuracy of the identified biomarkers, both individually and combined, in discriminating between lung cancer patients and controls.

All statistical analyses were conducted using R program language v2.10.

Results

(i) We compared mRNA expression from PWB in stage I adenocarcinoma cases *versus* controls, in the overall sample and stratified by smoking categories (Table 2). Two significant gene signatures in stage I cases were detected: one in the combined smokers and non smokers ($FDR=0.10$) and the second among current smokers only ($FDR=0.15$). No significant results were found within subsets of former or never smokers ($FDR=0.97$ and 1.00 , respectively). However, gene expression changes significant in the analysis among current smokers showed similar, although not significant, trends in the analyses among never and former smokers (data not shown). At the same time, the analysis among current smokers revealed distinct alterations, which might be particularly important for individuals who smoke. Thus, for the comparison of stage I cases *vs.* controls, we considered both results from all subjects and from current smokers only (221 and 144 gene-probes, respectively, 81 overlapping between the two). To increase specificity, we restricted the successive analyses to gene-probes with $FC \leq 0.66$ or ≥ 1.5 . The resulting 25 down-regulated gene-probes (20 genes), and 36 up-regulated genes-probes (30 genes), are shown in the heatmap of Figure 1 and Supplemental Material 2. In general, fold changes were stronger in the analysis restricted to current smokers than in the overall analysis. Since there was no significant difference between cigarette per day or cumulative pack-years between cases and controls (Table 1) and the analysis adjusted by these covariates provided almost identical results, our findings are unlikely due to differences in smoking quantity between cases and controls. We verified whether the identified 61 gene-probes were also differentially expressed between cases and controls in late stage disease. FCs were consistently stronger in the analysis limited to stage I cases, but had concordant directions in all groups analyzed (Figure 3).

(ii) We aimed to identify changes in gene expression related to early stage lung cancer that are detectable in both blood cells and lung tissue cells. Thus, for the 61 gene-probes (50 genes) in the analysis of stage I patients and controls (Figure 1 and Supplemental Material 2), we compared gene expression in tumor *versus* paired non-involved lung tissue samples in 15 stage I adenocarcinoma cases. We found that 10 probes from 8 genes (*TGFBR3*, *RUNX3*, *TRGC2*, *TRGV9*, *TARP*, *ACPI*, *VCAN*, and *TSTA3*) were differentially expressed (p -values ≤ 0.003) in tumor compared to non-involved lung tissue samples and in the same direction and intensity as in stage I adenocarcinoma cases compared to controls (Figure 2).

(iii) We validated the PWB-based gene expression differences in stage I cases compared to controls using qRT-PCR measurements of RNA extracted from PWB of additional 82 stage I adenocarcinoma patients and 130 controls from EAGLE. Each gene was covered by a single ABI probe with the exception of *TARP*, covered by both the *TRGC2* and the *TRGV9* ABI probes, due to overlap between these 3 genes. Results were strongly confirmed for all examined genes: *RUNX3*, *TGFBR3*, *TRGC2/TARP*, and *TRGV9/TARP* were significantly down-regulated in stage I lung cancer patients compared to controls (FCs = 0.6, 0.5, 0.5, 0.6, P -values = 1.0×10^{-7} , 1.4×10^{-8} , 3.4×10^{-7} , 2.6×10^{-6} , respectively) and *VCAN*, *ACPI* and *TSTA3* were significantly up-regulated in stage I lung cancer patients compared to controls (FCs = 1.2, 1.2, 1.3, P -values = 5.0×10^{-3} , 5.0×10^{-3} , 3.0×10^{-3} , respectively). We then validated gene expression differences between tumor and non-involved lung tissue samples for all 8 genes using microarray gene expression data from a previously published dataset (14), which included 27 adenocarcinoma and non-involved paired lung tissue samples. The direction of changes was 100% consistent with our original finding: *RUNX3*, *TGFBR3*, *TRGV9*, *TARP*, and *TRGC2* were significantly down-regulated in tumor compared to non-involved tissues (FCs = 0.7, 0.2, 0.3, 0.6, 0.5, P -values = 0.06, 3.0×10^{-11} , 4.0×10^{-7} , 3.4×10^{-5} , 2.5×10^{-6} , respectively) and *VCAN*, *ACPI* and *TSTA3* were significantly up-regulated in tumor compared to non-involved tissues (FCs = 2.6, 1.5, 2.5, P -values = 0.002, 5.7×10^{-5} , 3.8×10^{-9} , respectively). We evaluated the ability of PWB-based expression of each gene to discriminate lung cancer patients from controls in the validation set by means of receiver operating characteristic (ROC) curves (Figure 4). The area under the curve (AUC) ranged from 0.55 (95% CI = 0.46–0.64) for *ACPI* to 0.73 (95% CI = 0.66–0.81) for *TGFBR3* (Figure 4), thus indicating a reasonable discrimination power between lung cancer cases and controls for most genes when considered individually. In addition, a combination of all markers based on a logistic regression model showed the best diagnostic accuracy with an AUC of 0.81 (95% CI=0.74–0.87, red ROC curve in Figure 4).

Discussion

We identified a gene expression signature from blood samples consisting of 8 genes (*RUNX3*, *TGFBR3*, *TRGC2*, *TRGV9*, *TARP*, *ACPI*, *VCAN* and *TSTA3*) that differentiates stage I lung adenocarcinoma cases from controls and mirrors cancer-related gene expression changes in the target tissue. Results were validated in additional independent sets of tissue-based and blood-based gene expression analyses of adenocarcinoma cases and controls. Although present in all stages, expression changes were weaker in advanced stages, possibly because of secondary changes due to the spread of the disease. Similarly, the changes were stronger in current smokers but present in all smoking categories. The accuracy in discriminating between stage I lung adenocarcinoma cases and controls was good for most genes when considered separately, in particular those that were down-regulated between cases and controls. A multiplex model based on the expression of all 8 genes combined showed a high diagnostic accuracy of 81% (Figure 4). If further validated in prospective studies using PWB of cases drawn prior to lung cancer diagnosis (28), this gene expression signature may be used as a blood-based biomarker for early detection of lung adenocarcinoma in heavy smokers at high risk of lung cancer. We validated its use in

current smokers. Further study in never and former smokers is warranted. Moreover, it will be important to test the identified biomarkers in other lung cancer histologies.

The identified genes are promising with regard to potential mechanistic relevance. *RUNX3* (runt-related transcription factor 3), down-regulated in our analyses and with an AUC of 0.69, is involved in the negative regulation of epithelial cell proliferation, functions as a tumor suppressor, and is frequently deleted or transcriptionally silenced in cancer. Hypermethylation of *RUNX3* has also been associated with the evolution of lung cancer (29) and specifically of lung adenocarcinoma (30). In addition, higher protein expression of *RUNX3* has been associated with increased survival from lung adenocarcinoma (31). *TGFBR3* (transforming growth factor beta receptor III) encodes a glycoprotein that binds TGF β , a cytokine that modulates several tissue development and repair processes. *TGFBR3* is the TGF- β component most commonly down-regulated at both the message and protein levels in several cancers (32–36), including non-small cell lung cancer (37). Our study is the first to show down-regulation of *TGFBR3* mRNA expression in both blood and tumor tissue cells of lung adenocarcinoma patients. *TGFBR3* showed the highest accuracy among the single gene models in discriminating cases from controls (AUC = 0.73). *TRGC2* (T cell receptor *gamma* constant 2), *TRGV9* (T cell receptor *gamma* variable 9), and *TARP* (T cell receptor *gamma* alternate reading frame protein) are colocalized at chromosome locus 7p14.1, close to the 7p14.3 chromosomal region that frequently shows allelic loss in non-small cell lung cancer (38). *TARP* is embedded within the TCR *gamma* locus and cDNA that detect TCR *gamma* mRNA also detect *TARP* mRNA. Accordingly, probes in *TRGC2*, *TRGV9*, and *TARP* showed very similar results in our study. *TRGV9* cells have been shown to contribute to the natural immune surveillance against colon cancers (39). *TARP* has been previously studied as a prostate-specific gene and an androgen-regulated protein that may carry out its biological functions via action on mitochondria (40). Down-regulation in cases with respect to controls and in tumor compared to non-involved tissues of *TRGC2*, *TRGV9*, and *TARP* points to an immune-related alteration as a possible contribution to lung adenocarcinoma development. Case-control discrimination based on *TRGC2*, *TRGV9*, and *TARP* was also good (average AUC = 0.70). *ACPI* (acid phosphatase 1) gene, up-regulated in our analysis, is polymorphic and encodes at least two electrophoretically different isozymes. An increase of fast isozyme concentration increases cancer cells' invasiveness, whereas a decrease of slow isozyme concentration in cancer results in cancer cell proliferation (41). In the validation set *ACPI* showed the poorest accuracy in discriminating cases and controls (AUC = 0.55). *VCAN* (versican) encodes a protein involved in cell adhesion, proliferation, migration, angiogenesis, tissue morphogenesis and maintenance. *VCAN* was initially identified in cultures of lung fibroblasts (42) and has been recognized to play a role in the invasion of several cancers (43) including lung cancer (44). *VCAN* mRNA expression was up-regulated in both lung tumor tissue and PWB of adenocarcinoma cases in our study. *TSTA3* (tissue specific transplantation antigen P35B) gene, also up-regulated in our analysis, is involved in the expression of many glycoconjugates. Intriguingly, *TSTA3* is located in chromosomal region 8q24, which contains several polymorphic variants recently associated with several cancers (45–47). *VCAN* and *TSTA3* also showed a reasonable performance in discriminating between cases and controls (AUC = 0.61 and 0.59, respectively). In addition to the described eight genes, we also identified 42 additional genes whose expression in PWB distinguishes stage I lung adenocarcinoma from controls (Figure 1 and Supplemental Material 2) and was stronger among subjects who currently smoked. If further confirmed in additional blood-based analyses, these genes could also contribute to the detection of early lung adenocarcinoma lesions.

In conclusion, gene expression changes from peripheral blood samples can differentiate early stage lung adenocarcinoma cases from controls and resemble gene expression changes in early stage lung adenocarcinoma tissue. This finding suggests that early processes of lung

adenocarcinoma development may lead to systemic alterations that can be detected in peripheral blood tests. Gene expression from PWB can provide an important tool for the identification of early detection markers of cancer in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support: This research was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, MD.

We would like to thank the EAGLE participants and study collaborators listed on the EAGLE website (<http://eagle.cancer.gov/>).

References

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, et al. Cancer statistics, 2008. *CA Cancer J Clin*. 2008; 58:71–96. [PubMed: 18287387]
2. Charles PC, Alder BD, Hilliard EG, Schisler JC, Lineberger RE, Parker JS, et al. Tobacco use induces anti-apoptotic, proliferative patterns of gene expression in circulating leukocytes of Caucasian males. *BMC Med Genomics*. 2008; 1:38. [PubMed: 18710571]
3. Forrest MS, Lan Q, Hubbard AE, Zhang L, Vermeulen R, Zhao X, et al. Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers. *Environ Health Perspect*. 2005; 113:801–807. [PubMed: 15929907]
4. Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Ortmann WA, Espe KJ, et al. Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc Natl Acad Sci U S A*. 2003; 100:2610–2615. [PubMed: 12604793]
5. Bompreszi R, Ringner M, Kim S, Bittner ML, Khan J, Chen Y, et al. Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. *Hum Mol Genet*. 2003; 12:2191–2199. [PubMed: 12915464]
6. Bovin LF, Rieneck K, Workman C, Nielsen H, Sorensen SF, Skjodt H, et al. Blood cell gene expression profiling in rheumatoid arthritis. Discriminative genes and effect of rheumatoid factor. *Immunol Lett*. 2004; 93:217–226. [PubMed: 15158620]
7. Oudijk EJ, Nijhuis EH, Zwank MD, van de Graaf EA, Mager HJ, Coffey PJ, et al. Systemic inflammation in COPD visualised by gene profiling in peripheral blood neutrophils. *Thorax*. 2005; 60:538–544. [PubMed: 15994259]
8. Burczynski ME, Twine NC, Dukart G, Marshall B, Hidalgo M, Stadler WM, et al. Transcriptional profiles in peripheral blood mononuclear cells prognostic of clinical outcomes in patients with advanced renal cell carcinoma. *Clin Cancer Res*. 2005; 11:1181–1189. [PubMed: 15709187]
9. Moos PJ, Raetz EA, Carlson MA, Szabo A, Smith FE, Willman C, et al. Identification of gene expression profiles that segregate patients with childhood leukemia. *Clin Cancer Res*. 2002; 8:3118–3130. [PubMed: 12374679]
10. Showe MK, Vachani A, Kossenkova EV, Yousef M, Nichols C, Nikonova EV, et al. Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease. *Cancer Res*. 2009; 69:9202–9210. [PubMed: 19951989]
11. Xu T, Shu CT, Purdom E, Dang D, Ilsley D, Guo Y, et al. Microarray analysis reveals differences in gene expression of circulating CD8(+) T cells in melanoma patients and healthy donors. *Cancer Res*. 2004; 64:3661–3667. [PubMed: 15150126]
12. Youlten DR, Cramb SM, Baade PD. The International Epidemiology of Lung Cancer: geographical distribution and secular trends. *J Thorac Oncol*. 2008; 3:819–831. [PubMed: 18670299]

13. Guo NL, Tosun K, Horn K. Impact and interactions between smoking and traditional prognostic factors in lung cancer progression. *Lung Cancer*. 2009; 66:386–392. [PubMed: 19304339]
14. Su LJ, Chang CW, Wu YC, Chen KC, Lin CJ, Liang SC, et al. Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics*. 2007; 8:140. [PubMed: 17540040]
15. Bagnardi V, Randi G, Lubin J, Consonni D, Lam TK, Subar AF, et al. Alcohol consumption and lung cancer risk in the Environment and Genetics in Lung Cancer Etiology (EAGLE) study. *Am J Epidemiol*. 2010; 171:36–44. [PubMed: 19933698]
16. Consonni D, De Matteis S, Lubin JH, Wacholder S, Tucker M, Pesatori AC, et al. Lung cancer and occupation in a population-based case-control study. *Am J Epidemiol*. 2010; 171:323–333. [PubMed: 20047975]
17. Gao Y, Goldstein AM, Consonni D, Pesatori AC, Wacholder S, Tucker MA, et al. Family history of cancer and nonmalignant lung diseases as risk factors for lung cancer. *Int J Cancer*. 2009; 125:146–152. [PubMed: 19350630]
18. Koshiol J, Rotunno M, Consonni D, Pesatori AC, De MS, Goldstein AM, et al. Chronic obstructive pulmonary disease and altered risk of lung cancer in a population-based case-control study. *PLoS One*. 2009; 4:e7380. [PubMed: 19812684]
19. Lam TK, Rotunno M, Lubin JH, Wacholder S, Consonni D, Pesatori AC, et al. Dietary quercetin, quercetin-gene interaction, metabolic gene expression in lung tissue and lung cancer risk. *Carcinogenesis*. 2010; 31:634–642. [PubMed: 20044584]
20. Landi MT, Consonni D, Rotunno M, Bergen AW, Goldstein AM, Lubin JH, et al. Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. *BMC Public Health*. 2008; 8:203. [PubMed: 18538025]
21. Rotunno M, Yu K, Lubin JH, Consonni D, Pesatori AC, Goldstein AM, et al. Phase I metabolic genes and risk of lung cancer: multiple polymorphisms and mRNA expression. *PLoS One*. 2009; 4:e5652. [PubMed: 19479063]
22. Muller MC, Merx K, Weisser A, Kreil S, Lahaye T, Hehlmann R, et al. Improvement of molecular monitoring of residual disease in leukemias by bedside RNA stabilization. *Leukemia*. 2002; 16:2395–2399. [PubMed: 12454744]
23. Rainen L, Oelmueller U, Jurgensen S, Wyrich R, Ballas C, Schram J, et al. Stabilization of mRNA expression in whole blood samples. *Clin Chem*. 2002; 48:1883–1890. [PubMed: 12406972]
24. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One*. 2008; 3:e1651. [PubMed: 18297132]
25. Hu N, Qian L, Hu Y, Shou J, Wang C, Giffen C, et al. Quantitative real-time RT-PCR validation of differential mRNA expression of SPARC, FADD, Fascin, COL7A1, CK4, TGM3, ECM1, PPL and EVPL in esophageal squamous cell carcinoma. *BMC Cancer*. 2006; 6:33.
26. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B*. 2005; 57:289–300.
27. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2⁻(Delta Delta C(T)) Method. *Methods*. 2001 Dec; 25(4):402–408. [PubMed: 11846609]
28. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst*. 2001; 93:1054–1061. [PubMed: 11459866]
29. Li QL, Kim HR, Kim WJ, Choi JK, Lee YH, Kim HM, et al. Transcriptional silencing of the RUNX3 gene by CpG hypermethylation is associated with lung cancer. *Biochem Biophys Res Commun*. 2004; 314:223–228. [PubMed: 14715269]
30. Sato K, Tomizawa Y, Iijima H, Saito R, Ishizuka T, Nakajima T, et al. Epigenetic inactivation of the RUNX3 gene in lung cancer. *Oncol Rep*. 2006; 15:129–135. [PubMed: 16328045]
31. Araki K, Osaki M, Nagahama Y, Hiramatsu T, Nakamura H, Ohgi S, et al. Expression of RUNX3 protein in human lung adenocarcinoma: implications for tumor progression and prognosis. *Cancer Sci*. 2005; 96:227–231. [PubMed: 15819721]

32. Copland JA, Davies PJ, Shipley GL, Wood CG, Luxon BA, Urban RJ. The use of DNA microarrays to assess clinical samples: the transition from bedside to bench to bedside. *Recent Prog Horm Res.* 2003; 58:25–53. [PubMed: 12795413]
33. Dong M, How T, Kirkbride KC, Gordon KJ, Lee JD, Hempel N, et al. The type III TGF-beta receptor suppresses breast cancer progression. *J Clin Invest.* 2007; 117:206–217. [PubMed: 17160136]
34. Gordon KJ, Dong M, Chislock EM, Fields TA, Blobe GC. Loss of type III transforming growth factor beta receptor expression increases motility and invasiveness associated with epithelial to mesenchymal transition during pancreatic cancer progression. *Carcinogenesis.* 2008; 29:252–262. [PubMed: 17999987]
35. Hempel N, How T, Dong M, Murphy SK, Fields TA, Blobe GC. Loss of betaglycan expression in ovarian cancer: role in motility and invasion. *Cancer Res.* 2007; 67:5231–5238. [PubMed: 17522389]
36. Sharifi N, Hurt EM, Kawasaki BT, Farrar WL. TGFBR3 loss and consequences in prostate cancer. *Prostate.* 2007; 67:301–311. [PubMed: 17192875]
37. Finger EC, Turley RS, Dong M, How T, Fields TA, Blobe GC. TbetaRIII suppresses non-small cell lung cancer invasiveness and tumorigenicity. *Carcinogenesis.* 2008; 29:528–535. [PubMed: 18174241]
38. Tseng RC, Chang JW, Hsien FJ, Chang YH, Hsiao CF, Chen JT, et al. Genomewide loss of heterozygosity and its clinical associations in non small cell lung cancer. *Int J Cancer.* 2005; 117:241–247. [PubMed: 15900585]
39. Corvaisier M, Moreau-Aubry A, Diez E, Bennouna J, Mosnier JF, Scotet E, et al. V gamma 9V delta 2 T cell response to colon carcinoma cells. *J Immunol.* 2005; 175:5481–5488. [PubMed: 16210656]
40. Maeda H, Nagata S, Wolfgang CD, Bratthauer GL, Bera TK, Pastan I. The T cell receptor gamma chain alternate reading frame protein TARP., a prostate-specific protein localized in mitochondria. *J Biol Chem.* 2004; 279:24561–24568. [PubMed: 15150260]
41. Alho I, Clara BM, Carvalho R, da Silva AP, Costa L, Bicho M. Low molecular weight protein tyrosine phosphatase genetic polymorphism and susceptibility to cancer development. *Cancer Genet Cytogenet.* 2008; 181:20–24. [PubMed: 18262048]
42. Zimmermann DR, Ruoslahti E. Multiple domains of the large fibroblast proteoglycan, versican. *EMBO J.* 1989; 8:2975–2981. [PubMed: 2583089]
43. Ricciardelli C, Sakko AJ, Ween MP, Russell DL, Horsfall DJ. The biological role and regulation of versican levels in cancer. *Cancer Metastasis Rev.* 2009; 28:233–245. [PubMed: 19160015]
44. Pirinen R, Leinonen T, Bohm J, Johansson R, Ropponen K, Kumpulainen E, et al. Versican in nonsmall cell lung cancer: relation to hyaluronan, clinicopathologic factors, and prognosis. *Hum Pathol.* 2005; 36:44–50. [PubMed: 15712181]
45. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc Natl Acad Sci U S A.* 2006; 103:14068–14073. [PubMed: 16945910]
46. Schumacher FR, Feigelson HS, Cox DG, Haiman CA, Albanes D, Buring J, et al. A common 8q24 variant in prostate and breast cancer from a large nested case-control study. *Cancer Res.* 2007; 67:2951–2956. [PubMed: 17409400]
47. Yeager M, Chatterjee N, Ciampa J, Jacobs KB, Gonzalez-Bosquet J, Hayes RB, et al. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat Genet.* 2009; 41:1055–1057. [PubMed: 19767755]

Figure 1. Heat Map of Genes Differentially Expressed in Stage I Lung Adenocarcinoma Patients vs. Controls

Gene symbols and Affymetrix gene-probe names are reported on the vertical axis and sample color codes are shown on the horizontal axis. Samples and gene-probes are ordered according to unconditional clustering based on linkage method and Spearman correlation distance. Samples are color-coded as following: stage I cases (magenta), controls (cyan), current smokers (blue), former smokers (orange), and never smokers (yellow). The first 36 gene-probes from the top are up-regulated in stage I cases compared to controls. The last 25 gene-probes are down-regulated in stage I cases compared to controls. Subjects with different smoking status are homogeneously distributed along the horizontal axis, which

indicates that smoking status does not contribute to the differential expression between cases and controls in these genes. Detailed information on these gene-probes and related results are described in Supplemental Material 2.

Figure 2. Gene expression changes differentiating tumor status in both blood and tissue-based analyses

Boxplots of gene expression values (in log scale base 2) by case/control status and by Tumor/Non-involved (T/NI) tissue type for 10 gene-probes from Figure 1 corresponding to 8 genes whose expression changes by tumor status in PWB mirrored changes in the target organ. This analysis included 30 paired samples from 15 stage I adenocarcinoma cases. Data shown in figure included all smoking categories. Similar results were obtained for analyses restricted to current smokers.

Figure 3. Comparison of gene expression changes in stage I cases vs. in cases with advanced stages

Plot of fold changes (FC) by case-control status for the gene-probes that differed significantly (i.e., $FDR \leq 0.1$ based on $P\text{-value} \leq 0.001$) and had large fold change (i.e., $FC \leq 0.66$ or ≥ 1.5) in the analysis comparing stage I cases *versus* controls, among current smokers or among all smoking categories. Probes are listed by chromosomal location. FCs for down-regulated probes are shown in the lower side of the graph ($FC \leq 1$, open symbols) and corresponding gene symbols in the bottom x-axis. FCs for up-regulated probes are shown on the upper side of the graph ($FC > 1$, filled symbols) and corresponding gene symbols in the top x-axis. FCs were obtained comparing cases with stage I (black triangles), cases with stage II, III, and IV (dark grey squares), and cases with all stages (light grey circles) to controls. For consistency, FCs of gene-probes that were significant in the analysis comparing stage I cases to controls among current smokers were computed based on current smokers also in the analyses including later stages. Analogously, FCs of gene-probes that were significant in the analysis comparing stage I cases to controls among all smoking categories (i.e., *GNLY*, *PBX1*, *CYP1B1*, *SIAH2*, *PF4V1*, *VCAN*, *HIST1H3H*, *CD36*, *CALD1*, *TSTA3*, *ABCA1*, and *TPM1*) were computed based on all smoking categories also in the analyses including advanced stages. Gene-probes showed generally higher up-regulated FCs and lower down-regulated FCs in the analysis based on stage I cases (triangles) with respect to the analysis based on advanced stage cases (squares), although with concordant direction. As expected, FCs in the analyses based on all stage cases (circles) were in between those based on stage I (triangles) and those based on advanced stage (squares) cases.

Figure 4. Receiver operating characteristic (ROC) curves based on qRT-PCR results from the PWB validation set for the eight genes whose expression differentiated tumor status in both blood and tissue-based analyses

ROC curves show sensitivity *versus* specificity in discriminating between stage I lung adenocarcinoma patients and controls at different gene expression thresholds for the following models: individual *RUNX3*, *TGFBR3*, *TRGC2/TARP*, *TRGV9/TARP* (i.e., one ABI probe covered *TRGC2* and *TARP* and another ABI probe covered *TRGV9* and *TARP*), *ACP1*, *VCAN*, and *TSTA3* and a multiplex model combining all genes. The area under the curve (AUC) values and corresponding 95% confidence interval (in parentheses) are reported for each predictive model.

Characteristics of Lung Adenocarcinoma Patients and Controls used for the Gene Expression Analysis from Peripheral Whole Blood, Italy, 2002–2005.

Table 1

	Discovery Set		Validation Set	
	Controls	Cases	P-val*	Cases Controls P-val*
Sex			0.93	0.48
Male	40	37		111 67
Female	40	36		19 15
Age			0.73	0.46
40–44	2	2		2 1
45–49	3	2		2 2
50–54	7	7		5 5
55–59	11	9		16 14
60–64	13	13		27 15
65–69	13	16		31 17
70–74	21	19		33 18
75–79	10	5		14 10
Cigarette Smoking			0.71	
Never	21	22		0 0
Former	32	27		0 0
Current	27	24		130 82
Cumulative Pack Years			0.44	0.76
Below Median	31	23		63 40
Above Median	28	28		67 39
Missing	0	0		0 3
Mean (standard deviation)	33.7 (21.9)	34.7 (20.5)		51.2 (21.9) 50.1 (24.4)
Cigarette Per Day			0.93	0.56
Below Median	29	25		48 26
Above Median	30	25		82 53
Missing	0	1		0 3
Mean (standard deviation)	17.9 (10.1)	17.5 (9.1)		21.1 (7.8) 21.4 (9.3)

	Discovery Set		Validation Set	
	Controls	Cases	P-val*	Cases P-val*
Tumor Stage[†]				
IA		11		36
IB		15		46
IIA		3		0
IIB		15		0
IIIA		9		0
IIIB		8		0
IV		12		0
Tumor Grade				
1		6		8
2		28		28
3		33		38
missing		6		6

* P-value based on a two-sided Wald test.

[†] Stage definition was based on the International System for Staging Lung Cancer by the AJCC and UICC, Mountain CF 1997.

Summary of Gene Expression Results from Peripheral Whole Blood Comparing Lung Adenocarcinoma Cases and Controls, Italy, 2002–2005.

Table 2

	Controls	Stage I cases		All Stage cases	
All Subjects	n=80	n=26	n=73		
FDR ^{//}	Reference	0.10	0.03		
	Down	Up	Down	Up	
Gene probes significant by P-value [†]	82	139	435	353	
Gene probes significant by P-value and FC *	4	22	0	21	
Current Smoker	n=27	n=12	n=24		
FDR ^{//}	Reference	0.15	0.11		
	Down	Up	Down	Up	
Gene probes significant by P-value	100	44	117	78	
Gene probes significant by P-value and FC	23	23	5	45	
Former Smoker	n=32	n=5	n=27		
FDR ^{//}	Reference	0.97	1.00		
	Down	Up	Down	Up	
Gene probes significant by P-value	9	5	3	14	
Gene probes significant by P-value and FC	0	2	0	3	
Never Smoker	n=21	n=9	n=22		
FDR ^{//}	Reference	1.00	0.32		
	Down	Up	Down	Up	
Gene probes significant by P-value	1	11	21	43	
Gene probes significant by P-value and FC	0	2	3	20	

^{//} Maximum False Discovery Rate (FDR), shown as bolded if ≤0.1.

[†] P-value ≤0.001.

* FC ≤0.66 or ≥1.5.