Molecular Biomarkers for Quantitative and Discrete COPD Phenotypes

Soumyaroop Bhattacharya[1*‡], Sorachai Srisuma[1,4*], Dawn L. DeMeo[2], Steven D. Shapiro[1§], Raphael Bueno[3], Edwin K. Silverman[2], John J. Reilly[1§], and Thomas J. Mariani[1‡]

[1]Department of Pulmonary Medicine, [2]The Channing Laboratory, and [3]Thoracic Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; and [4]Department of Physiology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

Chronic obstructive pulmonary disease (COPD) is an inflammatory lung disorder with complex pathological features and largely unknown etiology. The identification of biomarkers for this disease could aid the development of methods to facilitate earlier diagnosis, the classification of disease subtypes, and provide a means to define therapeutic response. To identify gene expression biomarkers, we completed expression profiling of RNA derived from the lung tissue of 56 subjects with varying degrees of airflow obstruction using the Affymetrix U133 Plus 2.0 array. We applied multiple, independent analytical methods to define biomarkers for either discrete or quantitative disease phenotypes. Analysis of differential expression between cases ($n = 15$) and controls ($n = 18$) identified a set of 65 discrete biomarkers. Correlation of gene expression with quantitative measures of airflow obstruction (FEV_1%predicted or FEV_1/FVC) identified a set of 220 biomarkers. Biomarker genes were enriched in functions related to DNA binding and regulation of transcription. We used this group of biomarkers to predict disease in an unrelated data set, generated from patients with severe emphysema, with 97% accuracy. Our data contribute to the understanding of gene expression changes occurring in the lung tissue of patients with obstructive lung disease and provide additional insight into potential mechanisms involved in the disease process. Furthermore, we present the first gene expression biomarker for COPD validated in an independent data set.

Keywords: microarray; gene expression; emphysema; lung function

Chronic obstructive pulmonary disease (COPD), an inflammatory disorder that is characterized by a gradual loss of lung function, is currently the fourth leading cause of death in the United States (1). Strongly associated with cigarette smoking, COPD is expected to be the third most common cause of death and fifth most common cause of disability worldwide by 2020 (2). COPD is typically diagnosed late in the course of disease when the patient presents with significant physiologic impairment (3, 4). The need for improved early diagnosis and the identification of novel therapeutic targets, which may improve treatment options and reduce mortality, has recently gained heightened interest (5).

(*Received in original form March 24, 2008 and in final form August 8, 2008*)

* These authors contributed equally to this work.

‡ Present affiliation: Division of Neonatology and Center for Pediatric Biomedical Research, University of Rochester, Rochester, New York.

§ Present affiliation: Department of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania.

This work was supported by National Heart, Lung, and Blood Institute contract grant numbers HL72303 (to J.J.R.) and K08 HL072918(to D.L.D.).

Correspondence and requests for reprints should be addressed to Thomas J. Mariani, Ph.D., Division of Neonatology and Center for Pediatric Biomedical Research, Department of Pediatrics, University of Rochester, Box 703, 601 Elmwood Avenue, Rochester, NY 14642. E-mail: Tom_Mariani@URMC.Rochester.edu

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

Am J Respir Cell Mol Biol Vol 40. pp 359–367, 2009
Originally Published in Press as DOI: 10.1165/rcmb.2008-0114OC on October 10, 2008
Internet address: www.atsjournals.org

CLINICAL RELEVANCE

Chronic obstructive pulmonary disease includes a broad spectrum of histopathologic findings and respiratory symptoms, with a complex disease pathogenesis. We used genome-wide expression profiling to identify biomarkers for disease-related phenotypes.

COPD includes a broad spectrum of histopathologic findings and respiratory symptoms best characterized as a syndrome. Chronic obstructive bronchitis/bronchiolitis with peribronchiolar fibrosis (small airways disease), and abnormal enlargement of airspace distal to the terminal bronchioles with destruction of lung parenchyma (emphysema), are the pathological hallmarks of disease. Small airways disease and emphysema can present alone or in combination, with varying degrees of severity (6, 7). Genetic and environmental factors contribute to variable susceptibility in the general population. α_1-Antitrypsin deficiency is a proven genetic risk factor, which modifies disease susceptibility in response to environmental factors, most notably tobacco smoke exposure (8, 9). Given the complexity of disease pathogenesis, the presence of varying levels of susceptibility in the general population and the fact that patients rarely present early in disease pathogenesis, at a time when disease-modifying therapy may be more effective, the identification of biological markers of disease susceptibility and/or progression is needed.

Numerous previous studies have sought to identify disease biomarkers, such as genetic or expression variants. DNA microarrays have been proven to be a powerful tool capable of biomarker discovery for various disease states, including COPD (10–14). Spira and coworkers measured gene expression from lung tissue of 35 patients undergoing surgery for severe emphysema, and identified gene expression patterns associated with disease severity and surgical outcome (13). Golpon and colleagues used a similar approach to characterize global gene expression patterns in patients with and without α_1-antitrypsin deficiency (10). Both Zhang and coworkers and Ning and colleagues identified EGR1 overexpression in patients with COPD (12, 14). We have previously reported a multidisciplinary approach, using gene expression microarrays and genetic association studies, to identify serine protease inhibitor, clade E, member 2 (SERPINE2) as a novel COPD candidate susceptibility gene (15).

Here, we present a novel gene expression microarray data set generated from 56 subjects with mild to severe COPD as defined by airflow obstruction. We report the identification of gene expression biomarkers for both discrete (case versus control) and quantitative COPD-related phenotypes. Further, we identify a subset of these biomarkers that can reliably predict disease in an independent data set, derived from a distinct population with COPD. This gene set represents a robust gene expression biomarker for COPD. The identification of these differentially expressed genes may assist with the future de-

velopment of methods (such as genetic tests) for improved early diagnosis or the identification of therapeutic response.

MATERIALS AND METHODS

Subjects

Lung tissue was obtained, according to an approved IRB protocol, from 62 subjects undergoing surgical resection of a solitary nodule suspected to be cancer. Tissue samples were acquired and processed as previously described (16). Briefly, frozen samples of resected "normal" (grossly uninvolved) lung were obtained within 30 minutes of resection and subdivided into samples ($\sim$ 100 mg). Samples intended for nucleic acid extraction were snap frozen on powdered dry ice and individually stored in liquid nitrogen. Diagnosis was confirmed by surgical pathology. Written informed consent was provided and subjects underwent lung function testing by spirometry (without bronchodilator) and completed a lung health–related questionnaire before surgery. Age, height, weight, sex, and surgical pathology were obtained from subjects' medical charts. Predicted lung function values (FEV_1%predicted, FVC) were calculated in SAS v9.1 from SAS Institute (Cary, NC) using the Crapo equation for white subjects (17) and the Hankinson equation for African-American subjects (18).

RNA Isolation

Tissue used for RNA isolation was obtained from a histologically normal area of the lung distant from the tumor. Lung tissue was dissected and adjacent pieces were either fixed for histology or snap-frozen in liquid nitrogen and stored at less than −70°C (16). Frozen tissue was immediately pulverized upon removal from liquid nitrogen and transferred into Trizol reagent (Invitrogen, Carlsbad, CA). Samples were homogenized in Trizol using a rotor-stator homogenizer, and RNA was isolated following manufacturer's protocols. RNA was further purified using the solid-phase column method (RNeasy kit; Qiagen, Valencia, CA). A total of 64 RNA samples were generated from lung tissue obtained from 62 subjects.

Microarray Analysis

Labeled target was synthesized from purified RNA samples according to manufacturer's recommendations as previously described (19). These studies used the Affymetrix HG-U133Plus 2.0 array (Affymetrix, Santa Clara, CA), containing 54,675 probe sets interrogating over 47,000 human transcripts. Target hybridization, washing, and array scanning were performed according to standard protocols. Two independent versions of expression intensities were extracted from raw data files using either Robust Multichip Average (RMA [20]) or Affymetrix Microarray Suite (MAS) 5.0 algorithms. MAS 5.0 yields scaled, background-subtracted, nonnormalized signal intensities, while RMA provides background-subtracted, log-transformed signal intensities. Data extraction was performed using *affy* library in BioConductor, an R-based package. The annotation information of the selected probe sets was retrieved from the Affymetrix analysis portal (NetAffx, www.netaffx.com). Reliability of signal intensity measurement was determined using the Detection Call extracted using *affy* library in BioConductor, an R-based package. Unsupervised clustering with the nonparametric bootstrap (21) was applied to check for undesirable and unanticipated structure or associations among the samples. Of 64 RNA samples arrayed, 56 passed quality control criteria, including those recommended by the Best Practices Working Group (22), and were subjected to further analysis. The data described in this manuscript have been made available at Gene Expression Omnibus (accession number GSE8581). All samples have been annotated as per the requirements of MIAME/MAGE standards.

Microarray Data Analysis

For discrete phenotype analysis, cases were defined as subjects with $FEV_1 < 70\%$ predicted and $FEV_1/FVC < 0.7$, and controls as subjects with $FEV_1 > 80\%$ predicted and $FEV_1/FVC > 0.7$. We applied two independent tests for differential expression on both RMA and MAS5 versions of the data set: Bayesian analysis of differential gene expression (BADGE, http://genomethods.org/badge [23, 24]) at $P < 0.01$ and Significance Analysis of Microarrays (SAM [25]) at a False Discovery Rate (FDR) of 0 using MultiExperiment Viewer (MeV) 3.0 from TIGR (http://www.tm4.org/mev.html). Here FDR = 0 means that the median number of false positives calculated during the procedure is equal to 0, not that there are no false positives. For probe sets to be defined as differentially expressed, they were required to show significant changes in expression between groups in all four statistical comparisons (RMA-BADGE, MAS5-BADGE, RMA-SAM, MAS5-SAM). For quantitative phenotype analysis, correlation coefficients of signal intensity and lung function (FEV_1%predicted or FEV_1/FVC) were calculated. For each probe set, we calculated both the Pearson linear and Spearman rank correlation coefficients for both RMA and MAS5-derived expression intensities using SAS. For probe sets to be defined as significantly correlated, they were required to show significant correlation (at a P value threshold of ≤ 0.05; *see* RESULTS) between gene expression and lung function in all four statistical comparisons (RMA-Pearson, RMA-Spearman, MAS5-Pearson, MAS5-Spearman). Visualization of data for significantly regulated genes was generated using RMA-derived signal intensity data and plotted in MeV.

Class Prediction

Of the 254 probe sets that were either differentially expressed or significantly correlated with lung function parameters in the current data set using the Affymetrix U133 plus 2.0 array, 84 were also present on the Affymetrix U133A 2.0 (which is a subset of U133 plus 2.0 with a total of 22,275 probe sets) used by Spira and coworkers (13). Our gene expression biomarkers were used to distinguish emphysema cases from controls in the cohort of Spira and colleagues by average linkage hierarchical clustering with Euclidean distance. These analyses were performed using the Gene Expression Data Analyzer (http://bioinformatics2.pitt.edu/GE2/GEDA.html) (26). Cases were subjects who met the clinical criteria and underwent lung volume reduction surgery. Controls were former smokers with an FEV_1 greater than 45% predicted or $D_{L_{CO}}$ greater than 50% predicted, for which emphysema was ruled out by high-resolution computed tomography.

Expression Validation

We performed quantitative reverse transcriptase-polymerase chain reaction (qPCR) for a subset of the genes identified as discrete and/or quantitative disease markers. qPCR was performed on a Stratagene MX3000P (Stratagene, La Jolla, CA) using Taqman chemistry, essentially as previously described (27). Pre-developed, gene-specific assays for measuring gene expression were purchased from Applied Biosystems (Foster City, CA). Gene expression levels were calculated according to the relative expression analysis approach using GAPDH and/or PPIA (peptidyl prolyl isomerase A or cyclophilinA) as an internal, endogenous control. Primary validation was defined as a significant ($P < 0.05$) concordance in expression patterns between array data and qPCR as defined by correlation coefficient. For this analysis, we measured the linear (Pearson) and rank (Spearman) correlation between the dCt (cycle threshold of biomarker gene – endogenous control gene) of the biomarker and the RMA-derived relative signal intensity values, using GAPDH as the endogenous control. We also repeated the analysis using PPIA as an endogenous control. Differential expression analysis was performed on individual sample values of dCt for each gene using either the parametric Student's t test or nonparametric Mann-Whitney U-test. For differential expression analysis, we used either GAPDH or PPIA as en endogenous control.

Gene Ontology

Functional classification of genes was performed using Expression Analysis Systematic Explorer (EASE) v2.0 (http://david.abcc.ncifcrf.gov) (28). Entrez GeneIDs for the selected biomarker genes were used as the input list, while Entrez GeneIDs for all filtered probe set genes (16,452 always detected in either cases or controls) served as the background set. Gene Ontology categories with an EASE score of less than 0.05 were defined as significantly overrepresented.

RESULTS

Subject Demographics

We assessed genome-wide expression patterns in lung tissue specimens derived from 56 subjects. These subjects were undergoing

lobectomy for removal of a suspected lung tumor, and tissue for our studies was derived from histologically normal tissue distant from the tumor margin. Low values for both FEV_1%predicted and FEV_1/FVC are characteristic features of COPD and associated with disease severity. For our studies, we defined cases ($n = 15$) as subjects with $FEV_1 < 70\%$ predicted and $FEV_1/FVC < 0.7$ and controls ($n = 18$) as subjects with $FEV_1 > 80\%$ predicted and $FEV_1/FVC > 0.7$. The distribution of lung function in cases and controls is listed in Table 1. Individual subject characteristics are listed in Table E1 in the online supplement. Twenty-three subjects were not classified as cases or controls, and data derived from these subjects were used solely for quantitative analysis. A majority of the subjects were diagnosed with adenocarcinoma ($n = 26$) or squamous cell carcinoma ($n = 19$), while other tumor types or benign lesions were found in the remaining subjects ($n = 11$). There was a similar frequency of tumor incidence in cases or controls (80% versus 67%, respectively), although squamous cell carcinoma was more frequently observed in cases (53%), while adenocarcinoma was more frequent in controls (44%).

Identification of Gene Expression Markers for COPD Susceptibility

We used a highly stringent set of criteria to define differential expression in this data set, relying upon multiple data extraction and significance testing methods, and focusing on consistency of observations (*see* MATERIALS AND METHODS). We first removed data from all probe sets that were not reliably detectable in either all cases or all control samples. We then extracted signal intensity data using RMA and MAS5. Each data set was tested separately using both BADGE (23, 24) and SAM (25) to identify differential expression. Signal intensity values for a total of 293 probe sets were significantly different in BADGE analysis ($P < 0.01$). SAM analysis was more restrictive, identifying a total of 65 probe sets that were significantly different between COPD cases and controls. All probe sets identified in SAM analysis were also identified using BADGE. The relative expression levels for these 65 probe sets, representing 55 genes, are shown in Figures 1 and E1. Additional information regarding these probe sets is provided in Table E2. While expression of these genes clearly segregates a subset of the samples (note Cases 1–9 versus Controls 3–16), others appear to have intermediate levels of expression. This likely reflects the combined heterogeneity of tissue samples, disease subtypes, and a relatively small sample size. Interestingly, all genes identified using these highly stringent criteria were expressed at a lower level in cases as compared with controls. This is a result of our restrictive approach, as numerous genes were identified as expressed at a significantly higher level in cases than controls using individual tests. However, there is a clear trend toward significant reduction in gene expression in COPD tissue samples. This may reflect a "diseased state" of the tissue, but does not seem to be related to subject age (mean cases = 63 versus controls = 64; Table 1).

We assessed whether differences in the distribution of tumor type between cases and controls contributed to the identification of these gene expression biomarkers. We applied differential expression analysis (as described for cases and controls above) comparing all samples classified as adenocarcinoma and all samples classified as squamous cell carcinoma. No probe sets were identified as consistently differentially expressed between tumor types. Further, no probe sets identified as differentially expressed between tumor types in any single analysis were among the COPD biomarker gene set.

TABLE 1. GROUP DEMOGRAPHICS AND LUNG FUNCTION

	Age, yr	FEV_1%predicted	FVC %predicted	FEV_1/FVC (%)
Case ($n = 15$)	63 (39–82)	43.23 (10–67)	63.17 (18–99)	56.52 (41–63)
Control ($n = 18$)	64 (56–77)	101.39 (83–154)	105 (80–125)	77.17 (73–85)
Unclassified ($n = 23$)	69 (50–84)	71.52 (51–98)	82.33 (66–119)	73.15 (68–89)
All ($n = 56$)	66 (39–84)	72.24 (10–154)	83 (18–125)	69.53 (41–89)

Provided are group means and range. Complete information for individual subjects is provided in Table E1 in the online supplement.

Identification of Gene Expression Markers for COPD Severity

Case-control analysis identified a set of genes that did not clearly segregate disease and control samples, in part due to small sample size. A number of our subjects (40%) were not classified as either cases or controls using the defined criteria. In an effort to identify additional markers capable of disease prediction, we further analyzed our entire data set for gene expression correlation with lung function. The signal intensity for each probe set was correlated to FEV_1%predicted across all subjects ($n = 56$) to identify quantitative gene expression markers. Again, we used highly stringent criteria to confirm correlation of gene expression and lung function, relying upon multiple data extraction and significance testing methods, and focusing upon consistency of observations (*see* MATERIALS AND METHODS). These results are shown in Table 2. A total of 614 probe sets were significantly correlated with FEV_1 %predicted at $P < 0.05$. A subset of 65 probe sets, representing 47 known genes, were significant at $P < 0.01$. The relative expression levels for these 65 probe sets are shown in Figure 2. Correlation coefficient with FEV_1%predicted and P values for these probe sets are listed in Table E3A. There were an equal proportion of genes showing positive and negative correlations with FEV_1%predicted. Only two probe sets (PCDH10, KLF8) were significantly correlated with FEV_1 at $P < 0.001$.

We repeated quantitative gene expression analysis using FEV_1/FVC as a quantitative phenotype. We identified 1,649 probe sets significantly correlated with FEV_1/FVC at $P < 0.05$, 170 probe sets significant at $P < 0.01$ and 9 at $P < 0.001$ (Table 2). Correlation coefficients with FEV_1/FVC and P values are listed in Table E3B. We further considered whether any markers were consistently correlated with both FEV_1%predicted and FEV_1/FVC (Figure 3). At $P < 0.05$, 220 probe sets were significantly correlated with both lung function measures, while only 8 probe sets (representing KLF8, SEMA6D, ZNF30, LCMT1, RMST, PTCH, ZF, RP9) were correlated at $P < 0.01$. There was no overlap among probe sets at $P < 0.001$.

Identification of a Robust Gene Expression Biomarker for COPD

A total of 254 probe sets passed criteria as either discrete ($n = 65$) or quantitative ($n = 220$ at $P < 0.05$) gene expression markers of COPD. Among these, 43 probe sets representing 35 genes were significantly different in case-control analysis and significantly correlated with FEV_1 at $P < 0.05$. Further, 31 probe sets representing 22 known genes were significantly different in case-control analysis and significantly correlated with both FEV_1%predicted and FEV_1/FVC at $P < 0.05$. The relative expression levels for these 22 genes are shown in Figure 4.

These genes are highly informative within our data set. However, a major limitation of gene expression biomarker identification is the failure to replicate across studies or populations. Spira and coworkers reported a gene expression microarray data set from a distinct cohort of patients with severe COPD undergoing lung volume reduction surgery (13). This data set included a similar number of cases and controls (18 and 15, respectively) and was generated using Affymetrix microarrays. Unlike our data set, the cases in the data set of Spira and colleagues were undergoing lung volume reduction

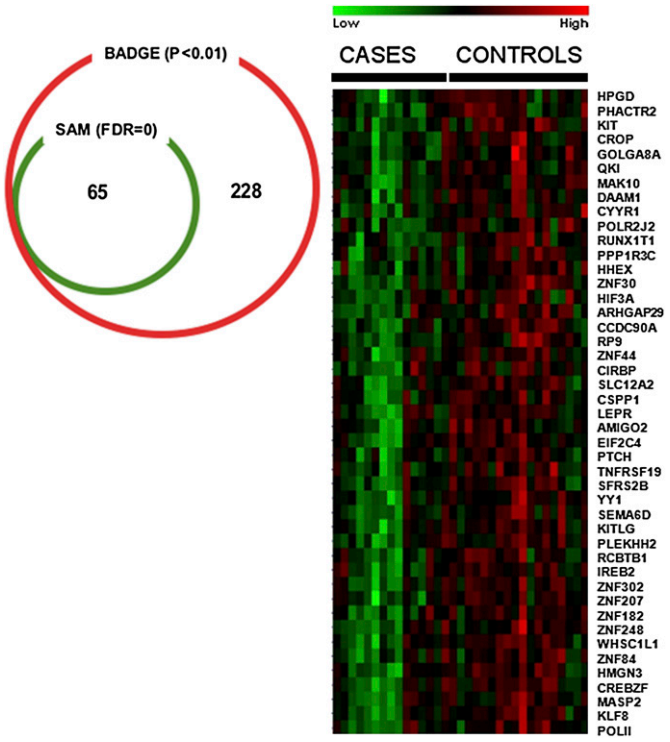


Figure 1. Discrete biomarkers. Shown are signal intensity measurements for each of the 65 probe sets identified as significantly differentially expressed between cases and controls using both Significance Analysis of Microarrays and Bayesian analysis of differential gene expression. Data from individual subjects are in *columns* and data for individual genes are in *rows*. Signal intensity data are color-coded such that the intensity of *red* indicates a relatively high level of expression, while the intensity of *green* represents a relatively low level of expression. Poorly annotated probe sets without an Entrez GeneID have been removed.

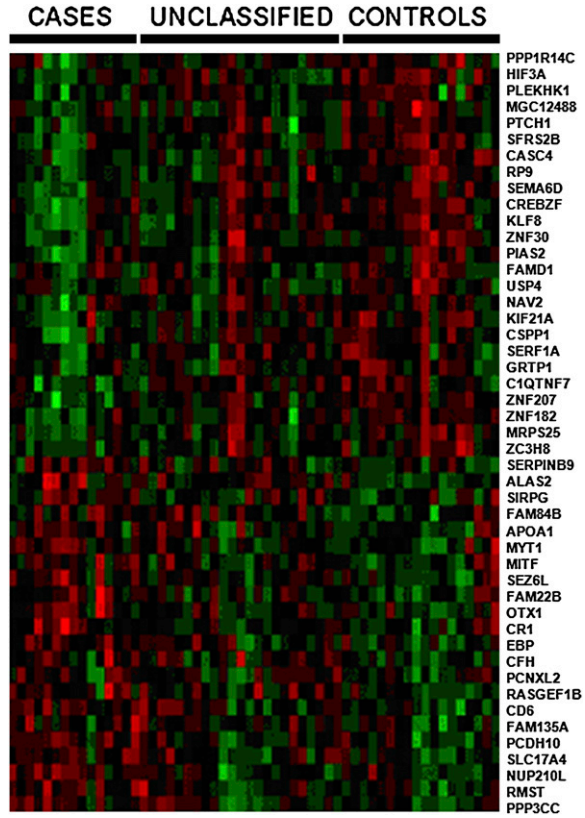


Figure 2. Quantitative biomarkers. Shown are signal intensity measurements for 57 probe sets representing well-annotated genes (among the 65 probe sets) identified as significantly correlated with FEV_1%predicted at $P < 0.01$. Poorly annotated probe sets without an Entrez GeneID have been removed. Data from individual subjects are in *columns* and data for individual genes are in *rows*. Signal intensity data are color-coded such that the intensity of *red* indicates a relatively high level of expression, while the intensity of *green* represents a relatively low level of expression.

surgery for severe emphysema. Furthermore, Spira and coworkers used a previous version of Affymetrix Human Genome arrays (Hu133Av2). We tested our 254 gene expression biomarkers for their ability to identify affection status in the data set of Spira and colleagues using class prediction methods. A total of 84 probe sets were identified on the Hu133Av2 platform corresponding to the 254 gene expression biomarkers we identified in our data set (*see* MATERIALS AND METHODS). We tested the ability of these 84 probe sets to discriminate cases from controls in the data set of Spira and coworkers (Table 3). Using all 84 probe sets, we achieved 97% predictive accuracy with 100% sensitivity and 93% specificity. Slightly reduced predictive accuracies were achieved using subsets derived solely from quantitative (85%) or discrete (88%) biomarkers. Further, we were able to identify a subset of 40 probe sets that achieved 100% accuracy in distinguishing cases from controls. Information for these 84 probe sets is provided in Table E4, and the corresponding dendrograms are shown in Figure E2.

Expression data for these biomarkers from the data set of Spira and colleagues are presented in Tables E5A–E5C. For these data, we extracted the signal intensity values and applied our analytical approach (as described in MATERIALS AND METHODS), specifically focusing upon consistency of results across multiple data extraction algorithms and statistical tests. Results from qualitative (A) and quantitative (B, C) analysis are included.

TABLE 2. DISTRIBUTION OF PROBE SETS CORRELATED WITH LUNG FUNCTION

Threshold	FEV_1% predicted	FEV_1/FVC	Overlap
$P < 0.05$	614	1649	220
$P < 0.01$	65	170	8
$P < 0.001$	2	9	0

Gene Ontology

In an effort to determine if any biological systems or functions were particularly defined by either discrete or quantitative COPD biomarkers, we performed gene ontology assessment using EASE (Figure 5). Strikingly, there was a universal and consistent overrepresentation of functions relating to transcriptional activity and nucleic acid binding for all sets of COPD biomarkers. A total of 24 genes, or 43% of biomarkers tested for ontology (a subset of each list lacked ontological annotation), were classified in one or more categories related to these functions. For discrete marker genes (case-control), 11 of 20 (55%; $P < 0.05$) were classified for Nucleic Acid Binding (GO:0003676) and 8 of 19 (42%; $P < 0.05$) were classified for DNA-dependent Transcription (GO:0006351). For quantitative marker genes, 19 of 47 (40%; $P < 0.05$) were classified for Nucleic Acid Binding (GO:0003676) and 16 of 45 (36%; $P < 0.05$) were classified for DNA-dependent Transcription (GO:0006351). For the 31 markers shared between discrete and quantitative phenotypes, 5 of the 10 genes ($P < 0.05$) with ontological information

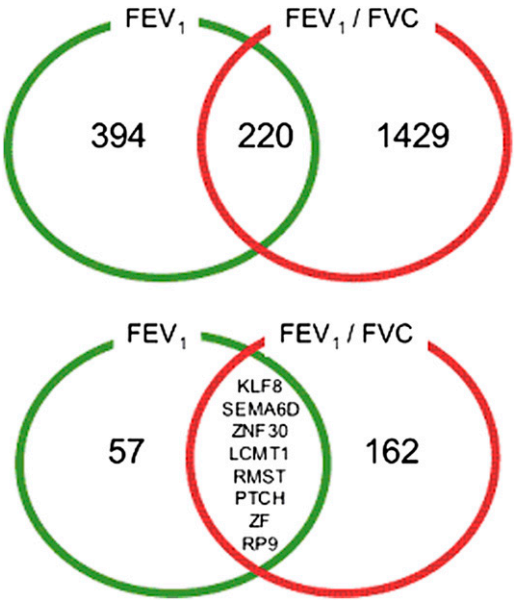


Figure 3. Distribution of quantitative biomarkers. Venn diagram showing relationships for biomarkers of quantitative COPD phenotypes (FEV_1%predicted and FEV_1/FVC) at significance levels of $P < 0.05$ (*top*) or $P < 0.01$ (*bottom*).

available were classified for DNA-dependent Regulation of Transcription (GO:0006355). In all Cases, the ontological over-representation included a group of zinc finger proteins whose biological functions have not been well defined.

Biological Validation

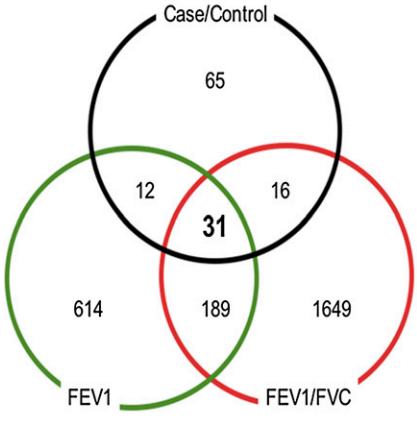
We measured the expression of 25 biomarker genes in a subset of our subjects ($n = 16$) by qPCR. We observed validation of the array data for 14 of the 25 genes, as defined by a significant correlation (Pearson or Spearman $P < 0.05$ using GAPDH) in expression between array-based and qPCR-based expression levels (Table 4). Another five genes showed a trend toward validation ($P < 0.10$). Similar results were obtained using an alternate endogenous control gene (PPIA; data not shown). In summary, we observe evidence for array expression validation for a total of 19 (76%) of the 25 biomarker genes.

Using either of the endogenous control genes, we were able to validate significant differences in expression between cases and controls for 6 of the 25 biomarker genes (Table 5; ARHGAP, C1QTNF7, CIRBP, HIF3A, HPGD, ZF10). Additional genes showed a trend toward significance and/or a substantial fold-change between cases and controls. By including qPCR data from the remaining 17 subjects ($n = 33$), we validated significant differences in expression between cases and controls for an additional four genes (CTSK, CYYR1, KLF8, SERPINB9). Another six genes (ARHGEF, KIT, KITL, PHACTR2, RUNX1T1, ZNF207) showed a trend ($P < 0.10$) toward differential expression. In summary, we observe evidence for differential expression for a total of 16 (64%) of the 25 biomarker genes (Table 5). In total, we were able to find some evidence of validation for 23 (92%) of the 25 biomarker genes.

DISCUSSION

Several approaches have been undertaken to discover biomarkers for COPD that may be useful for early diagnosis, prevention, therapeutic intervention, and prognosis. The first COPD biomarker was described by Eriksson, in that patients lacking α_1-antitrypsin, the principal inhibitor of neutrophil elastase, developed early-onset emphysema (29). Subsequent genetic studies have identified regions of the genome, and lists of gene variants, associated with COPD phenotypes (30, 31). DNA microarrays have been proven to be a major contributor in the discovery of biomarkers for various diseases. Microarray technology allows simultaneous comparison of expression of thousands of genes (32). Numerous studies on the use of DNA microarrays have supported the effectiveness of gene expression patterns for clustering diseased tissues apart from each other and from normal tissues. However, comparison of the observed gene expression data often reveals significant biases in classification schemes.

Recently, gene expression microarray analysis of human lung tissue has been used in an effort to identify biomarkers, distinguish disease subtypes, and generate candidates for further genetic and biological studies. Spira and colleagues reported genome-wide expression profiling of subjects with severe emphysema undergoing lung volume reduction surgery (13). These studies identified gene expression markers for severe emphysema as well as positive response to surgery. Golpon and coworkers used a similar approach and identified gene expression biomarkers distinguishing patients with α_1-antitrypsin deficiency (10). As with most disease-focused microarray studies, there has been a general lack of consistency in the identification of COPD gene expression biomarkers. One notable exception is EGR1. EGR1 was identified in a microarray study as a gene overexpressed in subjects with emphysema by Zhang and colleagues (14). Subsequently, Ning and coworkers, using a com-



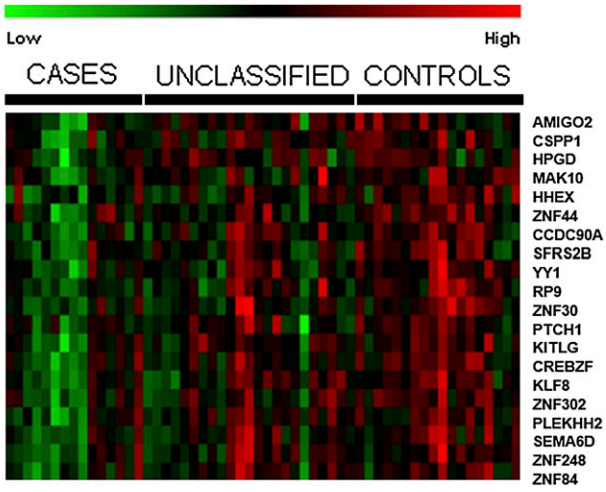


Figure 4. Biomarkers for both discrete and quantitative phenotypes. Shown are signal intensity measurements for 22 probe sets representing well-annotated genes (among 31 probe sets) identified as significantly different between cases and controls and significantly correlated with both quantitative phenotypes (FEV_1%predicted and FEV_1/FVC).

TABLE 3. CLASS PREDICTOR SIZE AND ACCURACY

No. of Probe sets	False Positive	False Negative	Sensitivity	Specificity
84*	1	0	100	93.33
74†	3	2	88.88	80
21‡	2	2	88.88	86.66
40	0	0	100	100

* Probe sets derived from union of markers of discrete and quantitative phenotypes.
† Probe sets derived from markers of quantitative phenotypes only.
‡ Probe sets derived from markers of discrete phenotype only.

bined microarray/SAGE approach, validated EGR1 induction associated with COPD severity (12). Ning and colleagues went on to show that EGR1 appears to contribute to disease pathogenesis, as it can regulate matrix remodeling potential through fibroblast protease production. Interestingly, we find no evidence of differential expression for EGR1 in our population with regard to either discrete or quantitative phenotypes.

We have recently used an integrated genomics approach to identify SERPINE2 as a candidate COPD susceptibility gene (15). These data indicated that SERPINE2 expression was significantly correlated with quantitative COPD phenotypes in the data set of Spira and coworkers (13). No probe sets for SERPINE2 passed the repeated criteria used in this current study to be defined as gene expression biomarkers. However, we did find significant association for each of the three SERPINE2 probe sets for individual quantitative traits (227487_s_at, $r_{FEV1\%predicted} = -0.36833$, $P = 0.0061$; 212190_at, $r_{FEV1\%predicted} = -0.28406$, $P = 0.037$; 236599_at, $r_{FEV1/FVC} =$

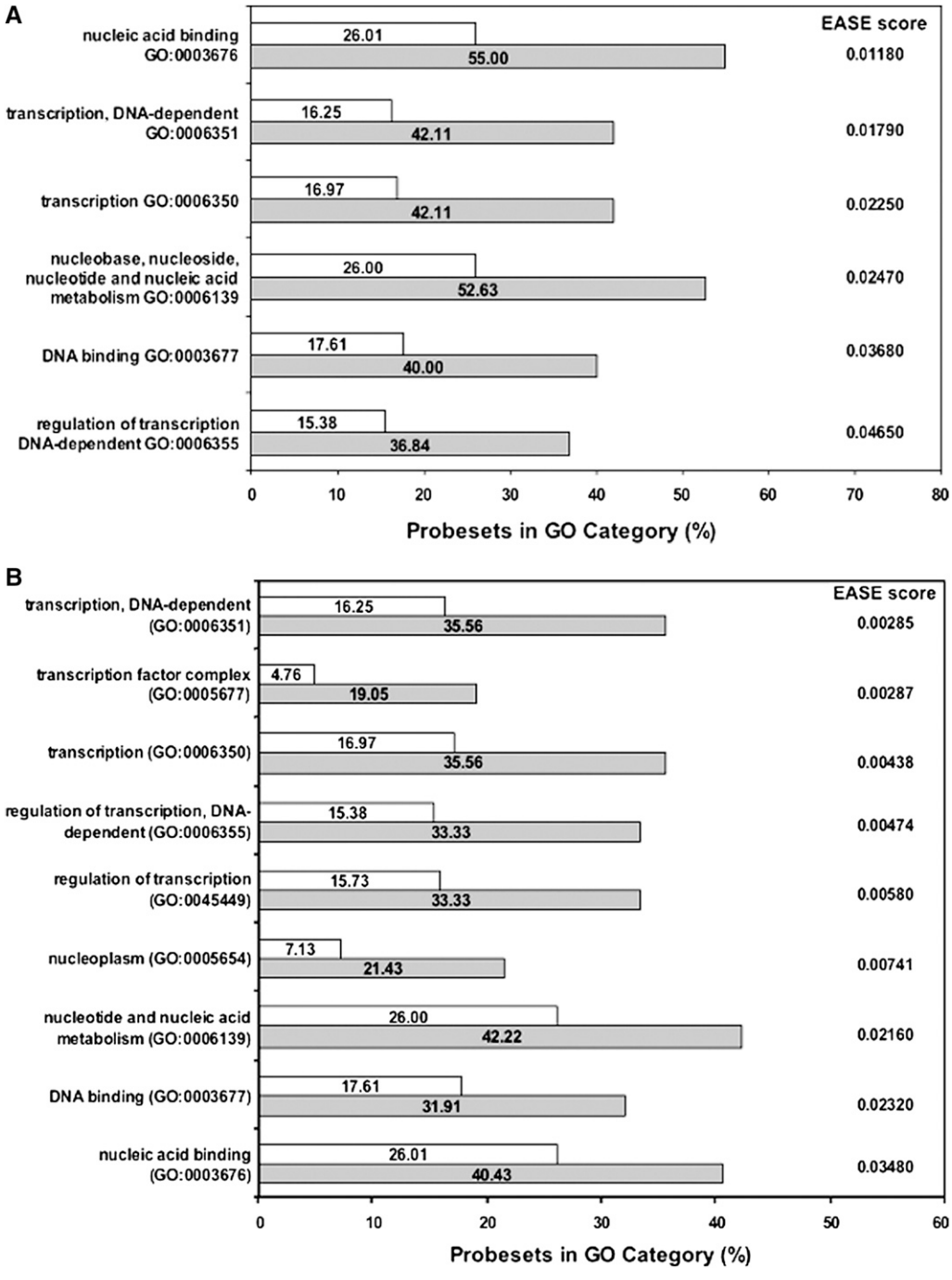


Figure 5. Gene ontology analysis. Ontologic categories significantly (EASE score < 0.05) overrepresented in (*A*) discrete, (*B*) quantitative, or a (*C*) union of COPD biomarkers. Given are GO category name and number, the percentage of genes within the category (plotted, with value) and the EASE score for the category. *Open bars*, all genes tested; *shaded bars*, COPD biomarkers.

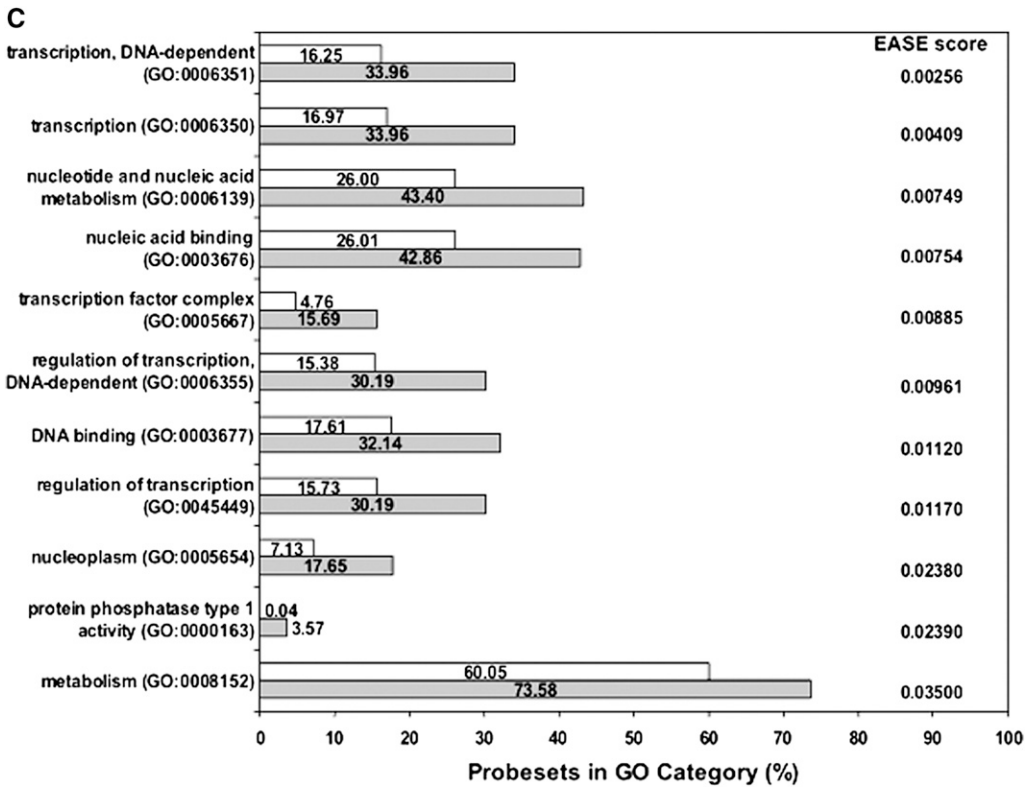


Figure 5. (*Continued*).

-0.28908, $P = 0.034$). These data are consistent with our previous observations, revealing robust and consistent increases in SERPINE2 gene expression in the lungs of subjects with airflow obstruction.

In the studies described here, we report the identification of a molecular signature for discrete and quantitative COPD phenotypes through the generation and analysis of microarray data from human lung tissue. We used a repeated approach for data analysis; gene expression level (signal intensity) values were extracted from raw data files using both nonnormalized (MAS5) and normalized (RMA) approaches, Frequentist (SAM) and Bayesian (BADGE) statistical methods were used to test for significant associations between gene expression and discriminate phenotypic variables (e.g., disease versus control), and linear (Pearson) and rank (Spearman) correlations were used to test for significance with continuous phenotypic variables (e.g., FEV_1%predicted, FEV_1/FVC). All analysis methods were repeated for each probe set and signal intensity data set. Results were summarized where data consistently implicated an association between gene expression and the disease variables. Initially, we identified genes differentially expressed between cases and controls, as has been performed in previous studies. In principle, differentially expressed genes (i.e., genes that are

TABLE 4. CORRELATION IN EXPRESSION PATTERNS BETWEEN MICROARRAY AND qPCR DATA

Gene	Correlation (r, GAPDH)	*P* Value
KLF8	0.793	$P < 0.05$
CYYR1	0.674	$P < 0.05$
AMIGO2	0.662	$P < 0.05$
HIF3A	0.612	$P < 0.05$
RUNX1T1	0.600	$P < 0.05$
KITLG	0.591	$P < 0.05$
SEMA6D	0.585	$P < 0.05$
DAAM1	0.545	$P < 0.05$
KIT	0.518	$P < 0.05$
HPGD	0.512	$P < 0.05$
IREB2	0.488	$P < 0.05$
MRPS25	0.468	$P < 0.05$
ZNF207	0.450	$P < 0.05$
ARHGAP	0.435	$P < 0.05$
CROP	0.416	$P < 0.10$
QKI	0.400	$P < 0.10$
SERPINB9	0.399	$P < 0.10$
PTCH	0.357	$P < 0.10$
C1QTNF7	0.341	$P < 0.10$

For a given gene, the RMA-derived signal intensity for each subject was correlated with the gene expression level determined by qPCR as defined by the dCt using GAPDH as the endogenous control gene. Listed is the maximum absolute value correlation coefficient and associated significance value for each gene.

TABLE 5. DIFFERENTIAL EXPRESSION ANALYSIS BETWEEN CASES AND CONTROLS BY qPCR

Gene	Fold-Change	*P* Value
RUNX1T1	0.126	$P < 0.10$
KITLG	0.130	$P < 0.10$
ZF10	0.211	$P < 0.05$
ARHGEF	0.266	$P < 0.10$
ARHGAP	0.333	$P < 0.05$
CYYR1	0.334	$P < 0.05$
KLF8	0.336	$P < 0.05$
HIF3A	0.384	$P < 0.05$
ZNF207	0.399	$P < 0.10$
KIT	0.434	$P < 0.10$
C1QTNF7	0.479	$P < 0.05$
CIRBP	0.479	$P < 0.05$
HPGD	0.547	$P < 0.05$
PHACTR2	1.716	$P < 0.10$
CTSK	2.198	$P < 0.05$
SERPINB9	2.398	$P < 0.05$

For a given gene, differential expression between cases and controls was calculated using the mean and standard deviation of the dCt values for each subject in a class. Listed is the highest significance level for each gene along with the associated fold-change.

expressed more in one group than another) should provide the highest predictive power, yet methods developed to date fall short in their ability to predict the status of known samples. The identification of genes differentially expressed in the presence or absence of COPD in our data set appeared to be driven by a subset of the subjects and was potentially biased due to the small sample size (used only 33 of 56 subjects) and phenotypic heterogeneity. In addition, we performed an assessment of gene expression changes associated with quantitative changes in lung function. This allowed us to use the entire data set and control for phenotypic heterogeneity as defined by FEV_1%predicted or FEV_1/FVC. We suggest that the combined set of genes identified in these studies represents a robust molecular signature for discrete and quantitative COPD phenotypes.

Finally, we assessed the utility of our methods and results to predict COPD in a separate data set. Biomarkers were developed using our heterogeneous subject population, containing individuals with wide-ranging levels of airflow obstruction. We tested these biomarkers in a more homogeneous population composed of subjects with severe emphysema (13). Using the 254 informative probe sets identified in our subjects, 84 of which were available in the data set of Spira and colleagues, we had 97% predictive accuracy and 100% sensitivity. This represents the first gene expression array biomarker for COPD validated in an independent population. In addition, we discovered a group of 40 of these probe sets (representing 38 genes) with 100% predictive accuracy.

Even though the establishment of a validated gene expression biomarker for COPD is a significant achievement, the current study has limitations. Due to the varying distribution of airflow obstruction in our study cohort, we chose to design analyses based on quantitative spirometry measures as opposed to GOLD criteria, as recently reported by others (12). We classified cases on the basis of general criteria for significant airflow obstruction characteristic of COPD, including $FEV_1 < 70\%$ predicted and $FEV_1/FVC < 0.7$, while controls showed no evidence of significant airflow obstruction ($FEV_1 > 80\%$ predicted, $FEV_1/FVC > 0.7$). Of course, one must consider variability in the measurement of quantitative traits such as lung function that may contribute to reliability of marker detection. Further, the presence of emphysema by radiology/surgical pathology was not thoroughly assessed in a majority of our subjects. The phenotypic heterogeneity of COPD may be the cause of limited replication of previous results in the current study, and in previous studies in general. Other confounding factors that limit the reliability of these types of studies include tissue sample heterogeneity and small number of samples relative to the number of genes tested. The effect of phenotypic heterogeneity upon marker identification, at least in theory, can be minimized by assessing quantitative variables of disease severity. We applied such an approach here to both offset the obvious disease heterogeneity in our subjects and to substantially increase our sample size ($n = 33$ versus 56). A sample size of 56 subjects makes this the largest gene expression biomarker study of COPD published to date. In addition, cigarette smoke can have broad and significant effects on gene expression. The genome-wide response to cigarette smoke exposure in airway epithelial cells has been reported (33). It will be of great interest to examine the relationships between gene expression changes resulting from cigarette smoke exposure and those consistently associated with COPD phenotypes as defined in the current study. Those genes that are responsive to smoke and differentially expressed in diseased individuals may represent true susceptibility factors.

Another potential limitation of the current study is the diagnosis of tumors in most subjects. Lung cancer and COPD are both typically found in smokers, and the diagnosis of lung cancer can serve as an independent predictor for COPD, independent of smoking history. In this study, the presence of malignant, or even benign, tumors may result in significant effects on gene expression in the distant, histologically normal lung tissue used for our gene expression studies (*see* Ref. 34). The vast majority of our subjects (80%) were diagnosed with either squamous cell carcinoma (34%) or adenocarcinoma (46%). We tested for and found no consistent differences in gene expression between tumor types within cases, within controls, or independent of lung function. Further, COPD biomarkers were not significantly differentially expressed between tumor types in any independent test. Finally, the potential influences of tumor upon gene expression did not limit the ability of our biomarkers to serve as successful class predictors in tumor-free patients with COPD. These data suggest that any effects of the tumor upon gene expression in distant, histologically normal tissue were not consistent or robust.

While there is no indication that the genes that we identified are etiological or causative in COPD pathology, an analysis of biomarker function using ontological assessment identifies an overrepresentation of genes involved in DNA binding and transcription factor activity. This was unanticipated and is independently observed for biomarkers of either discrete or quantitative COPD phenotypes. Historically, there has been only modest investigation of the involvement of transcriptional regulators in COPD pathogenesis. Notable exceptions include the previously identified and validated COPD expression biomarker gene EGR1 (12, 14), and the recent identification of Nrf2 as a genetic susceptibility factor for experimental emphysema in mice (35). Interestingly, histone deacetylase activity (HDAC2) has recently been implicated in gene dysregulation in human patients with COPD (36). The identity and regulatory function of individual biomarker genes identified in this study are not clear, but include a number of zinc finger–binding domain containing proteins.

We used a rigorous analytical approach for these studies, to identify the most robust and consistent set of biomarkers for discrete and quantitative COPD phenotypes. This strategy used multiple, independent microarray data extraction methods and repeated statistical testing. This approach is prompted by the limitations of any single analytical method when applied to complex, disease tissue–associated microarray data sets. This approach is supported by our successful validation using an independent COPD lung tissue data set. The genes we identified and validated have no previously described roles in processes relevant to disease pathogenesis, so they are more likely to be true markers rather than etiological. The identification of these markers may help to facilitate the development of noninvasive methods (such as genetic tests) that facilitate diagnosis, classification of disease subtypes, and/or provide a means to define response to therapeutic intervention. Further studies will be required to determine if any of these biomarker genes play a role in human COPD susceptibility or pathogenesis.

Conflict of Interest Statement: None of the authors has a financial relationship with a commercial entity that has an interest in the subject of this manuscript.

Acknowledgment: The authors thank Dr. William G. Richards for coordinating tissue procurement, handling, and processing as part of the Brigham and Women's Hospital Tumor Bank. The authors also thank Dr. Feng Tu for technical support, Adrienne Camp and Laura Hoffmeister for help with patient consent and documentation, Temana Andalcio and Debbie Bardi for data processing compilation, and Aditi Basu for assistance.

References

1. NHLBI. 2005. NHLBI Factbook. US Department of Health and Human Services, Public Health Service, National Institutes of Health, Bethesda, MD.
2. Lopez AD, Murray CC. The global burden of disease, 1990–2020. *Nat Med* 1998;4:1241–1243.

3. Chen JC, Mannino DM. Worldwide epidemiology of chronic obstructive pulmonary disease. *Curr Opin Pulm Med* 1999;5:93–99.
4. Murtagh E, Heaney L, Gingles J, Shepherd R, Kee F, Patterson C, MacMahon J. Prevalence of obstructive lung disease in a general population sample: the NICECOPD study. *Eur J Epidemiol* 2005;20: 443–453.
5. Barnes PJ, Hansel TT. Prospects for new drugs for chronic obstructive pulmonary disease. *Lancet* 2004;364:985–996.
6. Barnes PJ. Chronic obstructive pulmonary disease. *N Engl J Med* 2000; 343:269–280.
7. Hogg JC, Chu F, Utokaparch S, Woods R, Elliott WM, Buzatu L, Cherniack RM, Rogers RM, Sciurba FC, Coxson HO, *et al.* The nature of small-airway obstruction in chronic obstructive pulmonary disease. *N Engl J Med* 2004;350:2645–2653.
8. Janus ED, Phillips NT, Carrell RW. Smoking, lung function, and alpha 1-antitrypsin deficiency. *Lancet* 1985;1:152–154.
9. Larsson C. Natural history and life expectancy in severe alpha1-antitrypsin deficiency, Pi Z. *Acta Med Scand* 1978;204:345–351.
10. Golpon HA, Coldren CD, Zamora MR, Cosgrove GP, Moore MD, Tuder RM, Geraci MW, Voelkel NF. Emphysema lung tissue gene expression profiling. *Am J Respir Cell Mol Biol* 2004;31:595–600.
11. Golpon HA, Geraci MW, Moore MD, Miller HL, Miller GJ, Tuder RM, Voelkel NF. HOX genes in human lung: altered expression in primary pulmonary hypertension and emphysema. *Am J Pathol* 2001;158: 955–966.
12. Ning W, Li CJ, Kaminski N, Feghali-Bostwick CA, Alber SM, Di YP, Otterbein SL, Song R, Hayashi S, Zhou Z, *et al.* Comprehensive gene expression profiles reveal pathways related to the pathogenesis of chronic obstructive pulmonary disease. *Proc Natl Acad Sci USA* 2004; 101:14895–14900.
13. Spira A, Beane J, Pinto-Plata V, Kadar A, Liu G, Shah V, Celli B, Brody JS. Gene expression profiling of human lung tissue from smokers with severe emphysema. *Am J Respir Cell Mol Biol* 2004;31:601–610.
14. Zhang W, Yan SD, Zhu A, Zou YS, Williams M, Godman GC, Thomashow BM, Ginsburg ME, Stern DM, Yan SF. Expression of Egr-1 in late stage emphysema. *Am J Pathol* 2000;157:1311–1320.
15. DeMeo DL, Mariani TJ, Lange C, Srisuma S, Litonjua AA, Celedon JC, Lake SL, Reilly JJ, Chapman HA, Mecham BH, *et al.* The SERPINE2 gene is associated with chronic obstructive pulmonary disease. *Am J Hum Genet* 2006;78:253–264.
16. Bhattacharjee A, Richards W, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 2001;98:13790–13795.
17. Crapo R, Morris A, Gardner R. Reference spirometric values using techniques and equipment that meet ATS recommendations. *Am Rev Respir Dis* 1981;123:659–664.
18. Hankinson J, Odencrantz J, Fedan K. Spirometric reference values from a sample of the general US population. *Am J Respir Crit Care Med* 1999;159:179–187.
19. Mariani TJ, Reed JJ, Shapiro SD. Expression profiling of the developing mouse lung: insights into the establishment of the extracellular matrix. *Am J Respir Cell Mol Biol* 2002;26:541–548.
20. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003; 4:249–264.
21. Bhattacharya S, Long D, Lyons-Weiler J. Overcoming confounded controls in the analysis of gene expression data from microarray experiments. *Appl Bioinformatics* 2003;2:197–208.
22. The Tumor Analysis Best Practices Working Group. Expression profiling–best practices for data generation and interpretation in clinical trials. *Nat Rev Genet* 2004;5:229–237.
23. Klings ES, Safaya S, Adewoye AH, Odhiambo A, Frampton G, Lenburg M, Gerry N, Sebastiani P, Steinberg MH, Farber HW. Differential gene expression in pulmonary artery endothelial cells exposed to sickle cell plasma. *Physiol Genomics* 2005;21:293–298.
24. Sebastiani P, Yu YH, Ramoni MF. Bayesian machine learning and its potential applications to the genomic study of oral oncology. *Adv Dent Res* 2003;17:104–108.
25. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001;98:5116–5121.
26. Lyons-Weiler J, Patel S, Bhattacharya S. A classification-based machine learning approach for the analysis of genome-wide expression data. *Genome Res* 2003;13:503–512.
27. Simon DM, Arikan MC, Srisuma S, Bhattacharya S, Tsai LW, Ingenito EP, Gonzalez F, Shapiro SD, Mariani TJ. Epithelial cell PPAR [gamma] contributes to normal lung maturation. *FASEB J* 2006;20: 1507–1509.
28. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003;4:3.
29. Eriksson S. Pulmonary emphysema and alpha1-antitrypsin deficiency. *Acta Med Scand* 1964;175:197–205.
30. Silverman EK. Progress in chronic obstructive pulmonary disease genetics. *Proc Am Thorac Soc* 2006;3:405–408.
31. Silverman EK, Palmer LJ, Mosley JD, Barth M, Senter JM, Brown A, Drazen JM, Kwiatkowski DJ, Chapman HA, Campbell EJ, *et al.* Genomewide linkage analysis of quantitative spirometric phenotypes in severe early-onset chronic obstructive pulmonary disease. *Am J Hum Genet* 2002;70:1229–1239.
32. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 1996;93:10614–10619.
33. Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, Palma J, Brody JS. Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc Natl Acad Sci USA* 2004;101:10143–10148.
34. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, Gilman S, Dumas YM, Calner P, Sebastiani P, *et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med* 2007;13:361–366.
35. Rangasamy T, Cho CY, Thimmulappa RK, Zhen L, Srisuma SS, Kensler TW, Yamamoto M, Petrache I, Tuder RM, Biswal S. Genetic ablation of Nrf2 enhances susceptibility to cigarette smoke-induced emphysema in mice. *J Clin Invest* 2004;114:1248–1259.
36. Ito K, Ito M, Elliott WM, Cosio B, Caramori G, Kon OM, Barczyk A, Hayashi S, Adcock IM, Hogg JC, *et al.* Decreased histone deacetylase activity in chronic obstructive pulmonary disease. *N Engl J Med* 2005; 352:1967–1976.