Translating the COPD Transcriptome

Insights into Pathogenesis and Tools for Clinical Management

Julie E. Zeskind¹, Marc E. Lenburg^{1,2,3}, and Avrum Spira^{1,2,3}

¹Bioinformatics Program, Boston University, Boston, Massachusetts; ²Department of Pathology and Laboratory Medicine, and ³Pulmonary Center, Boston University School of Medicine, Boston, Massachusetts

While the role cigarette smoke plays in chronic obstructive pulmonary disease (COPD) is undisputed, the molecular mechanisms by which inhaled smoke contributes to disease pathogenesis remains unclear. One of the major barriers to effective approaches to diagnose and manage COPD is the remarkable heterogeneity displayed by patients with the disease. Whole-genome gene-expression studies of airway and lung tissue from patients with COPD provide an opportunity to gain insights into disease pathogenesis, allowing for both a molecular understanding of the pathogenic processes that contribute to this heterogeneity, and the ability to target therapies to these processes. This review focuses on synthesizing and integrating the limited numbers of high-throughput gene expression studies that have been conducted on lung tissue and airway samples from smokers with COPD. Comparing several lung tissue studies using computational approaches, we find that the results suggest fundamental similarities and identify common biological processes underlying COPD, despite each study having identified largely non-overlapping lists of differentially expressed genes. Given these similarities, we argue that additional lung tissue and airway gene-expression studies are warranted, and present a roadmap for how such studies could lead to clinically relevant tools that would impact COPD management.

Keywords: gene expression; microarray analysis; biomarkers; emphysema

Chronic obstructive pulmonary disease (COPD) is the fourth most common cause of death in the United States (1), and the mortality rate due to COPD continues to increase as heart disease and stroke mortality decline (2). Airflow limitation in COPD is usually progressive and is associated with an abnormal inflammatory response to cigarette smoking (3). COPD is diagnosed on the basis of fixed airflow obstruction that is not due to bronchiectasis, cystic fibrosis, or previous tuberculosis (3). This operational definition of COPD encompasses a broad spectrum of patients among whom there is substantial heterogeneity in terms of respiratory and systemic symptoms, pulmonary physiology, radiographic findings, histopathology, and, undoubtedly, underlying molecular disease processes.

Environmental exposures, most notably tobacco smoking (but also exposure to other particulate and gaseous pollutants), are associated with longitudinal decline of pulmonary function and are important causes of COPD. Several lines of evidence indicate that genetic factors also contribute to COPD. First, family studies have revealed increased risk of lung function impairment in smoking first-degree relatives of individuals with

COPD (4, 5), and population-based studies have shown substantial heritability of spirometry measures (6, 7). Second, severe α_1 -antitrypsin deficiency due to homozygous mutations in *SERPINA1* (*AAT*) gene is a documented cause of COPD (8). However, as the frequency of these deleterious variants in *AAT* is low, they explain only a small proportion of COPD prevalence.

The histopathology of COPD includes abnormalities of the central airways (subepithelial infiltration with macrophages and CD8⁺ T lymphocytes, mucous gland hypertrophy, and inflammation), peripheral airways (goblet cell hyperplasia, subepithelial infiltration with CD8⁺ T lymphocytes, smooth muscle hypertrophy, increased airway wall thickness, and loss of alveolar attachments), and lung parenchyma (centriacinar emphysema, panacinar emphysema, and infiltration of CD8⁺ T lymphocytes) (9). Among patients with COPD, there is great heterogeneity in the severity, relative predominance, and regional distribution of these pathologies. In some patients, parenchymal emphysema predominates, while others have a relative predominance of airway wall thickening. Although some data have suggested that the degree of parenchymal emphysema is the most important pathologic correlate of lung function impairment in patients with severe disease (10), the relative contributions of emphysema and airways disease are likely to differ among patients (11).

While the contribution of cigarette smoke to the pathogenesis of COPD is widely appreciated, the mechanism by which inhaled tobacco smoke contributes to disease pathogenesis remains unclear. Furthermore, we have very little insight as to why 10 to 20% of smokers develop COPD while the remainder does not. Whole-genome gene expression studies of diseased lung and airway tissue provide an opportunity to gain a detailed and unbiased portrait of the molecular consequences of COPD with the potential to gain insights into the processes that contribute to disease pathogenesis, and the relationship between the various pathologies that give rise to the clinical manifestations of COPD. The hope is that this detailed molecular understanding of COPD will both allow for patients to be categorized according to the specific disease processes responsible for their disease, and enable therapies to be targeted to these pathogenic processes. Realizing the promise of these studies has the potential to revolutionize the diagnosis and treatment of COPD.

We remain at the beginning of this path, as there are presently very few whole-genome expression studies of COPD in humans (12–16). Progress has been slow, and some of the initial enthusiasm for using genome-wide gene expression in the setting of COPD has been tempered by the apparent lack of overlap between these studies as to which genes are differentially expressed in response to COPD (15). In this review, we focus on synthesizing and integrating these studies. Using several computational approaches, we find that there is a high degree of overlap in the biological themes represented by the genes identified in these studies, and that while the lists of differentially expressed genes presented in each of the studies

(Received in original form July 30, 2008; accepted in final form August 28, 2008)

This work was supported by NIH/NHLBI 1R01HL095388 (AS, MEL), NIH/NIEHS U01ES016035 (AS, MEL), and NSF IGERT grant DGE-0654108 awarded to the Boston University Bioinformatics program (JEZ).

Correspondence and requests for reprints should be addressed to Avrum Spira, M.D., M.Sc., Boston University School of Medicine, Pulmonary Center, 715 Albany Street, R-304, Boston, MA 02118. E-mail: aspira@bu.edu

Proc Am Thorac Soc Vol 5, pp 834–841, 2008

DOI: 10.1513/pats.200807-074TH

Internet address: www.atsjournals.org

are largely nonoverlapping, there is considerable evidence that the gene-expression results of each paper share fundamental similarities. As a result, we believe that larger, better-powered studies are warranted. Future directions for such studies will be outlined to provide a roadmap for translating studies of COPD-related gene expression into clinically relevant tools that can impact disease management.

SUMMARY OF LUNG TISSUE GENE EXPRESSION STUDIES

As of June 2008, there have been four studies published on gene expression of lung tissue in COPD (12–15). Each study examined cohorts that were ascertained using slightly different inclusion and exclusion criteria, and identified genes that are differentially expressed as a function of different clinical variables. These differences between studies complicate comparisons of their findings. The cohorts, clinical samples, variables examined, and the gene expression technologies employed are summarized in Table 1.

Our group studied lung tissue from patients with COPD undergoing lung volume reduction surgery (LVRS) for severe emphysema ($n = 18$), as well as lung tissue from patients with little to no emphysema undergoing thoracotomy for suspicious pulmonary lesions ($n = 14$) (12), and identified genes that are differentially expressed as a function of four variables: presence of severe emphysema versus mild or no emphysema, diffusing capacity of the lung for carbon monoxide (DL_{CO}), forced expiratory volume in 1 second percent predicted ($FEV_1\%$), and the improvement or decline in BODE score after LVRS.

The BODE score integrates body mass index (BMI), degree of lung obstruction, perception of dyspnea, and 6-minute walk distance (17). This study also found that a number of genes that are differentially expressed in patients with severe emphysema map to chromosomal loci that have been shown to be associated with early-onset COPD in genome-wide linkage analyses. Genes found to be differentially expressed between lung tissue with severe emphysema and lung tissue with little to no emphysema included those involved in the extracellular matrix (up-regulated in patients with emphysema), immune and cell signaling (down-regulated in patients with emphysema), as well as those involved in stress-related processes including oxidoreductases, isomerases, and genes involved in complement activity.

Although this study employed a moderate sample size, it used leave-one-out cross-validation and chose only genes that were consensus picks among several algorithms as those that are differentially expressed in COPD. The decision to combine patients with no COPD and those with mild COPD as the reference group in this analysis may conceivably have limited the power of the study if individuals without emphysema have a different pattern of gene expression than individuals with mild early-stage disease. A perhaps more serious concern about this study is that the analysis may have been confounded by cancer-specific alterations in lung gene expression, as 64% of the individuals in the severe emphysema group had lung cancer compared with 0% of the control group. The relevance of this concern is highlighted by subsequent work from our group demonstrating that alterations in gene expression occur

TABLE 1. WHOLE-GENOME GENE EXPRESSION STUDIES IN LUNG TISSUE OF PATIENTS WITH CHRONIC OBSTRUCTIVE PULMONARY DISEASE

Study	Cohorts	Lung Tissue Sample	Clinical Variables Associated with Gene Expression	High-throughput Gene Expression Technology
Spira and colleagues, 2004	a) Severe emphysema: 18 patients undergoing LVRS b) Little to no emphysema: 14 current or former smokers undergoing thoracotomy for lesions suspicious for malignancy.	a) Severely emphysematous resected lung tissue b) Normal tissue adjacent to resected nodule	i) Severe emphysema vs. little to no emphysema ii) DL_{CO} iii) $FEV_1\%$ predicted iv) LVRS outcome	Affymetrix HG-U133A Genechip (> 14,500 genes), GPL96
Golpon and colleagues, 2004	a) Emphysema (smokers): – 5 regular emphysema – 6 AAT deficient emphysema Patients undergoing lung transplantation (94%) or LVRS (6%). b) No emphysema (nonsmokers): – 5 donor lungs not used for transplantation	Peripheral lung tissue section	i) No emphysema vs. AAT and regular emphysema ii) AAT-related emphysema vs. smoking-related emphysema	Affymetrix GeneChip Human Full Length Array HuGeneFL (> 5,600 genes), GPL80
Ning and colleagues, 2004	Smokers for pooled samples: a) Emphysema: 6 GOLD-2 b) No emphysema: 5 GOLD-0 Smokers for individually analyzed samples: a) Emphysema: 6 GOLD-2 b) No emphysema: 6 GOLD-0	Lung tissue samples from surgical specimens	GOLD-2 vs. GOLD-0	Pooled samples: i) SAGE ii) UniGEM-V cDNA microarrays from Incyte Pharmaceuticals (7,000 genes) Individually analyzed samples: i) Codelink cDNA microarrays from Amersham (9,000 genes) Agilent Functional ID v2.0 array (> 23,000 genes) from Rosetta/Merck, GPL3991
Wang and colleagues, 2008	Patients across a spectrum of emphysema severity: a) 3 GOLD-3 b) 10 GOLD-2 c) 9 GOLD-1 d) 21 GOLD-0 e) 5 nonsmokers with no emphysema Patients undergoing pneumonectomy or lobectomy for small peripheral lung nodules. All had lung cancer.	Tissue from any part of the lung except where the nodule was resected.	$FEF_{25-75\%}$ predicted	Agilent Functional ID v2.0 array (> 23,000 genes) from Rosetta/Merck, GPL3991

Definition of abbreviations: AAT = α_1 -antitrypsin; DL_{CO} = diffusion capacity for carbon monoxide; GPL = gene expression omnibus platform; $FEF_{25-75\%}$ = forced expiratory flow between 25 and 75% of forced expiratory volume; GOLD = global initiative for chronic obstructive lung disease; LVRS = lung volume reduction surgery; SAGE = serial analysis of gene expression.

throughout the respiratory tract in patients with lung cancer (27).

The study by Golpon and coworkers (13) took peripheral lung tissue samples from patients undergoing LVRS or transplantation for COPD. This included six smokers with emphysema associated with α_1 -antitrypsin deficiency and five smokers with emphysema of otherwise unknown etiology. Gene expression profiles in these diseased tissue samples were compared with those of lung tissue from nonsmoker organ donors. They compared gene expression in all emphysema samples versus the normal samples, α_1 -antitrypsin-deficient (AAT) emphysema versus regular emphysema, regular emphysema versus normal samples, and AAT emphysema versus normal samples. They found gene expression differences between AAT emphysema and non-AAT emphysema, with genes involved in protein biosynthesis, energy pathways, and electron transport being expressed at higher levels in AAT emphysematous lung tissue, and genes involved in the cellular defense response and small GTPase-mediated signal transduction being expressed at higher level in non-AAT emphysematous lung tissue. They also found differences between the emphysematous samples and the normal tissue.

The study of AAT emphysema versus normal emphysema is interesting and important, as AAT emphysema is often used as a COPD model. However, the ability to make firm conclusions about these differences is constrained by the small size of the cohort. Perhaps a more serious limitation is that the non-emphysematous tissue was from healthy nonsmoking lung donors, while 91% of the patients with emphysema were smokers. Thus, it is not possible to determine the relative contribution of smoking and emphysema to the gene expression differences between the two arms of the cohort.

Ning and colleagues (14) obtained surgical lung tissue specimens from smokers with no ($n = 12$) or moderate ($n = 14$) emphysema as staged using the Global Initiative for Chronic Obstructive Lung Disease criteria (GOLD-0 and GOLD-2, respectively). They did not specify why the patients were having surgery. It is therefore unclear if there are non-COPD pathologies that confound ascertainment to the various arms of the study. They pooled six GOLD-2 samples and five GOLD-0 samples and ran both SAGE and microarray analysis. They ran six each of the remaining GOLD-2 and GOLD-0 samples individually on microarrays. They also obtained fibroblasts from excised lungs from patients with COPD undergoing lung transplantation ($n = 3$) and from healthy donor lungs ($n = 3$). They performed additional immunofluorescence and quantitative reverse transcriptase-polymerase chain reaction (qRT-PCR) studies on these fibroblasts to confirm the COPD-related differential expression of EGR-1 expression, which they had identified in their microarray studies.

The use of SAGE is unique among the studies. SAGE relies on sequencing the 3' end of cDNA generated from each the samples (for review see Reference 18). Because differential expression is detected as differences in the number of times a particular sequence is detected in the collection of sequences from the healthy versus diseased pool, it can detect both expression and differential expression of relatively low abundance genes. Such genes are not usually measurable using microarrays due to the signal-to-noise characteristics of microarray technology. The results from the SAGE experiments and the microarray experiments in this study support each other, and the authors successfully validated a subset of their results in fibroblasts from additional patients. The sample size in this study was again quite small. However, the genes found to be differentially expressed between GOLD-2 and GOLD-0 patients have functions in processes that are thought to be involved in

COPD, which suggests that their results are in concordance with the biology of COPD.

Wang and coworkers (15) took lung-tissue samples from patients undergoing lung-nodule resection for suspected lung cancer; samples came from nonsmokers and smokers with COPD severity ranging from GOLD-0 to GOLD-4, for a total of 48 samples. They also examined the histopathology of each sample to determine percent parenchyma and other tissue types. They found 203 genes whose expression was associated with forced expiratory flow between 25 and 75 percent of forced expiratory volume (FEF_{25-75%} predicted), a measure of small airways function, while controlling for pack-years. They found that further controlling for the fraction of parenchyma in each tissue sample did not affect the ability to detect genes whose expression is associated with COPD. Using immunohistochemical methods, they additionally found no correlation between FEF_{25-75%} predicted and the number or percentage of alveolar macrophages positive for PLAU or PLAUR, suggesting that the gene expression differences for these genes are unlikely to be a result of inflammation or tissue remodeling.

The strengths of this study include its large sample size and the use of immunohistochemistry to explore the cell types that are responsible for COPD-related gene expression changes, as well as the contribution of inflammation and tissue remodeling to the lung function-dependent changes in gene expression. It is interesting that the authors associated gene expression with FEF_{25-75%} predicted, while FEV₁ predicted is a more standard measure of lung function in the context of COPD. While all samples from both smokers and nonsmokers are from patients with lung cancer, it is important to consider the possibility that the gene expression consequences of lung cancer may vary between sporadic and tobacco smoke-induced cancer.

INTEGRATION OF LUNG TISSUE GENE EXPRESSION STUDIES

A table published by Wang and colleagues (15) and reproduced in Table 2 shows that there is very little overlap between the genes that have been identified as being differentially expressed as a function of COPD-related markers in each study. This is likely due to the small sample size of each study and possibly to differences in the clinical markers of disease used for gene expression association. Differences in inclusion criteria, high-throughput gene expression platform, and analytical approaches may also contribute to this lack of overlap. One of the challenges of the meta-analyses that have been performed on these datasets is that each study relies on a P value threshold to determine whether a gene is significantly differentially expressed. However, due to the relatively small sample size in each study, the statistical power of each study to detect the complete set of genes that are affected by COPD is quite low. This complicates the interpretation of a gene not being detected as differentially expressed in any study. In the meta-analysis that we will describe

TABLE 2. A COMPARISON OF THE OVERLAP OF GENES IDENTIFIED TO BE DIFFERENTIALLY EXPRESSED IN FOUR CHRONIC OBSTRUCTIVE PULMONARY DISEASE GENE EXPRESSION STUDIES

Study	Wang and Colleagues	Spira and Colleagues	Golpon and Colleagues	Ning and Colleagues
Wang and colleagues	203			
Spira and colleagues	1	102		
Golpon and colleagues	0	1	74	
Ning and colleagues	1	5	3	325

Reprinted by permission from Reference 15.

below (Figure 1), we have employed a pathway-based analysis using Gene Ontology classifiers (19) to identify common biological themes in the genes identified as being differentially expressed in each study. In addition, we have used a method called Gene Set Enrichment Analysis (GSEA) (20, 21) to further explore the “*P* value threshold” problem. This approach ranks all genes according to their differential expression as a function of COPD (using a signal-to-noise metric) and then determines whether the genes identified as being differentially expressed in a second study (the “Gene Set”) are significantly skewed toward the top or bottom of that ranked list.

Gene Ontology Analysis

While the lack of overlap between the genes identified to be differentially expressed in patients with COPD in each of the studies has been described (15), the genes from each of the studies might be involved in biological processes that are generally perturbed in patients with COPD, and there might therefore be a higher degree of concordance between studies in the biological processes in which the identified genes participate. To test this hypothesis, we identified biological processes from the Gene Ontology project (19) that are significantly over-represented in the genes whose expression associates with COPD, and then examined whether the biological processes identified in this way were similar between the various studies. This analysis was performed with the Database for Annotation, Visualization, and Integrated Discovery (DAVID) and its associated program Expression Analysis Systematic Explorer (EASE) (22, 23), which is freely available (<http://david.abcc.ncifcrf.gov/>). Table 3 shows the GO biological processes that are over represented (Fisher’s Exact *P* value < 0.05) in the lists

of COPD-related differentially expressed genes identified by at least three of the four lung tissue gene expression studies. Although the gene lists overlap very little, many of the processes in which these genes participate are shared between studies.

GSEA Analysis

We used GSEA to determine if the gene sets from each paper are unevenly distributed in the ranked genes from the article by Spira and colleagues (12). GSEA software was obtained from the GSEA website (<http://www.broad.mit.edu/gsea/>). The Spira data is freely available through the Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) as series GSE1650. GSEA was performed using the 9,336 probesets that were reported by Spira and coworkers to be reliably detected, and the data were scaled and normalized as described (12). The genes corresponding to the probesets were ranked with a signal-to-noise metric according to the differential expression observed between samples from patients with severe COPD and those obtained from patients with no or mild COPD. The significance (*P* value) of the distribution of gene sets within the ranked list was determined by gene set permutation (20, 21) and corrected for multiple hypothesis testing (*q* value). This analysis shows that the genes identified in the studies by Ning and colleagues and by Golpon and coworkers as being differentially expressed as a function of COPD show similar patterns of differential expression in the Spira dataset. More specifically, the genes Ning and colleagues reported as more highly expressed in GOLD-2 patients versus GOLD-0 patients are expressed more highly in the severe emphysema samples from the study by Spira and colleagues relative to the mild and no emphysema samples in this dataset (*P* < 0.004 and *q* < 0.05). The same

Figure 1. Flow of analysis of chronic obstructive pulmonary disease (COPD) lung tissue gene expression study results performed in this review. Gray arrows indicate analysis flow; black arrows connect analysis descriptions to the data representations. GO categories enriched in each gene list were determined using the Database for Annotation, Visualization, and Integrated Discovery and its associated program Expression Analysis Systematic Explorer as described in the text. Overlap in genes whose expression is affected by COPD between datasets was determined using gene set enrichment analysis (GSEA) as described in the text. GO = Gene ontology. * Leading edge analysis is described in Reference 20.

TABLE 3. COMMON GENE ONTOLOGY BIOLOGICAL PROCESS CATEGORIES ENRICHED AMONG GENES IDENTIFIED TO BE DIFFERENTIALLY EXPRESSED IN LUNG TISSUE STUDIES

GO Biological Process Category	Spira and Colleagues (Severe Emphysema vs. Mild/None)	Spira and Colleagues (FEV ₁ % Predicted)	Golpon and Colleagues (Normal Lung vs. AAT and Regular Emphysema)	Golpon and Colleagues (Regular Emphysema vs. Normal Lung)	Ning and Colleagues (GOLD-2 vs. GOLD-0)	Wang and Colleagues (FEF ₂₅₋₇₅ % Predicted)
Cellular component organization and biogenesis		x		x	x	
Response to stress		x			x	x
Organ development	x				x	x
Biological adhesion/cell adhesion	x				x	x
Cellular macromolecule	x		x		x	
metabolic process/cellular protein metabolic process						

Definition of abbreviations: AAT = α_1 -antitrypsin; FEF₂₅₋₇₅% = forced expiratory flow between 25 and 75% of forced expiratory volume; GO = gene ontology; GOLD = global initiative for chronic obstructive lung disease

GO Biological Process categories enriched (Fisher Exact P value < 0.05) in the lists of COPD-related differentially expressed genes from three or more of the lung tissue gene-expression studies. A correction for multiple comparisons was not performed.

pattern is true for the genes Golpon reported as more highly expressed in patients with non-AAT emphysema versus control patients ($P < 0.05$, $q < 0.05$). The Wang results are not enriched in the Spira severe emphysema phenotype, which may reflect the fact that the Wang genes were chosen for their relationship to FEF₂₅₋₇₅% predicted, rather than their differential expression in severe emphysema.

“Leading Edge” genes in GSEA are the genes from each gene set that are most differentially expressed in the test set (20)—in this case, the data from Spira and coworkers. Using the GSEA software, we identified leading edge genes for the Ning and Golpon sets (containing genes each group identified to be expressed at a higher level in patients with emphysema). Biological categorization of these genes was performed with DAVID and its associated program EASE (22, 23) as described earlier. Biological categories that are significantly enriched in the leading edge genes from the analysis of the Ning gene list included GO biological processes involved in response to external stimulus and response to wounding. The expression of these genes in the Spira dataset was explored via hierarchical clustering (Figure 2). The leading edge of the Golpon genes fell into the GO biological process categories related to cell adhesion.

These analyses demonstrate the ability to find significant similarities between studies by examining the biological implications of each study's results, rather than the specific genes that met each study's criteria for significant differential expression. Small numbers of samples, different patient inclusion criteria, and different gene expression platforms ensure that the results of any two microarray studies are unlikely to ever be identical. Overall, the GO and GSEA analyses show that the four high-throughput gene expression studies implicate similar (though not identical) underlying biology. They lend support to the idea that gene expression studies can indeed detect COPD-specific differences and that such studies could be a worthwhile path to pursue for future research.

AIRWAY GENE EXPRESSION STUDIES

The studies that have been described up until now have focused on determining the gene expression consequences of COPD in lung tissue: a tissue in which disease-related processes can be expected to occur. In this section, we discuss the potential to study COPD-related processes by examining gene expression in more readily accessible proximal airway tissue. This work is motivated by the concept that smoking creates a “field of

molecular injury” in the epithelial cells that line the entire respiratory tract. In support of this, many groups (including our own), have studied gene expression in bronchial airway epithelial cells from smokers with and without smoking-related lung disease using high-throughput gene expression platforms. Initial studies showed that smoking effects airway epithelial cell gene expression (24, 25), and that while some of these gene expression changes revert toward baseline after smoking cessation, some genes remain irreversibly altered (26, 27). The effect of cigarette smoke on airway gene expression is heterogeneous between individuals, and we have found that some of this heterogeneity in large airway epithelial gene expression reflects the presence or absence of lung cancer (28). This cancer-specific gene expression heterogeneity can be used as a biomarker to distinguish smokers with and without lung cancer (28, 29).

The above studies support the hypothesis that airway gene expression reflects the host response to cigarette smoke and that heterogeneity in this response may associate with the presence of (or perhaps susceptibility to) tobacco-associated lung disease. Given that large airway epithelial cells can be collected in a less invasive fashion than lung tissue, profiling airway epithelium to gain insights into COPD-related disease processes would be extremely attractive if, as is the case with lung cancer, some of the heterogeneity in airway gene expression reflects differences associated with COPD-related processes. Using this site to study aspects of COPD is attractive, as it would make it possible to study large numbers of patients to gain insights into heterogeneity in the processes that contribute to disease pathogenesis. In addition, the ability to identify COPD-related processes in airway gene expression would raise the possibility of developing airway biomarkers that could be used clinically to identify smokers at higher risk for developing disease. Pierrou and colleagues have recently published a study comparing bronchial airway gene expression profiles in 38 smokers with COPD, 18 healthy smokers, and 14 healthy nonsmokers, in which they identified 200 oxidative stress-related genes that were differentially expressed between smokers with and without COPD (16). One of the major strengths of this study was the careful characterization of subjects using high-resolution computed tomography (HRCT) of the chest to select healthy smokers without emphysema. However, given the limited numbers of smokers with moderate to severe COPD, smokers with GOLD stage 2–4 were combined for the subgroup analysis within subjects with COPD. Interestingly, nonlinear gene expression patterns were identified for some these antioxidant genes across the spectrum of COPD stages/severity. GSEA analysis revealed

Figure 2. Hierarchical clustering of the Spira dataset according to the expression of genes identified in the Ning study (more highly expressed in GOLD-2 than GOLD-0 patients) that are present in the leading edge by GSEA. Gene expression was z-score normalized, and clustering was performed and visualized in R using the heatmap function. The tissue sample represented in each column is indicated by the gray and black bar. Black = severe COPD (lung volume reduction surgery samples); light gray = no COPD or mild COPD; red = higher expression; green = lower expression.

that these pathways are also among those most affected by smoking, with nonsmokers, healthy smokers, and smokers with COPD forming a continuum. Supporting the idea that COPD-related gene expression changes in the airway may be relevant to pathogenic processes occurring in the lung, the authors found that approximately 23% of the genes with COPD-specific patterns of differential expression in the airway were also found to have COPD-specific patterns of differential expression in the lung by Spira and coworkers (12).

FUTURE DIRECTIONS

The diverse clinical, radiographic, and pathologic findings in COPD represent a major challenge to molecular studies aimed at unraveling underlying molecular mechanisms. There is substantial heterogeneity among patients with COPD with regard to potentially confounding factors (e.g., medication use, comorbidities such as lung cancer, and degree of tobacco smoke exposure). There is also substantial heterogeneity with regard to the severity, relative predominance, and regional distribution of pathologic findings. As a result, future studies that combine detailed clinical and pathological phenotyping of subjects (e.g., HRCT of the chest, pulmonary function, chart review, and detailed questionnaires) with gene expression profiling will likely be required to dissect the various components of COPD. Furthermore, it is likely that it will be necessary to study the different tissues within the respiratory tract that are differentially affected by COPD-related pathologies. For example, separately profiling the small airways and parenchyma in tissue sections within subjects with COPD may lead to a better understanding of the potentially distinct disease processes that

underlie these major pathologic components of this disease. Gosselink and colleagues at the University of British Columbia have taken a first step in this direction by looking at expression of candidate genes via QRT-PCR in both the small airway and adjacent lung parenchyma from smokers with COPD (30).

Related to the above concern about the tissue specificity of different disease processes, the mixture of cell types in a given lung tissue sample represents a significant challenge to interpreting the biological significance of observed gene expression differences. In particular, there are significant concerns that COPD-associated gene expression changes may represent COPD-associated changes in the relative proportions of various cell types (e.g., COPD-specific neutrophil infiltration) as opposed to cell-intrinsic changes in gene expression. Localizing the cell type responsible for gene expression differences observed in whole lung tissue using *in situ* hybridization or immunohistochemistry, as done in the study by Wang and colleagues (15), is needed if gene expression studies of whole lung tissue are to lead to a better understanding of disease pathogenesis. While technically difficult, the use of laser capture microdissection to isolate various cell types from lung tissue before gene expression is another approach to address this issue.

Finally, one of the ultimate challenges in deriving mechanistic insights from a collection of genes that are differentially expressed as a function of any disease-related process is the difficulty in determining which genes are proximal to the disease-promoting process and which are secondary consequences of the activity of this process. One avenue that may identify at least a subset of the direct disease-promoting differences in gene expression involves integrating gene expression studies with large-scale genotyping studies of COPD: germ-line genetic

variants that are associated with altered COPD risk and are located within genes that are differentially expressed as a function of COPD are more likely to indicate gene-expression differences that are causally involved in the pathogenesis of COPD. This type of approach was used by Spira and coworkers to find genes differentially expressed in COPD that map to chromosomal regions linked to various phenotypes of early onset COPD (12). Computational approaches to identify gene regulatory networks are also likely to provide insights into the primary causes and secondary consequences of the complex biology that underlies COPD.

The study by Pierrou and colleagues (16) suggests the exciting possibility that gene expression changes in the proximal airway can be a useful surrogate for COPD-specific changes occurring in the lung. One hope is that such studies may ultimately yield biomarkers from this readily accessible tissue (which can be collected via bronchoscopy) that could be used clinically to provide information about which patients are most likely to develop COPD, as well as information about prognosis and disease subtype. Another exciting possibility is that treatment-induced changes in biomarker expression would serve as an intermediate endpoint to gauge both the efficacy of candidate COPD therapeutics and the clinical response of individuals to established COPD therapies. Such intermediate measures of disease progression and remission are needed for assessing the efficacy of candidate COPD therapeutics given the slow progression of COPD clinical symptoms.

SUMMARY

While limited numbers of high-throughput gene expression studies have been conducted on COPD lung tissue, they have begun to provide insights into disease pathogenesis and targets for therapy. Enthusiasm for these studies has been tempered somewhat by the limited overlap between the genes identified in each study as being differentially expressed as a function of COPD, but a more in-depth analysis reveals significant overlap between the studies and suggests that common biological pathways are seen to be altered by COPD across datasets. Gene expression profiling of proximal airway epithelium in the context of COPD offers the promise of using this more readily accessible tissue both for basic studies into COPD as well as clinically. There is hope that gene expression in the proximal airway will ultimately impact the diagnostic and prognostic evaluation of patients with COPD and for the potential use of gene expression to tailor anti-COPD therapy to the patient. There is also hope that proximal airway gene expression will be a useful tool for evaluating the response to novel and existing therapies. If gene expression studies of COPD are to ultimately yield clinical benefit, large well-designed microarray studies will need to be conducted to dissect and understand the various pathologic components of the disease.

Conflict of Interest Statement: J.E.Z. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript. M.E.L. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript. A.S. received an honorarium from AstraZeneca for a talk at the Lund Symposium in 2008.

Acknowledgments: The authors thank Adam Gower for his assistance in preparing Figure 2.

References

- Hoyert DL, Kochanek KD, Murphy SL. Deaths: final data for 1997. *Natl Vital Stat Rep* 1999;47:1–104.
- Hurd S. The impact of COPD on lung health worldwide: epidemiology and incidence. *Chest* 2000;117:1S–4S.
- Celli BR, MacNee W. Standards for the diagnosis and treatment of patients with COPD: a summary of the ATS/ERS position paper. *Eur Respir J* 2004;23:932–946.
- Kueppers F, Miller RD, Gordon H, Hepper NG, Offord K. Familial prevalence of chronic obstructive pulmonary disease in a matched pair study. *Am J Med* 1977;63:336–342.
- Sandford AJ, Silverman EK. Chronic obstructive pulmonary disease: 1. Susceptibility factors for COPD—the genotype-environment interaction. *Thorax* 2002;57:736–741.
- Givelber RJ, Couropmitree NN, Gottlieb DJ, Evans JC, Levy D, Myers RH, O'Connor GT. Segregation analysis of pulmonary function among families in the Framingham study. *Am J Respir Crit Care Med* 1998;157:1445–1451.
- Wilk JB, Djousse L, Arnett DK, Rich SS, Province MA, Hunt SC, Crapo RO, Higgins M, Myers RH. Evidence for major genes influencing pulmonary function in the NHLBI Family Heart Study. *Genet Epidemiol* 2000;19:81–94.
- Patel BD, Loo WJ, Tasker AD, Screaton NJ, Burrows NP, Silverman EK, Lomas DA. Smoking related COPD and facial wrinkling: is there a common susceptibility? *Thorax* 2006;61:568–571.
- Turato G, Zuin R, Saetta M. Pathogenesis and pathology of COPD. *Respiration* 2001;68:117–128.
- Nagai A, West WW, Thurlbeck WM. The National Institutes of Health Intermittent Positive-Pressure Breathing Trial: Pathology Studies. II: correlation between morphologic findings, clinical findings, and evidence of expiratory air-flow obstruction. *Am Rev Respir Dis* 1985;132:946–953.
- Snider GL. Chronic obstructive pulmonary disease: a continuing challenge. *Am Rev Respir Dis* 1986;133:942–944.
- 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.
- Golpon HA, Coldren CD, Zamora MR, Cosgrove GP, Moore MD, Tudor RM, Geraci MW, Voelkel NF. Emphysema lung tissue gene expression profiling. *Am J Respir Cell Mol Biol* 2004;31:595–600.
- 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.
- Wang IM, Stepaniants S, Boie Y, Mortimer JR, Kennedy B, Elliott M, Hayashi S, Loy L, Coulter S, Cervino S, et al. Gene expression profiling in patients with chronic obstructive pulmonary disease and lung cancer. *Am J Respir Crit Care Med* 2008;177:402–411.
- Pierrou S, Broberg P, O'Donnell RA, Pawlowski K, Virtala R, Lindqvist E, Richter A, Wilson SJ, Angco G, Moller S, et al. Expression of genes involved in oxidative stress responses in airway epithelial cells of smokers with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2007;175:577–586.
- Celli BR, Cote CG, Marin JM, Casanova C, Montes de OM, Mendez RA, Pinto P, V, Cabral HJ. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *N Engl J Med* 2004;350:1005–1012.
- Yamamoto M, Wakatsuki T, Hada A, Ryo A. Use of serial analysis of gene expression (SAGE) technology. *J Immunol Methods* 2001;250:45–66.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet* 2000;25:25–29.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;102:15545–15550.
- Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003;34:267–273.
- 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.
- Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. *Genome Biol* 2003;4:R70.
- Hackett NR, Heguy A, Harvey BG, O'Connor TP, Luetthich K, Flieder DB, Kaplan R, Crystal RG. Variability of antioxidant-related gene

- expression in the airway epithelium of cigarette smokers. *Am J Respir Cell Mol Biol* 2003;29:331–343.
25. 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.
26. Chari R, Loneragan KM, Ng RT, MacAulay C, Lam WL, Lam S. Effect of active smoking on the human bronchial epithelium transcriptome. *BMC Genomics* 2007;8:297.
27. Beane J, Sebastiani P, Liu G, Brody JS, Lenburg ME, Spira A. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol* 2007;8:R201.
28. 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.
29. Beane J, Sebastiani P, Whitfield T, Steiling K, Lenburg M, Spira A. A prediction model for lung cancer diagnosis that integrates clinical and genomic features. *Cancer Prev Res Phila Pa* 2008;1:56–64.
30. Gosselink JV, Hayashi S, Elliott WM, Chan B, Sin D, Pare PD, Pierce J, Pierce R, Patterson A, Cooper J, *et al.* Gene expression of peripheral lung remodeling in COPD [abstract]. *Am J Respir Crit Care Med* 2008;177:A656.