Biomarkers, 2010; 15(8): 715–730

informa
healthcare

ORIGINAL ARTICLE

Gene expression profiling of peripheral blood leukocytes identifies potential novel biomarkers of chronic obstructive pulmonary disease in current and former smokers

Jeffery S. Edmiston[1], Kellie J. Archer[2,3], Mariano J. Scian[4], Andrew R. Joyce[5], Barbara K. Zedler[5], and E. Lenn Murrelle[5]

[1]Research, Development & Engineering, Altria Client Services, Richmond, VA, USA, [2]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA, [3]StatSolvers, LLC, Richmond, VA, USA, [4]Rem X Speciality Staffing, c/o Altria Client Services, Richmond, VA, USA, and [5]Venebio Group, LLC, Richmond, VA, USA

Abstract

Background: Chronic obstructive pulmonary disease (COPD) is an inflammatory lung disease with associated systemic effects.

Objective: To use gene expression microarrays in peripheral blood leukocytes of current and former cigarette smokers to identify differences associated with COPD.

Materials and methods: Random forest modelling and a split-sample case–control approach were used to identify candidate predictors.

Results: We identified 1013 genes and one smoking exposure variable that differentiated current and former smokers with or without COPD. This predictor set was reduced to a nine-gene classifier (*IL6R*, *CCR2*, *PPP2CB*, *RASSF2*, *WTAP*, *DNTTIP2*, *GDAP1*, *LIPE* and *RPL14*).

Conclusion: These gene expression profiles represent potential biomarkers for COPD and may help increase mechanistic understanding of the disease.

Keywords: *Cigarette smoking; COPD; gene expression; microarray; leukocytes; random forest*

Introduction

Chronic obstructive pulmonary disease (COPD) is predicted to become the third leading cause of death worldwide by 2020 (Mannino & Braman 2007), and cigarette smoking is widely recognized as its primary environmental causative factor. The pulmonary component of COPD is primarily characterized by airway inflammation and incompletely reversible and usually progressive airflow obstruction (Barnes et al. 2003, Barnes 2003, Rabe et al. 2007). The pathophysiological mechanisms of COPD are thought to include an imbalance between protease and antiprotease activity in the lung, dysregulation of antioxidant activity and chronic abnormal inflammatory response to long-term exposure to noxious gases or particles leading to the destruction of the lung alveoli and connective tissue (Barnes et al. 2003, Barnes 2003, Rabe et al. 2007). However, COPD may be best characterized as a syndrome associated with significant systemic effects that are attributed to low-grade, chronic systemic inflammation (Agusti & Soriano 2008, Agusti et al. 2003, Fabbri&Rabe 2007, Rahman et al. 1996).

The opportunity to sample target tissue in order to study disease, such as lung in COPD, is often limited by the invasiveness of the procedures involved. Despite this difficulty, multiple groups have been successful

Address for Correspondence: Jeffery S. Edmiston, Research, Development & Engineering, Altria Client Services, 601 East Jackson Street, Richmond, VA 23219, USA. Tel: +1-804-335-2366. Fax: +1-804-335-2095. E-mail: Jeffery.S.Edmiston@altria.com

(Received 26 July 2010; accepted 27 July 2010)

ISSN 1354-750X print/ISSN 1366-5804 online © 2010 Informa UK, Ltd.
DOI: 10.3109/1354750X.2010.512091

in differentiating subjects with emphysema or COPD from control subjects using gene expression microarray analysis in lung tissue (Bhattacharya et al. 2009, Golpon et al. 2004, Ning et al. 2004, Pierrou et al. 2007, Spira et al. 2004, Wang et al. 2008). However, more readily accessible circulating leukocytes share more than 80% of the gene expression profile, or transcriptome, with many target tissues, including lung (Liew et al. 2006). In recent years, increasing evidence has shown that peripheral blood leukocytes (PBLs) can be used as target tissue 'surrogates' that accurately reflect disease or risk of disease (Mohr &Liew 2007). PBLs have been successfully used to identify gene expression differences associated with several inflammatory or autoimmune diseases, including asthma (Hansel et al. 2005), multiple sclerosis (Achiron&Gurevich 2006), systemic lupus erythematosus(Qing &Putterman 2004), pulmonary arterial hypertension (Bull et al. 2004), rheumatoid arthritis (Bovin et al. 2004) and osteoarthritis (Marshall et al. 2005). As shown by these studies, the generation of high-throughput data from PBLs may aid in the pathophysiological or mechanistic understanding of disease, as well as result in potential novel biomarkers for disease or disease risk.

The goal of the present study was to use cDNA microarray data to identify genes differentially expressed in PBLs between adult current and former cigarette smokers with or without COPD. In a case–control training set clearly defined by spirometric criteria, random forest statistical modelling was used to generate a list of variables that predicted COPD classification. This list was then subjected to a L_1 penalized logistic regression model to create a more parsimonious set of variables. Both lists were assessed in a case–control test set of subjects whose spirometric values more closely bordered the diagnostic cut-off value for COPD. The identified genes were analysed for their ontology assignment and pathway involvement. The gene expression profiles identified in this study have potential as novel biomarkers for COPD and may provide insight into disease mechanisms.

Materials and methods

Study design and subjects

The COPD Biomarker Discovery Study (CBD) was a cross-sectional study at the University of Utah to identify novel diagnostic, prognostic or therapeutic biomarkers of COPD in adult current or former cigarette smokers. The University of Utah Institutional Review Board approved the study protocol, and all subjects provided written informed consent. Male and female self-reported cigarette smokers, aged 45 years or older, with at least 10 pack-years smoking history were recruited from the University Health Sciences Network of local clinics and hospitals and from community physician offices. COPD was diagnosed in 300 subjects according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) spirometric guidelines as having a ratio of forced expiratory volume in 1 s (FEV_1) to forced vital capacity (FVC) <0.70 (Rabe et al. 2007). The control group included 425 sex- and age-matched (using 10-year bands), current or former cigarette smokers, without apparent lung disease who had FEV_1/FVC ≥0.70 and were recruited from the same clinical settings. Individuals who had recent exacerbation of COPD, uncontrolled angina, hypertension or allergy to albuterol, and female subjects who were pregnant or lactating were excluded. Demographic variables, respiratory symptoms and medical history, tobacco use history and concomitant medications were assessed. Pack-years were calculated as (maximum average number of cigarettes smoked daily over total smoking history/20) × (total years smoking). Body weight and height were measured. Spirometry was performed with a rolling seal spirometer by certified pulmonary function technicians according to American Thoracic Society guidelines (Miller et al. 2005). Measurements of FEV_1 and FVC were made before and at least 20 min after inhaled bronchodilator administration (albuterol 180 μg). The FEV_1/FVC ratio was calculated for each subject from the highest post-bronchodilator values of FEV_1 and FVC. A blood sample was collected for assessment of carboxyhaemoglobin (COHb) and complete blood cell counts.

A whole blood sample was also obtained for assessment of gene expression in PBLs in a subgroup consisting of the first 142 subjects enrolled in the study (81 with COPD and 61 unaffected controls). A power calculation supported this sample size. An extreme discordant phenotype design (Zhang et al. 2006), based on the FEV_1/FVC ratio, was used to select a training set for gene expression profiling in order to stratify maximally the case and control subgroups. Of the 142 subjects, 36 were clearly classified as having COPD (FEV_1/FVC <0.60), and 36 were clearly classified as unaffected controls (FEV_1/FVC >0.75) (Figure 1). Three COPD patients and two controls were excluded from the analyses due to missing key data, for a final training set of 67. The remaining 70 subjects with FEV_1/FVC values of 0.60–0.75 were used as the test set. One COPD patient was excluded due to missing key data, leaving 69 in the final test set.

Blood sample collection and processing

Whole blood samples were obtained from each subject by venipuncture using 10 ml EDTA Vacutainer® tubes (BD, Franklin Lakes, NJ, USA). COHb, haemoglobin, haematocrit and total and differential leukocyte counts were measured at ARUP Laboratories™, a national CLIA (Clinical Laboratory Improvement Amendments of 1988)-certified reference laboratory (Centers for Medicare & Medicaid Services 1992). Isolation of PBLs

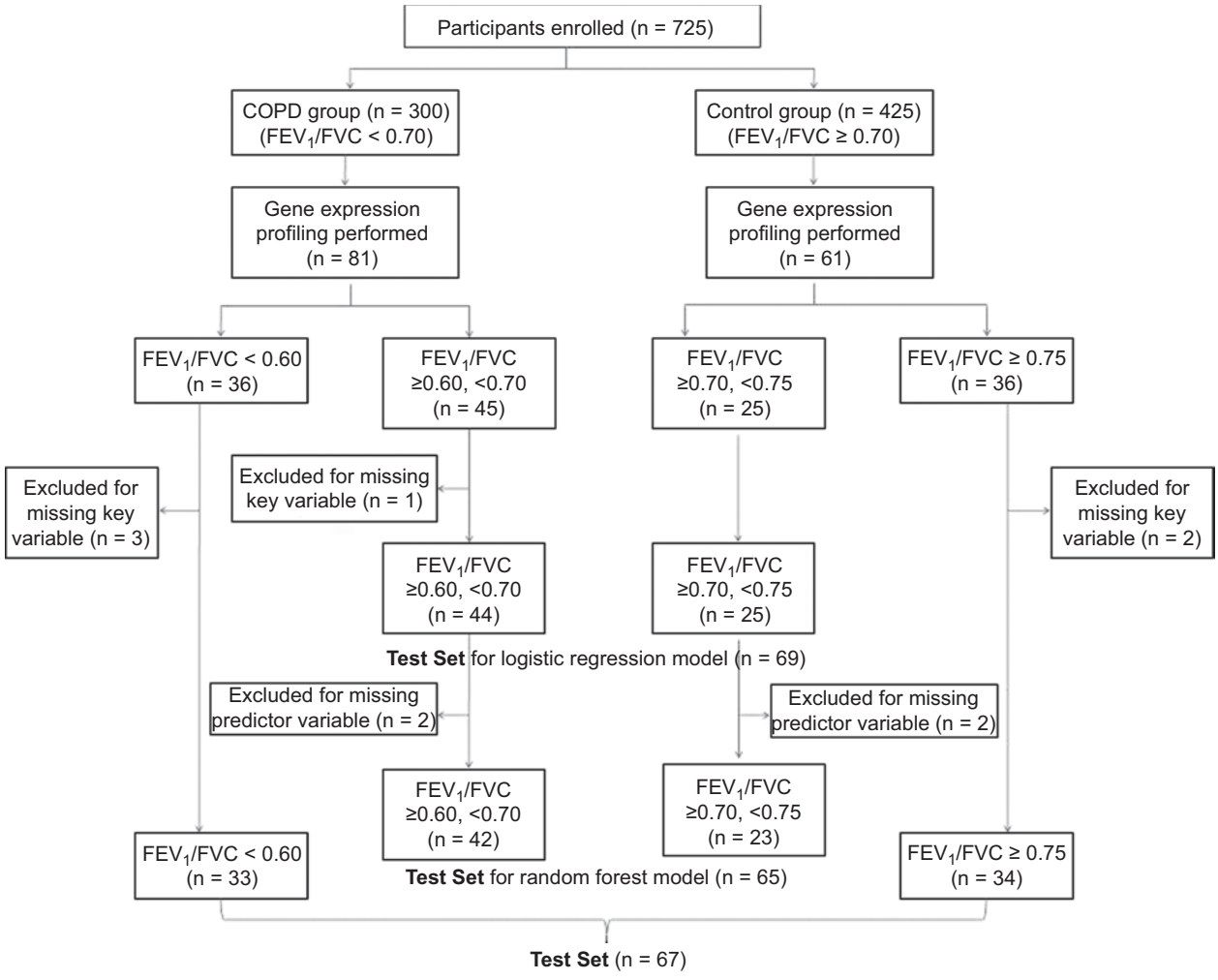

Figure 1. Study flow diagram. Due to the inability of the random forest algorithm to handle missing values among the predictor variables, four additional subjects with missing data were excluded from the test set for the random forest model (n=65) compared with the test set for the L_1-penalized logistic regression model (n=69).

was carried out using the LeukoLOCK™ Total RNA Isolation System (Ambion, Inc., Austin, TX, USA) following the manufacturer's protocol. Briefly, after isolation of PBLs, the filter was flushed with 3 ml of phosphate-buffered saline, to remove residual red blood cells, and then with RNAlater®, to stabilize the leukocyte RNA and frozen at -20ºC until processing for RNA. RNA isolation was then carried out using the mirVana™miRNA Isolation Kit (Ambion, Inc., Austin, TX, USA). The LeukoLOCK™ filter was flushed with 2.5 ml of mirVanamiRNA Lysis Solution, and the lysate was collected in a 15-ml conical tube. Then mirVanamiRNA homogenate additive (one-tenth volume) was added to the cell lysate. A volume of acid–phenol:chloroform, equal to the lysate volume, was used to flush the LeukoLOCK™ filter and collected into the same 15-ml conical tube as the lysate. The tube was shaken vigorously for 30 s and stored for 5 min at room temperature. The samples were centrifuged for 10 min at 10 000 *g* (maximum) in a table-top centrifuge. The aqueous phase was transferred into a new tube, mixed with 1.25 volumes of room-temperature 100% ethanol, and the mixture was filtered through the filter cartridge into the collection tubes supplied with the kit. The isolated RNA was then washed and eluted following the standard steps described in the kit's manual. Quality of the isolated RNA was checked using the Agilent 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA, USA) before use and storage at -80ºC.

Microarray data acquisition

Statistical procedures and analysis involved in preprocessing and identifying differential expression of microarray data were performed using Bead Studio® v3.0.14 (Illumina Inc., San Diego, CA, USA) and R-2.6.1 software (R Development Core Team 2007). cRNA from each sample following RNA isolation was hybridized to Sentrix® Human WG-6 BeadChips (Illumina Inc.). Hybridized BeadArrays™

were examined with respect to number of genes detected, average intensity, 95th percentile of signal intensity, signal-to-noise ratio and background signal intensity as a means of assessing quality. For each quality control (QC) measure, the BeadArray statistics were plotted and the mean ±3 standard deviations were overlaid on the plot as a method for identifying potentially outlying arrays. All BeadArrays were considered to be within acceptable limits for these QC measures. In addition, we examined the BeadArrays with respect to beadtypes labelled as hybridization, low and high stringency, biotin, housekeeping, and labelling controls (data not shown). All control beadtypes yielded intensities at the expected levels, therefore, each of the 142 hybridizations were considered to be of good quality.

Microarray data preprocessing

Prior to analysis, the gene expression data was $\log_2$ transformed. As negative control bead background correction was demonstrated to impact negatively, identifying differentially expressed genes (Dunning et al. 2008), the estimated background from the negative control beads was not subtracted from the mean beadtype signal intensities. The $\log_2$-transformed intensities were subsequently normalized using a global median scaling method. Specifically, the expression for each sample was scaled by an array-specific constant factor so that the median expression values were the same across all arrays. An arbitrarily selected array was set as the baseline against which all other arrays were normalized. For array array i and beadtype j, using the $\log_2$-transformed expression values $\log_2(x_{ij})$, global normalization was performed as follows: (1) the median expression for the baseline array $\tilde{x}_{base} = \underset{j}{\text{median}}\left(\log_2\left(x_{base,j}\right)\right)$, was calculated; (2) for the i^{th} array, the median expression, $\tilde{x}_i = \underset{j}{\text{median}}\left(\log_2\left(x_{ij}\right)\right)$, was also calculated; and (3) for the i^{th} array, $\beta_i = \tilde{x}_{\text{base}} / \tilde{x}_i$ was taken to be the global scaling factor and was applied to normalize the j expression values for array i so that the $\log_2$-transformed and scaled values for beadtype j and array i were $x_{ij}^{\text{norm}} = \beta_i \log_2(x_{ij})$.

Random forest analysis

The normalized gene expression data were combined with selected demographic, smoking history and clinical variables (see supplementary Table1 in the online version of this article). A random forest consisting of 10 000 trees was derived for predicting COPD-affected cases or unaffected controls, using a split-sample approach (training and test sets) and the *randomForest* package in the R programming environment (Breiman 2001, Liaw & Wiener 2002, R Development Core Team 2007).

The classification tree methodology is a heuristic algorithm whereby observations are recursively partitioned into disjoint subsets. Rather than using one classification tree (as classification trees are known to be unstable), multiple classification trees can be fitted where each tree is grown using a bootstrapped resample from the original dataset. When using the random forest methodology, in addition to using bootstrapped resamples for deriving each tree, at each node of the tree, m of the p independent variables are randomly selected from which to choose to split (Breiman 2001). The randomly selected covariate that results in the largest decrease in node impurity, as measured by the Gini impurity criterion, was selected for splitting the node. Moreover, for each classification tree in the random forest, the observations left out of the bootstrap resample (e.g. 'out-of-bag') were used as a natural test set for estimating prediction error. The out-of-bag observations were also used to estimate the importance of each variable for the classification task (Archer &Kimes 2008). The bootstrap method was used to estimate the null distribution for the mean decrease in Gini impurity by drawing a random sample with replacement from those variables with a non-zero mean decrease in Gini impurity, estimating the mean decrease of the resampled observations and repeating this procedure 2000 times. Candidate predictors with a Gini impurity >99.99795% were considered significant for the classification task.

L_1 penalized logistic regression

To identify a more parsimonious set of predictor variables that have a similar error rate as the random forest, an L_1 penalized logistic regression model was fitted to predict the dichotomous outcome variable (case/control status) using the significant candidate predictors identified by the random forest algorithm. This model was fitted using the same training set used to derive the random forest model, where the normalized gene expression and demographic, smoking history and clinical variables were standardized. L_1-penalized models, also referred to as least absolute shrinkage and selection operator (LASSO) models (Tibshirani 1996), are effective in performing automatic variable selection by shrinking some coefficients while setting other coefficients to zero by imposing a constraint on the estimated coefficients. Specifically, for p predictor variables, when fitting an L_1-penalized logistic regression model the likelihood is maximized subject to the constraint $\sum_{j=1}^{p} |\beta_j| \le t$. The *glmpath* library (Park & Hastie 2007) in the R programming environment (R Development Core Team 2007) was used for fitting the L_1-penalized models. The final model was selected as that model with minimum Akaike's information criterion (AIC) and was subsequently used to obtain fitted probabilities for all testable subjects. Those subjects with probabilities ≥0.5 were classified as cases, and all others were classified as controls.

Table 1. Characteristics of the spirometrically defined chronic obstructive pulmonary disease (COPD)-affected and unaffected (controls) subjects.

	All subjects			Training subset[a]		
Characteristic	COPD(n=81)	Controls(n=61)	p-Value[b]	COPD (n=36)	Controls (n=36)	p-Value[b]
Male (%)	67	62	0.60	64	61	1.00
Age (years)	61.2 (8.2)	54.8 (9.0)	<0.0001	63.3 (7.4)	52.6 (7.7)	<0.0001
Current smoker (%)	62	64	0.86	58	69	0.46
Cigarettes per day[c]	14.6 (17.0)	12.0 (12.3)	0.30	12.7 (14.1)	13.0 (13.4)	0.92
Pack-years	59.5 (38.0)	38.1 (19.8)	<0.0001	64.3 (38.8)	32.8 (19.3)	<0.0001
FEV_1 (l)	2.33 (1.01)	3.12 (0.79)	<0.0001	1.74 (0.94)	3.30 (0.75)	<0.0001
FEV_1 (% predicted)	70.6 (24.9)	94.6 (14.3)	<0.0001	54.2 (23.5)	99.0 (14.1)	<0.0001
FVC (l)	4.05 (1.32)	4.04 (1.01)	0.94	3.8 (1.47)	4.1 (0.97)	0.32
FEV_1/FVC (%)	56.3 (12.9)	77.4 (4.9)	<0.0001	44.7 (11.1)	80.8 (3.1)	<0.0001
WBC, total (10^3 μl^{-1})	7.4 (1.7)	7.6 (2.1)	0.57	7.6 (1.9)	7.3 (1.8)	0.51
Granulocytes (%)	64 (7)	59 (10)	0.004	66 (6)	57 (10)	<0.0001
Lymphocytes (%)	25 (7)	30 (9)	0.002	23 (6)	32 (10)	<0.0001
Monocytes (%)	6.2 (1.7)	5.9 (1.6)	0.19	6.4 (1.7)	5.7 (1.4)	0.06

FEV_1, forced expiratory volume in 1 s; FVC, forced vital capacity; PBL, peripheral blood leukocytes. Values are mean (±SD) unless otherwise indicated.[a]COPD subjects with %FEV_1/FVC <60 and control subjects with %FEV_1/FVC >75;[b]p-value for difference in mean values between the COPD/control groups was obtained by Welch's t-test for continuous variables and by Fisher's exact test for categorical variables; [c]average daily cigarette consumption of current smokers during the 3 months prior to study participation.

Gene ontology and pathway analysis

Genes identified statistically as having significant predictive value for the discrete case/control outcome were used as the input for subsequent gene ontology and pathway analysis. Gene ontology and functional categories were identified by analysing isolated gene lists using the Database for Annotation, Visualization and Integrated Discovery (DAVID; http://david.abcc.ncifcrf.gov/) (Dennis et al. 2003) and Pathway Studio V5.0 (Ariadne Inc., Rockville, MD, USA). EASE scores for gene-enrichment analysis were calculated using a 0.1 threshold from the program. The DAVID annotation tool was also used to probe the Kyoto Encyclopedia of Genes and Genomes (KEGG;http://www.genome.jp/kegg/kegg2.html), BioCarta (http://www.biocarta.com/genes/index.asp) and the Biological and Biochemical Image Database (BBID; http://bbid.grc.nia.nih.gov/) pathway databases to identify regulated pathways and to complement the gene ontology. 'Biological processes' and 'Pathways' with a p-value ≤0.05 were considered significant. The output analyses were manually filtered to remove overlapping and redundant categories to generate non-redundant lists. In addition to the gene ontology and pathway analysis performed with DAVID, protein–protein interactions of the genes were analysed with Pathway Studio™.

Quantitative real-time PCR

Quantitative real-time polymerase chain reaction (qRT-PCR) was performed on isolated RNA from 24 randomly selected subjects in the training set to confirm the microarray results in terms of differential expression and statistical significance. First-strand cDNA was synthesized from 1 μg of RNA in a 100 μl reaction volume with the TaqMan® Reverse Transcriptase Reaction Kit (Applied Biosystems, Carlsbad, CA, USA) using random hexamers as primers following the manufacturer's recommended protocol. After the synthesis was complete, the cDNA was diluted 1:3. Six microliters of diluted cDNA was then used for each qRT-PCR reaction in a final volume of 20 μl, using predesigned Gene Expression Assays (Applied Biosystems) for the genes of interest. All PCR reactions were carried out in triplicate. Relative expression levels were calculated using the ΔΔCt method algorithm provided by Applied Biosystems. The average intensity value obtained for the control subjects was used as the calibrator. All reactions were run in an Applied Biosystems 7500 Fast Sequence Detection System (Applied Biosystems). The gene expression assays used were: 18S (Hs99999901_s1), GAPDH (4310884E), DNTTIP2 (Hs00966646_m1), GDAP1 (Hs00184079_m1), IL6R (Hs01075667_m1), LIPE (Hs00943410_m1), WTAP (Hs00374488_m1), CCR2 (Hs00174150_m1), PPP2CB (Hs00602137_m1), RASSF2 (Hs00542460_m1) and RPL14 (Hs00427856_g1).

Results

Subject demographics

Characteristics of the spirometrically defined COPD-affected and unaffected control groups (overall and for the training set) are summarized in Table 1. The distribution of the COPD patients by severity of airflow obstruction, based on FEV_1 as a percentage of predicted by GOLD

spirometric guidelines (Rabe et al. 2007), was GOLD 1 (FEV_1<80% predicted, mild, *n*=30), GOLD 2 (FEV_1 50–80% predicted, moderate, *n*=38), GOLD 3 (FEV_1 30–50% predicted, severe, *n*=6) and GOLD 4 (FEV_1<30% predicted, very severe, *n*=7). Of note, ten subjects categorized as controls according to the GOLD guideline (FEV_1/FVC > 0.70) had subnormal FEV_1 (<80%predicted) and could be considered to have spirometrically indeterminate case/control status; three subjects were in the training set, and seven were in the test set. The clinical approach to such individuals is currently uncertain; therefore, these ten were retained in the control group. In the cohort overall and in the training and test sets, the COPD group was older and had at least 56% greater pack-years of cigarette smoking, on average, than the control group. However, the proportion of current smokers was similar across all groups, at 58–69%. Although the mean total PBL count did not differ significantly between the groups, the COPD groups had significantly higher mean neutrophils and lower mean lymphocytes, as percentages of the total PBL count, than did the control groups.

Identification of COPD predictors: random forest analysis

Due to the inability of the random forest algorithm to handle missing values among the predictor variables, the medication history of the subjects was not included in the analysis as a number of subjects had missing values. For example, 15/81 (19%) cases and 19/61 (31%) controls failed to indicate whether they were using glucocorticoids. The out-of-bag estimate of error associated with the random forest analysis in the training set was 6% overall, with a discordant classification rate of 3% for the spirometric controls and 9% for the spirometric cases (Table 2). The random forest algorithm identified 1014 candidate predictor variables, which included only one smoking exposure variable, 'years of daily smoking'. The top 30 candidate predictors using the mean decrease in Gini impurity, as well as the mean decrease in accuracy, are displayed in Figure 2. (The complete list of predictors can be found in supplementary Table 2 in the online version of this article.)

Table 2. Spirometric class versus random forest model-predicted class with associated class-specific discordance rates for the training set (FEV_1/FVC <0.60 or >0.75) and the test set (FEV_1/FVC 0.60–0.75).

	Spirometric class			
	Training set (*n*=67)		Test set (*n*=65)	
	COPD	Controls	COPD	Controls
Predicted class				
COPD	30	1	27	2
Controls	3	33	14	22
Discordance rate (%)	9	3	34	8

FEV_1, forced expiratory volume in 1 s; FVC, forced vital capacity.

The random forest model derived in the training set was then applied to the test set. Two COPD patients and two controls were excluded due to missing values for a predictor variable, leaving 65 subjects as a test set for evaluation of the random forest classifier. The overall discordant classification rate for the test set was 25% (16/65). Spirometric versus gene expression-predicted classifications for the training and test sets are shown in Table 2, along with discordant classification rates. Of the discordantly classified subjects in the test group, 14/16 (88%) were classified as cases by spirometry but not by their gene expression profile. For some observations, the posterior probabilities were near 0.5, indicating that the subject was difficult to classify as either case or control (see supplementary Table 3 in the online version of this article). For example, one of the spirometric controls in the test set that both gene expression-based models classified as a case could be considered to have spirometrically indeterminate case/control status as described in the Subject demographics section earlier in this paper.

Gene ontology and pathway analyses

In an effort to identify biological processes and pathways that were differentially affected in cases versus controls, we performed gene ontology assessment using the DAVID annotation tool (Dennis et al. 2003). A total of 784 genes (84% of the 1013 genes identified by random forest modelling) were represented in the DAVID gene ontology categories. The analysis output list was manually edited to remove redundant and overlapping gene ontologies. Biological processes that were enriched in the set of predictor genes between COPD cases and controls included regulation of apoptosis and cell growth, macromolecule (protein and RNA) transport, post-translational protein modification, cellular defence response, inflammatory response and RNA processing (Figure 3). Major pathways identified by DAVID included apoptosis (mitochondrial apoptotic signalling and caspase cascade), p38 MAPK, WNT and PPAR signalling, focal adhesion and leukocyte transendothelial migration (Figure 4).

The gene ontology analysis of the predictor genes revealed a number of upregulated genes involved in positive regulation of apoptosis (e.g. *BAD*, *CASP4*, *CASP6*, *CASP10*, *DIABLO*, *FAF1*, *FASTK* and *TRADD*) as well as a number of genes involved in inhibition of apoptosis (e.g. *BCL2L1*, *BIRC2*, *CDKN2D*, *MCL1*, *NAIP*, *SERPINB2*, *SGMS1* and *YWHAZ*). A similar situation is seen with cell cycle progression related genes. Several of the genes identified are involved in cell development and cell growth (e.g.*CCT7, CDC2L1, CDK2, CDC42, CDKN2D, MDM4, NEDD9, PCNA, PML, PMS1, RASSF2, RASSF4, RASSF5, RB1, TSC1, VEGFB* and *VHL*) with a number of them clearly involved in its negative regulation (e.g. *CDKN2D, PML, RASSF2, RASSF4, RB1* and *TSC1*).

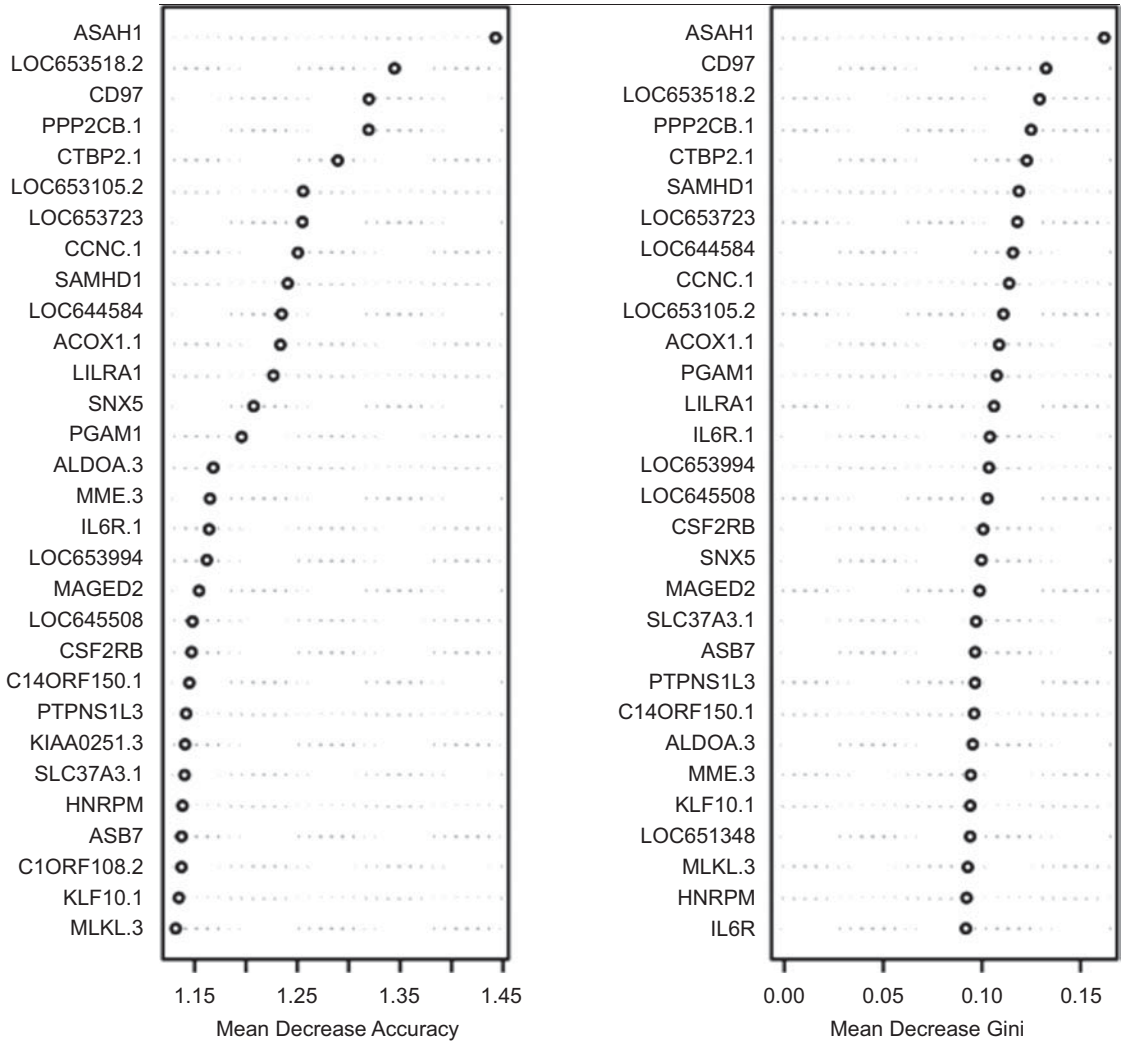

Figure 2. Chronic obstructive pulmonary disease (COPD) candidate predictors. Top 30 candidate predictors sorted in decreasing order by mean decrease in accuracy (left panel) and mean decrease in Gini impurity (right panel).

Table 3. Spirometric class versus L_1-penalized logistic regression model-predicted class with associated class-specific discordance rates for the training set (FEV_1/FVC <0.60 or >0.75) and the test set (FEV_1/FVC 0.60–0.75).

	Spirometric class			
	Training set (n=67)		Test set (n=69)	
	COPD	Controls	COPD	Controls
Predicted class				
COPD	32	1	31	2
Controls	1	33	13	23
Discordance rate (%)	3	3	30	8

FEV_1, forced expiratory volume in 1 s; FVC, forced vital capacity.

We also identified a number of genes involved in the intracellular signalling cascade (e.g. *ATF2, ATF4, DUSP6, DUSP10, IL1R2, MAP2K3, MAP4K3, MAPK14, MAX, MEF2A, PIK3R5, SOS1, SOS2* and *TGFBR2*) and in inflammatory response (e.g.*ALOX5, CCL7, CCR2, CCR4, CD97, CD163, NFRKB, NLRP3, PLAA, SPN, TLR4, TLR6, TLR8*) that differed between COPD cases and controls. This is consistent with prior reports in the literature and the systemic proinflammatory characteristics associated with COPD (Agusti& Soriano 2008, Agusti et al. 2003, Chung 2001, 2005, Fabbri&Rabe 2007, Mossman et al. 2006, Rahman et al. 1996, Rahman 2005). In addition, a summary of the protein–protein interactions of the genes and possible biological outcomes identified by Pathway Studio™ is shown in Figure 5.

L_1-penalized logistic regression model

In order to identify a more parsimonious set of variables having a similar predictive capability as the random forest, an L_1-penalized logistic regression model was fit to predict the dichotomous outcome variable (case/control status) using the 1014 variables identified by the random forest algorithm. L_1-penalized models are effective in performing automatic variable selection (Tibshirani 1996). The model

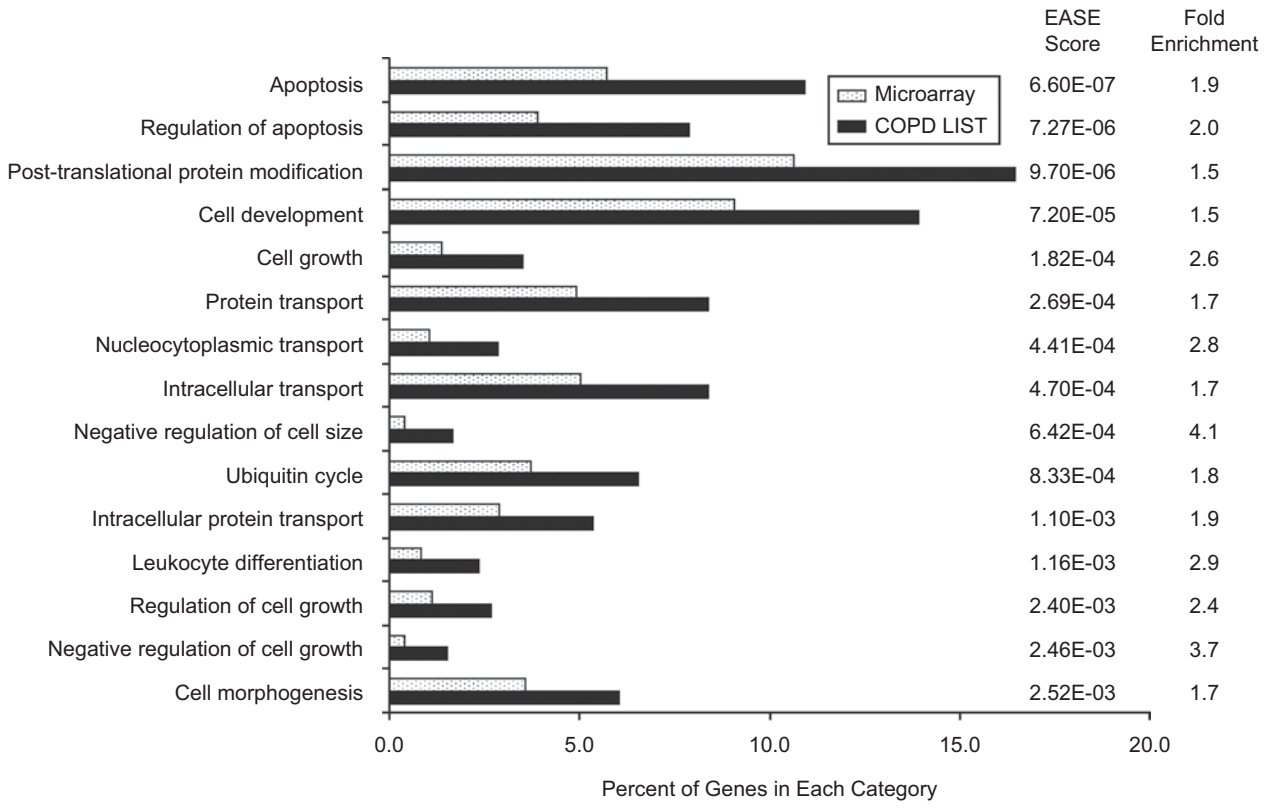

Figure 3. Top 15 DAVID annotated biological processes, including the gene ontology category name, percentage of genes within the category, EASE score and fold enrichment. All categories presented in the figure have an EASE score (*p*-value) <0.01 and a fold enrichment >1.5. 'COPD LIST' refers to genes identified as being enriched in chronic obstructive pulmonary disease cases by random forest analysis; 'Microarray' refers to all the genes represented in the array.

was first fitted using random forest algorithm data from the same training set of 33 cases and 34 controls used to derive the random forest model. The final model, selected as the L_1 logistic regression model with minimum AIC (data not shown), comprised nine predictor genes: *IL6R, CCR2, PPP2CB, RASSF2* and *WTAP* were upregulated and *DNTTIP2, GDAP1, LIPE* and *RPL14* were downregulated in cases compared with controls (Figure 6A). As shown in Table 3, the nine-gene model had an overall error rate of 3%, discordantly classifying one spirometric case and one spirometric control. The derived L_1-penalized logistic regression model was subsequently applied to classify the test set of 69 subjects with FEV_1/FVC of 0.60–0.75 (the two COPD and two controls excluded from the random forest analysis due to missing data were included in this analysis). The overall discordant classification rate was 22% (Table 3). The calculated sensitivity, specificity and positive and negative predictive values in the test set of samples for both models are shown in Table 4.

Biological validation

Results of real-time PCR performed to confirm the microarray results for the nine predictor genes are shown in Figure 6B. Similar to the study of Bhattacharya et al. (2009) using microarray analysis of lung tissue from subjects with COPD, not all of the predictors from the microarray data were confirmed by qRT-PCR. However, we observed a concordant directional trend in differential expression (Pearson correlation coefficient 0.795) between the two platforms for seven of the nine genes, although in some instances the magnitude of the difference between cases and controls by qRT-PCR varied from that detected by microarray. No statistically significant differences were observed for *PPP2CB* and *GDAP1* by qRT-PCR.

Discussion

Using microarray analysis of PBLs and random forest modelling, we identified 1013 genes and one smoking exposure variable (years of daily smoking) as candidate predictors capable of differentiating current or former smokers with or without COPD. Gene ontology analyses indicate that these genes are involved in various cellular processes including regulation of apoptosis, regulation of cell growth, macromolecule (protein and RNA) transport, post-translational protein modification, cellular defence

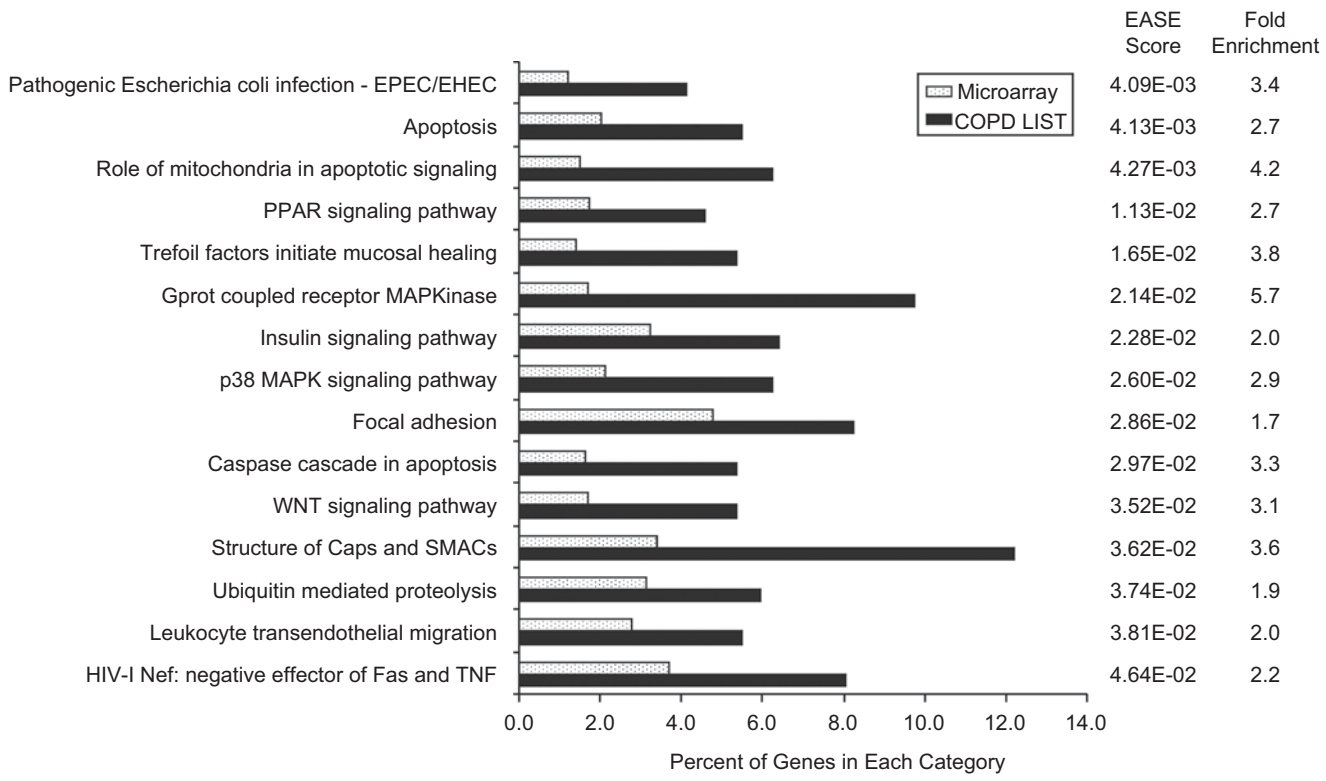

Figure 4. DAVID annotated biological pathways, including the percentage of genes identified, EASE score and fold enrichment. All pathways shown in the figure have an EASE score (p-value) < 0.01 and a fold enrichment >1.5. 'COPD LIST' refers to genes identified as being enriched in chronic obstructive pulmonary disease cases by random forest analysis; 'Microarray' refers to all the genes represented in the array.

response, inflammatory response and RNA processing. We further identified a nine-gene subset derived from the larger set of candidate predictors that reliably discriminated between COPD and non-COPD subjects. Differential expression of seven of the nine genes identified was confirmed by qRT-PCR, corroborating the microarray results.

The full random forest predictive model discordantly classified 6% of the spirometrically defined training set and 25% of the test set, and the nine-gene model differed from the spirometrically defined classification for 3% of the training set and 22% of the test set. Thus, these models performed well in the more phenotypically extreme training set and less well in the test set whose FEV_1/FVC values more closely bordered the accepted diagnostic case/control cut-off value of 0.70. Interestingly, the great majority of the discordantly classified subjects in the test set were classified as cases by spirometry but controls by their gene expression profile. Spirometry results depend partly on subject effort and even with good procedural QC, it is possible for an individual to have a spuriously low airflow measurement that could result in a misdiagnosis of COPD using the fixed, arbitrary GOLD cut-off value for FEV_1/FVC. Gene expression, on the other hand, is not affected by subject effort, and may therefore represent an alternative diagnostic methodology, particularly in individuals with spirometrically indeterminate case/control status. In fact, one such spirometrically indeterminate control in our test set was classified as a COPD case by both gene expression-based models.

Furthermore, although spirometric parameters are the traditional diagnostic and prognostic markers for COPD, it has become clear that they do not adequately represent all of its respiratory and systemic aspects (Celli 2006, Marin et al. 2009). FEV_1 correlates poorly with the degree of dyspnea, and the change in FEV_1 does not reflect the rate of decline in health status (Burge et al. 2000, Celli et al. 2004, Celli 2006). Other factors, such as emphysema and hyperinflation (Casanova et al. 2005), malnutrition (Schols et al. 1998), peripheral muscle dysfunction (Maltais et al. 2000) and dyspnea (Nishimura et al. 2002), are independent predictors of outcome. In fact, the multifactorial BODE index that includes body mass index (B), degree of airflow obstruction (O), dyspnea score (D), and exercise endurance (E), was a better predictor of mortality than FEV_1 alone (Celli et al. 2004). The PBL gene expression profile alone, or more likely in combination with currently accepted clinical markers such as the BODE components and/or lung parenchymal or airway changes on chest computed tomography (CT) scans (Omori et al. 2006), may be more predictive of the (early) presence, activity, and progression of the multicomponent syndrome that is COPD than the clinical parameters alone.

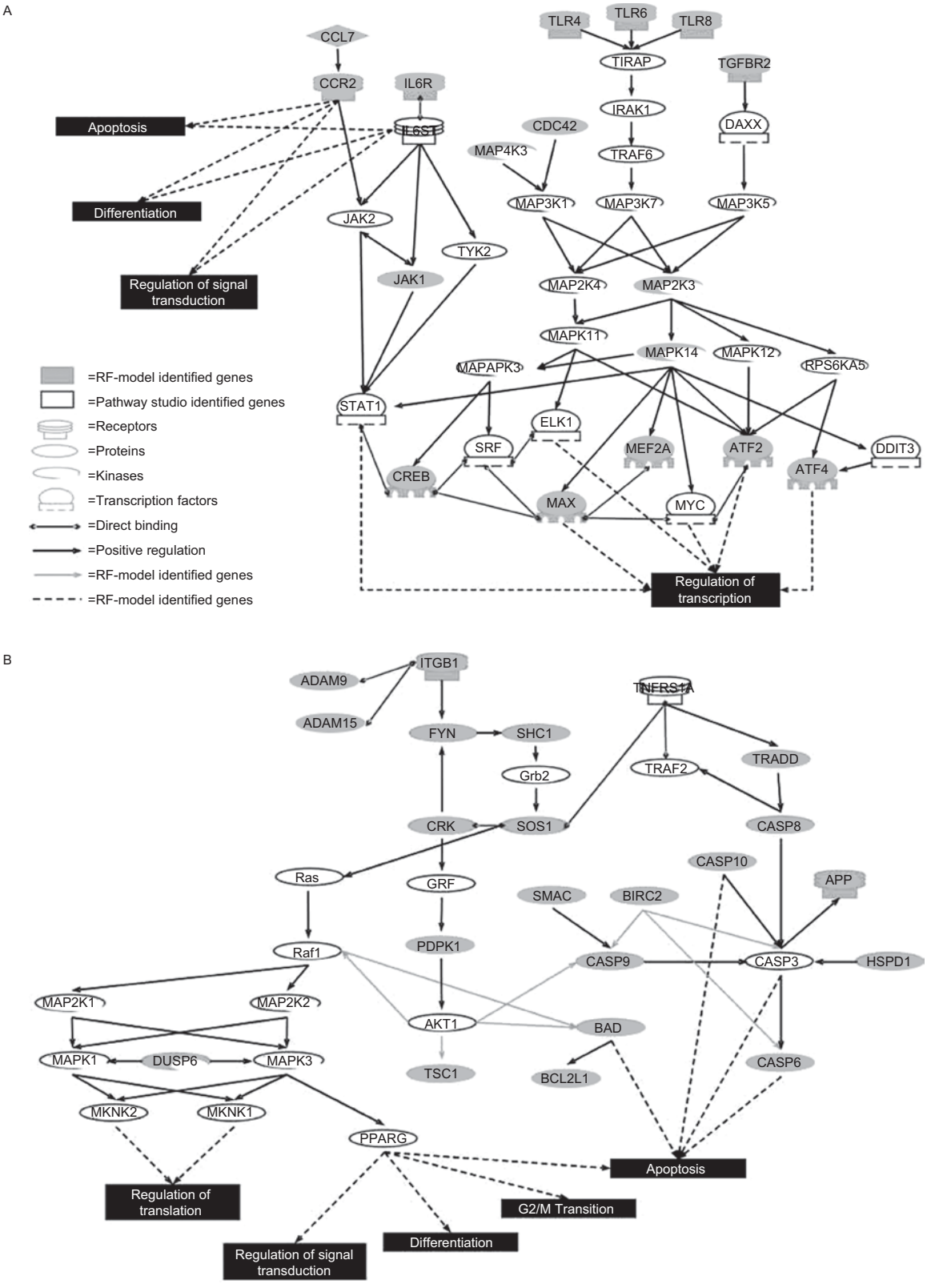

Figure 5. Summary of potential regulatory interactions and possible biological outcomes identified using the Pathway Studio software. (A) Protein–protein interactions primarily associated with the MAPK signalling cascade. (B) Protein–protein interactions associated with the apoptotic cascade. Note that MAP2K4 can phosphorylate and activate MAPK1 and that binding of MAP3K1 to TRAF2 results in their subsequent activation providing two potential links between the two pathways depicted in A and B (Chadee et al. 2002, Witowsky& Johnson 2003). Grey proteins represent random forest model-identified genes; white proteins represent Pathway Studio-identified genes.

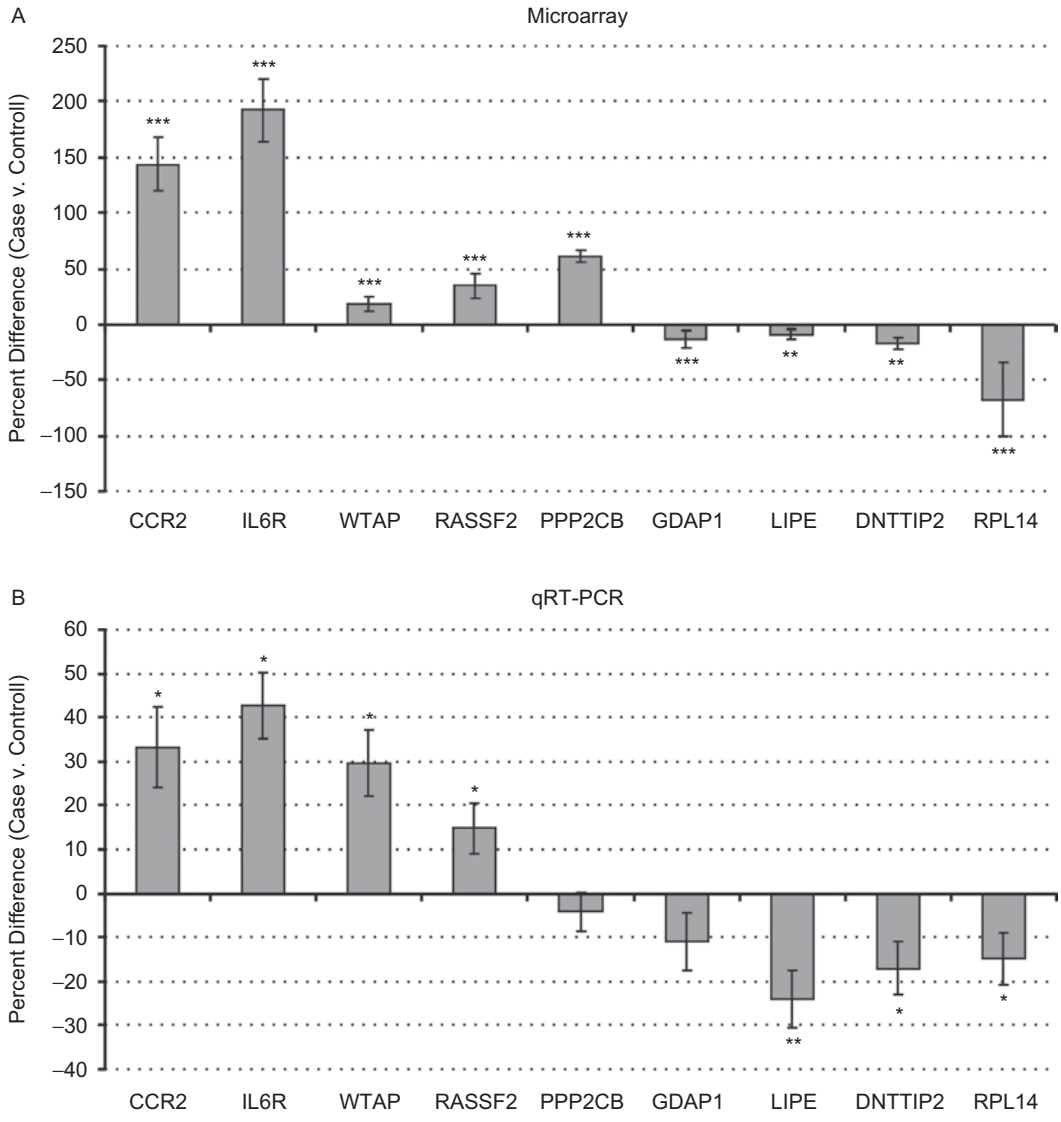

Figure 6. Gene expression forL_1-penalized logistic regression model. (A) Microarray results for the randomly selected samples from the training set (12 controls and 12 cases). Relative mRNA fold difference in expression was calculated using the control group as the comparator, and *p*-value for difference between the case/control group mean values was obtained by Student's t-test;*p<0.05, **p <0.01,***p <0.001. (B) Real-time polymerase chain reaction was conducted on the same samples as in (A). Relative mRNA expression levels were calculated using the ΔΔCt method algorithm; *p<0.05,**p<0.01.

Table 4. Performance characteristics of the model-based classifiers in the test set (n=65, FEV_1/FVC 0.60–0.75).

	Classifier performance in test set					
Model classifier	Number of variables	Discordant classification (%)	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)
Full random forest	1014	25	66	92	93	61
L_1-penalized logistic regression	9	22	71	92	94	64

FEV_1, forced expiratory volume in 1 s; FVC, forced vital capacity.

One of the major constraints of COPD biomarker discovery has been the accessibility of suitable samples. In the past, sputum, bronchoalveolar lavage fluid, exhaled breath condensate and bronchial biopsy tissue have been used (Sin & Man 2008). However, the sampling methodologies for such specimens are limited by their invasiveness and poor reproducibility. As COPD is accompanied by systemic changes, as well as increased serum levels of certain proteins (e.g. C-reactive protein (CRP), interleukin 6 (IL-6), IL-8, leukotriene B_4 (LTB4)

and tumour necrosis factor (TNF)-α), the use of PBLs as a surrogate biosample is an attractive alternative because they can be easily collected in large quantities at multiple time points using a relatively non-invasive procedure (Agusti et al. 2003, Celli 2006, Noguera et al. 1998, Rahman et al. 1996, Rahman&Biswas 2004, Schols et al. 1996, Vernooy et al. 2002). As noted earlier, PBL gene expression profiles have been successfully used to identify the presence or risk of other diseases having a prominent systemic component.

Due to the role of PBLs in inflammation, the gene expression differences between subjects with and without COPD in this population of cells may reflect the degree of systemic or pulmonary inflammation. Although the differences in gene expression may be due to other systemic inflammatory conditions, other aspects of COPD, and/or treatment, lung inflammation is known to correlate with the severity of the disease, as measured by the degree of airflow limitation (Hogg et al. 2004). Our gene expression-based classifier was derived from a training set of COPD patients with the most extreme airflow limitation, who probably also had the greatest degree of inflammation, while the test group with lesser airflow limitation, therefore, might be predicted to have less inflammation. This may also partially account for the poorer predictive ability between spirometric cases and controls in the test set compared with the training set.

Gene expression differences between COPD cases and controls in PBLs may provide mechanistic insight into COPD. Differences in activation state, surface marker expression, chemotaxis and extracellular proteolysis from specific populations of PBLs in subjects with COPD have been reported (Agusti 2005). In the present study, biological processes identified as over-represented in the set of COPD predictor genes included regulation of apoptosis, regulation of cell growth, macromolecule (protein and RNA) transport, post-translational protein modification, cellular defence response, inflammatory response and RNA processing. Major pathways identified included apoptosis, p38/MAPK signalling, focal adhesion and leukocyte transendothelial migration. Changes in these biological processes and pathways may reflect the changes in activation, differentiation and cellular composition of the samples analysed. The identification of leukocyte transendothelial migration seems to be an important change in this cell population due to the fact that COPD is characterized by leukocyte infiltration in the lung parenchyma (Panina et al. 2006). It is possible that differences in expression of these genes may result in a predisposition of leukocyte subpopulations to infiltrate the lung tissue, and perhaps other tissues. This observation is supported by previously reported changes in chemotaxis and extracellular proteolysis in neutrophils isolated from the blood of subjects with COPD (Burnett et al. 1987).

The subset of nine genes that we identified using L_1-penalized logistic regression modelling had similar predictive performance as the full set of candidate predictors identified by the random forest model. It included five upregulated genes (*CCR2, IL6R, PPP2CB, RASSF2* and *WTAP*) and four downregulated genes (*DNTTIP2, GDAP1, LIPE, RPL14*) in COPD patients compared with controls. *IL6R* and *CCR2* have been previously reported to play possible roles in COPD development and progression (Owen 2001, Wilk et al. 2007). However, there have been no prior reports of an association with COPD for *DNTTIP2, GDAP1, LIPE, PPP2CB, RASSF2, RPL14* and *WTAP*.

The *IL6R* gene codes for the IL-6 receptor, which is only reported to be expressed in subpopulations of leukocytes (monocytes, neutrophils and T and B lymphocytes) and hepatocytes (Chalaris et al. 2007, Hamid et al. 2004, Jones et al. 2001). Many cell types do not express IL6R and are not directly responsive to IL6 (Chalaris et al. 2007, Jones et al. 2001). However, these cell types can be stimulated by IL-6 bound to a soluble form of the IL-6 receptor in a process called trans-signalling (Chalaris et al. 2007, Jones et al. 2001). IL-6R shedding and subsequent release of the soluble form of the receptor results from cleavage of the membrane-bound receptor during apoptosis, a biological process and pathway identified in our gene expression signatures. This process is dependent on the metalloproteinases, ADAM17 and to a lesser extent ADAM10 (Chalaris et al. 2007, Matthews et al. 2003). ADAM17 was also found to be upregulated in our microarray and was identified as one of the candidate predictor genes. Reported inducers of IL-6R shedding include phorbol myristate acetate, cholesterol depletion, CRP, bacterial toxins, Fas stimulation and ultraviolet light (Chalaris et al. 2007, Jones et al. 1999, Matthews et al. 2003, Mullberg et al. 1992). Signalling through IL-6R has also been shown to play a role in both inflammation and apoptosis (Finotto et al. 2007). Furthermore, genome-wide association analyses have identified *IL6R* as a likely candidate gene for association with lung function (Wilk et al. 2007).

CCR2, which encodes the receptor for monocyte chemoattractant protein 1 and 3 (MCP1 and MCP3), is involved in inflammatory processes related to rheumatoid arthritis, alveolitis and tumour infiltration (Owen 2001). Higher levels of MCP1 mRNA and protein were detected in the bronchiolar epithelium in subjects with COPD, and increased levels of *CCR2* have been detected in macrophages, mast cells and epithelial cells of COPD subjects, suggesting that MCP1 and CCR2 may be involved in the recruitment of macrophages into the airway epithelium (de Boer et al. 2000, Owen 2001). This increased expression of *CCR2* also correlated with increased levels of mast cells and macrophages in the lungs of COPD subjects (de Boer et al. 2000). In addition, it has been demonstrated

that activated neutrophils migrate in response to MCP1 (Johnston et al. 1999). These findings support mechanistic roles of *IL6R* and *CCR2* in systemic and lung inflammation in COPD.

The seven other genes in our nine-gene profile have varied biological functions. *PPP2CB* encodes the beta-isoform of the catalytic subunit of protein phosphatase 2A (PP2A) (Cohen 1989, Hemmings et al. 1988). PP2A has been shown to regulate apoptosis in neutrophils by dephosphorylating both p38/MAPK and its substrate caspase 3, suggesting that PP2A plays a role in the induction of apoptosis and the resolution of inflammation (Alvarado-Kristensson&Andersson 2005). RASSF2 promotes apoptosis and cell cycle arrest (Vos et al. 2003). WTAP is involved in the expression of genes related to cell division cycle and the G2/M checkpoint (Horiuchi et al. 2006). The DNTT-interacting protein 2 (DNTTIP2), also known as estrogen receptor-binding protein, can bind the estrogen receptor-α and enhance its transcriptional activity in an estrogen-dependent manner (Bu et al. 2004). GDAP1, or ganglioside-induced differentiation-associated protein 1, is found localized in the mitochondrial outer membrane and regulates the mitochondrial network. Overexpression of *GDAP1* induces fragmentation of mitochondria without inducing apoptosis, affecting overall mitochondrial activity, or interfering with mitochondrial fusion (Cuesta et al. 2002, Niemann et al. 2005). LIPE, also known as HSL (hormone-sensitive lipase), plays a role in the mobilization of free fatty acids from adipose tissue by controlling the rate of lipolysis of the stored triglycerides (Holm et al. 1988). Finally, *RPL14* is a gene coding for a protein of the large ribosomal subunit (Robledo et al. 2008). The role of these genes in COPD is unclear, but some appear to be linked to the cellular processes and pathways, such as cell cycle regulation and apoptosis, associated with our full list of genes.

Previous gene expression microarray analysis studies investigating COPD have used lung tissue. Bhattacharya et al. (2009) analysed gene expression in lung tissue from COPD patients and showed an over-representation of transcription and nucleic acid binding protein categories. Ning et al. (2004) demonstrated that the majority of gene expression changes in lung tissue from COPD patients occurred in lipoprotein biosynthesis, NADH dehydrogenase activity, 26S proteasome, sodium ion transporter activity, mitochondrion, secretion, hormone activity, receptor binding, signal transduction, development and growth factor activity. Pierou et al. (2007) identified oxidative phosphorylation, oxidant response, ATP synthesis, proteasome, thioredoxins and glycolysis pathways in their analysis of COPD lungs with healthy smoker lungs. Although many of these categories were not identified in the current PBL study, it appears that cell growth/growth factory activity, proteasome/proteolysis and signal transduction were affected in COPD in both PBL and lung tissue samples.

Certain limitations of our study merit comment. Although many factors could influence the gene expression profiles detected by microarray in this study, the most obvious of these is the cellular composition of the samples. Although the average total PBL counts were similar between the groups with and without COPD, the mean lymphocyte and granulocyte counts as percentages of the total were significantly different (Table 1). These parameters were included in the random forest analysis but not retained in the final model, suggesting that the gene expression differences are more predictive of COPD status than lymphocyte and granulocyte percentages. However, this does not rule out the influence of these quantitative differences in cellular composition in the observed gene expression profile. Another factor that can influence gene expression is medication use, such as corticosteroids in COPD (Barnes 2006, MacRedmond et al. 2007). Due to the random forest algorithm's inability to handle missing values among the predictor variables, the medication history of the subjects was not included in the analysis as a substantial proportion of subjects had missing values. Although it is unclear how corticosteroids might affect gene expression in PBLs, it is known that the small airway inflammation responsible for airflow obstruction in COPD is poorly sensitive to the anti-inflammatory effects of corticosteroids (Barnes 2006, Hogg et al. 2004). Recent evidence has attributed this to oxidative and nitrative stress-induced reduction in histone deacetylase expression in inflammatory cells, thus preventing activated corticosteroid receptors from reversing the acetylation of activated inflammatory genes and turning off their transcription (Barnes 2006). Although this would suggest that corticosteroids do not have a large impact on the inflammatory response in COPD, the contribution of medications to the differential gene expression detected in the present study cannot be ruled out. In addition, we included in the analysis ten subjects with arguably indeterminate spirometric COPD case/control status based on their combination of FEV_1/FVC and FEV_1 %predicted, categorizing them spirometrically as controls by the GOLD-defined FEV_1/FVC cut-off value. As noted earlier in the Discussion, one of these spirometric controls in the test set was discordantly classified as having COPD by his gene expression profile in both the full and reduced models.

Cigarette smoke exposure can also influence gene expression. Of the 1013 predictor genes identified in our analysis, differential expression of *ATF4, MCL1, MAPK14, SERPINA1* and *SOD2* was also identified in a study by van Leeuwen et al. (2007) as strongly correlating with serum cotinine levels, a biomarker of recent exposure to tobacco smoke. Two additional genes in our list, *CCR2* and *EPB41*, were observed by Lampe et al. (2004) as part

of a cigarette smoke exposure molecular signature. Both the van Leeuwen and Lampe studies used PBLs isolated from current smokers and non-smokers, suggesting that the differential expression of some of the genes identified in our study may be related to tobacco smoke exposure. In a study of bronchial epithelial cells from never, current and former smokers, Beane et al. (2007) found 175 genes differentially expressed between never and current smokers, with irreversible changes in expression for 28 genes, slowly reversible for six genes and rapidly reversible for 139 genes. This indicates that duration and possibly intensity of cigarette smoking, and length of time since quitting, may be important confounding variables to gene expression analysis. The only smoking exposure variable identified as a candidate predictor in our study ('years of daily smoking') appears to support this possibility. However, importantly, no other demographic variables, such as age, nor smoking exposure variables, including pack years, current smoking status and current/recent (3 months) smoking intensity (number of cigarettes per day), were significant predictor variables in the random forest model. Although age was not a significant predictor in our analyses after simultaneous adjustment for years of daily smoking, age should always be carefully considered in this type of research, particularly in situations in which highly correlated, age-related variables are not available.

In conclusion, we have identified in a training set, and confirmed in a test set, differential gene expression for 1013 genes occurring in peripheral blood leukocytes that discriminated between current and former cigarette smokers with or without spirometrically defined COPD. Gene ontology and pathway analyses indicated that these genes are involved in regulation of apoptosis, regulation of cell growth, macromolecule (protein and RNA) transport, RNA processing, post-translational protein modification, cellular defence response and inflammatory response. This list of 1013 genes was subsequently reduced to a nine-gene subset with similar performance in differentiating current and former cigarette smokers with or without COPD. To our knowledge, this is the first study to use microarray analysis of PBLs to identify gene expression differences associated with COPD. PBL samples are easy to obtain and their gene expression profiles could complement and improve current clinical diagnostic and prognostic approaches for COPD. The gene expression profiles identified here represent potential biomarkers for COPD in current and former cigarette smokers. However, due to the limited sample size and other limitations of this study, further investigations are needed to determine the performance of these gene expression profiles in an independent group of subjects and to determine whether the differential gene expression is predictive or reflective of cigarette smoking-related COPD.

Acknowledgements

The authors gratefully acknowledge the contributions to this study and manuscript by Michael S. Paul, PhD and Alex Lindell from LineaGen, Inc., Salt Lake City, Utah and George J. Patskan, PhD and Willie J McKinney, PhD from Altria Client Services. The authors also acknowledge the comments of reviewers (Priyadarshi Basu, PhD and Robert McKallip, PhD), the editorial assistance of Eileen Y. Ivasauskas of Accuwrit Inc., data management by Zaigang Liu and microarray processing by Yankai Jia, PhD

Declaration of interest

Financial support was provided by Philip Morris USA Inc. and LineaGen, Inc. J.S.E., M.J.S., E.L.M., B.K.Z. and A.R.J. were employed through Altria Client Services at the time of manuscript preparation. The other authors declare no other potential conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Achiron A, Gurevich M. (2006). Peripheral blood gene expression signature mirrors central nervous system disease: the model of multiple sclerosis. *Autoimmun Rev* 5:517–22.

Agusti A, Soriano JB. (2008). Clinical review: COPD as a systemic disease. *COPD* 5:133–8.

Agusti AG. (2005). Systemic effects of chronic obstructive pulmonary disease. *Proc Am ThoracSoc* 2:367–70.

Agusti AG, Noguera A, Sauleda J, Sala E, Pons J, Busquets X. (2003). Systemic effects of chronic obstructive pulmonary disease. *EurRespir J* 21:347–60.

Alvarado-Kristensson M, Andersson T. (2005). Protein phosphatase 2A regulates apoptosis in neutrophils by dephosphorylating both p38 MAPK and its substrate caspase 3. *J BiolChem* 280:6238–44.

Archer KJ, Kimes RV. (2008). Empirical characterization of random forest variable importance measures. *Comput Stat Data Anal* 52:2249–60.

Barnes PJ. (2003). New concepts in chronic obstructive pulmonary disease. *Annu Rev Med* 54:113–29.

Barnes PJ. (2006). Reduced histone deacetylase in COPD – clinical implications. *Chest* 129:151–5.

Barnes PJ, Shapiro SD, Pauwels RA. (2003). Chronic obstructive pulmonary disease: molecular and cellular mechanisms. *EurRespir J* 22:672–88.

Beane J, Sebastiani P, Liu G, Brody JS, Lenburg ME, Spira A. (2007). Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol* 8:R201.

Bhattacharya S, Srisuma S, Demeo DL, Shapiro SD, Bueno R, Silverman EK, Reilly JJ, Mariani TJ. (2009). Molecular biomarkers for quantitative and discrete COPD phenotypes. *Am J Respir Cell MolBiol* 40:359–67

Bovin LF, Rieneck K, Workman C, Nielsen H, Sorensen SF, Skjodt H, Florescu A, Brunak S, Bendtzen K. (2004). Blood cell gene expression profiling in rheumatoid arthritis. Discriminative genes and effect of rheumatoid factor. *ImmunolLett* 93:217–26.

Breiman L. (2001). Random forests. *Mach Learn* 45:5–32.

Bu H, Kashireddy P, Chang J, Zhu YT, Zhang Z, Zheng W, Rao SM, Zhu YJ. (2004). ERBP, a novel estrogen receptor binding protein

enhancing the activity of estrogen receptor. *BiochemBiophys Res Commun* 317:54–9.

Bull TM, Coldren CD, Moore M, Sotto-Santiago SM, Pham DV, Nana-Sinkam SP, Voelkel NF, Geraci MW. (2004). Gene microarray analysis of peripheral blood cells in pulmonary arterial hypertension. *Am J RespirCrit Care Med* 170:911–19.

Burge PS, Calverley PM, Jones PW, Spencer S, Anderson JA, Maslen TK. (2000). Randomised, double blind, placebo controlled study of fluticasone propionate in patients with moderate to severe chronic obstructive pulmonary disease: the ISOLDE trial. *Br Med J* 320:1297–303.

Burnett D, Chamba A, Hill SL, Stockley RA. (1987). Neutrophils from subjects with chronic obstructive lung disease show enhanced chemotaxis and extracellular proteolysis. *Lancet* 2: 1043–6.

Casanova C, Cote C, de Torres JP, Aguirre-Jaime A, Marin JM, Pinto-Plata V, Celli BR. (2005). Inspiratory-to-total lung capacity ratio predicts mortality in patients with chronic obstructive pulmonary disease. *Am J RespirCrit Care Med* 171:591–7.

Celli BR. (2006). Roger S. Mitchell lecture. Chronic obstructive pulmonary disease phenotypes and their clinical relevance. *Proc Am ThoracSoc* 3:461–5.

Celli BR, Cote CG, Marin JM, Casanova C, Montes dO, Mendez RA, Pinto P, V, Cabral HJ. (2004). The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *New Engl J Med* 350:1005–12.

Centers for Medicare & Medicaid Services.(1992). Medicare, Medicaid and CLIA programs; regulations implementing the Clinical Laboratory Improvement Amendments of 1988 (CLIA). *Fed Regist* 40:7002–186.

Chadee DN, Yuasa T, Kyriakis JM. (2002). Direct activation of mitogen-activated protein kinase kinasekinase MEKK1 by the Ste20p homologue GCK and the adapter protein TRAF2. *Mol Cell Biol* 22:737–49.

Chalaris A, Rabe B, Paliga K, Lange H, Laskay T, Fielding CA, Jones SA, Rose-John S, Scheller J. (2007). Apoptosis is a natural stimulus of IL6R shedding and contributes to the proinflammatory trans-signaling function of neutrophils. *Blood* 110:1748–55.

Chung KF. (2001). Cytokines in chronic obstructive pulmonary disease. *EurRespir J* Suppl 34:50s–59s.

Chung KF. (2005). Inflammatory mediators in chronic obstructive pulmonary disease. *Curr Drug Targets Inflamm Allergy* 4:619–25.

Cohen P. (1989). The structure and regulation of protein phosphatases. *Annu Rev Biochem* 58:453–508.

Cuesta A, Pedrola L, Sevilla T, Garcia-Planells J, Chumillas MJ, Mayordomo F, LeGuern E, Marin I, Vilchez JJ, Palau F. (2002). The gene encoding ganglioside-induced differentiation-associated protein 1 is mutated in axonal Charcot-Marie-Tooth type 4A disease. *Nat Genet* 30:22–5.

de Boer WI, Sont JK, van Schadewijk A, Stolk J, van Krieken JH, Hiemstra PS. (2000). Monocyte chemoattractant protein 1, interleukin 8, and chronic airways inflammation in COPD. *J Pathol* 190:619–26.

Dennis G, Sherman B, Hosack D, Yang J, Gao W, Lane HC, Lempicki R. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4:3.

Dunning MJ, Barbosa-Morais NL, Lynch AG, Tavare S, Ritchie ME.(2008). Statistical issues in the analysis of Illumina data. *BMC Bioinformatics* 9:85.

Fabbri LM, Rabe KF. (2007). From COPD to chronic systemic inflammatory syndrome?*Lancet* 370:797–9.

Finotto S, Eigenbrod T, Karwot R, Boross I, Doganci A, Ito H, Nishimoto N, Yoshizaki K, Kishimoto T, Rose-John S, Galle PR, Neurath MF. (2007). Local blockade of IL-6R signaling induces lung CD4+ T cell apoptosis in a murine model of asthma via regulatory T cells. *IntImmunol* 19:685–93.

Golpon HA, Coldren CD, Zamora MR, Cosgrove GP, Moore MD, Tuder RM, Geraci MW, Voelkel NF.(2004). Emphysema lung tissue gene expression profiling. *Am J Respir Cell MolBiol* 31:595–600.

Hamid YH, Urhammer SA, Jensen DP, Glumer C, Borch-Johnsen K, Jorgensen T, Hansen T, Pedersen O. (2004). Variation in the interleukin-6 receptor gene associates with type 2 diabetes in Danish whites. *Diabetes* 53:3342–5.

Hansel NN, Hilmer SC, Cope LM, Guo J, Irizarry RA, Diette GB. (2005). Oligonucleotide-microarray analysis of peripheral-blood lymphocytes in severe asthma. *J Lab Clin Med* 145:263–74.

Hemmings BA, Wernet W, Mayer R, Maurer F, Hofsteenge J, Stone SR. (1988). The nucleotide sequence of the cDNA encoding the human lung protein phosphatase 2A beta catalytic subunit. *Nucleic Acids Res* 16:11366.

Hogg JC, Chu F, Utokaparch S, Woods R, Elliott WM, Buzatu L, Cherniack RM, Rogers RM, Sciurba FC, Coxson HO, Pare PD. (2004). The nature of small-airway obstruction in chronic obstructive pulmonary disease. *New Engl J Med* 350:2645–53.

Holm C, Kirchgessner TG, Svenson KL, Lusis AJ, Belfrage P, Schotz MC. (1988). Nucleotide sequence of rat adipose hormone sensitive lipase cDNA. *Nucleic Acids Res* 16:9879.

Horiuchi K, Umetani M, Minami T, Okayama H, Takada S, Yamamoto M, Aburatani H, Reid PC, Housman DE, Hamakubo T, Kodama T. (2006). Wilms' tumor 1-associating protein regulates G2/M transition through stabilization of cyclin A2 mRNA. *ProcNatlAcadSci USA* 103:17278–83.

Johnston B, Burns AR, Suematsu M, Issekutz TB, Woodman RC, Kubes P. (1999). Chronic inflammation upregulates chemokine receptors and induces neutrophil migration to monocyte chemoattractant protein-1. *J Clin Invest* 103:1269–76.

Jones SA, Horiuchi S, Topley N, Yamamoto N, Fuller GM. (2001). The soluble interleukin 6 receptor: mechanisms of production and implications in disease. *FASEB J* 15:43–58.

Jones SA, Novick D, Horiuchi S, Yamamoto N, Szalai AJ, Fuller GM. (1999). C-reactive protein: a physiological activator of interleukin 6 receptor shedding. *J Exp Med* 189:599–604.

Lampe JW, Stepaniants SB, Mao M, Radich JP, Dai H, Linsley PS, Friend SH, Potter JD. (2004). Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. *Cancer Epidemiol Biomarkers Prev* 13:445–53.

Liaw A, Wiener M. (2002).Classification and regression by randomForest. *R News* 2:18–22.

Liew CC, Ma J, Tang HC, Zheng R, Dempsey AA. (2006). The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. *J Lab Clin Med* 147:126–32.

MacRedmond RE, Greene CM, Dorscheid DR, McElvaney NG, O'Neill SJ. (2007). Epithelial expression of TLR4 is modulated in COPD and by steroids, salmeterol and cigarette smoke. *Resp Res* 8:84.

Maltais F, LeBlanc P, Jobin J, Casaburi R. (2000). Peripheral muscle dysfunction in chronic obstructive pulmonary disease. *Clin Chest Med* 21:665–77.

Mannino DM, Braman S. (2007). The epidemiology and economics of chronic obstructive pulmonary disease. *Proc Am ThoracSoc* 4:502–6.

Marin JM, Carrizo SJ, Casanova C, Martinez-Camblor P, Soriano JB, Agusti AG, Celli BR. (2009). Prediction of risk of COPD exacerbations by the BODE index. *Respir Med* 103:373–8.

Marshall KW, Zhang H, Yager TD, Nossova N, Dempsey A, Zheng R, Han M, Tang H, Chao S, Liew CC. (2005). Blood-based biomarkers for detecting mild osteoarthritis in the human knee. *Osteoarthritis Cartilage* 13:861–71.

Matthews V, Schuster B, Schutze S, Bussmeyer I, Ludwig A, Hundhausen C, Sadowski T, Saftig P, Hartmann D, Kallen KJ, Rose-John S. (2003). Cellular cholesterol depletion triggers shedding of the human interleukin-6 receptor by ADAM10 and ADAM17 (TACE). *J BiolChem* 278:38829–39.

Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, Crapo R, Enright P, van der Grinten CPM, Gustafsson P, Jensen R, Johnson DC, MacIntyre N, McKay R, Navajas O, Pedersen F, Pellegrino R, Viegi G, Wanger J. (2005). ATS/ERS standardization of lung function testing: standardization of spirometry. *EurRespir J* 26:319–38.

Mohr S, Liew CC. (2007). The peripheral-blood transcriptome: new insights into disease and risk assessment. *Trends Mol Med* 13:422–32.

Mossman BT, Lounsbury KM, Reddy SP. (2006). Oxidants and signaling by mitogen-activated protein kinases in lung epithelium. *Am J Respir Cell MolBiol* 34:666–9.

Mullberg J, Schooltink H, Stoyan T, Heinrich PC, Rose-John S. (1992). Protein kinase C activity is rate limiting for shedding of the interleukin-6 receptor. *BiochemBiophys Res Commun* 189:794–800.

Niemann A, Ruegg M, La P, V, Schenone A, Suter U. (2005). Ganglioside-induced differentiation associated protein 1 is a regulator of the mitochondrial network: new implications for Charcot–Marie–Tooth disease. *J Cell Biol* 170:1067–78.

Ning W, Li CJ, Kaminski N, Feghali-Bostwick CA, Alber SM, Di YP, Otterbein SL, Song R, Hayashi S, Zhou Z, Pinsky DJ, Watkins SC, Pilewski JM, Sciurba FC, Peters DG, Hogg JC, Choi AM. (2004). Comprehensive gene expression profiles reveal pathways related to the pathogenesis of chronic obstructive pulmonary disease. *ProcNatlAcadSci USA* 101:14895–900.

Nishimura K, Izumi T, Tsukino M, Oga T. (2002). Dyspnea is a better predictor of 5-year survival than airway obstruction in patients with COPD. *Chest* 121:1434–40.

Noguera A, Busquets X, Sauleda J, Villaverde JM, MacNee W, Agusti AG. (1998). Expression of adhesion molecules and G proteins in circulating neutrophils in chronic obstructive pulmonary disease. *Am J RespirCrit Care Med* 158:1664–8.

Omori H, Nakashima R, Otsuka N, Mishima Y, Tomiguchi S, Narimatsu A, Nonami Y, Mihara S, Koyama W, Marubayashi T, Morimoto Y. (2006). Emphysema detected by lung cancer screening with low-dose spiral CT: prevalence and correlation with smoking habits and pulmonary function in Japanese male subjects. *Respirology* 11:205–10.

Owen C. (2001). Chemokine receptors in airway disease: which receptors to target? *PulmPharmacolTher* 14:193–202.

Panina P, Mariani M, D'Ambrosio D. (2006). Chemokine receptors in chronic obstructive pulmonary disease (COPD). *Curr Drug Targets* 7:669–74.

Park MY, Hastie T. (2007). L1 regularization path algorithm for generalized linear models. *J R Stat Soc B* 69:659–77.

Pierrou S, Broberg P, O'Donnell RA, Pawlowski K, Virtala R, Lindqvist E, Richter A, Wilson SJ, Angco G, Moller S, Bergstrand H, Koopmann W, Wieslander E, Stromstedt PE, Holgate ST, Davies DE, Lund J, Djukanovic R. (2007). Expression of genes involved in oxidative stress responses in airway epithelial cells of smokers with chronic obstructive pulmonary disease. *Am J RespirCrit Care Med* 175:577–86.

Qing X, Putterman C. (2004). Gene expression profiling in the study of the pathogenesis of systemic lupus erythematosus. *Autoimmun Rev* 3:505–9.

R Development Core Team. (2007). R: a language and environment for statistical computing. Available at: http://www R-project org.

Rabe KF, Hurd S, Anzueto A, Barnes PJ, Buist SA, Calverley P, Fukuchi Y, Jenkins C, Rodriguez-Roisin R, van Weel C, Zielinski J. (2007). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J RespirCrit Care Med* 176:532–55.

Rahman I. (2005). The role of oxidative stress in the pathogenesis of COPD: implications for therapy. *Treat Respir Med* 4:175–200.

Rahman I, Biswas SK. (2004). Non-invasive biomarkers of oxidative stress: reproducibility and methodological issues. *Redox Rep* 9:125–43.

Rahman I, Morrison D, Donaldson K, MacNee W. (1996). Systemic oxidative stress in asthma, COPD, and smokers. *Am J RespirCrit Care Med* 154:1055–60.

Robledo S, Idol RA, Crimmins DL, Ladenson JH, Mason PJ, Bessler M. (2008). The role of human ribosomal proteins in the maturation of rRNA and ribosome production. *RNA* 14:1918–29.

Schols AM, Buurman WA, Staal van den Brekel AJ, Dentener MA, Wouters EF. (1996). Evidence for a relation between metabolic derangements and increased levels of inflammatory mediators in a subgroup of patients with chronic obstructive pulmonary disease. *Thorax* 51:819–24.

Schols AM, Slangen J, Volovics L, Wouters EF. (1998). Weight loss is a reversible factor in the prognosis of chronic obstructive pulmonary disease. *Am J RespirCrit Care Med* 157:1791–7.

Sin DD, Man SF. (2008). Biomarkers in COPD: are we there yet? *Chest* 133:1296–8.

Spira A, Beane J, Pinto-Plata V, Kadar A, Liu G, Shah V, Celli B, Brody JS. (2004). Gene expression profiling of human lung tissue from smokers with severe emphysema. *Am J Respir Cell MolBiol* 31:601–10.

Tibshirani R. (1996). Regression shrinkage and selection via the Lasso. *J R Stat Soc B* 58:267–88.

van Leeuwen DM, van Agen E, Gottschalk RW, Vlietinck R, Gielen M, van Herwijnen MH, Maas LM, Kleinjans JC, van Delft JH. (2007). Cigarette smoke-induced differential gene expression in blood cells from monozygotic twin pairs. *Carcinogenesis* 28:691–7.

Vernooy JH, Kucukaycan M, Jacobs JA, Chavannes NH, Buurman WA, Dentener MA, Wouters EF. (2002). Local and systemic inflammation in patients with chronic obstructive pulmonary disease: soluble tumor necrosis factor receptors are increased in sputum. *Am J RespirCrit Care Med* 166:1218–24.

Vos MD, Ellis CA, Elam C, Ulku AS, Taylor BJ, Clark GJ. (2003). RASSF2 is a novel K-Ras-specific effector and potential tumor suppressor. *J BiolChem* 278:28045–51.

Wang IM, Stepaniants S, Boie Y, Mortimer JR, Kennedy B, Elliott M, Hayashi S, Loy L, Coulter S, Cervino S, Harris J, Thornton M, Raubertas R, Roberts C, Hogg JC, Crackower M, O'Neill G, Pare PD. (2008). Gene expression profiling in patients with chronic obstructive pulmonary disease and lung cancer. *Am J RespirCrit Care Med* 177:402–11.

Wilk JB, Walter RE, Laramie JM, Gottlieb DJ, O'Connor GT. (2007). Framingham Heart Study genome-wide association: results for pulmonary function measures. *BMC Med Genet* 8 (Suppl. 1):*S8.*

Witowsky JA, Johnson GL. (2003). Ubiquitylation of MEKK1 inhibits its phosphorylation of MKK1 and MKK4 and activation of the ERK1/2 and JNK pathways. *J BiolChem* 278:1403–6.

Zhang G, Nebert DW, Chakraborty R, Jin L. (2006). Statistical power of association using the extreme discordant phenotype design. *Pharmacogenet Genomics* 16:401–13.