Nicotine & Tobacco Research, 2016, 1903–1909
doi:10.1093/ntr/ntw129
Original investigation
Advance Access publication May 6, 2016

Original investigation

The Effect of Different Case Definitions of Current Smoking on the Discovery of Smoking-Related Blood Gene Expression Signatures in Chronic Obstructive Pulmonary Disease

Ma'en Obeidat PhD[1], Xiaoting Ding BSc[1], Nick Fishbane MSc[1], Zsuzsanna Hollander PhD[1,2], Raymond T. Ng PhD[2,3], Bruce McManus MD[1,2], Scott J. Tebbutt PhD[1,2], Bruce E. Miller PhD[4], Stephen Rennard MD[5], Peter D. Paré MD[1,6], Don D. Sin MD[1,6]

[1]University of British Columbia Centre for Heart Lung Innovation, St Paul's Hospital, Vancouver, BC, Canada; [2]Prevention of Organ Failure (PROOF) Centre of Excellence, Vancouver, BC, Canada; [3]Department of Computer Science, University of British Columbia Centre, Vancouver, BC, Canada; [4]GlaxoSmithKline, King of Prussia, PA; [5]Division of Pulmonary and Critical Care Medicine, University of Nebraska Medical Center, Omaha, NE; [6]Respiratory Division, Department of Medicine, University of British Columbia, Vancouver, BC, Canada

Corresponding Author: Don D. Sin, MD, UBC Centre for Heart Lung Innovation, St Paul's Hospital, 1081 Burrard Street, Vancouver, BC V6Z 1Y6, Canada. Telephone: 604-806-8346 ext 68395; Fax: 604-806-8351; E-mail: don.sin@hli.ubc.ca

Abstract

Introduction: Smoking is the number one modifiable environmental risk factor for chronic obstructive pulmonary disease (COPD). Clinical, epidemiological and increasingly "omics" studies assess or adjust for current smoking status using only self-report, which may be inaccurate. Objective measures such as exhaled carbon monoxide (eCO) may also be problematic owing to limitations in the measurements and the relatively short half life of the molecule. In this study, we determined the impact of different case definitions of current cigarette smoking on gene expression in peripheral blood of patients with COPD.

Methods: Peripheral blood gene expression from 573 former- and current-smokers with COPD in the ECLIPSE study was used to find genes whose expression was associated with smoking status. Current smoking was defined using self-report, eCO concentrations, or both. Linear regression was used to determine the association of current smoking status with gene expression adjusting for age, sex and propensity score. Pathway enrichment analyses were performed on genes with $P < .001$.

Result: Using self-report or eCO, only two genes were differentially expressed between current and ex-smokers, with no enrichment in biological processes. When current smoking was defined using both eCO and self-report, four genes were differentially expressed (LRRN3, PID1, FUCA1, GPR15) with enrichment in 40 biological pathways related to metabolic processes, response to hypoxia and hormonal stimulus. Additionally, the combined definition provided better distributions of test statistics for differential gene expression.

Conclusion: A combined phenotype of eCO and self report allows for better discovery of genes and pathways related to current smoking.

Implications: Studies relying only on self report of smoking status to assess or adjust for the impact of smoking may not fully capture its effect and will lead to residual confounding of results.

Introduction

The smoking epidemic remains one of the biggest public health threats in modern history.[1] With the current estimates of 50% of young men and 10% of young women becoming smokers and a smaller percentage quitting in many developing countries, tobacco-attributable deaths will rise from about 6 million a year currently to more than 10 million globally by 2030.[2–4] Smoking is the principal modifiable environmental risk factor for chronic obstructive pulmonary disease (COPD), ischemic heart disease, and lung cancer.[5] COPD, for instance, affects 300 million people and is the third leading cause of death worldwide.[6] In Canada, COPD is the number one cause of hospital admissions, accounting for 80 000 admissions per year.[7]

Self-reported smoking status is widely used in epidemiological, interventional, genetic and genomic studies as a covariate to "control" for the harmful effects of tobacco exposure. However, self-report is subject to reporting bias and has been shown to underestimate smoking prevalence and intensity.[8,9] Objective biochemical measures for smoking status include exhaled carbon monoxide (eCO) and cotinine levels in serum, urine or saliva. Measuring eCO is attractive because it is relatively inexpensive, noninvasive and the measurements are well-standardized and can accurately identify recent smokers.[10,11]

Many blood gene expression profiling studies have been published for COPD.[12–15] Smoking is an important driver and confounder of these studies. The studies typically compare the genome-wide expression profiles between COPD cases and smoking controls, and include self-reported smoking status and/or pack-years as covariates in the analysis. Using this approach, pathway analyses of differentially expressed genes have revealed enrichment in diverse processes including apoptosis, cell growth, cellular defense response and inflammatory response,[12] as well as sphingolipid (ceramide) metabolism,[14] and cancer.[15] However, the differences in peripheral blood gene expression profiles between current and former smokers in individuals with COPD are not clearly understood. The objectives of the current study are to (1) unravel the molecular signature in blood between current and former smokers with COPD and (2) examine the impact of differential phenotyping of smoking status using subjective and objective measures on gene expression.

Materials and Methods

Study Subjects

The study population was a subset of the parent ECLIPSE (Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints) study.[16] ECLIPSE was a 3-year noninterventional, multicentre, longitudinal prospective study. The ECLIPSE study was approved by the relevant ethics review boards at each of the participating centers. ECLIPSE included 2164 COPD patients aged 40–75 years (smoking history ≥10 pack-years with a post-bronchodilator FEV_1/FVC < 0.70 and FEV_1 < 80% predicted) and 337 smokers and 245 nonsmokers who acted as control subjects (FEV_1/FVC > 0.70 and FEV_1 > 90% predicted). The inclusion criteria included individuals with >10 pack-years of smoking. Former smokers were defined as those who had quit smoking at least 6 months prior to study entry. Blood was collected in PAXgene RNA tubes and frozen at −80°C. The gene expression sub-study of ECLIPSE was designed to determine bio-signatures of exacerbation in peripheral blood of patients with COPD.[17] Two groups of individuals were chosen from the parent ECLIPSE study for this purpose: patients who were frequent exacerbators defined as having two or more exacerbations per year and patients who did not experience any exacerbation during follow-up. The two groups were matched with respect to age, sex, and smoking status (current and former smokers). A total of 531 former and current- smokers were selected and had both gene expression and phenotypic data available. Study participants provided written informed consent, and participants' information was de-identified. ECLIPSE study was funded by GlaxoSmithKline, under http://ClinicalTrials.gov identifier NCT00292552 and GSK No. SCO104960. This gene expression sub-study was funded by Genome British Columbia and was approved by the Providence Health Care Research Ethics Board (REB) of the University of British Columbia (UBC) (H11-00786).

Microarray Data Processing

Total RNA was extracted using the PAXgene Blood miRNA kit from PreAnalytix (Cat. #763134). RNA was hybridized to the Affymetrix Human Gene 1.1 ST array. Affymetrix GeneTitan MC Scanner (Affymetrix Inc) was used to scan the array plates. The oligo[18] and RMA Express[19] packages from Bioconductor were used to perform quality control on the microarray data. Background correction, normalization, and summarization of the data and filtering out non-informative probe sets was undertaken using the Factor Analysis for Robust Microarray Summarization (FARMS Bioconductor package).[20] The gene expression data are available on the NCBI Gene Expression Omnibus (GEO) under www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71220.

Smoking Status

Smoking status was analyzed using three different case definitions: (1) self-reported smoking status at the time of the blood draw; (2) eCO concentration; and (3) the combination of these two. Receiver Operating Characteristic (ROC) curves were used to determine the optimal eCO cut-off point that discriminated between self-reported former and current smokers according to Youden's criterion.[21] As illustrated in Supplementary Figure 1, the optimal eCO cut-off value was 8 ppm, which was associated with a maximal area under the curve (AUC) of 0.881 (95% CI, 0.847–0.916), a sensitivity of 83.3% and a specificity of 84.8%.

For the combined phenotype definition, smoking status was assigned using the self-reported smoking status with the additional exclusion of self-reported former smokers whose eCO concentrations were greater than 8 ppm (n = 54) and current smokers with eCO ≤ 8 ppm (n = 29). A boxplot of eCO concentrations stratified by self-reported smoking status is shown in Figure 1.

Differential Gene Expression Analysis

The Linear Models for Microarray Data (Limma) Bioconductor package[22] was used to evaluate genome-wide differential gene

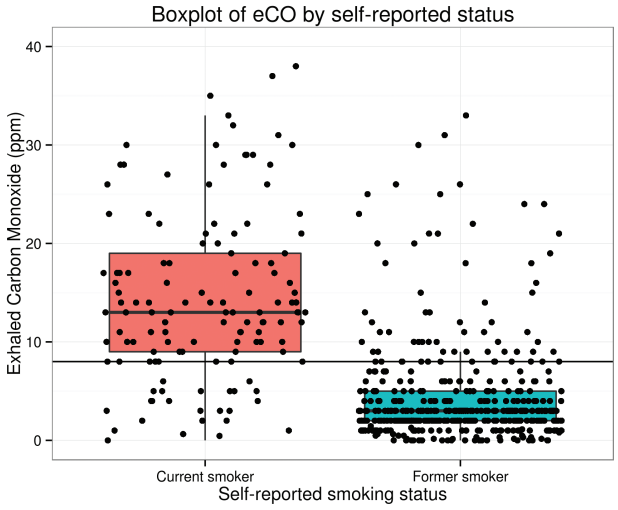

Figure 1. Boxplot of exhaled carbon monoxide (eCO) by self-reported status. eCO is shown in ppm on theY axis. Self-reported smoking status is shown on the X axis. The horizontal line at 8.1 ppm represents the cut-off point applied to remove subjects discordant for the two measures of smoking status. Self reported smokers with a eCO < 8.1 were excluded from the analysis as were self reported ex-smokers with an eCO > 8.1. Fifty-four and 29 subjects fell into these categories, respectively.

expression. A propensity score was used to adjust for potential confounders.[23] The propensity score was generated by using regression modeling in which smoking status was the dependent variable and the baseline characteristics were independent variables. Significant variables ($P < .05$ on univariate analysis) that were associated with smoking status were applied in stepwise selection to a logistic/linear regression model to identify predictive independent variables based on Akaike information criterion (AIC). This approach was taken for all three case definitions of smoking status to calculate a propensity score specific for each case definition. Limma was then used to test for differential gene expression with adjustments for age, sex, and the propensity score for each case definition. The Benjamini-Hochberg method was applied to correct for multiple testing.[24]

Quantile–Quantile Plot

To compare the distributions of *P* values for differential gene expression, quantile–quantile (QQ) plots were used. In a QQ plot, the observed *P* values from a particular analysis are plotted on the Y axis and a uniform distribution of *P* values obtained for a certain number of tests is plotted on the X axis. If the observed relationships are stronger than what would be expected by chance, then the line will deviate upwards from the diagonal line of identity.

Pathway Enrichment Analysis

Differentially expressed genes (unadjusted $P < .001$) were tested for enrichment in Gene Ontology (GO) biological processes and pathways using the WEB-based GEne SeT AnaLysis Toolkit (WebGestalt).[25] Enrichment was undertaken using hypergeometric tests[26] and corrected for multiple testing using the Benjamini-Hochberg False Discovery Rate (FDR).

Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA)[27] was used to compare overlap in results from the current study with published studies that have reported on the effects of smoking on gene expression. GSEA uses 1000 permutations and weighted enrichment statistics to identify Enrichment Score (ES) that evaluates if the genes are randomly distributed or found at the extremes (top or bottom) of the ranked list. The probability of the ES being false positive is assessed using FDR. A total of seven studies were identified in our search and were used in the GSEA analysis as summarized in Supplementary Table 1.

Statistical Analysis Software

All analyses were performed with R version 3.1.2 and Bioconductor packages.[28]

Results

The current study included 573 subjects. A total of 12 381 probesets that mapped to 7366 unique genes were tested for differential expression with respect to three different definitions of smoking. The results of differential gene expression using the three different definitions of current smoking are presented below. For the dichotomous definitions based on self-report and the combined phenotype, gene expression changes and test statistics in former smokers versus current smokers are reported, while the eCO analysis reports expression changes and test statistics for each unit increase in eCO concentration.

Self-Reported Smoking Status

Based on self-report, there were 126 current and 447 former smokers. The characteristics of study subjects using self reported smoking status are shown in Supplementary Table 2. The propensity score constructed for this model included the following variables: years smoked, body mass index, cigarettes/day, FEV_1% predicted, FEV_1/FVC, white blood cell (WBC) count, chronic wheeze, chronic cough, exacerbations in the year prior to blood withdrawal, and % monocytes.

At a nominal $P < .05$, 195 genes were differentially expressed following adjustments for propensity score, age, and sex. Of these genes, 137 were up-regulated and 58 were down-regulated. Two genes were differentially expressed at FDR < 0.1. Table 1 shows that the two most significantly different genes were both down-regulated in former smokers, namely LRRN3 (log2FC = −0.532, FDR = 0.0018) and PID1 (log2FC = −0.198, FDR = 0.0816). Figure 2A shows a volcano plot for differentially expressed genes using self-report of smoking, annotating genes differentially expressed at FDR < 10%. For the GO pathway enrichment analysis, genes with unadjusted $P < .001$ were selected. However, there was no enrichment for significant biological pathways following adjustments for multiple testing (FDR < 0.1).

eCO Analysis

This analysis included 573 subjects with eCO data available. The propensity score for this analysis included the following variables: years smoked, body mass index, FEV_1% predicted, FEV_1/FVC, WBC count, cough, exacerbations in the year prior to blood withdrawal, % neutrophils, and % monocytes. In this analysis, 282 genes were differentially expressed at a $P < .05$, with 194 genes that were down-regulated and 88 that were up-regulated in those with raised eCO concentrations. Two genes were significant at an FDR < 0.1; the most significant gene was LRRN3, similar

to the analysis based on self reported smoking status, with an FDR = 5.95E-06 and a log2 fold change of 0.0255. The second most significant gene was GPR15, with an FDR of 0.0017 and a log2 fold change of 0.0210. These two genes were up-regulated with increased eCO concentrations (Table 1). The volcano plot for these results is shown in Figure 2B. There were 10 unique genes at an unadjusted $P < .001$, which were used for the GO pathway analysis. However, there were no significant pathways detected at an FDR < 0.1.

Combined Smoking Status Phenotype

The use of the combined phenotype (eCO and self report) yielded 97 current and 393 former smokers. The characteristics of these subjects are shown in Supplementary Table 3. In this population, males represented 68% and 64% of current and former smokers, respectively. The average FEV_1% predicted was 51% and 54% in current and former smokers, respectively, while FEV_1/FVC was similar at 0.47. In addition, current smokers had a lower body mass index (25 kg/m^2 compared to 28 kg/m^2), had a higher WBC (8.1 GI/L compared to 7.4 GI/L), and had a lower number of exacerbations in the year prior to screening (0.34 compared to 0.84).

In this model, the propensity score included the following variables: years smoked, body mass index, cigarettes/day, FEV_1% predicted, FEV_1/FVC, WBC count, chronic cough, exacerbations in the year prior to blood draw, and % monocytes. A total of 454 genes were significantly associated at a $P < .05$; of these genes, 219 were down-regulated and 235 up-regulated in current smokers. Using an FDR < 0.1, four genes were differentially expressed: LRRN3 (log2FC = −0.7274, FDR = 2.31E-06), PID1 (log2FC = −0.2489, FDR = 0.0116), FUCA1 (log2FC = −0.1621, FDR = 0.0284), and GPR15 (log2FC = −0.5091, FDR = 0.0284; Table 1). The volcano plot of differential expression is shown in Figure 2C. For the pathway enrichment analysis, nine genes with a $P < .001$ were enriched in 40 significant pathways. The top 10 significant pathways are shown in Table 2. These pathways were related to metabolic processes, response to nutrients, and response to decreased oxygen levels, hypoxia, and hormones. The complete list of significant pathways is shown in Supplementary Table 4.

In a sensitivity analysis, a lower eCO cut-off of 7 ppm was used. In this analysis, 105 current and 379 former smokers were included. Compared to the 8 ppm cut-off, this analysis included eight more current and 14 fewer former smokers. A total of 620 genes showed association at $P < .05$. Of these, 355 were down-regulated and 265 were up-regulated. Using an FDR < 0.1, three genes were differentially expressed: LRRN3 (log2FC = −0.8051, FDR = 7.00E-09), GPR15 (log2FC = −0.5492, FDR = 0.0079), and PID1 (log2FC = −0.2301, FDR = 0.03177). A total of 12 genes with $P < .001$ were enriched in 34 biological processes (data not shown).

To visually compare the P value distributions of differentially expressed genes across the three different case definitions of smoking status, we created QQ plots, which are shown in Figure 3. The distribution of P values from the combined phenotype was parallel to the diagonal (expected) line and deviated towards lower P values indicating associations beyond what is expected by chance. The QQ plots of self report or eCO concentrations showed slightly deflated distributions.

GSEA Results

Results from the GSEA analyses revealed strong support for the signature identified in this study (Supplementary Table 1). The two strongest enrichments were discovered in two studies that

Table 1. Significant Differentially Expressed Genes at an FDR < 0.1

Definition of smoking status	Gene symbol	Gene name	LogFC	*P*	Adjusted *P* value
Self-reported	LRRN3	Leucine rich repeat neuronal 3	-0.5322	1.48E-07	.0018
	PID1	Phosphotyrosine interaction domain containing 1	-0.1978	1.32E-05	.0816
Exhaled carbon monoxide	LRRN3	Leucine rich repeat neuronal 3	0.0255	4.81E-10	5.95E-06
	GPR15	G protein-coupled receptor 15	0.0210	2.79E-07	.0017
Combined smoking status phenotype	LRRN3	Leucine rich repeat neuronal 3	-0.7274	1.87E-10	2.31E-06
	PID1	Phosphotyrosine interaction domain containing 1	-0.2489	1.88E-06	.0116
	FUCA1	Fucosidase, alpha-L-1, tissue	-0.1621	9.11E-06	.0284
	GPR15	G protein-coupled receptor 15	-0.5091	9.16E-06	.0284

FDR = False Discovery Rate. Adjusted *P*-Value: Benjamini Hochberg Adjusted *P*-value. LogFC: log2 fold change.

Table 2. The Top 10 Most Significant Pathways Identified Using the Combined Smoking Phenotype to Represent Smoking Status

Biological process	*P*	Adjusted *P*-value	Genes in this pathway
Positive regulation of macromolecule metabolic process	8.82E-05	.005	PTK2B/PID1/DBNL/USF1/CD38/LRRN3
Positive regulation of cellular metabolic process	9.92E-05	.005	PTK2B/PID1/DBNL/USF1/CD38/LRRN3
Positive regulation of metabolic process	.0001	.005	PTK2B/PID1/DBNL/USF1/CD38/LRRN3
Response to nutrient	.0002	.005	PTK2B/USF1/CD38
Positive regulation of phosphate metabolic process	.0002	.005	PTK2B/PID1/DBNL/LRRN3
Positive regulation of phosphorus metabolic process	.0002	.005	PTK2B/PID1/DBNL/LRRN3
Response to decreased oxygen levels	.0003	.005	PTK2B/USF1/CD38
Response to hypoxia	.0003	.005	PTK2B/USF1/CD38
Response to oxygen levels	.0003	.005	PTK2B/USF1/CD38
Response to hormone stimulus	.0004	.006	PTK2B/PID1/USF1/CD38

Adjusted *P*-value: Benjamini Hochberg Adjusted *P*-value.

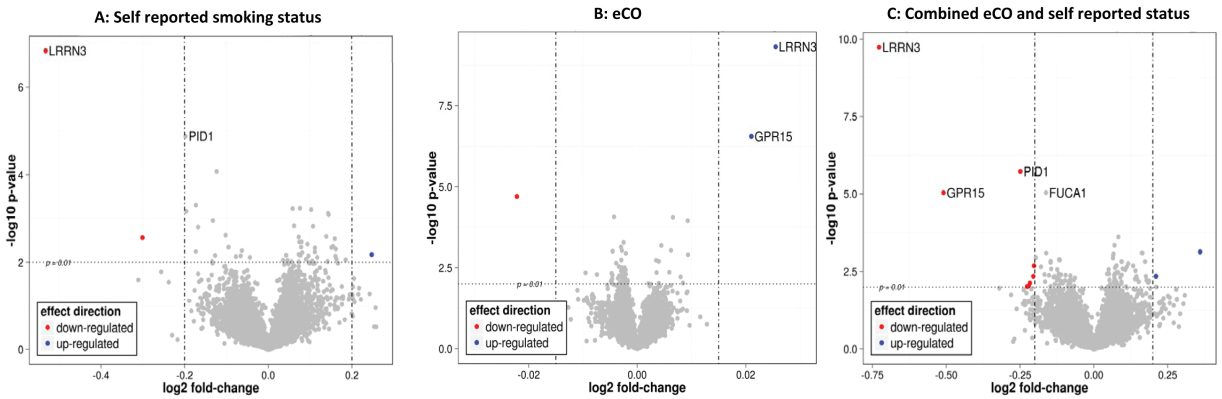

Figure 2. Volcano Plots of differential gene expression using three case definitions. The plot shows the log2 fold difference in gene expression on the X axis versus the unadjusted *P* values (on the –log10 scale) on the Y axis. For the self-reported status and the combined phenotype, the blue and red dots represent genes that showed fold change in either direction greater than 0.2 and have unadjusted *P* value < .01. Genes that had a False Discovery Rate (FDR) adjusted *P* values less than .1 for differential expression are annotated on the graph. (A) Self-reported smoking status; (B) exhaled carbon monoxide (eCO); (C) Combined phenotype. The eCO plot (B) shows gene expression differences per unit increase in eCO concentration, hence the X scale is different to A and C.

reported the effects of smoking on gene expression and methylation profiles. Fifty genes in the study by Charlesworth et al.,[29] 15 CpG sites corresponding to 14 unique genes in the study by Wan et al.,[30] and five genes in the study by Beineke et al.[31] showed significant enrichment against the 7366 unique genes pre-ranked by their absolute *t*-statistics in our gene set from the combined phenotype analysis using FDR *q*-values of 0.00058, 0.0012 and 0.0385, respectively.

Discussion

In the current study, we used three different case definitions to identify peripheral blood gene expression signatures related to current smoking. Using FDR <0.1 for genes and pathway enrichment, we found two genes that were differentially expressed and no enriched pathway using self report or eCO concentrations by themselves. By combining self-report with eCO concentrations, however, we discovered four differentially expressed genes and 40 pathways that were enriched in the current smokers. The QQ plots in Figure 3 suggest that the combined phenotype produced a "better" *P* value distribution, providing added confidence to this approach. The plot also shows that the combined phenotype approach produced more significant (smaller) *P* values, thus reducing the risk of false negatives. Interestingly, the increase in statistical significance was achieved despite the fact that the sample size was reduced in the combined analysis, indicating the improved resolution of this approach in ascertaining significant gene expression changes related to current smoking.

The genes that were differentially expressed in all three analyses are very well supported by data from previous studies on smoking and gene expression and/or methylation. In the study of Wan et al.,[30] hypomethylation of the CpG sites at LRRN3 in peripheral blood correlated with current smoking and a shorter time since quitting in former smokers, and hypomethylation of the GPR15 site correlated with current smoking, higher cumulative smoke exposure (more pack-years) and shorter time since quitting in former smokers. In the study of Charlesworth et al.,[29] smoking status was associated with gene expression of LRRN3, PID1, and FUCA1 in lymphocytes. In lung tissue, Bossé et al. reported up-regulation of FUCA1 among smokers.[32] The study of Beineke et al.[31] reported up-regulation of

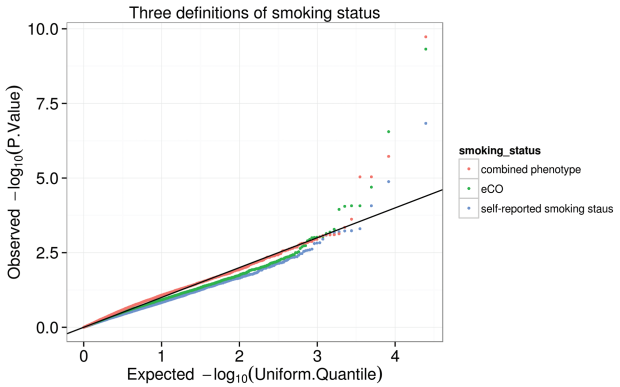

Figure 3. A quantile–quantile (QQ) plot for the three case definitions of smoking status. The X axis is –log10 of the expected *P*-values, and the Y axis is –log10 of the actual *P*-values in QQ plot. Under the null hypothesis, the points should fall approximately along the 45-degree reference line. Genes with low *P* values deviate from the reference line, indicating significant association.

LRRN3, PID1, GPR15 and FUCA1 gene expression in the whole blood of current smokers with LRRN3 being the most statistically significant gene. In a recent study by Guida et al.,[33] CpG sites at both LRRN3 and GPR15 were found to be hypomethylated in blood. LRRN3 is particularly interesting because it has been shown to be over-expressed in current smokers and in the study by Guida et al.,[33] it was the only gene associated with smoking at the level of both methylation and gene expression. In genetic association studies, variants in PID1 have been associated with lung function in the Korean population.[34]

The exact biological roles of LRRN3, PID1, FUCA1, and GPR15 in smoking are not clear. Leucine rich repeat neuronal 3 (LRRN3) is thought to play a role in neural development and regeneration[35] and is up-regulated during cortical neuronal injury.[36] Its genetic variants have been associated with autism.[37] Phosphotyrosine interaction domain containing one (PID1) function as a growth-inhibitory gene in brain tumors[38] and is a potent intracellular inhibitor of the insulin signaling pathway during obesity in humans and mice.[39] Fucosidase, alpha-L-1, tissue (FUCA1) is a liposomal enzyme that degrades a variety of fucose-containing fucoglycoconjugates and has been proposed as a promising

tumor marker in the diagnosis[40] and prognosis[41] of hepatocellular carcinoma. Finally, G protein-coupled receptor 15 (GPR15) acts as a chemokine receptor for human immunodeficiency virus.[42]

Biological pathway enrichment was detected for the self-report/eCO combined phenotype, and revealed intriguing biology about the role of smoking related genes. The pathways were related to regulation of metabolic processes, response to decreased oxygen levels and hypoxia, response to hormones and lipids, response to cytokines and cell proliferation. The effect of smoking in reducing tissue oxygenation is well established.[43–45] Smoking leads to oxidative stress which induces systemic inflammation.[46] Accordingly, smokers consistently demonstrate elevated levels of proinflammatory cytokines in the blood.[47] Additionally, epidemiological studies have shown that smokers have higher levels of total cholesterol, triglycerides, and low-density lipoprotein cholesterol (LDLC) compared with nonsmokers.[48] Smoking has also been shown to affect the metabolic and biological processes including secretion of hormones.[49] These data are consistent with the biological processes identified in the combined smoking definition analysis.

In this study, defining smoking status using a combination of self report and eCO concentrations produced better distribution of test statistics for gene expression and identified more genes that showed stronger enrichment in biological pathways when compared to using self-report or eCO alone.

Smoking is one of the most common variables used to adjust for confounders in genetic, epidemiological and interventional studies. Our data suggest that self-report is probably insufficient to fully capture the effects of smoking in these studies, leading to some degree of residual confounding by smoking. A genome-wide association study of genetic variation underlying eCO levels while adjusting for self report arrived at a similar conclusion where eCO was found to capture aspects of cigarette smoke exposure in current smokers beyond the number of cigarettes smoked per day.[50] Thus for future -omics and other studies, our data strongly suggest the need to complement self-report with an objective measure of active smoking such as eCO to more accurately capture the effects of smoking in these studies.

The current study has a number of limitations. First, the sample size of the study may have been too small to detect modest changes in gene expression. On the other hand, the top genes identified in our study have been previously associated with smoking status suggesting that with the current sample size we were able to detect a reproducible smoking-related gene signature. Second, we did not assess the impact of different case definitions on hard clinical endpoints such as mortality as this was beyond the purview of the present study. Third, we also did not evaluate other objective measures of smoking status such as cotinine. Thus, the impact of using urinary, blood or salivary cotinine in lieu of or in addition to eCO is not known. Fourth, although the differential gene expression analysis was adjusted for a number of variables including cell percentages, this may not be able to fully account for the cellular changes between former and current smokers. Finally, the use of eCO is not without limitations. eCO has a short half-life and studies have shown that eCO may be a poor biomarker of smoking that occurred more than 8 hours prior to the eCO measurement.[51] A number of studies also showed eCO to be elevated in individuals with respiratory diseases such as asthma or COPD.[52,53] Current and former smokers phenotyped using the combined definition were not statistically different with respect to their lung function and the propensity score included FEV_1 as a percentage of predicted, chronic cough and exacerbations, providing some assurances that the results were unlikely to have been confounded by COPD status or severity.

In conclusion, blood from current smokers exhibits a differential expression profile when compared to former smokers. Combining both eCO and self reported smoking status to define current and former smokers improved the discovery of differentially expressed genes and enriched pathways. For future studies, combining self-reported smoking status with eCO will enhance the statistical power of these studies to discover (or adjust for) the impact of smoking.

Supplementary Material

Supplementary Tables 1–4 and Supplementary Figure 1 can be found online at http://www.ntr.oxfordjournals.org

Funding

ECLIPSE study (www.eclipse-copd.com/) was funded by GlaxoSmithKline (GSK) http://ClinicalTrials.gov identifier NCT00292552 and GSK No. SCO104960. Funders have no role in study design, data collection, analysis, interpretation or preparing the manuscript. Gene expression analysis was funded by Genome Canada and the Canadian Respiratory Research Network (CRRN). MO is a Postdoctoral Fellow of the Michael Smith Foundation for Health Research (MSFHR) and the Canadian Institute for Health Research (CIHR) Integrated and Mentored Pulmonary and Cardiovascular Training program (IMPACT). DDS is a Canada Research Chair in COPD.

Declaration of Interests

BEM is an employee and shareholder of GSK. SR has served as a consultant, participated in advisory boards, received honorarium for speaking or grant support from: American Board of Internal Medicine, Advantage Healthcare, Almirall, American Thoracic Society, AstraZeneca, Baxter, Boehringer Ingelheim, Chiesi, ClearView Healthcare, Cleveland Clinic, Complete Medical Group, CSL, Dailchi Sankyo, Decision Resources, Forest, Gerson Lehman, Grifols, GroupH, Guidepoint Global, Haymarket, Huron Consulting, Inthought, Johnson and Johnson, Methodist Health System—Dallas, NCI Consulting, Novartis, Pearl, Penn Technology, Pfizer, PlanningShop, PSL FirstWord, Qwessential, Takeda, Theron and WebMD. Since August 10, 2015 he has served as chief clinical scientist, new clinical development, AstraZeneca, UK. DDS: Over the past 3 years, DDS has served as a consultant on AstraZeneca (AZ) and Novartis Advisory Boards for COPD. He has been a consultant with Amgen and Almirall. He has received research funding from AZ and Boehringer Ingelheim (BI). He has given lectures sponsored by BI and AZ.

Acknowledgments

MO and XD contributed equally for this article.

References

1. WHO. *Tobacco Fact Sheet*. Geneva, Switzerland: World Health Organization; 2015.
2. Peto RLA, Boreham J, Thun M. Mortality from smoking in developed countries 1950–2020. 2015. www.ctsu.ox.ac.uk/~tobacco. Accessed January 15, 2015.
3. Jha P, Peto R. Global effects of smoking, of quitting, and of taxing tobacco. *NEJM*. 2014;370(1):60–68.
4. Jha P. Avoidable global cancer deaths and total deaths from smoking. *Nat Rev Cancer*. 2009;9(9):655–664.
5. Doll R, Peto R. Mortality in relation to smoking: 20 years' observations on male British doctors. *BMJ*. 1976;2(6051):1525–1536.
6. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859):2095–2128.

7. (CIHI)" CIfHI. *A Snapshot of Health Care in Canada as Demonstrated by Top 10 Lists, 2011*. Ottawa, Canada: CIHI; 2011.
8. Stelmach R, Fernandes FLA, Carvalho-Pinto RM, et al. Comparison between objective measures of smoking and self-reported smoking status in patients with asthma or COPD: are our patients telling us the truth? *J Brasileiro de Pneum*. 2015;41(2):124–132.
9. Fendrich M, Mackesy-Amiti ME, Johnson TP, Hubbell A, Wislar JS. Tobacco-reporting validity in an epidemiological drug-use survey. *Add Beh*. 2005;30(1):175–181.
10. Sato S, Nishimura K, Koyama H, et al. Optimal cutoff level of breath carbon monoxide for assessing smoking status in patients with asthma and COPD. *Chest*. 2003;124(5):1749–1754.
11. Jarvis MJ, Russell MA, Saloojee Y. Expired air carbon monoxide: a simple breath test of tobacco smoke intake. *BMJ*. 1980;281(6238):484–485.
12. Edmiston JS, Archer KJ, Scian MJ, Joyce AR, Zedler BK, Murrelle EL. Gene expression profiling of peripheral blood leukocytes identifies potential novel biomarkers of chronic obstructive pulmonary disease in current and former smokers. *Biomarkers*. 2010;15(8):715–730.
13. Poliska S, Csanky E, Szanto A, et al. Chronic obstructive pulmonary disease-specific gene expression signatures of alveolar macrophages as well as peripheral blood monocytes overlap and correlate with lung function. *Respiration*. 2011;81(6):499–510.
14. Bahr TM, Hughes GJ, Armstrong M, et al. Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease. *Am J Respir Cel Mol Biol*. 2013;49(2):316–323.
15. Bhattacharya S, Tyagi S, Srisuma S, et al. Peripheral blood gene expression profiles in COPD subjects. *J Clin Bio*. 2011;1(1):12.
16. Vestbo J, Anderson W, Coxson HO, et al. Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *ERJ*. 2008;31(4):869–873.
17. Obeidat M, Fishbane N, Nie Y, et al. The Effect of Statins on Blood Gene Expression in COPD. *PLoS One*. 2015;10(10):e0140022.
18. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*. 2010;26(19):2363–2367.
19. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–264.
20. Hochreiter S, Clevert D-A, Obermayer K. A new summarization method for affymetrix probe level data. *Bioinformatics*. 2006;22(8):943–949.
21. YOUDEN WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–35.
22. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Gen Mol Bio*. 2004;3(1):1–25.
23. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multi Beh Res*. 2011;46(3):399–424.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc Series B (Methodological)*. 1995;57(1):289–300.
25. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nuc Aci Res*. 2013;41(W1):W77–W83.
26. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. *Bioinformatics*. 2007;23(2):257–258.
27. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550.
28. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Gen Bio*. 2004;5(10):R80.
29. Charlesworth J, Curran J, Johnson M, et al. Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. *BMC Med Gen*. 2010;3(1):29.
30. Wan ES, Qiu W, Baccarelli A, et al. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum Mol Gen*. 2012;21(13):3073–3082.
31. Beineke P, Fitch K, Tao H, et al. A whole blood gene expression-based signature for smoking status. *BMC Med Gen*. 2012;5(1):58.
32. Bosse Y, Postma DS, Sin DD, et al. Molecular Signature of Smoking in Human Lung Tissues. *Can Res*. 2012.
33. Guida F, Sandanger TM, Castagné R, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum Mol Gen*. 2015;24(8):2349–2359.
34. Kim WJ, Lim MN, Hong Y, et al. Association of lung function genes with chronic obstructive pulmonary disease. *Lung*. 2014;192(4):473–480.
35. Haines BP, Gupta R, Jones CM, Summerbell D, Rigby PW. The NLRR gene family and mouse development: modified differential display PCR identifies NLRR-1 as a gene expressed in early somitic myoblasts. *Dev Biol*. 2005;281(2):145–159.
36. Chen Y, Aulia S, Li L, Tang BL. AMIGO and friends: an emerging family of brain-enriched, neuronal growth modulating, type I transmembrane proteins with leucine-rich repeats (LRR) and cell adhesion molecule motifs. *Brain Res Rev*. 2006;51(2):265–274.
37. Sousa I, Clark T, Holt R, et al. Polymorphisms in leucine-rich repeat genes are associated with autism spectrum disorder susceptibility in populations of European ancestry. *Mol Auti*. 2010;1(1):7.
38. Erdreich-Epstein A, Robison N, Ren X, et al. PID1 (NYGGF4), a new growth-inhibitory gene in embryonal brain tumors and gliomas. *Clin Cancer Res*. 2014;20(4):827–836.
39. Sabeera B, Craig M, Jackie A, et al. Pid1 induces insulin resistance in both human and mouse skeletal muscle during obesity. *Mol Endo*. 2013;27(9):1518–1535.
40. Leray G, Deugnier Y, Jouanolle AM, et al. Biochemical aspects of alpha-L-fucosidase in hepatocellular carcinoma. *Hepatology*. 1989;9(2):249–252.
41. Wang K, Guo W, Li N, et al. Alpha-1-fucosidase as a prognostic indicator for hepatocellular carcinoma following hepatectomy: a large-scale, long-term study. *Brit J Can*. 2014;110(7):1811–1819.
42. Blaak H, Boers PH, Gruters RA, Schuitemaker H, van der Ende ME, Osterhaus AD. CCR5, GPR15, and CXCR6 are major coreceptors of human immunodeficiency virus type 2 variants isolated from individuals with and without plasma viremia. *J Virol*. 2005;79(3):1686–1700.
43. Jensen JA, Goodson WH, Hopf HW, Hunt TK. Cigarette smoking decreases tissue oxygen. *Arch Surg*. 1991;126(9):1131–1134.
44. Morecraft R, Blair WF, Brown TD, Gable RH. Acute effects of smoking on digital artery blood flow in humans. *J Hand Surg Am*. 1994;19(1):1–7.
45. Sagone AL Jr, Lawrence T, Balcerzak SP. Effect of smoking on tissue oxygen supply. *Blood*. 1973;41(6):845–851.
46. Grassi D, Desideri G, Ferri L, Aggio A, Tiberti S, Ferri C. Oxidative stress and endothelial dysfunction: say NO to cigarette smoking! *Curr Pharm Des*. 2010;16(23):2539–2550.
47. van Zyl-Smit RN, Binder A, Meldau R, et al. Cigarette smoke impairs cytokine responses and BCG containment in alveolar macrophages. *Thorax*. 2014;69(4):363–370.
48. Craig WY, Palomaki GE, Haddow JE. Cigarette smoking and serum lipid and lipoprotein concentrations: an analysis of published data. *BMJ*. 1989;298(6676):784–788.
49. Kapoor D, Jones TH. Smoking and hormones in health and endocrine disorders. *Euro J Endo*. 2005;152(4):491–499.
50. Bloom AJ, Hartz SM, Baker TB, et al. Beyond cigarettes per day. A genome-wide association study of the biomarker carbon monoxide. *Ann Am Thorac Soc*. 2014;11(7):1003–1010.
51. Sandberg A, Sköld CM, Grunewald J, Eklund A, Wheelock ÅM. Assessing recent smoking status by measuring exhaled carbon monoxide levels. *PLoS One*. 2011;6(12):e28864.
52. Ryter SW, Sethi JM. Exhaled carbon monoxide as a biomarker of inflammatory lung disease. *J Breath Res*. 2007;1(2):026004.
53. Horvath I, Loukides S, Wodehouse T, Kharitonov SA, Cole PJ, Barnes PJ. Increased levels of exhaled carbon monoxide in bronchiectasis: a new marker of oxidative stress. *Thorax*. 1998;53(10):867–870.