Molecular Signature of Smoking in Human Lung Tissues

Yohan Bossé1,2, Dirkje S. Postma3, Don D. Sin4,5, Maxime Lamontagne2, Christian Couture2, Nathalie Gaudreault2, Philippe Joubert2, Vivien Wong4, Mark Elliott4, Maarten van den Berge3, Corry A. Brandsma6, Catherine Tribouley7, Vladislav Malkov7, Jeffrey A. Tsou7, Gregory J. Opiteck7, James C. Hogg4,8, Andrew J. Sandford4,5, Wim Timens6, Peter D. Paré4,5, Michel Laviolette2

1Department of Molecular Medicine, Laval University, Québec, Canada;
2Institut universitaire de cardiologie et de pneumologie de Québec, Québec, Canada;
3Department of Pulmonology, University Medical Center Groningen, University of Groningen, GRIAC research institute, Groningen, The Netherlands;
4UBC James Hogg Research Center, Providence Heart + Lung Institute, St. Paul’s Hospital, Vancouver, BC, Canada;
5Respiratory Division, Department of Medicine, The University of British Columbia, Vancouver, BC, Canada;
6Department of Pathology and Medical Biology, University Medical Center Groningen, University of Groningen, GRIAC research institute, Groningen, The Netherlands;
7Merck & Co. Inc., Rahway, New Jersey, USA;
8Department of Pathology and Laboratory Medicine, The University of British Columbia, Vancouver, BC, Canada.

Running title: The lung transcriptome in smokers

Keywords: gene expression; tobacco; former-smokers; lung cancer; genomics
Financial support:
Merck Research Laboratories, Chaire de pneumologie de la Fondation JD Bégin de l’Université Laval, Fondation de l’Institut universitaire de cardiologie et de pneumologie de Québec, Respiratory Health Network of the FRQS, Cancer Research Society.

Corresponding author:
Yohan Bossé, Ph.D.
Assistant Professor, Laval University
Institut universitaire de cardiologie et de pneumologie de Québec
Pavillon Marguerite-d'Youville, Y4190
2725, chemin Sainte-Foy
Québec (Québec), Canada, G1V 4G5
Tel: 418-656-8711 ext. 3725
Fax: 418-656-4602
email: yohan.bosse@crhl.ulaval.ca

Conflicts of interest:
Four authors are full time employees of Merck & Co Inc.

Abstract word count: 216

Word count text: 5,499

Total numbers of figures and tables: 7

Number of references: 48
Abstract

Cigarette smoking is the leading risk factor for lung cancer. To identify genes deregulated by smoking and to distinguish gene expression changes that are reversible and persistent following smoking cessation, we performed genome-wide gene expression profiling on non-tumor lung tissue from 853 patients with lung cancer. Gene expression levels were compared between never- and current-smokers, and time-dependent changes in gene expression were studied in former-smokers. A total of 3,223 transcripts were differentially expressed between smoking groups in the discovery set (n=344, p<1.29x10^{-6}). A substantial number of smoking-induced genes also were validated in two replication sets (n=285 and 224), and a gene expression signature of 599 transcripts consistently segregated never- from current-smokers across all three sets. The expression of the majority of these genes reverted to never-smoker levels following smoking cessation although the time course of normalization differed widely among transcripts. Moreover, some genes showed very slow or no reversibility in expression, including SERPIND1, which was found to be the most consistent gene permanently altered by smoking in the three sets. Our findings therefore indicate that smoking deregulates many genes, many of which reverse to normal following smoking cessation. However, a subset of genes remains altered even decades following smoking cessation and may account, at least in part, for the residual risk of lung cancer among former-smokers.

Precis

This study utilizes whole-genome gene expression to demonstrate the long-term impact of smoking on gene expression in non-tumor lung tissues from patients with lung cancer.
Introduction

Globally, more than one billion people smoke tobacco products (1). Tobacco smoke contains about 5000 compounds including many carcinogens (2). Smoking increases the risk of lung cancer 20-fold (3). Smoking cessation reduces the relative risk of lung cancer, but individuals who quit smoking still have an increased risk even decades following smoking cessation compared to never-smokers (4, 5).

The molecular effects of smoking and smoking cessation in the human lung are still not well understood. Previous transcriptomic studies indicate that a large number of genes are modulated by smoking in airways (6-14), alveolar macrophages (15-17), lung tumors (7, 18-21), and peripheral leukocytes (22-27). Fewer transcriptomic studies have also examined the effect of the duration of smoking cessation on gene expression profiles (6, 8, 10, 12). All these studies were performed in epithelial cells obtained from bronchial brushings with sample sizes that range from 24 to 104 individuals. These studies showed that the gene expression pattern is similar between former and never-smokers, suggesting that most smoking-induced gene expression reverse following smoking cessation. However, the same studies also demonstrated that the levels of expression of a small group of genes are permanently altered by cigarette smoke, but little overlap was found among studies. In addition, it is unknown whether these findings in human airway epithelial cells can be extrapolated to human lung parenchyma. Only a few transcriptomic studies have evaluated the impact of tobacco smoking on gene expression in histologically normal pulmonary parenchyma (18, 20, 21). The largest among the later studies was performed with 15 never, 18 former, and 16 current smokers (18).
Here we reported the results of a large-scale whole-genome gene expression study evaluating the impact of smoking and smoking cessation in non-tumor lung tissue. We hypothesized that the gene expression profile of the human lung would be altered by cigarette smoking and that most of these changes would revert to never-smoking levels following smoking cessation. We also hypothesized that the expression of a smaller number of genes would be permanently altered by smoking. To examine these hypotheses, we collected non-tumor lung samples from a large group of patients who underwent lung surgery. The lung transcriptomes of never and current-smokers were then compared. Genes differentially expressed among smoking groups were then validated in two independent replication sets of lung specimens to obtain a list of genes consistently altered by smoking. The genes reproducibly induced by smoking were further studied in relationship to years of smoking cessation in former-smokers in order to identify early, late, and never reversible genes.
Materials and Methods

Discovery set

Lung tissue was obtained from patients undergoing lung cancer surgery between April 2004 and December 2008. Specimens for the discovery set were taken from the Institut universitaire de cardiologie et de pneumologie de Québec (IUCPQ) site of the Respiratory Health Network Tissue Bank of the Fonds de la recherche en santé du Québec. Lung tissue samples were obtained in accordance with Institutional Review Board guidelines. All patients provided written informed consent and the study was approved by the ethics committee of the IUCPQ.

Smoking status

Smoking history included self-reported current smoking status, number of pack-years, and year of smoking cessation (for former-smokers). Plasma cotinine quantification was determined by HPLC-MS-MS (ACQUITY UPLC® System and the Quattro Premier XE, Waters, Milford, Mass, USA). Current-smokers were defined as subjects self-reported as smokers who also had a plasma cotinine concentration greater than or equal to 15 ng/mL (28). Never and former-smokers were defined as subjects self-reported as lifelong non-smokers and ex-smokers, respectively, whose plasma cotinine concentrations were below detection (<0.4 ng/mL). Never and former-smokers with cotinine levels >0.4 ng/mL were excluded from analysis.

Tissue processing

After surgical removal, lung specimens were immediately examined by a pulmonary pathologist (C.C.). After processing for pathological diagnosis and staging, a non-neoplastic pulmonary parenchyma sample (2-5 cm³) was harvested from a site as far distant as possible from the tumor. The research specimens were immediately divided into smaller fragments (~0.5 cm³) placed in 5
ml cryovials and snap-frozen in liquid nitrogen. The cryovials were then transported in dry ice to the IUCPQ biobank where they were stored at -80°C until further processing. The time from surgical removal to storage was between 15 to 30 minutes.

**Replication sets**

Lung specimens collected at two other sites (University of British Columbia, Vancouver, Canada and University of Groningen, Groningen, The Netherlands) were used as replication sets. At the UBC site, the study was approved by the ethics committee of the UBC-Providence Health Care Research Institute Ethics Board. At the Groningen site, the study protocol was consistent with the Research Code of the University Medical Center Groningen and Dutch national ethical and professional guidelines (“Code of conduct; Dutch federation of biomedical scientific societies”; http://www.federa.org). To obtain replication sets similar to the discovery set, only subjects with lung neoplasm were considered. The final analyses were conducted with 285 and 224 lung specimens from UBC and Groningen, respectively.

**Assays**

Whole-genome gene expression profiling was performed using Affymetrix arrays at the same facility using the same methods for the discovery and replication sets. Microarray platform and preprocessing, quality controls, RNA extraction, quantitative real-time PCR, and immunohistochemistry are described in the supplementary data. Gene expression data are available through GSE23546.
Statistical analyses

All analyses were performed with the R statistical software version 2.9.0 and Bioconductor packages (29).

Gene filtering. The MAS5 call from the R affy package was used to remove genes that could not be adequately detected in the human lung tissues. Only probe sets called “present” in at least 20% of the samples were kept for subsequent analyses. A total 38,820 probe sets were available following the application of this filter.

Selection of covariates. The impact of standard demographic variables and possible confounders, such as COPD and types of cancer, on gene expression in the lung were tested independently to select covariates. Smoking had a much bigger impact on gene expression than COPD and lung cancer considered independently (see Supplementary Fig. 1), which is consistent with the idea that smoking induces molecular damages that in turn lead to lung diseases. Several clinical variables had an effect size similar to COPD or types of lung cancer on gene expression and adjusting for all these variables would have made our model unstable. Accordingly expression traits were only adjusted for age, sex and height. Please note that we further rely on replication sets to identify genes reproducibly modulated by smoking.

Expression trait processing. The distribution of expression traits was largely non-Gaussian. Accordingly, we used robust residuals and nonparametric tests to perform the association analyses. The rlm function in the R statistical package MASS was used to calculate residuals (M-estimation with Tukey’s bisquare weights). Residual values deviating from the median by more than three standard deviations were filtered out as outliers.

Association tests. Wilcoxon tests were used to compare adjusted expression traits between never- and current-smokers. We applied a Bonferroni correction to correct for multiple testing (0.05/38,820 probe sets, p value < 1.29x10^-6). The fold changes were obtained by raising 2 to the power of the difference in log2 expression levels.
power of the mean difference in expression between never- and current smokers. Transcripts differentially expressed were those that passed the Bonferroni correction and had a fold change greater than or equal to 1.2.

**Replication sets.** Tests of association between gene expression and smoking status were performed as described above. However association tests were limited to the 3,223 probe sets that were significantly altered by smoking in the discovery set. To adjust for multiple testing, we used both the Benjamini-Hochberg procedure to calculate the false discovery rate (30) and the Bonferroni correction (0.05/3,223 probe sets, \(p<1.55\times10^{-5}\)).

**Smoking cessation and normalization of gene expression**

We estimated the time it takes for gene expression to revert to never-smoker levels after smoking cessation. These reversibility analyses were only performed on probe sets significantly altered by smoking in the three data sets. To do this, we first determined the mean expression values for each probe set for both never and current-smokers. Next, we determined the mean expression values for former-smokers in relation to the elapsed time between smoking cessation and the time of surgery. The elapsed time intervals were divided such that there were at least 25 individuals per bin. The time intervals in the discovery set were: a) within two years before surgery, b) 2-5 years, c) 6-9 years, d) 10-14 years, e) 15-19 years, f) 20-24 years, and g) 25-49 years, resulting in bin sizes of 26, 29, 32, 26, 31, 28, and 37 subjects respectively. Probe sets were considered to have reverted to the “normal” value if, within a given time interval, the Wilcoxon test comparing former-smokers to never-smokers was no longer significant \((p>0.01)\). These time course analyses generated eight clusters of up- (U1-U8) and eight clusters of down-regulated (D1-D8) probe sets based on the duration of smoking cessation to return to normal. U8 and D8 represent probe sets that do not return to normal.
Similar to the discovery set, the time intervals for the replication sets were divided such that there were at least 25 individuals per bin. The final time intervals strata differ across populations owing to differences in sample sizes and the distribution in the number of years of smoking cessation. The time intervals for the UBC set were: a) 0-2 years, b) 3-9 years, c) 10-19 years, and d) 20-50 years, resulting in bin sizes of 36, 38, 30, and 25 subjects, respectively. Five clusters of up- (U1-U5) and five clusters of down-regulated (D1-D5) probes sets were formed for the UBC analyses. The time intervals for the Groningen set were: a) 0-2 years, b) 3-9 years, and c) 10-30 years, resulting in bin sizes of 36, 71, and 49 subjects, respectively. Four clusters of up- (U1-U4) and four clusters of down-regulated (D1-D4) probes sets were formed for the Groningen analyses.

In order to identify slowly or never reversible genes that were consistent across populations, we performed chi-square tests (2x2 frequency tables) comparing the distribution of probe sets returning or not returning to never-smoker levels within ten years of smoking cessation between the discovery set and the replication sets. Probe sets not returning to never-smoker levels within 10 years of smoking cessation in the discovery set and in at least one of the two replication sets were identified.
Results

Discovery set population

The clinical characteristics of the 344 patients who passed microarray quality control and blood cotinine filters in the discovery set are shown in Table 1. Most patients were former (61.3%) or current (26.2%) smokers. The vast majority of these patients underwent lung surgery for non-small-cell lung cancer (Table 1), predominantly adenocarcinoma (55.5%) and squamous cell carcinoma (27.6%).

Transcripts associated with smoking status in the discovery set

A total of 38,820 adjusted expression traits were compared between never- and current-smokers. 3,223 probe sets were differentially expressed between the two groups (i.e. p value<1.29x10^{-6} and fold change > 1.2). These included 1,943 differentially up-regulated and 1,280 down-regulated probe sets in smokers (Supplementary Table S1). Fig. 1 shows a heat map of these probe sets. Former-smokers were ordered based on the duration of smoking cessation (Fig. 1, upper panel). The date of smoking cessation was not available for two former-smokers. These two patients were excluded from the heat map. A clear separation in the pattern of gene expression was observed between never and current-smokers. The most significant probe set interrogated the aryl-hydrocarbon receptor repressor (AHRR) gene, which had a p value of 3.3x10^{-20} and a fold change of 6.1. The next three most significant probe sets were all testing the CYP1B1 gene (p values<1.1x10^{-19} and fold changes > 4.0). SERPIND1 was also among the top up-regulated genes in smokers with a p value of 5.2x10^{-17} and fold change of 13.3. The changes in expression of AHRR, SERPIND1, and CYP1B1 were confirmed by quantitative real-time PCR in a subset of samples (Fig. 2).
Replication sets

The UBC set included 30 never-smokers, 158 former-smokers, and 97 current-smokers. The clinical characteristics of these subjects by smoking group are provided in Supplementary Table S2. Of the 3,223 probe sets found to be significant in the discovery set, 1,696 probe sets were significantly associated with smoking following the Benjamini-Hochberg correction. Considering the large number of probe sets, we also applied a more stringent multiple testing correction factor in order to focus on the most strongly replicated genes. Following Bonferroni correction (p<1.55x10^{-5}), 144 probe sets were significantly associated with smoking in the UBC data set. There was 100% concordance in the orientation of the altered gene expression for the 1,696 probe sets between the two cohorts (i.e. genes up-regulated by smoking in the discovery samples were also up-regulated in the UBC samples and similarly for genes that were down-regulated).

The Groningen set included 16 never-smokers, 164 former-smokers, and 44 current-smokers. The clinical characteristics of these subjects by smoking group are provided in Supplementary Table S3. Association tests were performed between smoking status and 3,223 probe sets. 910 and 30 probe sets were significantly associated with smoking status following Benjamini-Hochberg and Bonferroni (p<1.55x10^{-5}) corrections, respectively. The orientation of the effect for the 30 probes sets was 100% concordant between the discovery and Groningen samples. For the Benjamini-Hochberg threshold 906 out of 910 probe sets were concordant. The p values for the 3,223 probe sets in both replication samples are provided in Supplementary Table S1.
Genes associated with smoking across the three populations

Using stringent Bonferroni correction, seven probe sets overlapped among significant probe sets in the discovery (n=3,223), UBC (n=144), and Groningen (n=30) populations. Table 2 shows these probe sets with annotation and association test results for the three populations. These seven highly reproducible probe sets were all up-regulated by smoking. Using a less stringent correction for multiple testing (i.e. Benjamini-Hochberg), 599 probe sets overlapped among significant probe sets in the discovery (n=3,223), UBC (n=1,696), and Groningen (n=910) populations. These include 558 and 41 up- and down-regulated probe sets, respectively. The direction of expression is 100% concordant across the three populations for these probe sets. Supplementary Table S1 shows these reproducible probe sets.

Time course analyses

Fig. 3 shows the time course analyses for the most significant probe set associated with smoking in the discovery set (i.e. AHRR). The expression of the AHRR gene falls substantially following smoking cessation, but remained significantly higher in former-smokers who had quit smoking for more than 25 years compared to never-smokers (p=0.0001). The same analyses were conducted for all of the reproducible probe sets (n=599) in order to identify genes that are slowly or never-reversible following smoking cessation. Among up-regulated probe sets (n=558), 135 returned to never-smoker levels within two years of smoking cessation (cluster U1), 202 probe sets returned to normal after 2-5 years (U2), 29 after 6-9 years (U3), 57 after 10-14 years (U4), 44 after 15-19 years (U5), 59 after 20-24 years (U6), and 12 after 25 years (U7). Finally the level of expression of 20 probe sets (U8) remained significantly different (p>0.01) between former-smokers who quit smoking for more than 25 years and never-smokers. Among these 20 never reversed probe sets, the three that interrogate the CYP1B1 gene were the most significant. For
down-regulated probe sets (n=41), 11 returned to never-smoker levels within two years of smoking cessation (D1), 14 after 2-5 years (D2), six after 6-9 years (D3), 0 after 10-14 years (D4), three after 15-19 years (D5), six after 20-24 years (D6), and 0 after 25 years (D7). One probe set down-regulated by smoking did not returned to never-smoker levels (D8). For each cluster, the list of probe sets with gene annotation is provided in Supplementary Table S1.

The time course analyses of gene expression were also performed in the two replication sets for the probe sets reproduced across the three populations. It should be noted that because of a lesser number of subjects the number of clusters for the duration of smoking cessation differ across the three populations. Accordingly, a head-to-head comparison with the discovery set is difficult for these analyses. Nonetheless, fast and slow responding genes were significantly concordant between the discovery and UBC sets. This was shown by chi-squared analysis for probe sets dichotomized into early and late reversibility clusters in the two patient groups ($\chi^2=14.4$, p value=0.0001, Supplementary Fig. 2). Interestingly we observed a faster rate of gene expression recovery following smoking cessation in the Groningen data set. Out of 558 probe sets up-regulated by smoking and replicated across the three data sets, only seven probe sets including six known genes were still significantly higher than never-smokers after 10 years of smoking cessation (SERPIND1, AHRR, FASN, PI4K2A, ACSL5, and GANC). Five out of these seven probe sets did not return to normal within 10 years of smoking cessation in the discovery set ($\chi^2=4.3$, p value=0.038, Supplementary Fig. 3). The list of slowly reversible probe sets found in the discovery set and validated in at least one of the replication cohort can be found in Table 3. SERPIND1 was the only known gene overlapping among slowly or never reversible genes in the three populations (Table 3 and Fig. 4). The up-regulation of SERPIND1 in smokers compared to never-smokers was confirmed at the protein level by immunochemistry (Supplementary Fig. 4).
Gene set enrichment analysis (31) was used to test whether slowly reversible genes in the replication sets were enriched among the 599 reproduced genes altered by smoking in the discovery set (see supplementary data for more details). The 599 reproduced probe sets were pre-ranked by their degree of reversibility in the discovery set. Slowly reversible genes in the two replication sets were then tested for enrichment against this pre-ranked list. Supplementary Figs. 5 and 6 show the enrichment plots for both gene sets (i.e. UBC and Groningen). Genes slowly reversible in UBC (n=49 probe sets) cluster among the top slowly reversible genes in the discovery set (FDR q-value=0.002, Supplementary Fig. 5). Although only six known genes were considered slowly reversible in the Groningen set, they also tend to cluster among slowly reversible genes in the discovery set (FDR q-value=0.053, Supplementary Fig. 6).
Discussion

This study provides a comprehensive description of the molecular signature of smoking in the human lung and the effects of smoking cessation on this signature. Thousands of genes were significantly associated with smoking status in the discovery set and the gene expression patterns of never and current-smokers were clearly distinguishable. The impact of smoking was validated for 599 probe sets in two replication data sets. Although most of the differentially expressed genes returned to levels similar to never-smokers following smoking cessation, many slowly reversible genes were identified and some never normalized despite prolonged smoking cessation. Some of the genes that showed enduring alterations in expression have never been reported to be involved in lung cancer, while others are targets of current clinical trials or are used as biomarkers for the progression or prognosis of this disease.

Most whole-genome gene expression studies that have been done to investigate the impact of tobacco smoking were performed on airways, (6-14) alveolar macrophages, (15-17) lung tumors, (7, 18-21) or peripheral leukocytes (22-27). The impact of tobacco smoking on gene expression in non-tumor lung tissue has been evaluated in only a few studies (18, 20, 21). Perhaps the best comparison with our study is the report by Landi et al. (18) who studied the gene expression profile of non-tumor lungs of 15 never, 18 former, and 16 current-smokers. They identified 25 up- and 73 down-regulated genes between never and current-smokers. Sixteen up-regulated (CD1A, CRTAM, CYBB, CYP1B1, DNASE2B, ELF5, FABP3, FGG, IGSF6, ITGAX, KMO, RRAGD, SERPIND1, SPINK5, TM7SF4, TREM2) and two down-regulated (TMSL8 and GHR) genes overlapped with our list of 599 replicated transcripts. When compared to the list of significant transcripts in our discovery set, 24 out of 25 up-regulated genes and 56 out of 73 down-regulated genes identified by Landi et al. (18) overlapped. They also confirmed five up-
regulated genes (CEACAM5, CYPBB, CYLT1, FGG, TM7SF4) by real-time PCR in technical and biological replicates. These five genes were significantly up-regulated by smoking in our discovery set. This study provides strong external validation of our results.

The gene most strongly up-regulated by smoking was the aryl-hydrocarbon receptor repressor (AHRR), which is a negative regulator of the aryl-hydrocarbon receptor (AHR) signaling pathway (32, 33). This pathway, also known as the xenobiotic or dioxin signaling pathway, is an important regulator of cellular responses to exogenous ligands. Cigarette smoke contains agonists of the AHR signaling pathway (34) that in turn regulate the expression of genes encoding cytochrome P450 enzymes (35). In the current study, members of the CYP1 family, including CYP1A1, CYP1A2, and CYP1B1 were among the top genes up-regulated by smoking. These enzymes are essential for detoxification of environmental chemicals, but are also known to generate mutagenic and toxic intermediates that are believed to be carcinogenic (34). Murine models suggest that the AHR pathway is required for carcinogen-induced cancer (36) and that the constitutive activation of this pathway may be responsible for the development of cancer (37, 38). Treatment of mice with benzopyrene, a major carcinogen found in cigarette smoke, resulted in an increase in AHRR expression in the lung (39). In humans, AHRR is ubiquitously expressed, but is more abundant in some tissues including the lung (40). Recent evidence suggests that AHRR is a tumor suppressor gene (41). It is thus tempting to speculate that the up-regulation of AHRR in the lung tissues related to cigarette smoking may be a defence response against possible malignant transformation.

There was large variation in the time since smoking cessation among former-smokers in our data sets. This provided a unique opportunity to differentiate genes that respond favorably from those
that are resistant to smoking cessation. Our data suggest that the smoking-related gene expression signature largely disappears within a few years of smoking cessation. More importantly, the time for gene expression to revert to never-smokers values is gene specific. While groups of genes return to normal within the first years of smoking cessation, others only slowly revert or their expression levels are still higher than in never-smoker after decades of smoking cessation. The latter genes are likely to be important markers to aid in the understanding of the residual risk of lung cancer among former-smokers and may also explain why some individuals with COPD have persistent accelerated lung function decline even after smoking cessation. Particularly relevant for lung cancer is \textit{SERPIND1}, the most enduringly altered gene in our study. This gene encodes a protein that has been used as a biomarker for progression of non-small-cell lung cancer (NCT00155116, ClinicalTrials.gov). \textit{SERPIND1} and the other slowly reversible genes may thus be extremely valuable in improving our understanding of the initial and persistent processes that eventually lead to lung diseases in smokers.

There are certain limitations to the study. Firstly, we did not have longitudinal clinical or gene expression data on these patients. As such, we do not know the robustness of the molecular signature related to smoking within the same individual over time. Moreover, we could not adequately assess possible interactions of smoking and other time-dependent factors such as aging on gene expression. Secondly, the smoking related gene expression pattern could be influenced by a different proportion of cells of different types in the lungs of smokers. There is cellular heterogeneity in lung tissue as in other organs and this influences the gene expression pattern (42). Lung samples contain many cell types including organ specific cells (i.e. pneumocytes), endothelial cells and migratory inflammatory cells, and our analysis represents the amalgam of all these cells. Smoking could influence the relative proportion of cell types in the
lung and since gene expression patterns are cell-type specific this could contribute to the smoking signals that we observed. For example it is well known that the total number of alveolar macrophages is markedly increased in the lungs of smokers (43). We have also shown that inflammation persists in former-smokers and COPD patients years after smoking cessation (44, 45). Accordingly our results must be interpreted with caution and follow-up experiments using specific cell types will be required. Considering the large number of future follow-up experiments that will be required, we believe that sharing our results with the scientific community is likely to accelerate the confirmation of preventive and therapeutic targets. Thirdly, we do not know which of the genes altered by cigarette smoking and smoking cessation are critically important in the genesis of lung cancer. Future work is clearly needed to elucidate the salient pathways. In addition, gene expression is only one aspect of the molecular alterations induced by smoking in the lung (46-48). Multi-dimentional genomic profiling of lung specimens including somatic mutations and epigenetic marks will be required to understand lung carcinogenesis and the residual risk of lung cancer among former-smokers. Finally, the list of slowly reversible genes following smoking cessation is derived from patients with lung cancer. It would be interesting to know whether these genes are also permanently deregulated in current and former-smokers that never develop lung cancer. Identifying genes which do not normalize following smoking cessation in subjects who develop lung cancer compared to those who do not develop lung cancer will be an important step to uncovering the drivers of disease.

In conclusion, we have demonstrated that smoking has a major impact on the lung parenchymal transcriptome and that the altered molecular signature can return to near normal levels over time. However, there are genes affected by smoking which do not normalize or are slow to normalize even after decades of smoking cessation. It also provides molecular evidence of the devastating
health impact of smoking in human lung and reinforces the importance of smoking cessation. The current study is an important step to pinpoint the specific genes and pathways altered by smoking in the lung and provides new candidates that are likely involved in lung cancer and other respiratory diseases.
Disclosure of Potential Conflicts of Interest

Four authors are full time employees of Merck & Co Inc.

Acknowledgments

The authors would like to thank Christine Racine and Sabrina Biardel at the IUCPQ site of the Respiratory Health Network Tissue Bank of the FRSQ for her valuable assistance. We thank Rémy Thériault and the information technology team at the Institut universitaire de cardiologie et de pneumologie de Québec for the secure server-based data repository. We also acknowledge the staff at the Consortium Laval, Université du Québec, McGill and Eastern Quebec (CLUMEQ) for IT support with the high performance computer clusters.

Grant Support

Y. Bossé is a research scholar from the Heart and Stroke Foundation of Canada. D.S. Postma is the holder of a KNAW (Royal Academy of Arts and Sciences) chair for research. D. Sin holds a Canadian Research Chair in COPD and is a Michael Smith Foundation Scholar. A.J. Sandford is the recipient of a Canada Research Chair in genetics and a Michael Smith Foundation for Health Research Senior Scholar Award. P.D. Paré is a Michael Smith Foundation Distinguished Scholar and the Jacob Churg Distinguished Researcher. This study was funded by Merck Research Laboratories. It was also partly supported by the Chaire de pneumologie de la Fondation JD Bégin de l’Université Laval, the Fondation de l’Institut universitaire de cardiologie et de pneumologie de Québec, the Respiratory Health Network of the FRSQ, and the Cancer Research Society.
Contributors

YB, DSP, DDS, WT, PDP, and MLaviolette conceptualized and designed the study. CT, VM, JAT, and GJO coordinated and performed the microarray experiment for all patients. YB performed the statistical analyses, interpreted the results, and wrote the manuscript. PDP, VM, and CT contributed to the analytic strategy. MLamontagne and VW contributed to data analysis. NG performed the qPCR experiment. CC and MLaviolette collected the Laval samples including lung specimens and corresponding clinical data. PDP and JCH collected the UBC samples including lung specimens and corresponding clinical data. DSP and WT collected the Groningen samples including lung specimens and corresponding clinical data. ME, MB, and CAB coordinated sample collection and managed clinical data at their respective institutions. DSP, DDS, CT, VM, GJO, AJS, WT, PDP, and MLaviolette contributed to the clinical and biological interpretation of the data and edited the manuscript. All authors approved the latest version of the manuscript.
References


Table 1. Clinical characteristic of patients in the discovery set that passed microarray quality control filters grouped by smoking status

<table>
<thead>
<tr>
<th></th>
<th>All subjects (n=344)</th>
<th>Never-smokers (n=43)</th>
<th>Former-smokers (n=211)</th>
<th>Current-smokers (n=90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (male:female)</td>
<td>195:149 (56.7% male)</td>
<td>13:30 (30.2% male)</td>
<td>136:75 (64.5% male)</td>
<td>46:44 (51.1% male)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>63.9 ± 10.1</td>
<td>56.6 ± 11.8</td>
<td>65.9 ± 9.1</td>
<td>62.5 ± 9.5</td>
</tr>
<tr>
<td>Body mass index (kg/m²)</td>
<td>27.2 ± 5.3</td>
<td>27.9 ± 6.4</td>
<td>27.7 ± 5.3</td>
<td>25.9 ± 4.6</td>
</tr>
<tr>
<td>FEV₁ % predicted</td>
<td>81.4 ± 18.9</td>
<td>91.3 ± 15.4</td>
<td>82.3 ± 19.9</td>
<td>75.1 ± 15.5</td>
</tr>
<tr>
<td>FVC % predicted</td>
<td>90.4 ± 16.0</td>
<td>95.5 ± 15.2</td>
<td>90.6 ± 16.8</td>
<td>87.7 ± 14.1</td>
</tr>
<tr>
<td>Cardiac diseases</td>
<td>107 (31.1%)</td>
<td>9 (20.9%)</td>
<td>72 (34.1%)</td>
<td>26 (28.9%)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>43 (12.5%)</td>
<td>2 (4.7%)</td>
<td>30 (14.2%)</td>
<td>11 (12.2%)</td>
</tr>
<tr>
<td>COPD</td>
<td>164 (52.6%)</td>
<td>7 (19.4%)</td>
<td>97 (50.5%)</td>
<td>60 (71.4%)</td>
</tr>
<tr>
<td>Asthma</td>
<td>13 (3.8%)</td>
<td>1 (2.3%)</td>
<td>10 (4.7%)</td>
<td>2 (2.2%)</td>
</tr>
<tr>
<td>Primary diagnostic (n)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>adenocarcinoma</td>
<td>191 (55.5%)</td>
<td>25 (58.1%)</td>
<td>115 (54.5%)</td>
<td>51 (56.7%)</td>
</tr>
<tr>
<td>squamous cell carcinoma</td>
<td>95 (27.6%)</td>
<td>2 (4.7%)</td>
<td>63 (29.9%)</td>
<td>30 (33.3%)</td>
</tr>
<tr>
<td>NSCLC other</td>
<td>11 (3.2%)</td>
<td>1 (2.3%)</td>
<td>7 (3.3%)</td>
<td>3 (3.3%)</td>
</tr>
<tr>
<td>carcinoid</td>
<td>17 (4.9%)</td>
<td>7 (16.3%)</td>
<td>9 (4.3%)</td>
<td>1 (1.1%)</td>
</tr>
<tr>
<td>large cell carcinoma</td>
<td>12 (3.5%)</td>
<td>0 (0%)</td>
<td>9 (4.3%)</td>
<td>3 (3.3%)</td>
</tr>
<tr>
<td>small cell lung carcinoma</td>
<td>4 (1.2%)</td>
<td>0 (0%)</td>
<td>2 (0.9%)</td>
<td>2 (2.2%)</td>
</tr>
<tr>
<td>others</td>
<td>14 (4.1%)</td>
<td>8 (18.6%)</td>
<td>6 (2.8%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

Continuous variables are mean ± standard deviation. The numbers of missing values are shown in brackets [ ].
BMI, body mass index; FEV₁, forced expiratory value in one second; FVC, forced vital capacity.
Table 2. Highly replicated probe sets up-regulated by smoking

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>p Laval</th>
<th>p UBC</th>
<th>p GRNG</th>
</tr>
</thead>
<tbody>
<tr>
<td>100123705_TGI_at</td>
<td>PI4K2A</td>
<td>phosphatidylinositol 4-kinase type 2 alpha</td>
<td>3.18E-12</td>
<td>1.51E-05</td>
<td>7.19E-06</td>
</tr>
<tr>
<td>100128259_TGI_at</td>
<td>UBASH3B</td>
<td>ubiquitin associated and SH3 domain containing</td>
<td>1.62E-12</td>
<td>9.17E-06</td>
<td>2.53E-06</td>
</tr>
<tr>
<td>100130524_TGI_at*</td>
<td>NA</td>
<td>NA</td>
<td>3.32E-16</td>
<td>1.62E-12</td>
<td>3.44E-06</td>
</tr>
<tr>
<td>100133335_TGI_at</td>
<td>SERPIND1</td>
<td>serpin peptidase inhibitor, clade D (heparin cofactor), member 1</td>
<td>5.20E-17</td>
<td>9.60E-10</td>
<td>1.82E-06</td>
</tr>
<tr>
<td>100136024_TGI_at</td>
<td>DNASE2B</td>
<td>deoxyribonuclease II beta</td>
<td>2.56E-15</td>
<td>6.92E-07</td>
<td>3.83E-06</td>
</tr>
<tr>
<td>100149468_TGI_at</td>
<td>FUCA1</td>
<td>fucosidase, alpha-L-1, tissue</td>
<td>1.01E-15</td>
<td>5.74E-09</td>
<td>7.99E-06</td>
</tr>
<tr>
<td>100155403_TGI_at</td>
<td>AHRR</td>
<td>aryl-hydrocarbon receptor repressor</td>
<td>3.28E-20</td>
<td>5.03E-12</td>
<td>1.68E-07</td>
</tr>
</tbody>
</table>

*Human mRNAs from GenBank AK056826.
Table 3. Replicated probe sets up-regulated by smoking and not returning to never-smoker levels within 10 years of smoking cessation in the discovery set and validated in at least one of the two replication sets

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>p Laval*</th>
<th>p UBC*</th>
<th>p GRNG*</th>
</tr>
</thead>
<tbody>
<tr>
<td>100123700_TGI_at†</td>
<td>ACSL5</td>
<td>acyl-CoA synthetase long-chain family member 5</td>
<td>6.66E-13</td>
<td>2.05E-02</td>
<td>1.45E-06</td>
</tr>
<tr>
<td>100155403_TGI_at†</td>
<td>AHRR</td>
<td>aryl-hydrocarbon receptor repressor</td>
<td>3.28E-20</td>
<td>5.03E-12</td>
<td>1.68E-07</td>
</tr>
<tr>
<td>100159958_TGI_at</td>
<td>AMICA1</td>
<td>adhesion molecule, interacts with CXADR antigen 1</td>
<td>5.31E-15</td>
<td>7.36E-07</td>
<td>6.07E-03</td>
</tr>
<tr>
<td>100312275_TGI_at</td>
<td>AMICA1</td>
<td>adhesion molecule, interacts with CXADR antigen 1</td>
<td>9.12E-15</td>
<td>1.01E-06</td>
<td>8.24E-03</td>
</tr>
<tr>
<td>100154796_TGI_at</td>
<td>AMICA1</td>
<td>adhesion molecule, interacts with CXADR antigen 1</td>
<td>4.54E-15</td>
<td>2.02E-06</td>
<td>6.22E-03</td>
</tr>
<tr>
<td>100307462_TGI_at</td>
<td>ATP13A4</td>
<td>ATPase type 13A4</td>
<td>3.18E-17</td>
<td>2.67E-07</td>
<td>3.58E-05</td>
</tr>
<tr>
<td>100150912_TGI_at</td>
<td>C2orf58</td>
<td>chromosome 2 open reading frame 58</td>
<td>3.48E-17</td>
<td>3.63E-10</td>
<td>1.89E-05</td>
</tr>
<tr>
<td>100153378_TGI_at</td>
<td>CD163L1</td>
<td>CD163 molecule-like 1</td>
<td>2.23E-14</td>
<td>2.72E-06</td>
<td>6.07E-03</td>
</tr>
<tr>
<td>100130932_TGI_at</td>
<td>CD86</td>
<td>CD86 molecule</td>
<td>6.29E-14</td>
<td>1.10E-04</td>
<td>8.71E-03</td>
</tr>
<tr>
<td>100132353_TGI_at</td>
<td>CLIP4</td>
<td>CAP-GLY domain containing linker protein family, member 4</td>
<td>7.06E-16</td>
<td>4.37E-06</td>
<td>3.21E-04</td>
</tr>
<tr>
<td>100125484_TGI_at</td>
<td>CYP1B1</td>
<td>cytochrome P450, family 1, subfamily B, polypeptide 1</td>
<td>1.14E-19</td>
<td>3.33E-10</td>
<td>2.81E-03</td>
</tr>
<tr>
<td>100131143_TGI_at</td>
<td>CYP1B1</td>
<td>cytochrome P450, family 1, subfamily B, polypeptide 1</td>
<td>6.12E-20</td>
<td>3.30E-11</td>
<td>3.37E-03</td>
</tr>
<tr>
<td>100303658_TGI_at</td>
<td>CYP1B1</td>
<td>cytochrome P450, family 1, subfamily B, polypeptide 1</td>
<td>6.69E-20</td>
<td>4.16E-10</td>
<td>1.70E-03</td>
</tr>
<tr>
<td>100134409_TGI_at</td>
<td>DNAJC5B</td>
<td>DnaJ (Hsp40) homolog, subfamily C, member 5 beta</td>
<td>2.32E-14</td>
<td>2.17E-06</td>
<td>4.67E-05</td>
</tr>
<tr>
<td>100309944_TGI_at†</td>
<td>FASN</td>
<td>fatty acid synthase</td>
<td>4.29E-11</td>
<td>5.18E-03</td>
<td>2.21E-05</td>
</tr>
<tr>
<td>100136438_TGI_at</td>
<td>GPR110</td>
<td>G protein-coupled receptor 110</td>
<td>3.62E-18</td>
<td>1.07E-06</td>
<td>1.37E-03</td>
</tr>
<tr>
<td>100302763_TGI_at</td>
<td>HABP2</td>
<td>hyaluronan binding protein 2</td>
<td>1.13E-15</td>
<td>3.53E-08</td>
<td>1.43E-03</td>
</tr>
<tr>
<td>100305401_TGI_at</td>
<td>ITPR2</td>
<td>inositol 1,4,5-triphosphate receptor,</td>
<td>6.79E-15</td>
<td>6.37E-05</td>
<td>5.32E-04</td>
</tr>
<tr>
<td>Probe Set</td>
<td>Symbol</td>
<td>Description</td>
<td>Fold Change</td>
<td>p-value</td>
<td>Corrected p-value</td>
</tr>
<tr>
<td>-----------</td>
<td>--------</td>
<td>-------------</td>
<td>-------------</td>
<td>---------</td>
<td>------------------</td>
</tr>
<tr>
<td>100122391_TGI_at</td>
<td>MT1F</td>
<td>metallothionein 1F</td>
<td>1.35E-09</td>
<td>1.98E-05</td>
<td>5.14E-03</td>
</tr>
<tr>
<td>100162545_TGI_at</td>
<td>PLA2G4E</td>
<td>phospholipase A2, group IVE</td>
<td>1.28E-17</td>
<td>1.80E-11</td>
<td>2.13E-05</td>
</tr>
<tr>
<td>100121991_TGI_at</td>
<td>PLAUR</td>
<td>plasminogen activator, urokinase receptor</td>
<td>1.13E-10</td>
<td>2.96E-04</td>
<td>1.20E-02</td>
</tr>
<tr>
<td>100303500_TGI_at</td>
<td>PLAUR</td>
<td>plasminogen activator, urokinase receptor</td>
<td>1.76E-11</td>
<td>1.99E-04</td>
<td>4.34E-03</td>
</tr>
<tr>
<td>100138457_TGI_at</td>
<td>PLAUR</td>
<td>plasminogen activator, urokinase receptor</td>
<td>5.52E-15</td>
<td>1.80E-05</td>
<td>1.38E-04</td>
</tr>
<tr>
<td>100300362_TGI_at</td>
<td>PLAUR</td>
<td>plasminogen activator, urokinase receptor</td>
<td>1.53E-14</td>
<td>9.17E-06</td>
<td>1.61E-04</td>
</tr>
<tr>
<td>100143476_TGI_at</td>
<td>PON1</td>
<td>paraoxonase 1</td>
<td>4.09E-08</td>
<td>1.16E-05</td>
<td>1.37E-03</td>
</tr>
<tr>
<td>100133335_TGI_at</td>
<td>SERPIND1</td>
<td>serpin peptidase inhibitor, clade D (heparin cofactor), member 1</td>
<td>5.20E-17</td>
<td>9.60E-10</td>
<td>1.82E-06</td>
</tr>
<tr>
<td>100144500_TGI_at</td>
<td>SFN</td>
<td>stratifin</td>
<td>6.77E-10</td>
<td>4.41E-04</td>
<td>2.41E-03</td>
</tr>
<tr>
<td>100160780_TGI_at</td>
<td>SPINK5</td>
<td>serine peptidase inhibitor, Kazal type 5</td>
<td>3.81E-15</td>
<td>9.92E-06</td>
<td>3.21E-04</td>
</tr>
<tr>
<td>100139182_TGI_at</td>
<td>similar to phosphodiesterase 4D, cAMP specific</td>
<td>2.34E-11</td>
<td>1.22E-05</td>
<td>3.49E-05</td>
<td></td>
</tr>
<tr>
<td>100156798_TGI_at</td>
<td></td>
<td></td>
<td>2.12E-12</td>
<td>2.09E-08</td>
<td>1.20E-03</td>
</tr>
</tbody>
</table>

*p values for the discovery set (Laval) and the two replication sets (UBC and Groningen). P values are derived from a Wilcoxon test comparing current-smokers to never-smokers.
†Probe sets not returning to normal within 10 years of smoking cessation in the discovery and Groningen sets.
Probe sets in bold are those not returning to normal within 10 years of smoking cessation in the three populations.
‡Human mRNAs from GenBank BQ225423.
§Human mRNAs from GenBank BX648604.
||Human mRNAs from GenBank AK056826.
Figure legends

Figure 1. The molecular signature of smoking in the discovery set. The samples (n=342) and probe sets (n=3,223) are illustrated in columns and rows, respectively. Yellow indicates a high level of expression; blue indicates a low level of expression. Never-smokers are on the left and current-smokers on the right. The vertical white lines separate the three smoking groups. The former-smokers are ordered based on the number of years since cessation of smoking (upper panel). A clear dichotomy of gene expression pattern is observed between never and current-smokers.

Figure 2. Confirmation of AHRR, CYP1B1, and SERPIND1 by quantitative real-time PCR. Panels A, B, and C show the qPCR results for AHRR, CYP1B1, and SERPIND1, respectively. Results are derived from 10 never-smokers, 10 former-smokers, and 10 current-smokers. P values are derived from Kruskal-Wallis tests comparing the number of copies in the three smoking groups, relative to GAPDH. Bars are means ± standard errors. NS, never-smokers; FS, former-smokers; CS, current-smokers.

Figure 3. Expression of the AHRR gene in smoking groups. The y-axis represents mean residual gene expression values. The average and standard error of gene expression values for current-smokers (“S”, n=90) and never-smokers (“NS”, n=43) are shown at the two extremes of the x-axis. Former-smokers were grouped by years of smoking cessation in a way that ensured at least 25 individuals per bin (n for each bin=26, 29, 32, 26, 31, 28, and 37). P values are Wilcoxon tests comparing current-smokers and each group of former-smokers to never-smokers.
Figure 4. Expression of SERPIND1 in the three populations by smoking groups and duration of smoking cessation among former-smokers. Dots are means ± standard errors. S, current-smokers; NS, never-smokers. In the three panels, former-smokers were grouped based on the duration of smoking cessation (see Methods).
Molecular Signature of Smoking in Human Lung Tissues

Yohan Bosse, Dirkje S. Postma, Don D Sin, et al.

Cancer Res  Published OnlineFirst June 1, 2012.

Updated version  Access the most recent version of this article at: doi:10.1158/0008-5472.CAN-12-1160

Supplementary Material  Access the most recent supplemental material at: http://cancerres.aacrjournals.org/content/suppl/2012/06/01/0008-5472.CAN-12-1160.DC1

Author Manuscript  Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.