Supplementary data

Title:
Molecular signature of smoking in human lung tissues

Authors:

Yohan Bossé, Dirkje S. Postma, Don D. Sin, Maxime Lamontagne, Christian Couture, Nathalie Gaudreault, Philippe Joubert, Vivien Wong, Mark Elliott, Maarten van den Berge, Corry A. Brandsma, Catherine Tribouley, Vladislav Malkov, Jeffrey A. Tsou, Gregory J. Opiteck, James C. Hogg, Andrew J. Sandford, Wim Timens, Peter D. Paré, Michel Laviolette

Methods

The discovery set

Lung tissue samples

Preoperatively, patients underwent pulmonary function testing in which lung volumes, forced expiratory volume in 1 sec (FEV_1), forced vital capacity (FVC), and diffusion capacity for carbon monoxide (D_{Lco}) were determined according to the American Thoracic Society guidelines (1, 2). FEV_1 and FVC values were used to define COPD in accordance with the recommendations of the Global initiative for chronic Obstructive Lung Disease (GOLD) (3). Patient's medical charts were abstracted for co-morbidities including asthma, cardiac diseases and type II diabetes.

RNA extraction

500 pieces of lung tissues free of tumor (~200 mg) were sent to Rosetta Inpharmatics for RNA isolation. Total RNA was extracted using the SV96 Total RNA Isolation System (Promega) according to standard protocols employed at Rosetta Gene Expression Laboratory. The integrity of the RNA was confirmed using the Bioanalyzer 2100 (Agilent). Samples with either RIN < 5 or 28s/18s < 0.75 were discarded. Nine samples failed RNA quality control and were replaced by lung specimens derived from nine other patients. A total of 500 RNAs were of sufficient quality for microarray analysis.

Microarrays

Expression profiling was performed using an Affymetrix custom array designed by Rosetta Inpharmatics which tested 51,562 probe sets. The accuracy of sample processing was monitored by sequential quality checks throughout the Affymetrix protocol. A total of 21 samples failed quality control. Arrays were scanned on GeneChip Scanner 3000 to acquire DAT file images. The Affymetrix GeneChip Operating Software (GCOS) was used to generate .CEL files. Expression values were extracted using the Robust Multichip Average (RMA) method (4) as implemented in the Affymetrix Power Tools (APT) software. The quality of the arrays was judged using standard quality control parameters (5). A total of 4 arrays were excluded based on quality control filters, leaving 475 unique samples for analyses. The complete data set, including RMA expression values and raw .CEL files, has been deposited in the National Center for Biotechnology Information's Gene Expression Omnibus (6) and is accessible through GEO Series accession number GSE23546 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=xbobfamguyoewze&acc=GSE23546).

Quantitative real-time PCR (qPCR)

Thirty subjects were selected for validation by qPCR including 10 subjects in each smoking group (i.e. never, former and current-smokers). Never- and current-smokers were selected randomly. Because of the influence of the duration of smoking cessation on gene expression, former-smokers were selected using a random-stratified sampling method. They were first sorted by the number of years of smoking cessation and then one subject per bin size of 32 was randomly selected. RNA for qPCR validation was isolated from 30 mg of frozen non-neoplastic pulmonary parenchyma using the RNeasy® Mini Kit (QIAGEN, Mississauga, Ontario) according to manufacturer's instructions. RNA concentration and purity was measured by the Synergy HT Multi-Mode Microplate Reader (BioTek®, Winooski, USA). Three genes significantly associated with smoking in the discovery set were evaluated including *AHRR*, *SERPIND1* and *CYP1B1*. The QuantiTect Reverse Transcription Kit (QIAGEN) was used to synthesize cDNA from 2 μg of

RNA from each sample as described by the manufacturer. *GAPDH* was utilized as a reference gene (7). The primers were designed using the software Primer3 v.0.4.0 (http://frodo.wi.mit.edu/primer3) and synthesized by Integrated DNA Technologies (Toronto, Ontario). The genes, forward (F) and reverse (R) primer sequences used for qPCR were *GAPDH* (F: 5’- ATGTTCGTCATGGGTGTGAA and R: 5’-GGTGCTAAGCAGTTGGTGGT), *AHRR* (F: 5’ AACTTATATTTTTGCAGTTTCTACTGG- and R: 5’-AGCAGTAGAGAAAGTTGCATTTA), *SERPIND1* (F: 5’- GACCTGTTCAAGCACCAAGG and R: 5’- GTCGACAGTGAAGCGGACTT), *CYP1B1* (F: 5’-CGGCTGGATTTGGAGAACGTA and R: 5’- TGATCCAATTCTGCCTGCACT). The lengths of the amplicons were between 89 and 146 bp. The same cDNA sample was used to prepare the standard curves for each gene and was made from a pool of 30 cDNA samples. For each gene, the samples were tested in triplicate using the Rotor-Gene 6000 (Corbett Life Science, Concorde, Australia) in a final reaction volume of 20 µl containing 5 µl of 100X diluted cDNA, 10µl of 2X QuantiTect SYBR Green PCR Kit (QIAGEN), and 0.3 µM of each primer. The final concentration of $MgCl_2$ was 3 mM for *AHRR* and 2.5 mM for the other genes. The qPCR conditions were the same for all genes, 95°C for 15 min, and then 45 cycles at 94°C for 15 sec, annealing temperature for 30 sec and 72°C for 30 sec. The annealing temperatures used were 59°C for *GAPDH*, 53 °C for *AHRR*, and 58 °C for *SERPIND1* and *CYP1B1*. A melting curve analysis was performed at the end of each run and all showed a single peak, indicating specificity of the amplified products. For each gene, the quantification cycle (Cq) of three replicates was averaged and normalized to *GAPDH* standard curve using the Rotor-gene 6000 series software, and the mRNA levels were expressed as the absolute number of copies.

Immunohistochemistry

Twenty subjects were selected for immunohistochemistry including 10 never-smokers and 10 current-smokers. These were the same subjects used in the qPCR validation. Immunohistochemistry was performed using Dako EnVision®+ System-HRP (DAB) together with Dako’s Autostainer Link 48 instrument. Deparaffinization, rehydration and heat-induced epitope retrieval was performed on 4 mm thick formalin fixed paraffin-embedded sections using EnVision™ FLEX, High pH solution for 20 minutes at 97°C. Slides were immediately rinsed and incubated with EnVision™ FLEX Peroxidase Blocking Reagent for 5 minutes followed by 30 minutes incubation with monoclonal mouse anti-human SERPIND1 antibody (Santa Cruz Biotechnology, Inc., CA, USA) diluted 1:50. Slides were then rinse, incubated 30 minutes with EnVision™ FLEX horseradish peroxidase-conjugated mouse antibody, stained with diaminobenzidine reagent for 10 minutes and finally counterstained with DakoCytomation Maye’s Hematoxylin (Lillie’s modification) Histological Staining reagent.

Gene set enrichment analysis (GSEA)

GSEA (8) was used to test the overlap among gene slowly reversible between the discovery set and the replication sets. The 599 probe sets significantly altered by smoking in the three data sets were ranked by their elapse time to revert to never smoker levels in the discovery set. Probe sets not returning to normal were on the top of the list, while fast returning probe sets were at the bottom. Slowly reversible genes (i.e. those not returning to never-smokers levels before 10 years of smoking cessation) in the two replication sets were then tested against this pre-ranked list of genes. 49 and 6 slowly reversible probe sets with known gene symbols were included in the UBC and Groningen gene sets, respectively. These two gene sets were tested using default analysis options in GSEA after adjusting the minimum size gene set parameter for allowing smaller gene sets to be considered and setting the collapse dataset to gene symbols to false in order to consider identifiers as they are in the gene set file (.gmt) and the ranked list (.rnk).

References

1. Standardization of Spirometry, 1994 Update. American Thoracic Society. Am J Respir Crit Care Med. 1995;152:1107-36.
2. American Thoracic Society. Single-breath carbon monoxide diffusing capacity (transfer factor). Recommendations for a standard technique--1995 update. Am J Respir Crit Care Med. 1995;152:2185-98.
3. Rabe KF, Hurd S, Anzueto A, Barnes PJ, Buist SA, Calverley P, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. Am J Respir Crit Care Med. 2007;176:532-55.
4. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4:249-64.
5. Heber S, Sick B. Quality assessment of Affymetrix GeneChip data. OMICS. 2006;10:358-68.
6. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009;37:D885-90.
7. Gorzelniak K, Janke J, Engeli S, Sharma AM. Validation of endogenous controls for gene expression studies in human adipocytes and preadipocytes. Horm Metab Res. 2001;33:625-7.
8. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545-50.

Supplementary Table S2. Clinical characteristic of patients in the first replication set (UBC) that passed microarray quality control filters grouped by smoking status

	All subjects (n=285)	Never-smokers (n=30)	Former-smokers (n=158)	Current-smokers (n=97)
Gender (male:female)	153:132 (53.7% male)	13:17 (43.3% male)	85:73 (53.4% male)	55:42 (56.7% male)
Age (years)	62.1 ± 12.7 [0]	53.2 ± 21.9 [0]	64.3 ± 10.1 [0]	61.4 ± 11.5 [0]
Body mass index (kg/m^2)	25.6 ± 5.3 [3]	24.6 ± 4.8 [0]	26.0 ± 5.4 [1]	25.2 ± 5.4 [2]
FEV$_1$ % predicted	78.9 ± 22.9 [23]	86.2 ± 23.3 [10]	78.7 ± 25.0 [5]	77.6 ± 18.6 [8]
FVC % predicted	87.1 ± 19.0 [21]	87.7 ± 20.0 [9]	86.8 ± 20.4 [4]	87.6 ± 16.2 [8]
Cardiac diseases	44 (20.9%) [74]	4 (16.7%) [6]	29 (22.7%) [30]	11 (18.6%) [38]
Diabetes	13 (26.5%) [236]	2 (22.2%) [21]	10 (38.5%) [132]	1 (7.1%) [83]
COPD	113 (46.3%) [41]	3 (15.8%) [11]	62 (43.4%) [15]	48 (58.5%) [15]
Asthma	21 (10.2%) [79]	3 (13.6%) [8]	8 (6.3%) [32]	10 (17.2%) [39]
Primary diagnostic (n)				
adenocarcinoma	86 (30.2%)	7 (23.3%)	54 (34.2%)	25 (25.8%)
squamous cell carcinoma	83 (29.1%)	1 (3.3%)	50 (31.6%)	32 (33.0%)
NSCLC other	7 (2.5%)	0 (0.0%)	3 (1.9%)	4 (4.1%)
carcinoid	15 (5.3%)	5 (16.7%)	6 (3.8%)	4 (4.1%)
large cell carcinoma	20 (7.0%)	0 (0.0%)	10 (6.3%)	10 (10.3%)
small cell lung carcinoma	26 (9.1%)	3 (10.0%)	12 (7.6%)	11 (11.3%)
others	48 (16.8%)	14 (46.7%)	23 (14.6%)	11 (11.3%)

Values are mean ± standard deviation. The numbers of missing values are shown in brackets []. BMI, body mass index; FEV$_1$, forced expiratory value in one second; FVC, forced vital capacity.

Supplementary Table S3. Clinical characteristic of patients in the second replication set (Groningen) that passed microarray quality control filters grouped by smoking status

	All subjects (n=224)	Never-smokers (n=16)	Former-smokers (n=164)	Current-smokers (n=44)
Gender (male:female)	116:108 (51.8% male)	8:8 (50.0% male)	86:78 (52.4% male)	22:22 (50.0% male)
Age (years)	57.4 ± 9.7 [0]	54.9 ± 12.4 [0]	57.4 ± 9.7 [0]	58.4 ± 8.9 [0]
Body mass index (kg/m^2)	23.9 ± 4.0 [0]	23.5 ± 4.0 [0]	23.7 ± 3.9 [0]	24.4 ± 4.5 [0]
FEV$_1$ % predicted	70.1 ± 29.3 [111]	75.9 ± 31.3 [6]	65.8 ± 32.8 [99]	75.8 ± 20.6 [6]
FVC % predicted	86.6 ± 19.6 [120]	85.7 ± 18.0 [7]	85.0 ± 21.8 [106]	89.4 ± 16.1 [7]
Cardiac diseases	15 (6.7%) [0]	1 (6.3%) [0]	9 (5.5%) [0]	5 (11.4%) [0]
Diabetes	11 (4.9%) [0]	0 (0.0%) [0]	9 (5.5%) [0]	2 (4.5%) [0]
COPD	139 (74.3%) [37]	7 (53.8%) [3]	106 (80.9%) [32]	26 (60.5%) [1]
Asthma	NA [224]	NA [16]	NA [164]	NA [44]
Primary diagnostic (n)				
adenocarcinoma	32 (14.3%)	3 (18.8%)	18 (11.0%)	11 (25.0%)
squamous cell carcinoma	41 (18.3%)	0 (0.0%)	25 (15.2%)	16 (36.4%)
NSCLC other	3 (1.3%)	0 (0.0%)	2 (1.2%)	1 (2.3%)
carcinoid	1 (0.4%)	1 (6.3%)	0 (0.0%)	0 (0.0%)
large cell carcinoma	15 (6.7%)	1 (6.3%)	7 (4.3%)	7 (15.9%)
small cell lung carcinoma	1 (0.4%)	0 (0.0%)	0 (0.0%)	1 (2.3%)
others	131 (58.5%)	11 (68.8%)	112 (68.3%)*	8 (18.2%)

Values are mean ± standard deviation. The numbers of missing values are shown in brackets []. BMI, body mass index; FEV$_1$, forced expiratory value in one second; FVC, forced vital capacity.
*40 are alpha-1 antitrypsin deficiency and 68 are COPD.

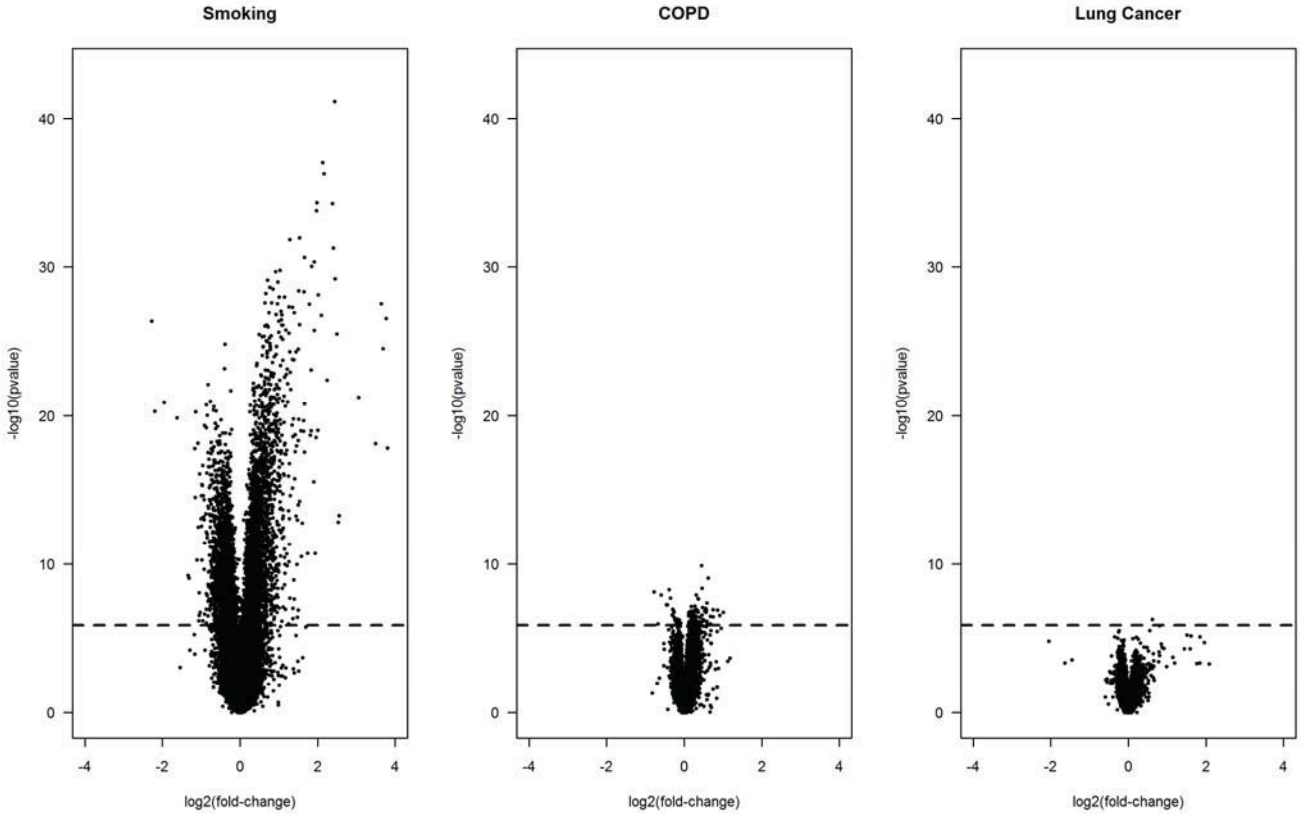
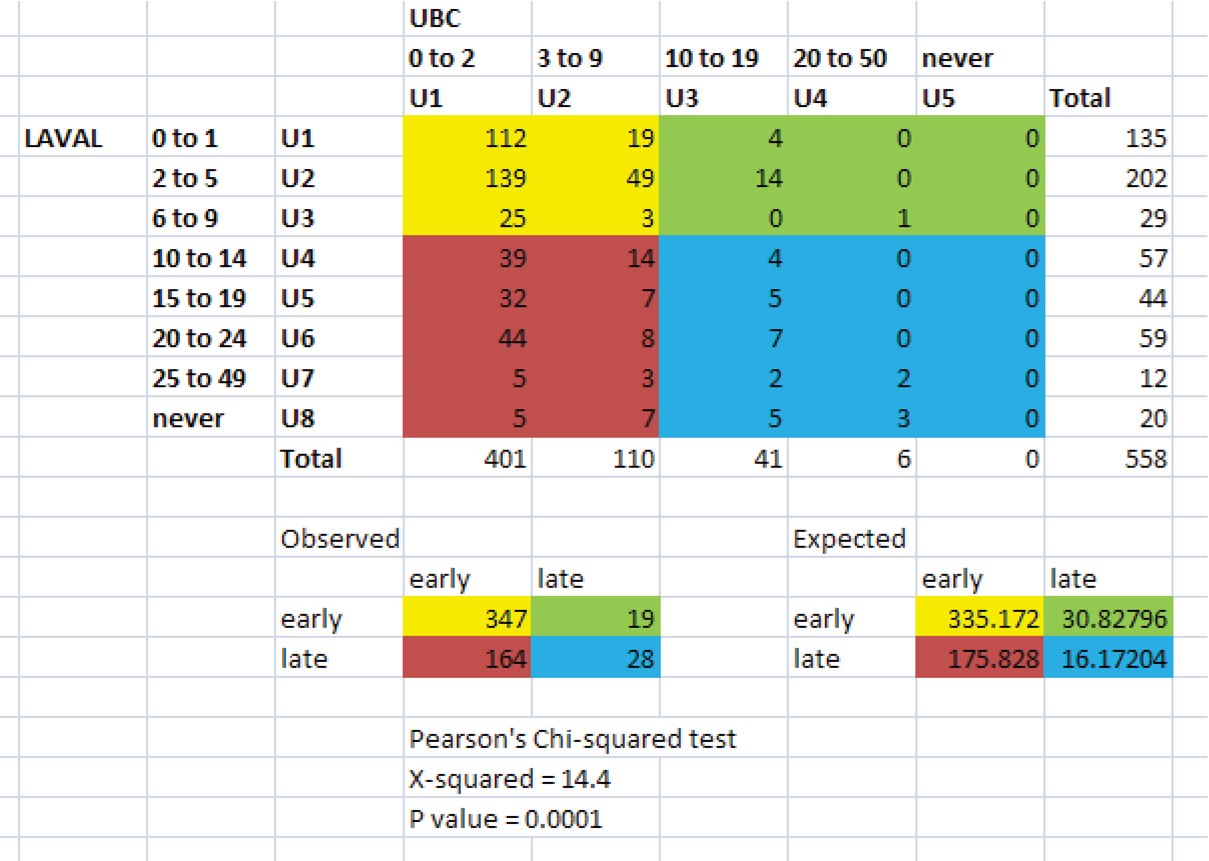


Figure S1. Volcano plots showing the impact of smoking (left), COPD (middle) and lung cancer (right) on gene expression in the lung in the discovery set. The three panels are drawn using the same scale for ease of comparison. The x- and y-axes represent fold changes and –log10 p values, respectively. The horizontal dashed line represents the Bonferroni correction threshold (0.05/38,820 = 1.29 x 10^{-6}). Raw RMA expression values were used in the analyses to assess the impact of smoking, COPD and lung cancer on gene expression (i.e. for these analyses, gene expression was not adjusted for covariates). The left panel shows the impact of smoking status (three group comparison: never, former and current smokers) on gene expression. The 344 subjects presented in the manuscript are considered in this analysis. Fold-changes are obtained by comparing current smokers to never smokers (former smokers were not considered in the calculation of fold-change). The middle panel shows the impact of COPD. Gene expression of 164 patients with COPD was compared with 148 patients without COPD defined by GOLD criteria. The right panel shows the impact of lung cancer on gene expression. Only the two major lung cancer subtypes were compared including 191 patients with adenocarcinoma and 95 patients with squamous cell carcinoma.



			UBC					
			0 to 2	3 to 9	10 to 19	20 to 50	never	
			U1	U2	U3	U4	U5	Total
LAVAL	0 to 1	U1	112	19	4	0	0	135
	2 to 5	U2	139	49	14	0	0	202
	6 to 9	U3	25	3	0	1	0	29
	10 to 14	U4	39	14	4	0	0	57
	15 to 19	U5	32	7	5	0	0	44
	20 to 24	U6	44	8	7	0	0	59
	25 to 49	U7	5	3	2	2	0	12
	never	U8	5	7	5	3	0	20
		Total	401	110	41	6	0	558

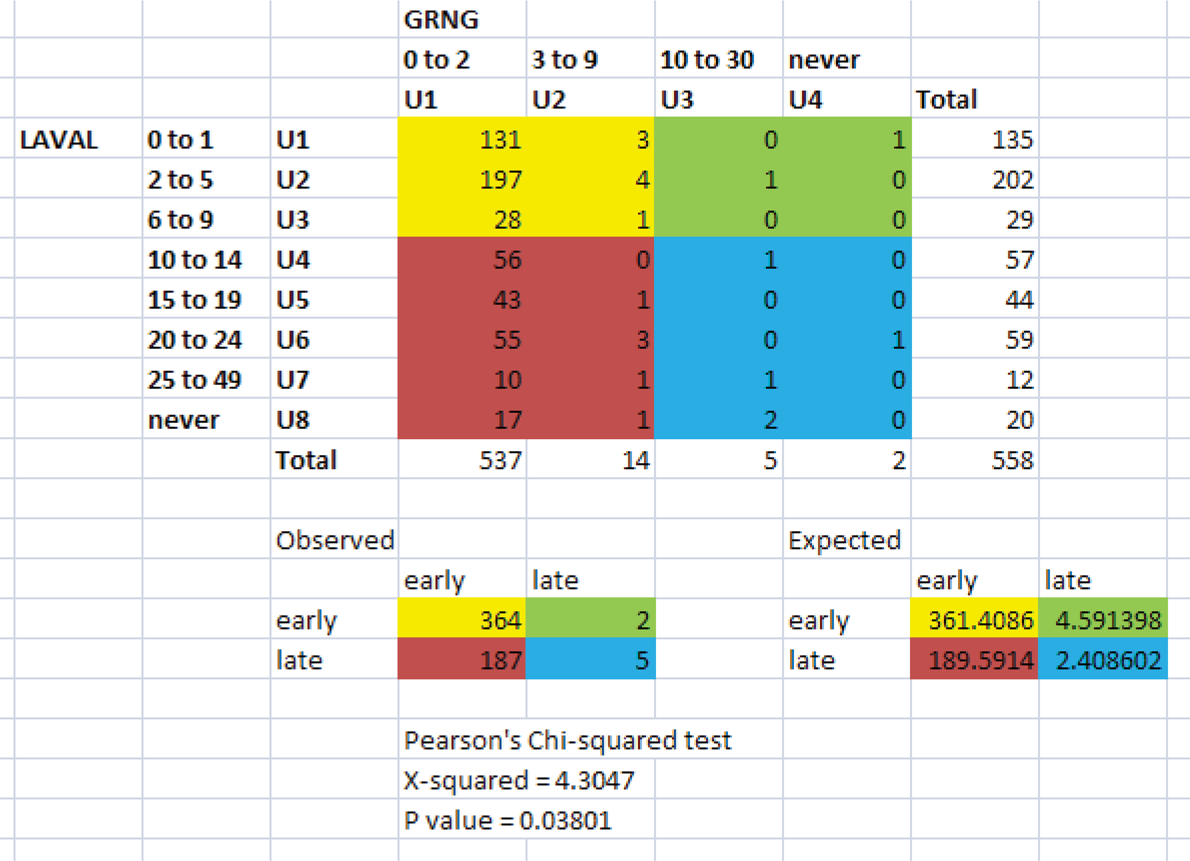
Observed				Expected		
	early	late			early	late
early	347	19		early	335.172	30.82796
late	164	28		late	175.828	16.17204

Pearson's Chi-squared test

X-squared = 14.4

P value = 0.0001

Figure S2. Comparison of the time for gene expression recovery following smoking cessation between the discovery set (Laval) and UBC. Only the 558 probe sets up-regulated by smoking and replicated across the three data sets were considered. Slowly or never reversible genes in Laval (U4-U8) were more likely to be found among the latest clusters (U3-U5) in UBC (blue area). The specific probe sets found in the Laval and UBC clusters are provided in Supplementary Table 1. Table 3 provides the list of the 28 slow responding probe sets found in both populations (blue area).



			GRNG					
			0 to 2	3 to 9	10 to 30	never		
			U1	U2	U3	U4	Total	
LAVAL	0 to 1	U1	131	3	0	1	135	
	2 to 5	U2	197	4	1	0	202	
	6 to 9	U3	28	1	0	0	29	
	10 to 14	U4	56	0	1	0	57	
	15 to 19	U5	43	1	0	0	44	
	20 to 24	U6	55	3	0	1	59	
	25 to 49	U7	10	1	1	0	12	
	never	U8	17	1	2	0	20	
		Total	537	14	5	2	558	
		Observed				Expected		
			early	late			early	late
		early	364	2		early	361.4086	4.591398
		late	187	5		late	189.5914	2.408602
			Pearson's Chi-squared test					
			X-squared = 4.3047					
			P value = 0.03801					

Figure S3. Comparison of the time for gene expression recovery following smoking cessation between the discovery set (Laval) and Groningen. Only the 558 probe sets up-regulated by smoking and replicated across the three data sets were considered. Slowly or never reversible genes in Laval (U4-U8) were more likely to be found among the latest clusters (U3-U4) in Groningen (blue area). The specific probe sets found in the Laval and Groningen clusters are provided in Supplementary Table 1. Table 3 provides the list of the 5 slow responding probe sets found in both populations (blue area).

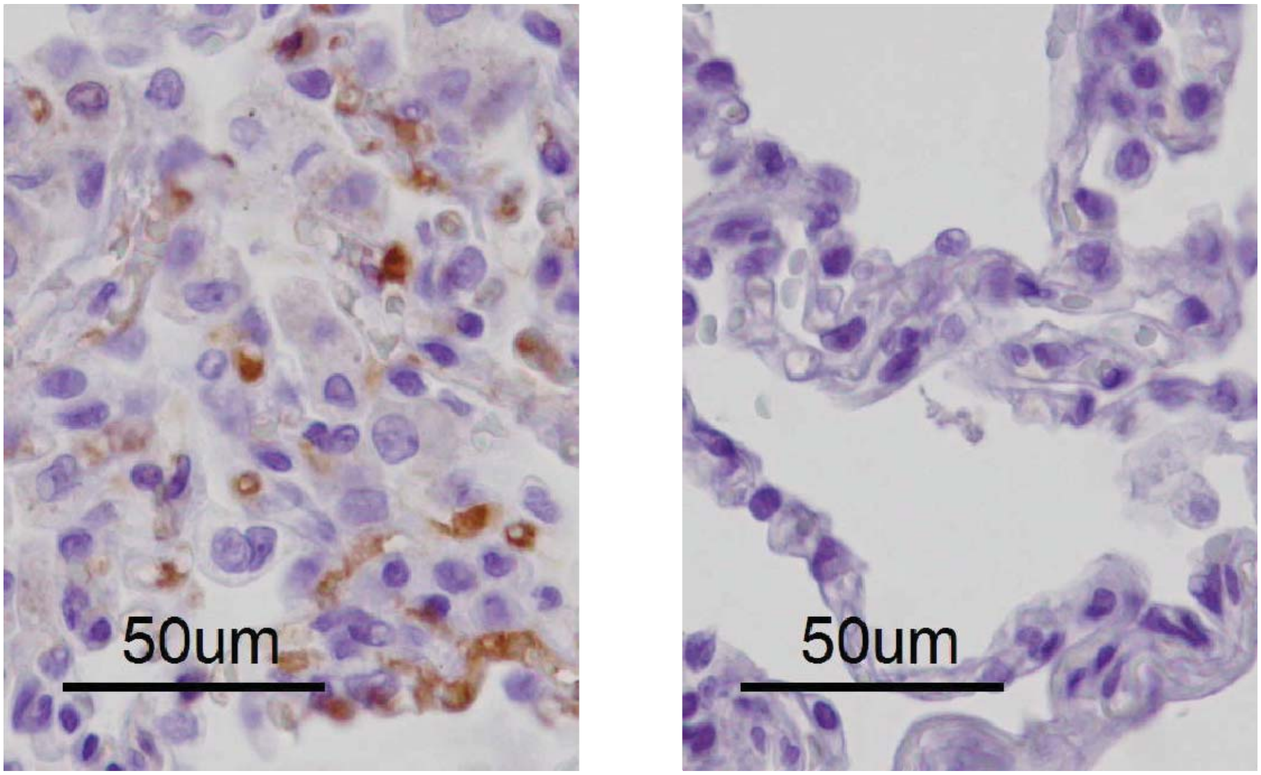


Figure S4. Representative images of serpin peptidase inhibitor clade D member 1 (SERPIND1) immunohistochemistry in a smoker (left panel) and a never-smoker (right panel) lung parenchyma. Alveolar capillary endothelial staining is apparent in smoker.

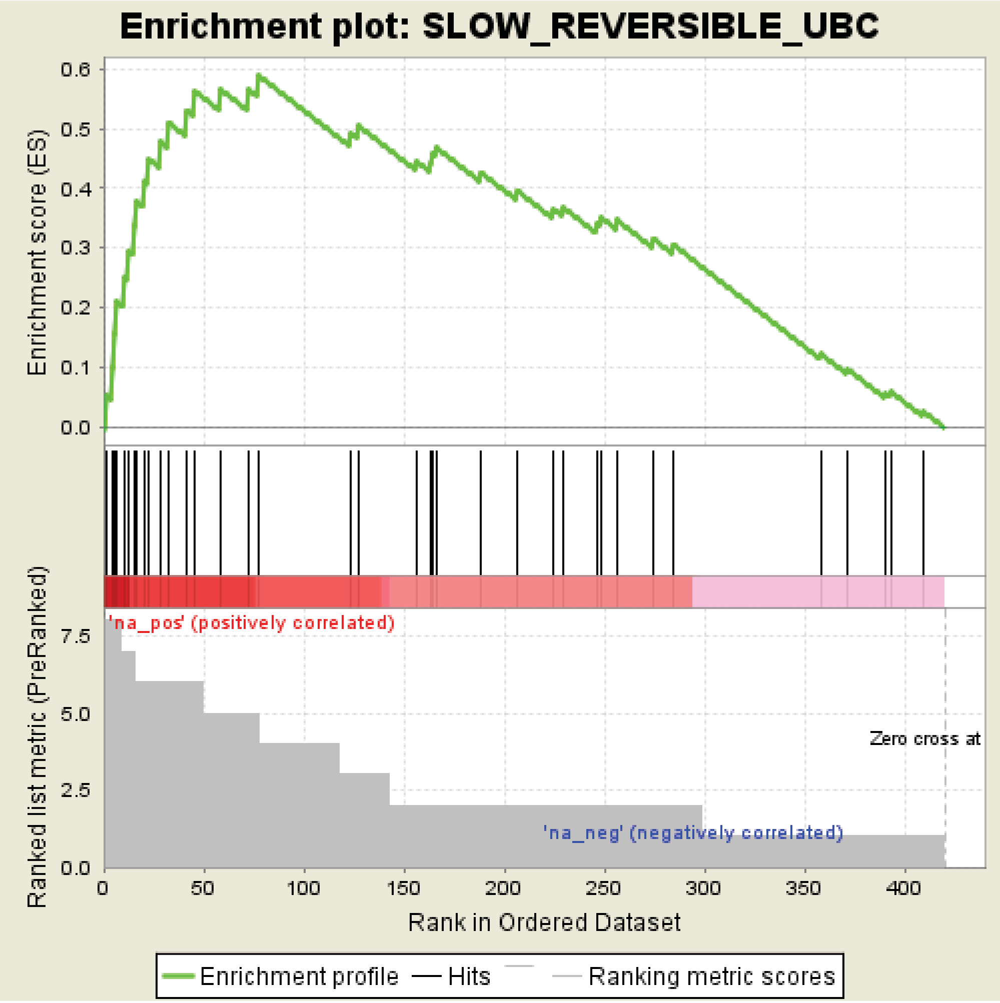


Figure S5. Enrichment plot for slowly reversible gene in UBC. Slowly reversible genes in UBC (n=49) were tested for enrichment against the 599 reproduced probe sets pre-ranked by their degree of reversibility following smoking cessation in the discovery set.

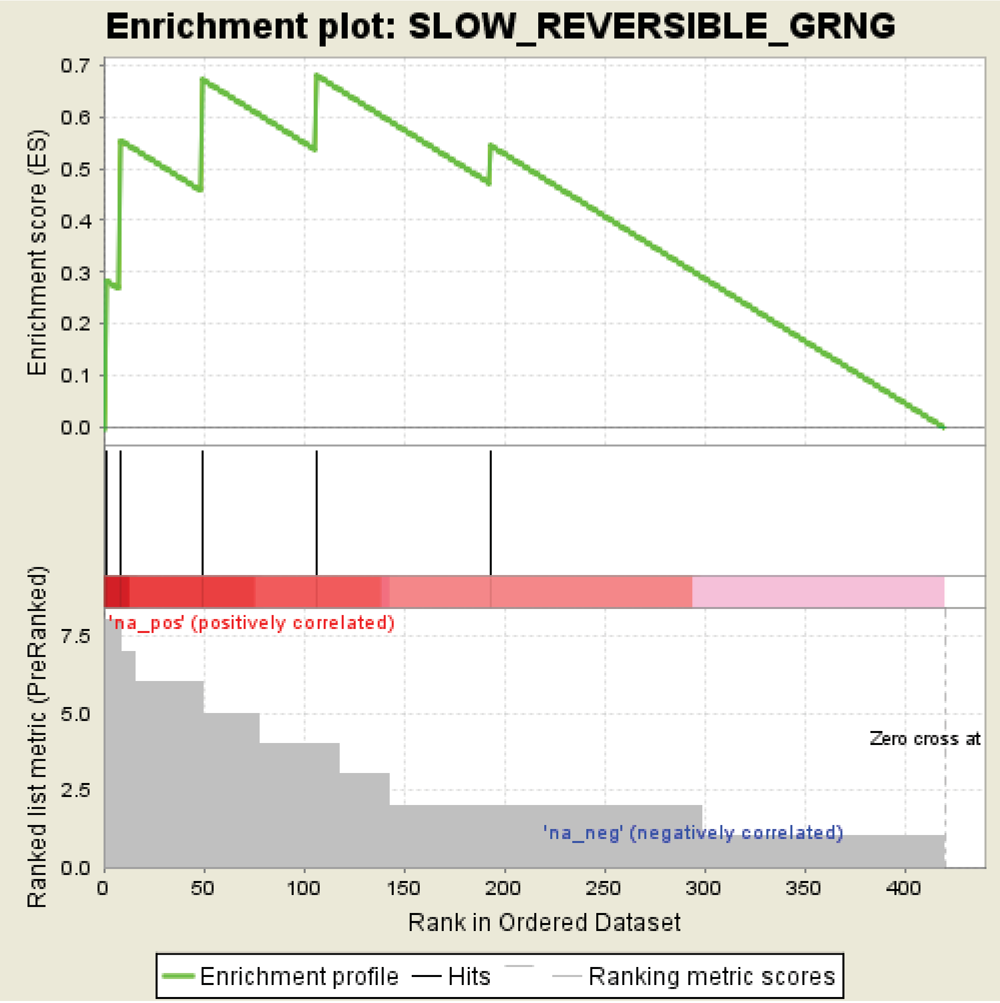


Figure S6. Enrichment plot for slowly reversible gene in Groningen. Slowly reversible genes in Groningen (n=6) were tested for enrichment against the 599 reproduced probe sets pre-ranked by their degree of reversibility following smoking cessation in the discovery set.