Supplementary Materials and Methods

Inclusion and Exclusion Criteria for Healthy Nonsmokers, Healthy Smokers, and COPD Smokers

Healthy nonsmokers

Inclusion criteria

- Males and females, at least 18 yr old
- Provide informed consent
- Good health without history of chronic lung disease, including asthma, and without recurrent or recent (within 3 months) acute pulmonary disease
- Normal physical examination
- Normal routine laboratory evaluation, including general hematologic studies, general serologic / immunologic studies, general biochemical analyses, and urine analysis
- HIV1 negative
- α1-antitrypsin level normal
- Normal PA and lateral chest X-ray
- Normal lung function; FVC - forced vital capacity; FEV1 - forced expiratory volume in 1 sec; FEV1/FVC; TLC - total lung capacity; and DLCO - diffusing capacity
- Normal electrocardiogram (sinus bradycardia, premature atrial contractions permissible)
- Not pregnant (females)
- No history of allergies to medications used in the bronchoscopy procedure
- Not taking any medications relevant to lung disease or having an effect on the airway epithelium
- Willingness to participate in the study
- Self-reported non smokers, with smoking status validated by the absence of nicotine and cotinine in urine

Exclusion criteria

- Unable to meet the inclusion criteria
- Current active infection or acute illness of any kind
- Alcohol or drug abuse within the past 6 months
- Evidence of malignancy within the past 5 yr

Healthy smokers

Inclusion criteria

- Males and females, at least 18 yr old
- Provide informed consent
- Good health without history of chronic lung disease, including asthma, and without recurrent or recent (within 3 months) acute pulmonary disease
- Normal physical examination
- Normal routine laboratory evaluation, including general hematologic studies, general serologic / immunologic studies, general biochemical analyses, and urine analysis
- HIV1 negative

- α1-antitrypsin level normal
- Normal PA and lateral chest X-ray
- FVC - forced vital capacity; FEV1 - forced expiratory volume in 1 sec; FEV1/FVC; TLC - total lung capacity; and DLCO - diffusing capacity
- Normal electrocardiogram (sinus bradycardia, premature atrial contractions are permissible)
- Not pregnant (females)
- No history of allergies to medications used in the bronchoscopy procedure
- Not taking any medications relevant to lung disease or having an effect on the airway epithelium
- Willingness to participate in the study
- Self-reported current daily smokers with any number of pack-yr, validated by urine nicotine 30 ng/ml or cotinine 50 ng/ml

Exclusion criteria

- Unable to meet the inclusion criteria
- Current active infection or acute illness of any kind
- Alcohol or drug abuse within the past 6 months
- Evidence of malignancy within the past 5 yr

COPD smokers

Inclusion criteria

- Must be capable of providing informed consent
- Males and females, age 18 or older
- Current daily smokers with any number of pack-yr, validated by urine nicotine 30 ng/ml or cotinine 50 ng/ml
- Meeting GOLD stages I-III criteria for chronic obstructive lung disease (COPD) based on post-bronchodilator spirometry
- Taking any or no pulmonary-related medication, including beta-agonists, anticholinergics, or inhaled corticosteroids
- Normal routine laboratory evaluation, including general hematologic studies, general serologic / immunologic studies, general biochemical analyses, and urine analysis
- Females - not pregnant
- Negative HIV serology
- Normal electrocardiogram (sinus bradycardia, premature atrial contractions are permissible)
- Chest X-ray PA and lateral consistent with COPD
- No history of allergies to medications to be used in the bronchoscopy procedure
- Willingness to participate in the study

Exclusion criteria

- Unable to meet the inclusion criteria
- Individuals in whom participation in the study would compromise the normal care and expected progression of their disease
- Current active infection or acute illness of any kind
- Current alcohol or drug abuse
- Evidence of malignancy within the past 5 yr other than localized skin malignancy

Collection of Airway Epithelium and Assessment of Gene Expression

Fiberoptic bronchoscopy was used to obtain small (and for a random subset, large) airway epithelial cells. To do this, subjects who had previously provided informed consent, were sedated with intravenous midazolam and fentanyl while vital signs, ECG and pulse oximetry were monitored carefully throughout. Supplemental oxygen was administered by nasal cannulae. A 2 mm disposable brush (Wiltek Medical, Winston-Salem, NC) was advanced through the working channel of the bronchoscope into the airways beyond the orifice of the desired lobar bronchus. Samples were then obtained by gently advancing the brush 7 to 10 cm distal to the 3rd generation bronchial airway (i.e., 10th to 12th order bronchi), and sliding the brush back and forth on the epithelium approximately 20 times in multiple adjacent locations. LAE samples were obtained by brushing more proximal in the bronchial tree, at the level of the 3rd to 4th generation airways. The epithelium was collected by repeatedly flicking the brush tip in 5 ml of ice-cold bronchial epithelial basal medium (BEBM, Clonetics, Walkersville, MD). An aliquot of 0.5 ml of all airway epithelial samples was used to estimate differential airway epithelial cell counts on a hemocytometer slide using Diffquik staining reagents (Dade Behring, Newark, NJ), and the remaining 4.5 ml of sample was processed for RNA extraction.

Preparation of cDNA and Hybridization of Labeled cRNA

CSTA gene expression in the airway epithelial samples was assessed using the Affymetrix HG-U133 Plus 2.0 array in accordance with Affymetrix (Santa Clara, CA) protocols. Total RNA was extracted from pelleted cells using TRIzol (Invitrogen, Carlsbad, CA) followed by chloroform extraction. The aqueous phase containing total RNA was purified by RNeasy MinElute RNA purification kit (Qiagen, Valencia, CA). The resulting yield of RNA (2 to 4 μg from 10^6 cells) was stored in RNA Secure (Ambion, Austin, TX), with the NanoDrop-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE) used to ascertain concentration. RNA quality was confirmed by running an aliquot on an Agilent Bioanalyzer (Agilent Technologies, Palo Alto, CA). Double-stranded complementary DNA was created from 3 µg of total RNA using the GeneChip One-Cycle cDNA Synthesis kit, and was followed by a cleanup step using GeneChip Sample Cleanup Module. The samples were then transcribed into biotin-labeled cRNA with GeneChip IVT Labelling Kit, followed by additional cleanup and quantification by spectrophotometry (all reagents from Affymetrix, Santa Clara, CA). Affymetrix protocols then required a test chip hybridization step and, if quality control was acceptable, hybridization to the microarrays was performed. The hybridized arrays were processed by the Affymetrix GeneChip Fluidics Station 450, and scanned with an Affymetrix GeneChip Scanner 3000 7G (http://affymetrix.com/support/technical/manual/expression_manual.affx). For additional quality assurance, all experiments were expected to meet the criteria: (1) cRNA transcript integrity, assessed by signal intensity ratio of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) 3' to 5' probe sets ≤3.0; and (2) multi-chip normalization scaling factor ≤10.0 (1).

TaqMan RT-PCR Confirmation of Microarray-based CSTA Expression Levels

TaqMan RT-PCR was used to quantify the relative gene expression of CSTA in RNA samples collected from the small airways of 23 randomly selected normal healthy nonsmokers, 28 healthy smokers and 13 COPD smokers for whom SAE CSTA expression levels had been determined by microarray. cDNA was synthesized from 2 µg of RNA using the TaqMan Reverse Transcriptase Reaction kit (Applied Biosystems, Foster City, CA), carried out in a 100 µl reaction volume, with random hexamers as primers. Triplicates of 2 dilutions (1:10 and 1:100) were prepared from each subjects sample. The TaqMan reactions (Applied Biosystems Sequence Detection System 7500), were optimized to demonstrate equal amplification efficacy compared to an endogenous control (18s rRNA), and with the average level for the nonsmokers as the calibrator value, relative expression levels were determined using the $\Delta\Delta C_t$ method (Applied Biosystems, Foster City, CA) (2).

Exclusion of Copy Number Variation (CNV) Effects on Associations of Genotype with Gene Expression

Copy number variation polymorphisms could potentially account for observed association of genotype with SAE gene expression. Examination of previously reported regions of CNV showed the presence of a single variant overlapping the boundaries of the CSTA gene [Variation_ 50916, Database of Genomic Variants (3)]. The CNV status of the CSTA locus was inspected for CNVs using the Affymetrix Human SNP Array 5.0 data using Partek Genomics Suite software version 6.5 (Partek Inc., St Louis, MO) relative to HapMap subject, NA12249, a sample chosen from 30 other random HapMap subjects on the basis of possessing the highest copy number probe set intensity values for this region. Search parameters (p<10-5, >10 probe sets, fold-change>1.7 to <2.3, signal/noise ratio=0.5) were chosen to allow detection of Variation_50916 located at chromosome 3 and encompassing the entire CSTA gene within its reported boundaries.

Data Analysis and Statistics

Gene expression analyses on all samples was carried out using the Microarray Suite 5.0 software. The Affymetrix HGU133 Plus 2.0 microarray only possesses a single probe set for CSTA, and data from this probe set (204971_at) was normalized per chip and per gene across all samples using GeneSpring version 7.2 software (Agilent Technologies). To do this, measurements were set at <0.01 to 0.01 and were initially normalized per array to the median expression value on the array (i.e., raw data divided by the 50th percentile of all measurements), and for the comparison of differentially expressed genes, data were additionally normalized per gene, by dividing the raw data by the median expression level for all the genes across all arrays in a dataset. All microarray data have been deposited at the Gene Expression Omnibus (GEO) site (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=jjqvhwyeyegqmty&acc=GSE22047) Squamous cell carcinoma and adenocarcinoma data from Kuner et al is available at http://www.ncbi.nlm.nih.gov/geo/; under accession number GSE10245 (4). Genomic data from the Affymetrix Human SNP 5.0 were assessed using the BRLMM-P Analysis Tool 2.0 software (Affymetrix, Santa Clara, CA) to determine genotype for *cis*-SNPs in the vicinity of CSTA on chromosome 3. For quality control purposes, the gender derived from the microarray was compared to the gender of the subject as reported in the phenotypic database, identity-by-state analysis using the PLINK software program was undertaken to estimate identity-by-descent, allowing detection of duplicates, and a STRUCTURE software analysis using ancestry-informative markers verified the self-reported genetic ancestry. The average BRLMM-P call rate was 98.9% of genotypes successfully called for each subject (range 93.1 to 99.5%).

PLINK was used to examine the associations between CSTA gene expression levels in SAE and the 48 *cis*-SNPs located within 100 kb of the gene. The most significant associations were confirmed using TaqMan expression data for CSTA. In addition, PLINK analyses were used to ensure the main effect remained significant following 10^3 permutations within ancestral clusters, where the self-reported ancestry was verified by STRUCTURE. A linear regression model was applied by PLINK to the data comparing CSTA gene expression levels in SAE with local CSTA haplotypes. Haplotype data were generated in Haploview and in PLINK for the ancestral subpopulation analyses, following exclusion of related subjects using identity-by-state analysis in PLINK (5). Functional information on SNPs was obtained from the National Center for Biotechnology (NCBI) databases. Statistical testing for association between genotypes of a SNP of interest and expression levels was also done with a one way analysis of variance (ANOVA) using a F-test for significance testing. Two factor ANOVAs were used to assess for interaction among variables (ancestry and phenotype, genotype and phenotype).

Supplementary Results

Study Population

Small airway epithelial samples from 178 individuals evaluated included 60 healthy nonsmokers and 82 healthy smokers, and 36 smokers with COPD. LAE samples were obtained from a subset of 21 healthy nonsmokers and 31 healthy smokers (Supplementary Table I). All of the healthy subjects had no significant prior medical history, a normal physical examination, normal urine and blood studies, normal chest imaging, and normal pulmonary function tests. COPD smokers had an FEV1/FVC ratio <70% predicted, and using the GOLD classification of severity, 20 subjects were GOLD I, 14 were GOLD II, and 2 were GOLD III. The healthy nonsmokers did not differ significantly from the healthy smokers in terms of demographic composition (Supplementary Table I), including age ($p>0.3$), gender ($p>0.5$), genetic ancestry ($p>0.1$) or any of the pulmonary function indices ($p>0.05$, all indices) except for DLCO which was lower in smokers compared to nonsmokers ($p<0.03$), although on the average all indices were still within the range of normal values. The group of smokers with COPD were older with a higher pack-yr ($p<0.03$) and had abnormally reduced FEV1, FVC and DLCO relative to the healthy smokers ($p<10^{-3}$, all parameters). There was no difference from healthy smokers with respect to gender ($p>0.5$) or genetic ancestry ($p>0.1$).

Sampling of the Small and Large Small Airway Epithelium

Bronchoscopic brushings were obtained from the small (10th to 12th order) airways yielding from 2.4 to 17.3 x 10^6 cells, of which an average of 98% displayed epithelial morphology, with the typical epithelial cell types (Supplementary Table I) (6). SAE samples from healthy smokers had significantly fewer ciliated cells ($p<10^{-5}$), and significantly more basal ($p<0.04$), undifferentiated ($p<10^{-5}$) and secretory cells ($p<0.03$) compared to healthy nonsmokers. COPD smokers had the greatest proportion of secretory cells of the 3 subject groups (11.9%, $p<0.02$ *vs* healthy smokers), but otherwise had similar proportions of epithelial cell types to what was observed for the healthy smokers.

In addition to the SAE samples, brushings were also performed to collect LAE samples in a subset of the healthy subjects, resulting in an average of 6.9 x 10^6 cells per subject (range 2.5 x 10^6 cells to 18.2 x 10^6 cells), with 99.3% purity for epithelial origin (Supplementary Table I). LAE samples from healthy smokers had significantly more basal cells than were observed in healthy nonsmokers ($p<0.05$), but were otherwise not significantly different in differential cell composition ($p>0.06$, all other comparisons).

Assessment of CSTA Gene Expression by Microarray, and Comparison with Potential Regional Copy Number Variation

The single Affymetrix HGU133 Plus 2.0 microarray probe set for CSTA, 204971_at, was called "Present" in 100% of samples (by Affymetrix "P" call). CSTA was expressed in all phenotypic subgroups. To rule out any contribution of a known copy number variation (CNV) polymorphism overlapping the CSTA gene (7) to the association of genotype with CSTA gene expression, Affymetrix Human SNP Array 5.0 probe set intensities were compared for all genotyped subjects (n=112) to a HapMap reference sample. There was no CNV detected (p<0.00001) in this region in any of the subjects.

Association of CSTA Regional Haplotypes with CSTA Small Airway Epithelium Gene Expression

Cis-haplotypes that encompass the CSTA genomic region could also explain some of the observed pattern of association of genomic variants with CSTA gene expression. To investigate the role of haplotypes and minimize bias from population stratification, subjects of the 2 largest ancestral subpopulations, African American ancestry and European origin were separately phased and analyzed using the program Haploview. For the subset of 56 subjects of African American ancestry with available genotype data, haplotypes derived from the 48 consecutive Affymetrix Human SNP Array 5.0 SNPs located within 100 kbp of the CSTA gene were assessed. All subjects were unrelated as ascertained using identity-by-state analysis, and haplotypes were phased in PLINK. The analyses showed that, in the case of subjects with African American ancestry, 4 haplotypes from 2 adjacent haplotype blocks (out of 6 blocks identified) were significantly associated with small airway epithelium CSTA expression levels. The strongest of these associations was seen for the haplotypes GG and AA of haplotype block 3 (Figure 2B, Supplementary Table II, p<0.02), upstream of the CSTA gene, and rs16832956 was in strong LD with haplotype block 3 (Figure 2B). In the case of subjects of European genetic ancestry, 4 haplotype blocks were identified from among the 48 Affymetrix-genotyped SNPs located within 100 kb of the CSTA gene. The haplotype GAGGGACCCGCT was significantly associated with CSTA small airway epithelium gene expression (Figure 2C, Supplemental Table II, p<0.009).

Effect of Smoking Status and COPD on CSTA Gene Expression in the SAE

SAE CSTA gene expression levels were significantly higher in the group of healthy smokers (n=82) compared to healthy nonsmokers (n=60, p<0.04, pairwise Student's t-test), with evidence of further up-regulation in smokers with COPD (n=36) compared to the healthy smokers (Figure 3A, $p<10^{-3}$, pairwise Student's t-test; $p<10^{-4}$ by analysis of variance for all 3 groups). The up-regulation of SAE CSTA gene expression in healthy smokers *vs* nonsmokers (p<0.05, pairwise Student`s t-test) and in COPD smokers compared to healthy smokers was confirmed by RT-PCR (Figure 3B, p<0.03, pairwise Student's t-test; $p<10^{-3}$ by ANOVA). While phenotype was significantly associated with CSTA gene expression levels as assessed by ANOVA ($p<10^{-4}$), there was no significant difference in CSTA small airway gene expression levels attributable to genetic ancestry (p>0.09) and there was no significant interaction among phenotype and ancestry in relation to gene expression level (Figure 3C, p>0.5). As further evidence of the smoke-responsiveness of CSTA gene expression, the effect of cigarette pack-yr on gene expression was assessed as a function of age within the 3 phenotypic groups. There was no significant difference in SAE CSTA gene expression levels among the older and younger nonsmokers (p>0.07), although healthy smokers older than that subgroup's median age of 43 yr displayed higher CSTA gene expression than the younger healthy smokers (Figure 3D, p<0.03). In parallel with this observation, a difference in cigarette pack-yr was seen among older compared to younger healthy smokers, with the older smokers having a mean pack-yr history of 32.5 ± 2.7 compared to the younger healthy smokers mean pack-yr of 18.6 ± 1.9 (Figure 3D, $p<10^{-4}$). The magnitude of significant difference in pack-yr by age group was smaller in COPD smokers (p<0.03) than what was seen for the healthy smokers. There was a corresponding trend to smaller difference in CSTA gene expression among older compared to younger COPD smokers, although this did not attain statistical significance (Figure 3D, p>0.4). Together, these observations are evidence of a dose-response relationship between smoking and CSTA gene expression. Of note, CSTA expression still trended higher in younger COPD smokers compared to older healthy smokers, despite an opposing trend in their mean pack-yr, consistent with the observed effect of COPD on CSTA gene expression independent of smoking status (Figure 3A, B, D).

Effect of Genetic Variability on the Influence of Smoking and COPD on Small Airway Epithelium CSTA Expression

Because COPD is a complex disease of varied phenotypes that develops in only a minority of smokers, there is agreement that many genetic factors are likely to contribute to the development of disease. Variable gene expression is 1 mechanism whereby genetics can play an interactive role in such complex diseases, but the apparent genetic modulation of CSTA small airway epithelial gene expression levels might not be apparent within certain phenotypes compared to others. We therefore asked: Does genetic variation modulate the influence of smoking and COPD on CSTA gene expression when each of these factors is considered separately? For the SNP displaying the greatest overall correlation of CSTA gene expression with genotype, rs16832956, when phenotypic subgroups were examined apart from each other, healthy smokers with the CC genotype of rs16832956 had significantly higher CSTA expression than the healthy smokers with G_ genotypes ($p<10^{-2}$) and the same pattern was observed among the COPD smokers (Supplementary Figure 1A, $p<0.02$). Analysis of variance showed no significant interaction among the factors of phenotype and genotype in their correlation with small airway epithelium CSTA gene expression ($p>0.3$). For the next most significant SNP overall, rs5008830, the GG genotype had significantly higher expression than A_ genotype subjects in all 3 phenotypic groups examined separately, i.e. the healthy nonsmokers, healthy smokers and smokers with COPD (Supplementary Figure 1B; $p<0.04$, all comparisons).

CSTA Gene Expression in LAE, SCC and Adenocarcinoma

The majority of lung cancers are derived from the SAE (especially adenocarcinoma, which is also the most prevalent histological subtype), and yet a number of reports have indicated that smoking and COPD are more strongly linked to SCC than to adenocarcinoma (8-10). SCCs exhibit a predilection towards the more central airways of the lung (11). Therefore, we performed additional analyses using gene expression data generated using the same microarrays on samples of large airway epithelium (LAE), obtained at bronchoscopy on a subset of subjects (n=21 healthy nonsmokers and n=31 healthy smokers). In addition, publicly available gene expression data obtained from experiments in lung cancer subjects using the same Affymetrix gene expression platform used in the present study were selected and normalized to examine relative differences in CSTA gene expression among lung cancer and the LAE and SAE samples (4). The study population demographic findings and characteristics of brushed large airway epithelium samples are detailed elsewhere in Supplementary Results, and population characteristics pertaining to the Kuner et al (4) data set are publicly available via the internet. CSTA expression levels in LAE of healthy nonsmokers were similar to the upregulated levels seen in the SAE of COPD smokers. The LAE of healthy smokers revealed higher CSTA expression than in nonsmoker LAE (Figure 5A, $p<10^{-3}$). CSTA levels in squamous cell lung carcinoma from the group of 18 individuals reported by Kuner et al (4), were substantially elevated compared to levels observed in the non-malignant SAE or LAE samples (Figure 5A, $p<10^{-2}$, all comparisons). In the 40 lung adenocarcinoma individuals, CSTA gene expression levels were significantly down-regulated compared to levels in SCC and in all airway epithelial sample groups (Figure 5A, $p<0.02$, all comparisons) with the exception of CSTA levels in the SAE of healthy nonsmokers (Figure 5A, $p >0.4$).

In view of the marked up-regulation of CSTA observed in SCC subjects, a comparison was made of the ratio of CSTA gene expression to expression levels of the 3 known cathepsin targets (B, H and L) of cystatin A within individuals. The SAE, LAE and SCC gene expression data show an imbalance in the ratio of CSTA to cathepsins. For example, in the case of all 3 cathepsins, there is a progressive rise in the ratios of CSTA expression level to the cathepsin expression level in healthy smokers compared to healthy nonsmokers, and in COPD smokers compared to healthy smokers. These ratios ("excesses" of CSTA) are most different from the unstimulated (nonsmoker) state, in SCC (Figure 5B-D, $p<10^{-2}$, all comparisons). The progressive rise in CSTA expression observed in these disease states was not offset by a corresponding rise in cognate cathepsin levels in the same tissue from a given individual (Figure 5B-D).

References for Supplemental Data

1. Raman T, O'Connor TP, Hackett NR, et al. Quality control in microarray assessment of gene expression in human airway epithelium. BMC Genomics 2009;10:493.

2. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 2001;25:402-8.

3. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. Nat Genet 2004;36:949-51.

4. Kuner R, Muley T, Meister M, et al. Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. Lung Cancer 2009;63:32-8.

5. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263-5.

6. Harvey BG, Heguy A, Leopold PL, Carolan BJ, Ferris B, Crystal RG. Modification of gene expression of the small airway epithelium in response to cigarette smoking. J Mol Med 2007;85:39-53.

7. de Smith AJ, Tsalenko A, Sampas N, et al. Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. Hum Mol Genet 2007;16:2783-94.

8. Khuder SA. Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis. Lung Cancer 2001;31:139-48.

9. Rosado-de-Christenson ML, Templeton PA, Moran CA. Bronchogenic carcinoma: radiologic-pathologic correlation. Radiographics 1994;14:429-46.

10. Wasswa-Kintu S, Gan WQ, Man SF, Pare PD, Sin DD. Relationship between reduced forced expiratory volume in one second and the risk of lung cancer: a systematic review and meta-analysis. Thorax 2005;60:570-5.

11. Toh CK. The changing epidemiology of lung cancer. Methods Mol Biol 2009;472:397-411.

Supplementary Table I. Characteristics of the Study Population and Airway Epithelial Samples[1]

	SAE[2]			LAE[3]	
Parameter	Healthy nonsmokers	Healthy smokers	COPD smokers	Healthy nonsmokers	Healthy smokers
n	60	82	36	21	31
Gender (male/female)	38/22	56/26	28/8	15/6	21/10
Age (yr)	41 ± 12	42 ± 8	51 ± 8	40 ± 8	44 ± 7
Ancestry (Afr/Eur/Oth)[4]	27/23/10	47/20/15	13/13/10	10/7/4	19/6/6
Smoking (pack-yr)	0	26 ± 17	38 ± 25	0	27 ± 18
Urine nicotine (ng/ml)	negative	1247 ± 239	1025 ± 318	negative	1005 ± 937
Urine cotinine (ng/ml)	negative	1312 ± 164	1327 ± 173	negative	931 ± 410
S/carboxyHb (%)[5]	0.5 ± 0.3	1.8 ± 0.8	2.8 ± 0.9	0.5 ± 0.7	2.2 ± 1.7
Pulmonary function[6]					
FVC	107 ± 13	110 ± 14	94 ± 28	108 ± 13	112 ± 13
FEV1	107 ± 13	108 ± 20	79 ± 22	109 ±17	112 ± 14
FEV1/FVC	83 ± 6	81 ± 11	63 ± 7	83 ± 5	81 ± 4
TLC	101 ± 13	101 ± 12	99 ± 30	99 ± 14	104 ± 11
DLCO	99 ± 14	93 ± 15	73 ± 19	101 ± 17	95 ± 11
Epithelial cells					
Total no. recovered x 10^6	6.3 ± 0.8	7.3 ± 0.9	6.1 ± 0.8	6.6 ± 1.8	7.1 ± 1.6
% inflammatory cells	0.9 ± 1.2	1.5 ± 4.3	2.4 ± 5.6	0.5 ± 0.7	0.3 ± 0.9
% epithelial cells	99.1 ± 1.2	98.5 ± 4.3	97.7 ± 5.6	99.5 ± 0.7	99.2 ± 0.6
% ciliated	72.3 ± 8.9	63.1 ± 14.1	62.3 ± 12.3	53.2 ± 8.8	48.2 ± 13.6
% secretory	6.8 ± 3.8	8.4 ± 4.3	11.9 ± 5.4	8.5 ± 3.6	9.7 ± 4.1
% basal	12.4 ± 6.6	15.2 ± 8.7	12.2 ± 6.6	21.1 ± 4.2	25.9 ± 9.9
% undifferentiated	7.7 ± 3.5	11.9 ± 6.3	11.2 ± 3.7	17.0 ± 8.5	16.2 ± 9.1

[1] Data are presented as mean ± SD.

[2] SAE, small airway epithelium

[3] LAE, large airway epithelium

[4] Afr, African ancestry; Eur, European ancestry; Oth, Other ancestries (mainly Hispanic of Latin American, and Asian ancestry).

[5] S/carboxyHb, serum carboxyhemoglobin.

[6] FVC, forced vital capacity; FEV1, forced expiratory volume in 1 sec; FVC, FEV and FEV1/FVC are all post-bronchodilator values; TLC, total lung capacity. DLCO, diffusing coefficient of the lung for carbon monoxide. FVC, FEV1, TLC and DLCO are presented as % predicted, FEV1/FVC is expressed as % observed.

Supplementary Table II. Association of Regional CSTA Haplotypes with SAE CSTA Gene Expression Levels in Unrelated Individuals of African American and European Genetic Ancestry[1]

Genetic ancestry[2]	Haplotype block #[3]	Haplotype	Frequency[4]	r^2	p value[5]
Afr	3	AA	0.068	0.090	0.010
Afr	3	GG	0.932	0.090	0.010
Afr	4	ATTCG	0.342	0.084	0.013
Afr	4	ACCCG	0.342	0.078	0.017
Eur	2	GAGGGACCCGCT	0.143	0.162	0.008

[1] SNPs located within 100 kbp of the CSTA gene on chromosome 3 were genotyped using the Affymetrix Human SNP Array 5.0. Unrelated subjects of African American ancestry (n=56) and of European ancestry (n=35) were identified and phased by PLINK and haplotypes were identified based on linkage disequilibrium using Haploview software.

[2] Afr, African American ancestry. Eur, European ancestry.

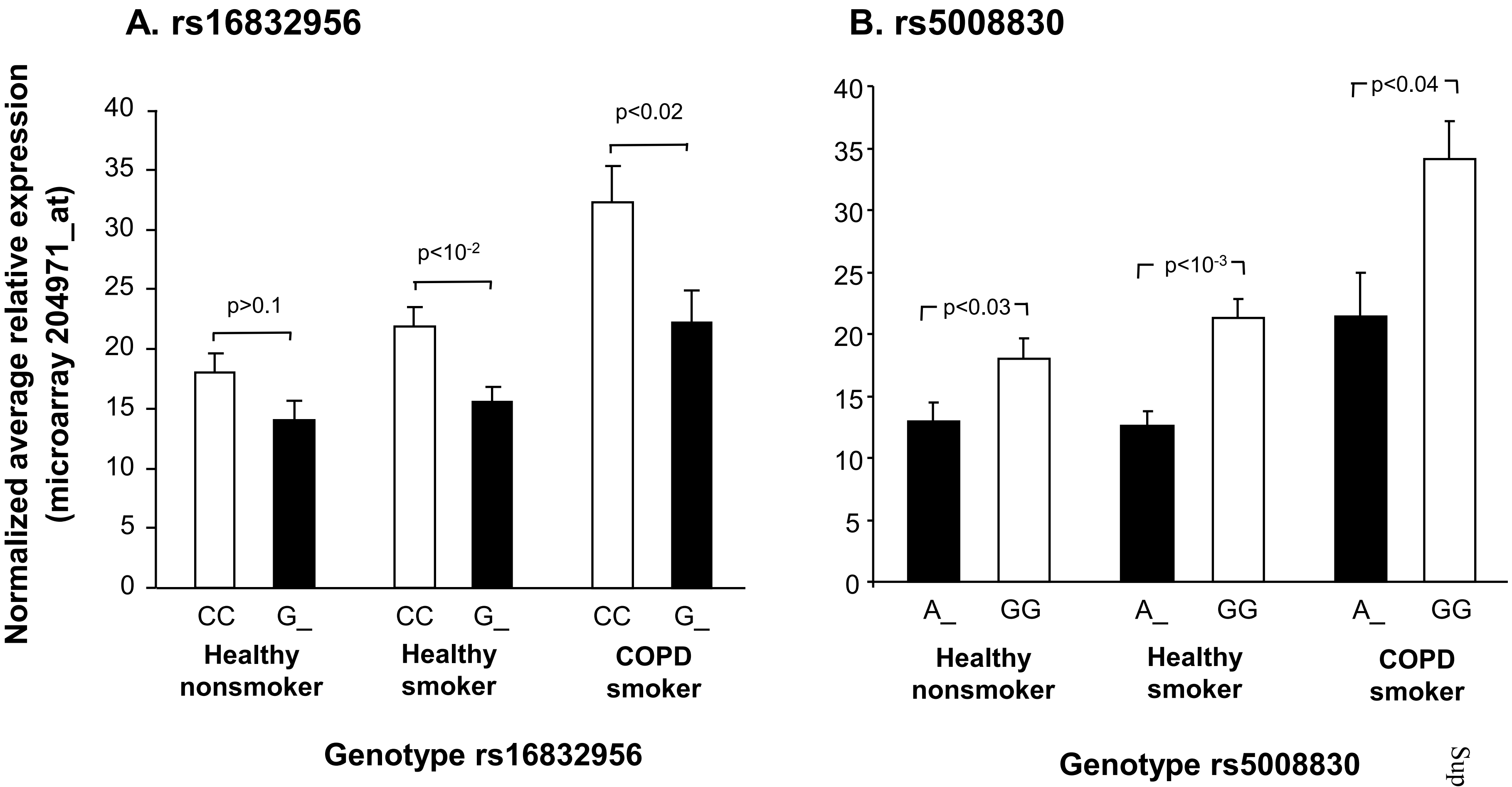
[3] Haplotype block numbers correspond to those shown in Figure 2.

[4] Haplotype frequencies are those observed in the present study.

[5] p values represent Wald statistic based on the t-distribution.

Supplementary Figure Legends

Supplementary Figure 1. Association of genotype with small airway epithelial CSTA expression within phenotypes. **A**. Ordinate: average relative normalized microarray SAE CSTA gene expression levels. Genotypes of rs16832956 are shown for each of the 3 phenotypic groups, i.e. healthy nonsmokers (n=44), healthy smokers (n=48) and smokers with COPD (n=20). **B.** Shown on the ordinate is the average relative normalized microarray small airway epithelial gene expression levels of CSTA. Genotypes of rs5008830 are shown on the abscissa for each of the 3 phenotypic groups. P values shown are from pairwise student t-tests. G_=data for the combined genotypes GG and CG. A_=data for the combined genotypes AA and AG.



A. rs16832956

B. rs5008830

Butler et al.
Supplementary Figure 1