Supplementary Methods

Outline:

I. Study Patients and Sample Collection

A. Primary sample set:

We recruited current and former smokers undergoing flexible bronchoscopy for clinical suspicion of lung cancer at four tertiary medical centers between January 2003 and April 2005. All subjects were greater than 21 years of age and had no contraindications to flexible bronchoscopy including hemodynamic instability, severe obstructive airway disease, unstable cardiac or pulmonary disease (i.e. unstable angina, congestive heart failure, respiratory failure) inability to protect airway or altered level of consciousness and inability to provide informed consent. Never smokers and subjects who only smoked cigars were excluded from the study. For each consented subject, we collected data regarding their age, gender, race, and a detailed smoking history including age started, age quit (if applicable), and cumulative tobacco exposure. Former smokers were defined as patients who had not smoked a cigarette for at least one month prior to entering our study. All subjects were followed, post-bronchoscopy, until a final diagnosis of lung cancer or an alternative diagnosis was made (mean *follow-up time* = 52 days). For those patients diagnosed with lung cancer, the stage and cell type of their tumor was recorded. Only subjects with a final diagnosis available as of May 1st, 2005 were included in the primary sample set. The clinical data collected from each subject in this study including age, gender, smoking history and radiographic findings can be accessed at: http://pulm.bumc.bu.edu/CancerDx/.

While our study recruited patients whose indication for bronchoscopy included a suspicion for lung cancer, each patient's clinical pre-test probability for disease varied. None of the patients had a definitive diagnosis of lung cancer prior to bronchoscopy. In order to ensure that our class prediction model was trained on samples representing a spectrum of lung cancer risk, three independent pulmonary clinicians, blinded to the final diagnoses, evaluated each patient's clinical history (including age, smoking status, cumulative tobacco exposure, co-morbidities, symptoms and radiographic findings) and assigned a pre-bronchoscopy probability for lung cancer. Each patient was classified into one of three risk groups: low (< 10% probability of lung cancer), medium (10–50% probability of lung cancer) and high (> 50% probability of lung cancer). The final risk assignment for each patient was decided by the majority opinion.

B. Prospective sample set:

After completion of the primary study in April 2005, a second set of samples was collected from smokers undergoing flexible bronchoscopy for clinical suspicion of lung cancer at 5 medical centers (St. Elizabeth's Hospital in Boston, MA was added to the 4 institutions used for the primary dataset) between May 2005 and December 2005. Inclusion and exclusion criteria were identical to the primary sample set. Forty subjects who achieved a final diagnosis by January 1st, 2006 were included in this second validation set (including several subjects recruited with the primary sample set whose diagnoses were still pending at the completion of the primary study on May 1st, 2005).

Thirty-five subjects had microarrays that passed our quality-control filter. Demographic data on these subjects, including 18 smokers with primary lung cancer and 17 smokers without lung cancer, is presented in **Supplementary Table 1**. There was no statistical difference in age or cumulative tobacco exposure between case and controls in this prospective cohort (as opposed to our primary dataset).

C. Airway epithelial cell collection:
Bronchial airway epithelial cells were obtained from subjects via flexible bronchoscopy. Following local anesthesia with 2% topical lidocaine to the oropharynx, flexible bronchoscopy was performed via the mouth or nose. Following completion of the standard diagnostic bronchoscopy studies (i.e. bronchoalveolar lavage, brushing and endo/transbronchial biopsy of the affected region), brushings were obtained via 3 endoscopic cytobrushes from the right mainstem bronchus (the number of brushes increased from two to three during the course of the study). If a suspicious lesion (endobronchial or submucosal) was seen in the right mainstem bronchus, brushings were obtained from the uninvolved left mainstem bronchus. Each cytobrush was rubbed over the surface of the airway several times and then retracted from the bronchoscope so that epithelial cells could be placed immediately in TRIzol solution and kept at –80 °C until RNA isolation was performed. Cells were vortexed off the brush into TRIzol solution prior to freezing in order to minimize RNA degradation. The procedure used to obtain these cells is identical to standard diagnostic brushings that are performed, involving the same cytobrush (Cellebrity Cytobrush, Boston Scientific, Boston, MA) brushed, back and forth 3–5 times, against the inside of the airway wall. The only difference relates to numbers of brushes used; our procedure involves 3 cytobrushes per patient as opposed to the 1–2 brushes used for diagnostic cytopathology alone. We obtained approximately 1–2 million epithelial cells per subject.

Given that these patients were undergoing bronchoscopy for clinical indications, the risks from our study were minimal, with less than a 5% risk of a small amount of bleeding from these additional brushings. The clinical bronchoscopy was prolonged by approximately 3–5 minutes in order to obtain the research samples. All participating subjects were recruited by IRB-approved protocols for informed consent, and participation in the study did not affect subsequent treatment. Patient samples were given random identification numbers to protect patient privacy.

II. Microarray Data Acquisition and Preprocessing

A. Microarray data acquisition
6–8 μg of total RNA from bronchial epithelial cells were converted into double-stranded cDNA with SuperScript II reverse transcriptase (Invitrogen) using an oligo-dT primer containing a T7 RNA polymerase promoter (Genset). The ENZO Bioarray RNA transcript labeling kit (Enzo Life Sciences, Inc) was used for in vitro transcription (IVT) of the purified double stranded cDNA. The biotin-labeled cRNA was then purified using the RNeasy kit (Qiagen) and fragmented into fragments of approximately 200 base pairs by alkaline treatment. cDNA and cRNA gels as well as IVT yields were evaluated in

order to assess the quality of RNA for each sample. Each cRNA sample was then hybridized overnight onto the Affymetrix HG-U133A array followed by a washing and staining protocol. Confocal laser scanning (Agilent) was then performed to detect the streptavidin-labeled fluor.

B. Preprocessing of array data via RMA

The Robust Multichip Average (RMA) algorithm was used for background adjustment, normalization, and probe-level summarization of the microarray samples in this study[1]. RMA expression measures were computed using the R statistical package and the justRMA function in the 'affy' Bioconductor package. A total of 296 CEL files from airway epithelial samples included in this study as well as those previously processed in our lab were analyzed using RMA. RMA was chosen for probe-level analysis instead of Microarray Suite 5.0 (MAS 5.0) because it maximized the correlation observed between seven pairs of technical replicates. The average (±SD) Pearson correlation between seven pairs of technical replicates was 0.972 (± 0.012) and 0.985 (± 0.009) for data preprocessed with MAS 5.0 and RMA respectively.

C. Sample filter:

We removed very few samples due to subjective quality metrics during sample processing, and sought an objective and independent quality metric that could be used to exclude microarrays from samples of low quality. A number of quality metrics for Affymetrix arrays have been proposed that utilize summary statistics calculated during the processing of probe-level data by the MAS 5.0 algorithm[2] for sample-quality filtering. As we use data processed by the RMA algorithm for biomarker discovery and validation, we sought to develop a quality metric for sample filtering that did not require also processing the data in MAS 5.0 but which would be well-correlated with the generally accepted MAS 5.0 metrics. For this purpose, we calculated an average *z*-score statistic where each probeset was *z*-score normalized to have a mean of zero and a standard deviation of one across all 152 samples. These normalized gene-expression values were then averaged across all probe-sets for each sample to derive an average *z*-score. This average *z*-score correlates with Affymetrix MAS 5.0 quality metrics such as percent present (Pearson r^2= 0.82) and *GAPDH* 3'/5' ratio (data not shown) both in our data set and a number of other datasets. For the purposes of this proof-of-concept study demonstrating that gene expression in normal airway epithelium can diagnose smokers with lung cancer, we chose to eliminate the 15% samples with the highest (worst) average *z*-score (average *z*-score > 0.129). The resulting sample set consisted of 60 smokers with cancer and 69 smokers without cancer. This average-*z*-score threshold effectively eliminates most samples with less than 30% present probesets in our dataset, which represents a reasonable cutoff for MAS 5.0-processed data[2]. More specifically, the worst 15% of samples had an average percent present of 19.5 %(*SD* = 8.3%), and the remaining samples had an average percent present of 39.5% (*SD* = 6.5%).

D. Prospective validation test set:

CEL files for the additional 40 samples were analyzed using RMA to derive expression values for the new samples. Microarrays that had an average z-score with a value greater than 0.129 (5 of the 40 samples) were excluded. Class prediction of the 35 remaining

prospective samples was conducted using the vote weights for the 80-predictive probesets derived from the training set of 77 samples (described in detail below) using expression values computed in section B above. 28/35 (80%) of samples were accurately classified in this cohort (**Fig. 2b**).

As described in the Methods section of the manuscript, the technique to collect airway epithelium was optimized during the course of the primary study. In order to evaluate our gene-expression biomarker across all subjects entered into the study, we also calculated the diagnostic yield of our biomarker across all prospective subjects regardless of sample quality. A total of 40 subjects were recruited into the series and five of these subjects were excluded due to poor quality array data. As a result, our biomarker achieved an accurate classification of 28/40 (70%) samples. Further optimization of the sample collection protocol may increase the ultimate diagnostic yield of the gene-expression biomarker.

III. Microarray Data Analysis

A. Class Prediction Algorithm

The weighted voting algorithm[3] was implemented as the class prediction method, with modifications to the gene-selection methodology. Genes that varied between smokers with and without cancer in the training set samples after adjusting for tobacco-smoke exposure ($P < 0.05$) were identified using an ANCOVA with pack-years as the covariate. Age also differed between the two groups, but was not included as a covariate because of its strong correlation with pack years ($r_{spearman} = 0.65$, $P < 10^{-16}$). Further gene selection was performed using the signal to noise metric and internal cross-validation where the 40 most consistently up- and the 40 most consistently down-regulated probesets were identified. The internal cross validation involved leaving 30% of the training samples out of each round of cross-validation, and selecting genes based on the remaining 70% of training set samples. The final gene committee consisted of eighty probesets that were identified as being most frequently up- or down- regulated across 50 rounds of internal cross-validation. The parameters of this gene-selection algorithm were chosen to maximize the average accuracy, sensitivity and specificity obtained from fifty runs. We chose 80 probesets as the optimal number for the predictor based on an exploration of parameters affecting the performance of the algorithm across 1000 runs where different training/test sets were chosen. An 80 probeset biomarker had the highest accuracy in the test sets across these 1000 runs when compared to biomarkers containing 2, 20, 40, 60 and 100 genes. This algorithm was implemented in R and yields results that are comparable to the original implementation of the weighted-voted algorithm in GenePattern when a specific training, test and gene set are given as input.

After determination of the optimal gene-selection parameters, the algorithm was run using a training set of 77 samples to arrive at a final set of genes capable of distinguishing between smokers with and without lung cancer. The accuracy, sensitivity and specificity of this classifier were tested against 52 samples that were not included in the training set. The performance of this classifier in predicting the class of each test-set

sample was assessed by comparing it to runs of the algorithm where either: 1) different training/tests sets were used; 2) the cancer status of the training set of 77 samples was randomized; or 3) the genes in the classifier were randomly chosen (see below for details).

B. Randomization

The accuracy, sensitivity, specificity, and area under the Receiver Operating Characteristics (ROC) curve (*AUC*) (using the signed prediction strength as a continuous cancer predictor) for the 80-probeset predictor were compared to 1000 runs of the algorithm using three different types of randomization (**Fig. 1b** and **Supplementary Table 4**). First, the class labels of the training set samples were permuted and the entire algorithm, including gene selection, was re-run. These steps were performed 1000 times (referred to as Random 1). The second randomization used the 80 genes in the original predictor but permuted the class labels of the training set samples over 1000 runs to randomize the gene weights used in the classification of the test set samples (referred to as Random 2). In both of these randomization methods, the class labels were permuted such that half of the training set samples were labeled correctly. The third randomization method involved randomly selecting 80 probesets for each of 1000 random classifiers (referred to as Random 3). The *P*-value for each metric and randomization method indicate the percentage of runs using that randomization method that had the same or better performance than the actual classifier.

In addition to the above analyses, the actual classifier was compared to 1000 runs of the algorithm where different training/test sets were chosen and the correct sample labels were retained. Empirically derived p-values were also computed to compare the actual classifier to the 1000 runs of the algorithm (**Supplementary Table 4**).

Finally, these 1000 runs of the algorithm from above were compared to 1000 runs were the class labels of *different* training sets had been randomized (**Supplementary Fig. 3)**. This randomization is in contrast to the randomizations conducted above which were restricted to the training set of 77 samples. Empirically derived p-values were computed to compare 1000 runs to 1000 random runs (**Supplementary Table 4**).

C. Characteristics of the 1000 additional runs of the algorithm:

In order to further assess the stability of the biomarker gene committee, the number of times each of the 80 probesets used in the biomarker was selected across 1000 runs was examined (**Supplementary Fig. 3**). The majority of the 80-biomarker probesets were chosen frequently over the 1000 runs (37 probesets were present in over 800 runs, and 58 of the probesets were present in over half of the runs). For purposes of comparison, when the cancer status of the training set samples are randomized over 1000 runs, the most frequently selected probeset is chosen 66 times, and the average is 7.3 times.

D, Comparison of RMA vs. MAS 5.0 and weighted voting vs. PAM

To evaluate the robustness of airway gene expression to classify smokers with and without lung cancer, we examined the effect of different class-prediction and data

preprocessing algorithms. We tested the 80-probesets in our classifier to generate predictive models using the Prediction Analysis of Microarrays (PAM) algorithm[4], and also tested the ability of the WV algorithm to use probeset level data that had been derived using the MAS 5.0 algorithm instead of RMA. The accuracy, sensitivity and specificity of the weighted voting algorithm on data preprocessed in MAS 5.0 were identical were identical to the performance on data preprocessed in RMA. The accuracy of our biomarker using the PAM algorithm was 87% for data preprocessed in RMA and MAS 5.0.

E. Prediction strength:
The Weighted voting algorithm predicts a sample's class by summing the votes each gene on the class prediction committee gives to one class versus the other. The level of confidence with which a prediction is made is captured by the Prediction Strength (*PS*) and is calculated as follows:

$$PS = \frac{V_{winning} - V_{losing}}{V_{winning} + V_{losing}}$$

$V_{winning}$ refers to the total gene committee votes for the winning class and V_{losing} refers to the total gene committee votes for the losing class. Since $V_{winning}$ is always greater than V_{losing}, *PS* varies from 0 (arbitrary) to 1 (complete confidence) for any given sample.

In our test set, the average *PS* for our gene profile's correct predictions (43 / 52 test samples) is 0.73 (± 0.27), while the average *PS* for the incorrect predictions (9 / 52 test samples) is much lower: 0.49 (± 0.33). This result shows that, on average, the Weighted Voting algorithm is more confident when it is making a correct prediction than when it is making an incorrect prediction. This result holds across 1000 different training and test set pairs: the average prediction strength when a sample was classified correctly was 0.64 for cancers and 0.75 for no cancers, and the average prediction strength when a sample was misclassified was 0.53 for cancer and 0.58 for no cancers.

F. Tobacco exposure as a potential confounding variable
Several analyses were performed to address the potential confounding effects of tobacco exposure. The biomarker's ability to accurately classify the test set of 52 samples does not appear to be correlated with the subject's tobacco smoke exposure. Although there is a pack year difference between cancer and no cancer samples in our dataset, these data suggest that incorrect predictions are not the result of the subject's pack years.

As described above, a gene filter was applied upstream of the class prediction algorithm using the training set samples to identify genes differentially expressed between smokers with and without cancer ($P < 0.05$) using an ANCOVA with pack-years as the covariate. The ANCOVA was implemented to control for the pack year difference between the two groups. There were two probesets in the classifier where the *P*-value for the pack-years covariate was also < 0.05. To account for possible non-linear relationships between gene

expression and pack years, we examined the Spearman correlation between pack years and gene expression for each of the 80 probesets. 11 of the 80 probesets had a $P < 0.05$. One of these probesets overlapped with the two identified by ANCOVA, resulting in 12 unique probesets that show pack-year correlated gene expression. We removed these 12 probesets from the classifier, and derived a new classifier using the training set samples and the 68 remaining probesets to predict the 52 test samples. The new classifier using the 68 probesets had a similar accuracy (80.8%) to the original 80 probeset classifier with only a single no cancer sample predicted differently.

Furthermore, we tested the performance of a classifier derived from a smaller training set where smokers with and without cancer were more closely matched for degree of tobacco exposure. The 77 samples in our original training set were divided into quartiles based on cumulative smoke exposure and then separated by class (cancer vs. no cancer). The class with the smallest number of samples falling into a given quartile were matched with an equivalent number of randomly chosen samples falling into the same quartile from the alternate class. This resulted in a subset of 50 matched samples (25 in each class) that was used to train a new 80 probeset classifier using the probeset selection methodology described above. When tested against our original test set of 52 samples, the new classifier had an overall accuracy of 77% (40 / 52), a sensitivity of 75% (15 / 20) and a specificity of 78% (25 / 32). This slight reduction in overall accuracy likely reflects the use of a smaller training set to derive the classifier, as 100 classifiers using 50 sample subsets of the original training set (where 25 cancer samples and 25 noncancer samples were randomly chosen and smoking history was not matched) yielded a similar overall accuracy, sensitivity and specificity as the matched 50 sample training set (data not shown).

G. Cancer cell type and stage:

To determine if tumor cell subtype or stage affects the expression of genes that distinguish large airway epithelium samples from smokers with and without lung cancer, Principal Component Analysis (PCA) was performed on the gene-expression measurements for the 80 probesets in our predictor and all of the airway epithelium samples from patients with lung cancer. Gene expression measurements were *z*-score normalized prior to PCA. Samples did not separate by cell type or stage in this analysis (data not shown).

The accuracy of the biomarker cancer predictions appears similar as a function of stage and cell type (**Supplementary Fig. 4**). While squamous cell carcinomas were the predominant cell type in our primary lung cancer dataset, the biomarker also appears robust in identifying smokers with adenocarcinomas as smokers with lung cancer. All four adenocarcinomas in the initial test set were correctly classified. Furthermore, 7 / 18 cancers in our prospective series were adenocarcinomas (a proportion that is more reflective of the epidemiology of the disease), and we correctly classified all seven cases as smokers with lung cancer. Given that we correctly classified all 11 adenocarcinoma cases in the two independent sets, we used a Bayesian analysis to compute a confidence interval around our false-positive error rate (α) estimate. We use the standard Bayesian

conjugate analysis that assumes, before seeing the data, that all values of α are equally likely. The 11 cases that are correctly classified were used to update this uniform distribution into a β distribution with shape parameters 12 and 1. Under this distribution, we are 95% confident that the accuracy for adenocarcinomas is greater than 78%

IV. Link to Lung Cancer Tissue Microarray Datasets

A. Preprocessing of Bhattacharjee dataset

RMA-derived expression measurements were calculated from the 254 HgU95Av2 CEL files from Bhattacharjee *et al.*[5] as described above. Technical replicates were filtered by choosing one at random to represent each patient. In addition, arrays from carcinoid samples and arrays from patients with no smoking history were excluded, leaving 151 samples. The *z*-score quality filter described above was applied to this data set resulting in 128 samples for further analysis (88 adenocarcinomas, three small cell, 20 squamous, and 17 normal lung samples).

Probesets were mapped between the HGU133A array and HGU95Av2 array using Chip Comparer (http://tenero.duhs.duke.edu/genearray/perl/chip/chipcomparer.pl). 64 probesets on the HGU95Av2 array mapped to the 80-biomarker probesets. The 64 probesets on the HGU95Av2 correspond to 48 out of the 80 biomarker probesets (meaning that 16 of 80 biomarker probesets map to multiple HGU95Av2 probesets and that 32 of 80 biomarker probesets have no clear corresponding probe on the HGU95Av2 array).

B. Analyses of Bhattacharjee dataset

In order to explore the expression of biomarker genes in the lung tumors profiled by Bhattacharjee, two different analyses were performed. In the first analysis, PCA was used to examine variability in biomarker gene expression between the 128 Bhattacharjee samples. PCA was conducted in R using the package prcomp on the *z*-score normalized 128 sample by 64 probeset matrix. The normal and malignant samples in the Bhattacharjee dataset appear to vary along the first principal component (**Fig. 3**). To assess the significance of this difference, PCA of the samples was repeated using 1000 randomly chosen sets of 64 probesets. The mean difference between normal and malignant samples was calculated for the first principal component of the actual 64 probesets and for each of the 1000 random sets of 64 probesets. The difference between normal and malignant from the 1000 random gene sets was used to generate a null distribution. The observed difference between the normal and malignant samples using the biomarker probesets was greater than the difference observed using randomly selected genes ($P = 0.026$ for mean difference and $P = 0.034$ for median difference).

The second analysis involved using the weighted voted algorithm to predict the class of 108 samples in the Bhattacharjee dataset using the 64 probesets and a training set of 10 randomly chosen normal tissues and 10 randomly chosen tumor tissues. The samples were classified with 89.8% accuracy, 89.1% sensitivity, and 100% specificity. To

examine the significance of these results, the weighted voted algorithm was re-run using two types of data randomization. First, the class labels of the training set of 20 samples were permuted and the algorithm, including gene selection, was re-run 1000 times. The second randomization involved permuting the class labels of the training set of 20 samples and re-running the algorithm 1000 times keeping the list of 64-probsets constant). In the above two types of randomization, the class labels were permuted such that half the samples were correctly labeled. *P*-values for the classification accuracy of the single run relative to the accuracies obtained from each randomization method were calculated as the percentage of 1000 runs that had the same or better performance than the actual classifier. Biomarker genes are significantly better able to distinguish lung cancer tissue from normal lung tissue when the sample classes are correct ($P < 0.01$ for both randomization methods).

C. *Analyses of Wachi Dataset*

RMA probe level expression values derived from the 10 HGU133A arrays in the Wachi dataset[6] were downloaded from Gene Expression Omnibus (GEO Dataset Accession ID: GDS1312). The 10 arrays represent five squamous cell lung cancer tissue samples from smokers and five matched adjacent histologically normal lung tissue samples taken from the same patients.

To evaluate our lung cancer biomarker in both malignant and matched normal lung tissue profiled by Wachi, two different analyses were performed. First, an unsupervised clustering analysis was used to explore the relatedness of the 10 Wachi samples with the 52 large airway samples from our test set. Expression values for each of our 80-classifier probesets were *z*-score normalized across all 62 samples and clustering was performed using Average Linkage and Pearson Correlation as the distance metric. For the 80-classifier probesets, squamous carcinoma and tumor-adjacent tissue exhibit a pattern of gene expression more similar to bronchial airway epithelium of smokers with cancer than smokers without cancer (data not shown). The second analysis involved using our biomarker to classify the cancer status of the 10 Wachi samples using the probeset weights generated from our original 77 sample training set. This resulted in all 10 of the samples, including the five adjacent normal lung samples, being predicted as coming from patients with cancer. To assess the significance of this result, the cancer status of the 10 Wachi samples was predicted using randomly chosen sets of 80 probesets (data not shown). Analysis of the randomly chosen 80 probeset classifiers over multiple, independent weighted voting runs showed that on average only 54% of the five phenotypically normal Wachi samples were predicted as being cancerous.

D. Analyses of *Raponi and Potti datasets:*

MAS 5.0 derived probe level expression values from the 130 microarrays in the Raponi *et al*. dataset[7] were downloaded from the Gene Expression Omnibus (GEO Series Accession ID: GSE4573). These 130 samples represent fresh frozen, surgically resected malignant lung tissue from 129 individual patients with different stages of Squamous Cell Carcinoma. To evaluate the performance of our lung cancer biomarker on these malignant lung tissue samples, we used our biomarker to classify the cancer status of all 130 samples using the probeset weights generated by expression values in our original 77

sample training set when processed by MAS 5.0. This resulted in 129 / 130 of the samples being predicted as coming from smokers with lung cancer. Similarly, we downloaded the RMA derived probe level expression values from the 198 lung cancer microarrays in the Potti *et al.* study[8] from the Gene Expression Omnibus (GEO Series Accession ID: GSE3593). These 198 samples represent resected malignant lung tissue from 198 individual patients with different stages of non-small cell lung cancer. Our biomarker classified 178 of 198 samples as coming from smokers with lung cancer.

V. Real Time PCR:

Quantitative RT-PCR analysis was used to confirm the differential expression of a seven genes from our classifier. Primer sequences for the candidate genes and a housekeeping gene, the 18S ribosomal subunit, were designed with PRIMER EXPRESS software (Applied Biosystems) (**Supplementary Table 5**). Primer sequences for five other housekeeping genes (*HPRT1*, *SDHA*, *YWHAZ*, *GAPDH*, and *TBP*) were adopted from Vandesompele *et* al[6]. RNA samples (1 μg of the RNA used in the microarray experiment) were treated with DNAfree (Ambion, Austin, TX), according to the manufacturer's protocol, to remove contaminating genomic DNA. Total RNA was reverse-transcribed using random hexamers (Applied Biosystems) and SuperScript II reverse transcriptase (Invitrogen). The resulting first-strand cDNA was diluted with nuclease-free water (Ambion) to 5 ng/μl. PCR amplification mixtures (25 μl) contained 10 ng template cDNA, 12.5 μl of 2X SYBR Green PCR master mix (Applied Biosystems) and 300 nM forward and reverse primers. Forty cycles of amplification and data acquisition were carried out in an Applied Biosystems 7500 Real Time PCR System. Threshold determinations were automatically performed by Sequence Detection Software (version 1.2.3) (Applied Biosystems) for each reaction. All real-time PCR experiments were carried out in triplicate on each sample (six samples total: three smokers with lung cancer and three smokers without lung cancer).

Data analysis was performed using the geNorm tool[9]. Three genes (*YWHAZ*, *GAPDH*, and *TBP*) were determined to be the most stable housekeeping genes and were used to normalize all samples. Data from the QRT-PCR for seven biomarker genes showed changes that were consistent with those observed by microarray (**Supplementary Fig. 1**).

VI. Investigating the cell of origin for biomarker genes:

A.Inflammatory cell gene filter

While cytologic review of select airway brushings from smokers with and without lung cancer revealed that greater than 90% of cells are epithelial in origin and that there was no difference in proportion of inflammatory cells between the two groups, we wanted to further explore at the gene-expression level whether any of the observed differences in gene expression might be due to differences in the presence of inflammatory cells. We had previously developed and validated a gene expression filter to identify samples that

were potentially contaminated with significant numbers of inflammatory cells[10]. The filter consists of 11 probesets representing genes that are specific for various lineages of white blood cells[11] and thus should not be expressed in bronchial epithelial tissue. Expression profiles of these 11 probesets across multiple tissue types in the Human Gene Atlas study[12] confirmed that these 11 probesets were expressed below the median gene expression level (across all tissue types) in bronchial epithelial tissue while they were expressed at levels greater than 3 × the median in various immune cells, blood, lymphoid and myeloid tissues. In our primary dataset of 129 samples, only one sample had two or more of these inflammatory-cell specific probesets expressed at detectable levels (detection *P*-value < 0.05). These results suggest that inflammatory cells are present at levels which are too low to significantly impact the gene expression measurements from airway bronchoscopy samples.

In order to further determine if there was a systematic difference in inflammatory cell number between bronchoscopy samples from smokers with or without cancer, we performed a Principal Components Analysis using the gene expression values of the 11 inflammatory cell-specific probesets across the 129 samples in the primary dataset (**Supplementary Fig. 2**). There is no separation of the training set samples with regard to sample class, suggesting that the number of inflammatory cells is similar between the two groups.

B.Expression of biomarker inflammatory genes in Human Gene Atlas Study
Eight (*TPD52*, *CD55*, *CD164*, *C6*, *DEFB1*, *IL8*, *FCGR3A*, and *PLA2G4A*) out of the 12 genes immune/inflammation related genes in the biomarker (**Supplementary Table 3**) are expressed in bronchial epithelial cells at levels equal to or greater than the median expression level compared with all other samples (*n* = 78 different types of cells or tissues) in a recent version of the Human Gene Atlas known as SymAtlas (http://symatlas.gnf.org/SymAtlas/). In addition, *IL8* is expressed in bronchial epithelial cells in SymAtlas at a level higher than 75% of the other samples.

C.Immunohistochemisty:
Immunohistochemisty studies were performed to detect protein products of two inflammatory markers in our biomarker, CD55 and IL-8, using airway bronchoscopy samples from our study. These genes were selected because they are the inflammatory genes that contribute most strongly to the biomarker (**Supplementary Table 3**).

Bronchial brushes from smokers with and without lung cancer (*n* = 4) were rinsed immediately in ice-cold saline, and the cells were pelleted (300 *g* x 10 min), resuspended in phosphate buffered saline and attached onto glass sides by cytospin (ThermoShandon). The cell monolayers were fixed with freshly prepared paraformaldehyde (2%) for 10 minutes and used for indirect immunofluorescence. Nonspecific binding sites were blocked with normal goat serum for 30 minutes. Rabbit polyclonal anti-human CD55 antibody (Santa Cruz Biotechnology) or rabbit polyclonal anti-human IL-8 antibody (Santa Cruz) was applied at a dilution of 1:50 for 1 h. Slides were rinsed three times with phosphate buffered saline before applying a 1:100 dilution of Alexa-Fluor 568 goat anti-rabbit IgG (Molecular Probes) for 1 h. All applications were

at room temperature. Nuclei were stained with DAPI and cells visualized by fluorescence microscopy. Human peripheral blood mononuclear cells treated in the same fashion were used as positive controls for CD55 antibody concentrations. Signal specificity was confirmed using separate cell preparations and eliminating primary or secondary antibody from the protocol and by application of non-immune rabbit serum as negative controls. Results from these studies demonstrate that CD55 and IL-8 are expressed in bronchial airway epithelium (**Supplementary Fig. 2**).

VII. References:

(1) Irizarry, R.A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15 (2003).

(2) Hoffman, A., et al. Expression profiling — best practices for data generation and interpretation in clinical trials. *Nature Reviews Genetics*. **5**: 229-38 (2004).

(3) Golub, T.R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).

(4) Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567-6572 (2002).

(5) Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **98**, 13790-13795 (2001).

(6) Wachi, S., Yoneda, K., Wu, R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* **21**, 4205-8 (2005)

(7) Raponi, M., et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res*. **66**, 7466-72 (2006).

(8) Potti, A., et al. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N Engl J Med*. **355**:570-80 (2006).

(9) Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3**, RESEARCH0034 (2002).

(10) Spira, A. *et al.* Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl. Acad. Sci. USA* **101**, 10143-10148 (2004).

(11) Hoffman, Benz, Shattil, Furie, Cohen, Silberstein, McGlave. Hematology, Basic Principles and Practice. 3rd Edition, p2474 (2000).

(12) Su, A.I., et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc *Natl Acad Sci U S A*. **101**:6062-7 (2004)