

RESEARCH ARTICLE

Open Access

# Derivation of a bronchial genomic classifier for lung cancer in a prospective study of patients undergoing diagnostic bronchoscopy

Duncan H Whitney<sup>1,7\*</sup>, Michael R Elashoff<sup>2</sup>, Kate Porta-Smith<sup>1,7</sup>, Adam C Gower<sup>3</sup>, Anil Vachani<sup>4</sup>, J Scott Ferguson<sup>5</sup>, Gerard A Silvestri<sup>6</sup>, Jerome S Brody<sup>3</sup>, Marc E Lenburg<sup>3</sup> and Avrum Spira<sup>3</sup>

## Abstract

**Background:** The gene expression profile of cytologically-normal bronchial airway epithelial cells has previously been shown to be altered in patients with lung cancer. Although bronchoscopy is often used for the diagnosis of lung cancer, its sensitivity is imperfect, especially for small and peripheral suspicious lesions. In this study, we derived a gene expression classifier from airway epithelial cells that detects the presence of cancer in current and former smokers undergoing bronchoscopy for suspect lung cancer and evaluated its sensitivity to detect lung cancer among patients from an independent cohort.

**Methods:** We collected bronchial epithelial cells (BECs) from the mainstem bronchus of 299 current or former smokers (223 cancer-positive and 76 cancer-free subjects) undergoing bronchoscopy for suspected lung cancer in a prospective, multi-center study. RNA from these samples was run on gene expression microarrays for training a gene-expression classifier. A logistic regression model was built to predict cancer status, and the finalized classifier was validated in an independent cohort from a previous study.

**Results:** We found 232 genes whose expression levels in the bronchial airway are associated with lung cancer. We then built a classifier based on the combination of 17 cancer genes, gene expression predictors of smoking status, smoking history, and gender, plus patient age. This classifier had a ROC curve AUC of 0.78 (95% CI, 0.70-0.86) in patients whose bronchoscopy did not lead to a diagnosis of lung cancer ( $n = 134$ ). In the validation cohort, the classifier had a similar AUC of 0.81 (95% CI, 0.73-0.88) in this same subgroup ( $n = 118$ ). The classifier performed similarly across a range of mass sizes, cancer histologies and stages. The negative predictive value was 94% (95% CI, 83-99%) in subjects with a non-diagnostic bronchoscopy.

**Conclusion:** We developed a gene expression classifier measured in bronchial airway epithelial cells that is able to detect lung cancer in current and former smokers who have undergone bronchoscopy for suspicion of lung cancer. Due to the high NPV of the classifier, it could potentially inform clinical decisions regarding the need for further invasive testing in patients whose bronchoscopy is non diagnostic.

## Background

Lung cancer remains the leading cause of cancer mortality in the United States, with an estimated 224,000 new diagnoses, and 160,000 deaths in 2014, 90% of which are due to smoking [1]. Recently, the National Lung Cancer Screening Trial showed that low dose Computed Tomography (CT) screening results in a 20% relative mortality

reduction in high risk individuals [2]. The mortality reduction, however, was accompanied by a high rate (~96%) of false-positive CT findings, which in turn has generated concern for the overutilization of invasive diagnostic procedures [3].

Patients with suspected lung cancer are often referred for bronchoscopy where the primary aim is to sample a suspicious pulmonary lesion for pathological analysis. It is estimated that 500,000 bronchoscopies are performed per year in the U.S. [4], of which roughly half are for the diagnosis of lung cancer. Bronchoscopy is considered to

\* Correspondence: Duncan.whitney@veracyte.com

<sup>1</sup> Allegro Diagnostics, Corp, Maynard, MA, USA

<sup>7</sup> Present affiliation: Veracyte, Inc, South San Francisco, CA, USA

Full list of author information is available at the end of the article

be safer than other invasive sampling methods, such as transthoracic needle biopsy (TTNB), or surgical techniques. However the diagnostic sensitivity of bronchoscopy is sub-optimal, ranging from 34% (for <2 cm peripheral nodules) to 88% (for larger, centrally located lesions) [5]. Adoption of guidance techniques has expanded the applicability of bronchoscopy to more challenging suspicious lesions (i.e., solitary pulmonary nodules which are often peripheral in the lung), but the overall clinical sensitivity of bronchoscopy for lung cancer has not improved substantially [6,7]. When bronchoscopy is non-diagnostic, physicians are often left with the ambiguity of whether to pursue further invasive diagnostic procedures, with associated complications [8,9], or choose imaging surveillance. In current practice when these invasive procedures are performed, approximately a third of patients are determined to have benign disease [10,11], suggesting that these procedures are avoidable. Methods that reduce this ambiguity by substantially improving the diagnostic yield of bronchoscopy could improve patient care.

It has previously been demonstrated that cigarette smoke creates a molecular field of injury in airway epithelial cells that line the entire respiratory tract [12]. The reversible and irreversible impact of cigarette smoke on the bronchial airway transcriptome has been characterized and a set of gene-expression alterations in the bronchial epithelium have been identified in current and former smokers with lung cancer [13]. These cancer-associated gene expression profiles have previously been shown to yield a sensitive classifier for detecting lung cancer when bronchoscopy is non-diagnostic. The high sensitivity of this classifier, measured in a biospecimen readily accessible during bronchoscopy, results in a very low probability of lung cancer when the test result is negative, and suggests that physicians might be enabled to confidently pursue active surveillance and reduce risky invasive procedures in subjects without lung cancer.

We have expanded upon these proof of concept studies and conducted a prospective, multi-center study to derive a gene-expression classifier that could directly impact management of current and former smokers undergoing bronchoscopy for suspicion of lung cancer. We then validated the classifier in an independent cohort.

## Methods

### Training set patient population

Patients were enrolled in the AEGIS trials (Airway Epithelium Gene Expression In the DiagnosiS of Lung Cancer), designed as prospective, observational, cohort studies (registered as NCT01309087 and NCT00746759) of current and former cigarette smokers with a suspicion of lung cancer undergoing bronchoscopy as part of their diagnostic workup. A set of patients from one of the cohorts ("AEGIS 1") was selected for the exclusive purpose

of training a gene expression classifier. All enrolled patients were followed post-bronchoscopy until a final diagnosis was made, or for 12 months. Patients were diagnosed as having primary lung cancer based on cytopathology obtained at bronchoscopy or upon subsequent lung biopsy (such as TTNB or surgical lung biopsy (SLB) when bronchoscopy did not lead to a diagnosis of lung cancer). Patients were diagnosed as having benign disease based on a review of medical records and follow-up procedures at 12 months post-bronchoscopy (described in more detail in Additional file 1). Bronchoscopy was considered "diagnostic" when clinical samples collected at the time of the bronchoscopy procedure yielded a confirmed lung cancer diagnosis via cytology or pathology. The study was approved by IRB at each of the participating medical centers (the ethics committees and the study protocol numbers for each of the centers is listed separately; Additional file 2), and all patients signed an informed consent prior to enrollment.

### Sample collection

Physicians at each of 25 participating medical centers (see Additional file 3) were instructed to collect normal appearing bronchial epithelial cells (BEC) from the right mainstem bronchus (or the left side if any abnormalities were observed on the right) during bronchoscopy using standard bronchoscopic cytology brushes. Following collection, the cytology brushes were cut and placed in an RNA preservative (Qiagen RNeasy Protect, Cat. 76526) immediately after collection and stored at 4°C. Specimens were then shipped at 4-20°C to a central laboratory for further processing.

### RNA isolation

BECs were separated from cytology brushes using a vortex mixer and were then pelleted and processed using QIAzol lysis reagent (Qiagen). RNA was isolated by phenol/chloroform extractions and purified on a silica membrane spin-column (Qiagen miRNeasy kit, Cat. #217004) according to manufacturer's recommendations. RNA was analyzed on a NanoDrop ND-1000 spectrophotometer (Thermo Scientific) to determine concentration and purity, and RNA integrity (RIN) was measured on a 2100 Bioanalyzer (Agilent Technologies). Each sample was then stored at -80°C until processing further on microarrays.

### Microarray processing

Total RNA (200 ng) was converted to sense strand cDNA, amplified using the Ambion WT Expression kit (Life Technologies Cat. #4440536), and labeled with Affymetrix GeneChip WT terminal labeling kit (Affymetrix Cat. #900671), (described in more detail in Additional file 1). The labeled cDNA was hybridized to Gene 1.0 ST microarrays (Affymetrix Cat. #901085) and analyzed on an

Affymetrix GeneChip Scanner. Individual CEL files for each of the patient samples were normalized using the standard Affymetrix Gene 1.0 ST CDF and RMA [14].

### Classifier development

A gene expression classifier was derived in a multi-step process. Initial modeling consisted of using the training data to select genes which were associated with three clinical covariates (gender, tobacco use, and smoking history) to identify gene expression correlates of these clinical variables. Lung cancer-associated genes were then selected, and finally a classifier for predicting the likelihood of lung cancer based on the combination of the cancer genes, the gene expression correlates, and patient age was derived. All aspects of this classifier development procedure were determined using cross validation and using only data from the training set samples.

### Clinical Factor Gene Expression Correlates (CFGC)

Covariates of lung cancer in this study population, including sex (male/female), smoking status (current/former), and pack years (<10/>10), were modeled to identify gene expression correlates for the clinical factors. Empirical Bayes t-tests were used to identify genes whose expression was significantly associated with each of the clinical factors. Next, the significant genes were used to build three models, one for predicting each clinical factor, using penalized logistic regression (LASSO) [15]. Finally, the predicted values from the gene expression models for gender (GG), smoking status (GS), and pack-years (GPY) were computed, yielding genomic sex, genomic smoking status, and genomic pack year measures for each patient. These three genomic measures were used as new covariates to help in selecting genes with lung-cancer associated gene expression and in the lung cancer classifier (described below).

### Selection of lung cancer genes

A logistic regression model with lung cancer status (1 = cancer-positive and 0 = cancer-negative) as the dependent variable was fit using the training data, CFGC's, and patient age as predictors. This model served as the "baseline" for subsequent gene expression analysis.

Next an empirical Bayes linear model was fit using gene expression values as the independent variable and the logistic regression baseline model residuals as the dependent variable. The residuals from this baseline model are a measure of patient cancer status that could not be predicted on the basis of clinical factors or their genomic correlates alone. That is, the empirical Bayes linear model was used to select genes with predictive potential for lung cancer independent or additive to that represented by clinical covariates. We note that a gene

associated with both clinical factors and cancer could still be selected if the cancer association retained significance in this model. The top lung cancer-associated genes from this analysis were grouped using hierarchical clustering. To reduce the number of genes, for each cluster we selected a small number (2–4) of genes whose average was highly correlated to the average of all genes in the cluster. Subsequent modeling used these "reduced" cluster mean expression values rather than individual gene expression values. Cross validation was used to select which cluster means were independently significantly associated with lung cancer in the context of the other clusters. Overall, this served to select clusters that cumulatively provided the best classifier performance, and specific genes that best represented each of these clusters in a parsimonious manner. Functional analysis of genes within each of the cancer clusters was performed using DAVID [16] to identify biological terms describing the cancer-associated genes in the classifier.

### Lung cancer classifier

A lung cancer classifier was developed using lung cancer status as the outcome variable and a) the cancer associated gene expression cluster means, b) patient age, c) genomic gender (GG), d) genomic smoking status (GS), and e) genomic pack years (GPY) as predictors. The model was fit using a penalized logistic regression model; the penalization factor ( $\lambda$ ) was 0 for the clinical/ gene expression correlates and 10 for each of the gene expression cluster means. The resulting model score is on a 0 to 1 scale. A score threshold for predicting lung cancer status was established to achieve a sensitivity of approximately 90% for patients with a non-diagnostic bronchoscopy. An evaluation of the benefit of the gene expression classifier to predict lung cancer compared to clinical factors alone was performed by generating a "clinical model" that included age, gender, smoking status, and pack-years (all determined clinically) in a logistic regression model to predict lung cancer status. The difference in performance between the complete gene expression classifier and the clinical factors classifier to predict lung cancer status was assessed by comparing the AUC's of each model in the training set.

### Analysis of an independent test set

Data from a prior study [13] were used as an independent test set to assess the performance of the locked classifier derived in this study. In that study BECs were collected at bronchoscopy from patients undergoing bronchoscopy for suspicion of lung cancer, and RNA was analyzed on microarrays (Affymetrix HG-U133A). CEL files from that study ( $n = 163$ ) were re-normalized to produce gene-level expression values using Robust Multiarray Average (RMA) [14] in the Bioconductor R

package (version 1.28.1). This processing used the Entrez Gene-specific probeset chip definition file (CDF) [17] in place of the standard U133A CDF provided by Affymetrix in order to facilitate cross-platform analyses. Analyses were performed using the R environment for statistical computing (version 2.9.2).

The classifier was applied to patients in the test set with two modifications to account for the difference in microarray platforms. First, the HG-U133A RMA expression values were adjusted by adding a gene-wise constant defined as the difference between the mean of the test set samples and the mean of the training set samples, separately for each gene. This procedure functioned to shift the mean of each gene's expression levels in the test set to the mean observed in the training set. Second, for the classifier genes where a corresponding HG-U133A probeset was not available (LYPD2 and RNF150), the gene's mean expression value in the training set was used for all of the test set samples.

### Statistical methods

Classifier accuracy was assessed using standard measures of prediction accuracy: the area under the curve (AUC), sensitivity, specificity, NPV and PPV. Cross-validation, using a 10% sample hold-out set, was used in the training set to estimate the performance of the prediction classifiers generated using these approaches [18]. These performance estimates were used to guide the development of the classifier discovery procedure. A final model was set prior to performing a one-time analysis of the test set. Fisher's exact test was used to calculate statistical significance of all categorical variables (i.e., sex, smoking status, race, mass size, and mass location) and a t-test was used for continuous variables (i.e., age and smoking history).

## Results

### Study populations

A set of 299 patients from AEGIS 1 consisting of 223 patients diagnosed with lung cancer and 76 patients diagnosed with benign disease (Table 1) were used to derive our gene expression classifier. Characteristics of the independent test set have been previously described [13], and are summarized here (Additional file 4). Although the study design was similar to the one described here, there were some differences in the study populations. The patients were older on average in the training set compared to the test set ( $p < 0.001$ ) (although there was no significant difference in age ( $p = 0.959$ ) for patients diagnosed with lung cancer). The training set also consisted of fewer current smokers ( $p = 0.050$ ); and a lower proportion of patients with  $< 3$  cm lesions ( $p < 0.001$ ). In addition, the prevalence of lung cancer was higher in the training set (75% versus 48%;  $p < 0.001$ ).

### Derivation of the classifier and evaluation of performance

Gene expression was associated with current smoking status for a large fraction of the genes on the array (6477 genes with  $p < 0.001$ ; top 10 genes reported in Additional file 5). Three of the top ranked genes (SLC7A11, TKT, and CLND10) were selected to serve as a logistic regression-based smoking status classifier based on cross-validation. This smoking status classifier had an AUC of 0.93 within the training set. An additional CFGC was derived for smoking history, independent of smoking status, and was based on cumulative smoke exposure, measured in pack-years. Smoking history ( $< 10$  PY vs  $> 10$  PY) was significantly associated with the expression of 531 genes ( $p < 0.001$ ; top 10 genes reported in Additional file 6). Two of the top genes were selected to serve as a logistic regression-based smoking history classifier (RUNX1T1, AKR1C2) which had an AUC of 0.78 within the training set. Sex was significantly associated with 339 genes ( $p < 0.001$ ; top 10 genes reported in Additional file 7). The top ranked gene (RPS4Y1) was a perfect classifier (AUC = 1) of sex within the training set.

As described in the methods, we identified genes whose expression is significantly associated with the residuals from the CFGC model for lung cancer. A total of 232 cancer associated genes (Additional file 8) met the significance criteria ( $T$  score  $> 2.7$ ). A pairwise correlation of the 232 genes followed by hierarchical clustering was examined to identify genes with similar expression patterns and partitioned the genes into 11 clusters (Figure 1). Since genes were correlated within each cluster, we hypothesized that the mean of a small set of genes within each cluster could be used to represent the cluster in a sparse manner. We optimized the classifier, using cross validation to estimate the AUC. We selected genes to represent the gene clusters whose expression was most strongly associated with lung cancer and determined that inclusion of clusters 1, 2, 4, 7, 9 and 10 gave the best AUC. We also determined that beyond 2–4 genes per cluster the performance of the test did not improve. In cross-validation, AUC = 0.80 (95% CI 0.75 – 0.84) for all patients in the training set ( $n = 299$ ); for the subset of patients with non-diagnostic bronchoscopy ( $n = 134$ ) the performance was similar (AUC = 0.81; 95% CI 0.74 – 0.87).

The final lung cancer classifier was then determined using the finalized classifier discovery procedure on the entire training set. The classifier consisted of a combination of the six cancer gene clusters (represented by 17 genes in total), patient age, and the gene expression correlates (GG, GS, GPY) (Table 2) as predictors. Dichotomous classification was performed using a score threshold of 0.65 (patients with scores  $\geq 0.65$  were predicted as cancer-positive and  $< 0.65$ , cancer-negative). The classifier had a sensitivity of 93% and specificity of 57% in the training set and there was no difference in the AUC of the

**Table 1 Clinical and demographic characteristics of the patients used to train the classifier**

Category	Sub-category	Lung cancer	Benign disease	p
N		223	76	
Sex	Female	97	26	0.178
	Male	126	50	
Age (median years)		65	56	<0.001
Race	Caucasian	168	59	0.757
	African-American	47	13	
	Other	5	3	
	Unknown	3	1	
Smoking status	Current	101	26	0.107
	Former	122	50	
Smoking history (median PY)		43	30	<0.001
Mass size	<2 cm	46	23	<0.001
	>2 to <3 cm	30	12	
	≥3 cm	122	19	
	ill-defined infiltrate	10	13	
	Unknown	15	9	
Mass location	Central	86	16	0.018
	Peripheral	60	30	
	Central & peripheral	60	18	
	Unknown	17	12	
Histology	Sub-type			
SCLC		40		
NSCLC		180		
	Adenocarcinoma	83		
	Squamous	73		
	Large cell	6		
	Mixed/undefined	18		
Unknown		3		
Histology	Stage			
SCLC	Limited	16		
	Extensive	18		
	Unknown	6		
NSCLC	1	28		
	2	16		
	3	42		
	4	62		
	Unknown	32		
Benign disease	Sub-category			

**Table 1 Clinical and demographic characteristics of the patients used to train the classifier (Continued)**

Alternative diagnosis	54
Infection	23
Sarcoid	14
Inflammation	7
Fibrosis	4
Other	4
Benign growths	2
Resolution/Stability	22

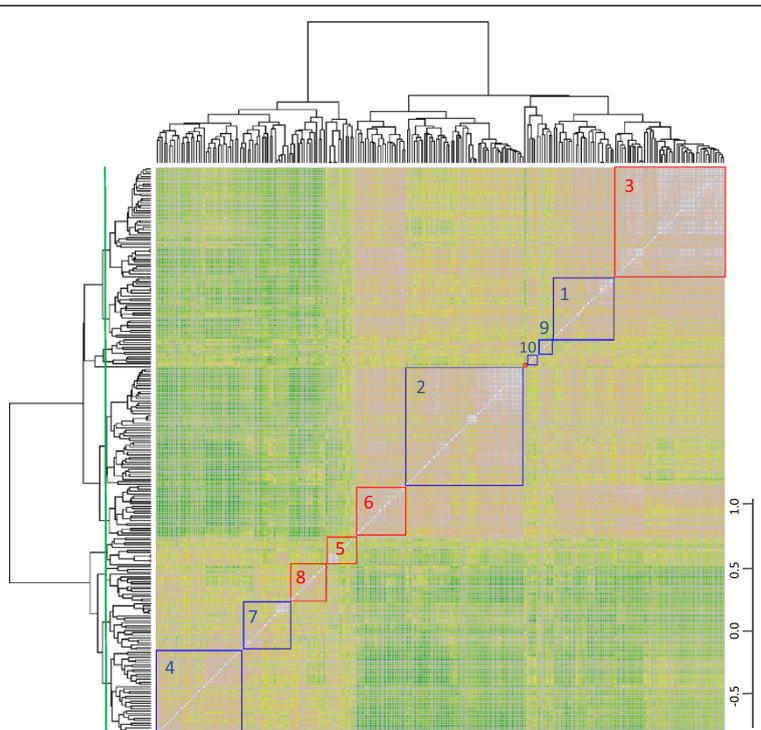
The classifier training set included 223 patients diagnosed with lung cancer and 76 patients diagnosed with benign disease. The table lists clinical and demographic factors for all patients in the training set as well as characteristics of the lung cancer positive and patients with benign disease. The p-value for race is calculated for Caucasian versus non-caucasian.

classifier for the entire training set (0.78; 95% CI, 0.73-0.82), compared with the subset of patients whose bronchoscopy was non-diagnostic for lung cancer (AUC = 0.78; 95% 0.71-0.85), (see, Additional file 9). We also found that there was no difference in the AUC ( $p = 0.62$ ) comparing Caucasians and African-Americans (the two predominant races in the training set), although the former smoked significantly more ( $p = 0.03$ ), with a mean PY difference of 46 versus 38, respectively (Additional file 10).

The gene expression classifier performed significantly better (AUC = 0.78; 95% CI, 0.73-0.82) than a model using clinical factors alone (AUC = 0.72; 95% CI, 0.67-0.77) in the training set ( $p < 0.001$ ). Functional analysis of the 17 cancer genes is summarized separately (Additional file 11). Nine of the genes are down-regulated and eight are up-regulated in association with cancer.

#### Validation in an independent test set

In the patients with non-diagnostic bronchoscopy ( $n = 123$ ) of the independent test set, the AUC of the classifier was 0.81 (95% CI, 0.73 – 0.88), (Figure 2) which was similar to the performance in patients with non-diagnostic bronchoscopy in the training set (AUC = 0.78; 95% 0.71-0.85;  $p = 0.495$ ). The sensitivity was 92% and with a specificity of 53%, the NPV was 94% (95% CI, 83-99%), (see Table 3). Interestingly we did not observe any effect of cancer histology, cancer stage (Table 4), or lesion size (Table 5) on the classifier's sensitivity for cancer. Moreover, in the test set the classifier had an AUC of 0.79 in current smokers and 0.82 in former smokers, suggesting that smoking status does not have a significant effect on classifier performance ( $p = 0.710$ ). When compared with bronchoscopy alone, the combination of the gene expression classifier with bronchoscopy improved the sensitivity from 51% to 96% ( $p < 0.001$ ).



**Figure 1** Pairwise correlation of genes with cancer-associated gene expression. The correlation between all possible pairs of genes with cancer-associated gene expression (n = 232) were assessed to identify groups of genes that share a similar pattern of gene expression. Unsupervised hierarchical clustering was used to group correlated genes into 11 clusters, with the dendrogram threshold level to establish clusters indicated on the y-axis (green line). Genes were selected from the clusters in a parsimonious manner to predict lung cancer status using linear regression. The classifier genes came from specific clusters (outlined in blue), using 2–4 genes from each cluster. Clusters 4 and 7 contain genes which were up-regulated in lung cancer, and clusters 1, 2, 9, and 10 were down-regulated in lung cancer.

**Table 2 Description of the gene expression classifier<sup>a</sup>**

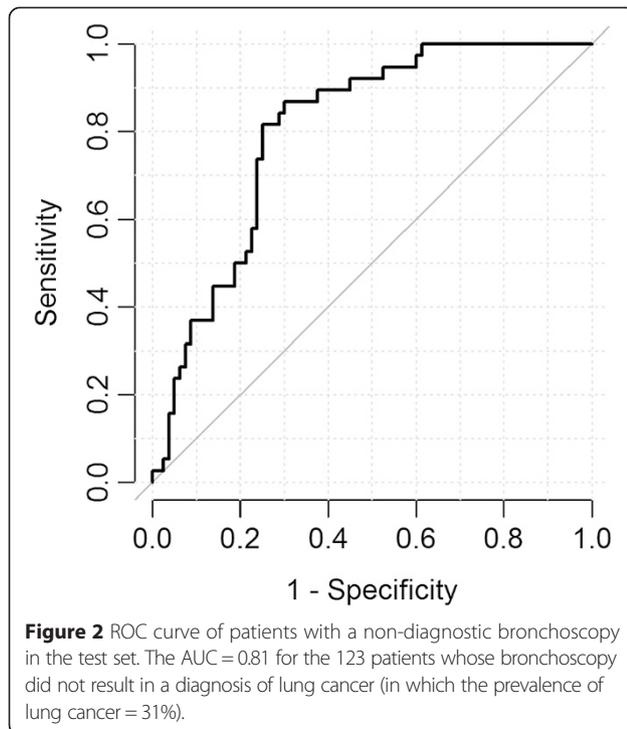
Feature <sup>b</sup> , (x <sub>i</sub> )	Coefficient, (b <sub>i</sub> )	Genes within features
Age	0.0623	
GG	0.5450	RPS4Y1
GS	0.1661	SLC7A11 CLND10 TKT
GPY	3.0205	RUNX1T1 AKR1C2
CA (1)	-0.4406	BST1 CD177.1 CD177.2
CA (2)	-0.3402	ATP12A TSPAN2
CA (4)	0.1725	GABBR1 MCAM NOVA1 SDC2
CA (7)	0.5670	CDR1 CGREF1 CLND22 NKX3-1
CA (9)	-0.3160	EPHX3 LYPD2
CA (10)	-0.3791	MIA RNF150
Intercept (b <sub>0</sub> )	3.3173	

<sup>a</sup>Genomic gender was defined as GG = 1 (female) if RPS4Y1 < 7.5, 0 (male) otherwise. The predicted genomic smoking (GS) value was derived, where  $x = 40.8579 - 0.4462 * SLC7A11 - 2.1298 * CLND10 - 1.8256 * TKT$ , and genomic smoking  $GS = e^x / (1 + e^x)$ . The predicted genomic pack years (GPY) value was derived, where  $x = -5.1429 + 2.1891 * RUNX1T1 - 0.9506 * AKR1C2$ , and genomic pack years  $GPY = \exp(x) / (1 + \exp(x))$ . The generalized equation for the prediction classifier was:  $Score = e^y / (1 + e^y)$ , where,  $y = b_0 + \sum(b_i * x_i)$ , where  $b_0$  is the intercept,  $b_i$  is the coefficient, and  $x_i$  is the feature (as shown).

<sup>b</sup>Features include patient age (as reported), GG, GS, GPY as described in the methods, and CA (i), the lung cancer gene clusters (shown in Figure 1).

## Discussion

Previous work has demonstrated that there are persistent gene-expression alterations in normal epithelial cells from the bronchial airway that are associated with exposure to cigarette smoke and the presence of lung cancer in current and former smokers [12,19-21]. These cancer-associated differences can be used to derive classifiers capable of accurately detecting lung cancer in these relatively non-invasively collected biospecimens obtained during bronchoscopy [13]. In current practice it is challenging to rule out lung cancer when bronchoscopy does not lead to a finding of malignancy, and the false-negative rate can range from 20-70% [5]. Current guidelines suggest that patients with elevated risk of disease should be pursued with more invasive follow-up diagnostic procedures [5], which carry increased risk of complications [8]. However due to uncertainty these procedures often performed in patients found to have benign disease [10,11]. Therefore our goal was to derive a gene-expression classifier using epithelial cells collected from the normal-appearing proximal airway during bronchoscopy that could be used in combination with bronchoscopy to increase the overall sensitivity and negative predictive value for lung cancer diagnosis. A classifier with



high sensitivity and high NPV among current or former smokers with a non-diagnostic bronchoscopy could serve to significantly reduce the probability of lung cancer in this clinical setting, reducing the use of additional unnecessary invasive procedures in smokers with benign lung lesions.

In this study, we leveraged a cohort of current and former smokers undergoing bronchoscopy for suspected lung cancer from a larger multicenter study to derive a gene-expression classifier for lung cancer. The classifier is a multivariate logistic regression model that has high sensitivity and high NPV. Importantly, we have validated the performance of the classifier in an independent cohort, using data from a previously published study of airway samples collected from smokers undergoing bronchoscopy for suspected lung cancer. The sensitivity is 92% in patients whose bronchoscopy is non-diagnostic in the test set with a specificity of 53%. The NPV is 94%

in the test set compared to an NPV of 69% for bronchoscopy alone suggesting that the classifier could help physicians reliably identify patients unlikely to have lung cancer after a non-diagnostic bronchoscopy. Given the different microarray platform (Affymetrix HU133A) used in the test set, the analysis was done with a classifier lacking *LYPD2* and *RNF150* since these genes were not measured on that microarray platform. However, the two genes added to classifier performance during cross-validation in the training data and the full classifier has subsequently been validated in two additional datasets (manuscript in preparation).

The functions of the differentially expressed genes in the normal appearing airway epithelium in current and former smokers with lung cancer provide insight into the biology underlying the field of injury (see Additional file 8). Among genes that are suppressed, there are a number involved in the immune response, including *CD177* and *BST1*, suggesting an impaired immune response in the airway of smokers with lung cancer. The gene *TSPAN2*, whose expression is depressed by p53 knockdown and is associated with poor prognosis in lung adenocarcinomas [22], was also expressed at lower levels in patients with cancer. Also *EPHX3*, a gene involved in xenobiotic metabolism, processing of carcinogens in tobacco smoke, and carcinogenesis in other epithelial cancers is down-regulated [23]. Among the classifier genes that are up-regulated in lung cancer, *NOVA1* and *CDR1* are predominantly expressed in neurons, but are also expressed in tumors and are associated with para-neoplastic antibodies in several malignancies, including small-cell lung cancer [24-28]. Furthermore, *MCAM* which is up-regulated in lung cancer, is expressed in basal bronchial epithelial cells [29]. *MCAM* is also strongly and transiently up-regulated in tracheal epithelium during repair [30], is required for tracheal epithelial regeneration [31], and is up-regulated in the bronchial epithelium of patients with COPD [32] and asthma [33]. A number of classifier genes that regulate cell growth and proliferation are up-regulated in patients with lung cancer, including *SDC2*, and *NKX3-1* as well as the cell-cycle-arrest mediator *CGREF1*. Finally the *CFG* genes selected

**Table 3 Performance of bronchoscopy, classifier, and the combined procedures in the test set**

Category	Bronchoscopy	Classifier*	Classifier & bronchoscopy combined
N, total	163	123	163
N, Lung cancer	78	38	78
N, Benign disease	85	85	85
Sensitivity (95% CI)	51% (40-62%)	92% (78-98%)	96% (89-99%)
Specificity (95% CI)	100% (95-100%)	53% (42-63%)	53% (42-63%)
NPV (95% CI)	69% (60-77%)	94% (83-98%)	94% (83-98%)
PPV (95% CI)	100% (90-100%)	47% (36-58%)	65% (56-73%)

\*The performance of the classifier was evaluated in patients in which bronchoscopy did not result in a finding of cancer (n = 123).

**Table 4 Sensitivity of bronchoscopy, the classifier, and the combined procedures for patients with lung cancer in the test set**

Histology	Sub-type	N	Bronchoscopy sensitivity	Classifier sensitivity	Combined sensitivity
All cancers		78	51% <sup>a</sup>	92% <sup>b</sup>	96% <sup>c</sup>
SCLC		14	64%	100%	100%
NSCLC		64	48%	91%	95%
	Adenocarcinoma	18	33%	83%	89%
	Squamous	27	56%	92%	96%
	Large cell	4	25%	100%	100%
	Undefined	15	60%	83%	93%
Histology	Stage				
SCLC	Limited	9	78%	100%	100%
	Extensive	5	40%	100%	100%
NSCLC					
	1	14	36%	100%	100%
	2	2	50%	100%	100%
	3	25	52%	92%	96%
	4	22	55%	80%	91%
	Unknown	1	0%	100%	100%

Of 163 patients who underwent a diagnostic bronchoscopy procedure for suspicion of lung cancer, 78 were diagnosed with cancer. A lung cancer diagnosis was made at bronchoscopy (a) in 40 patients (51%; 95% CI, 40-62%), and in the remaining lung cancer patients where no diagnosis was made at bronchoscopy, (b) the classifier correctly predicted 34 of them (89%; 95% CI, 75-96%). The classifier combined with bronchoscopy (c) yielded a detection of 74 of 78 (95%; 95% CI, 87-98%) patients with lung cancer. The sensitivities of bronchoscopy, the classifier, and the combined procedures are also shown for lung cancers according to sub-type and stage.

to predict smoking status (SLC7A11, CLDN10, TKT) and smoking history (RUNX1T1, AKR1C2) in our classifier have been previously reported as being altered by tobacco smoke exposure, confirming the robust effect of smoking on airway epithelium biology [12,19,34].

Our discovery approach extends earlier work on gene-expression based lung cancer diagnostics [13] primarily in the explicit modeling of clinical covariates as components of the predictive model prior to selection of features with lung cancer-associated expression. It is known that the response to environmental insults and other clinical factors can vary substantially between individuals. Therefore our approach was to use gene expression to capture the patient-level physiological response to an environmental insult (e.g., cumulative smoke exposure), as this response

**Table 5 Sensitivity of bronchoscopy, the classifier, and the combined procedures in the test set stratified by size of suspicious lesions**

Mass size*	N	Bronchoscopy sensitivity	Classifier sensitivity	Combined sensitivity
<3 cm	99	44%	87%	93%
>3 cm	48	58%	94%	98%
Ill-def Infiltrate	16	38%	100%	100%

\*Includes patients diagnosed with lung cancer and those with benign disease.

may be more reflective of disease risk than the actual reported values [35]. Additionally, the use of gene expression data to predict critical data inputs (such as patient sex) minimizes potential for data entry errors in clinical practice. Another component of our approach was selecting genes whose expression is associated with cancer after accounting for the modeled clinical factors. We hypothesized that this approach would help ensure that the information about the likelihood of cancer captured by the genes with cancer-associated gene expression is independent from the information about cancer captured by the modeled clinical factors. An additional important aspect of our classifier discovery approach was our methodology to identify patterns of independent cancer-associated gene expression through clustering and then to model cancer as the additive effects of each of the cancer-associated gene expression modules. This is in contrast to selecting only genes that are globally top-ranked according to their association with cancer which could potentially result in selecting an entire panel of genes that reflect a single cancer-associated molecular process. Previous studies to derive a gene expression classifier to predict risk of lung cancer in normal appearing airway epithelial cells have described similar results with high sensitivity and NPV when bronchoscopy is non-diagnostic [13]. While there are no common genes in that

classifier compared to the one described here, we believe that our new classifier represents similar mechanisms of action given the strong performance in the independent test set. However, the differences in the specific genes selected in the classifier described here may be due to differences in the feature selection process, specifically, the method of accounting for gene expression strongly associated with clinical covariates while selecting cancer genes.

## Conclusion

We have derived a gene expression classifier for lung cancer in current and former smokers using cells from the proximal airway that can be used in conjunction with bronchoscopy for suspected lung cancer. We have validated the performance of this classifier in an independent test set. The classifier adds substantial sensitivity to the bronchoscopy procedure resulting in high NPV. This classifier can be used to aid in decision-making when bronchoscopy is non-diagnostic by identifying patients who are at low risk of having lung cancer.

## Additional files

**Additional file 1:** Supplemental Methods.

**Additional file 2:** Ethics committees of the AEGIS 1 study.

**Additional file 3:** Medical centers in which the AEGIS 1 clinical cohort was enrolled.

**Additional file 4:** Demographic and clinical characteristics of the test set patients.

**Additional file 5:** Top differentially expressed genes associated with smoking status.

**Additional file 6:** Top differentially expressed genes associated with smoking history.

**Additional file 7:** Top differentially expressed genes associated with gender.

**Additional file 8:** Genes associated with cancer which are included in the classifier.

**Additional file 9:** ROC curve analysis of the training set cohort using the finalized gene expression classifier.

**Additional file 10:** Analysis of clinical characteristics and classifier performance according to race.

**Additional file 11:** Biological characterization of classifier genes.

## Competing interests

DW and KP were formerly employed by Allegro Diagnostics (the sponsor of the AEGIS trials) and are now employees of Veracyte, Inc. AS and JB were founders of and consultants to Allegro and inventors on patents licensed to the company and AS has received in-kind research support from Allegro. ML has received consulting fees and in-kind research support from Allegro and is an inventor on patents that have been licensed to Allegro. ME has received consulting fees from Allegro. AV, JSF, GS, ML and AS have previously served on a scientific advisory board to Allegro.

## Authors' contributions

DW, ME and ML designed the analysis plan. DW, ME, ML, and AG performed data analysis. KP, AS, and JB designed the clinical study. SF, AV, GS, and AS provided clinical oversight and case review. DW, ML, AV, GS, AS, ME, and AG wrote the manuscript. All authors read and approved the final manuscript.

## Acknowledgements

We thank all of the Principal Investigators (listed in Additional file 3) and coordinators of the participating AEGIS trial sites for patient recruitment and collection of specimens and clinical data. We thank Giulia Kennedy, Michael Rosenbluth, Edward Parsons, James Diggans, and Jing Huang for critical review of the data and the manuscript. Finally we thank all of the patients in the AEGIS trials for participating in the study.

The authors also acknowledge funding support for a portion of the study from NIH/NCI U01 CA152751.

## Author details

<sup>1</sup>Allegro Diagnostics, Corp, Maynard, MA, USA. <sup>2</sup>Elashoff Consulting, LLC, Redwood City, CA, USA. <sup>3</sup>Boston University School of Medicine, Boston, MA, USA. <sup>4</sup>University of Pennsylvania School of Medicine, Philadelphia, PA, USA. <sup>5</sup>University of Wisconsin School of Medicine and Public Health, Madison, WI, USA. <sup>6</sup>Medical University of South Carolina, Charleston, SC, USA. <sup>7</sup>Present affiliation: Veracyte, Inc, South San Francisco, CA, USA.

Received: 11 February 2015 Accepted: 23 April 2015

Published online: 06 May 2015

## References

- Howlader N, Noone AM, Krapcho M, et al. SEER stat fact sheets: lung and bronchus. Available online: <http://seer.cancer.gov/statfacts/html/lungb.html>.
- The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with Low-dose computed tomographic screening. *N Engl J Med*. 2011;365:395–409.
- Tanoue LT, Tanner NT, Gould MK, Silvestri GA. Lung Cancer Screening. *Am J Respir Crit Care Med*. published online 04 Nov 2014
- Ernst A, Silvestri G, Johnstone D. Interventional pulmonary procedures: guidelines from the American college of chest physicians. *Chest*. 2003;123:1693–717.
- Rivera MP, Mehta AC, Wahidi MM. Establishing the diagnosis of lung cancer diagnosis and management of lung cancer, 3rd ed: American college of chest physicians evidence-based clinical practice guidelines. *CHEST Supplement*. 2013;143, e143S.
- Silvestri GA, Feller-Kopman D, Chen A, Wahidi M, Yasufuku K, Ernst A. Latest advances in advanced diagnostic and therapeutic pulmonary procedures update on pulmonary procedures. *CHEST J*. 2012;142(6):1636–44.
- Gildea TR, Mazzone PJ, Karnak D, Meziane M, Mehta AC. Electromagnetic navigation diagnostic bronchoscopy: a prospective study. *Am J Respir Crit Care Med*. 2006;174(9):982–9.
- Wiener RS, Schwartz LM, Woloshin S, Welch HG. Population-based risk for complications after transthoracic needle lung biopsy of a pulmonary nodule: an analysis of discharge records. *Ann Intern Med*. 2011;155(3):137–44.
- Fontaine-Delaruelle C, Ferretti G, Gamondes D, Pradat E, Souquet PJ, Couraud S. Is transthoracic core needle biopsy under CT scan a good deal for benign diseases' diagnosis? *Eur Res J*. 2014;44 Suppl 58:P679.
- Smith MA, Battafarano RJ, Meyers BF, Zoole JB, Cooper JD, Patterson GA. Prevalence of benign disease in patients undergoing resection for suspected lung cancer. *Ann Thorac Surg*. 2006;81:1824–8. discussion 1828–1829.
- Detterbeck FC, Mazzone PJ, Naidich DP, Bach PB. Screening for lung cancer: diagnosis and management of lung cancer: American College of Chest Physicians evidence-based clinical practice guidelines. *CHEST J*. 2013;143(5\_suppl):e78S–92.
- Beane J, Sebastiani P, Liu G, Brody JS, Lenburg ME, Spira A. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol*. 2007;8(9):R201.
- Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med*. 2007;13(3):361–6.
- 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(2):249–64.
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996;58(1):267–88.
- Da Wei Huang BTS, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2008;4(1):44–57.
- Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*. 2005;33(20):e175.

18. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286(5439):531–7.
19. Kadara H, Fujimoto J, Yoo SY, Maki Y, Gower AC, Kabbout M, et al. Transcriptomic architecture of the adjacent airway field cancerization in Non-small cell lung cancer. *J Natl Cancer Inst*. 2014;106:3.
20. Steiling K, Ryan J, Brody JS, Spira A. The field of tissue injury in the lung and airway. *Cancer Preven Res*. 2008;1(6):396–403.
21. Bossé Y, Postma DS, Sin DD, Lamontagne M, Couture C, Gaudreault N, et al. Molecular signature of smoking in human lung tissues. *Cancer Res*. 2012;72(15):3753–63.
22. Otsubo C, Otomo R, Miyazaki M, Matsushima-Hibiya Y, Kohno T, Iwakawa R, et al. TSPAN2 is involved in cell invasion and motility during lung cancer progression. *Cell Reports*. 2014;7(2):527–38.
23. Oster B, Thorsen K, Lamy P, Wojdacz TK, Hansen LL, Birkenkamp-Demtröder K, et al. Identification and validation of highly frequent CpG island hypermethylation in colorectal adenomas and carcinomas. *Int J Cancer*. 2011;129(12):2855–66.
24. Knudsen A, Monstad SE, Dørum A, Lønning PE, Salvesen HB, Drivsholm L, et al. Ri antibodies in patients with breast, ovarian or small cell lung cancer determined by a sensitive immunoprecipitation technique. *Cancer Immunol Immunother*. 2006;55(10):1280–4.
25. Buckanovich RJ, Posner JB, Darnell RB. Nova, the paraneoplastic Ri antigen, is homologous to an RNA-binding protein and is specifically expressed in the developing motor system. *Neuron*. 1993;11(4):657–72.
26. Salemi M, Fraggetta F, Galia A, Pepe P, Cimino L, Condorelli RA, et al. Cerebellar degeneration-related autoantigen 1 (CDR1) gene expression in prostate cancer cell lines. *Int J Biol Markers*. 2013;29(3):e288–90.
27. Tanaka M, Tanaka K, Onodera O, Tsuji S. Trial to establish an animal model of paraneoplastic cerebellar degeneration with anti-Yo antibody: 1. Mouse strains bearing different MHC molecules produce antibodies on immunization with recombinant Yo protein, but do not cause Purkinje cell loss. *Clin Neurol Neurosurg*. 1995;97(1):95–100.
28. Furneaux HM, Rosenblum MK, Dalmau J, Wong E, Woodruff P, Graus F, et al. Selective expression of Purkinje-cell antigens in tumor tissue from patients with paraneoplastic cerebellar degeneration. *New Engl J Med*. 1990;322(26):1844–51.
29. Shih IM, Nesbit M, Herlyn M, Kurman RJ. A new Mel-CAM (CD146)-specific monoclonal antibody, MN-4, on paraffin-embedded tissue. *Modern Pathol*. 1998;11(11):1098–106.
30. Tsukamoto Y, Taira E, Miki N, Sasaki F. The role of gicerin, a novel cell adhesion molecule, in development, regeneration and neoplasia. *Histol Histopathol*. 2001;16(2):563–71.
31. Tsukamoto Y, Taira E, Kotani T, Yamate J, Wada S, Takaha N, et al. Involvement of gicerin, a cell adhesion molecule, in tracheal development and regeneration. *Cell Growth Differ*. 1996;7(12):1761–7.
32. Schulz C, Petrig V, Wolf K, Krätzel K, Köhler M, Becker B, et al. Upregulation of MCAM in primary bronchial epithelial cells from patients with COPD. *Eur Respir J*. 2003;22(3):450–6.
33. Simon GC, Martin RJ, Smith S, Thaikootathil J, Bowler RP, Barenkamp SJ, et al. Up-regulation of MUC18 in airway epithelial cells by IL-13: implications in bacterial adherence. *Am J Resp Cell Mol Biol*. 2011;44(5):606–13.
34. Penning TM, Lerman C. Genomics of smoking exposure and cessation: lessons for cancer prevention and treatment. *Cancer Preven Res*. 2008;1(2):80–3.
35. Lampe JW, Stepaniants SB, Mao M, Radich JP, Dai H, Linsley PS, et al. Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. *Cancer Epidemiol Biomarkers Prev*. 2004;13(3):445–53.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

