HHS Public Access

Author manuscript

Cancer Prev Res (Phila). Author manuscript; available in PMC 2019 September 24.

Published in final edited form as:

Cancer Prev Res (Phila). 2017 November ; 10(11): 651–659. doi:10.1158/1940-6207.CAPR-17-0098.

†To whom correspondence should be addressed. **Corresponding authors:** Avrum Spira, aspira@bu.edu; Marc Lenburg, mlenburg@bu.edu; Ana Pavel, anapavel@bu.edu; Section of Computational Biomedicine, Boston University School of Medicine, 72 East Concord Street, E601, Boston, MA.

*Contributed equally as co-senior authors

‡A complete list of investigators in the AEGIS (Airway Epithelium Gene Expression in the Diagnosis of Lung Cancer) Study Team is provided at the end of the manuscript.

Author's contribution

A.S., M.E.L. conceived the study. A.B.P. performed the primary analyses. J.D.C. contributed to the sequencing data analysis. G.L. performed the experiments. M.E.L., D.E. advised statistical analyses. K.S., D.W. provided the samples. A.B.P., J.D.C, G.L., D.E., S.D., K.S., D.W., M.E.L., A.S. contributed to the writing and revising of the manuscript.

The AEGIS Study Team

Beth Israel Deaconess Medical Center, Boston MA

Principal Investigator: Armin Ernst and Gaetane Michaud; Co-Investigators: Adnan Majid, Sidharta Pena Gangadharan, Andres Sosa, Renelle Myers, Michael Kent, Malcolm DeCamp, Dilip Nataraj, Samaan Rafeq, David Berkowitz, Saleh Alazemi, Robert Garland; Study Coordinators: Arthur Dea, Paula Mulkern, Christina Carbone

Cleveland Clinic, Cleveland OH

Principal Investigator: Tom Gildea; Co-Investigators: Francisco Almeida, Joseph Cicenia, Mike Machuzak; Study Coordinators: Meredith Seeley

Columbia University, New York, NY

Principal Investigator: Charles Powell; Co-Investigators: William Bulman, Joshua Sonett, Rebecca Toonkel; Study Coordinators: Kivilcim Sungur-Stasik

Georgia Clinical Research, Austell GA

Principal Investigator: Stuart Simon; Co-Investigators: Chad Case, Alexander Gluzman, Charles Hartley, Steven Harris, Aristidis Iatridis, Jermaine Jackson, C. Coy Lassiter, Brion Lock, Kathryn McMinn, Chad Miller, Sriram Paramesh, Craig Patterson, Brett Sandifer, Samuel Szumstein, James Waldron, Christy Wilson, Paul Zolty, Stephen Strazay, Ashley Waddell; Study Coordinators: Betsy Rambo, Monica Haughton, Stacy Beasley, Penny Murray, Debra Yeager

Indiana University, Indianapolis, IN

Principal Investigator: Francis Sheski; Co-Investigators: Praveen Mathur; Study Coordinators: MaryAnn Caldwell, Annette Hempfling

Jamaica Hospital Medical Center, Jamaica, NY

Principal Investigator: Craig Thurm; Co-Investigators: Aradhana Agarwal, Akash Ferdaus; Study Coordinators: Kelly Cervellione

Louisiana State University, New Orleans, LA

Principal Investigator: Stephen Kantrow; Co-Investigators: Susan Gunn, David Welsh, Jennifer Ramsey, Jaime Palomino, Richard Tejedor; Study Coordinators: Connie Romaine

Medical University of South Carolina, Charleston, SC

Principal Investigator: Gerard Silvestri; Co-Investigators: Nicholas James Pastis, Nichole Tripician Tanner, Peter Doelken, John Terrill Huggins: Study Coordinators: Jeffrey Waltz, Katherine Taylor, Kalon Eways

National Jewish Health, Denver, CO

Principal Investigator: Ali Musani; Co-Investigators: David Hsia, Joseph Seaman, Justin Thomas: Study Coordinators: Phillip Lopez, Jami Henriksen

New York University, New York, NY

Principal Investigator: William Rom; Co-Investigators: Eric Leibert, Derrick Raptis, James Tsay, Robert Lee, Eric Bonura; Study Coordinators: Katie Schliessman, Ellen Eylers

North Florida/South Georgia Veterans Health System, Gainesville, FL

Principal Investigator: Peruvemba Sriram; Study Coordinators: Ana Thomas, Katherine Herring, Carmen Lowell

Overlake Hospital, Bellevue, WA

Principal Investigator: Amy Markezich; Co-Investigators: James Copeland, Eric Gottesman, Todd Freudenberger, William Watts: Study Coordinators: Tina Fortney

Pulmonary Associates, P.A., Phoenix, Arizona

Principal Investigator: Mark Gotfried; Co-Investigators: Robert Comp, Andreas Kyprianou, James Ross, Ronald Servi; Study Coordinators: Li Yi Fu, Sherry Harker

Pulmonary and Allergy Associates, P.A., Summit, NJ

Principal Investigator: Robert Sussman; Co-Investigators: Donatella Graffino, Mark Zimmerman, Robert Restifo, Vincent Donnabella, Federico Cerrone, John Oppenheimer, Robert Capone, Jaime Cancel, Edward Dimitry, Matthew Epstein, Sue Fessler, Erwin Oei, Frederic Scoopo, Chirag Shah; Study Coordinators: Virginia Hala, Kathy Izzo, Marissa Reinton-Lim, Hazel Scherb, MaryAnn Constantino

St. Elizabeth's Medical Center, Brighton, MA

Principal Investigator: Samaan Rafeq and Armin Ernst; Co-Investigators: Ali Ashraf, Antonio DeGorordo Arzamende, Deirdre Keogh, Ryan Chua, Ali Khodabandeh; Study Coordinators: Arthur Dea, Paula Mulkern

St. James's Hospital, Trinity College, Dublin, Ireland

Principal Investigator: Joe Keane; Study Coordinators: Jennifer Winkles, Eliot Woodward

Temple University, Philadelphia, PA

Alterations in bronchial airway microRNA expression for lung cancer detection

Ana B. Pavel[†,1,2], **Joshua D. Campbell**[1,2], **Gang Liu**[2], **David Elashoff**[3], **Steven Dubinett**[3], **Kate Smith**[4], **Duncan Whitney**[4], **Marc E. Lenburg**[*,†,1,2], **Avrum Spira**[*,†,1,2] **AEGIS Study Team**[‡]

[1]The Graduate Program in Bioinformatics, Boston University, Boston, MA

[2]Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA

[3]University of California Los Angeles, Los Angeles, CA

[4]Veracyte, South San Francisco, CA

Abstract

We have previously shown that gene-expression alterations in normal-appearing bronchial epithelial cells can serve as a lung cancer detection biomarker in smokers. Given that microRNAs regulate airway gene-expression responses to smoking, we evaluated whether microRNA-

Principal Investigator: John Travaline; Co-Investigators: Peter Bercz, Wissam Chatila, Brian Civic, Francis Cordova, Gerard Criner, Gilbert D'Alonzo, Victor Kim, Samuel Krachman, Albert Mamary, Nathaniel Marchetti, Aditi Satti, Kartik Shenoy, Irene Permet Swift, Maria-Elena Vega Sanchez, Sheila Weaver, Nicholas Panetta, Parag Desai, Fred Kueppers, Namrata Patel, Kathleeen Brennan, Alex Swift, David Ciccolella, Fred Jaffe, Jamie Lee Garfield; Study Coordinators: Carla Grabianowski, Carolina Aguiar
University of Alabama, Birmingham, AL
Principal Investigator: Mark Dransfield; Study Coordinators: Sherry Tidwell
University of British Columbia, Vancouver, BC, Canada
Principal Investigator: Stephen Lam; Co-Investigators: Annette McWilliams; Study Coordinators: Sharon Gee
University of California- Davis, Sacramento, CA
Principal Investigator: Richart Harper; Co-Investigators: Ken Yoneda, Jason Adams, Katherine Cayetano, Andrew Chan, Heba Ismail, Charles Poon, Rokhsara Rafii, Christian Sebat, Yasmeen Shaw, Matthew Sisitki, Will Tseng; Study Coordinators: Maya Juarez, Kaitlyn Kirk, Claire O'Brien
University of Missouri, Columbia, MO
Principal Investigator: Vamsi Guntur; Co-Investigators: Normand Caron, Harjyot Sohal, Casey Stahlheber, Danish Thameem, Shilpa Patel, Ousama Dabbagh, Rajiv Dhand, Rachel Kingree, Yuji Oba, Jason Goodin; Study Coordinators: Marta Fuemmeler, Angie Vick, Michel O'Donnell
University of Pennsylvania, Philadelphia, PA
Principal Investigator: Anil Vachani; Co-Investigators: Andrew Haas Colin Gillespie, Daniel Sterman,; Study Coordinators: Kristina Maletteri, Karen Dengel
University of Virginia, Charlottesville, VA
Principal Investigator: George Verghese; Co-Investigators: Cynthia Brown, Elizabeth Gay, Borna Mehrad, Manojkumar Patel, Mark Robbins, C. Edward Rose, Max Weder, Kyle Enfield; Study Coordinators: Peggy Doherty
University of Wisconsin, Madison, WI
Principal Investigator: Scott Ferguson; Co-Investigators: Mark Regan, Jennifer Bierach; Study Coordinators: Michele Wolff
Vanderbilt University, Nashville, TN
Principal Investigator: Pierre Massion; Co-Investigators: Alison Miller; Study Coordinators: Gabe Garcia, Anna Ostrander, Wendy Cooper, Willie Hudson
Virginia Commonwealth University, Richmond, VA
Principal Investigator: Wes Shepherd; Co-Investigators: Hans Lee, Rajiv Malhotra, Ashutosh Sachdeva; Study Coordinators: Christine DeWilde, Anna Priday
William Jennings Bryan Dorn VAMC
Principal Investigator: Brian Smith and Andrea Mass; Study Coordinators: Justin Reynolds, Andrea Peterson, Isaac Holmes
Yale University, New Haven, CT
Principal Investigator: Gaetane Michaud; Co-Investigators: Daniel Boffa, Frank Detterbeck, Kelsey Johnson, Anthony Kim, Jonathan Puchalski, Lynn Tanoue, Kyle Bramley; Study Coordinators: Christina Carbone

expression is also altered in the bronchial epithelium of smokers with lung cancer. Using epithelial brushings from the mainstem bronchus of patients undergoing bronchoscopy for suspected lung cancer (as part of the AEGIS-1/2; clinical trials), we profiled microRNA-expression via small-RNA sequencing from 347 current and former smokers for which gene-expression data was also available. Patients were followed for one year post-bronchoscopy until a final diagnosis of lung cancer (n=194) or benign disease (n=153) was made. Following removal of 6 low-quality samples, we used 138 patients (AEGIS-1) as a discovery set to identify four microRNAs (miR-146a-5p, miR-324–5p, miR-223–3p, miR-223–5p) that were down-regulated in the bronchial airway of lung cancer patients (ANOVA $p<0.002$, FDR<0.2). The expression of these microRNAs is significantly more negatively-correlated with the expression of their mRNA-targets than with the expression of other non-target genes (K-S $p<0.05$). Further, these mRNA-targets are enriched among genes whose expression is elevated in cancer patients (GSEA FDR<0.001). Finally, we found that the addition of miR-146a-5p to an existing mRNA-biomarker for lung cancer significantly improves its performance (AUC) in the 203 samples (AEGIS-1/2) serving an independent test set (DeLong $p<0.05$). Our findings suggest that there are microRNAs whose expression is altered in the cytologically-normal bronchial epithelium of smokers with lung cancer, and that they may regulate cancer-associated gene expression differences.

Keywords

Lung cancer; Airway; Biomarker; microRNA; Sequencing

Introduction

Lung cancer remains the leading cause of cancer death in the US and the world due, in large part, to our inability to detect the disease at its earliest and curable stage. Once a pulmonary lesion is identified, physicians must decide between CT surveillance vs. airway/lung biopsy. When biopsy is required, the approach can include bronchoscopy, transthoracic needle biopsy (TTNB), or surgical lung biopsy (SLB). The choice among these procedures is determined on the basis of considerations such as lesion size and location, the presence of adenopathy, the risk associated with the procedure, and local expertise. While bronchoscopy is relatively safe (less than 1% of procedures complicated by pneumothorax[1]), this procedure is limited by its sensitivity (from 34 to 88%), depending on the location and size of the lesion[2]. Even with newer bronchoscopic guidance techniques, the sensitivity for the detection of lung cancer is below 70% for peripheral lesions[3].

A nondiagnostic bronchoscopy in this setting leads to a clinical dilemma as to which of these patients should undergo further invasive diagnostic testing (TTNB or SLB). To facilitate this clinical decision, we recently developed and validated a gene expression-based classifier that distinguishes between smokers with and without lung cancer using mRNA isolated from cytologically normal cells in the mainstem bronchus[4,5]. We demonstrated that this biomarker can improve the diagnostic sensitivity of bronchoscopy for lung cancer detection.

The ability to identify gene expression changes associated with cancer status in the normal appearing airway supports the idea of an *airway molecular field of injury* spanning the respiratory tract[6]. In this current study, we extend the *field of injury* concept to microRNAs. MicroRNAs are a class of small, noncoding RNAs that repress gene expression and protein translation of their targets by complementary binding to the 3' UTR of RNA transcripts. In addition, compared to mRNAs, microRNAs are thought to be more stable molecules, making them more easily measured in degraded tissues[7]. Previous studies have shown that smoking alters the expression of microRNAs in the bronchial airway epithelium[8,9]. We hypothesize that similar to mRNA, there might also be microRNA expression changes associated with the presence of lung cancer in bronchial epithelium from the mainstem bronchus that may play a role in regulating cancer-associated gene expression differences, and that integrating microRNA with gene expression could improve lung cancer detection.

Materials and Methods

Selection of patients

As previously described, over 1000 current and former smokers undergoing bronchoscopy for suspected lung cancer were enrolled in the Airway Epithelial Gene Expression in the Diagnosis of Lung Cancer (AEGIS) trials, two independent, prospective, multicenter, observational studies (registered as NCT01309087 and NCT00746759)[4,5]. Exclusion criteria for patients enrolled in AEGIS trials were age less than 21 years, no history of smoking (defined as having smoked <100 cigarettes), and a concurrent cancer diagnosis or history of lung cancer. All study protocols were approved by the institutional review board at each medical center and written informed consent was obtained from all patients prior to enrollment. Patients were followed prospectively for up to one year post bronchoscopy until a final diagnosis was obtained.

In this study, we profiled microRNA expression via small RNA sequencing for 347 AEGIS patients. In choosing patients to include in our study, we were limited by patients with a benign diagnosis and matched them approximately 1:1 with patients diagnosed with lung cancer. Moreover, we attempted to balance the cases and controls for smoking status, cumulative smoke exposure (pack-years), gender, and age. For all of the samples selected for small RNA sequencing, gene expression profiling of the large RNA fraction had been performed previously using Affymetrix Human Gene 1.0 ST arrays[4,5] and was available for data integration.

We assigned 138 (~ 40%) samples from AEGIS-1 to be used as a discovery set (Table 1); these samples were drawn exclusively from the training set previously used to develop the gene expression classifier[4,5]. The remaining 203 samples comprise our test set (Table 1) and consist exclusively of samples from the AEGIS-1 (n = 133) and AEGIS-2 (n=70) test sets that were previously used to validate the gene expression classifier[5].

High-throughput sequencing of small RNA

Based on our previous work on the effect of multiplexing on microRNA expression quantitation[10], we sequenced 347 samples in three batches by multiplexing 12 samples per

lane on an Illumina HiSeq 2000. 200 ng of total RNA from each sample was used for library preparation. The TruSeq Small RNA Sample Prep Kit (Illumina) was used for the first batch, while the NEBNext Multiplex Small RNA Library Prep Set (Illumina) was used for the second and third batches. RNA adapters were ligated to 3' and 5' ends of the RNA and the adapter-ligated RNA was reverse transcribed into single-stranded cDNA. The RNA 3' adapter was designed to target microRNAs and other small RNAs that have a 3' hydroxyl group resulting from enzymatic cleavage by Dicer or other RNA processing enzymes. The cDNA was then amplified by PCR, using a common primer and a primer containing one of 12 index sequences. The introduction of the six-base index tag at the PCR step allowed multiplexed sequencing of different samples in a single lane of a flowcell. A 0.5% PhiX spike-in was also added in all lanes for quality control. Each multiplexed library was hybridized to one lane of the four 8-lane High-Output single-read flow cells on a cBot Cluster Generation System (Illumina) using TruSeq Single-Read Cluster Kit (Illumina). The clustered flowcell was loaded onto a HiSeq 2000 sequencer for a multiplexed sequencing run which consists of a standard 36-cycle sequencing read with the addition of a 7-cycle index read.

MicroRNA alignment and quality control

To estimate microRNA expression we used a small RNA sequencing pipeline previously described[10]. Briefly, the 3′ adapter sequence was trimmed using the FASTX toolkit. Reads longer than 15 nt were aligned to hg19 using Bowtie v0.12.7 [11] allowing up to one mismatch and alignment to up to 10 genomic locations. MicroRNA expression was quantified by counting the number of reads aligning to mature microRNA loci (miRBase v20) using Bedtools v2.9.0[12,13]. MicroRNA counts within each sample were normalized to $\log_2$ RPM values by adding a pseudocount of one to each microRNA, dividing by the total number of reads that aligned to all microRNA loci within that sample, multiplying by 1×10^6, and then applying a $\log_2$ transformation[10]. The $\log_2$ RPM expression values follow a normal distribution by an Anderson-Darling test[14] ($p = 2.2 \times 10^{-16}$).

Next, we examined the distribution of read lengths present in each sample to ensure that the sequences we observed were of the proper length for microRNA. The read length distribution ought to follow a normal distribution with a mean of 22 bases. We filtered out samples whose distribution had an abundance of reads well below or above the mean of 22 bases (with less than one million reads aligned to 22 read length), indicating that the sample was not properly sequenced, the adapters were improperly trimmed, or the sample was of poor quality. Six such samples were removed, leaving 341 samples included in the downstream analysis. Additionally, we removed microRNA loci with a low number of aligned reads (less than 20 on average). A total of 463 microRNA loci passed our filter and were included in the analysis. Lastly, we applied ComBat[15] to normalize the microRNA expression in the three different batches. Large scale variability in microRNA expression was examined by Principal Components Analysis (PCA). No outlier samples were detected using the first two principal components, and there were no apparent global differences in miRNA expression between samples from AEGIS-1 and AEGIS-2 (Supplementary Figure 1).

Data availability

Raw FASTQ files as well as the normalized microRNA expression data are available on Gene Expression Omnibus (GEO) under the GEO accession number GSE93284. We used mRNA data from Whitney et al.[4,5] and Silvestri et al.[4,5] (GSE66499).

Differential expression analysis

To identify smoking-associated microRNAs, while correcting for covariates, we applied an F-test (*anova* R function)[16] between a multiple linear regression (*lm* R function), with microRNA expression as the response variable, and smoking status, age, gender, cancer status, and pack-years as independent variables, and another multiple linear regression that did not include the smoking status as an independent variable.

Similarly, to identify microRNAs with cancer-associated expression patterns in the discovery cohort, while correcting for covariates, we applied an F-test between a multiple linear regression, with microRNA expression as the response variable, and cancer status, age, gender, smoking status, and pack-years as independent variables, and another multiple linear regression that did not include the cancer status as an independent variable.

The p-values were adjusted for false discovery rate using Benjamini-Hochberg FDR[17], and are denoted with *q-value*.

Identifying microRNA-mRNA relationships

We analyzed the correlations between the differentially expressed microRNAs and their targets as predicted in the Targetscan database[18]. We included the conserved targets as defined in TargetScan 5 and 6 (8mer >= 0.8; 7mer-m8 >= 1.3; 7mer-1A >= 1.6). The probability of conserved targeting[19] has the advantage of identifying targeting interactions that are not only more likely to be effective but also those that are more likely to be consequential. Correlation coefficients were calculated using Pearson's product-moment coefficient. For each microRNA, we compared the resulting distribution of correlation coefficients to the distribution of correlation coefficients between the microRNA and all the genes that have not been predicted to be targeted by it in Targetscan, using the Kolmogorov-Smirnov (KS) test. Next, we tested whether the negatively correlated targets (correlation FDR<0.1) of each differentially expressed microRNA were enriched among the genes whose expression is associated with cancer status by Gene Set Enrichment Analysis (GSEA)[20]. For this enrichment analysis, genes were ranked by the t-statistic of a multiple linear regression, with microRNA expression as the response variable, and cancer status, age, gender, smoking status, and pack-years as independent variables.

Incorporating microRNA expression into the mRNA classifier

First, we calculated the prediction score of the mRNA classifier[4,5]. Then, for each cancer-associated microRNA, we integrated the mRNA classifier score with the microRNA's expression using logistic regression (*glmnet* R package). The coefficients of the logistic regression, corresponding to the intercept (α_0 = 1.8480041), weight of the classifier score (α_1 = 4.3879703), and weight of the microRNA's expression (α_2 = −0.3724577), were determined in the discovery set and the performance of the fully specified model was

evaluated in the independent test set samples. Classification performance was assessed using the area under the receiver operating characteristic curve (ROC AUC). The statistical significance of the AUC improvement was computed by DeLong test[21] from the *pROC* R package[22].

Results

Patient Population

microRNA expression was profiled via small RNA sequencing for 347 patients (194 cancer-positive and 153 cancer-negative subjects) participating in the AEGIS-1 and AEGIS-2 trials. Of the 347 microRNA samples, 341 passed the sequencing quality control filter[10]. The characteristics of the discovery set (138 samples) and the test set (203 samples) are shown in Table 1. Except for cancer status, the other clinical variables are not significantly different between the training and test sets. We also found significant associations between cancer status and age and lesion size in the discovery set and with pack-years and lesion size in the test set (Supplementary Table 1).

Identifying smoking-associated microRNAs in airway epithelium

Previous work has shown that cigarette smoke creates a molecular field of injury throughout the airway, and specifically that microRNA expression is altered with tobacco smoke exposure[9,23–28]. We therefore used the ability to detect microRNAs with smoking-status associated expression as a positive control for the quality of the microRNA expression data.

A set of 28 microRNAs were previously identified as modulators of smoking-related gene expression changes in airway epithelium[9], with most of them (n=23) being down-regulated in current smokers compared to never smokers. We found that the microRNAs previously identified as being repressed by smoking were significantly enriched among the microRNAs that were most down-regulated in current smokers from AEGIS (GSEA $q<0.001$), as shown in Figure 1.

In addition, using our data we identified significantly differentially expressed microRNAs between current and former smokers by linear regression. We found 135 smoking-associated microRNAs by $p<0.05$ (Supplementary Table 2). The top 30 differentially expressed microRNAs in the discovery set ($q<0.01$) are shown in Supplementary Figure 2. Among these, we found microRNAs whose expression has been previously associated with smoking, such as miR-218, miR-365, miR-30a and miR-99a[9].

We also evaluated the relationship between bronchial microRNA expression and other potentially relevant clinical variables such as gender, age and pack years (Supplementary Tables 3, 4, 5). We found that in addition to smoking status, gender is also associated with microRNA expression (85 differentially expressed microRNAs, $p<0.05$).

Cancer-associated microRNA alterations in the bronchial airway epithelium

Using the discovery set (n=138), we identified 42 microRNAs that showed differential expression between patients with and without cancer by linear regression at a liberal p-value threshold of $p < 0.05$ (Supplementary Table 6). Of these, four microRNA isoforms showed

evidence of differential expression at FDR < 0.2 ($p<0.002$). These four are: miR-146a-5p, miR-324–5p, miR-223–3p, miR-223–5p. The expression profiles of these four microRNAs are shown in Figure 2. Each of these miRNA has previously been reported to have tumor-suppressor-like activity[29–32]. Consistent with the potential for these microRNA to function as tumor suppressors, we find that the four differentially expressed microRNA isoforms are down-regulated in the bronchial airway of patients with lung cancer.

Cancer associated microRNAs as potential regulators of the airway gene-expression alterations

MicroRNAs often lead to the degradation of the mRNAs to which they bind. Therefore, we sought to determine if the expression of these microRNAs was negatively correlated with the expression of their gene targets. We found that the distribution of the correlation coefficients of each cancer-associated microRNA and its predicted mRNA targets (binding site predicted targets from Targetscan) is significantly more negative than the distribution of correlation coefficients for non-target genes ($p<10^{-9}$ for each microRNA) (Figure 3).

To begin to understand the potential biological impact of the cancer-associated expression of these microRNAs, we investigated whether the expression of their gene targets are associated with cancer. From the predicted targets (Targetscan), we identified the genes whose expression is significantly negatively correlated (correlation $q<0.1$) with the cognate microRNA. The negatively correlated predicted targets of each of the four microRNAs were significantly enriched amongst the genes whose expression increased in the airway epithelium of patients with cancer relative to those with a benign diagnosis (GSEA $q<0.001$) (Figure 4).

In addition, the set of genes predicted to be regulated by these four microRNAs (n=254 in total; Targetscan binding-site predicted targets and negatively correlated microRNA – mRNA expression) is enriched by DAVID[33] for cancer-associated pathways, such as signaling pathways regulating pluripotency of stem cells ($p=0.001$), pathways in cancer ($p=0.007$), the TGF-beta signaling pathway ($p=0.035$) and the Ras signaling pathway ($p=0.043$).

microRNA expression adds to mRNA in the detection of lung cancer

We next sought to assess whether bronchial microRNA expression could add to the performance of a mRNA biomarker for lung cancer we previously identified[4]. Using the training set samples, we used logistic regression to build five cancer-prediction models: one model contained the mRNA biomarker score alone, the other four models contained the mRNA biomarker score in combination with one of the four microRNAs we identified as having significant cancer-associated expression. Next, we compared the ROC-curve AUC of the mRNA biomarker alone to the four microRNA-containing models using a test set (Table 1; Supplementary Table 1) comprised of AEGIS-1 and AEGIS-2 samples that are independent of the AEGIS-1 samples used to identify the four microRNAs with cancer associated expression and independent of the samples used to develop the mRNA biomarker. We found that adding miR-146a-5p to the mRNA biomarker significantly improved the AUC in the test set, from 0.66 to 0.71 ($p=0.025$). The AUC of biomarkers incorporating

either miR-324–5p or either of the two isoforms of miR-223 was not significantly different than the AUC of the mRNA biomarker alone ($p>0.25$) in the test set. The performance metrics of each microRNA combined with the mRNA biomarker are provided in Supplementary Table 7.

Discussion

We have previously identified bronchial airway gene expression differences between patients with and without lung cancer and shown that they can be used as a biomarker with clinical utility in the setting of patients with inconclusive results following bronchoscopy for suspect lung cancer[4–6]. In this study, we wished to determine if microRNA expression might also be altered in the normal-appearing epithelium of the mainstem bronchus, whether these microRNA expression differences might play a role in regulating the observed gene expression differences, and whether lung cancer associated microRNAs might have the potential to aid in the detection of disease.

We identified four microRNA isoforms (miR-146a-5p, miR-324–5p, miR-223–3p, miR-223–5p) that have altered expression in the airway epithelium of patients with lung cancer. That all four microRNAs have decreased expression in the bronchial airway of lung cancer patients is consistent with prior studies which have found microRNAs with cancer-specific expression, mostly down-regulated, in tumors compared to normal tissue[34]. Intriguingly, all four of the microRNAs we identified have previously been implicated in tumor suppressive pathways. Specifically, miR-146a has been previously shown to inhibit cell growth, migration and EGFR signaling[29,30,35], while inducing apoptosis. Furthermore, miR-146a/b expression levels have been shown to be significantly elevated during senescence (a cellular program that irreversibly arrests the proliferation of damaged cells)[36]. miR-223 has been shown to function as a tumor suppressor in the Lewis lung carcinoma cell line by targeting insulin-like growth factor-1 receptor and cyclin-dependent kinase-2[32]; and miR-324 has been associated with nasopharyngeal cancer[31]. While microRNA expression differences have been well documented in tumors, our results are the first to demonstrate altered expression of not just these cancer-related microRNAs, but any microRNA in the bronchial airway of lung cancer patients.

We found that the expression of mRNAs which are predicted targets of these microRNAs is significantly negatively correlated suggesting that the expression of downstream genes is induced as a consequence of the cancer-dependent loss of microRNA expression. Moreover, predicted targets with negatively correlated expression profiles are enriched for genes involved in processes important for cancer, such as the pluripotency of stem cells, TGF-beta and Ras signaling pathways. Among the 50 significantly negatively correlated predicted targets of miR-146a-5p, we found APPL1. The protein encoded by APPL1 gene binds to many other proteins, including PIK3CA, RAB5A, DCC, AKT2, and adiponectin receptors, as well as proteins of the NuRD/MeCP1 complex – which are involved in cell proliferation and crosstalk between adiponectin and insulin signaling pathways[37,38]. Interestingly, we also observed a significantly negative correlation between miR-146a-5p and PIK3CA, suggesting that miR-146a-5p might modulate the PI3K/AKT pathway. In addition to the important role of PI3K/AKT pathway in cell death/survival, increased PI3K activity has been observed in

lung cancer[39] and has been shown to occur early and potentially be reversible in the airway of smokers with premalignancy[39,40]. The anti-correlation of these differentially expressed bronchial microRNAs with cancer-associated mRNA targets suggest their role as lung cancer-associated regulators of gene expression, and that they could potentially serve as biomarkers of disease.

We assessed each differentially expressed microRNA's ability to enhance the performance of an mRNA-based lung cancer biomarker and found that miR-146a-5p significantly improves performance. One possible explanation for why miR-223–3p and miR-223–5p did not improve biomarker performance is that one of their targets (SNCA) is already a component of the mRNA classifier, thus miR-223 expression might be substantially redundant with SNCA expression levels. If this hypothesis is correct, it would suggest that miR-146a adds to the biomarker's performance because the mRNA biomarker does not already capture miR-146a-related expression information.

In this study, we demonstrate for the first time the presence of a microRNA *field of injury* in the bronchial airway for lung cancer. We identify microRNAs that are known to play a role in cancer-related processes, and importantly, we demonstrate that a multi 'omics data integration approach may improve lung cancer detection.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported by grants from the NIH Early Detection Research Network (5U01CA152751 to A.S and M.E.L.), the Department of Defense (DOD W81XWH-11–2-0161 to A.S. and M.L), and the Boston University Coulter Award (0–057-281-A594–5 to A.S. and M.E.L.). This study was sponsored by Veracyte Inc. Boston University owns patents related to the subject matter of this manuscript.

We would also like to thank Jake Kantrowitz and Jessica Vick for their help in revising the early draft of the manuscript.

Financial support: This research was supported by grants from the NIH Early Detection Research Network (5U01CA152751 to A.S and M.E.L.), the Department of Defense (DOD W81XWH-11–2-0161 to A.S. and M.L), and the Boston University Coulter Award (0–057-281-A594–5 to A.S. and M.E.L.). This study was sponsored by Veracyte Inc.

References

1. Tukey MH, Wiener RS. Population-based estimates of transbronchial lung biopsy utilization and complications. Respir Med. 2012;106(11):1559–1565. doi:10.1016/j.rmed.2012.08.008. [PubMed: 22938740]
2. 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. 2013;143(5 Suppl):e142S–65S. doi:10.1378/chest.12-2353. [PubMed: 23649436]
3. Wang Memoli JS, Nietert PJ, Silvestri GA. Meta-analysis of guided bronchoscopy for the evaluation of the pulmonary nodule. Chest. 2012;142(2):385–393. doi:10.1378/chest.11-1764. [PubMed: 21980059]

4. Whitney DH, Elashoff MR, Porta-Smith K, et al. Derivation of a bronchial genomic classifier for lung cancer in a prospective study of patients undergoing diagnostic bronchoscopy. BMC Med Genomics. 2015;8(1):18. doi:10.1186/s12920-015-0091-3. [PubMed: 25944280]
5. Silvestri GA, Vachani A, Whitney D, et al. A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. N Engl J Med. 2015;373(3):243–251. doi:10.1056/NEJMoa1504601. [PubMed: 25981554]
6. Spira A, Beane JE, Shah V, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nat Med. 2007;13(3):361–366. doi:10.1038/nm1556. [PubMed: 17334370]
7. Etheridge A, Lee I, Hood L, Galas D, Wang K. Extracellular microRNA: a new source of biomarkers. Mutat Res. 2011;717(1–2):85–90. doi:10.1016/j.mrfmmm.2011.03.004. [PubMed: 21402084]
8. Perdomo C, Campbell JD, Gerrein J, et al. MicroRNA 4423 is a primate-specific regulator of airway epithelial cell differentiation and lung carcinogenesis. Proc Natl Acad Sci U S A. 2013;110(47): 18946–18951. doi:10.1073/pnas.1220319110. [PubMed: 24158479]
9. Schembri F, Sridhar S, Perdomo C, et al. MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium. Proc Natl Acad Sci U S A. 2009;106(7):2319–2324. doi:10.1073/pnas.0806383106. [PubMed: 19168627]
10. Campbell JD, Liu G, Luo L, et al. Assessment of microRNA differential expression and detection in multiplexed small RNA sequencing data. RNA. 2015;21(2):164–171. doi:10.1261/rna.046060.114. [PubMed: 25519487]
11. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25. [PubMed: 19261174]
12. Griffiths-Jones S The microRNA Registry. Nucleic Acids Res. 2004;32(Database issue):D109–11. doi:10.1093/nar/gkh023. [PubMed: 14681370]
13. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–842. doi:10.1093/bioinformatics/btq033. [PubMed: 20110278]
14. Thode HC. Testing for Normality. New York: CRC Press; 2002.
15. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–127. doi:10.1093/biostatistics/kxj037. [PubMed: 16632515]
16. Chambers J Linear Models. (Chambers J, Hastie T, eds.). Pacific Grove: Wadsworth & Brooks/Cole; 1992.
17. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57 1995:289–300.
18. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120(1):15–20. doi:10.1016/j.cell.2004.12.035. [PubMed: 15652477]
19. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2008;19(1):92–105. doi:10.1101/gr.082701.108. [PubMed: 18955434]
20. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–15550. doi:10.1073/pnas.0506580102. [PubMed: 16199517]
21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837. doi:10.2307/2531595. [PubMed: 3203132]
22. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):77. doi:10.1186/1471-2105-12-77. [PubMed: 21414208]
23. Powell CA, Klares S, O'Connor G, Brody JS. Loss of heterozygosity in epithelial cells obtained by bronchial brushing: clinical utility in lung cancer. Clin Cancer Res. 1999;5(8):2025–2034. http://www.ncbi.nlm.nih.gov/pubmed/10473082. Accessed September 30, 2016. [PubMed: 10473082]

24. Wistuba II, Lam S, Behrens C, et al. Molecular damage in the bronchial epithelium of current and former smokers. J Natl Cancer Inst. 1997;89(18):1366–1373. http://www.ncbi.nlm.nih.gov/pubmed/9308707. Accessed September 30, 2016. [PubMed: 9308707]
25. Franklin WA, Gazdar AF, Haney J, et al. Widely dispersed p53 mutation in respiratory epithelium. A novel mechanism for field carcinogenesis. J Clin Invest. 1997;100(8):2133–2137. doi:10.1172/JCI119748. [PubMed: 9329980]
26. Guo M, House MG, Hooker C, et al. Promoter hypermethylation of resected bronchial margins: a field defect of changes? Clin Cancer Res. 2004;10(15):5131–5136. doi:10.1158/1078-0432.CCR-03-0763. [PubMed: 15297416]
27. Miyazu YM. Telomerase Expression in Noncancerous Bronchial Epithelia Is a Possible Marker of Early Development of Lung Cancer. Cancer Res. 2005;65(21):9623–9627. doi:10.1158/0008-5472.CAN-05-0976. [PubMed: 16266979]
28. Spira A, Beane J, Shah V, et al. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci U S A. 2004;101(27):10143–10148. doi:10.1073/pnas.0401422101. [PubMed: 15210990]
29. Chen G, Umelo IA, Lv S, et al. miR-146a inhibits cell growth, cell migration and induces apoptosis in non-small cell lung cancer cells. PLoS One. 2013;8(3):e60317. doi:10.1371/journal.pone.0060317. [PubMed: 23555954]
30. Labbaye C, Testa U. The emerging role of MIR-146A in the control of hematopoiesis, immune function and cancer. J Hematol Oncol. 2012;5:13. doi:10.1186/1756-8722-5-13. [PubMed: 22453030]
31. Li G, Liu Y, Su Z, et al. MicroRNA-324–3p regulates nasopharyngeal carcinoma radioresistance by directly targeting WNT2B. Eur J Cancer. 2013;49(11):2596–2607. doi:10.1016/j.ejca.2013.03.001. [PubMed: 23583221]
32. Nian W, Ao X, Wu Y, et al. miR-223 functions as a potent tumor suppressor of the Lewis lung carcinoma cell line by targeting insulin-like growth factor-1 receptor and cyclin-dependent kinase 2. Oncol Lett. 2013;6(2):359–366. doi:10.3892/ol.2013.1375. [PubMed: 24137330]
33. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57. doi:10.1038/nprot.2008.211. [PubMed: 19131956]
34. Lu J, Getz G, Miska EA, et al. MicroRNA expression profiles classify human cancers. Nature. 2005;435(7043):834–838. doi:10.1038/nature03702. [PubMed: 15944708]
35. Kumaraswamy E, Wendt KL, Augustine LA, et al. BRCA1 regulation of epidermal growth factor receptor (EGFR) expression in human breast cancer cells involves microRNA-146a and is critical for its tumor suppressor function. Oncogene. 2015;34(33):4333–4346. doi:10.1038/onc.2014.363. [PubMed: 25417703]
36. Bhaumik D, Scott GK, Schokrpur S, et al. MicroRNAs miR-146a/b negatively modulate the senescence-associated inflammatory mediators IL-6 and IL-8. Aging (Albany NY). 2009;1(4):402–411. doi:10.18632/aging.100042. [PubMed: 20148189]
37. Mao X, Kikani CK, Riojas RA, et al. APPL1 binds to adiponectin receptors and mediates adiponectin signalling and function. Nat Cell Biol. 2006;8(5):516–523. doi:10.1038/ncb1404. [PubMed: 16622416]
38. Buechler C, Wanninger J, Neumeier M. Adiponectin receptor binding proteins – recent advances in elucidating adiponectin signalling pathways. FEBS Lett. 2010;584(20):4280–4286. doi:10.1016/j.febslet.2010.09.035. [PubMed: 20875820]
39. Singh B, Reddy PG, Goberdhan A, et al. p53 regulates cell survival by inhibiting PIK3CA in squamous cell carcinomas. Genes Dev. 2002;16(8):984–993. doi:10.1101/gad.973602. [PubMed: 11959846]
40. Gustafson AM, Soldi R, Anderlind C, et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. Sci Transl Med. 2010;2(26):26ra25. doi:10.1126/scitranslmed.3000251.

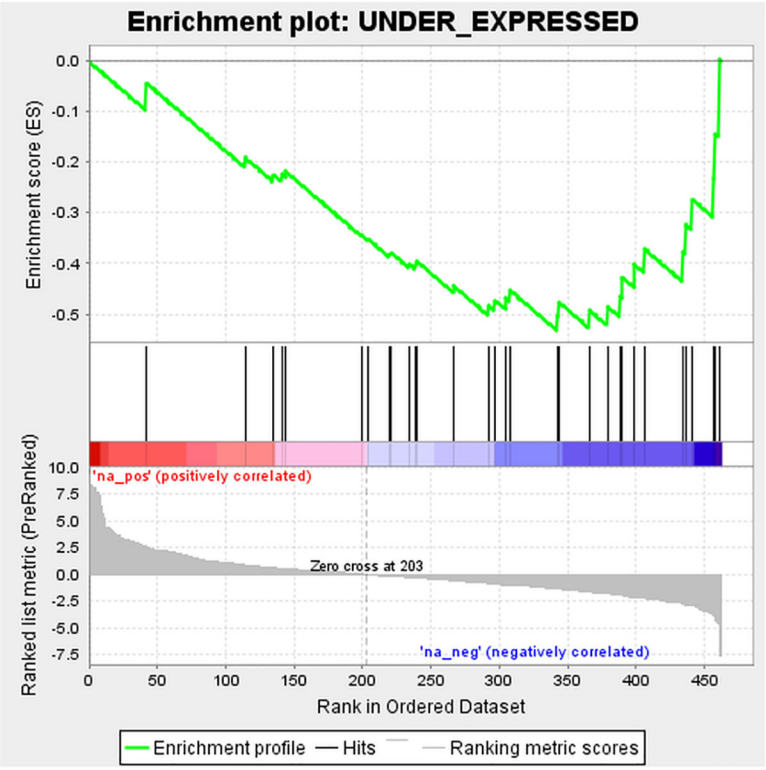

Figure 1. Enrichment of known smoking related microRNAs by GSEA.

A set of 23 previously described microRNAs that are expressed at lower levels in bronchial airway samples from current smokers are significantly enriched among the microRNAs most repressed among current smokers in the current dataset ($q<0.001$). The red to blue bar shows all 463 miRNA ranked from most induced in smokers to most repressed (as shown in the distribution of t-statistics at the bottom); while the vertical black lines show the position, within this ranked list, of the 23 microRNAs previously found by microarray to have decreased expression in the bronchial airway of smokers. The green line is the running enrichment score which has a significantly negative minimum, indicating that the previously reported microRNAs are among the microRNA most repressed among current smokers in the current small RNA sequencing dataset.

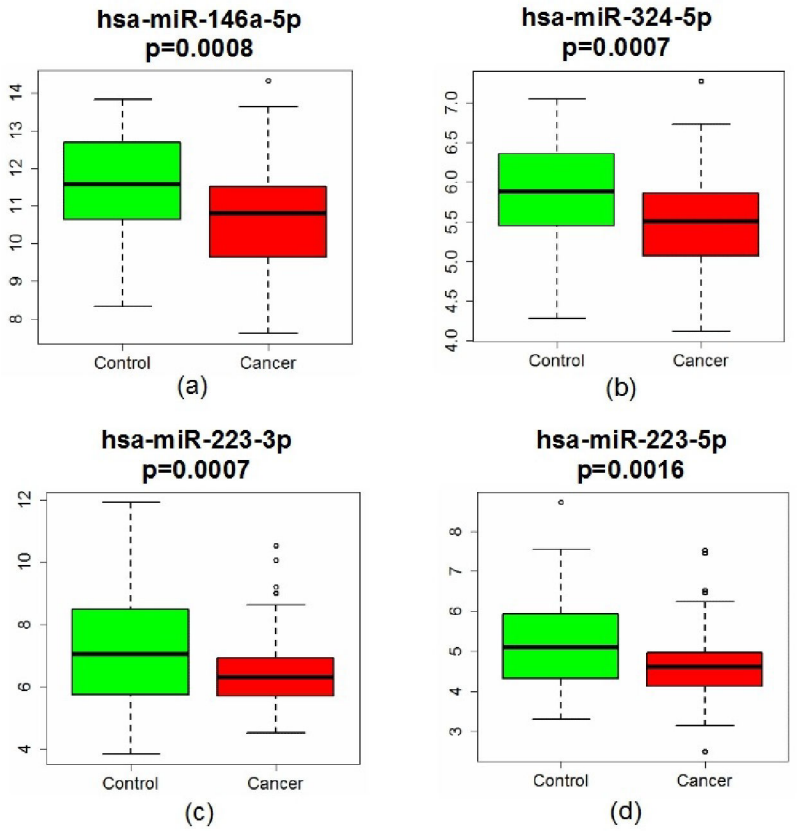

Figure 2. microRNAs significantly differentially expressed in bronchial epithelium between patients with and without lung cancer.

(a) Expression of hsa-miR-146a-5p (p=0.0008, q=0.125) **(b)** Expression of hsa-miR-324–5p (p=0.0007, q=0.125) **(c)** Expression of hsa-miR-223–3p (p=0.0007, q=0.125) **(d)** Expression of hsa-miR-223–5p (p=0.0016, q=0.184).

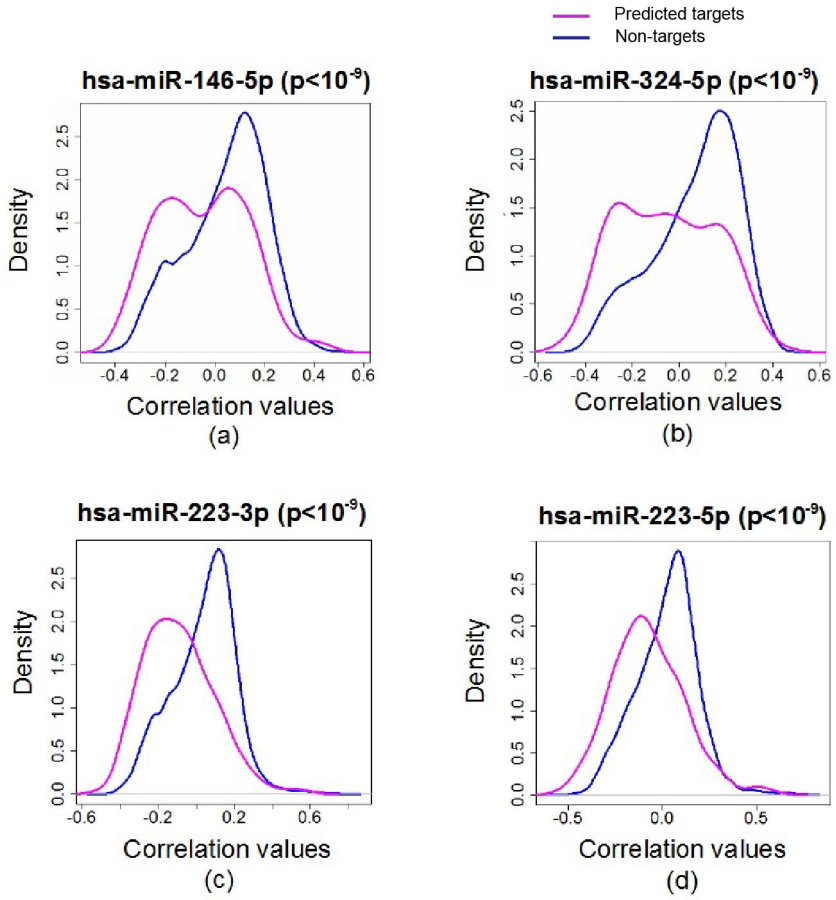

Figure 3. MicroRNAs with cancer-associated expression are negatively correlated with their predicted targets.

The distribution of microRNA-mRNA correlations for each microRNA and its predicted targets is shown with a solid line. The null distribution of microRNA-mRNA correlations for each microRNA and all non-targets is shown with a dashed line. The difference between the two distributions was tested using the Kolmogorov-Smirnov test.

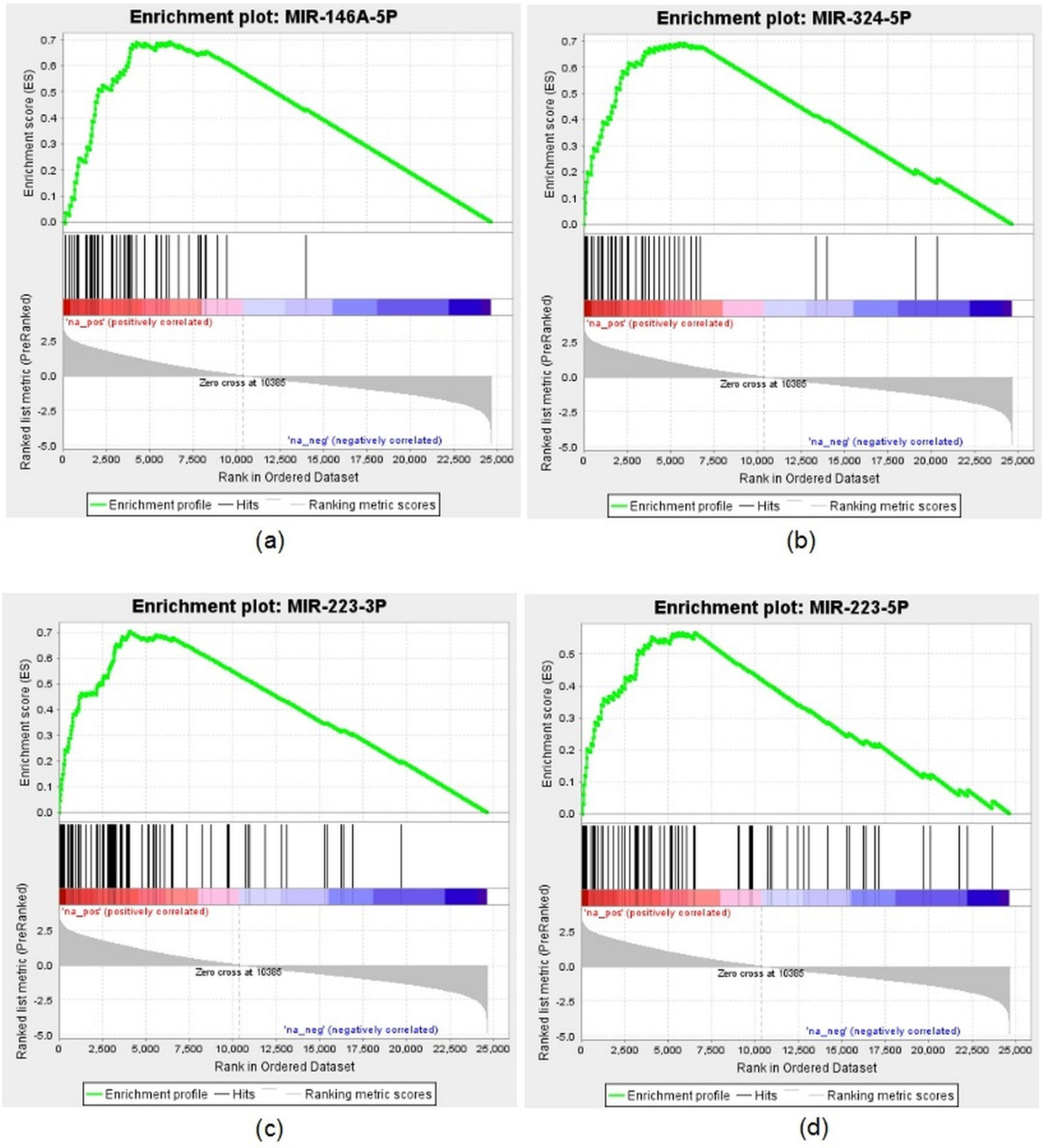

Figure 4. The negatively correlated and predicted gene targets of the four differentially expressed microRNAs are enriched among genes that are expressed more highly in the bronchial airway of patients with cancer:

the distribution of gene sets consisting of negatively correlated and predicted targets of **(a)** miR-146a-5p (50 genes); **(b)** miR-324–5p (43 genes) **(c)** miR-223–3p (89 genes) **(d)** miR-223–5p (72 genes) were examined in a list of genes ranked in the discovery set (n=138) by the association of their expression levels with cancer status in bronchial airway samples using GSEA. All of these gene sets are significantly enriched among the genes most induced in the bronchial airway of patients with lung cancer (GSEA q<0.001 for each). The red to blue bar shows all genes ranked from most induced in the bronchial airway of patients with

cancer to most repressed (as shown in the distribution of t-statistics at the bottom); while the set of vertical black lines in each panel shows the position of the predicted gene targets of each of the microRNAs whose expression is significantly negatively correlated with that microRNA. The green line is the running enrichment score which has a significantly positive maximum in each panel, indicating that these genes are enriched among the genes most induced in patients with cancer.

Table 1.

Patient demographics.

			Discovery set n=138	Test set n=203
Cancer Status (n) *	Lung Cancer		88	103
	Benign Disease		50	100
Gender (n)	Females		62	84
	Males		76	119
Age (SD; n)			59 (11; 138)	59 (10; 203)
Smoking Status (n)	Current		46	88
	Former		92	115
Cumulative Smoke Exposure - pack-yr. (SD; n)			36 (24; 137)	37 (29; 199)
Race (n)	White		109	149
	Black		24	46
	Unknown		5	8
Lesion Size (n)	<3cm		52	71
	>=3cm		58	91
	Infiltrate		15	31
	Unknown		13	10
Histology (n)	NSCLC		72	79
	NSCLC Stage	I	11	16
		II	3	5
		III	15	19
		IV	29	26
		Not specified	14	13
	NSCLC Subtype	Adenocarcinoma	31	34
		Squamous	27	25
		Large-cell	2	4
		Not specified	12	16
	SCLC		16	21
	SCLC Stage	Limited	4	8
		Extensive	8	12
		Not specified	4	1
	Uncertain Histology		0	3
Diagnosis of Benign Disease (n)	Resolution or Stability		11	26
	Alternative Diagnosis		39	74
	Type of Alternative Diagnosis	Sarcoidosis	9	17
		Inflammation	3	2
		Fibrosis	1	1
		Infection	8	14
		Other	18	40

n indicates number of patients with available clinical data; SD indicates standard deviation

* p-value < 0.05.