2017

Multi-omics data integration for the detection and characterization of smoking related lung diseases

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

MULTI-OMICS DATA INTEGRATION FOR THE DETECTION AND

CHARACTERIZATION OF SMOKING RELATED LUNG DISEASES

by

ANA-BRÂNDUŞA PAVEL

B.Eng., Politehnica University of Bucharest, 2009
B.S., University of Bucharest, 2011
M.Eng., Politehnica University of Bucharest, 2011

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2017

Approved by

First Reader   _____
                 Avrum Spira, M.D., M.Sc.
                 Professor of Medicine
                 Professor of Pathology & Laboratory Medicine

Second Reader   _____
                 Marc Lenburg, Ph.D.
                 Professor of Medicine
                 Professor of Pathology & Laboratory Medicine

DEDICATION

For my parents,

Who inspired me to become a scientist,

Who will always be in my heart and everything I do.

ACKNOWLEDGMENTS

MULTI-OMICS DATA INTEGRATION FOR THE DETECTION AND

CHARACTERIZATION OF SMOKING RELATED LUNG DISEASES

ANA-BRÂNDUŞA PAVEL

Boston University Graduate School of Arts and Sciences and

College of Engineering, 2017

Major Professor: Avrum Spira, Professor of Medicine, Professor of Pathology & Laboratory Medicine

ABSTRACT

Lung cancer is the leading cause of death from cancer in the world. First, we hypothesized that microRNA expression is altered in the bronchial epithelium of patients with lung cancer and that incorporating microRNA expression into an existing mRNA biomarker may improve its performance.

Using bronchial brushings collected from current and former smokers, we profiled microRNA expression via small RNA sequencing for 347 patients with available mRNA data. We found that four microRNAs were under-expressed in cancer patients compared to controls ($p<0.002$, $FDR<0.2$). We explored the role of these microRNAs and their gene targets in cancer. In addition, we found that adding a microRNA feature to an existing 23-gene biomarker significantly improves its performance (AUC) in a test set ($p<0.05$).

Next, we generalized the biomarker discovery process, and developed a visualization tool for biomarker selection. We built upon an existing biomarker discovery pipeline and created a web-based interface to visualize the performance of multiple predictors. The "visualization" component is the key to

sorting through a thousand potential biomarkers, and developing clinically useful molecular predictors.

Finally, we explored the molecular events leading to the development of COPD and ILD, two heterogeneous diseases with high mortality. We hypothesized that integrative genetic and expression networks can help identify drivers and elucidate mechanisms of genetic susceptibility.

We utilized 262 lung tissue specimens profiled with microRNA sequencing, microarray gene expression and SNP chip genotyping. Next, we built condition specific integrative networks using a causality inference test for predicting SNP-microRNA-mRNA associations, where the microRNA is a predicted mediator of the SNP's effect on gene expression. We identified the microRNAs predicted to affect the most genes within each network. Members of miR-34/449 family, known to promote airway differentiation by repressing the Notch pathway, were among the top ranked microRNAs in COPD and ILD networks, but not in the non-disease network. In addition, the miR-34/449 gene module was enriched among genes that increase in expression over time when airway basal cells are differentiated at an air-liquid interface and among genes that increase in expression with the airway wall thickening in patients with emphysema.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ALI .... Air-Liquid Interface

ANOVA ....Analysis of Variance

AUC .... Area Under the Curve

CLUAP1 .... Clusterin associated protein 1

COPD .... Chronic Obstructive Pulmonary Disease

CSV .... Comma Separated Values

DC ....Differential Connectivity

DCC..Deleted in Colorectal Carcinoma

DLCO .... Diffusing Capacity of the Lungs for Carbon Monoxide

eQTL ....Expression Quantitative Trait Loci

FC...Fold Chance

FDR .... False Discovery Rate

FEV1 .... Forced Expiratory Volume in 1 Second

FVC .... Forced Vital Capacity

GSEA ....Gene Set Enrichment Analysis

GSVA ....Gene Set Variation Analysis

GUI..Graphical User Interface

HS .... Heparan Sulfate

HS6ST3.. Heparan Sulfate Sulfotransferase 3

HSD.. Honest Significance Difference

ILD .... Interstitial Lung Disease

IPF .... Idiopathic Pulmonary Fibrosis

KNN .... K Nearest Neighbor

LCC .... Large Cell Carcinoma

LDA .... Linear discriminant analysis

LGRC ....Lung Genomics Research Consortium

MAD .... Median Absolute Deviation

MB .... Mega Bases

MDC .... Module Differential Connectivity

MeCP1 .... Methyl-CpG Binding Protein 1

microRNA .... Micro Ribonucleic Acid

miRNA .... Micro Ribonucleic Acid

mRNA .... Messenger Ribonucleic Acid

NSCLC .... Non-Small Cell Lung Cancer

NuRP ....Nucleosome Remodeling Deacetylase

pAUC .... Partial Area Under the Curve

PCR .... Polymerase chain reaction

PI3K.. … Phosphoinositide 3-kinase

PIK3CA............................ …Phosphatidylinositol 3-Kinase Catalytic Subunit Alpha

QQ-plot.. Quantile-Quantile Plot

RABBIT .... R Application for Building Biomarkers in Transcriptomic data

RNA.. Ribonucleic Acid

ROC .... Receiver operating characteristic

SAM .... Significance Analysis of Microarrays

SCC .... Squamous Cell Carcinoma

SCLC .... Small Cell Lung Cancer

SGE .... Sun Grid Engine

SLB .... Surgical Lung Biopsy

SVM .... Support Vector Machines

TTNB .... Trans-Thoracic Needle Biopsy

UTR .... Untranslated Region

CHAPTER ONE

Introduction

1.1 Lung cancer

Lung cancer remains the leading cause of cancer death in the world due, in large part, to our inability to detect the disease at its earliest and curable stage. The high mortality rate (80–85% within 5 years) (Siegel, Naishadham, and Jemal 2013) results, in part, from a lack of effective diagnostic options to detect this disease at an early stage. Symptoms of early stage lung cancer are mild and non-specific, such as a cough, shortness in breathing and tiredness, which can be associated with other benign conditions. Therefore, most patients are diagnosed at late stages associated with poor prognosis (Novaes et al. 2008). About 224,000 new diagnoses and 160,000 deaths were recorded in 2014, 90% of which are due to smoking ("Cancer of the Lung and Bronchus - SEER Stat Fact Sheets" 2016). Lung cancer is classified in two main histological subtypes, such as small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). SCLC develops in the upper airways and it is the most aggressive type of lung cancer that metastasizes quickly to other parts of the body ("Lung Cancer - Small Cell: MedlinePlus Medical Encyclopedia" 2016). However, SCLC represents only 15% of all lung cancer cases. Most lung cancers are NSCLC, and they are further classified into adenocarcinoma, squamous cell carcinoma and large cell carcinoma (Ginsberg, Grewal, and Heelan 2007).

Like most of other cancers, lung cancer is a heterogeneous disease with complex molecular profiles (Collisson et al. 2014; Hammerman et al. 2012). Recently, targeted therapies have worked successfully in patients with activated somatic oncogenes. For example, patients with EGFR mutations are showing response to EGFR targeted compounds, such as Erlotinib and Gefitinib (Greenhalgh et al. 2015; Wang, Schmid-Bindert, and Zhou 2012; Kim et al. 2011). Also, BRAF and ERBB2 genes are currently being investigated as potential therapeutic targets (Collisson et al. 2014; Stephens et al. 2004). Mutational profiles of lung cancer may also depend on the exposure to different carcinogenic compounds. For example, EGFR mutations are more common in never-smokers (Govindan et al. 2012). The complex molecular mechanisms of lung cancer are still poorly understood and there is a tremendous need to develop better diagnostic and therapeutic strategies.

1.2 Using the airway molecular field of injury to predict lung cancer

Chronic inflammation has been previously associated with tumorigenesis in different tissue types, such as lung (Zhai et al. 2008; Fujimoto et al. 2012), colon (Terzić et al. 2010) and skin (Maru et al. 2014). The airway is constantly affected by the exposure to different carcinogens and toxins, leading to chronic inflammation and ultimately to lung disease. Our laboratory has previously shown that the alterations that may occur in the distal part of the lung tissue are also reflected in the normal airway epithelial cells. The ability to identify gene

expression changes associated with smoking and cancer status in the normal appearing airway supports the idea of an *airway molecular field of injury* spanning the respiratory tract (Spira et al. 2007; Brody 2012; Beane et al. 2011; Steiling, Lenburg, and Spira 2009). Recently, a gene expression biomarker for lung cancer detection has been developed, a test that is now used clinically (Whitney et al. 2015; Silvestri et al. 2015).

In this work, we extend the *field of injury* concept to microRNAs. We characterize the microRNA expression changes associated with the presence of lung cancer in bronchial epithelium from the mainstem bronchus and show that these alterations can be used to improve lung cancer detection.

1.3 Biomarker discovery

High throughput technologies have been used to profile genes in multiple different dimensions, such as gene and protein expression, genetic variation, copy number and epigenetics. An important use of gene expression data is the classification of cancer patients with respect to genes that are either up or down regulated in a specific tissue. Gene expression classifiers have been developed for lung cancer detection (Silvestri et al. 2015; Whitney et al. 2015), breast cancer tumors (van 't Veer et al. 2002; Popovici et al. 2010), or prognosis of colo-rectal cancer (Bertucci et al. 2004). In 2006, Micro Array Quality Control (MAQC) project (MAQC Consortium et al. 2006), a community-wide effort involving 137 participants from 51 organizations, established the best practices of developing

molecular classifiers. Multiple statistical and machine learning algorithms have been tested and compared. Using bootstrapping, they have shown that a genomic predictor's accuracy is determined largely by an interplay between sample size and data heterogeneity, and that multiple feature selection and classification algorithms may produce statistically equally good predictors (Popovici et al. 2010; MAQC Consortium et al. 2006). Based on the methods described by MAQC project, our lab has developed a new tool for biomarker discovery. This thesis presents a methodology of visually selecting the best combination of methods that can leverage a clinically useful biomarker in a given dataset.

1.4 Chronic obstructive pulmonary disease and interstitial lung disease

Chronic obstructive pulmonary disease (COPD) is a progressive lung disease and the fourth leading cause of death worldwide (Osei et al. 2015), with an incidence of 2.8 cases per 1,000 population per year (Raherison and Girodet 2009). COPD is a very heterogeneous disease, with the two most common types of COPD being chronic bronchitis and emphysema. COPD consists of narrowing of the small airways and breakdown of lung tissue and it is mainly caused by tobacco smoking. Although biological processes, such as chronic inflammation, apoptosis, and oxidative stress, have been found to play a role in COPD pathogenesis, knowledge remains limited about the molecular mechanisms of this disease (Steiling et al. 2013).

4

Interstitial lung disease (ILD) is another heterogeneous group of chronic respiratory disorders, with the most common ILD being idiopathic pulmonary fibrosis (IPF). ILD has a lower incidence (6.8-8.8 per 100,000 population per year (Nalysnyk et al. 2012)) than COPD, but it is a disease with high mortality characterized by an interstitial fibrotic process (Gribbin et al. 2006; Raghu et al. 2006; Raghu et al. 2011). ILD is characterized by a progressive scarring of lung tissue, that may cause lung stiffness (Nathan et al. 2015). The most common symptom of ILD is shortness of breath. These diseases may be caused by an infection with bacteria (Mycoplasma pneumonia), viruses or fungi, or the cause may be unknown, as it is the case of IPF. There are no effective therapies for IPF (Bjoraker et al. 1998; Carrington et al. 1978; Stack, Choo-Kang, and Heard 1972), therefore understanding the molecular drivers underlying this condition may improve therapeutic strategies and patients outcome.

This thesis addresses three chronic lung diseases, such as ILD, COPD and lung cancer, by leveraging expression profiles and complex molecular interactions.

1.5 Integrative genetic and genomic networks to identify drivers of disease

Integrative network approaches have been used extensively to study complex diseases, such as cancer, diabetes, neurological and respiratory disorders, and other pathologies with underlying genetic causes. Models of regulatory networks have been developed to identify disease specific drivers and

recover the broken molecular pathways. For example, an integrative network approach has been used to identify genetic nodes important to late-onset Alzheimer's disease, highlighting an immune- and microglia-specific module (B. Zhang et al. 2013). Furthermore, regulators of genetic risk of breast cancer have been discovered by integrative network analysis (Castro et al. 2016) and the regulatory landscape of cancer hallmarks has been previously explored (Emmert-Streib et al. 2014). In addition, integrative networks have been used to improve tumor stratification (Hofree et al. 2013) or identify hyper-mutated pathways (Vandin, Upfal, and Raphael 2011; Leiserson et al. 2015). Inferring causality from molecular data is a difficult problem, particularly because correlation does not imply causality. However, incorporating multiple sources of information may lead to a better understanding of the biological mechanisms.

MicroRNAs are a class of small, noncoding RNAs that repress gene expression and protein translation. MicroRNAs (miRNAs) play important roles in complex cellular pathways by targeting multiple messenger RNAs (mRNAs) of protein coding genes (J. Zhang et al. 2014; Sass et al. 2011; Zafari et al. 2015). Inferring condition-specific microRNA activity, may reveal new drivers of disease. We aim to characterize the microRNA-mRNA disease-specific regulatory networks and identify the underlying genetic factors that lead to this dysregulation.

1.6 microRNA Sequencing

This thesis provides new data generated via Next Generation Sequencing technology. Particularly, by profiling microRNA expression, new disease associated microRNAs are revealed. microRNAs are small RNA molecules (~22 nucleotides), that are highly conserved across species. In addition, these molecules are less degraded than mRNAs due to their shorter length, making them a good source of biomarkers (Etheridge et al. 2011).

Mature microRNAs are single stranded. They originate from a double-stranded stem-loop structure with two arms, 3p and 5p (Figure 1). The endonuclease Dicer cleaves the precursor, generating two mature microRNAs that may target different mRNAs and provide different biological functions.
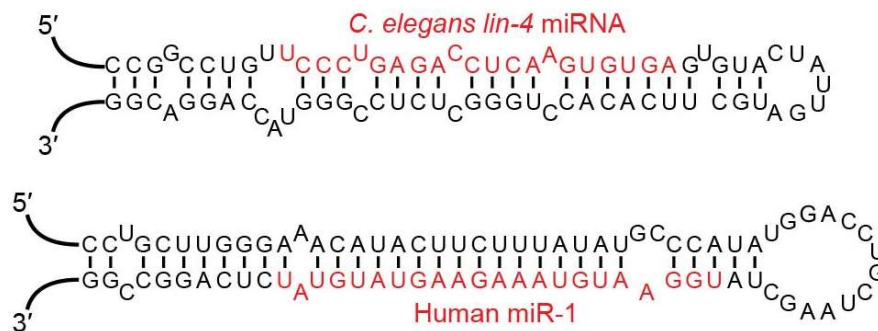
Figure 1. The microRNA stem-loop structure. This figure was imported from https://en.wikipedia.org/wiki/MicroRNA.

microRNAs present a seed sequence that binds specifically to the 3' UTR region of their mRNA targets, inhibiting protein translation. The seed sequence is

a conserved heptametrical sequence, located at positions 2-8 from the mature microRNA 5´-end (Figure 2).

Using the sequencing technology, the abundance level of known microRNAs can be estimated. The microRNA reads are aligned to human genome using short reads aligners, such as Bowtie (Langmead et al. 2009), and annotated using existing databases, such as miRBase (Kozomara and Griffiths-Jones 2011). However, the microRNA sequences can present small variations that can also be detected by sequencing (Figure 2).
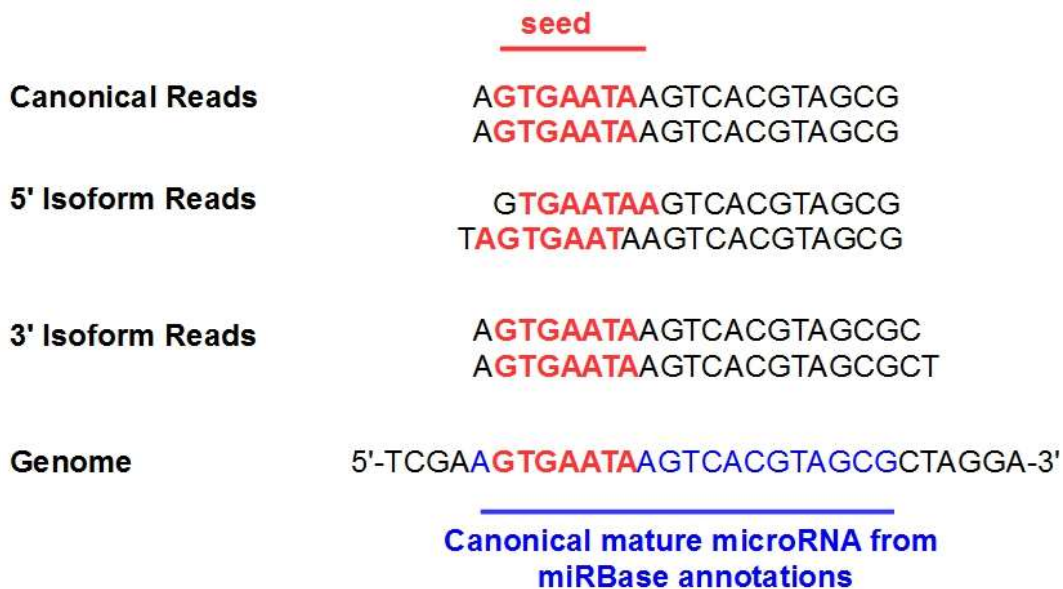
Figure 2. Examples of microRNA sequences: the seed sequence and potential position variants.

A sequencing alignment pipeline has been proposed by (Campbell et al. 2015). This tool aligns and normalized the short reads, providing several

statistics to evaluate the quality of the data. Details about the sequencing

protocol and sequencing data analysis are provided in the following chapters.

However variability in microRNA expression can be caused by different technical

artifacts of the sequencing protocol or platform. Therefore, it is very important to

use proper control in microRNA expression analysis (Baker 2010).

Previous studies have also used small RNA sequencing to reveal disease

associated microRNAs and characterized their role in cancer and other important

cellular processes (Farazi et al. 2011; Sandhu and Garzon 2011; He and Hannon

2004; Bartel 2004). This thesis provides new molecular insights of microRNA

profiles in lung cancer, ILD and COPD.

1.7 Dissertation Aims

1.7.1 Aim 1: Alterations in bronchial airway microRNA expression for lung cancer
detection

First, we aim to characterize the microRNA expression field in the airways

of ever smokers with lung cancer, and identify disease associated microRNAs.

In addition, we explore the clinical utility of bronchial microRNA data for

lung cancer detection in ever smokers. By incorporating microRNA expression

into an existing mRNA biomarker we significantly improve the performance

(AUC) in an independent test set.

The results and the methods of aim 2 are presented in detail in Chapter 2.

1.7.2 Aim 2: Biomarker discovery and visualization

We generalized the biomarker discovery process, and developed a visualization tool for biomarker selection. We built upon an existing biomarker discovery pipeline, *rabbit: an R Application for Building Biomarkers in Transcriptomic data*, and created a web-based interface to visualize the performance of multiple predictors.

The "visualization" component is the key to sorting through a thousand potential biomarkers, and developing clinically useful molecular predictors.

The proposed visualization software was developed using R-Shiny and it is currently available as an open source tool (https://github.com/anabrandusa/rabbitGUI). This project is presented in detail in Chapter 3.

1.7.3 Aim 3: Integrative analysis to identify microRNA drivers of COPD and ILD

The molecular events leading to the development of COPD and ILD are poorly understood. We hypothesized that integrative genetic and expression networks may reveal novel miRNA mediators of genetic factors.

We built condition specific integrative networks using a causality inference test for predicting SNP-miRNA-mRNA interactions, and identified potential microRNA drivers of these lung diseases.

Chapter 4 describes the methodology and the results of this aim.

CHAPTER TWO

Alterations in bronchial airway microRNA expression for lung cancer detection

2.1 Introduction

Based on the National Lung Cancer Screening Trial (NLST) results (Team 2011), we are currently screening high-risk smokers with annual CT scans of the chest, which is leading to an increase in the number of pulmonary lesions being discovered. Once a pulmonary lesion is identified, physicians must decide between CT surveillance vs. airway/lung biopsy, an assessment that is based on pretest risk of disease, comorbidities and patient preference. 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 (Tukey and Wiener 2012)), this procedure is limited by its sensitivity (from 34 to 88%), depending on the location and size of the lesion (Rivera, Mehta, and Wahidi 2013). Even with newer bronchoscopic guidance techniques, the sensitivity for the detection of lung cancer is below 70% for peripheral lesions (Wang Memoli, Nietert, and Silvestri 2012).

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, 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 has been proposed (Whitney et al. 2015; Silvestri et al. 2015).

In this work, we extend the *airway molecular field of injury* concept to microRNA expression. MicroRNAs are a class of small, noncoding RNAs that repress gene expression and protein translation of their target genes by binding to the 3' UTR complementary strands. This regulatory role is key to cellular function and can be leveraged to gain insight into the response to exposures and even pathogenesis of disease. In addition, compared to mRNAs, microRNAs are thought to be more stable molecules, making them more easily measured in degraded tissues (Etheridge et al. 2011).

Previous studies have shown that smoking alters the expression of microRNAs in the bronchial airway epithelium (Perdomo et al. 2013; Schembri et al. 2009). We hypothesize that bronchial microRNA expression changes may also be associated with the presence of lung cancer and that integrating microRNA with gene expression could improve lung cancer detection.

2.2 Results

2.2.1 Patient population

Over 1000 current and former smokers with suspected lung cancer were enrolled in the Airway Epithelial Gene Expression in the Diagnosis of Lung Cancer (AEGIS) trials (Whitney et al. 2015; Silvestri et al. 2015).

microRNA expression was profiled via small RNA sequencing for 347 patients (194 cancer-positive and 153 cancer-negative subjects) from AEGIS-1 and AEGIS-2 trials. Of the 347 samples, 341 passed the sequencing quality control filter. Details about the sequencing protocol and quality control are provided in subsections 2.3.2 and 2.3.3.

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 (Whitney et al. 2015; Silvestri et al. 2015). 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 (Silvestri et al. 2015).

The demographics data for the discovery cohort (138 samples) and the test set (203 samples) is presented in Table 1. Except for cancer status, the other clinical variables are not significantly different between the two datasets. Furthermore, cancer status is significantly associated with age in the discovery set and pack-years in the test set (Table 2).

13

			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	78	140
	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 Alternative Diagnosis	18	40

Table 1. Patient demographics table. n indicates number of patients with

available clinical data; SD indicates standard deviation; * p-value < 0.05 by

Fisher's Exact Test.

		Discovery set n=138			Test set n=203		
		Lung Cancer n=88	Benign n=50	p	Lung Cancer n=103	Benign n=100	p
Gender	Females	25	37	0.84	38	46	0.2
	Males	63	13		65	54	
Age (SD; n)		61 (10; 88)	56 (13; 50)	0.01	60 (9; 103)	58 (12; 100)	0.29
Smoking	Current	32	14	0.35	47	41	0.57
	Former	56	36		56	59	
Cumulative Smoke Exposure - pack-yr. (SD; n)		38 (22; 88)	33 (27; 49)	0.2	40 (28; 102)	32 (30; 97)	0.05
Race	White	69	40	0.98	74	75	0.8
	Black	15	9		26	20	
	Unknown	4	1		3	5	
Lesion Size	<3cm	30	22	$4 \cdot 10^{-4}$	20	51	$8 \cdot 10^{-14}$
	>=3cm	47	11		73	18	
	Infiltrate	4	11		6	25	
	Unknown	7	6		4	6	

Table 2. The association of cancer status with other clinical variables. p-values indicating the association of cancer with gender and smoking status were computing using a Fisher's exact test; p-values indicating the association of cancer with age and cumulative smoke exposure were computed using a Student's t-test; n indicates number of patients with clinical data available; SD indicates standard deviation.

2.2.2 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 (Schembri et al. 2009; Powell et al. 1999; Wistuba et al. 1997; Franklin et al. 1997; Guo et al. 2004; Miyazu 2005; Spira et al. 2004).

We therefore used the ability to detect microRNAs with smoking associated expression as a positive control for the quality of the microRNA expression data.

A set of 28 microRNAs were previously proposed as modulators of gene expression changes in airway epithelium (Schembri et al. 2009), with most of them (n=23) being down-regulated in current smokers compared to never smokers. We found that these down-regulated microRNAs induced by smoking were significantly negatively enriched among current smokers (q<0.001), in both the discovery (n=138) and the test (n=203) sets (Figure 3).
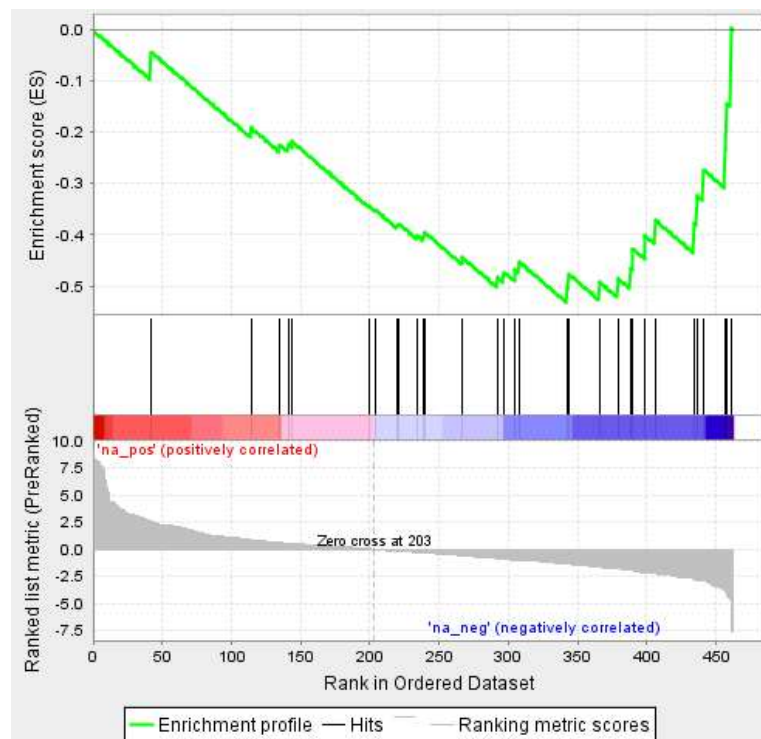
Figure 3. Enrichment of known smoking related microRNAs by GSEA. The set of 23 known smoking-induced down-regulated microRNAs are significantly negatively enriched among current smokers (q<0.001).

In addition, we identified significantly differentially expressed microRNAs between current and former smokers by linear regression. The top 30 differentially expressed microRNAs in the discovery set (q<0.01) are shown in Figure 4. Among these, we found microRNAs whose expression has been previously associated with smoking, such as miR-218, miR-365, miR-30a and miR-99a (Schembri et al. 2009).
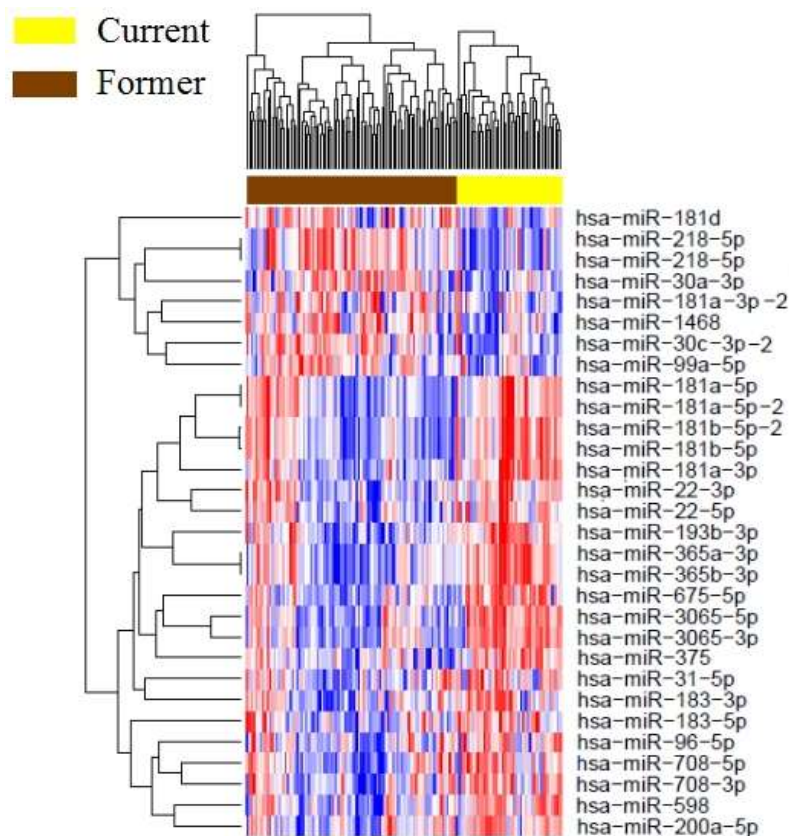
Figure 4. Significantly differentially expressed microRNAs between current and former smokers (q<0.01). Some of these microRNAs have been previously associated with smoking status, such as miR-218, miR-365, miR-30 and miR-99a.

2.2.3 Identifying cancer-associated microRNAs in airway epithelium

Using the discovery set (n=138), we identified four significantly differentially expressed microRNA isoforms between patients with and without cancer by linear regression (p<0.002, q<0.2), miR-146a-5p, miR-324-5p, miR-223-3p, 5p. The expression profiles of these microRNAs are shown in Table 3 and Figure 5. Each of these miRNA has previously been shown to have tumor-suppressor-like activity (Chen et al. 2013; Labbaye and Testa 2012; Li et al. 2013; Nian et al. 2013). Consistent with these previous observations, we find that these microRNAs were down-regulated in the bronchial airway of patients with cancer.

microRNA	miRBase ID	p-value	q-value	t-statistic	direction in cancer
hsa-miR-324-5p	MI0000813_MIMAT0000761	0.0007	0.125	-3.49	DOWN
hsa-miR-223-3p	MI0000300_MIMAT0000280	0.0007	0.125	-3.47	DOWN
hsa-miR-146a-5p	MI0000477_MIMAT0000449	0.0008	0.125	-3.43	DOWN
hsa-miR-223-5p	MI0000300_MIMAT0004570	0.0016	0.184	-3.23	DOWN

Table 3. Cancer-associated bronchial microRNAs (p<0.002, q<0.2).
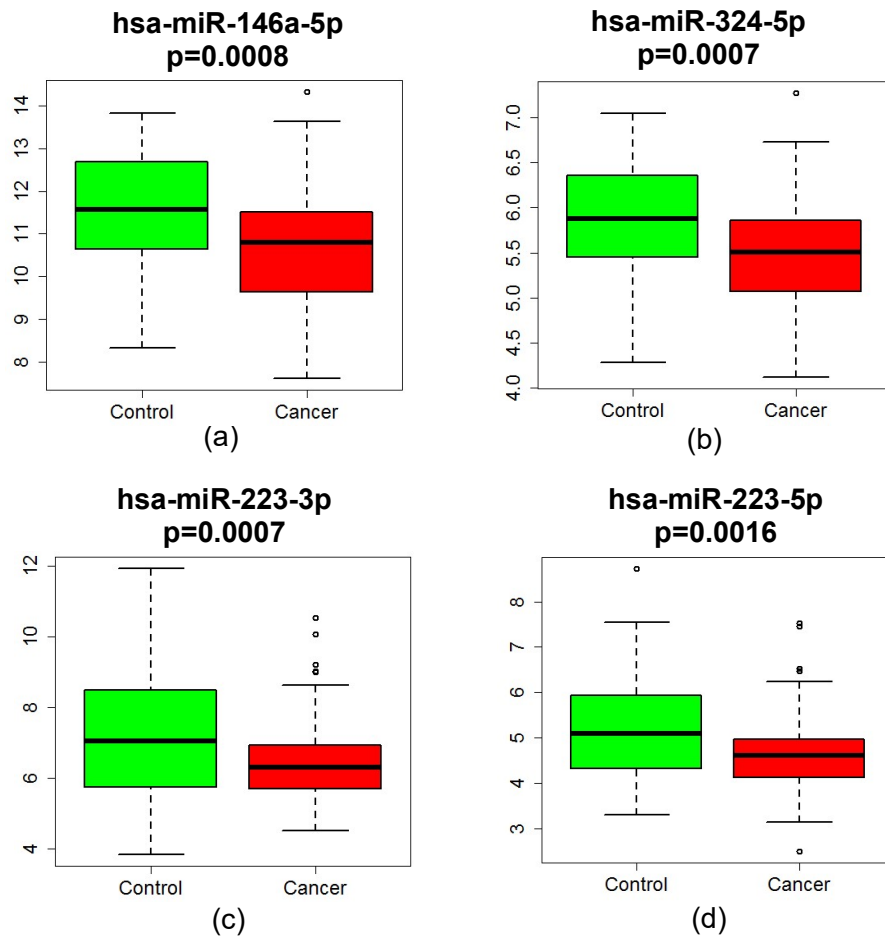
Figure 5. Bronchial microRNAs significantly differentially expressed between cancer-positive and cancer-negative patients. (a) Expression of hsa-miR-146a-5p (p=0.0008, q=0.125) (b) Expression of hsa-miR-324-5p (p=0.0007, q=0.125) (c) Expression of hsa-miR-223-3p (p=0.0007, q=0.125) (d) Expression of hsa-miR-223-5p (p=0.0016, q=0.184).

2.2.4 Identifying microRNA-mRNA relationships

MicroRNAs generally lead to the degradation of the mRNAs to which they bind. Therefore, we expect a microRNA with functional variation in expression to

be negatively correlated with the expression of its 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 null distribution ($p<10^{-9}$) (Figure 6).

To begin to understand the potential biological impact of the cancer-associated expression of the microRNA, we further evaluated the relationships between their gene targets and cancer. From the binding site predicted targets (Targetscan), we identified the genes whose expression is significantly negatively correlated (correlation $q<0.1$) with the cognate microRNA. These negatively correlated targets of each of the four microRNA isoforms were significantly positively enriched with cancer status ($q<0.001$) by Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005) (Figure 7). In addition, the set of genes regulated by these microRNAs (254 in total) is enriched by DAVID (Huang, Sherman, and Lempicki 2009) for cancer-associated pathways, such as signaling pathways regulating pluripotency of stem cells ($p=0.001$), pathways in cancer ($p=0.007$), TGF-beta signaling pathway ($p=0.035$), Ras signaling pathway ($p=0.043$).

Furthermore, we were able to validate in the test set the microRNA-mRNA relationships identified in the discovery set for all four isoforms. The correlation coefficients of each cancer-associated microRNA and its predicted mRNA targets (binding site predicted targets from Targetscan) are significantly more negative

than the null distribution ($p<10^{-7}$) (Figure 8). In addition, the negatively correlated

and predicted targets identified in the discovery set were significantly positively

enriched with cancer status in the test set for each of the four isoforms; GSEA

results are shown in Figure 9.
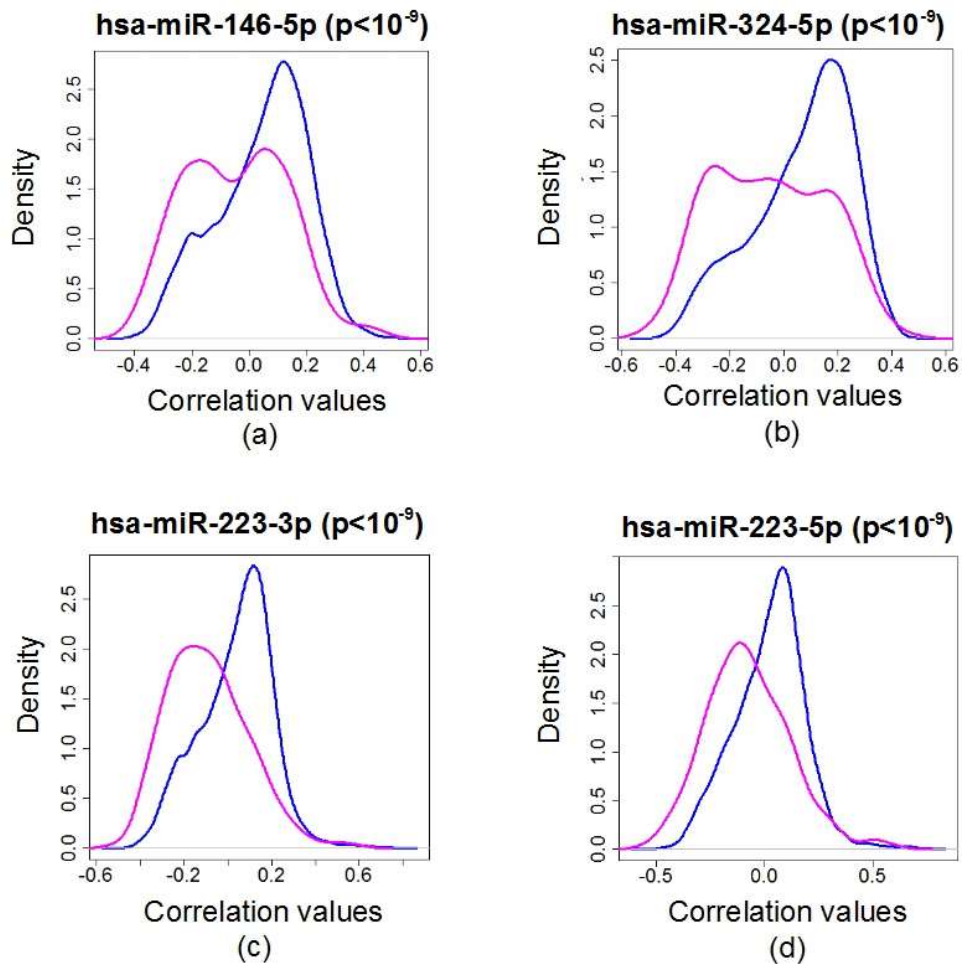
Figure 6. The correlation with the predicted targets in the discovery set is

significantly negative by a Kolmogorov-Smirnov test. The null distribution is

represented in blue; the distribution of microRNA-mRNA correlations for each

microRNA is represented in magenta.
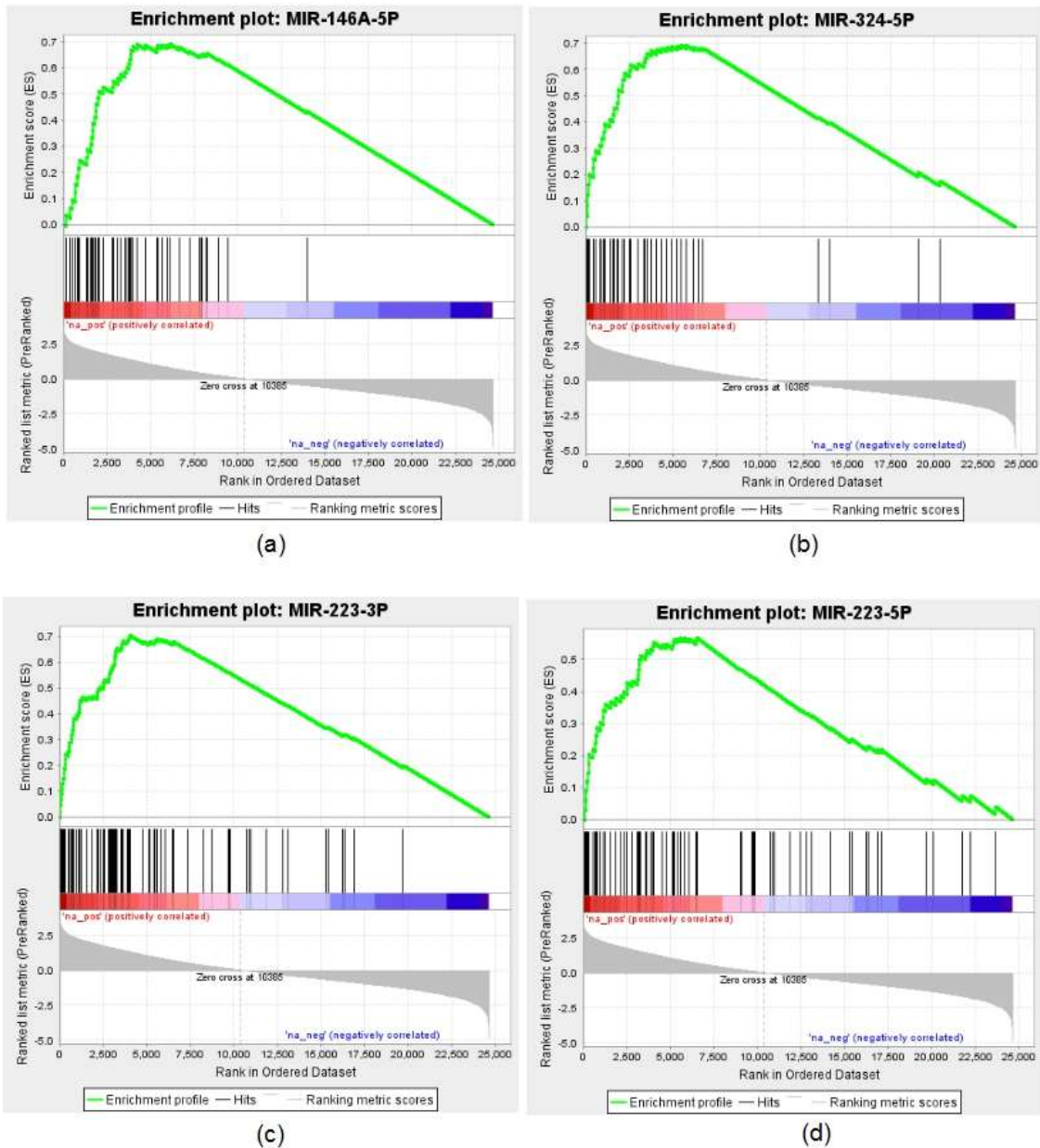
Figure 7. The negatively correlated and predicted gene targets of the four differentially expressed microRNA isoforms are enriched in the discovery set by GSEA. (a) miR-146a-5p (50 genes, GSEA q<0.001); (b) miR-324-5p (43 genes, GSEA q<0.001) (c) miR-223-3p (89 genes, GSEA q<0.001) (d) miR-223-5p (72 genes, GSEA q<0.001).
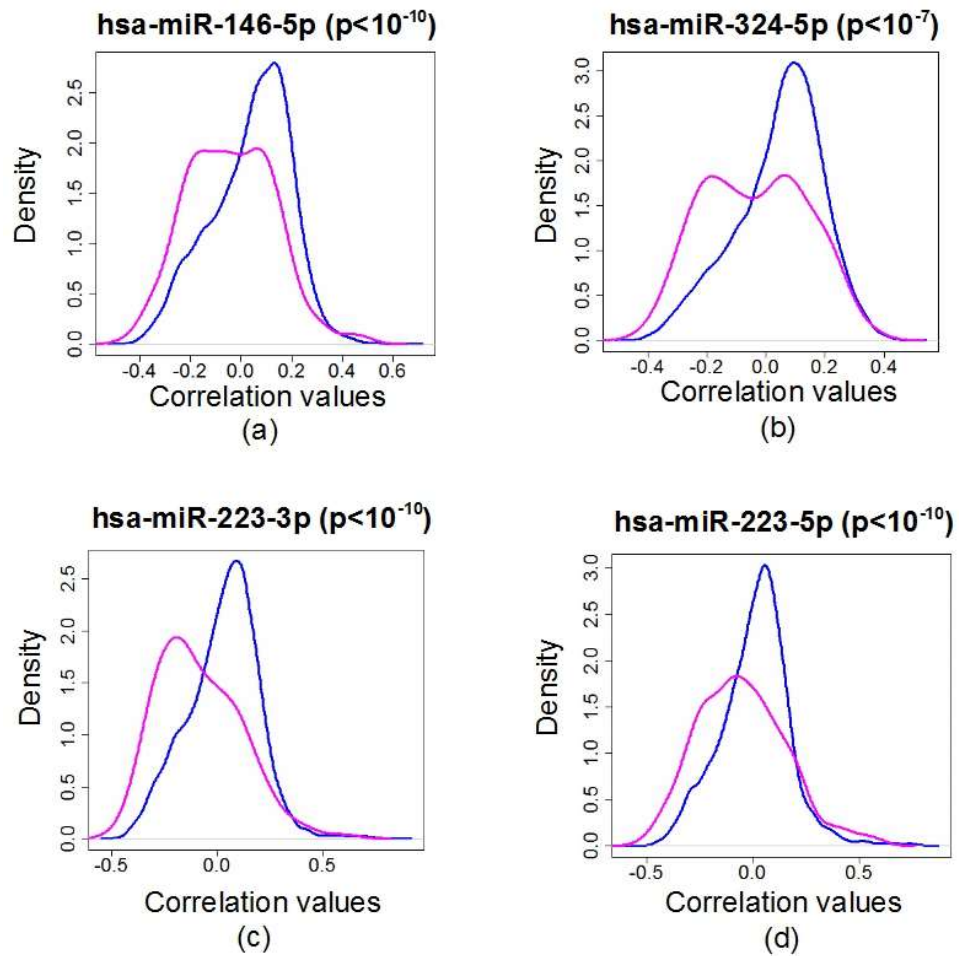
Figure 8. The correlation with the predicted targets in the test set is significantly negative by a Kolmogorov-Smirnov test. The null distribution is represented in blue; the distribution of microRNA-mRNA correlations for each microRNA is represented in magenta.
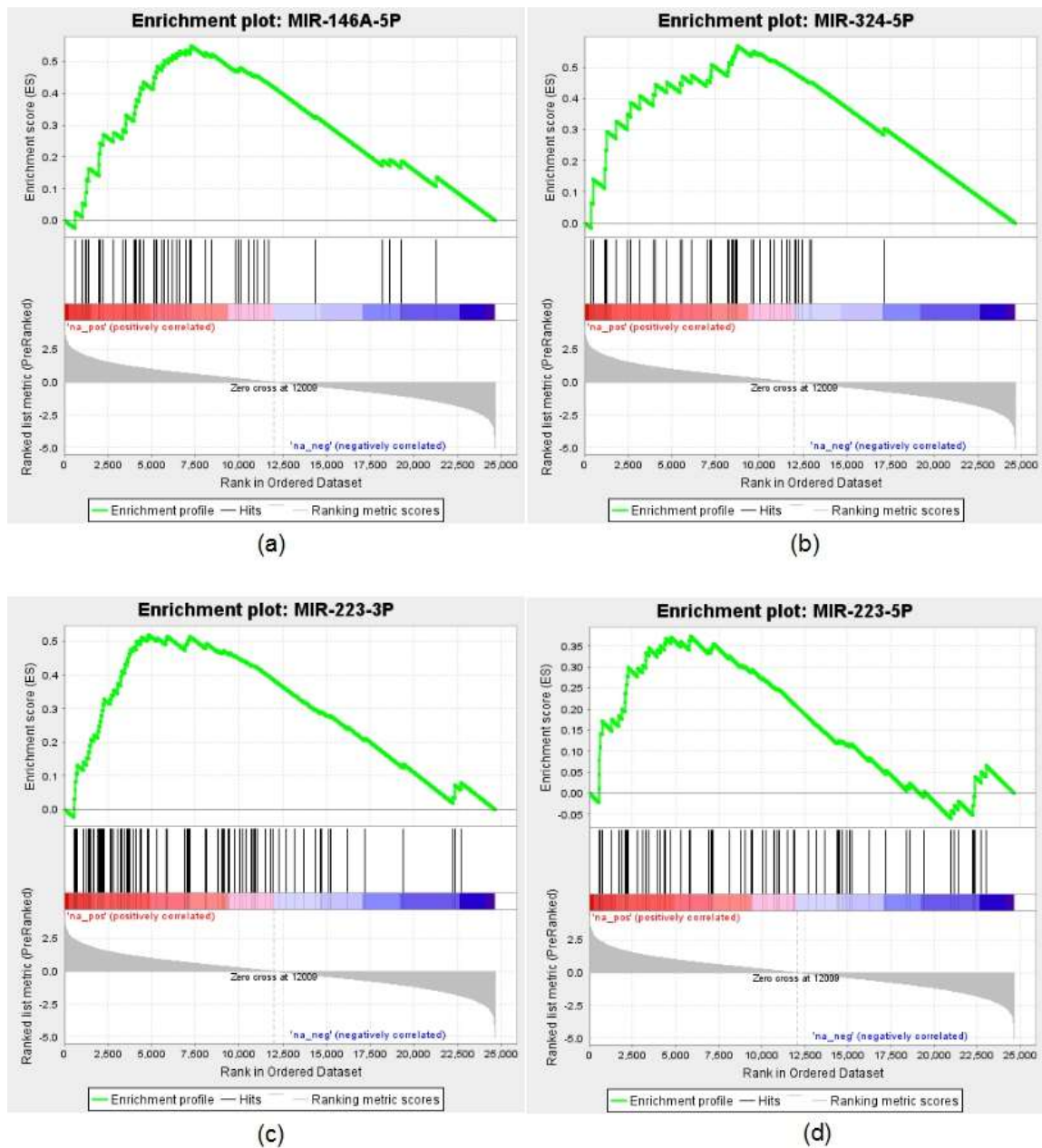
Figure 9. The negatively correlated and predicted gene targets of the four differentially expressed microRNA isoforms in the discovery set are also enriched in the test set by GSEA. (a) miR-146a-5p (50 genes, GSEA q<0.001); (b) miR-324-5p (43 genes, GSEA q<0.001) (c) miR-223-3p (89 genes, GSEA q<0.001) (d) miR-223-5p (72 genes, GSEA q<0.01).

2.2.5 Bronchial miR-146a-5p improves lung cancer diagnosis

We next sought to assess whether bronchial microRNA expression could add to the performance of a mRNA biomarker for lung cancer we previously identified (Whitney et al. 2015). 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 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. The demographic data of the test cohort is provided in Table 1 and Supplementary Table 1. We found that adding miR-146a-5p to the mRNA biomarker improved the AUC from 0.66 to 0.71 (p=0.025), as shown in Figure 10.

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).
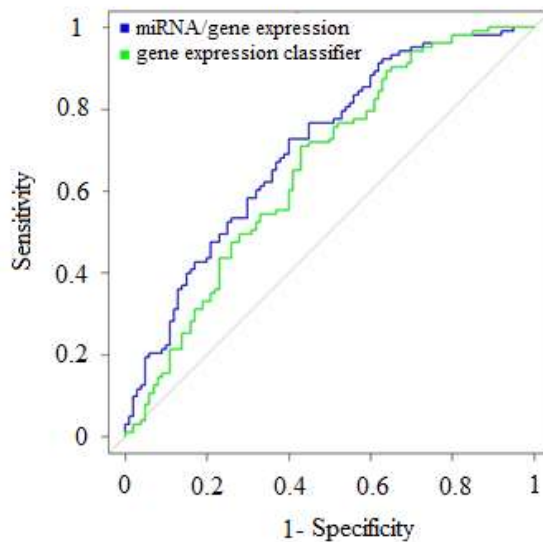
Figure 10. ROC AUC. miR-146a-5p significantly improves prediction of the gene-expression biomarker (p=0.025). The AUC increases from 0.66 (green) to 0.71 (blue).

2.3 Methods

2.3.1 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) (Whitney et al. 2015; Silvestri et al. 2015). Exclusion criteria for patients enrolled in AEGIS trials were age less than 21 years, no history of smoking (defined as having ever 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.

In this study, we profiled microRNA expression via small RNA sequencing for 347 patients. 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 groups 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 (Whitney et al. 2015; Silvestri et al. 2015) and was available for data integration.

2.3.2 High-throughput sequencing of small RNA

Based on previous work on the effect of multiplexing on microRNA expression quantitation (Campbell et al. 2015), 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

27

cleavage by Dicer or other RNA processing enzymes. The adapter used for the first batch has the following sequence: TGGAATTCTCGGGTGCCAAGG, while the one used for the second and the third batches has the following sequence: AGATCGGAAGAGCACACGTCT.

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 two 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.

2.3.3 MicroRNA alignment and quality control

To estimate microRNA expression we used a small RNA sequencing pipeline previously described (Campbell et al. 2015). 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 (Langmead et al. 2009) 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 (Griffiths-Jones 2004; Quinlan and Hall 2010).

Figure 11 illustrates the number of reads and Figure 12 shows the mismatch distribution of aligned reads.
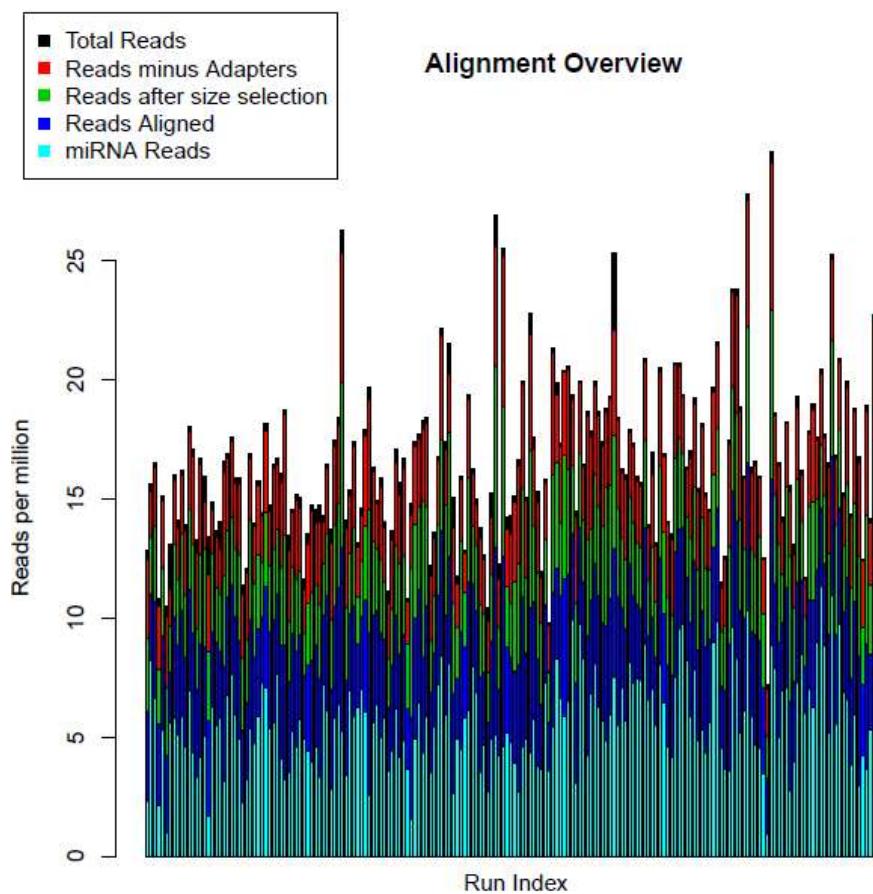
Figure 11. Alignment overview. An overview plot showing the total number of reads, the number of reads after filtering out adapter-only reads, the number of reads after size selection, the number of reads aligned, and the number of reads aligning to microRNA precursors for each sample.
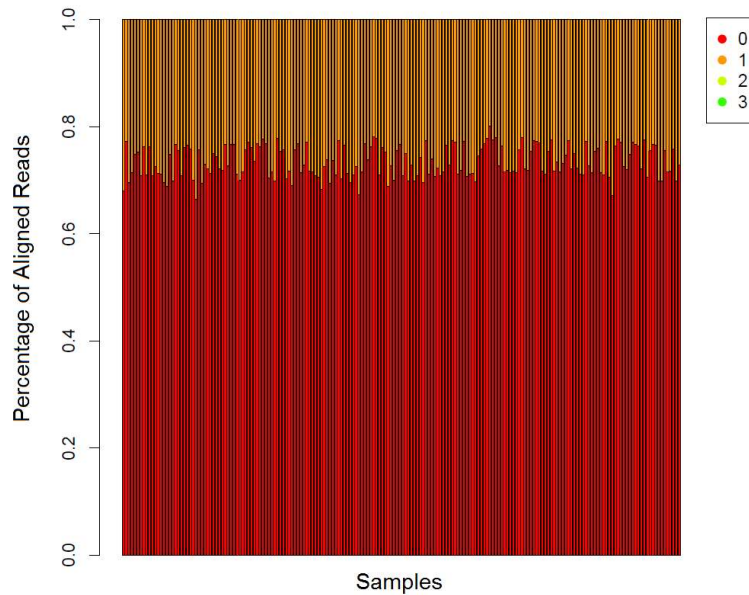
Figure 12. Mismatch distribution. The different number of mismatches are indicated by the different colors (red is 0; orange is 1).

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 (Campbell et al. 2015).

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 (Figure 13). Six such samples were removed, leaving 341 samples

included in the downstream analysis.
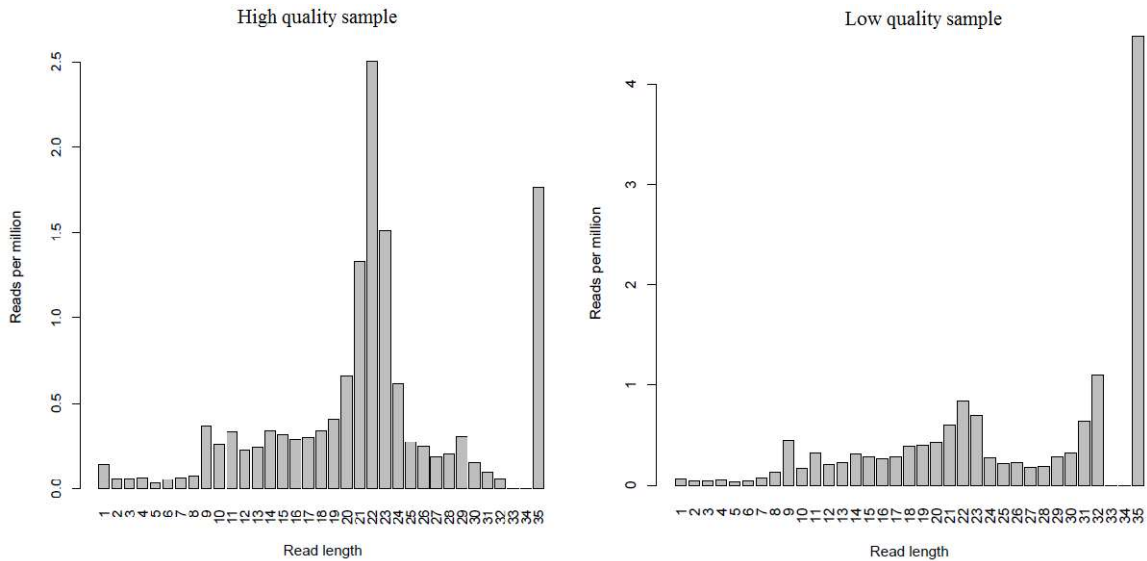
Figure 13. The distribution of lengths of aligned reads. Most aligned reads are

between 20 and 24 nucleotides long and predominantly align to microRNA loci.

Reads that were not trimmed (35 nucleotides) aligned to mostly to predicted

tRNA or snoRNA loci. The distribution of lengths of aligned reads of the high

quality samples shows that most reads are ~22 nucleotides long which is the

average microRNA length. For the low quality samples, the distribution of aligned

reads is not centered on the 22 nucleotides length.

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 (Johnson, Li, and Rabinovic 2007) 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.

2.3.4 Differential expression analysis

To identify smoking-associated microRNAs, while correcting for covariates, we applied an F-test (anova R function) (Chambers 1992) 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 (Benjamini and Hochberg 1995), and were denoted with q-value.

2.3.5 Identifying microRNA-mRNA relationships

We analyzed the correlations between the differentially expressed microRNAs and their targets as predicted in the Targetscan database (Lewis, Burge, and Bartel 2005). 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 are enriched among the genes whose expression is associated with cancer status by Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005). The genes were ranked by the t-statistic of a linear regression, with gene expression as the response variable and cancer status, age, gender, smoking status, and pack-years as the independent variables.

2.3.6 Improving the gene-expression classifier by incorporating the expression of microRNA

First, we calculated the prediction scores of the mRNA classifier (Whitney et al. 2015; Silvestri et al. 2015) using the subset of samples with matched mRNA and microRNA data. Then, for each cancer-associated microRNA, we integrated the mRNA classifier score with the microRNA's expression using logistic regression (*cv.glmnet* function from *glmnet* R package,

). Figure 14 summarizes the method.

The coefficients of the logistic regression were determined in the discovery set and the performance of fully specified models 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 (DeLong, DeLong, and Clarke-Pearson 1988) from the *pROC* R package (Robin et al. 2011).

$$Prediction_{Score} = \frac{1}{1+e^{-\left(\beta_0 + \beta_1\, ClinicoGenomic_{Score} + \beta_2\, miRNA_{Expression}\right)}}$$
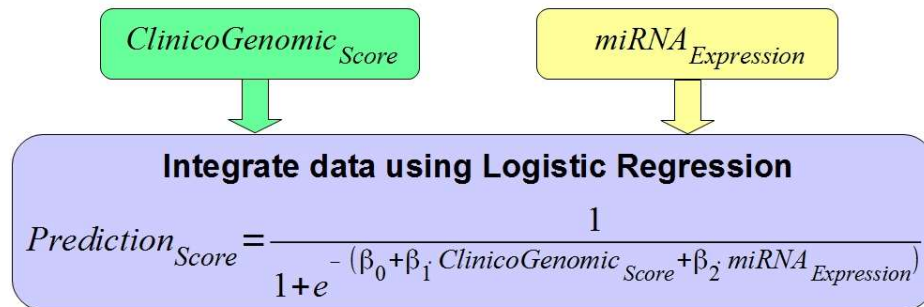
Figure 14. miR-146a-5p expression is integrated with the clinico-genomic score of the existing gene expression classifier by logistic regression. The logistic regression model was implemented using *cv.glmnet()* function from *glmnet* R package. By training the weights of the logistic regression in the discovery set, we obtained the following values: $\beta_0 = 1.85$, $\beta_1 = 4.39$, $\beta_2 = -0.37$.

2.4 Discussion

CT screening of high-risk smokers for lung cancer has led to an increase in the number of lesions detected. When routine clinical diagnostic workup is inconclusive, profiling mRNA in the bronchial airway epithelium has been shown to improve detection (Whitney et al. 2015; Silvestri et al. 2015; Spira et al. 2007). In this study, we expanded on this concept by profiling microRNA expression in the bronchial airway to identify lung cancer associated microRNAs that, in combination with mRNA, have the potential to aid in the detection of disease.

Prior studies have demonstrated that microRNAs have aberrant expression, mostly down-regulated, in tumors compared to normal tissue, and have been associated with tumor suppression, cell differentiation, cell signaling, and apoptosis (Lu et al. 2005). Furthermore, profiling microRNA in the bronchial airway, a less invasive site than tumor tissue, has revealed microRNA expression alterations associated with exposure to tobacco cigarette smoke[10]. In this study, we confirm that microRNA expression changes occur in the airway of current smokers when compared to formers, and importantly, show that the airway field of injury for lung cancer is reflected in microRNA expression differences.

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. Similar to findings in tumor tissue, these microRNAs were all down-regulated in cancer patients. Intriguingly, all these microRNAs have previously been implicated in tumor suppressive pathways. Specifically,

miR-146a has been previously shown to inhibit cell growth, migration and EGFR signaling (Labbaye and Testa 2012; Kumaraswamy et al. 2015; Chen et al. 2013), 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) (Bhaumik et al. 2009). 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 (Nian et al. 2013); and miR-324 has been associated with nasopharyngeal cancer (Li et al. 2013). While microRNA expression differences have already been well documented in tumors, we are showing for the first time that the expression of microRNAs with cancer-related functions is altered in the bronchial airway of lung cancer patients.

To begin to determine if the altered expression of these cancer-associated microRNAs has a functional impact on the airway epithelium, we demonstrated that the expression of mRNAs which are predicted targets of these microRNAs is significantly negatively correlated with the expression of the cancer-associated microRNAs 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

36

APPL1 (adaptor protein, phosphotyrosine interaction, PH domain and leucine zipper, DCC-interacting protein 13-alpha). The protein encoded by APPL1 gene binds to many other proteins, including PIK3CA, RAB5A, DCC, AKT2, adiponectin receptors, and proteins of the NuRD/MeCP1 complex, which are involved in cell proliferation and crosstalk between adiponectin and insulin signaling pathways. Interestingly, we also observed a significantly negative correlation between miR-146a-5p and PIK3CA, suggesting that miR-146a-5p interacts with PI3K/AKT pathway. In addition to the important role of PI3K/AKT pathway in cell death/survival, an increased activity of PI3K has been shown to be an early and potentially reversible event in the airway of smokers with premalignancy and lung cancer (Singh et al. 2002; Gustafson et al. 2010).

The correlation of these differentially expressed bronchial microRNAs with cancer-associated mRNA targets suggest their role as lung cancer-associated regulators of gene expression, and potentially could serve as biomarkers of disease.

We assessed each differentially expressed microRNA's ability to enhance the performance of an mRNA biomarker that had been developed and validated using samples from the same cohort[5]. We integrated miR-146a-5p expression into the mRNA classifier from Whitney et al. (Whitney et al. 2015), and have shown that it significantly improves the performance of the lung cancer biomarker.

In addition, we examined the added value of the other three cancer-associated microRNA isoforms and found that they did not improve the performance. Interestingly, we believe that miR-223-3p and miR-223-5p did not add to the biomarker performance because one of their targets (SNCA) is a member of the mRNA classifier, thus miR-223 expression might be redundant with SNCA expression levels and not capable of adding new information about the likelihood of lung cancer to the biomarker. 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 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 identified microRNA that are known to play a role in cancer-related processes, and importantly, we demonstrate that a multi 'omics data integration approach may improve prediction. Future work includes extending this biomarker development approach to even less invasive sampling sites (e.g. nasal brushings), which has the potential to expand the clinical impact of molecular biomarkers.

CHAPTER THREE

Biomarker discovery and visualization

3.1 Introduction

Based on the methods described by MAQC project (MAQC Consortium et al. 2006), our group has proposed a pipeline for biomarker discovery, *rabbit: an R Application for Building Biomarkers in Transcriptomic data* (J. Perez-Rogers, PhD Thesis, 2016; https://github.com/jperezrogers/rabbit). The software runs several combinations of binary class predictors in cross-validation on a given normalized gene-expression dataset. Figure 15 illustrates the four modules of the biomarker pipeline that correspond to the main steps of the biomarker discovery process:

1. feature filtering (unsupervised);

2. feature ranking (supervised);

3. biomarker size selection;

4. classification.

This tool has been developed using a modular approach based on object oriented programming paradigm. Therefore, the framework can be extended to other algorithms as well. The current version of the pipeline includes 840 combinations of methods, tested in cross-validation. However selecting the best biomarkers from almost a thousand potential predictors remains an open question. Some combinations of models may be biased towards noisy patterns

that are specific to a particular dataset. By simply selecting the predictor with the highest ROC AUC, the user may deal with overfitting and not necessarily with the most robust biomarker. In this work we propose a methodology to sort through almost a thousand potential biomarkers. Our approach is based on a graphical interface that visually guides the user through the entire biomarker selection process.
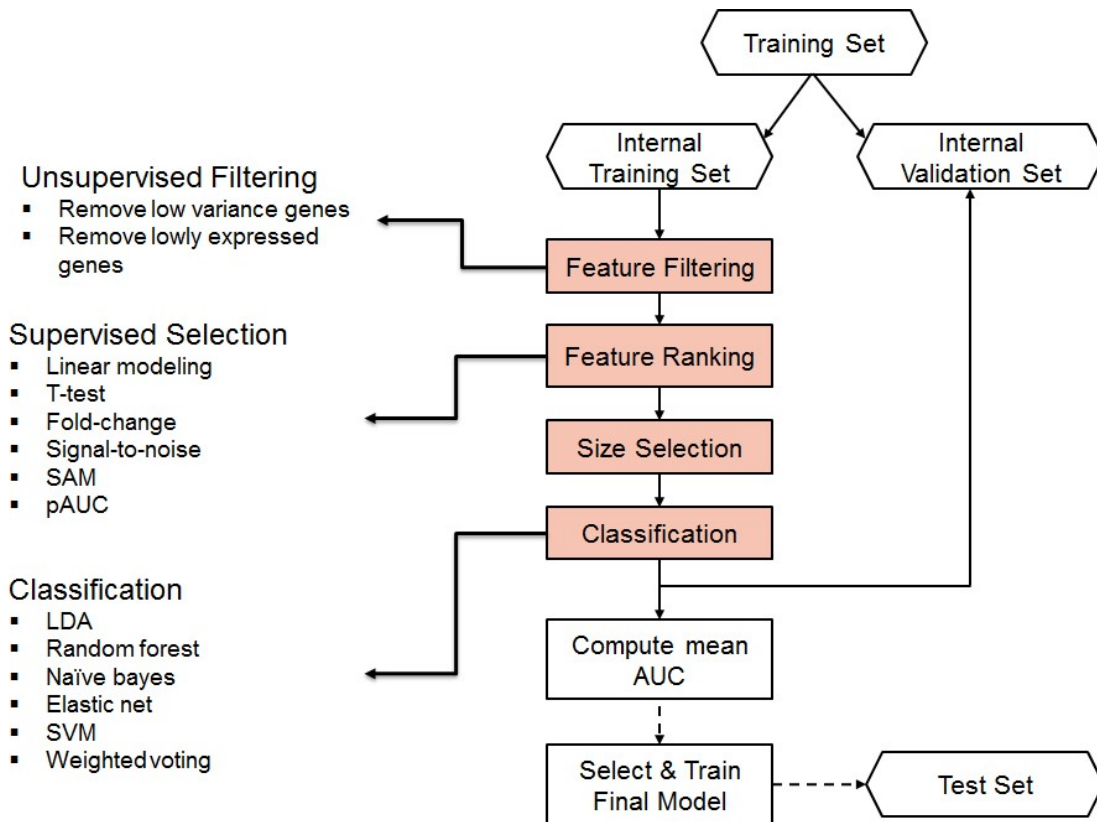

Figure 15. The biomarker discovery pipeline runs all available combinations of feature filters, feature ranking, biomarker sizes and classifiers in cross-validation.

40

We propose *rabbitGUI* a new web-based graphical interface that helps to evaluate the performance of all statistical and machine learning methods tested in cross-validation by *rabbit*. This tool has been implemented as an R-Shiny application and the code is available as open-source (https://github.com/anabrandusa/rabbitGUI). The functionalities of *rabbitGUI* will be detailed in the following sections.

3.2 Results

3.2.1 rabbitGUI: a web-based interface for biomarker discovery

In this section, a detailed description of rabbitGUI functionalities is provided. As a case study, a publicly available dataset from MAQC project (ER+/- breast cancer patients (Popovici et al. 2010)) is used.

3.2.1.1 Model selection

Model selection tab is designed to guide the user through the selection of the best combination of models in a step-wise manner, following the four main steps in the biomarker discovery process: feature filtering, feature ranking, biomarker size selection and classification.

The user can navigate through the four steps using a *radio button* menu. In each step, each method is evaluated across all the possible predictors that incorporate that method. If a method performs generally well across all the methods in all the other steps, it is considered robust and less likely to over-fit.

To provide an evaluation of the prediction results and compare them across the different models, the following statistical procedure is applied.

In each step, we compare the performance (mean ROC AUC) of each model across all the possible predictors by ANOVA. If the p-value is significant, then we apply a Tukey HSD test to identify the top model groups. The models that perform significantly lower than any other models are excluded, and the top remaining models are colored in red. The adjusted p-values from the Tukey HSD test are provided in a table displayed below the boxplots.

Figures 16, 17, 18 and 19 illustrate each of the four steps along with a summary of the results from the Tukey HSD test.

Based on these four steps, multiple combination of models may be selected. Previous studies have shown that multiple feature selection and classification algorithms may produce statistically equally good predictors (Popovici et al. 2010; MAQC Consortium et al. 2006). Therefore, the GUI provides a table with all selected biomarkers (Figure 20).

In addition, the GUI displays a summary of the boxplots displayed at each step, including the mean, standard deviation, median, minimum and maximum values, and 1st and 3rd interquartile values. This information is displayed if the user accesses the link below the boxplots: "Boxplot summary (click to display)". This information can be used to further filter the final list of model combinations.

Visualize the best option for each step
- ● Feature filter
- ○ Feature ranking
- ○ Size selection
- ○ Classification

Cross validation mean AUC
Anova p - value: 0.877

mean AUC

mad (fraction = 0.25)   mad (fraction = 0.5)   mad (fraction = 0.75)   mean-expression

■ Best mod
■ Other mo
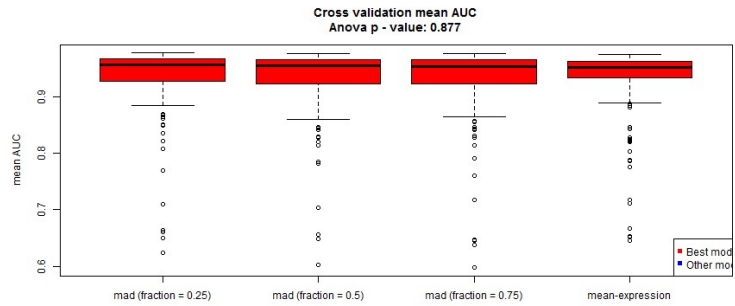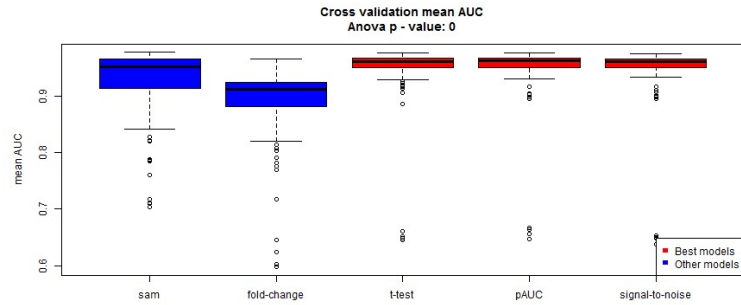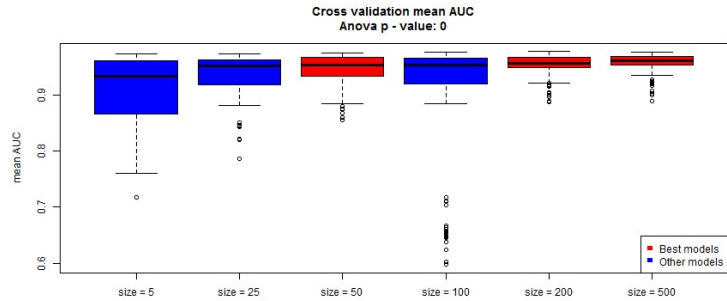
Boxplot summary (click to display)

Show 10 entries   Search:

	mad (fraction = 0.25)	mad (fraction = 0.5)	mad (fraction = 0.75)	mean-expression
Mean	0.938	0.935	0.935	0.934
SD	0.056	0.058	0.058	0.058
Min.	0.884	0.86	0.864	0.89
1st Qu.	0.928	0.924	0.923	0.934
Median	0.958	0.956	0.954	0.953
3rd Qu.	0.968	0.967	0.966	0.963
Max.	0.978	0.977	0.977	0.976

Showing 1 to 7 of 7 entries   Previous   1   Next

TukeyHSD comparison

Show 10 entries   Search:

	difference in means	adjusted p-value
(mad (fraction = 0.5)) -- (mad (fraction = 0.25))	-0.003	0.952
(mad (fraction = 0.75)) -- (mad (fraction = 0.25))	-0.004	0.923
(mean-expression) -- (mad (fraction = 0.25))	-0.004	0.865
(mad (fraction = 0.75)) -- (mad (fraction = 0.5))	-0.001	1
(mean-expression) -- (mad (fraction = 0.5))	-0.001	0.995
(mean-expression) -- (mad (fraction = 0.75))	-0.001	0.999

Showing 1 to 6 of 6 entries   Previous   1   Next

Figure 16. Feature filtering. In this example, all four feature filtering methods perform similarly well across all the predictors that incorporate them (ANOVA p=0.87).

Figure 17. Feature ranking. The top feature ranking methods in this case are t-test, pAUC and signal-to-noise (colored in red).

Figure 18. Biomarker size selection. The best biomarker sizes in this case are 50, 200, 500 features (colored in red). However, we recommend the user to consider the minimal size biomarker, since fewer features can be more easily translated in a clinically useful test.

Figure 19. Selection of the classifier. The best classifiers in this case are random forest, naïve bayes, svm and weighted voting (colored in red).

Best combination of models

Show [10 ∨] entries   Search: [   ]

| Model	Feature filter	Feature ranking	Size selection	Classification	Mean AUC
356	mad (fraction = 0.5)	pAUC	size selection (size = 50)	weighted voting	0.975
566	mad (fraction = 0.75)	pAUC	size selection (size = 50)	weighted voting	0.975
146	mad (fraction = 0.25)	pAUC	size selection (size = 50)	weighted voting	0.974
734	mean-expression	t-test	size selection (size = 50)	weighted voting	0.974
188	mad (fraction = 0.25)	signal-to-noise	size selection (size = 50)	weighted voting	0.972
398	mad (fraction = 0.5)	signal-to-noise	size selection (size = 50)	weighted voting	0.972
608	mad (fraction = 0.75)	signal-to-noise	size selection (size = 50)	weighted voting	0.972
818	mean-expression	signal-to-noise	size selection (size = 50)	weighted voting	0.972
104	mad (fraction = 0.25)	t-test	size selection (size = 50)	weighted voting	0.971
314	mad (fraction = 0.5)	t-test	size selection (size = 50)	weighted voting	0.971

Showing 1 to 10 of 48 entries   Previous [1] 2 3 4 5 Next

Best combination of models

Show [10 ∨] entries   Search: [   ]

Model	Feature filter	Feature ranking	Size selection	Classification	Mean AUC
524	mad (fraction = 0.75)	t-test	size selection (size = 50)	weighted voting	0.971
142	mad (fraction = 0.25)	pAUC	size selection (size = 50)	random forest	0.97
353	mad (fraction = 0.5)	pAUC	size selection (size = 50)	naive bayes	0.97
776	mean-expression	pAUC	size selection (size = 50)	weighted voting	0.97
143	mad (fraction = 0.25)	pAUC	size selection (size = 50)	naive bayes	0.969
145	mad (fraction = 0.25)	pAUC	size selection (size = 50)	svm	0.969
563	mad (fraction = 0.75)	pAUC	size selection (size = 50)	naive bayes	0.969
731	mean-expression	t-test	size selection (size = 50)	naive bayes	0.969
101	mad (fraction = 0.25)	t-test	size selection (size = 50)	naive bayes	0.968
185	mad (fraction = 0.25)	signal-to-noise	size selection (size = 50)	naive bayes	0.968

Showing 11 to 20 of 48 entries   Previous 1 [2] 3 4 5 Next

Figure 20. Best selected predictors. In this case we found 48 equivalently good biomarkers (20 of them are shown above). For the biomarker size we include only the smallest set of features that performs the best (in this case 50 features).

3.2.1.2 Comparison with random predictions

The *Random mean AUC* tab displays the performance of each method in each step compared to a random. For the random experiments we apply the same methods, using a random shuffle class label of the samples. The results generated using the random class label represent a quality control step, confirming that the real performance is not a technical artifact. For each step, we display the random results below the real performance for each method (Figure 21).
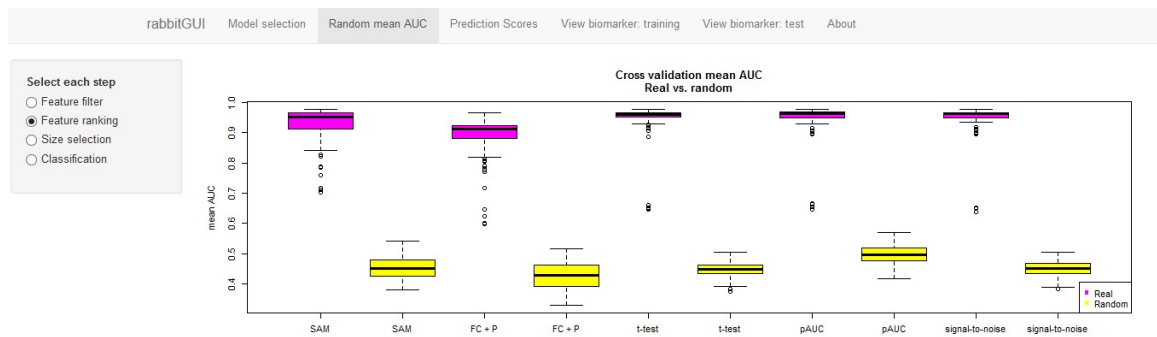
Figure 21. Real performance vs. random performance for each method in a step (this example shows the *feature ranking* step).

3.2.1.3 Visualize sample-level prediction scores

The *Prediction scores* tab allows the user to navigate through all potential biomarkers and visualize the sample-level prediction scores of all cross-validation runs (Figure 22). The models are sorted by the highest ROC AUC and the distribution of the scores are colored differently for the two classes. A

selection menu is available for the user to visualize a particular predictor by its index or name.
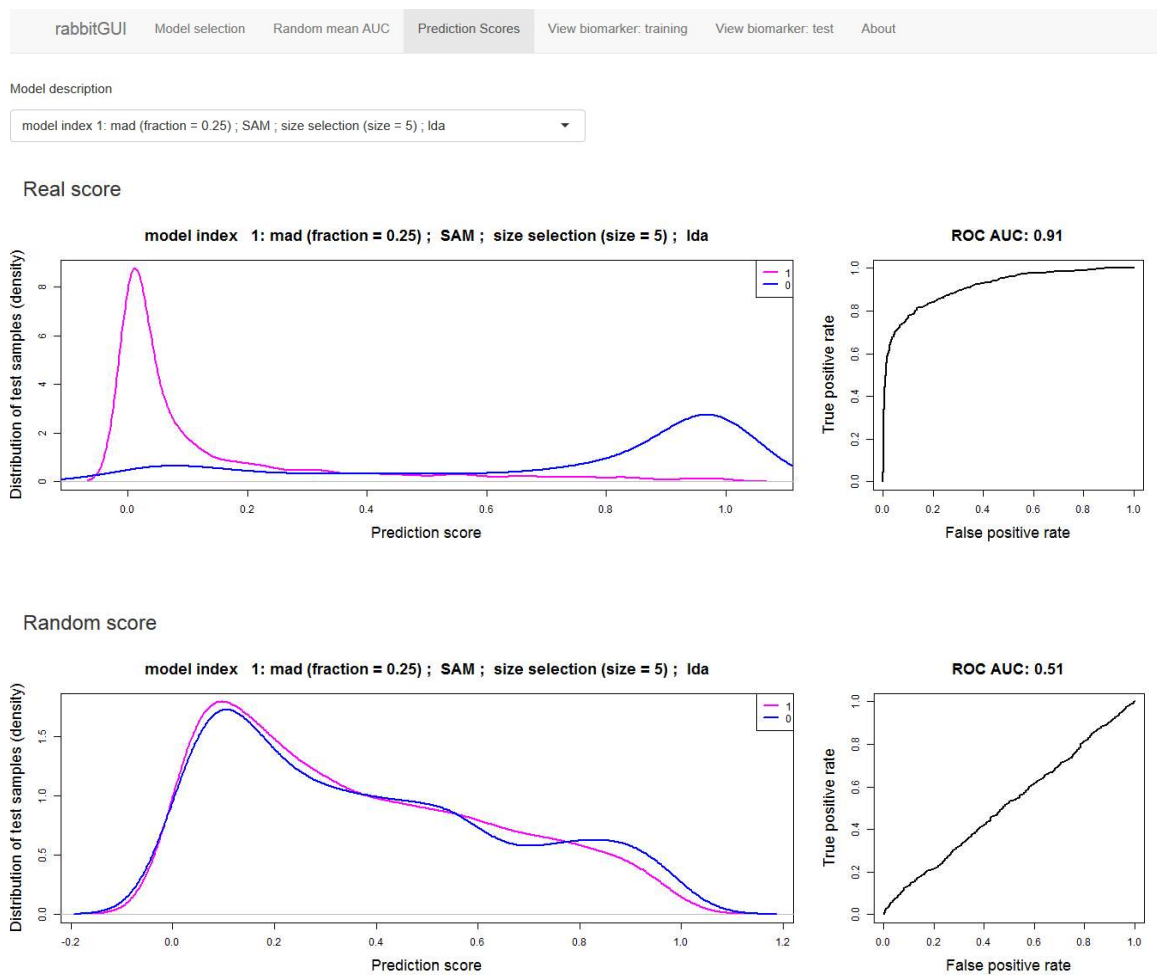
Figure 22. Visualize sample-level prediction scores for each predictor, for both the real and the random shuffle class label tests.

3.2.1.4 Visualize heatmap

This tab allows the user to visualize any heatmap, by uploading the following files:

- a .csv file with the expression matrix (samples on the columns and genes on the rows);

- a .csv file indicating the sample phenotype (a column vector with 0/1 for each sample, assuming the samples are in the same order as in the expression matrix);

- a .txt file with a list of features (the same format as the feature files generated by *rabbit*).

Bash scripts that collect the features for each model, across all cross-validation runs, have been implemented. To visualize all potentially relevant features based on the cross-validation results, the union of all selected features is considered. More details can be found in the section 3.3.2.

However, this tab is general and any heatmap can be displayed by the user, if the proper inputs are provided (Figure 23).
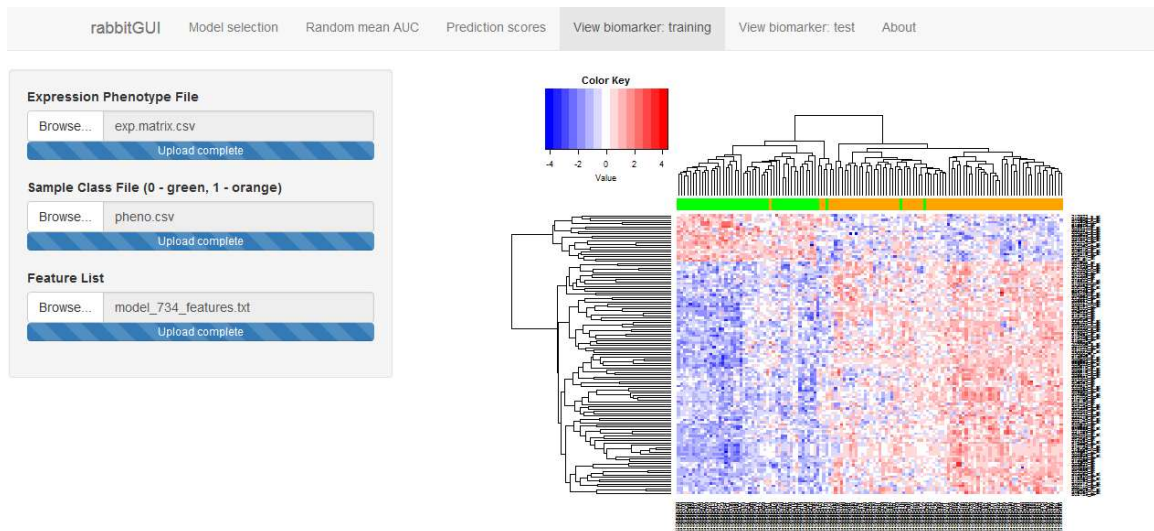
Figure 23. Visualize the heatmap of biomarker features.

3.3 Methods

3.3.1 Shiny applications

Shiny is a new package from *RStudio* that makes it easy to build interactive web applications with R (http://rstudio.github.io/shiny/tutorial/). It provides automatic binding of inputs and outputs and pre-built widgets that facilitates the process of developing user-friendly, interactive, and powerful applications.

The main features of *Shiny* apps are the following:

- Build useful web applications with only a few lines of code;

- Shiny applications are automatically interactive;

- Outputs change instantly as users modify inputs, without requiring a reload of the browser;

51

- Shiny user interfaces can be built entirely using R, or can be written directly in HTML, CSS, and JavaScript for more flexibility;

- Shiny works in any R environment (Unix, Windows or Mac);

- Pre-built output widgets for displaying plots, tables, and R objects.

A Shiny app is based on client-server architecture, where the client issues requests to the server based on the user inputs. The information provided by the server is then displayed by the client.

In this work leverage the flexibility of Shiny package and develop a usr-friendly GUI to evaluate and interact with the outputs of *rabbit* biomarker discovery package.

3.3.2 Installing rabbit, rabbitGUI and dependencies

rabbit pipeline is based on caret package, and it requires R>=3.2.3. In order to run *rabbit* pipeline with all the default models, the following R dependencies are required: *pbkrtest (R >= 3.2.3); car (R >= 3.2.0); nlme (R >= 3.0.2); devtools; multtest; impute; samr; e1071; randomForest; klaR; kernlab; glmnet; limma; genefilter.*

After these dependencies are installed, the user can install *rabbit* from github, as following*: install_github("jperezrogers/rabbit", ref="master").*

rabbitGUI does not require the installation of *rabbit* package, as long as the proper inputs are provided (see subsection 3.3.3 for more details). However, it requires that the following database of model names and indices is present in the directory *specs,* under the local directory of rabbit source code.

52

This database can be saved from rabbit: *stockPipeline$getModelSpecs()*.

Alternatively, if *rabbit* package is not installed, the *file specs.csv* can be found from the following web address:

https://github.com/anabrandusa/rabbitGUI/tree/master/rabbitGUI_code/specs.

In addition, in order to run *rabbitGUI*, the following R dependencies need to be installed: *shiny; DT; pROC; ROCR; markdown; gplots.*

The Shiny app can be downloaded from the following web address:

https://github.com/anabrandusa/rabbitGUI/tree/master/rabbitGUI_code, and launched by running *app.R* script.

The application can run on any system with an R environment (Unix, Windows or Mac). Online details about the GUI are available at the following web address: https://github.com/anabrandusa/rabbitGUI/blob/master/README.md.

3.3.3 Processing and aggregating the classification results from rabbit pipeline

rabbit biomarker discovery tool is able to run multiple models in cross-validation in parallel environment. It has been configured to run in parallel on an SGE cluster using *qsub*. The function *run* takes as inputs an object that defines the pipeline's configuration, an expression matrix, a binary phenotype vector, the current cross-validation iteration, the output directory, and other parameters (https://github.com/jperezrogers/rabbit/tree/master/vignettes). The function is submitted an available computing node for each iteration, using *qsub*.

The output consists of n directories (corresponding to n cross-validation iterations). Each of these directories contain m other directories (corresponding

to the outputs of m tested predictors). For each predictor two output files are generated: predictions.txt (the sample-level prediction scores) and features.txt (the list of features used for classification).

The role of *rabbitGUI* is to summarize the overall performance of all models and iterations. To standardize the inputs of the web-based interface, the outputs from rabbit pipeline are processed as following. For each iteration and each model, ROC AUC is computed. This can be done in parallel for each iteration, by running *process.rabbit.outputs.sh*. The results are then merged into one file by *create.shiny.inputs.sh*. The ROC AUC across n cross-validation iterations is computed in two different ways:

- mean AUC across all iterations;

- AUC of all test samples in all iterations.

The *getFeaturesInParallel.sh* submits a qsub job for each model, collecting all features from all cross-validation iterations for that model. Then by running *getFeatureUnion.sh*, a file containing the union of all features from all iterations is generated for each model. This file can be uploaded into *rabbitGUI* to visualize the heatmap all features selected in cross validation for a particular model.

All the Bash and R scripts that collect the results from *rabbit*, compute the performance and prepare the inputs for *rabbitGUI* are available as open source (https://github.com/anabrandusa/rabbitGUI/tree/master/prepare_inputs_rabbitGUI).

rabbitGUI inputs are stored under the local directory of the application, in data_clasified and data_random folders. Each of these two directories contain alldata.csv and aucmeans.csv files.

The results from data_random are obtained by running rabbit with a randomly assigned class of the samples, serving as a quality control of the data and methods used. Examples of rabbitGUI input files are provided on github and in Figures 24 and 25.

Iteration	Model	ModNames	Sample	Score	Classification	TrueClass	Direction
100	1	model_1	GSM505327	0.010404611	1	1	0
100	1	model_1	GSM505338	0.398484521	1	1	0
100	1	model_1	GSM505340	0.997606848	0	0	0
100	1	model_1	GSM505350	0.990142108	0	0	0
100	1	model_1	GSM505358	0.988532163	0	0	0
100	1	model_1	GSM505359	0.966351073	0	0	0
100	1	model_1	GSM505363	0.003459271	1	1	0
100	1	model_1	GSM505366	0.003736881	1	1	0
100	1	model_1	GSM505369	0.999815489	0	0	0
100	1	model_1	GSM505370	0.99933046	0	0	0
100	1	model_1	GSM505380	0.842934354	0	0	0
100	1	model_1	GSM505386	0.888871157	0	0	0
100	1	model_1	GSM505418	0.983552443	0	0	0
100	1	model_1	GSM505420	0.108742245	1	1	0
100	1	model_1	GSM505423	0.949451897	0	0	0
100	1	model_1	GSM505448	0.164592802	1	1	0
100	1	model_1	GSM505450	0.998652466	0	0	0
100	1	model_1	GSM505452	0.000900092	1	1	0
100	1	model_1	GSM505459	0.613685574	0	0	0
100	1	model_1	GSM505463	0.837544473	0	1	0
100	1	model_1	GSM505464	0.014022906	1	1	0
100	1	model_1	GSM505468	0.395794169	1	1	0
100	1	model_1	GSM505477	0.006256961	1	1	0
100	1	model_1	GSM505480	0.001095851	1	1	0
100	2	model_2	GSM505327	0.03	1	1	0
100	2	model_2	GSM505338	0.468	1	1	0

Figure 24. alldata.csv file is an input for rabbitGUI and contains the sample-level prediction scores of all predictors and all iterations merged in one csv file.

Model	ROCRAUCMEAN	ROCRAUC
34	0.979425771	0.9779416
27	0.979016165	0.9776332
377	0.978953606	0.9770676
167	0.9794241	0.9769643
244	0.978451653	0.9769577
454	0.979288583	0.9767694
587	0.978660604	0.9767584
118	0.978536566	0.9763899
237	0.977485068	0.9762947
202	0.978999006	0.9761804
160	0.977059286	0.9760287
419	0.977754385	0.9759899
545	0.978307166	0.9757943
335	0.977776521	0.9757034

Figure 25. *aucmeans.csv* file is an input for *rabbitGUI* and contains the ROC AUC values of all models, computed across all cross-validation iterations as a mean AUC and as the AUC of all test samples in all iterations. The models in the table are sorted in descending order by ROCRAUC values.

3.3.4 Biomarker discovery methods available from rabbit and rabbitGUI

   rabbitGUI displays and evaluates the results from all methods tested by *rabbit* pipeline. *rabbit* is flexible and allows the user to activate and deactivate specific methods. In addition, new algorithms can be added to the collection. However, this section briefly describes the default models integrated in *rabbit* package and evaluated by *rabbitGUI*.

3.3.4.1 Feature filtering

This step consists of unsupervised feature filtering methods. The following methods select the features that present the most variability across samples, having the highest potential to discriminate the data.

- Median Absolute Deviation (MAD)
    - ranks each feature X by the absolute median deviation across samples i=1,…,n:

        MAD(X) = median(abs(Xi-median(X)));

    - selects the top q features;
    - MAD is ran three times independently by default, assigning the following q values: 25%, 50% and 75%.
- Mean-expression
    - compares each feature vector with the vector of mean expression across all features;
    - genes with p<0.05 are selected.

3.3.4.2 Feature selection

This step is a supervised gene ranking. The features are ranked by their ability to discriminate between the two classes.

- Significance Analysis of Microarrays (SAM)
    - non-parametric statistics;

- computes a test statistic for the relative difference in gene expression between the two groups, based on permutation analysis of expression data, and calculates a false discovery rate;
- *samr()* function from *samr* R package.

- Moderated t-test and fold change (FC+P)
  - genes are first scored by moderated t-statistic (Smyth 2004), using *eBayes* function from *limma* R package;
  - genes with a p-value less than 0.05 are selected and then ranked by $\log_2$ fold-change.

- T-test
  - standard two-sample t-test assuming equal variances.

- Partial AUC (pAUC)
  - integrates AUC to a limit p given as parameter (default p=0.1);
  - *rowpAUCs()* function from *genefilter* R package;

- Signal-to-noise
  - defined by (Golub et al. 1999): each feature is ranked by the difference in means between the two groups relative to the standard deviations within the two groups;

3.3.4.3 Biomarker size selection

Using the features ranking described previously, this step selects the top n features. In case there are fewer than n features, all features are included.

3.3.4.4 Classification

This step uses the set of features selected by the previous steps applies different supervised algorithms for binary classification.

- Linear discriminant analysis (LDA)
  - o finds a linear combination of features that characterizes or separates two or more classes;
  - o assumes continuous independent variables and categorical dependent variable;
  - o included in *caret* R package (*method="lda"*).
- Weighted Voting
  - o defined by (Golub et al. 1999): this procedure uses a fixed subset of genes and makes a prediction on the basis of the expression level of these genes in a new sample; each informative gene generates a "weighted vote" for one of the classes, with the magnitude of each vote dependent on the expression level in the new sample and the degree of that gene's correlation with the class distinction; the votes are summed to determine the winning class, as well as the prediction score;
  - o included in *caret* R package (function *wv.model()*).
- Support Vector Machines (SVM)
  - o discriminative classifier defined by a separating hyperplane; an SVM model is a representation of the samples as points in space,

59

mapped so that the separate categories are divided by a clear gap that is as wide as possible;

- o included in *caret* R package (*method="svmRadial"*).

- Naïve Nayes

  - o probabilistic classifier based on applying Bayes' theorem, under the assumption that features are independent;

  - o included in *caret* R package (*method="nb"*).

- Elastic Net

  - o regularized regression method that linearly combines the penalties of the lasso and ridge methods for a logistic regression model;

  - o included in *caret* R package (*method="glmnet"*).

- K-Nearest Neighbors

  - o assigns a sample to the class of its closest neighbor in the feature space, based on a defined distance, such as the Euclidian distance;

  - o included in *caret* R package (*method="knn"*).

- Random Forest

  - o selects random subsets of the feature-variables and creates a decision tree on each subset by maximizing information gain

  - o it classifies the test samples using all trees; the class is assigned based on the prediction of the majority of the trees;

  - o included in *caret* R package (*method="rf"*).

3.4 Discussion

rabbitGUI is an open source GUI for biomarker discovery. The visual component is designed to guide the user to select the best predictors from a pool of about one thousand model combinations.

rabbit and *rabbitGUI* can serve as a user-friendly framework for biomarker discovery. The main advantages of this software system are the following:

- standardized tool for biomarker discovery;

- provides an easy-to-use GUI for evaluating and interpreting the results;

- availability: free and open source;

- flexibility: it can be extended by adding new methods and visualization components;

- multi-platform: it can run on different platforms (Unix, Windows or Mac);

- parallelized for SGE cluster.

The framework has been designed to be a user-friendly resource for researchers in all fields, including both computational and experimental biologists. We plan to further extend both *rabbit* and *rabbitGUI* by integrating new methods and visualization features.

Currently, *rabbitGUI* is a powerful visual resource to interpret and display the results generated by *rabbit* pipeline.

In future work, we plan to extend the web-based interface to provide a configuration menu for *rabbit* pipeline and the option to run the application interactively from the GUI.

In addition, we plan to extend this software system to run on a cloud computing platform, such as Amazon or Google Cloud Platform, providing quick and user-friendly access. We believe this tool will be a valuable resource for the translational bioinformatics research community.

CHAPTER FOUR

Integrative microRNA networks reveal potential roles for miR-449/34 family in COPD and ILD

4.1 Introduction

Complex diseases arise from a heterogeneous molecular interplay between genetic and genomic alterations. Although biological processes, such as chronic inflammation, apoptosis, and oxidative stress, have been found to play a role in COPD and ILD pathogenesis, knowledge remains limited about the key molecular interactions of these diseases (Steiling et al. 2013).

Several computational approaches have been applied to infer causality from biological data, including Bayesian networks (Vignes et al. 2011; Aliferis et al. 2010; Dondelinger, Husmeier, and Lèbre 2012), factor graphs (Ng et al. 2012; Vaske et al. 2010) and ridge and least absolute shrinkage and selection operator (Omranian et al. 2016).

In 2009, E. Schadt's group proposed a data driven statistical framework to infer mediators of genetic factors associated with quantitative traits. The causality is modeled as a "chain" of mathematical conditions that test the strength of the associations (Millstein et al. 2009). This approach has been applied to characterize the role of microRNAs (miRNAs) within gene regulatory networks (Su, Kleinhanz, and Schadt 2011). In this work, we take a similar approach to

unravel the miRNA dysregulations that mediate the genetic factors of COPD and ILD.

We profiled miRNA expression from 351 patients from the Lung Genomics Research Consortium (LGRC). Previous LGRC GWAS studies have found associations of genetic factors with adult lung function (Tang et al. 2014) and idiopathic pulmonary fibrosis (Noth et al. 2013). We integrate miRNA expression with publicly available SNP and mRNA data. We first infer causality of the molecular associations between SNP, miRNA and mRNA using show that there is a difference in the connectivity between the disease networks compared to control. The networks capture the differences in miRNA regulation, revealing new miRNA drivers of disease.

4.2 Results

4.2.1 eQTL analysis

Using the LGRC cohort, we profiled miRNA expression via small-RNA sequencing from 351 lung tissue samples from patients with COPD, ILD and controls (Table 4).

By "anchoring" expression data with genetic information, we can identify key miRNA regulators of gene expression associated with COPD. Therefore, we utilized a subset of the 262 lung tissue samples with profiled miRNA expression, as well as publicly available SNP and mRNA data (Table 5).

Covariates	Control (n=62)	ILD (n=144)	COPD (n=145)
Smoking Status [1‡]	2 current, 38 former, 19 never, 3 NA	5 current, 85 former, 50 never, 4 NA	8 current, 129 former, 6 never, 2 NA
Age [‡]	63.1 (12.0)	61.2 (10.2)	64.4 (9.9)
Pack Years [*1‡]	41.1 (36.6)	26.3 (19.9)	55.9 (39.0)
Gender	31 males, 31 females	78 males, 66 females	86 males, 59 females
FEV1/FVC [*1‡]	0.77 (0.1)	0.83 (0.1)	0.5 (0.2)
Percent Emphysema [1‡]	0.7 (1.0)	0.8 (1.7)	16.6 (18.0)

Table 4. Demographics table of samples with available miRNA and mRNA data. [*]Significantly different between ILD and Control (p<0.05); [1]Significantly different between COPD and Control (p<0.05); [‡]Significantly different between ILD and COPD (p<0.05); p-values for gender and smoking status were calculated by using *Fisher's exact test*; p-values for age, pack years, FEV1/FVC and Percent Emphysema were calculated by using *Student's t-test*.

Covariates	Control (n=38)	ILD (n=113)	COPD (n=111)
Smoking Status [1‡]	2 current, 24 former, 12 never	4 current, 71 former, 38 never	7 current, 99 former, 5 never
Age	65.5 (11.5)	62.2 (9.2)	63.8 (9.2)
Pack Years [*‡]	49.9 (40.8)	26.3 (20.4)	55.1 (37.8)
Gender	22 males, 16 females	61 males, 52 females	65 males, 46 females
FEV1/FVC [*1‡]	0.76 (0.06)	0.83 (0.07)	0.49 (0.24)
Percent Emphysema [1‡]	0.7 (1.0)	0.74 (1.7)	17.0 (18.3)

Table 5. Demographics table of samples with available miRNA and mRNA data. [*]Significantly different between ILD and Control (p<0.05); [1]Significantly different between COPD and Control (p<0.05); [‡]Significantly different between ILD and COPD (p<0.05); p-values for gender and smoking status were calculated by using *Fisher's exact test*; p-values for age, pack years, FEV1/FVC and Percent Emphysema were calculated by using *Student's t-test*.

Using *PLINK* (Purcell et al. 2007) we performed an eQTL (expression quantitative trait loci) analysis, considering both CIS and TRANS interactions, where CIS was defined as <1MB, and including both miRNA and mRNA features.

We identified all genes and miRNAs associated with a SNP by ANOVA while correcting for age, gender, smoking status, and population structure ($p<0.0005$). The number of significant CIS and TRANS eQTLs ($p<0.05$) identified within each group are presented in Figure 26. Most of the associations between SNPs and miRNAs are TRANS, and very few are CIS (see section 4.3.4 for the definition of CIS/TRANS associations).

The QQ-plots show significant p-values for both CIS (local) and TRANS (distant) associations, in all three groups: COPD (Figure 27), and ILD (Figure 28), control (Figure 29).

		All samples (n=262)	Control (n=38)	COPD (n=111)	ILD (n=113)
Gene-SNP Pairs	Cis	34450	328	8248	10993
	Trans	864206	546	110981	68412
miR-SNP Pairs	Cis	302	0	53	52
	Trans	3378	0	557	362
Unique Genes	Cis	3093	54	955	1156
	Trans	14881	408	6850	4393
Unique microRNAs	Cis	63	0	17	14
	Trans	615	0	221	144
Unique SNPs	Cis	29390	324	7667	10113
	Trans	275796	253	62335	48385

Figure 26. Number of significant eQTLs ($p<0.05$).
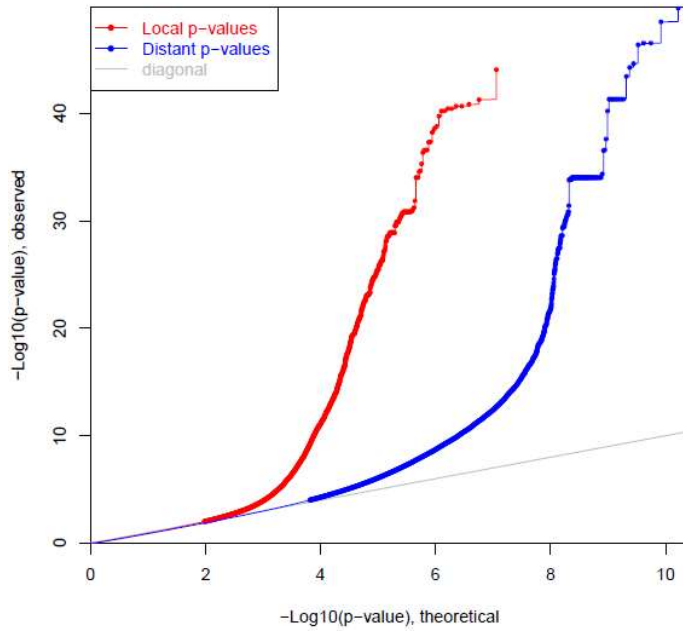
Figure 27. QQ-plot in COPD patients.
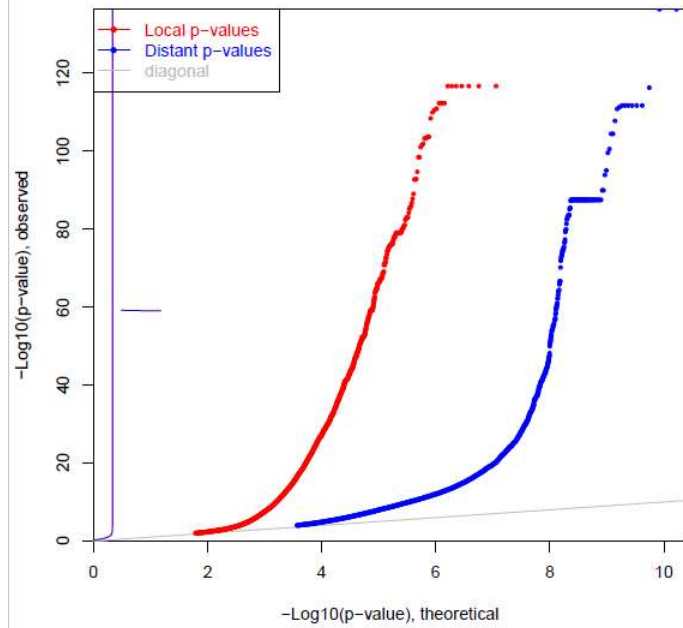
Figure 28. QQ-plot in ILD patients.
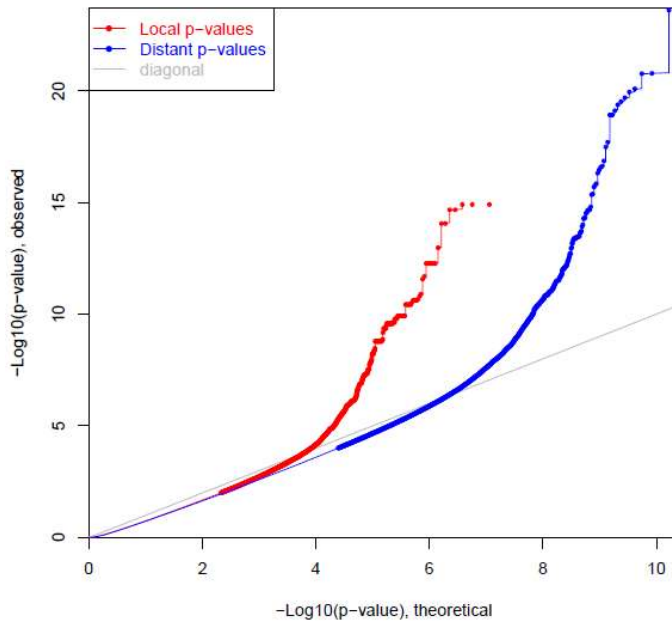
Figure 29. QQ-plot in control patients.

4.2.1 miR-34/449 family is differentially connected in disease compared to control

Next, we built integrative networks within the COPD, ILD, and control patients using the causality inference test (CIT) (Millstein et al. 2009). CIT assesses the hypothesis that a potential mediator between an initial randomized variable and an outcome variable is causal for that outcome. Causal and independent relationships are defined as series of conditions of associations between the three variables, corresponding to SNP, microRNA and mRNA nodes (Figure 30).

Then, we explored the *scale-free* property of the networks (Barabási and Oltvai 2004; Barabasi and Bonabeau 2003) by computing the frequency of node

degree in log-scale. As expected, the networks are biologically meaningful, presenting a negative linear relationship between the node degree and the frequency of node degree in log scale (Figure 31).
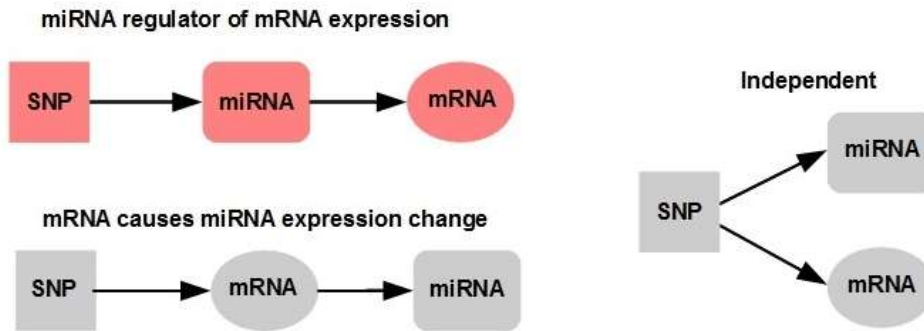
Figure 30. Network construction; we select those SNP-miRNA-mRNA triplets where the SNP-mRNA relationship is defined by a miRNA mediator. We filter out independent relationships and those triplets where the SNP is not associated with the miRNA.
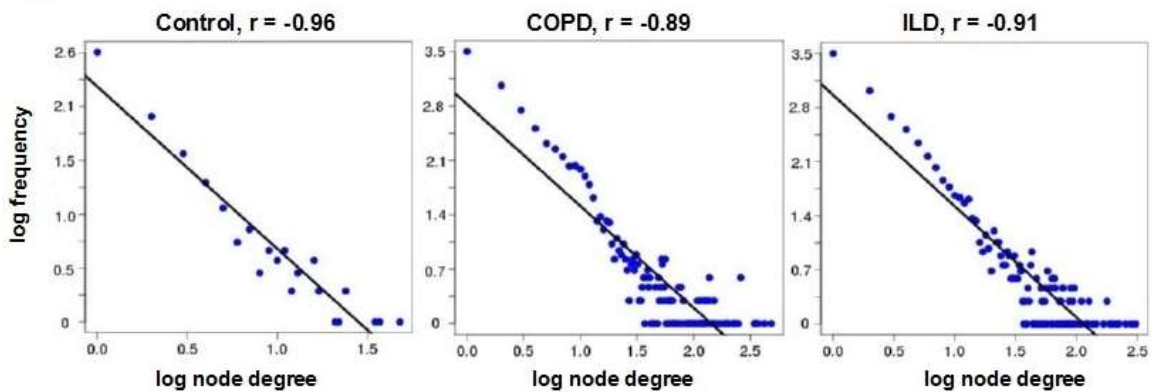
Figure 31. The CIT networks follow a power law. The negative correlation between the frequency of node degree and the node degree indicates that the networks are *scale-free*.

Furthermore, we identified the miRNAs predicted to interact with the most genes in each network (Table 6).

Top mostly connected miRNAs in COPD	Top mostly connected miRNAs in ILD	Top mostly connected miRNAs in Control
hsa-miR-27a-5p *	hsa-miR-92b-3p *	hsa-miR-21-5p *
hsa-miR-190b *	hsa-miR-449a *	hsa-miR-4802-3p *
hsa-miR-449b-5p *	hsa-miR-200a-5p *	hsa-miR-146a-5p *
hsa-miR-449a *	hsa-miR-31-5p *	hsa-miR-378c *
hsa-miR-449c-5p *	hsa-miR-92b-5p *	hsa-miR-142-3p *
hsa-miR-4423-5p *	hsa-miR-449c-5p	hsa-miR-146b-5p *
hsa-miR-92b-3p *	hsa-miR-200b-3p *	hsa-miR-421 *
hsa-miR-34c-3p	hsa-miR-31-3p *	hsa-miR-30a-3p
hsa-miR-205-5p	hsa-miR-190b *	hsa-miR-378a-5p *
hsa-miR-23a-5p *	hsa-miR-449b-5p *	hsa-miR-378a-3p *
hsa-miR-509-3p-2	hsa-miR-511-1 *	hsa-miR-330-5p *
hsa-miR-509-3p-3	hsa-miR-511-2 *	hsa-miR-425-5p *
hsa-miR-509-3p-1	hsa-miR-34c-5p	hsa-miR-378i *
hsa-miR-30a-3p	hsa-miR-34c-3p	hsa-miR-26a-5p-1 *
hsa-miR-34b-3p *	hsa-miR-146b-5p *	hsa-miR-26a-5p-2 *
hsa-miR-1185-1-3p	hsa-miR-2110 *	hsa-miR-223-5p *
hsa-miR-125b-1-3p	hsa-miR-34b-5p	hsa-miR-191-5p *
hsa-miR-654-5p	hsa-miR-450b-5p	hsa-miR-30a-5p
hsa-miR-485-5p	hsa-miR-200a-3p	hsa-miR-509-3p-2
hsa-miR-34c-5p	hsa-miR-34b-3p	hsa-miR-509-3p-3
		hsa-miR-509-3p-1
		hsa-miR-5571-3p
		hsa-miR-301b *
		hsa-miR-766-3p *
		hsa-miR-199b-5p *
		hsa-miR-34a-5p *

Table 6. Top 20 mostly connected miRNAs in each phenotype. * indicates the significant FDR-adjusted p-values (q<0.2) by a Fisher's exact test that determines the difference in the connectivity frequencies between the two phenotypes for each miRNA.

Dysregulated miRNAs that interact with large gene modules are more likely to cause important phenotypic changes. Members of the miR-449/34 family were found to be among the top ranked in COPD and ILD networks (Figure 32), indicating that miR-449/34 family has a greater impact on gene expression regulation in disease compared to control group.

miR-449 members are located on chromosome 5 and miR-34 members, on chromosome 11. However, these miRNAs are known to be co-expressed, presenting similar biological functions (Fededa et al. 2016). Members of miR-34/449 family can promote airway differentiation by repressing the Notch pathway (Chevalier et al. 2015; Bae et al. 2012; Lizé, Klimke, and Dobbelstein 2011). In addition, these miRNAs were found to share an increased number of associated genes, as illustrated in Figure 33.
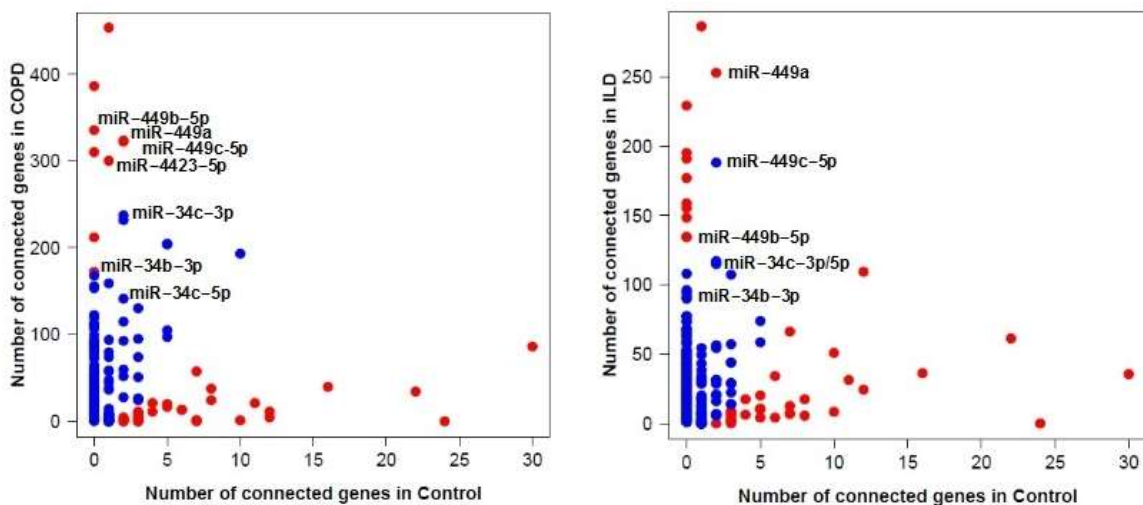
Figure 32. Top differentially connected microRNAs in COPD (right) and ILD (left).
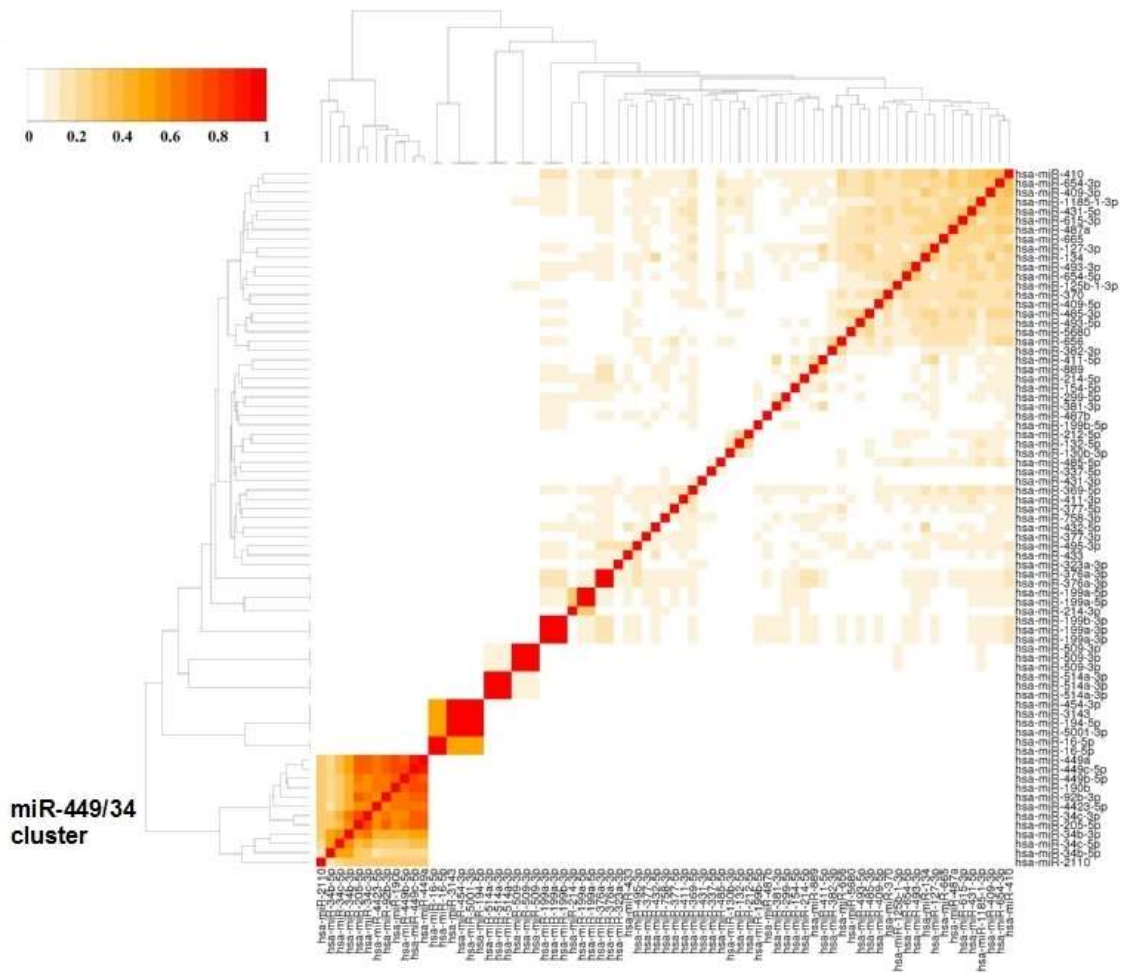
Figure 33. miR-449/34 modules present an increased number of shared genes by Jaccard index.

Furthermore, we observed that the combined set of genes that positively correlated with these miRNAs were enriched among genes that increase in expression over time when airway basal cells are differentiated at an air-liquid interface (ALI) (Ross et al. 2007). Gene enrichment results were significant by both GSVA, $p<10^{-3}$ (Figure 34a) and GSEA, $q<0.001$ (Figure 35).

This set of genes was also associated with genes that are positively correlated with airway wall thickening in patients with emphysema, by both GSVA, $p<10^{-4}$ (Figure 34b) and GSEA, $q<0.001$ (Figure 36). All these results suggest that the miR-449/34 family is playing a role in differentiation associated with the airway wall thickening phenotypes.
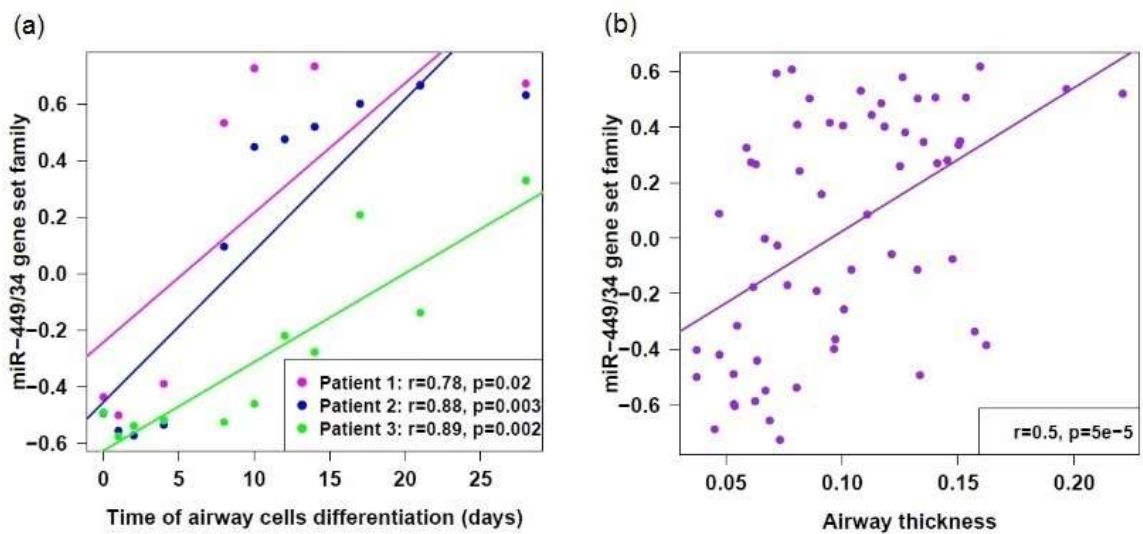
Figure 34. Enrichment of miR-449/34 modules by GSVA. (a) Enrichment of miR-449/34 gene set family with airway cells differentiation by GSVA. The set of genes that positively correlated with miR-449/34 family (406 genes) were enriched among genes that increase in expression with the airway epithelial cells differentiation in COPD; (b) Enrichment of miR-449/34 gene set family with increasing airway wall thickness in patients with emphysema by GSVA. The set of genes that positively correlated with miR-449/34 family were enriched among genes that increase in expression with airway wall thickening of patients with emphysema in COPD.

73
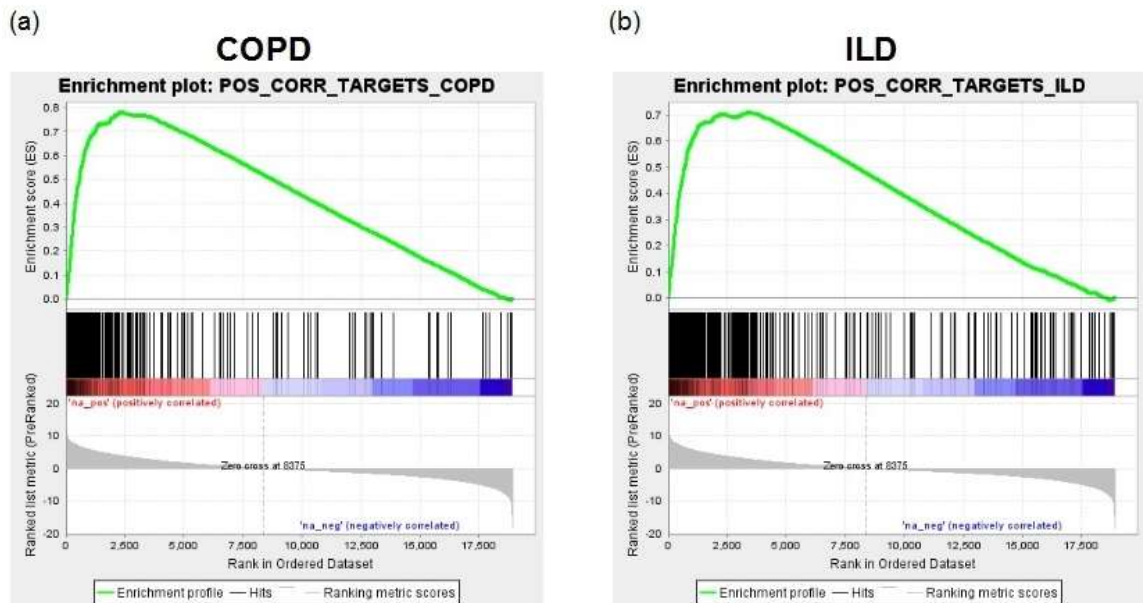
Figure 35. Enrichment of miR-449/34 associated genes with airway differentiation by GSEA. The set of genes that positively correlated with miR-449/34 family were enriched among genes that increase in expression with the airway epithelial cells differentiation (q≈0), both in (a) COPD and (b) ILD. Genes were ranked from those that increased in expression with time of differentiation (red) to those that decreased in expression with time of differentiation (blue) in a publicly available dataset (Ross et al. 2007).
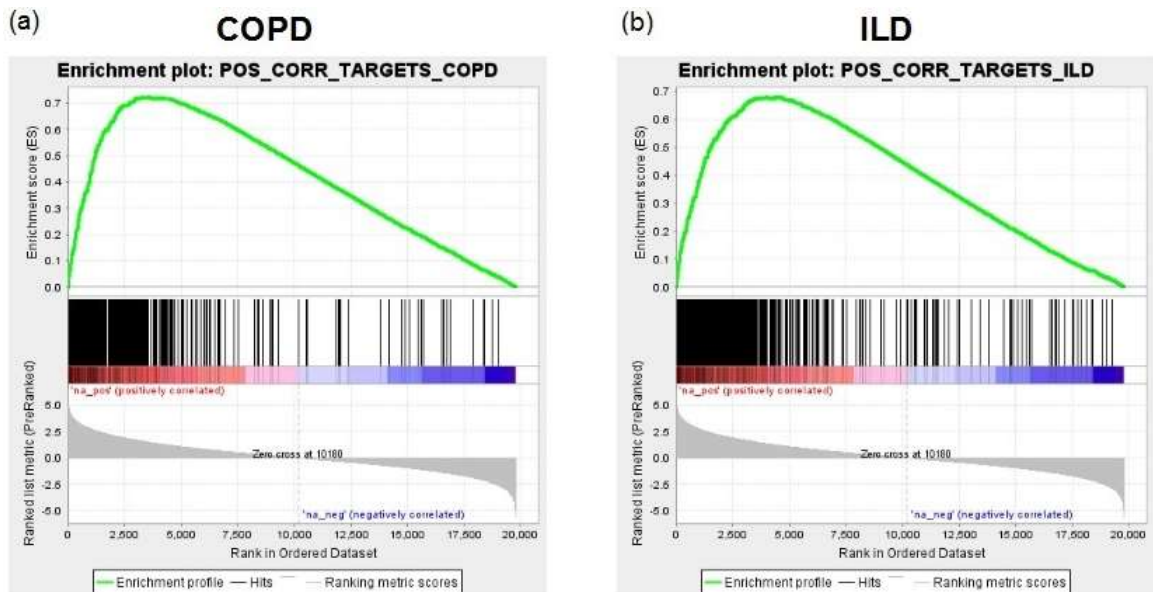
Figure 36. Enrichment of miR-449/34 associated genes with increasing airway wall thickness in patients with emphysema by GSEA. The set of genes that positively correlated with miR-449/34 family were enriched among genes that increase in expression with airway wall thickening of patients with emphysema (q≈0), both in (a) COPD and (b) ILD. Genes were ranked from those that increased in expression with thicker airway walls (red) to those that decreased in expression with thicker airway walls (blue) in an independent dataset of 60 airway samples from 8 different patients with emphysema.

4.2.2 SNPs associated with disease that regulate miR-34/449

Using the causality inference test (Millstein et al. 2009), we found 75 SNPs in COPD and 60 SNPs in ILD that may regulate miR-449/34 family. Some of these SNPs have been previously associated with asthma, inflammation, cancer and other degenerative diseases by GRASP (Leslie, O'Donnell, and Johnson 2014). Top significantly associated SNPs with COPD or ILD by a Fisher's exact test (q<0.25) are shown in Figure 37.

To illustrate the association of the SNP data with miRNA and mRNA expression, an example is provided. We considered the following triplet obtained by the causality inference test: rs525770_C → miR-449a → CLUAP1. Figure 38 shows the association of miR-449a with CLUAP1 expression (p<0.001). Figure 39 illustrate the association of rs525770_C variant with miR-449a expression (p<1e-05) and CLUAP1 expression (p<0.002), respectively.

Interestingly, using UCSC Genome Browser (Kent et al. 2002), we determined that rs525770_C falls in the genomic region of HS6ST3 gene. This gene is a Heparan sulfate (HS) sulfotransferase that modifies HS to generate structures required for interactions of HS with a variety of proteins. The protein coded by HS6ST3 is implicated in proliferation and differentiation, adhesion, migration, inflammation, blood coagulation, and other diverse processes. In addition, CLUAP1 (Clusterin associated protein 1) is known to be associated with immunoglobulin IgG1 (Gardin and White 2011).
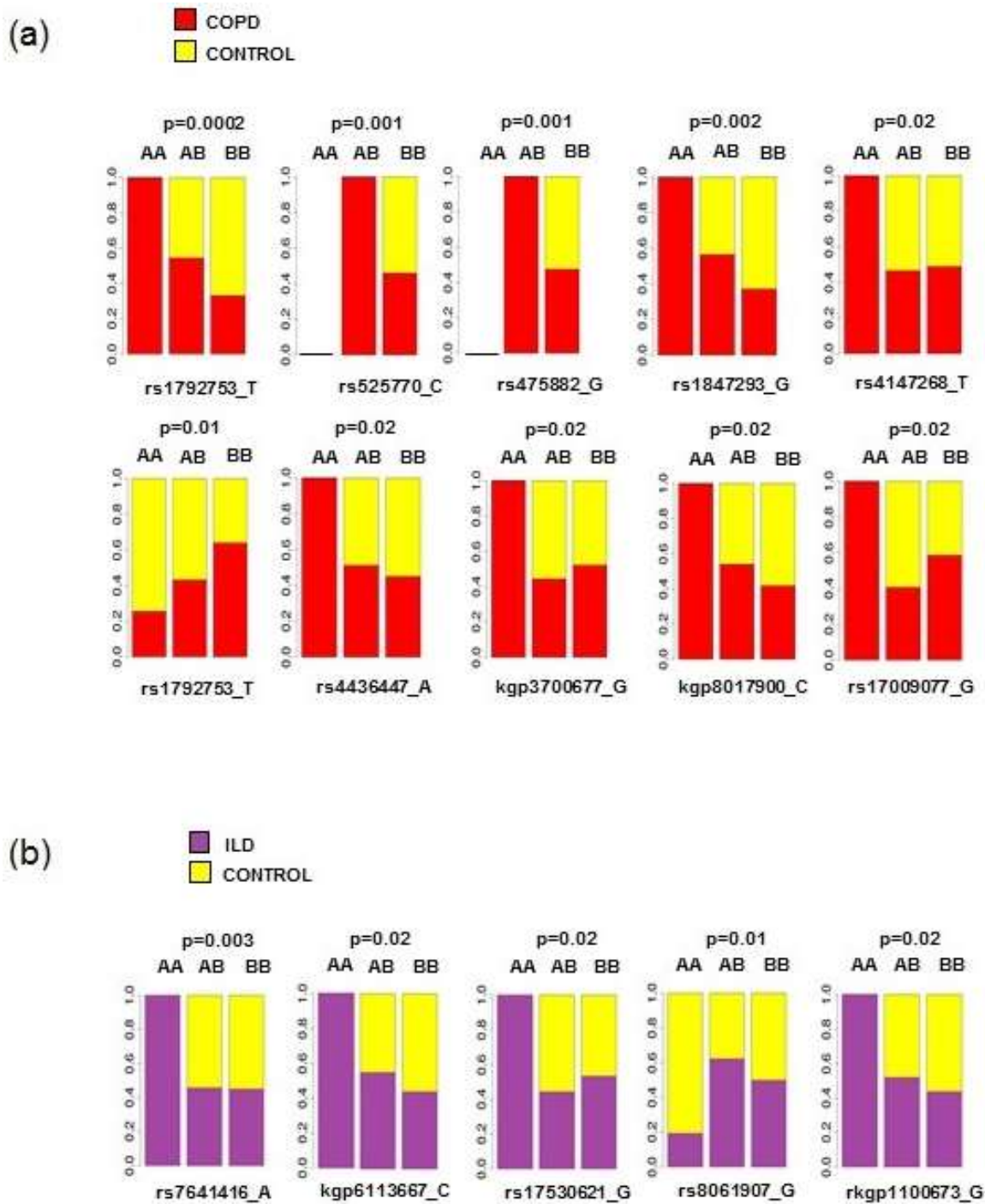
Figure 37. SNPs that are significantly associated with COPD and ILD. (a) SNPs that are significantly associated with COPD by a Fisher's exact test (q<0.25). (b) SNPs that are significantly associated with ILD by a Fisher's exact test (q<0.25).
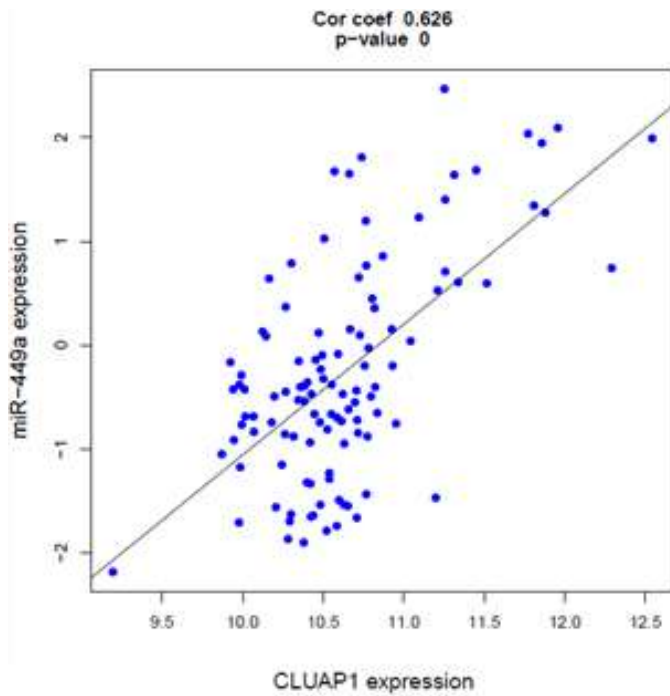
Figure 38. The association between miR-449a and CLUAP1 expression.
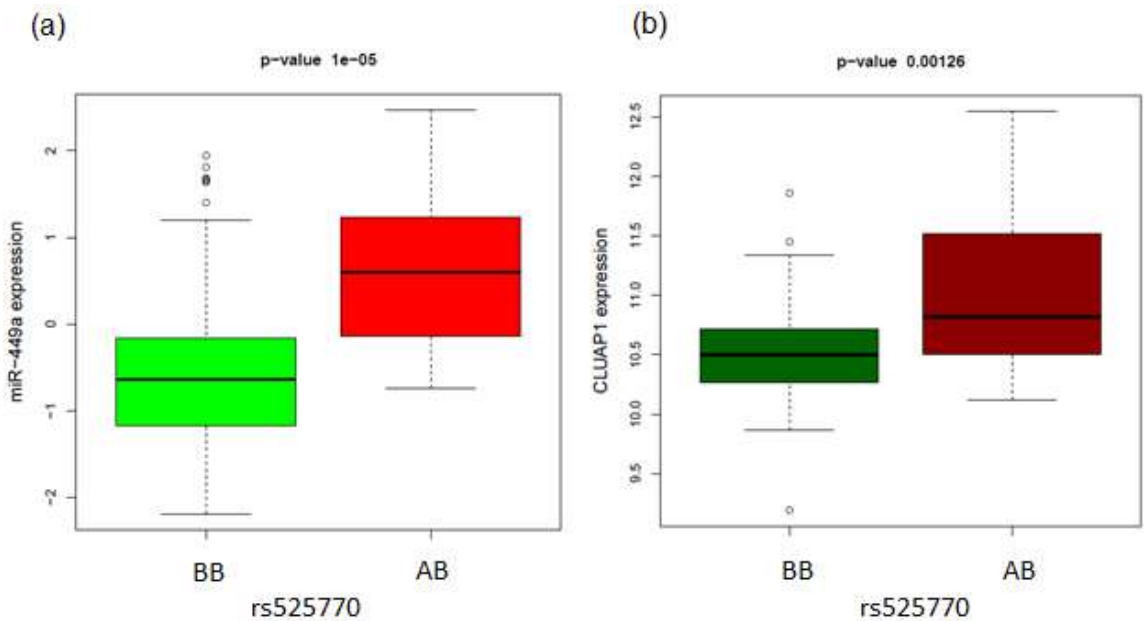
Figure 39. The association of rs525770_C variant with (a) miR-449a expression

and (b) CLUAP1 expression.

4.2.3 Differential connectivity of miRNA-mRNA regulatory networks

Next, we hypothesized that miRNA-mRNA regulatory networks are also differently connected between disease and normal states, independently of the associations with the SNP data. To test this hypothesis, a new approach to compute the module differential connectivity (MDC) of miRNA/mRNA association networks, is proposed. Each miRNA is assigned an MDC score, that captures the overall difference in the pairwise microRNA-gene correlation strengths between case and control networks (Figure 40). Then, we applied a permutation test by shuffling the class labels of the samples to determine the significance of real MDC scores. The computed p-values were adjusted by FDR correction.

$$MDC_{miRNA} = \frac{(W_A^C - W_A^{NC})^2 + (W_B^C - W_B^{NC})^2 + (W_C^C - W_C^{NC})^2 + (W_D^C - W_D^{NC})^2}{\text{number of the miRNA's neighbors}}$$
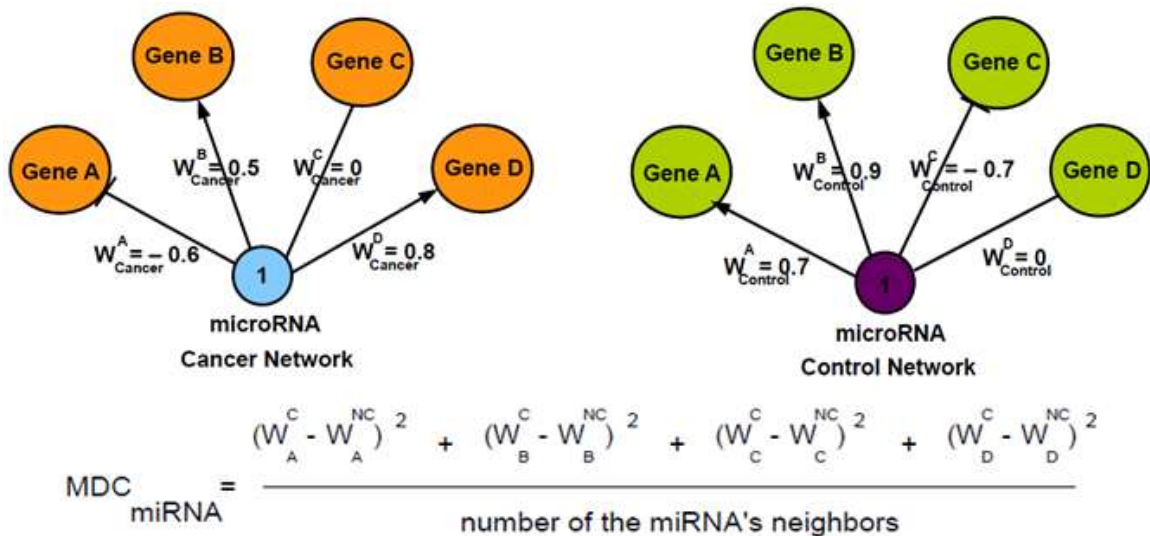
Figure 40. The differential connectivity of a miRNA is computed as the total squared difference between the edge weights of the two networks, scaled by the number of edges.

Using a 500-permutation test we identified 159 DC miRNAs in COPD and 595 in ILD (FDR<0.1). Therefore, we found significant differences between case and control miRNA-mRNA regulatory networks in both COPD and ILD. Figure 41 illustrates the significant difference between the real MDC score and the distribution of the random MDC scores obtained by 500 permutations, for miR-30a-5p (FDR=0.002), the top DC miRNA in COPD. This miRNA has been previously associated with COPD (Steiling, Lenburg, and Spira 2009; Stephanie A Christenson et al. 2013).
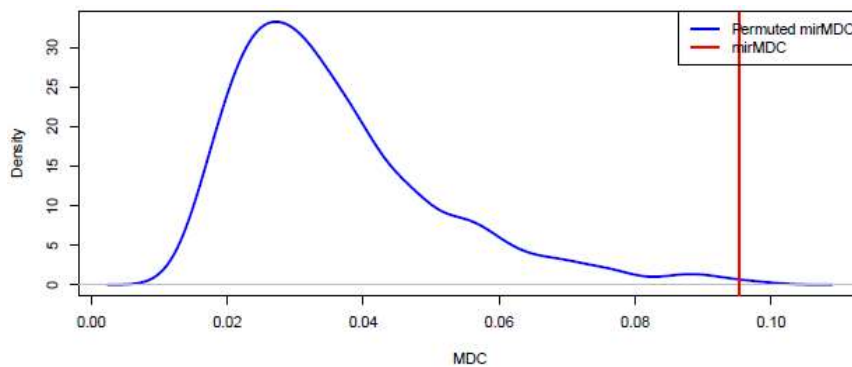
Figure 41. The permutation test assigns a p-value to each miRNA, by counting how many times the random MDC score is greater than the real MDC score. This example shows the real MDC score vs. the distribution of 500 random permutations for miR-30a-5p (FDR=0.002).

Interestingly, members of miR-449/34 family were also found to be among the top ranked DC miRNAs (Table 7).

	miRNA	DC FDR
	miR-449c-5p	0.004
	miR-449b-5p	0.01
	miR-34c-5p	0.05
ILD	miR-449a	0.07
	miR-34b-3p	0.09
	miR-34b-5p	0.1
	miR-34c-3p	0.13
	miR-34c-5p	0.08
COPD	miR-449c-5p	0.14
	miR-449b-5p	0.18
	miR-34b-3p	0.19

Table 7. Differentially connected members of miR-449/34 family.

In addition, the scale-free property of the disease-specific miRNA-mRNA regulatory networks was also evaluated. All three networks are scale-free (Figure 42), however the correlation coefficient of the log-log plots increases when SNP data is incorporated as previously described in subsection 4.2.1 (Figure 29).
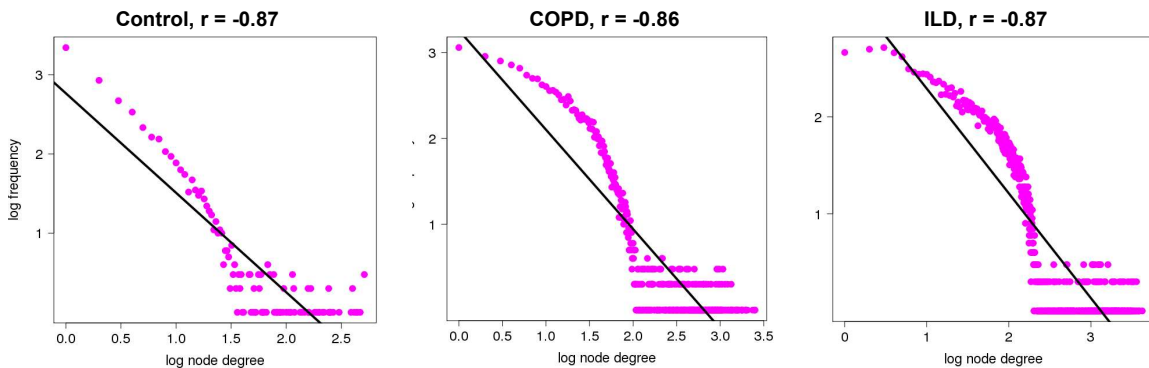
Figure 42. The miRNA-mRNA regulatory networks follow a power law. The negative correlation between the frequency of node degree and the node degree indicates that the networks are *scale-free*.

4.3 Methods

4.3.1 High-throughput sequencing of small RNA

45 samples were prepared with Small RNA Sample Prep Kit v1.5 (Illumina) and sequenced on the Genome Analyzer IIx (Illumina) according to the manufacturer's protocol. Multiplexed small RNA sequencing was conducted on the Illumina HiSeq 2000 for 319 lung tissue samples. Briefly, one microgram of total RNA from each sample was used for library preparation with a TruSeq Small RNA Sample Prep Kit (Illumina).

RNA adapters were ligated to 3' and 5' end of the RNA molecule and the adapter-ligated RNA was reverse transcribed into single-stranded cDNA. The RNA 3' adapter was specifically designed to target miRNAs 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 PCR amplified using a common primer and a primer containing one of 10 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. Ten individual PCR-enriched cDNA libraries with unique indices in equal amount were pooled and gel purified together. A 0.5% PhiX spike-in was also added in all lanes for quality control.

Each library was hybridized to one lane of the 8-lane single-read flowcell on a cBot Cluster Generation System (Illumina) using TruSeq Single-Read Cluster Kit (Illumina). The clustered flowcell was loaded onto HiSeq 2000

sequencer for a multiplexed sequencing run that consists of a standard 36-cycle sequencing read with the addition of a 7-cycle index read.

4.3.2 miRNA alignment and quality control

To estimate miRNA expression we used a small RNA sequencing pipeline previously described (Campbell et al. 2015). Similarly to the procedure described in 2.3.3, the 3′ adapter sequence was trimmed using the FASTX toolkit. Reads were aligned to hg19 using Bowtie v0.12.7 (Langmead et al. 2009).

miRNA expression was quantified by the number of reads aligned to mature miRNA loci (miRBase v20) using Bedtools v2.9.0 (Griffiths-Jones 2004; Quinlan and Hall 2010).

miRNA counts within each sample were RPM normalized, as previously described in section 2.3.3.

The batch effects of the two sequencing protocols were removed by Combat (Johnson, Li, and Rabinovic 2007) and 13 outliers were removed by PCA; 351 patients were included in the downstream analysis.

4.3.3 Quality control of the SNP data

Using *flashpca* (Abraham et al. 2014) principal components of the SNP data were computed, for those patients with overlapping miRNA and mRNA data. As expected, race is clearly separated into two groups, corresponding to African-American and Caucasian (Figure 43). This observation confirms that the SNP data is sound.

83

In addition, we can observe that the population structure may be a significant covariate of the eQTL analysis. However, this can be corrected by including the principal components into the model.
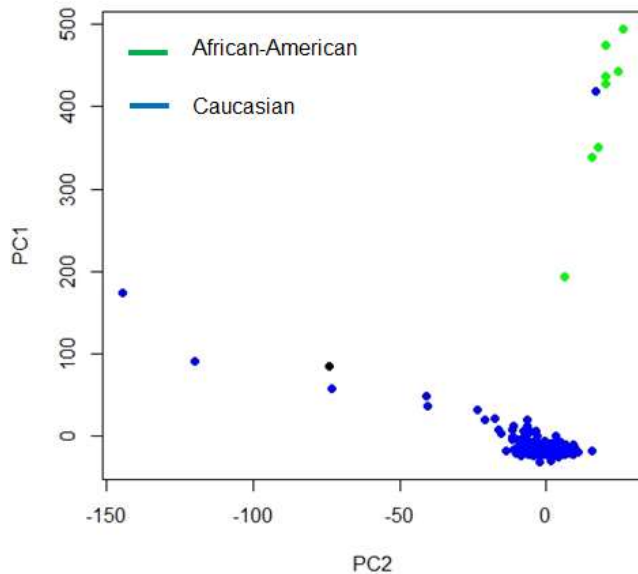
Figure 43. PCA of the SNP data shows the separation of the African-American and Caucasian groups.

4.3.4 eQTL analysis

eQTLs (expression quantitative trait loci) are regions of the genome containing DNA sequence variants that cause expression changes of one or more transcripts. eQTL interactions can be either CIS or TRANS, based on the proximity between the SNP and the transcript (Figure 44).

We utilized the subset of 262 lung tissue samples with miRNA expression profiled by sequencing, as well as publicly available Agilent gene expression array and Affymetrix SNP chip.

Genes and miRNAs associated with a SNP were identified by ANOVA (p<0.0005), while correcting for age, gender, smoking status, and population structure (the first three principal components). We considered both CIS and TRANS interactions, where CIS was defined as <1MB.
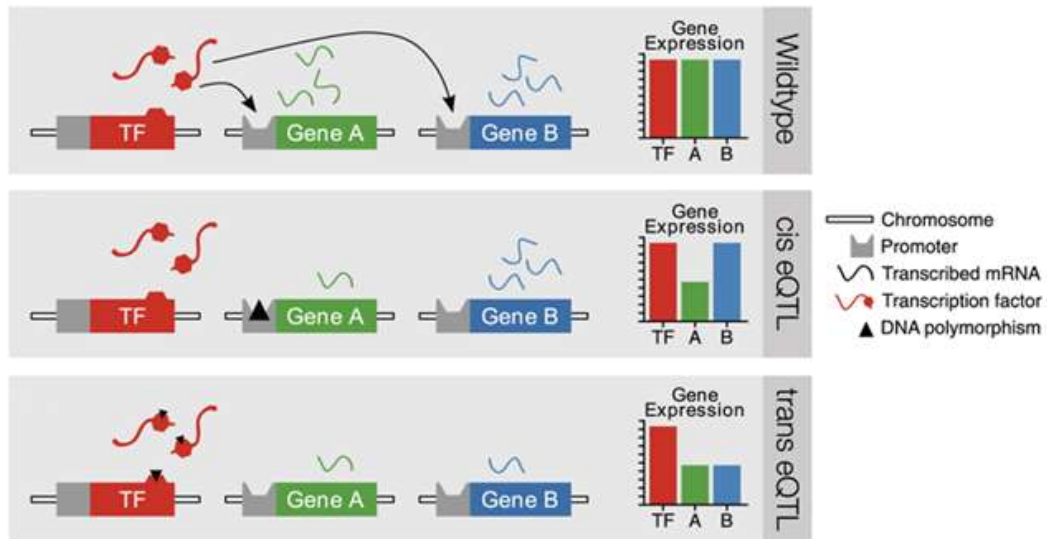
Figure 44. This figure illustrates the molecular interaction of CIS and TRANS SNPs with an RNA transcript. In this example, the CIS SNP affects the promoter of gene A, while the TRANS SNP affects a transcription factor located upstream gene A. The figure was imported from (Wolen and Miles 2006),

http://pubs.niaaa.nih.gov/publications/arcr343/306-317.htm

4.3.5 Building causal disease specific networks using SNP, microRNA and mRNA data

After we identified all genes and miRNAs associated with a SNP, as described in section 4.3.4, we built integrative networks within the COPD, ILD, and control patients using the causality inference test (CIT) (Millstein et al. 2009).

This test is a previously established method for predicting SNP-miRNA-mRNA triplets where the SNP is regulating the expression of the miRNA and the miRNA is regulating the expression of the gene (Millstein et al. 2009).

CIT assesses the hypothesis that a potential mediator between an initial randomized variable and an outcome variable is causal for that outcome. Causal and reactive models are defined as series of conditions of associations between the three variables, corresponding to SNP, microRNA and mRNA nodes. The significance of the test is computed for both the causal and reactive models. If the causal p-value is lower than 0.05 and the reactive higher than 0.05 then the call is considered causal. If both p-values are greater than 0.05 then the call is independent, and if both p-values are lower than 0.05, then the causality cannot be inferred.

We selected those SNP-miRNA-mRNA triplets where the SNP-mRNA relationship is defined by a miRNA mediator, filtering out independent relationships and those triplets where the SNP is not associated with the miRNA. The number of connections at each step of the network construction are shown in Figure 45.

Next, we compared the disease networks with the control network and evaluated those miRNAs that were differentially connected with their targets between the two states. For each microRNA in a CIT triplet, we counted how many genes were connected in each state. To assign significance, we applied a Fishers's exact test the ratio of connected genes between disease and control.

Number of patients		
Control	COPD	ILD
38	111	113

Number of miRNA-gene pairs		
Control	COPD	ILD
17024	254708	701188

Number of miRNA-gene pairs associated with a common SNP		
Control	COPD	ILD
5870	148837	107191

Causality test triplets		
Control	COPD	ILD
722	21959	15743
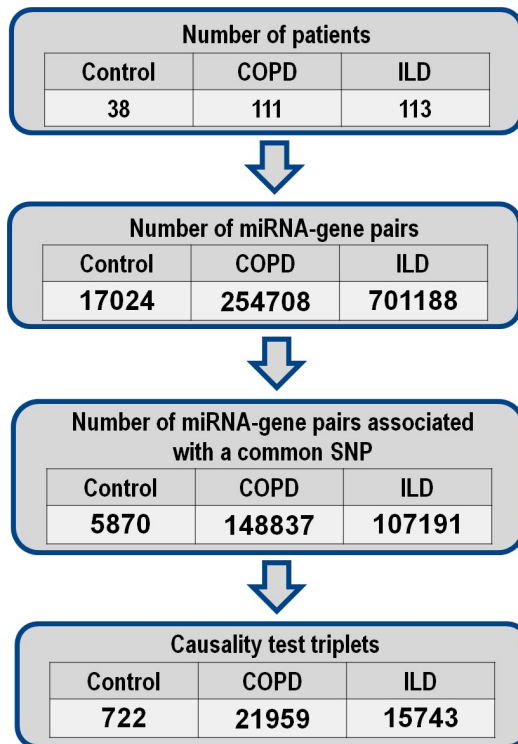
Figure 45. Number of significant interactions at each step of network construction in COPD, ILD and control groups.

4.3.6 Validation of the gene modules by gene enrichment

Using two independent datasets, we validated the miR-449/34 gene module by gene enrichment using GSVA (Hänzelmann, Castelo, and Guinney 2013) and GSEA (Subramanian et al. 2005).

First we performed gene enrichment of the gene module in a publically available dataset that provides gene expression measurements over time when airway basal cells are differentiated at an air-liquid interface (ALI) (Ross et al. 2007). First, the time points were correlated with the GSVA scores of the gene set corresponding to miR-449/34 gene module. The results were confirmed by

GSEA, where genes were ranked from those that increased in expression with time of differentiation to those that decreased in expression with time of differentiation in the ALI data, using a linear mixed-effects model (*lme()* R function).

The second validation dataset was generated by our laboratory, providing 60 airway gene expression samples profiled from 8 patients with emphysema (Campbell et al. 2012). We performed a GSVA analysis by correlating the airway wall thickness measure with the GSVA scores of the gene set that corresponds to the miR-449/34 gene module. These results were also confirmed by GSEA, where the genes were ranked from those that increased in expression with thicker airway walls to those that decreased in expression with thicker airway walls, using a linear mixed-effects model (*lme()* R function).

4.4 Discussion

This work presents novel insights about the complex mechanisms of lung pathogenesis. Using the causality inference test, we identified potential miRNA drivers of COPD and ILD. We show that there exists a significant number of different miRNA-mRNA interactions between disease and normal states. In addition, the networks obtained by this methodology are biologically meaningful, with a significant scale-free profile.

Among the top differentially connected miRNAs in COPD and ILD we found miR-449/34 family. Members of this family have been previously

associated with inflammatory lung diseases. Previous studies have shown that miR-449/34 family regulates mucociliary differentiation by directly targeting the NOTCH pathway (Marcet et al. 2011; Lizé, Klimke, and Dobbelstein 2011; Bae et al. 2012; Liu et al. 2015; Chevalier et al. 2015). In addition, we validated the association of these miRNAs with gene modules implicated in the airway epithelial cell differentiation. Besides miR449/34 family, we also found miR-4423 to be differentially connected in COPD. Expression of this miRNA has been previously associated with airway differentiation in smokers with lung cancer (Perdomo et al. 2013).

We generated a novel small-RNA sequencing dataset and highlighted the most dysregulated miRNAs in COPD and ILD by an integrative network approach. This work is a step forward in understanding the complex mechanisms of lung pathogenesis and developing new therapeutic strategies.

CHAPTER FIVE

Conclusions and future directions

This work provides new data and methods for biomarker discovery. We demonstrate for the first time the presence of a microRNA expression field in the bronchial epithelium of patients with lung cancer. While microRNA expression changes have already been associated with human cancers, we show for the first time that these alterations are measurable from bronchial epithelial tissue.

A bronchial biomarker for lung cancer detection has been developed by integrating microRNA and mRNA expression. By incorporating microRNA data, the proposed biomarker improves the performance of an existing bronchial mRNA predictor in an independent test set. This study is a proof of concept showing that bronchial microRNAs can be used to predict the presence of lung cancer. Building upon an existing clinical test, this work has important clinical implications. In future work we propose to profile nasal microRNAs and integrate them with nasal mRNAs to develop a robust and less invasive biomarker for lung cancer detection.

In addition, miR-146a has been characterized in many cancer tissues and its role as a tumor suppressor has been previously established (Labbaye and Testa 2012). Knockdown experiments in human breast cancer cells have shown that BRCA1/EGFR interaction is regulated by miR-146a, miR-146a expression being positively correlated with the expression of BRCA1 tumor suppressor. We

plan to further investigate the role of this miRNA in bronchial tissue and lung cancer development.

The second part of this work proposes a new graphical tool for binary classification problems, with application in biomarker discovery. Clinical data is limited, and most of the times the number of samples is much smaller than the number of molecular features, making it difficult to train different prediction algorithms. The results of different classifiers may differ based on the number of samples, the number of features or the class prevalence of a dataset. The goal of this software tool is to standardize the biomarker discovery process, when multiple algorithms are tested in cross-validation on the same dataset.

A web-based user-interface has been developed to facilitate the biomarker selection process and sort through a thousand potential biomarkers. This interface is user-friendly and guides the user through the entire process of biomarker selection and interpretation. This software may serve as a useful resource for the translational bioinformatics research community. Future work includes incorporating the biomarker pipeline and the GUI into a more accessible web-based system that can run on a cloud computing platform. In addition, other methods will be added to the existing collection of algorithms. Ensemble methods will be tested and compared with individual methods.

Finally, COPD and ILD disease-specific networks were analyzed. Interestingly, both the microRNA-mRNA regulatory networks and those that

incorporate genetic data (SNP) with microRNA and mRNA expression were significantly differentially connected between disease and normal states.

Using a causality inference test to infer SNP-microRNA-mRNA causal relationships, potential drivers of COPD and ILD, such as miR-449/34 family, were identified. These interesting molecular associations will further be investigated and validated by in-vitro experiments. This study is a step forward understanding the complex molecular mechanisms underlying chronic inflammatory lung diseases.

This thesis addresses three major health problems, such as lung cancer, COPD and ILD. In addition, it proposes new integrative approaches for biomarker discovery and provides new insights into the complex molecular mechanisms underlying smoking related lung diseases.

BIBLIOGRAPHY

Abraham, Gad, Michael Inouye, AL Price, NJ Patterson, RM Plenge, ME
Weinblatt, N Patterson, et al. 2014. "Fast Principal Component Analysis of
Large-Scale Genome-Wide Data." Edited by Yu Zhang. *PLoS ONE* 9 (4).
Public Library of Science: e93766. doi:10.1371/journal.pone.0093766.

Aliferis, Constantin F., Alexander Statnikov, Ioannis Tsamardinos, Subramani
Mani, and Xenofon D. Koutsoukos. 2010. "Local Causal and Markov Blanket
Induction for Causal Discovery and Feature Selection for Classification Part
I: Algorithms and Empirical Evaluation." *The Journal of Machine Learning
Research* 11. JMLR.org: 171–234.

Bae, Yangjin, Tao Yang, Huan-Chang Zeng, Philippe M Campeau, Yuqing Chen,
Terry Bertin, Brian C Dawson, Elda Munivez, Jianning Tao, and Brendan H
Lee. 2012. "miRNA-34c Regulates Notch Signaling during Bone
Development." *Human Molecular Genetics* 21 (13): 2991–3000.
doi:10.1093/hmg/dds129.

Baker, Monya. 2010. "MicroRNA Profiling: Separating Signal from Noise." *Nature
Methods* 7 (9). Nature Research: 687–92. doi:10.1038/nmeth0910-687.

Barabasi, A L, and E Bonabeau. 2003. "Scale-Free Networks." *Scientific American*. http://www.ncbi.nlm.nih.gov/pubmed/12701331.

Barabási, Albert-László, and Zoltán N. Oltvai. 2004. "Network Biology: Understanding the Cell's Functional Organization." *Nature Reviews. Genetics* 5 (2). Nature Publishing Group: 101–13. doi:10.1038/nrg1272.

Bartel, David P. 2004. "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function." *Cell* 116 (2): 281–97. doi:10.1016/S0092-8674(04)00045-5.

Beane, Jennifer, Jessica Vick, Frank Schembri, Christina Anderlind, Adam Gower, Joshua Campbell, Lingqi Luo, et al. 2011. "Characterizing the Impact of Smoking and Lung Cancer on the Airway Transcriptome Using RNA-Seq." *Cancer Prevention Research* 4 (6): 803–17. doi:10.1158/1940-6207.CAPR-11-0212.

Benjamini, Y, and Y Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society Series B 57*, 289–300.

Bertucci, François, Sébastien Salas, Séverine Eysteries, Valéry Nasser, Pascal Finetti, Christophe Ginestier, Emmanuelle Charafe-Jauffret, et al. 2004.

"Gene Expression Profiling of Colon Cancer by DNA Microarrays and Correlation with Histoclinical Parameters." *Oncogene* 23 (7): 1377–91. doi:10.1038/sj.onc.1207262.

Bhaumik, Dipa, Gary K Scott, Shiruyeh Schokrpur, Christopher K Patil, Arturo V Orjalo, Francis Rodier, Gordon J Lithgow, and Judith Campisi. 2009. "MicroRNAs miR-146a/b Negatively Modulate the Senescence-Associated Inflammatory Mediators IL-6 and IL-8." *Aging* 1 (4). Impact Journals, LLC: 402–11. doi:10.18632/aging.100042.

Bjoraker, J. A., J. H. Ryu, M. K. Edwin, J. L. Myers, H. D. Tazelaar, D. R. Schroeder, and K. P. Offord. 1998. "Prognostic Significance of Histopathologic Subsets in Idiopathic Pulmonary Fibrosis." *American Journal of Respiratory and Critical Care Medicine* 157 (1): 199–203. doi:10.1164/ajrccm.157.1.9704130.

Brody, Jerome S. 2012. "Transcriptome Alterations Induced by Cigarette Smoke." *International Journal of Cancer. Journal International Du Cancer* 131 (12): 2754–62. doi:10.1002/ijc.27829.

Campbell, Joshua D, Gang Liu, Lingqi Luo, Ji Xiao, Joseph Gerrein, Brenda Juan-Guardela, John Tedrow, et al. 2015. "Assessment of microRNA

Differential Expression and Detection in Multiplexed Small RNA Sequencing Data." *RNA* 21 (2). Cold Spring Harbor Laboratory Press: 164–171. doi:10.1261/rna.046060.114.

Campbell, Joshua D, John E McDonough, Julie E Zeskind, Tillie L Hackett, Dmitri V Pechkovsky, Corry-Anke Brandsma, John V Masaru Suzuki, et al. 2012. "A Gene Expression Signature of Emphysema-Related Lung Destruction and Its Reversal by the Tripeptide GHK." *Genome Medicine* 4 (67). doi:10.1186/gm367.

"Cancer of the Lung and Bronchus - SEER Stat Fact Sheets." 2016. Accessed December 18. https://seer.cancer.gov/statfacts/html/lungb.html.

Carrington, C. B., E. A. Gaensler, R. E. Coutu, M. X. FitzGerald, and R. G. Gupta. 1978. "Natural History and Treated Course of Usual and Desquamative Interstitial Pneumonia." *The New England Journal of Medicine* 298 (15): 801–9. doi:10.1056/NEJM197804132981501.

Castro, Mauro A A, Ines de Santiago, Thomas M Campbell, Courtney Vaughn, Theresa E Hickey, Edith Ross, Wayne D Tilley, Florian Markowetz, Bruce A J Ponder, and Kerstin B Meyer. 2016. "Regulators of Genetic Risk of Breast Cancer Identified by Integrative Network Analysis." *Nature Genetics* 48 (1):

12–21. doi:10.1038/ng.3458.

Chambers, JM. 1992. *Linear Models*. Edited by JM Chambers and T Hastie. Pacific Grove: Wadsworth & Brooks/Cole.

Chen, Gang, Ijeoma Adaku Umelo, Shasha Lv, Erik Teugels, Karel Fostier, Peter Kronenberger, Alex Dewaele, Jan Sadones, Caroline Geers, and Jacques De Grève. 2013. "miR-146a Inhibits Cell Growth, Cell Migration and Induces Apoptosis in Non-Small Cell Lung Cancer Cells." *PLoS One* 8 (3): e60317. doi:10.1371/journal.pone.0060317.

Chevalier, Benoît, Anna Adamiok, Olivier Mercey, Diego R Revinski, Laure-Emmanuelle Zaragosi, Andrea Pasini, Laurent Kodjabachian, Pascal Barbry, and Brice Marcet. 2015. "miR-34/449 Control Apical Actin Network Formation during Multiciliogenesis through Small GTPase Pathways." *Nature Communications* 6. Nature Publishing Group: 8386. doi:10.1038/ncomms9386.

Collisson, Eric A., Joshua D. Campbell, Angela N. Brooks, Alice H. Berger, William Lee, Juliann Chmielecki, David G. Beer, et al. 2014. "Comprehensive Molecular Profiling of Lung Adenocarcinoma." *Nature* 511 (7511). Nature Research: 543–50. doi:10.1038/nature13385.

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. 1988.
"Comparing the Areas under Two or More Correlated Receiver Operating
Characteristic Curves: A Nonparametric Approach." *Biometrics* 44 (3): 837.
doi:10.2307/2531595.

Dondelinger, Frank, Dirk Husmeier, and Sophie Lèbre. 2012. "Dynamic Bayesian
Networks in Molecular Plant Science: Inferring Gene Regulatory Networks
from Multiple Gene Expression Time Series." *Euphytica* 183 (3). Springer
Netherlands: 361–77. doi:10.1007/s10681-011-0538-3.

Emmert-Streib, Frank, Ricardo de Matos Simoes, Paul Mullan, Benjamin Haibe-
Kains, and Matthias Dehmer. 2014. "The Gene Regulatory Network for
Breast Cancer: Integrated Regulatory Landscape of Cancer Hallmarks."
Frontiers in Genetics 5. Frontiers: 15. doi:10.3389/fgene.2014.00015.

Etheridge, Alton, Inyoul Lee, Leroy Hood, David Galas, and Kai Wang. 2011.
"Extracellular microRNA: A New Source of Biomarkers." *Mutation Research*
717 (1–2): 85–90. doi:10.1016/j.mrfmmm.2011.03.004.

Farazi, Thalia A, Jessica I Spitzer, Pavel Morozov, and Thomas Tuschl. 2011.
"miRNAs in Human Cancer." *Journal of Pathology* 223 (2). NIH Public

Access: 102–15. doi:10.1002/path.2806.

Fededa, Juan Pablo, Christopher Esk, Beata Mierzwa, Rugile Stanyte, Shuiqiao
Yuan, Huili Zheng, Klaus Ebnet, Wei Yan, Juergen A Knoblich, and Daniel
W Gerlich. 2016. "MicroRNA□34/449 Controls Mitotic Spindle Orientation
during Mammalian Cortex Development." *EMBO Journal* 35 (22): 2386–
2398. doi:10.15252/embj.201694056.

Franklin, W A, A F Gazdar, J Haney, I I Wistuba, F G La Rosa, T Kennedy, D M
Ritchey, and Y E Miller. 1997. "Widely Dispersed p53 Mutation in
Respiratory Epithelium. A Novel Mechanism for Field Carcinogenesis." *The
Journal of Clinical Investigation* 100 (8): 2133–2137. doi:10.1172/JCI119748.

Fujimoto, Junya, Humam Kadara, Melinda M Garcia, Mohamed Kabbout,
Carmen Behrens, Diane D Liu, J Jack Lee, et al. 2012. "G-Protein Coupled
Receptor Family C, Group 5, Member A (GPRC5A) Expression Is
Decreased in the Adjacent Field and Normal Bronchial Epithelia of Patients
with Chronic Obstructive Pulmonary Disease and Non-Small-Cell Lung
Cancer." *Journal of Thoracic Oncology* 7 (12): 1747–1754.
doi:10.1097/JTO.0b013e31826bb1ff.

Gardin, A, and J White. 2011. "The Sanger Mouse Genetics Programme: High

Throughput Characterisation of Knockout Mice." *Acta Ophthalmologica* 89
(s248). Blackwell Publishing Ltd: 0–0. doi:10.1111/j.1755-3768.2011.4451.x.

Ginsberg, Michelle S., Ravinder K. Grewal, and Robert T. Heelan. 2007. "Lung
Cancer." *Radiologic Clinics of North America*, Update on Radiologic
Evaluation of Common Malignancies, 45 (1): 21–43.
doi:10.1016/j.rcl.2006.10.004.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov,
H. Coller, et al. 1999. "Molecular Classification of Cancer: Class Discovery
and Class Prediction by Gene Expression Monitoring." *Science* 286 (5439).

Govindan, Ramaswamy, Li Ding, Malachi Griffith, Janakiraman Subramanian,
Nathan D. Dees, Krishna L. Kanchi, Christopher A. Maher, et al. 2012.
"Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never
Smokers." *Cell* 150 (6): 1121–1134. doi:10.1016/j.cell.2012.08.024.

Greenhalgh, Janette, Adrian Bagust, Angela Boland, Kerry Dwan, Sophie Beale,
Juliet Hockenhull, Christine Proudlove, et al. 2015. "Erlotinib and Gefitinib for
Treating Non-Small Cell Lung Cancer That Has Progressed Following Prior
Chemotherapy (Review of NICE Technology Appraisals 162 and 175): A
Systematic Review and Economic Evaluation." *Health Technology*

Assessment 19 (47): 1–134. doi:10.3310/hta19470.

Gribbin, J, R B Hubbard, I Le Jeune, C J P Smith, J West, and L J Tata. 2006. "Incidence and Mortality of Idiopathic Pulmonary Fibrosis and Sarcoidosis in the UK." *Thorax* 61 (11): 980–985. doi:10.1136/thx.2006.062836.

Griffiths-Jones, Sam. 2004. "The microRNA Registry." *Nucleic Acids Research* 32 (Database issue): D109-11. doi:10.1093/nar/gkh023.

Guo, Mingzhou, Michael G House, Craig Hooker, Yu Han, Elizabeth Heath, Edward Gabrielson, Stephen C Yang, Stephen B Baylin, James G Herman, and Malcolm V Brock. 2004. "Promoter Hypermethylation of Resected Bronchial Margins: A Field Defect of Changes?" *Clinical Cancer Research* 10 (15): 5131–5136. doi:10.1158/1078-0432.CCR-03-0763.

Gustafson, Adam M, Raffaella Soldi, Christina Anderlind, Mary Beth Scholand, Jun Qian, Xiaohui Zhang, Kendal Cooper, et al. 2010. "Airway PI3K Pathway Activation Is an Early and Reversible Event in Lung Cancer Development." *Science Translational Medicine* 2 (26): 26ra25. doi:10.1126/scitranslmed.3000251.

Hammerman, Peter S., Michael S. Lawrence, Douglas Voet, Rui Jing, Kristian

Cibulskis, Andrey Sivachenko, Petar Stojanov, et al. 2012. "Comprehensive Genomic Characterization of Squamous Cell Lung Cancers." *Nature* 489 (7417). Nature Research: 519–525. doi:10.1038/nature11404.

Hänzelmann, Sonja, Robert Castelo, and Justin Guinney. 2013. "GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data." *BMC Bioinformatics* 14. BioMed Central: 7. doi:10.1186/1471-2105-14-7.

He, Lin, and Gregory J. Hannon. 2004. "MicroRNAs: Small RNAs with a Big Role in Gene Regulation." *Nature Reviews. Genetics* 5 (7). Nature Publishing Group: 522–531. doi:10.1038/nrg1379.

Hofree, M, J P Shen, H Carter, A Gross, and T Ideker. 2013. "Network-Based Stratification of Tumor Mutations." *Nature Methods*. doi:10.1038/nmeth.2651.

Huang, Da Wei, Brad T Sherman, and Richard A Lempicki. 2009. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." *Nature Protocols* 4 (1): 44–57. doi:10.1038/nprot.2008.211.

Johnson, W Evan, Cheng Li, and Ariel Rabinovic. 2007. "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods." *Biostatistics*

8 (1): 118–127. doi:10.1093/biostatistics/kxj037.

Kent, W James, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H
Pringle, Alan M Zahler, and David Haussler. 2002. "The Human Genome
Browser at UCSC." *Genome Research* 12 (6). Cold Spring Harbor
Laboratory Press: 996–1006. doi:10.1101/gr.229102. Article published
online before print in May 2002.

Kim, E S, R S Herbst, Wistuba II, J J Lee, G R Blumenschein Jr., A Tsao, D J
Stewart, et al. 2011. "The BATTLE Trial: Personalizing Therapy for Lung
Cancer." *Cancer Discovery*. doi:10.1158/2159-8274.CD-10-0010.

Kozomara, Ana, and Sam Griffiths-Jones. 2011. "miRBase: Integrating
microRNA Annotation and Deep-Sequencing Data." *Nucleic Acids Research*
39 (Database issue). Oxford University Press: D152-157.
doi:10.1093/nar/gkq1027.

Kumaraswamy, E, K L Wendt, L A Augustine, S R Stecklein, E C Sibala, D Li, S
Gunewardena, and R A Jensen. 2015. "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 34 (33): 4333–4346. doi:10.1038/onc.2014.363.

Labbaye, Catherine, and Ugo Testa. 2012. "The Emerging Role of MIR-146A in the Control of Hematopoiesis, Immune Function and Cancer." *Journal of Hematology & Oncology* 5: 13. doi:10.1186/1756-8722-5-13.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3): R25. doi:10.1186/gb-2009-10-3-r25.

Leiserson, M D, F Vandin, H T Wu, J R Dobson, J V Eldridge, J L Thomas, A Papoutsaki, et al. 2015. "Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes." *Nature Genetics*. doi:10.1038/ng.3168.

Leslie, R., C. J. O'Donnell, and A. D. Johnson. 2014. "GRASP: Analysis of Genotype-Phenotype Results from 1390 Genome-Wide Association Studies and Corresponding Open Access Database." *Bioinformatics* 30 (12): i185–194. doi:10.1093/bioinformatics/btu273.

Lewis, Benjamin P, Christopher B Burge, and David P Bartel. 2005. "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of

Human Genes Are microRNA Targets." *Cell* 120 (1): 15–20. doi:10.1016/j.cell.2004.12.035.

Li, Guo, Yong Liu, Zhongwu Su, Shuling Ren, Gangcai Zhu, Yongquan Tian, and Yuanzheng Qiu. 2013. "MicroRNA-324-3p Regulates Nasopharyngeal Carcinoma Radioresistance by Directly Targeting WNT2B." *European Journal of Cancer* 49 (11): 2596–2607. doi:10.1016/j.ejca.2013.03.001.

Liu, Xiang-Dong, Lian-Yun Zhang, Tie-Chui Zhu, Rui-Fang Zhang, Shu-Long Wang, and Yan Bao. 2015. "Overexpression of miR-34c Inhibits High Glucose-Induced Apoptosis in Podocytes by Targeting Notch Signaling Pathways." *International Journal of Clinical and Experimental Pathology* 8 (5): 4525–4534. http://www.ncbi.nlm.nih.gov/pubmed/26191142.

Lizé, Muriel, Alexander Klimke, and Matthias Dobbelstein. 2011. "MicroRNA-449 in Cell Fate Determination." *Cell Cycle*. doi:10.4161/cc.10.17.17181.

Lu, Jun, Gad Getz, Eric A. Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, et al. 2005. "MicroRNA Expression Profiles Classify Human Cancers." *Nature* 435 (7043). Nature Publishing Group: 834–838. doi:10.1038/nature03702.

"Lung Cancer - Small Cell: MedlinePlus Medical Encyclopedia." 2016. Accessed December 10. https://medlineplus.gov/ency/article/000122.htm.

MAQC Consortium, MAQC, Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, et al. 2006. "The MicroArray Quality Control (MAQC) Project Shows Inter- and Intraplatform Reproducibility of Gene Expression Measurements." *Nature Biotechnology* 24 (9). NIH Public Access: 1151–1161. doi:10.1038/nbt1239.

Marcet, Brice, Benoît Chevalier, Guillaume Luxardi, Christelle Coraux, Laure-Emmanuelle Zaragosi, Marie Cibois, Karine Robbe-Sermesant, et al. 2011. "Control of Vertebrate Multiciliogenesis by miR-449 through Direct Repression of the Delta/Notch Pathway." *Nature Cell Biology* 13 (6): 693–699. doi:10.1038/ncb2241.

Maru, Girish B., Khushboo Gandhi, Asha Ramchandani, and Gaurav Kumar. 2014. "The Role of Inflammation in Skin Cancer." In *Advances in Experimental Medicine and Biology*, 816:437–469. doi:10.1007/978-3-0348-0837-8_17.

Millstein, Joshua, Bin Zhang, Jun Zhu, and Eric E Schadt. 2009. "Disentangling Molecular Relationships with a Causal Inference Test." *BMC Genetics* 10

(1). BioMed Central: 23. doi:10.1186/1471-2156-10-23.

Miyazu, Y. M. 2005. "Telomerase Expression in Noncancerous Bronchial
Epithelia Is a Possible Marker of Early Development of Lung Cancer."
Cancer Research 65 (21): 9623–9627. doi:10.1158/0008-5472.CAN-05-
0976.

Nalysnyk, Luba, Javier Cid-Ruzafa, Philip Rotella, and Dirk Esser. 2012.
"Incidence and Prevalence of Idiopathic Pulmonary Fibrosis: Review of the
Literature." *European Respiratory Review* 21 (126): 355–361.
doi:10.1183/09059180.00002512.

Nathan, Nadia, Harriet Corvol, Serge Amselem, and Annick Clement. 2015.
"Biomarkers in Interstitial Lung Diseases." *Paediatric Respiratory Reviews*
16 (4): 219–224. doi:10.1016/j.prrv.2015.05.002.

Ng, Sam, Eric A Collisson, Artem Sokolov, Theodore Goldstein, Abel Gonzalez-
Perez, Nuria Lopez-Bigas, Christopher Benz, David Haussler, and Joshua M
Stuart. 2012. "PARADIGM-SHIFT Predicts the Function of Mutations in
Multiple Cancers Using Pathway Impact Analysis." *Bioinformatics* 28 (18).
Oxford University Press: i640–646. doi:10.1093/bioinformatics/bts402.

Nian, Weiqi, Xujun Ao, Yongzhong Wu, Yi Huang, Jianghe Shao, Yiming Wang,
Zhengtang Chen, Fanglin Chen, and Donglin Wang. 2013. "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." *Oncology Letters* 6 (2): 359–366.
doi:10.3892/ol.2013.1375.

Noth, Imre, Yingze Zhang, Shwu-Fan Ma, Carlos Flores, Mathew Barber, Yong
Huang, Steven M Broderick, et al. 2013. "Genetic Variants Associated with
Idiopathic Pulmonary Fibrosis Susceptibility and Mortality: A Genome-Wide
Association Study." *The Lancet. Respiratory Medicine* 1 (4). NIH Public
Access: 309–317. doi:10.1016/S2213-2600(13)70045-6.

Novaes, Fabiola Trocoli, Daniele Cristina Cataneo, Ruiz Junior, Raul Lopes, Júlio
Defaveri, Odair Carlito Michelin, and Antonio José Maria Cataneo. 2008.
"Lung Cancer: Histology, Staging, Treatment and Survival." *Jornal Brasileiro
de Pneumologia* 34 (8): 595–600. doi:10.1590/S1806-37132008000800009.

Omranian, Nooshin, Jeanne M. O. Eloundou-Mbebi, Bernd Mueller-Roeber,
Zoran Nikoloski, J. López-Barneo, R. Pardal, P. Ortega-Sáenz, et al. 2016.
"Gene Regulatory Network Inference Using Fused LASSO on Multiple Data

Sets." *Scientific Reports* 6 (February). Nature Publishing Group: 20533.

doi:10.1038/srep20533.

Osei, Emmanuel T., Laura Florez-Sampedro, Wim Timens, Dirkje S. Postma,

Irene H. Heijink, and Corry-Anke Brandsma. 2015. "Unravelling the

Complexity of COPD by microRNAs: It's a Small World after All." *European*

Respiratory Journal 46 (3).

Perdomo, Catalina, Joshua D Campbell, Joseph Gerrein, Carmen S Tellez, Carly

B Garrison, Tonya C Walser, Eduard Drizik, et al. 2013. "MicroRNA 4423 Is

a Primate-Specific Regulator of Airway Epithelial Cell Differentiation and

Lung Carcinogenesis." *Proceedings of the National Academy of Sciences of*

the United States of America 110 (47): 18946–18951.

doi:10.1073/pnas.1220319110.

Popovici, Vlad, Weijie Chen, Brandon G Gallas, Christos Hatzis, Weiwei Shi,

Frank W Samuelson, Yuri Nikolsky, et al. 2010. "Effect of Training-Sample

Size and Classification Difficulty on the Accuracy of Genomic Predictors."

Breast Cancer Research 12 (1): R5. doi:10.1186/bcr2468.

Powell, C A, S Klares, G O'Connor, and J S Brody. 1999. "Loss of

Heterozygosity in Epithelial Cells Obtained by Bronchial Brushing: Clinical

Utility in Lung Cancer." *Clinical Cancer Research* 5 (8): 2025–2034. http://www.ncbi.nlm.nih.gov/pubmed/10473082.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *American Journal of Human Genetics* 81 (3): 559–575. doi:10.1086/519795.

Quinlan, Aaron R, and Ira M Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–842. doi:10.1093/bioinformatics/btq033.

Raghu, Ganesh, Harold R Collard, Jim J Egan, Fernando J Martinez, Juergen Behr, Kevin K Brown, Thomas V Colby, et al. 2011. "An Official ATS/ERS/JRS/ALAT Statement: Idiopathic Pulmonary Fibrosis: Evidence-Based Guidelines for Diagnosis and Management." *American Journal of Respiratory and Critical Care Medicine* 183 (6): 788–824. doi:10.1164/rccm.2009-040GL.

Raghu, Ganesh, Derek Weycker, John Edelsberg, Williamson Z Bradford, and Gerry Oster. 2006. "Incidence and Prevalence of Idiopathic Pulmonary Fibrosis." *American Journal of Respiratory and Critical Care Medicine* 174

(7): 810–816. doi:10.1164/rccm.200602-163OC.

Raherison, C., and P-O Girodet. 2009. "Epidemiology of COPD." *European Respiratory Review* 18 (114).

Rivera, M Patricia, Atul C Mehta, and Momen M Wahidi. 2013. "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* 143 (5 Suppl): e142S–165S. doi:10.1378/chest.12-2353.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves." *BMC Bioinformatics* 12 (1): 77. doi:10.1186/1471-2105-12-77.

Ross, Andrea J, Lisa A Dailey, Luisa E Brighton, and Robert B Devlin. 2007. "Transcriptional Profiling of Mucociliary Differentiation in Human Airway Epithelial Cells." *American Journal of Respiratory Cell and Molecular Biology* 37 (2): 169–185. doi:10.1165/rcmb.2006-0466OC.

Sandhu, Sukhinder, and Ramiro Garzon. 2011. "Potential Applications of MicroRNAs in Cancer Diagnosis, Prognosis, and Treatment." *Seminars in*

Oncology 38 (6): 781–787. doi:10.1053/j.seminoncol.2011.08.007.

Sass, Steffen, Sabine Dietmann, Ulrike Burk, Simone Brabletz, Dominik Lutter, Andreas Kowarsch, Klaus F Mayer, et al. 2011. "MicroRNAs Coordinately Regulate Protein Complexes." *BMC Systems Biology* 5 (1): 136. doi:10.1186/1752-0509-5-136.

Schembri, Frank, Sriram Sridhar, Catalina Perdomo, Adam M Gustafson, Xiaoling Zhang, Ayla Ergun, Jining Lu, et al. 2009. "MicroRNAs as Modulators of Smoking-Induced Gene Expression Changes in Human Airway Epithelium." *Proceedings of the National Academy of Sciences of the United States of America* 106 (7): 2319–2324. doi:10.1073/pnas.0806383106.

Siegel, Rebecca, Deepa Naishadham, and Ahmedin Jemal. 2013. "Cancer Statistics, 2013." *CA: A Cancer Journal for Clinicians* 63 (1): 11–30. doi:10.3322/caac.21166.

Silvestri, Gerard A., Anil Vachani, Duncan Whitney, Michael Elashoff, Kate Porta Smith, J. Scott Ferguson, Ed Parsons, et al. 2015. "A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer." *New England Journal of Medicine* 373 (3): 243–251. doi:10.1056/NEJMoa1504601.

Singh, Bhuvanesh, Pabbathi G Reddy, Andy Goberdhan, Christine Walsh, Su Dao, Ivan Ngai, Ting Chao Chou, et al. 2002. "p53 Regulates Cell Survival by Inhibiting PIK3CA in Squamous Cell Carcinomas." *Genes & Development* 16 (8): 984–993. doi:10.1101/gad.973602.

Smyth, Gordon K. 2004. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." *Statistical Applications in Genetics and Molecular Biology* 3 (1): 1–25. doi:10.2202/1544-6115.1027.

Spira, Avrum, Jennifer E Beane, Vishal Shah, Katrina Steiling, Gang Liu, Frank Schembri, Sean Gilman, et al. 2007. "Airway Epithelial Gene Expression in the Diagnostic Evaluation of Smokers with Suspect Lung Cancer." *Nature Medicine* 13 (3): 361–366. doi:10.1038/nm1556.

Spira, Avrum, Jennifer Beane, Vishal Shah, Gang Liu, Frank Schembri, Xuemei Yang, John Palma, and Jerome S Brody. 2004. "Effects of Cigarette Smoke on the Human Airway Epithelial Cell Transcriptome." *Proceedings of the National Academy of Sciences of the United States of America* 101 (27): 10143–10148. doi:10.1073/pnas.0401422101.

Stack, B. H., Y. F. Choo-Kang, and B. E. Heard. 1972. "The Prognosis of

    Cryptogenic Fibrosing Alveolitis." *Thorax* 27 (5): 535–542.

Steiling, Katrina, Marc E Lenburg, and Avrum Spira. 2009. "Airway Gene

    Expression in Chronic Obstructive Pulmonary Disease." *Proceedings of the*

    American Thoracic Society 6 (8): 697–700. doi:10.1513/pats.200907-076DP.

Steiling, Katrina, Maarten van den Berge, Kahkeshan Hijazi, Roberta Florido,

    Joshua Campbell, Gang Liu, Ji Xiao, et al. 2013. "A Dynamic Bronchial

    Airway Gene Expression Signature of Chronic Obstructive Pulmonary

    Disease and Lung Function Impairment." *American Journal of Respiratory*

    and Critical Care Medicine 187 (9): 933–942. doi:10.1164/rccm.201208-

    1449OC.

Stephanie A Christenson, Corry-Anke Brandsma, Joshua D Campbell, Darryl A

    Knight, Dmitri V Pechkovsky, James C Hogg, Wim Timens, Dirkje S Postma,

    Marc Lenburg, and Avrum Spira. 2013. "miR-638 Regulates Gene

    Expression Networks Associated with Emphysematous Lung Destruction."

    Genome Medicine 5 (114). doi:10.1186/gm519.

Stephens, Philip, Chris Hunter, Graham Bignell, Sarah Edkins, Helen Davies,

Jon Teague, Claire Stevens, et al. 2004. "Lung Cancer: Intragenic ERBB2 Kinase Mutations in Tumours." *Nature* 431 (7008): 525–526. doi:10.1038/431525b.

Su, Wan-Lin, Robert R Kleinhanz, and Eric E Schadt. 2011. "Characterizing the Role of miRNAs within Gene Regulatory Networks Using Integrative Genomics Techniques." *Molecular Systems Biology* 7 (May): 490. doi:10.1038/msb.2011.23.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–15550. doi:10.1073/pnas.0506580102.

Tang, Wenbo, Matthew Kowgier, Daan W Loth, María Soler Artigas, Bonnie R Joubert, Emily Hodge, Sina A Gharib, et al. 2014. "Large-Scale Genome-Wide Association Studies and Meta-Analyses of Longitudinal Change in Adult Lung Function." *PLoS One* 9 (7): e100776. doi:10.1371/journal.pone.0100776.

Team, The National Lung Screening Trial Research. 2011. "Reduced Lung-

Cancer Mortality with Low-Dose Computed Tomographic Screening." *New

England Journal of Medicine* 365 (5): 395–409.

doi:10.1056/NEJMoa1102873.

Terzić, Janoš, Sergei Grivennikov, Eliad Karin, and Michael Karin. 2010.

"Inflammation and Colon Cancer." *Gastroenterology* 138 (6): 2101–2114.e5.

doi:10.1053/j.gastro.2010.01.058.

Tukey, Melissa H, and Renda Soylemez Wiener. 2012. "Population-Based

Estimates of Transbronchial Lung Biopsy Utilization and Complications."

Respiratory Medicine 106 (11): 1559–1565. doi:10.1016/j.rmed.2012.08.008.

van 't Veer, Laura J., Hongyue Dai, Marc J. van de Vijver, Yudong D. He,

Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, et al. 2002. "Gene

Expression Profiling Predicts Clinical Outcome of Breast Cancer." *Nature*

415 (6871): 530–36. doi:10.1038/415530a.

Vandin, F, E Upfal, and B J Raphael. 2011. "Algorithms for Detecting

Significantly Mutated Pathways in Cancer." *Journal of Computational

Biology*. doi:10.1089/cmb.2010.0265.

Vaske, Charles J, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher
Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. 2010.
"Inference of Patient-Specific Pathway Activities from Multi-Dimensional
Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26 (12): i237-
245. doi:10.1093/bioinformatics/btq182.

Vignes, Matthieu, Jimmy Vandel, David Allouche, Nidal Ramadan-Alban,
Christine Cierco-Ayrolles, Thomas Schiex, Brigitte Mangin, et al. 2011.
"Gene Regulatory Network Reconstruction Using Bayesian Networks, the
Dantzig Selector, the Lasso and Their Meta-Analysis." Edited by Magnus
Rattray. *PLoS One* 6 (12): e29165. doi:10.1371/journal.pone.0029165.

Wang, Yongsheng, Gerald Schmid-Bindert, and Caicun Zhou. 2012. "Erlotinib in
the Treatment of Advanced Non-Small Cell Lung Cancer: An Update for
Clinicians." *Therapeutic Advances in Medical Oncology* 4 (1). SAGE
Publications: 19–29. doi:10.1177/1758834011427927.

Wang Memoli, Jessica S, Paul J Nietert, and Gerard A Silvestri. 2012. "Meta-
Analysis of Guided Bronchoscopy for the Evaluation of the Pulmonary
Nodule." *Chest* 142 (2): 385–393. doi:10.1378/chest.11-1764.

Whitney, Duncan H, Michael R Elashoff, Kate Porta-Smith, Adam C Gower, Anil

Vachani, J Scott Ferguson, Gerard A Silvestri, Jerome S Brody, Marc E

Lenburg, and Avrum Spira. 2015. "Derivation of a Bronchial Genomic

Classifier for Lung Cancer in a Prospective Study of Patients Undergoing

Diagnostic Bronchoscopy." *BMC Medical Genomics* 8 (1): 18.

doi:10.1186/s12920-015-0091-3.

Wistuba, I I, S Lam, C Behrens, A K Virmani, K M Fong, J LeRiche, J M Samet,

S Srivastava, J D Minna, and A F Gazdar. 1997. "Molecular Damage in the

Bronchial Epithelium of Current and Former Smokers." *Journal of the*

National Cancer Institute 89 (18): 1366–1373.

http://www.ncbi.nlm.nih.gov/pubmed/9308707.

Wolen, Aaron, and Michael Miles. 2006. "Identifying Gene Networks Underlying

the Neurobiology of Ethanol and Alcoholism." *Alcohol Research: Current*

Reviews 34 (3).

Zafari, Sachli, Christina Backes, Petra Leidinger, Eckart Meese, and Andreas

Keller. 2015. "Regulatory MicroRNA Networks: Complex Patterns of Target

Pathways for Disease-Related and Housekeeping MicroRNAs." *Genomics,*

Proteomics & Bioinformatics 13 (3): 159–168.

doi:10.1016/j.gpb.2015.02.004.

Zhai, Yuxin, Zhenping Zhong, Chyi-Ying A Chen, Zhenfang Xia, Ling Song, Michael R Blackburn, and Ann-Bin Shyu. 2008. "Coordinated Changes in mRNA Turnover, Translation, and RNA Processing Bodies in Bronchial Epithelial Cells Following Inflammatory Stimulation." *Molecular and Cellular Biology* 28 (24): 7414–7426. doi:10.1128/MCB.01237-08.

Zhang, Bin, Chris Gaiteri, Liviu-Gabriel Bodea, Zhi Wang, Joshua McElwee, Alexei A. Podtelezhnikov, Chunsheng Zhang, et al. 2013. "Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease." *Cell* 153 (3): 707–720. doi:10.1016/j.cell.2013.03.030.

Zhang, Junpeng, Thuc Duy Le, Lin Liu, Bing Liu, Jianfeng He, Gregory J Goodall, and Jiuyong Li. 2014. "Inferring Condition-Specific miRNA Activity from Matched miRNA and mRNA Expression Data." *Bioinformatics* 30 (21): 3070–3077. doi:10.1093/bioinformatics/btu489.

CURRICULUM VITAE