2015

Integrative transcriptomics in smoking related lung diseases

BOSTON UNIVERSITY

SCHOOL OF MEDICINE

Dissertation

INTEGRATIVE TRANSCRIPTOMICS IN SMOKING RELATED LUNG DISEASES

by

REBECCA KUSKO

S.B., Massachusetts Institute of Technology, 2009

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2015

Approved by

First Reader

Avrum Spira, M.D., M.Sc.
Professor of Medicine

Second Reader

Yuriy Alekseyev, Ph.D.
Research Associate Professor

ACKNOWLEDGMENTS

I would like to take this opportunity to thank everyone who inspired and supported me throughout my thesis. Firstly, I would like to thank my thesis adviser Avrum Spira for bringing me into the lab and giving me room to grow scientifically and personally. I am also sincerely grateful to my committee chair Monty Montano for mentoring me and pushing me to achieve. The rest of my thesis committee also provided an incredible amount of guidance throughout my journey to complete my PhD for which I am very grateful.

Next I would like to thank the members of the Spira-Lenburg lab for all the helpful discussions, reviews of my work, pleasant conversations, and adventures.

I am eternally grateful to my family, especially my parents who tirelessly pushed me to believe in myself and to continue to try harder. I am also thankful to my entire family, both the family that has been supportive throughout my entire life and all the new family that I gained more recently.

Lastly I greatly appreciate all my friends and Arthur who played a critical role in keeping me sane and helping me to see the big picture. My capacity to endure comes from the support of these people.

INTEGRATIVE TRANSCRIPTOMICS IN SMOKING RELATED LUNG DISEASES

REBECCA KUSKO

Boston University School of Medicine, 2015

Major Professor: Avrum Spira, M.D., M.Sc., Professor of Medicine

ABSTRACT

Chronic lung diseases including Chronic Obstructive Pulmonary Disease (COPD), Idiopathic Pulmonary Fibrosis (IPF) and lung cancer are major causes of morbidity and mortality in the United States due to high incidence and limited therapeutic options. In order to address this critical issue, I have leveraged RNA sequencing and integrative genomics to define disease-associated transcriptomic changes which could be potentially targeted to lead to new therapeutics.

We sequenced the lung transcriptome of subjects with IPF (n=19), emphysema (n=19, a subtype of COPD), or neither (n=20). The expression levels of 1770 genes differed between IPF and control lung, and 220 genes differed between emphysema and control lung ($p<0.001$). Upregulated genes in both emphysema and IPF were enriched for the p53/hypoxia pathway. These results were validated by immunohistochemistry of select p53/hypoxia proteins and by GSEA analysis of independent expression microarray experiments. To identify regulatory events, I constructed an integrative miRNA target prediction and anticorrelation miRNA-mRNA network, which highlighted several miRNA whose expression levels were the opposite of genes differentially expressed in both IPF and emphysema. MiR-96 was a highly connected hub in this network and was subsequently overexpressed in cell lines to validate several potential regulatory connections.

Building upon these successful experiments, I next sought to define gene expression changes and the miRNA-mRNA regulatory network in never smoker lung cancer. Large and small RNA was sequenced from matched lung adenocarcinoma tumor and adjacent normal lung tissue obtained from 22 subjects (8 never, 14 current and former smokers). I identified 120 genes whose expression was modified uniquely in never smoker lung tumors. Using a repository of gene-expression profiles associated with small bioactive molecules, several compounds which counter the never smoker tumor signature were identified *in silico*. Leveraging differential expression information, I again constructed an mRNA-miRNA regulatory network, and subsequently identified a potential never smoker oncomir has-mir-424 and its transcription factor target FOXP2.

In this thesis, I have identified genes, pathways and the miRNA-mRNA regulatory network that is altered in COPD, IPF, and lung adenocarcinoma among never smokers. My findings may ultimately lead to improved treatment options by identifying targetable pathways, regulators, and therapeutic drug candidates.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABCC3	ATP-Binding Cassette, Sub-Family C (CTFR/MRP), Member 3
AFA	African American
ALK	Anaplastic Lymphoma Receptor Tyrosine Kinase
ARHGAP24	Rho GTPase Activating Protein 24
ASN	Asian
BAX	BCL2-associated X protein
Beas-2B	Bronchial Epithelium + Adenovirus 12SV40 virus hybrid Version 2B
BTK	Bruton Agammaglobulinemia Tyrosine Kinase
BU	Boston University
CABIN1	Calcineurin Binding Protein 1
CAU	Caucasian
CDKN1A	Cyclin Dependent Kinase Inhibitor 1A
CMAP	Connectivity Map
COPD	Chronic Obstructive Pulmonary Disease
CT	Computed Tomography
DAVID	Database for Annotation, Visualization and Integrated Discovery
DLCO	Diffusing capacity of Lung for Carbon Dioxide
DNA	Deoxyribo Nucleic Acid
EGFR	Epidermal Growth Factor Receptor
EGFR-TK	Epidermal Growth Factor Receptor Tyrosine Kinase
EML4	Echinoderm Microtubule Associated Protein Like 4

FAM83B	Family With Sequence Similarity 83, Member B
FDA	Food and Drug Administration
FEV1	Forced Expiratory Volume after one second
FPKM	Fragments Per Kilobase mapped per Million reads
FVC	Forced Vital Capacity
GDF15	Growth Differentiation Factor 15
GEO	Gene Expression Omnibus
GOLD	Global Initiative for Chronic Obstructive Lung Disease
GPR110	G Protein Coupled Receptor 110
GSEA	Gene Set Enrichment Analysis
GSVA	Gene Set Variation Analysis
HIF1A	Hypoxia Induced Factor 1 Alpha
HIMF	Hypoxia Induced Mitogenic Factor
hMHL1	Human MutL Homolog 1
hMHL2	Human MutL Homolog 2
ILD	Interstitial Lung Disease
IPF	Idiopathic Pulmonary Fibrosis
KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog
LDB2	LIM Domain Binding Protein 2
lme	Linear Mixed Effects
lncRNA	Long NonCoding RNA
LGRC	Lung Genome Research Consortium

LTRC	Lung Tissue Research Consortium
MDM2	Mouse Double Minute 2
miRNA	microRNA
mRNA	messenger RNA
ncRNA	NonCoding RNA
NES	Normalized Enrichment Score
NFKBiB2	Nuclear Factor of Kappa Gene Enhancer in B-Cells Inhibitor, Beta
NG	Nanogram
NHLBI	National Heart, Lung, and Blood Institute
NHLF	Normal Human Lung Fibroblast
nt	Nucleotide
OTH	Other
PCA	Principle Component Analysis
PET	Positron Emission Tomography
PLXNB1	Plexin B1
QC	Quality Control
RIN	RNA Integrity Number
RNA	RiboNucleic Acid
RNAseq	RNA sequencing
RPKM	Reads Per Kilobase of Transcript Length per Million Mapped Reads
RPM	Reads Per Million Mapped Reads
SH3BP5	SH3-Domain Binding Protein 5 (BTK associated)

SLC1A1	Solute Carrier Family 1 Member 1
TCGA	The Cancer Genome Atlas
TGF-β	Transforming Growth Factor, Beta
TP53	Tumor Protein 53
UIP	Usual Interstitial Pneumonia
UTR	Untranslated Region
WNT	Wingless-Type MMTV Integration Site Family

INTRODUCTION

Smoking Associated Lung Disease Clinical Overview

Smoking associated lung diseases such as Chronic Obstructive Pulmonary Disease (COPD), Idiopathic Pulmonary Fibrosis (IPF), and lung cancer are major causes of morbidity and mortality in the United States. Currently, they cause an estimated 300,000 deaths per year (158,318 from lung cancer, 134,676 from COPD, and 40,000 from IPF)[1,2]. This is likely an underestimate as COPD and IPF are frequently underdiagnosed clinically[3]. Lung cancer is the leading cause of cancer related death in men and women, causing more mortalities than pancreas, breast, and colorectal cancer combined[4]. Adding to the burden of smoking related lung diseases, COPD is the third leading cause of mortality in the United States, after heart disease and cancer. While IPF represents a much smaller disease burden, the incidence of this disease has doubled in the past decade[5].

All three smoking related lung diseases carry a large mortality burden because lung cancer has an exceptionally low 5 year survival rate (16.6%), and a treatment is not yet clinically available which halts the underlying pathogenesis of COPD or IPF[6]. Despite significant funding for pulmonary research in past decades, these diseases are persistently major public health problems. All three diseases have differing clinical pathology, but are associated with shared environmental factors such as cigarette smoking and air pollution[7–9].

Chronic Obstructive Pulmonary Disease

COPD limits a patient's pulmonary function and is considered a progressive disease, as severity increases with time. Patients with COPD will experience symptoms such as shortness of breath, wheezing, and productive cough. More than eighty percent of COPD patients are current or former smokers, making COPD a strongly smoking associated disease. Although it has been known since the 1964 Surgeon General's report on smoking that cigarette smoke causes COPD, the molecular processes underlying this connection remain unclear[10,11].

COPD is a disease constellation contains two disease subtypes: 1. chronic bronchitis, defined as "the presence of chronic productive cough for at least 3 months in two consecutive years, after excluding other causes of chronic cough." or 2. emphysema, defined as "a condition of the lung characterized by abnormal, permanent enlargement of airspaces distal to the terminal bronchiole, accompanied by the destruction of their walls, and without obvious fibrosis.[9]" Patients can be afflicted with one, the other, or both of these subtypes.

Patients diagnosed with COPD have airflow obstruction which is not permanently reversible with bronchodilator administration[12,13]. In order to arrive at this diagnosis or to classify COPD, physicians use clinical features such as forced expiratory volume in 1 second (FEV1, measured through spirometry), forced vital capacity (FVC, measured through spirometry), FEV1/FVC ratio, and the GOLD (Global Initiative for Chronic Obstructive Lung Disease) staging process [12]. However, these classification systems have limited clinical utility, as they do not reflect future rate of airway thickening or emphysematous destruction, histopathology, recommended treatment course, or other clinical variables.

Information about the clinical pathogenesis of COPD remains limited. It has long been observed that tobacco abstinence does not reverse or cure COPD, suggesting permanent and irreversible molecular damage. Moreover it has been observed that COPD is highly smoking associated but does not occur in all smokers, suggesting a shared genetic and environmental component. Specifically, some literature evidence suggests that the emphysema subtype is related to a mis-balance between elastin and anti-elastin production, causing the alveolar destruction phenotype observed in patients[14]. Adding support to this hypothesis, it is known that pro-inflammatory elastase producing macrophages are present at higher levels in the lungs of smokers[15,16]. Additional studies, including mouse work, have since implicated a complex immune mediated inflammation[17–19]. Most recently, selective apoptosis of structural cells are under suspicion for leading to the emphysema phenotype as well[20]. Despite these recent advances, the fact remains that no drug or treatment has been discovered that can target the molecular pathogenesis of emphysema, suggesting an urgent clinical need for improved understanding of the molecular mechanisms of disease development and progression.

Idiopathic Pulmonary Fibrosis

IPF is a devastating disease characterized by an overabundance of scar tissue in the parenchyma of the lung, which leads to shortness of breath, dry cough, weight loss, finger clubbing, and fatigue. This fibrosis often occurs in a heterogeneous manner, with select portions of the lung being effected and others not. Like COPD, IPF is a progressive disease, which worsens in a patient over time.

While both COPD and IPF are smoking associated, the molecular etiology of IPF is much more poorly understood and is under active investigation. Both genetics and the environment are guilty parties for driving this disease[21–25]. Based on observations of the pathology of IPF, including alveolar remodeling and invasion of fibroblasts, abnormal wound healing has been implicated[26]. When injured, the lung normally goes through a number of stages, one of which is fibrosis. Others have suggested that in IPF, wound healing becomes "stuck" in fibrosis and never progresses to the next stage[24]. While the final phenotype is very different, IPF does have certain molecular steps in common with emphysema, such as remodeling[21,22]. It is known that the fibrosis is driven specifically by myofibroblasts, which produce an excess of extracellular matrix materials which destroys the structure and function of the lung[27]. It is also hypothesized that abnormal epithelial cell activation could lead to IPF, in addition to wound healing[28].

IPF is diagnosed using high resolution CT imaging together with a surgical lung biopsy[29]. In patients with IPF, CT scans will reveal usual interstitial pneumonia (UIP) and/or honeycombing, where large lined air spaces form. Interestingly, it has been observed that some patients have variable emphysema and IPF across the lung[30]. Science has yet to prove whether these are distinct diseases occurring together in the same patient or the same underlying disease manifesting itself with heterogeneous phenotypes throughout the lung.

Lung Cancer

Lung cancer remains the leading cause of cancer-related death in the United States and worldwide[6]. While it is possible to cure lung cancer in a low percentage of patients, this disease remains a major cause of morbidity and mortality globally because of a lack of early detection and effective treatment options. Patients with lung cancer can experience fatigue, persistent cough, increased sputum production, cachexia, and shortness of breath. Lung cancer is normally diagnosed first with a CT or PET scan to discover the presence of a suspicious nodule, followed by bronchoscopy and/or fine needle biopsy to confirm the presence or absence of cancerous cells.

Although lung cancer is often thought of as a smoker's disease, 25% of all lung cancer patients worldwide are lifelong nonsmokers[31]. Lung cancer can be split into various subtypes. Adenocarcinoma is the most prevalent (40% of lung cancer cases) and occurs in the periphery of the lung[32]. Lung Adenocarcinoma arises from mucus secreting epithelial cells that line the airways in the lung. This is the most common form of lung cancer to occur in never smokers, and also occurs in current or former smokers. Studies suggest exposures (such as secondhand smoke[33–36], indoor cooking fumes[37–39], asbestos[40–43], radon[44,45], hormones[46–50]), previous IPF diagnosis[51–53], and genetics[54,55] as factors that can cause lung cancer to occur in never smokers while smoking remains the leading factor for lung cancer in current and former smokers.

Recent studies have described genetic[56,57] and genomic[57,58] differences between never smoker and ever smoker lung tumors. From the mutational perspective, it has been observed that there is a higher p53[57] and KRAS[59–61] mutation risk in active smokers and a higher EGFR mutation and EML4-ALK fusion[62] risk in never smokers. Relatedly, it has also been observed that ever and never smokers respond differently to therapy[63,64]. Moreover, the seminal observation that EGFR mutations are more prevalent in never smoker lung tumors[56] has caused a shift in lung cancer treatment through the use of EGFR inhibitor drugs[63,64]. Specifically, gefitinib and erlotinib are more effective in patients with mutations in the tyrosine kinase domain of EGFR[65]. Since the discovery of the EML4-ALK fusion (more common in never smoker lung adenocarcinoma), it has been shown that patients with this fusion respond well to crizotinib[66,67].

While all lung adenocarcinoma demonstrate chromosomal instability, one study observed that never smoker tumors as compared to ever smoker tumors tend to have 16p gains at a higher frequency[68]. In addition, never smoker lung adenocarcinoma has unique methylation aberrations, such as hypermethylation of the promoters of hMLH1 and hMSH2[69].

Despite these recent advances, lung cancer remains a leading cause of cancer related deaths in never smokers, and the molecular drivers within the never smoker lung tumor remain unclear. The above evidence supports the hypothesis that never smokers and ever smokers experience different molecular processes and events which drive the same tumor subtype. Thus, characterizing never smoker-specific molecular alterations in lung adenocarcinoma might be leveraged to identify processes that could be targeted by new therapeutics and existing compounds that could be repurposed to treat this disease.

In summary, science has yet to prove conclusively that lung adenocarcinoma arises through distinct molecular processes in never smokers as compared to ever smokers, and that these differing mechanisms of carcinogenesis require distinct therapy. Thus, understanding the molecular processes that contribute to lung carcinogenesis specifically in lifelong nonsmokers would allow us to identify regulators that might be targeted for therapy, and existing therapeutics that might be repurposed to treat never smoker lung adenocarcinoma.

High-Throughput Transcriptomics in COPD, IPF, and Lung Cancer

Microarray Studies

To date, there have been a number of microarray studies reported in COPD, IPF, and lung cancer. At the time, these studies represented major advances in the field as microarrays allow for translating the protein coding disease transcriptome. Broadly speaking, microarray technologies have generated hypotheses about which pathways could be important drivers for COPD, IPF, and lung cancer.

In the COPD space, a handful of transcriptomic gene expression studies have described genes associated with COPD[70–75]. However, there was little overlap in the exact genes reported by these studies[76]. More recent work using enrichment-based methods such as Gene Set Enrichment Analysis (GSEA)[77] has shown that there is more overlap in underlying pathways and gene ontology between these datasets than initially perceived, such as an overlap in cell adhesion pathways[76]. Since this analysis was done, several new microarray studies have come out and implicated additional pathways, such as TGF-β[78], WNT[79], and inflammation[80].

Although microarray studies in COPD are limited, microarray studies in IPF are even more so. Also unlike COPD, publications in the IPF space have converged around a set of pathways. Two gene expression studies comparing normal lung to IPF both reported perturbations in the extracellular matrix regulatory pathway[81]. A third microarray study has been reported but focused on a single differentially expressed gene (osteoporin). The authors profiled 13 IPF samples and 11 controls and reported that the osteoporin gene was the most upregulated gene in IPF samples as compared to control[82].

In the lung cancer space, a large number of gene expression studies have been reported[83], but only one group has published array studies in the never smoker lung cancer space. Landi et al.[84] profiled mRNA expression in 20 never smokers, 26 former smokers, and 28 current smokers. In this publication, the authors identify genes differentially expressed between the ever smoker versus never smoker tumor and ever smoker vs. never smoker adjacent normal. They found that the smoker tumors tend to upregulate mitotic spindle formation genes more than the never smoker tumors. The same group later profiled miRNA expression in a similar cohort but did not find any smoking associated changes in miRNA expression, perhaps due to the low number of never smoker patients in the study[85].

RNAseq Studies

As RNAseq is a relatively new platform and only recently has become more accessible in terms of cost, very limited studies using RNAseq have been reported in IPF, and COPD. To date, no studies have been reported which sequence COPD samples. One study with 3 IPF samples and 3 control samples observed IPF associated changes in splicing of senescence and oxidative stress genes[86].

In never smoker lung adenocarcinoma, multiple publications have reported RNAseq studies that have contributed to the hypothesis that never smoker lung adenocarcinoma is a molecularly unique disease. One study compared tumor to adjacent normal tissue in never smokers, and based on mutational and transcriptomic evidence implicated cell proliferation pathways[87]. However, this study lacked ever smokers as a control so it is unclear if their result is unique to never smokers. Another study reported transcriptome and whole genome sequencing in 6 never smoker and 11 ever smokers with lung adenocarcinoma. The tumors of smokers overall had more mutations, a different frequency of point mutations, and were more likely to express the gene containing a mutation[88]. These results greatly supports the hypothesis that never smoker lung cancer is a unique molecular entity, although the transcriptomic data was only used to support genomic data and was not analyzed independently.

Computational Approaches

RNAseq and Microarrays

The technology of gene expression microarrays has enabled scientists in many fields to rapidly and inexpensively profile the expression of protein coding genes. Gene expression arrays contain "probes", which are short DNA sequences that will bind to a corresponding protein coding gene transcript after conversion to cDNA. The intensity of fluorescence after imaging corresponds to the expression level of a gene. The intensity data is typically processed via Robust Multichip Average (RMA), which performs background correction, log2 transformation, and quantile normalization prior to analysis[89]. Microarray data has the advantage of being very fast to process, and having a very small memory and storage footprint. However, microarray technology is fundamentally limited by two factors: 1. only transcriptomic events with a corresponding probe will be profiled and 2. because array data is measured by intensity, the dynamic range measurement capabilities are small.

Recently next generation sequencing technology has developed, improved, and become cost effective for transcriptomic profiling. With this technology, isolated RNA is built into libraries, which are typically sequenced using reversible terminator chemistry, which emits a color depending on which nucleotide binds to the cDNA. Since RNA sequencing (RNAseq) is not limited to a specific probe set, it has enabled researchers to characterize more exotic members of the transcriptomic zoo, such as microRNAs (miRNAs), long noncoding RNAs (lncRNAs), and pseudogenes. Originally thought to be junk[90], these noncoding RNAs are now known to be highly important in regulating gene expression in development, cancer, and a spectrum of other states of health and disease[91–94]. Out of noncoding regulatory RNAs, microRNAs (miRNAs) are the best characterized. These small RNAs, only 22nt long, bind to the 3'UTR of protein coding genes using a specific sequence[95]. Binding of the miRNA blocks transcription and/or causes the RNA sequence to be degraded, thus stopping the message[96]. LncRNAs function through a much more complicated and less characterized mechanism by modifying chromatin[97], and/or positively regulating transcription initiation[98]. In the future, "lncRNA" will likely be regarded as a constellation term, as different lncRNAs appear to have different functions. Even today, pseudogenes are thought by some to be "junk" expression. A pseudogene is a copy of a protein coding gene which has certain mutations that prevent it from being translated into a protein. It has been observed that when expression levels of a pseudogene are increased, expression levels of the corresponding protein coding gene are increased too[99]. Since both the full protein coding gene and both the pseudogene have very similar 3' UTR regions, it has been hypothesized that the pseudogene may act as a molecular sponge which can "sop up" miRNAs[100]. Once a miRNA has bound to its target, it is not re-used elsewhere for further repression. Thus if a miRNA "finds" a pseudogene first it will not be able to target the protein coding version of the gene. To sum up, the RNAseq technological revolution has enabled us to posit braver and bolder interrogations into the disease transcriptome.

Data processing for RNAseq is much more complicated than for gene expression arrays. Images from the sequencer are processed into "reads", which represent strings of nucleotides called by the sequencing experiment. To determine which genes are responsible for generating the observed reads, alignment must be done. Since the transcriptome has a gapped structure due to alternative splicing, special aligners such as Tophat[101,102] and RNAstar[103] must be used. After alignment, expression levels can be quantified by calculating counts per gene/transcript and then normalizing with tools like RSEM[104] or with a method which integrates counting and normalizing such as Cufflinks[102]. While RNAseq has the advantage of being able to profile a wider pool of RNA and detect changes in a wider range, it is much more memory, time, and storage intensive than gene expression arrays.

Linear Modeling

Usually the purpose of a microarray or RNAseq experiment is to find changes in the transcriptome that associate with disease condition or phenotype. For log2 normalized gene expression array data or log normalized RNAseq data, differentially expressed genes or transcripts can be identified by fitting each gene or transcript individually to a model. For this approach to work, there must be sufficient sample size and the data must be normally distributed. If these conditions are met, a linear model is constructed including both the parameter of interest or an interaction of parameters of interest, as well as covariates that may also influence the transcriptome. For paired study designs, subject specific effects are accounted for with a random effect term. Each linear model produces a p-value for each gene or transcript, which is the chance that the observed difference in the means of two groups being compared due to chance.

Network Construction

Thanks to advances in RNA sequencing technologies, it is now possible to profile both mRNA and regulatory noncoding RNA, such as miRNA. In order to understand and characterize the interplay between these two types of RNA, networking approaches have been developed. Specifically, these algorithms predict which mRNAs will be targeted by which miRNAs based on seed sequence, flanking sequences, genomic context, binding energy and conservation[105–108]. Using this prior information, a static directional network can be constructed of predicted transcriptomic regulation, where each node is a miRNA or mRNA and each edge is a regulatory event.

Since it is known that miRNAs and lncRNAs regulate mRNAs, it is possible to build a regulatory association network using expression data from gene expression arrays or RNAseq. Calculating a correlation coefficient between mRNAs and regulatory RNAs of interest represents an easy and intuitive way of constructing this kind of network. This integrated network has RNAs as hubs and directed correlation as edges and is driven by two kinds of information. For miRNAs, it is much more interesting to look at negative correlations, as miRNAs should have opposite expression changes from the genes that they regulate. On the other hand, it is much more interesting to look at positively correlated lncRNAs, since lncRNAs usually regulate gene expression in a positive or enhancing manner.

By performing large and small RNA profiling on COPD, IPF, and lung cancer samples we are empowered to study the transcriptome in an unbiased manner. On first pass observation these three diseases seem very different, as they have different pathology and physical manifestations. However, all three diseases are smoking related and all three diseases lack satisfactory treatment options. One approach to move towards improved therapeutics is to supplement the understanding of molecular pathogenesis. COPD, IPF, and lung cancer are all smoking associated but not all smokers are diagnosed with COPD, IPF, and/or lung cancer, this suggests that these three diseases are regulated by both genetics and environmental factors. Given that the transcriptome represents readout of the interaction between genes and environment in a patient, the single nucleotide resolution of RNA sequencing represents the ideal platform to study these three diseases.

Integrated Genomics Approach Reveals Convergent Transcriptomic and Network Perturbations Underlying COPD and IPF

Background and Introduction

Chronic lung diseases affect a large portion of the US population and account for over 100,000 deaths per year[3]. The majority of these deaths can be attributed to chronic obstructive pulmonary disease (COPD), a frequently occurring smoking induced lung disease. A second major contributor to this high mortality rate is idiopathic pulmonary fibrosis (IPF), a fibrotic smoking associated lung disease with a nearly 100% fatality rate which results in more than 15,000 deaths annually[27]. COPD is defined by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) as a disease state characterized by exposure resulting in irreversible airflow limitation[12], and is thought to result from recruitment of inflammatory cells in response cigarette smoke. A subset of patients (those with COPD subtype emphysema) experience ECM protein and elastin destruction, alveolar cell apoptosis, and/or repair failure, which ultimately causes emphysematous airsac enlargement. Conversely, IPF has a very different physical phenotype from emphysema and is characterized by the findings of usual interstitial pneumonia (UIP), including the presence of inflammatory fibrotic patchy foci, excessive ECM activity and abnormal remodeling[27].

The recent development of high throughput transcript profiling has allowed investigators to discover mechanisms underlying human diseases. In chronic lung disease, limited studies have been reported in COPD [70–73] or IPF [81,82,109,110] and studies of miRNA expression in COPD [79,111] or IPF [112,113] remain underdeveloped. Despite having common risk factors such as cigarette smoking, no studies to date have directly queried if synergistic pathways exist in IPF and COPD. Recent publications hypothesize that parallel pathways may be at play in the development of these two chronic lung diseases and that direct comparisons of underlying disease biology may be informative[21]. In this study I examined IPF, COPD and normal lung tissue profiled together, with the intent to verify this hypothesis by identifying convergent transcriptional regulatory networks in COPD and IPF by leveraging integrative computational and functional transcriptomic approaches.

We performed mRNA sequencing and miRNA profiling using microarrays on 89 lung tissue samples from subjects with IPF, COPD, or without either disease. Samples were obtained through the NHLBI Lung Tissue Research Consortium (LTRC) as part of the Lung Genomic Research Consortium (LGRC). To glean disease-associated alterations in gene expression, we sequenced the lung transcriptome of each subject and identified molecular alterations shared by both chronic lung diseases. The p53/hypoxia pathway was up-regulated in both COPD and IPF compared to histologically normal controls, which was validated using a different gene expression technology in an independent sample cohort. My work provides the first RNA-seq study of chronic lung injury as a response to cigarette smoke as represented by both COPD and IPF. The overall study design and findings provide vision into a shared chronic lung disease response and highlight the central role of the p53/hypoxia pathway.

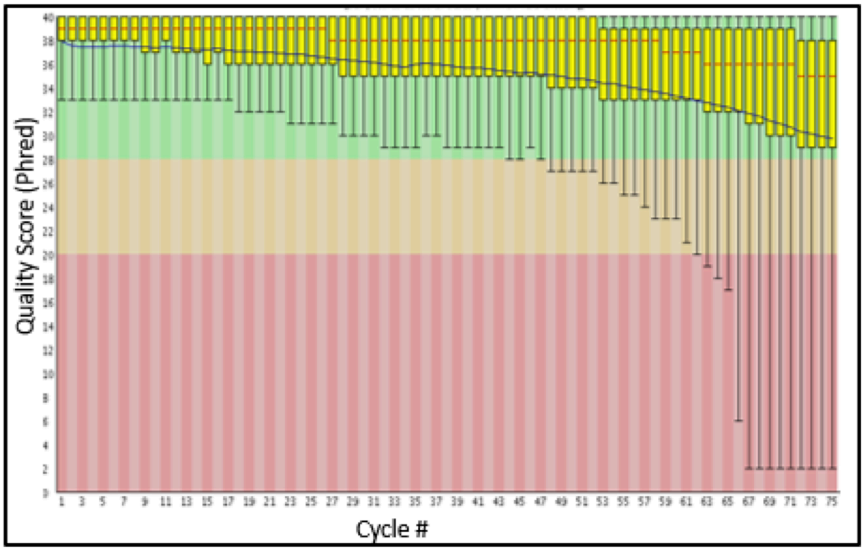
Results

Characteristics of Study Population and Samples Collected

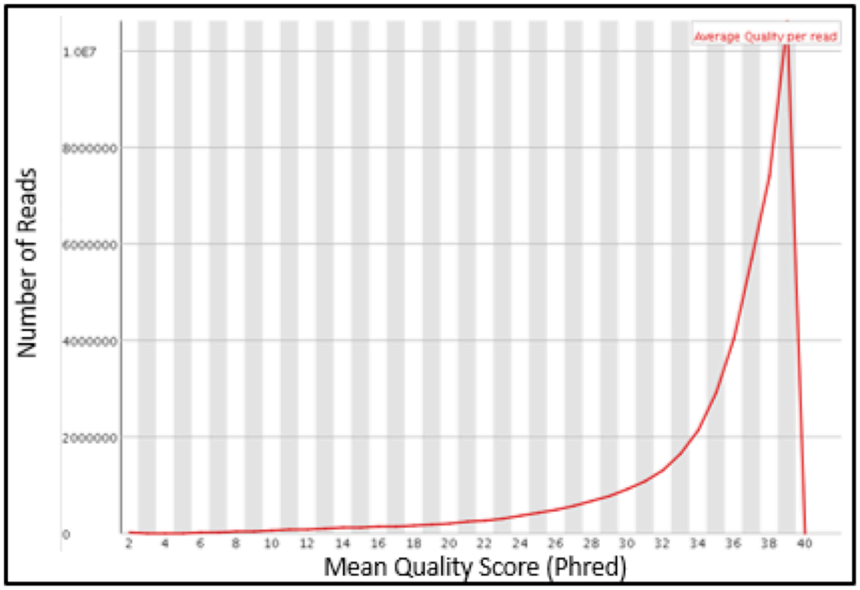
All lung samples were obtained from the Lung Tissue Research Consortium (LTRC) via the Lung Genomics Research Consortium (LGRC), both financed by the National Heart, Lung, and Blood Institute (NHLBI). In addition to tissue samples, the LTRC provided patient clinical information such as pulmonary functions, demographics, imaging results, pathology and clinical diagnoses. Seventy-five of the eighty-seven samples which exhibited distinct phenotypes of emphysema, COPD without emphysema, IPF or normal histology controls were deemed eligible for analysis (Table 1). Samples with intermediate percent emphysema were excluded from analysis. Samples were batched for sequencing and emphasis was placed on balancing age, smoking history, and gender across all batches. Institutional Review Boards approved all studies at participating collection and research institutions and all patients signed informed consent.

	Control	**IPF**	**COPD (Emphysema >= 30%)**	**COPD (Emphysema < 10%)**
Numbers	20	19	19	17
Age	63.3 +/- 10.0 (0)	64.0 +/- 9.7 (0)	56.3 +/- 8.7 (0)	68.4 +/- 10.4 (0) *
Sex	11 M, 9 F (0)	15 M, 4F (0)	10 M, 9 F (0)	12 M, 5 F (0)
Pack Years	27.3 +/- 22.6 (4)	31.24 +/- 23.4 (2)	47.9 +/- 27.4 (1)	50.7 +/- 20.4 (2)
Percent Emphysema	0.8 +/- 1.3 (2)	1.7 +/- 2.6 (7)	47.5 +/- 9.1 (0) ***	2.3 +/- 2.0 (2)
Cigarette Smoking Status	1 Current, 14 Former, 2 Never (2)	17 Former, 2 Never (0)	18 Former, 1 Never (0)	15 Former (2)

Table 1: Demographic Information of Samples Used: * = Significant with p < 0.05, * = Significant with p < 0.001, (#) = missing demographics.**



<u>Fig. 1: Quality Score Across All Bases in One Sample</u>. Blue line = the mean quality, Red line = median value, Yellow box = inter-quartile range (25-75%), Upper and lower whiskers = 10% and 90%.



<u>Fig. 2: Quality Score Distribution Across One Sample.</u>

RNA Sequencing

Each sample yielded approximately thirty million 75 nucleotide (nt) high quality paired-end reads (Figs. 1 and 2), and, on average, 28 million of these reads aligned to human genome build 19 using conservative alignment parameters. Specifically, 85.9% ± 6.9% of reads aligned to the genome, and 81.4% ± 3.1% aligned uniquely. Of the aligned reads, 90.3% ± 4.8% were aligned as paired ends (of which 88.7% ± 3.8% were properly paired), and 9.04% ± 4.8% were aligned as singletons. From these statistics I concluded that the RNA-seq data obtained from the LTRC tissue samples were of high quality.

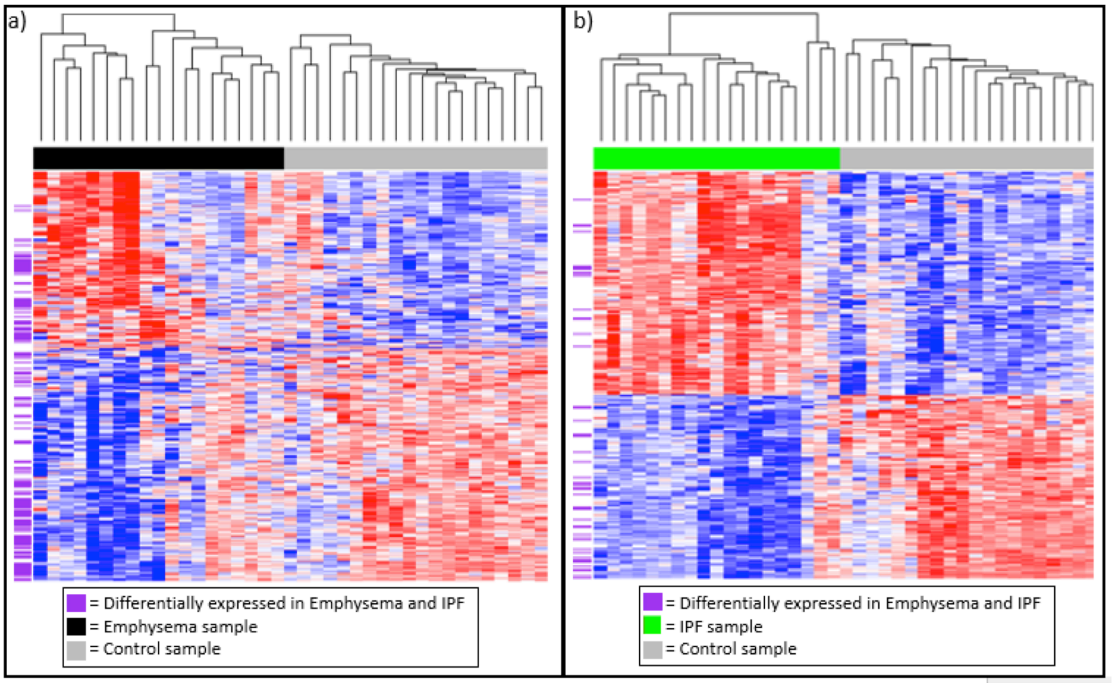
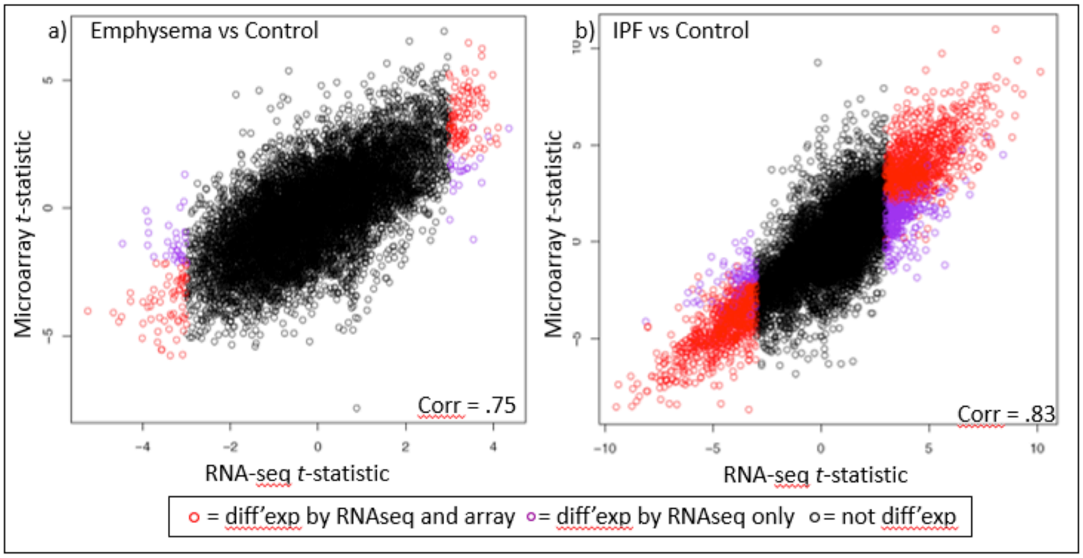


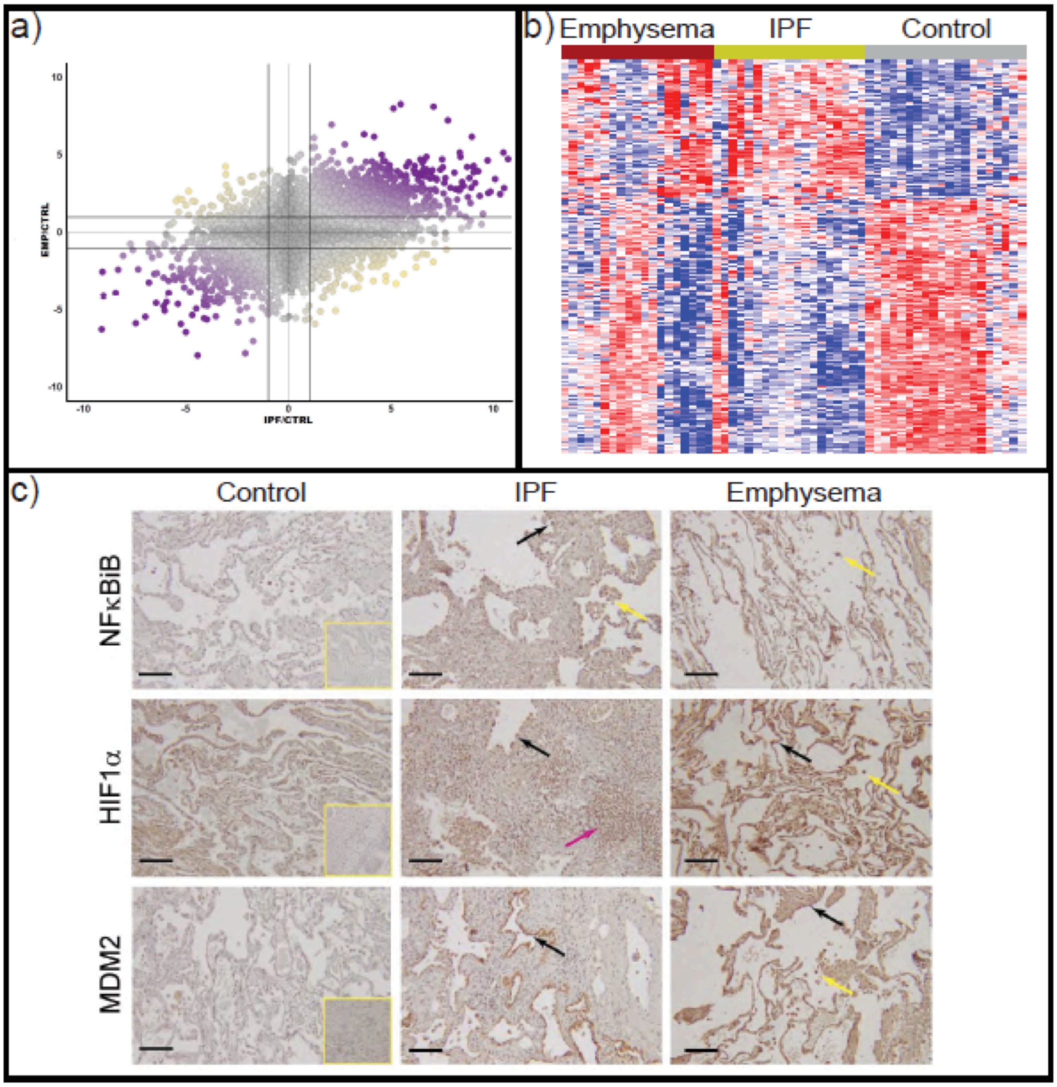
Fig. 3. Differentially Expressed Genes. Red indicates higher relative expression, blue indicates lower relative expression. A) Top 300 genes differentially expressed in emphysema vs. control B) Top 300 genes differentially expressed in IPF vs. control (pval<0.005).

Emphysema and IPF Differential Expression

I identified 2490 genes significantly differentially expressed (DE) between IPF and control subjects, and 337 genes DE between emphysema versus control subjects (P-value<0.005, 55 IPF genes and 53 emphysema genes expected by chance, Fig. 3). The number of genes differentially expressed between subjects with non-emphysema COPD and histologically normal controls at the same p-value threshold is less than by chance. These results were validated using gene expression microarrays run at a different university on the same 75 samples. The t-statistics were significantly correlated between RNAseq and gene expression microarrays (emphysema versus control R=.75 (Fig. 4A), p-value<0.001; IPF versus control r=.83, p-value<0.001 (Fig. 4B)).



<u>Fig. 4: T-statistic between Gene Expression Arrays and RNAseq.</u> A) Correlation of t-statistic of emphysema vs. control B) correlation of t-statistic of IPF vs. control.

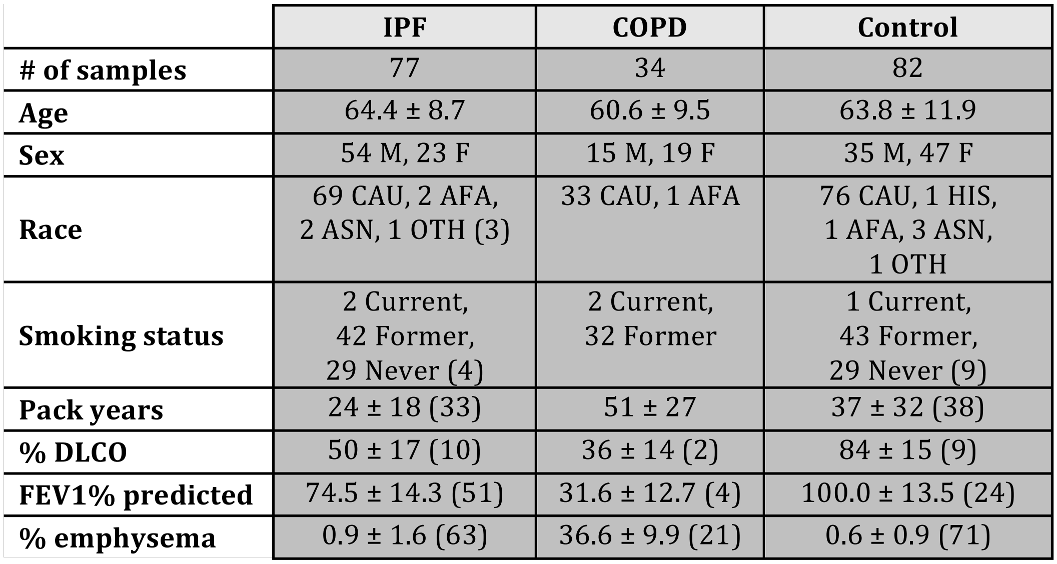


<u>Fig. 5: Differential Expression in Emphysema and IPF</u>. A) scatter plot showing correlation between emphysema vs. control (y axis) and IPF vs. control (x axis) t-statistic B) Genes which are commonly perturbed in emphysema and IPF. Red is higher relative expression, blue is lower relative expression. C) IHC of key p53 related genes. Black arrows = epithelial cells, yellow arrows = macrophages, pink arrows = lymphoid aggregates

Strikingly, the genes that distinguished IPF or emphysema from histologically normal controls revealed that, while not necessarily similar in magnitude, the overall change in gene-expression is concordant as shown in a scatter plot of all genes (Fig. 5A). Moreover, it was discovered that 214 genes shared statistically significant changes in expression between disease and normal histology controls, which can be seen in the heatmap (Fig. 5B). These common molecular alterations were significantly enriched for genes in the KEGG p53 pathway, Biocarta p53/Hypoxia pathway, Gene Ontology epidermis development, and other biological processes outlined in Table 2.

Up in Emphysema and IPF	**Down in Emphysema and IPF**
Biocarta p53/Hypoxia	Biocarta Myosin
KEGG Alanine Metabolism	Biocarta Par1
KEGG Autoimmune Thyroid	KEGG Endocytosis
KEGG p53	KEGG Long Term Potentiation
KEGG Ribosome	GO Anatomical Morphogenesis
GO Epidermis Development	GO Endocytosis
GO Tissue Development	

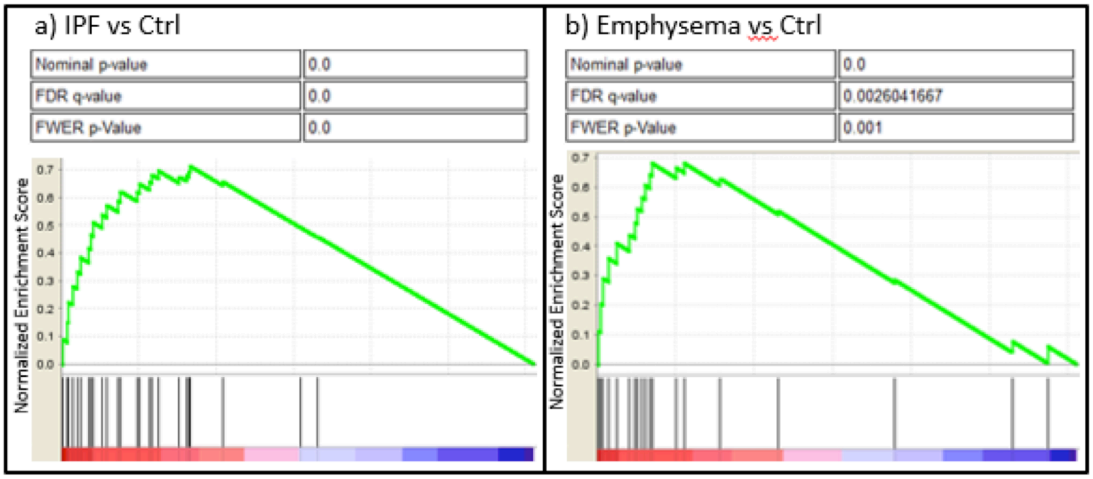
Table 2: Functional Enrichment of Shared Emphysema and IPF Genes. GSEA of Biocarta, KEGG, and GO gene sets were used against ranked lists of Emphysema vs. Control and IPF vs. Control. Results included are those with pval < .05 and concordance in the same direction.



	IPF	COPD	Control
# of samples	77	34	82
Age	64.4 ± 8.7	60.6 ± 9.5	63.8 ± 11.9
Sex	54 M, 23 F	15 M, 19 F	35 M, 47 F
Race	69 CAU, 2 AFA, 2 ASN, 1 OTH (3)	33 CAU, 1 AFA	76 CAU, 1 HIS, 1 AFA, 3 ASN, 1 OTH
Smoking status	2 Current, 42 Former, 29 Never (4)	2 Current, 32 Former	1 Current, 43 Former, 29 Never (9)
Pack years	24 ± 18 (33)	51 ± 27	37 ± 32 (38)
% DLCO	50 ± 17 (10)	36 ± 14 (2)	84 ± 15 (9)
FEV1% predicted	74.5 ± 14.3 (51)	31.6 ± 12.7 (4)	100.0 ± 13.5 (24)
% emphysema	0.9 ± 1.6 (63)	36.6 ± 9.9 (21)	0.6 ± 0.9 (71)

Table 3. Demographics of Independent Gene Expression Array Cohort: Parenthesis indicates missing data.

An identical analysis was performed on gene expression from a non-overlapping, independent cohort of lung tissue samples obtained from the LTRC (Emphysema N=34, control N=77, IPF N=82, Table 3) and profiled using microarrays run at the University of Pittsburgh. Samples were selected based on available clinical data, which was limited for certain patients. This analysis confirmed the up-regulation of the p53/hypoxia pathway in genes that distinguished emphysema or IPF from normal histology controls. Gene Set Enrichment Analysis (GSEA)[77] corroborated significant enrichment of the KEGG p53 and Biocarta p53/hypoxia leading edge from GSEA of the primary cohort among genes up-regulated in emphysema or IPF tissues compared to histologically normal controls (p-value < 0.001 , Fig. 6). Since cell type differences were a concern, we used immunohistochemistry to confirm the location of select differentially expressed genes in the p53 pathway in control, Emphysema, and IPF samples (N=5 for each). HIF1A, MDM2, and NFKBIB were all found to be expressed in the airway epithelium, suggesting that this is not a cell type effect (Fig. 3C).



a) IPF vs Ctrl

Nominal p-value	0.0
FDR q-value	0.0
FWER p-Value	0.0

b) Emphysema vs Ctrl

Nominal p-value	0.0
FDR q-value	0.0026041667
FWER p-Value	0.001

Fig. 6: Leading Edge of p53/Hypoxia Genes in Independent Sample Set. A) Ranked list was IPF vs. control B) Ranked list was emphysema vs. control.

miRNA Regulation of Shared Emphysema and IPF Differential Expression

Integrating mRNA-Seq and miRNA microarray expression data on the same samples uncovered additional insights into the transcriptomic regulation of the p53/hypoxia pathway in emphysema and IPF. Using miRconnX[114] I created a data-driven and prior-knowledge-based gene/miRNA regulatory network. Initially, I constructed a regulatory network using genes differentially expressed ($p < 0.05$) in the same direction in both emphysema and IPF to explore shared regulatory mechanisms between the two diseases (http://mirconnx.csb.pitt.edu/job_results?job_id=example10103). The network contains 15 miRNA, including miR-96, and 31 genes. We created two additional networks by submitting Emphysema vs. control genes (http://mirconnx.csb.pitt.edu/job_results?job_id=example10102) and IPF vs. control genes (Fig. 7, or http://mirconnx.csb.pitt.edu/job_results?job_id=example10101) with the same p-value cutoff. Both of these networks featured miR-96 as the most connected miRNA, suggesting that it plays an important regulatory role.

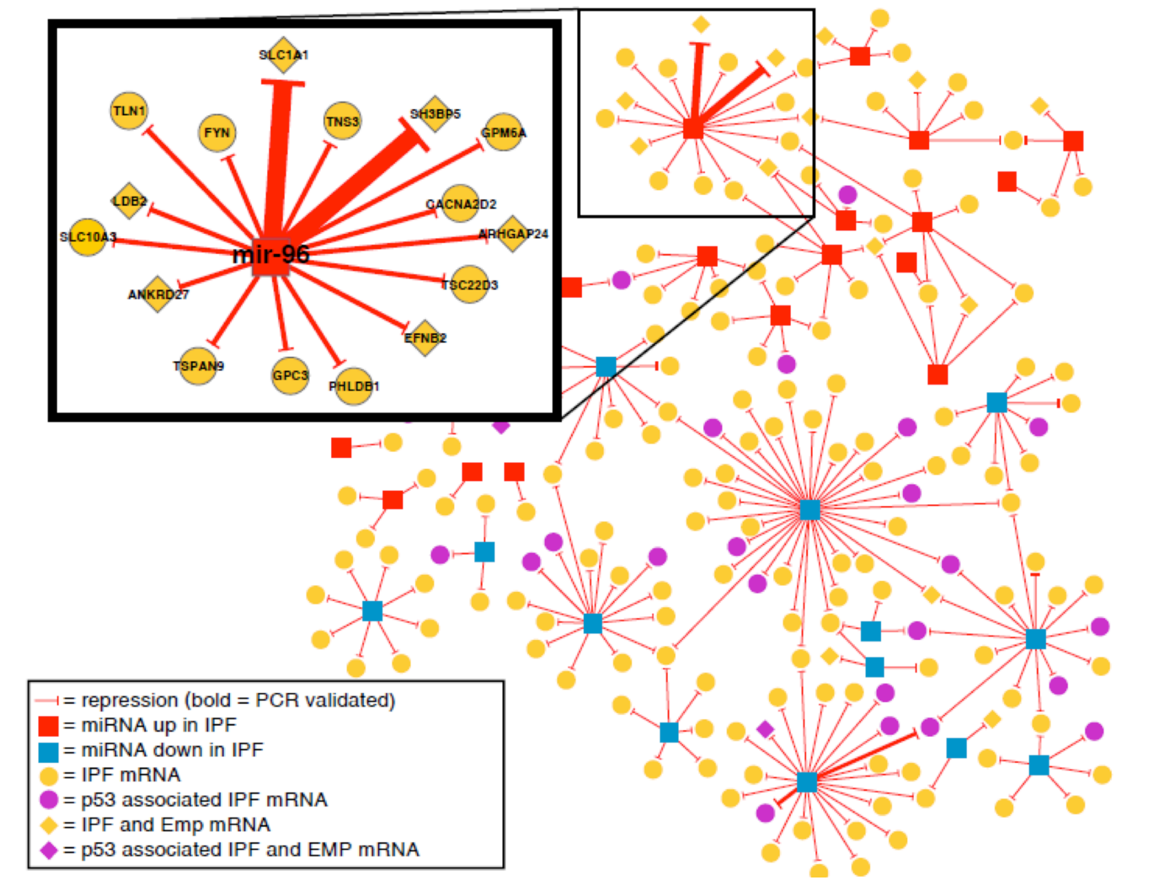


Fig. 7: Shared Emphysema and IPF miRNA Regulatory Network. Regulatory miRNA-mRNA network showing regulation in both diseases. Red lines indicate direction of repression. Bold red lines indicate interactions that were selected and validated by PCR. IPF mRNA = mRNA differentially expressed in IPF vs. control. Emp mRNA = mRNA differentially expressed in emphysema vs. control.

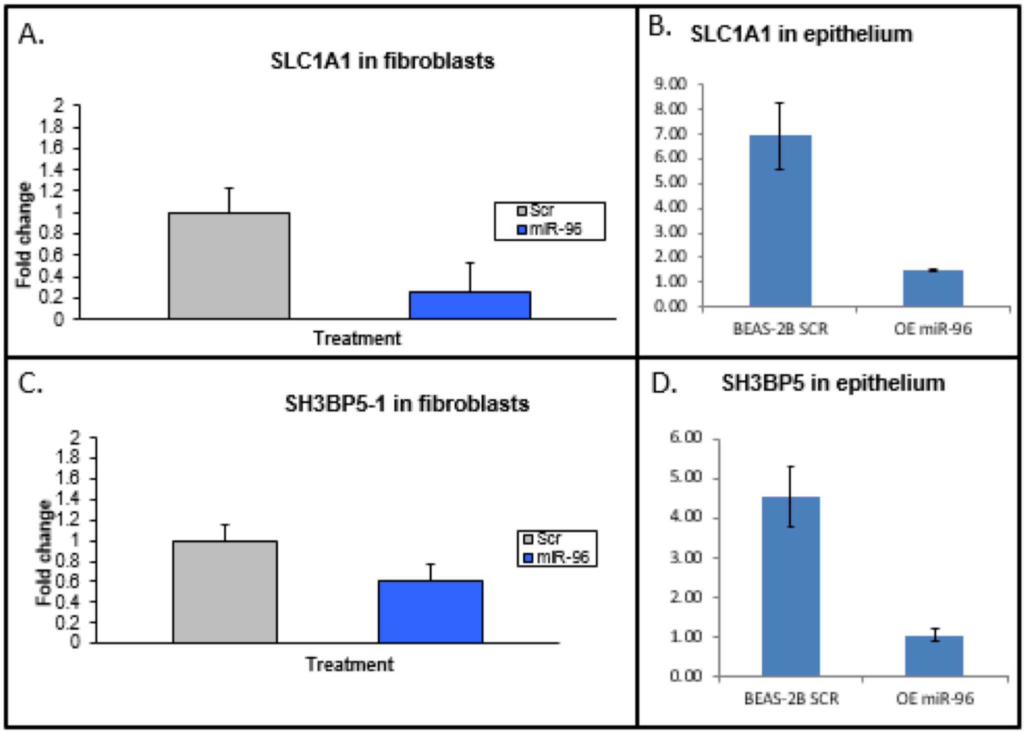
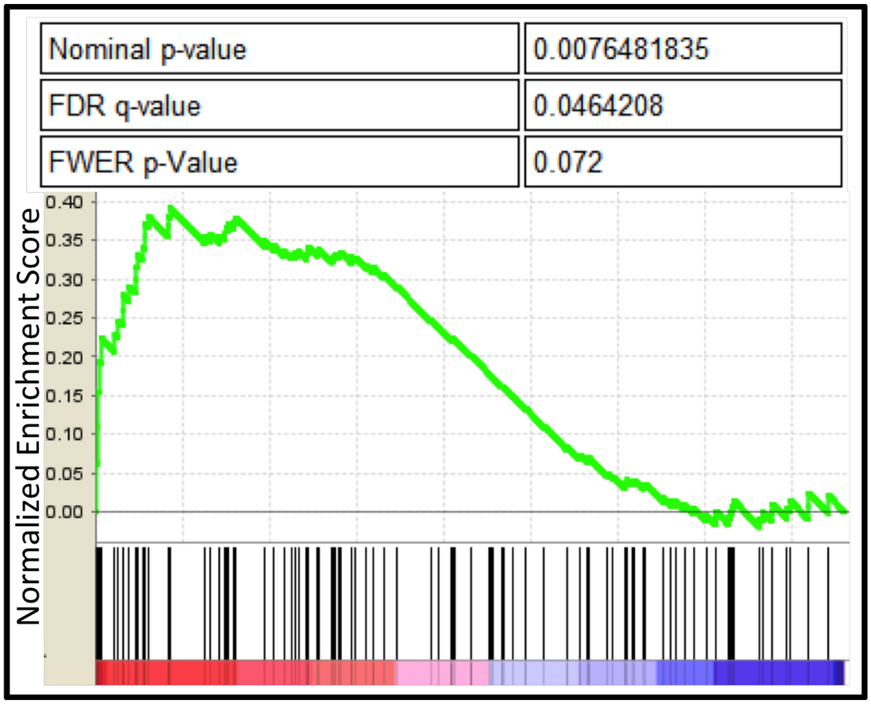


Fig. 8: Validation of SLC1A1 and SH3BP5 in Fibroblasts and Epithelial Cells. miR-96 was overexpressed in fibroblast and epithelial cell lines. Overexpression of miR-96 induced decreased expression of SLC1A1 and SH3BP5.

We validated the up-regulation of miR-96 in the emphysema and IPF regulatory network by qRT-PCR. Overexpression of miR-96 in fibroblasts and epithelial cells repressed the expression of glutamate transporter SCL1A1 and BTK inhibitor SH3BP5 (Fig. 8). These genes are down-regulated (as observed in RNAseq data) in both diseases and repressed by miR-96 in the shared regulatory network generated using miRconnX. To interrogate all genes that change with overexpression of miR-96, we ran gene expression arrays on RNA from our miR-96 overexpression studies. GSEA of these arrays revealed that genes that go up with overexpression of miR-96 in epithelial cells were enriched for genes that also go up in IPF relative to control (Fig. 9). Importantly, this result revealed that overexpression of miR-96 recapitulated some of IPF associated increases in gene expression and suggests its potential as a therapeutic target.



Nominal p-value	0.0076481835
FDR q-value	0.0464208
FWER p-Value	0.072

Fig. 9: Enrichment of Genes Up with miR-96 Overexpression in IPF vs. Control. **GSEA was used to test for enrichment of genes up with miR-96 in IPF vs. control genes. X-axis represents ranked list.**

Discussion

This study represents the most encompassing transcriptomic study of non-malignant chronic lung disease to date, and our particular study design empowered us to define the disease networks that are shared across two lung conditions: COPD subtype emphysema and IPF. While these two diseases have distinct clinical, radiographic and pathological manifestations, they share a common environmental exposure: cigarette smoke. My results suggest the presence of common transcriptional networks associated with both diseases which provide insight into the lung's response to chronic injury.

My initial analysis aimed at identifying the convergent molecular network in COPD and IPF at the gene-expression level. One of the striking findings from that analysis was the relatively large number of genes that were differentially expressed between IPF and normal lung, as compared to the number of genes that are differentially expressed between emphysema vs. normal lung. These findings may be driven by distinct cellular changes that characterize the fibrotic foci that were profiled compared to the more heterogeneous cell type composition in emphysema. Despite the cellular differences between conditions, I identified a shared molecular network enriched for the up-regulation of the p53/hypoxia pathway. When the p53 pathway is triggered by hypoxia instead of DNA damage, as I suspect happens in lungs with emphysema or IPF, apoptosis is not triggered. Specific members of the p53/hypoxia pathway have previously been shown to be up-regulated in one of the two diseases, including HIF1A, TP53, MDM2, CDKN1A, and BAX in IPF [115,116] and TP53 and BAX in emphysema [117]. Our work, however, provides an important advance by profiling both diseases simultaneously and characterizing additional components of this pathway that are similarly altered in both diseases. The upregulation of certain members of this pathway is of interest because these members are also upregulated in cancer, which emphysema and IPF are risk factors for. Specifically, HIF1A is transcribed at high levels in many tumors, and high expression is a marker of invasiveness and malignancy[118]. HIF1A activates p53. Activated p53 causes potent oncogene MDM2 to be transcribed, which then in turn ubiquinates p53[119]. Other shared p53 pathway members upregulated in both diseases includes SESN2, which is directly activated by p53 and is a critical part of the antioxidant response[120]. Further functional studies beyond the scope of this work are needed to pin down the exact role that the p53 pathway is playing, as current analysis provides a view only of transcript levels, not phosphorylation, ubiquination, or any number of other post-translational regulatory mechanisms.

A second approach explored the shared molecular network between IPF and COPD by synergizing the mRNA-seq and microRNA array data generated on the same lung tissue samples. The resulting network revealed that both diseases share common transcriptional regulatory motifs, with several microRNAs in common between regulatory networks. Of particular interest is MIR-96, which is up-regulated in both diseases and is suggested to regulate a number of genes differentially expressed in both IPF and COPD including SCL1A1, SH3BP5, LDB2, and ARHGAP24. SCL1A1 is a glutamate transporter, which is down-regulated under hypoxic conditions[121,122]. SH3BP5 inhibits BTK[123], which is a binding partner of Hypoxia Induced Mitogenic Factor (HIMF)[124]. LDB2 binds to LIM domain binding proteins[125], which inhibit HIF1A[126]. Importantly, we were able to demonstrate that overexpression of MIR-96 in both lung epithelial cells and fibroblasts in vitro recapitulates components of the shared COPD-IPF gene-expression network, providing further evidence that MIR-96 may be an important regulator of the shared disease gene-expression network.

While the unique study design and comprehensive transcriptional profiling provided an unprecedented resolution of the lung transcriptome in health and disease, there are a number of important limitations to this analysis. We profiled whole lung tissue and thus some of the differential gene-expression identified between emphysema or IPF as compared to control may simply reflect differing proportions of lung cell types. To address this concern, we pursued immunohistochemistry IHC to validate the cell type responsible for expression of a select number of genes. Despite the difference in cell types, the convergence of overall differential expression signals suggests the observed changes are due to disease biology. Given the cross-sectional nature of our study, we cannot readily distinguish gene expression changes that are causal versus consequential of the disease process. The integrative mRNA-miRNA network and the subsequent functional validation studies in vitro provide some evidence for a causal relationship between regulatory miRNA and the disease-associated gene expression network but do not prove direct first-degree regulation. Additionally, some of the control lungs were taken from smokers with lung cancer, suggesting that a portion of the differential expression could be driven by the influence of the tumor on the lung as a whole. To limit this potential effect, our collaborators used SNP arrays from the lung and blood to filter out samples with cytogenetic abnormalities.

In summary, the first aim of my thesis has exhibited the ability of next generation sequencing to provide unprecedented resolution of the lung in healthy and disease states. Importantly, by profiling distinct lung diseases in parallel within the same study, we uncovered molecular networks that are shared among smokers with IPF and emphysema. Our study also shows the necessity of integrating diverse genomic data from the same specimen in order to discover disease associated regulatory networks. With additional functional validation studies, these networks may not only provide insight into disease pathogenesis, but could eventually lead to novel diagnostic biomarkers and therapeutic targets for chronic lung disease.

Materials and Methods

Sample Collection

As part of the LTRC we collected lung tissue samples. We evaluated the initial 89 samples for potential field of cancerization effects, as some samples were collected from areas adjacent to lung cancer tumors. Two control samples contained between 12% and 25% abnormal cells by allele balance via the Illumina Infinium genotyping array and were thus removed from further analysis.

Patient Demographics

Remaining were n=19 COPD subjects with predominant Emphysema phenotype, n=17 COPD subjects without predominant emphysema phenotype (COPD airways disease), n=19 IPF subjects, n=13 COPD subjects with intermediate emphysema phenotype, and n=20 histologically normal tissue samples. 75 of the remaining 87 samples were selected by pathologists as displaying the most distinct phenotypes and were used for differential large and small RNA analysis (Supplemental Table 1). The COPD categories were defined based on the percent emphysema: samples with <10% Emphysema and >30% Emphysema were used to define the COPD airway and Emphysema phenotypes respectively. The ILD samples were subset down to those with “IPF” phenotype as per ATS criteria [29].

RNAseq Processing

We extracted total RNA from all lung samples using the QIAcube system (QIAGEN Inc., Valencia, CA) with the miRNeasy kit. RNA quality was determined using a Bioanalyzer 2100 (Agilent Technologies, Santa Clara CA) with a RNA Integrity Number (RIN) > 7.0 as the criterion for acceptable quality.

Large RNA-seq

Library preparation and mRNA sequencing was performed on each of the 89 LGRC samples. The mRNA was isolated using poly(A) selection, fragmented, and randomly primed for reverse transcription followed by second-strand synthesis to create double-stranded cDNA fragments. Ends were repaired, ligated to Illumina Paired-End sequencing adapters, and fragments of 300 bp were obtained through gel-based size selection. These fragments were PCR amplified, purified, and then subjected to cluster generation using Illumina Paired-End Cluster Generation Kit v4. Each sample was sequenced on 1 lane of an Illumina Genome Analyzer IIX to generate 30-40 million 75 nt paired-end reads having an average inner distance between mate pairs of 50 bp.

Initial data processing was done using Illumina GA pipeline version 1.3. The quality of each sample sequenced was assessed by examining several Illumina metrics such as the percent of clusters passing the filter, the density of the cluster passing the filter, and the number of sequencing cycles with a median phred quality score (log of base calling error probability) less than 30. In addition, we examined the distribution of the quality scores as a function of position, the nucleotide composition as a function of position, and histograms of the inner distance between paired end reads mapped to the genome using our own custom perl scripts as well as the FastQC java program by Simon Andrews at Babraham Bioinformatics (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

After samples were evaluated for quality, we aligned the samples to the human genome using Tophat version 1.2[127]. Tophat is a gapped aligner specifically optimized for RNA-seq data which identifies the reads that align to the genome as well as reads that span known and novel exon-exon splice junctions. We aligned the reads as pair ends. We allowed up to 2 mismatches per 25 nt segment and allowed the default number of multi-reads. Tophat was run by specifying Illumina GA pipeline version 1.3 or greater, unstranded library, a mate inner distance of 50, and a mate standard deviation of 100. After alignment, samples were again assessed for quality by examining the number of reads that aligned uniquely and the number of reads that aligned with varying number of mismatches. The first and second principle components were also assessed as part of post-alignment quality control. No outliers were found based on these analyses.

The results of the gapped alignment were used to quantify the number of reads mapping to each gene. We generated gene level expression estimates for each of the 87 samples using Cufflinks Version1.1.0[102]. The gene annotation for genes containing multiple transcripts was created by including common and unique regions of each transcript. Cufflinks was run to only quantify known transcripts through the use of a modified Ensembl59 GTF. Furthermore Cufflinks was run such that it performed upper quartile normalization, multi-read correction, and nucleotide bias correction.

First, we log2 transformed FPKM gene expression data from Cufflinks. Using their “on” or “off” status and coefficient of variance, we filtered genes. To determine a given gene’s status we used a modified version of the mixture model in the SCAN.UPC Bioconductor package[128]. For a gene to be included in differential expression analysis, it had to be classified as “on” in at least 25% of samples out of the two phenotypic groups being compared, but regardless of phenotype. Next, the bottom 20% of genes were filtered out based on their coefficient of variation.

mRNA and miRNA Array Processing

RNA from each subject was reverse transcribed, labeled with cyanine-5, and hybridized to Agilent V2 Human Whole Genome microarrays. The samples were randomized both by disease state and order in which the samples were hybridized to minimize batch effects. Immediately after hybridization and subsequent washing with Agilent Gene Expression Wash Buffer, the microarrays were scanned using the Agilent DNA Microarray Scanner. The resultant data was globally normalized using cyclic loess by in-house software built in the R programming environment. Differential gene expression was measured using BRB ArrayTools developed by Dr. Richard Simon and BRB-ArrayTools Development Team. RNA prepared as described above was also hybridized to the Agilent Human miRNA Microarray (V3). Samples were again randomized to avoid batch effects. Arrays were washed with Agilent miRNA Expression Wash Buffer and scanned using the Agilent DNA microarray Scanner. The data as then quantile normalized using GeneSpring.

Modeling Disease Associated Changes in Gene Expression

I identified differentially expressed genes with the limma R package. For Emphysema, we included only samples with greater than 30% emphysema. From the Interstitial Lung Disease population, we included only samples with pure Idiopathic Pulmonary Fibrosis. I included only genes annotated as "known" in Ensembl. Overall, we chose samples such that age, pack years, smoking status, and gender were not confounded with disease status. We acquired tissues from the Lung Tissue Research Consortium (LTRC). As previously described, we performed Immunohistochemistry employing mouse monoclonal antibodies directed against MDM2 (Millipore, Temecula, CA), HIF1A (Stressgen, Victoria, BC, Canada), and NFKBIB (ABD Serotec, Raleigh, NC), and a rabbit polyclonal antibody directed against PDGFA (Santa Cruz Biotechnology, Santa Cruz, CA). We took all brightfield images with an Olympus DP25 camera on an Olympus CH2 microscope[129]. We integrated miRNA array and mRNA-Seq data with MirConnX[114]. This tool combines a prior, static network created from miRNA binding predictions and literature validation with user submitted data to create a transcriptomic gene regulatory network. For each condition-control comparison, we filtered to only differentially expressed mRNAs. We inspected resulting regulatory networks for potential regulatory hubs (miRNAs with a high number of connected mRNAs).

P53 Pathway Validation with Gene Expression Arrays

Using a t-test in limma (same as for RNAseq analysis) I identified disease associated changes in gene expression. Genes in Ensembl without Agilent probe mappings were excluded from analyses. Platforms were compared by evaluating the overlap between genes identified as differentially expressed. Correlation between t-statistics on the two platforms was found using a Pearson correlation.

Overexpression of miR-96

NHLF p.6 (Lonza) were plated at 70% confluence (150,000 cells per well in 6-well plate) in FGM™-2 BulletKit™ medium (Lonza). After 6 hours media was replaced by Opti-MEM (Invitrogen) over night. 20nM of miR-96 and negative control (scramble) (Applied Bio Systems) were transfected using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. Media was replaced back to growth media after 6 hours. RNA was isolated 24h post-transfection.

Immortalized human bronchial epithelial cells (Beas-2B) were cultured in BEGM growth medium (Lonza) and plated at a 50% confluence in 6-well plates 24h before transfection. 100pmoles of pre-miR-96 or a Cy-3-labeled scrambled control (Ambion) were transfected into the cells using Lipofectamine RNAiMAX (Invitrogen) as per manufacturer's protocol. Cells were harvested at 48 hours post-transfection and total RNA was isolated using the miRNeasy mini kit (Qiagen).

To measure the expression of SLC1A and SH3BP5 in both of the above cell lines, total RNA was reverse transcribed using random hexamers (Applied Biosystems) and Superscript™II reverse transcriptase (Invitrogen). 20ng of starting cDNA product was added to SYBR® Green PCR master mix (Applied Biosystems). The primer sequences are as follow: SLC1A forward (5'- TAG GTA TTG TGC TGG TGG TGA G -3), SLC1A reverse (5'-TGA GAT CTA ACATGG CAT CCA C-3'), SH3BP5 forward (5'-CGA

GCA ACT GAA AAA GAC TGT G-3’) and SH3BP5 reverse (5’-TTC TTC AGG GCC ATC TTG TAC T-3’). Forty cycles of amplification were used and data acquisition was carried out with the StepOne Real Time PCR System (Applied Biosystems). The data was analyzed using the comparative CT method [130].

In order to perform mRNA microarray profiling, 50ng of total RNA from the BEAS2B cell line was extracted and used as a template for double stranded cDNA synthesis. The cDNA was used as a template to generate Cy3 labeled cRNA (using the low input kit) to be used after for hybridization. After purification and fragmentation of the samples was performed and were hybridized to Agilent SurePrint G3 Human Gene Expression 8x60K v2 Microarray at 65°C for 17h. Each array, after hybridization, was sequentially washed and scanned by Agilent Microarray Scanner. Images were processed using Agilent’s Feature Extraction software version 10.7.3.1.

Functional Enrichment of COPD, IPF, and COPD/IPF Gene Expression

DAVID was used to assess overrepresented functional categories or biological pathways in a given list of genes. In addition, GSEA was used to select enriched pathways in a ranked list (t-statistic) of COPD vs. control and IPF vs. control. Similarity of enrichment was assessed by reviewing concordance of the Normalized Enrichment Score (NES).

Never Smoker Lung Adenocarcinoma Exhibits Unique Transcriptomic Perturbations

Background and Introduction

Lung cancer is the leading cause of cancer-related death worldwide, claiming over 1 million lives annually[131]. While lung adenocarcinoma is predominantly considered a "smoker's disease", approximately 25% of these cancers arise in those who have never smoked, and it is the predominant cause of cancer related deaths among never smokers[31]. The number of lung cancer deaths among never smokers has increased annually and is currently estimated to be the seventh leading cause of cancer mortality[132].

Recent studies suggest that current or former (ever) smokers and never smokers who develop lung cancer harbor distinct profiles of somatic mutations and elicit disparate responses to targeted therapy[63,64]. EGFR oncogene mutations are present in 45% of never smoker lung cancers, but only 7% of ever smoker lung cancers[57], and can be exploited for targeted therapy with EGFR-TK (Tyrosine Kinase) inhibitors. Other molecular abnormalities, including p53 mutations, KRAS mutations, chromosomal aberrations, gene expression, and methylation profiles, all also vary between the lung tumors of never smokers and ever smokers[84,131,133]. This type of translational research has underscored the therapeutic value of identifying distinct molecular drivers and precipitates the need for a thorough comparison of the molecular differences between lung cancer cases in ever and never smokers.

Based on this evidence, we hypothesize that lung adenocarcinoma arises through distinct molecular processes in never smokers compared to ever smokers, and that these separate mechanisms of lung carcinogenesis necessitate different therapeutic approaches. By leveraging emergent transcriptomic sequencing technologies we are able to profile messenger RNAs as well as more exotic members of the transcriptomic zoo, such as noncoding RNA. In this study we sequence large and small RNA from the same patients. Thus, we are able to identify disease-associated changes in mRNA expression, miRNA expression, and the miRNA-mRNA regulatory network (Fig. 10, study overview). Network analysis in this study revealed a key miRNA-mRNA pair that appears to regulate tumorigenic gene expression and oncogenic phenotypes uniquely in never smoker tumors. Moreover, this transcriptomic analysis of never smoker lung cancer represents a critical first step in being able to extending our previously published bronchoscopy based lung cancer biomarker[134] from ever smokers into never smokers.

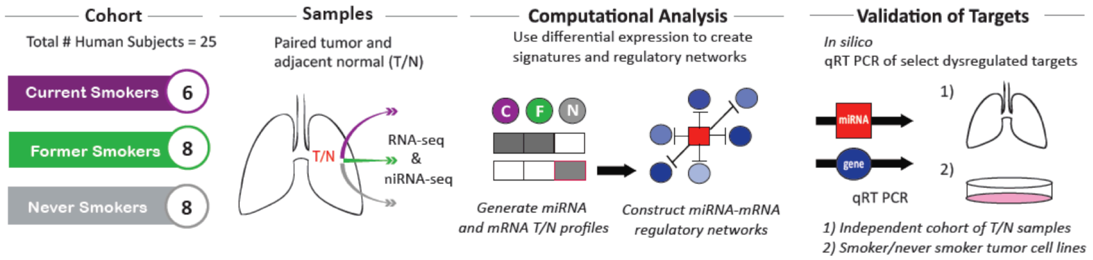
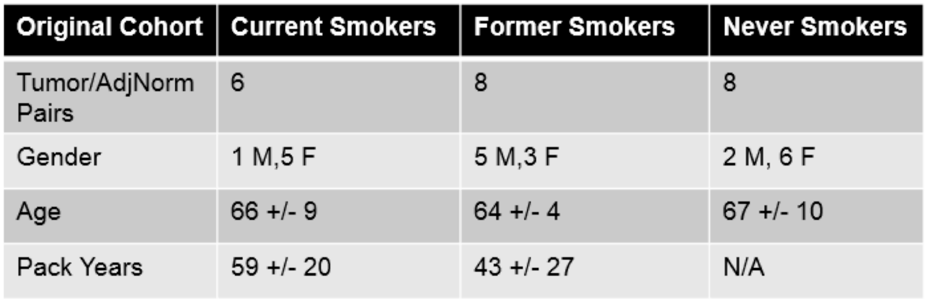

Cohort
Total # Human Subjects = 25
Current Smokers 6
Former Smokers 8
Never Smokers 8
Samples
Paired tumor and adjacent normal (T/N)
T/N
RNA-seq & niRNA-seq
Computational Analysis
Use differential expression to create signatures and regulatory networks
C F N
Generate miRNA and mRNA T/N profiles
Construct miRNA-mRNA regulatory networks
Validation of Targets
In silico
qRT PCR of select dysregulated targets
miRNA
gene
1)
2)
qRT PCR
1) Independent cohort of T/N samples
2) Smoker/never smoker tumor cell lines

Fig. 10: Overview of Study. Purple represents current smokers, green represents former smokers, light grey represents never smokers. Grey with a red box around it represents sought after gene expression patterns.

Results

Characteristics of Study Population and Samples Collected

Patients undergoing resection of lung adenocarcinoma, bronchioloalveolar carcinoma, or squamous cell carcinoma were recruited into this study. Samples originated at the Mayo clinic, and were sent to Lovelace Respiratory Research Institute where tumor purity was checked and RNA was isolated. The patient population contains a mixture of male and female patients, including current, former, secondhand, and never smokers (Table 4). All patients with squamous cell carcinoma are current or former smokers, all patients with bronchioloalveolar carcinoma are never smokers, and the patients with adenocarcinoma are a mix of all smoking statuses. For this analysis, I included only lung adenocarcinoma patients to compare the effect of smoking status.



Original Cohort	Current Smokers	Former Smokers	Never Smokers
Tumor/AdjNorm Pairs	6	8	8
Gender	1 M,5 F	5 M,3 F	2 M, 6 F
Age	66 +/- 9	64 +/- 4	67 +/- 10
Pack Years	59 +/- 20	43 +/- 27	N/A

Table 4: Demographics of Original RNAseq cohort.

Tumor Gene/miRNA Expression Profiles Associated Uniquely with Never Smoker Tumors

In order to identify changes in gene expression associated with the never smoker tumor, I utilized linear modeling. Strikingly, a linear mixed effects model with an interaction term followed by a post-hoc paired t-test pinpointed 120 large RNAs and 15 miRNAs uniquely changed in the never smoker tumor (Fig. 11, pval < 0.01 for lme and pval < 0.05 for paired t-test). Out of the 120 large RNAs, 49% are protein coding, 17% are lncRNAs and 34% are other ncRNA. Many of the noncoding RNA can easily be spotted in the heatmap by their distinctive “on” or “off” expression pattern. Some of the expression level alterations in protein coding genes have been previously reported to be important in other cancers, such as MDM2 being up and RBP1 being down in the never smoker tumor (Fig. 12A, 12B). The large RNA signature also revealed changes in potential key regulators such as the downregulation of lung development transcription factor FOXP2[135] (Fig 12D).

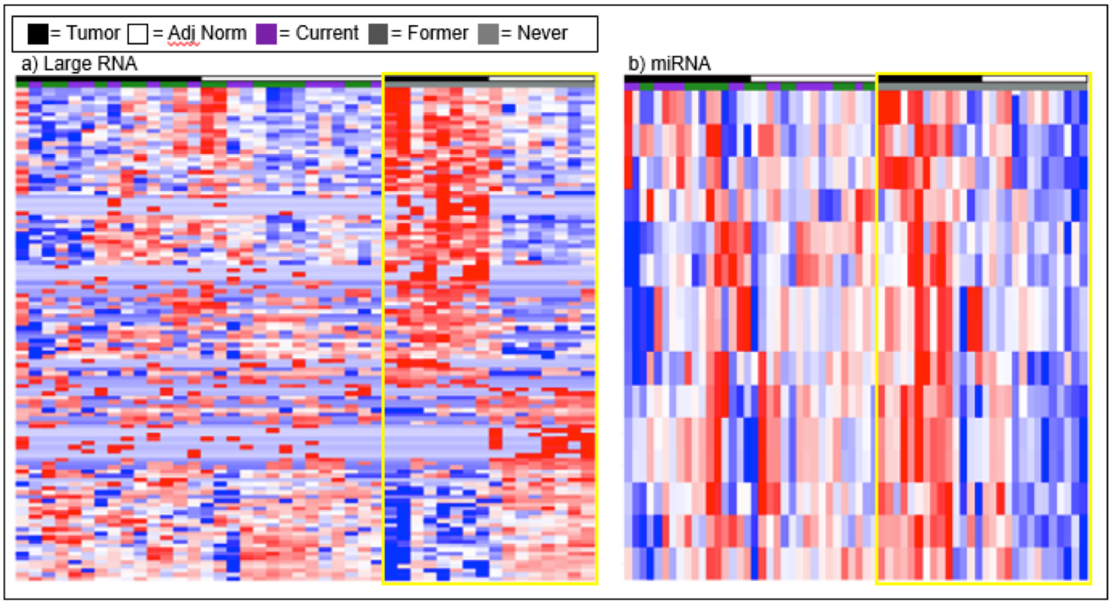


Fig. 11: Never Smoker Tumor Specific Changes in Large and Small RNA Expression. **Expression pattern of RNA differentially expressed in the tumor vs. adjacent normal of never smokers but not the tumor vs. adjacent normal of ever smokers A. 120 genes are differentially expressed uniquely in the never smoker tumor B. 15 miRNA are differentially expressed uniquely in the never smoker tumor.**

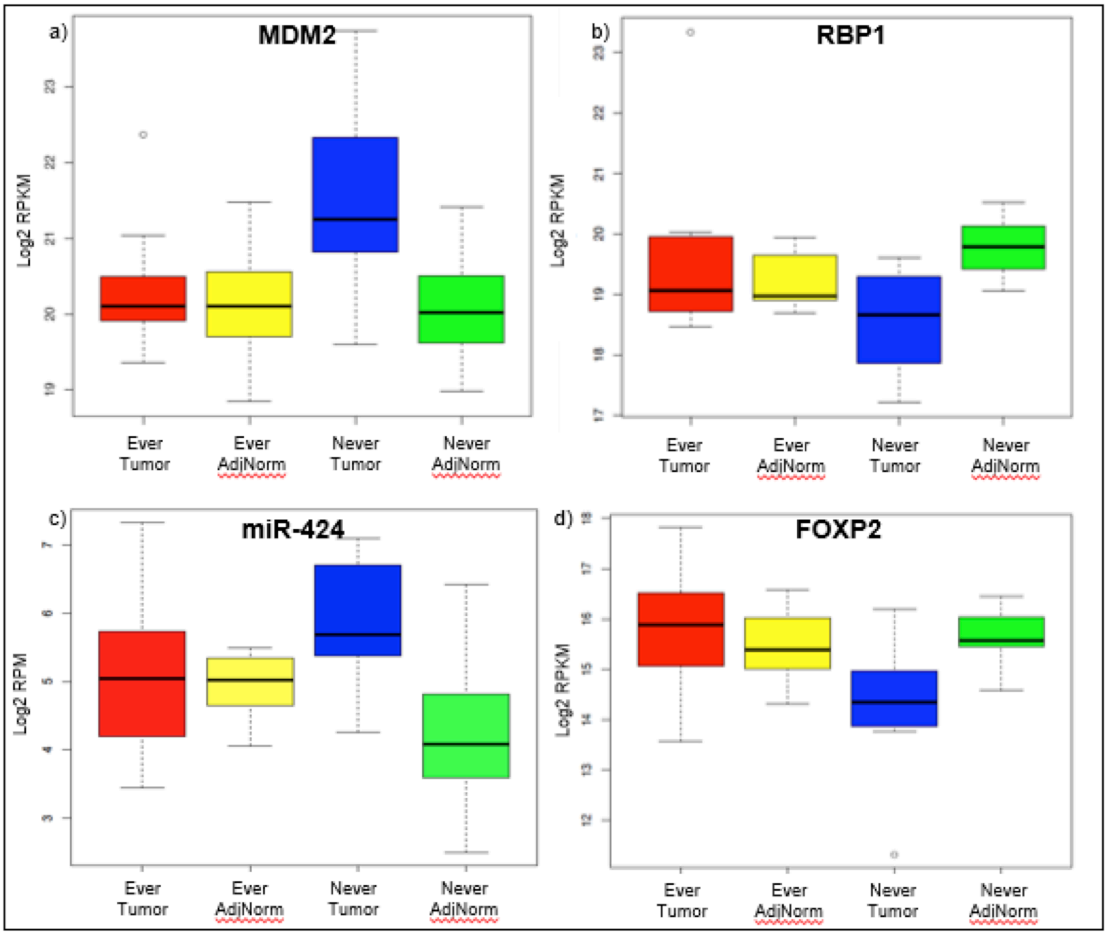
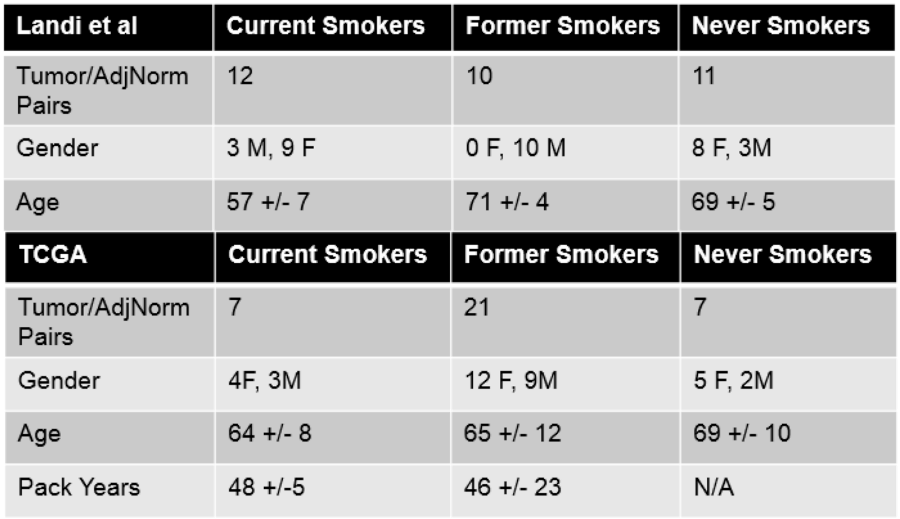


Fig. 12. Never Smoker Tumor Specific Gene Expression Changes.

In order to confirm our findings, we compared our 120 gene signature for never smoker tumor to RNAseq data from TCGA and microarray gene expression data from Landi et al. 2008[84] (Table 5). Both of these computational datasets contain paired tumor and adjacent normal samples from current, former, or never smokers with lung adenocarcinoma. The never smoker tumor specific gene expression signature was split into two gene sets based on direction of change. The activity of the two gene sets was scored per sample across expression data from the independent sequencing and array datasets using GSVA[136]. Running the same linear mixed effects model with an interaction term followed by a post-hoc t-test (only including genes significant in the first step) on the generated GSVA gene set scores yielded statistically significant results, showing that our signature is present in independently generated computational data. Specifically, when the set of genes down in the never smoker tumor were transformed into a single GSVA score per sample using data from Landi et al.[84], the same linear model approach and p-value cutoffs as used in the initial analysis yielded statistically significant results. (lme p-value = 0.01, post-hoc t-test p-value at 0.001 (Fig. 13A). In support of this finding in the Landi data, the set of genes up in the never smoker tumor, when tested in per sample scored TCGA data, had a p-value of 0.045 in the interaction model and 0.03 in the never smoker tumor vs. adjacent normal model (Fig. 13B). These results show that the never smoker tumor specific gene expression signature validated in independently collected and profiled samples.



Landi et al	Current Smokers	Former Smokers	Never Smokers
Tumor/AdjNorm Pairs	12	10	11
Gender	3 M, 9 F	0 F, 10 M	8 F, 3M
Age	57 +/- 7	71 +/- 4	69 +/- 5
TCGA	**Current Smokers**	**Former Smokers**	**Never Smokers**
Tumor/AdjNorm Pairs	7	21	7
Gender	4F, 3M	12 F, 9M	5 F, 2M
Age	64 +/- 8	65 +/- 12	69 +/- 10
Pack Years	48 +/-5	46 +/- 23	N/A

Table 5: Demographics of Independent Gene Expression Array and RNAseq Cohort from lung adenocarcinoma patients.

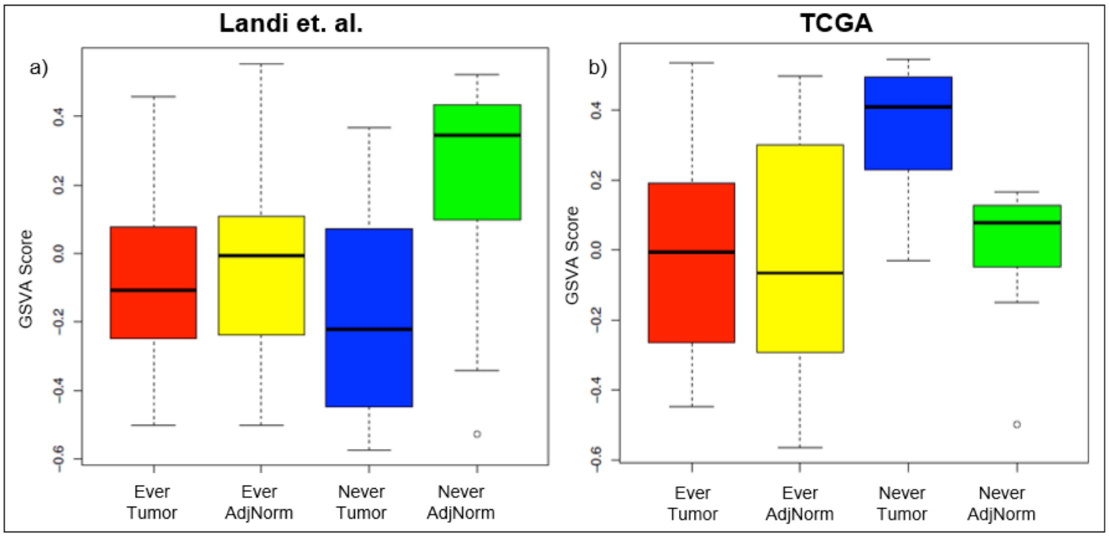


Fig. 13: Never Smoker Tumor Specific Genes Show Enrichment in Independent Gene Expression Array and RNAseq Dataset via GSVA. A) A per sample estimate was calculated for the genes which are down in the never smoker tumor but not the ever smoker tumor as compared to adjacent normal in array data from Landi et al., showing statistically significant enrichment. B) A per sample estimate was calculated for the genes which are up specifically in the never smoker tumor in RNAseq from TCGA, showing statistically significant enrichment.

Functional Enrichment of Never Smoker Specific Tumor Expression

Genes up in the never smoker tumor are enriched for pathways such as GO Epithelial Cell Differentiation, GO negative regulation of immune response, activation of Jun Kinase activity, and Biocarta Multidrug Resistant Proteins (p-value < 0.03). Moreover, several of the signature's genes upregulated in never smoker tumors have been implicated in other cancers. For instance, MDM2[137] and CABIN1[138] are known to inhibit p53 transcriptional activity and MDM2 has been suggested as a target for chemotherapy in ovarian cancer and others[119,139,140]. Additionally, high expression of four never smoker tumor specific genes causes breast cancer oncogenic phenotypes: ABCC3 (Human Multidrug Resistance Protein[141]) confers resistance to chemotherapy[142], PLXNB1 promotes metastasis, FAM83B drives epithelial cell transformation[143], and GPR110 increases anchorage independent growth[144] leading to metastasis. Lastly, PLXNB1[145] and GDF15[146] have also been implicated in ovarian cancer.

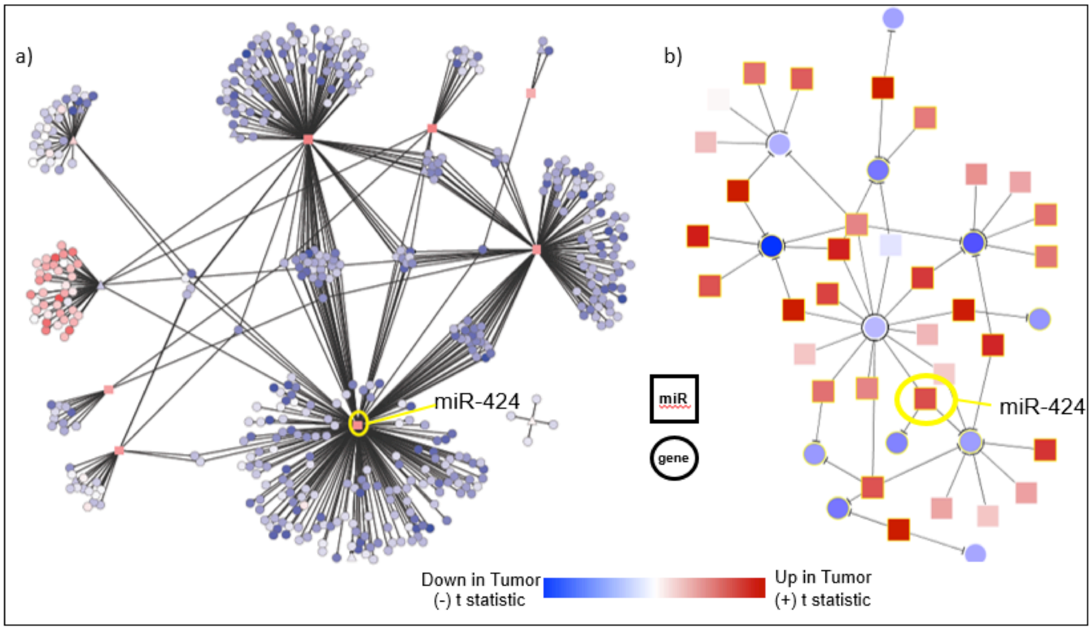


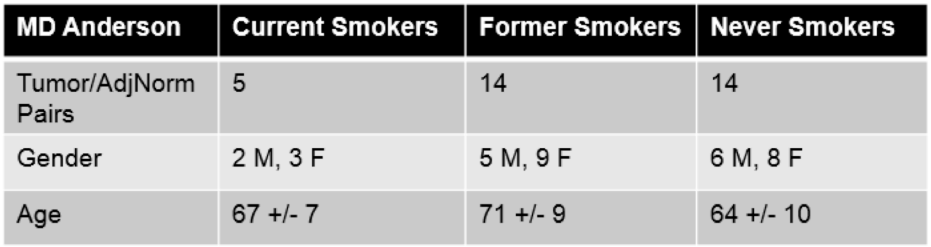
Fig. 14: Two Perspectives on the Never Smoker Tumor Specific Regulatory Network. Mir-424 is circled in yellow. A) All protein coding genes and 15 never smoker tumor specific miRNA were used to create this network B) Only 59 never smoker tumor specific mRNAs and all miRNAs were used to create this network.

Never Smoker Tumor Specific mRNA-miRNA Regulatory Network

To gain insight into the regulation of never smoker lung cancer, we used mirconnX to build a regulatory network two different ways using large RNAseq and miRNAseq data. First, a directed weighted network was constructed using never smoker expression of all protein coding genes and never smoker tumor specific miRNA. The network contains 7 miRNAs perturbing 591 protein coding genes, for a total of 592 interactions (Fig. 14A).

A second network was built including only 59 never smoker tumor specific protein coding genes (subset of 120 genes which have miRNA predictions in mirconnX) and all miRNA. This network contains 33 miRNAs perturbing 12 protein coding genes, for a total of 52 regulatory connections (Fig. 14B). Mir-424 is centrally connected in both of these networks, has previously been implicated as a driver of angiogenesis in other cancers[147–149], and is directly connected to key pulmonary developmental transcription factor of interest FOXP2. Functional enrichment of mRNAs connected to mir-424 (Fig. 14A) reveals KEGG Melanoma, pancreatic cancer, Biocarta RECK pathway, GO anatomical structure development all at pval < 0.01.

PCR Validation of mir-424 and its Target FOXP2 Original Samples and Independent Samples



MD Anderson	Current Smokers	Former Smokers	Never Smokers
Tumor/AdjNorm Pairs	5	14	14
Gender	2 M, 3 F	5 M, 9 F	6 M, 8 F
Age	67 +/- 7	71 +/- 9	64 +/- 10

Table 6: Demographics of Independent Cohort from MD Anderson.

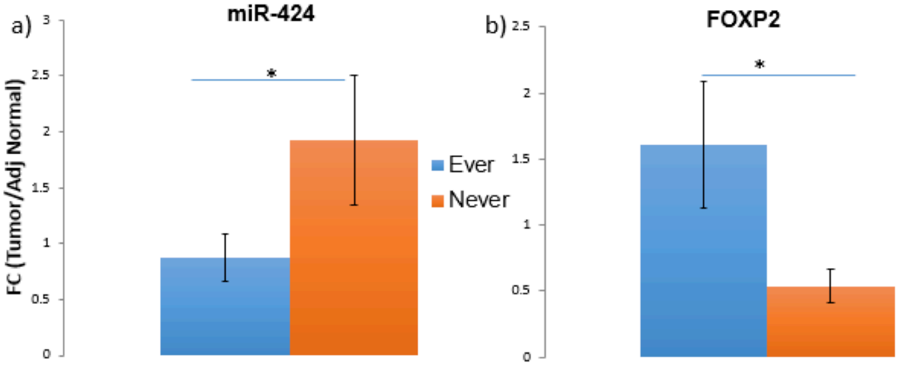


Fig. 15: qPCR Validation of miR-424 and its Target FOXP2. * = pval < 0.05.

Additional independent samples including tumor, adjacent normal, and small airways near tumor were collected from MD Anderson (5 current, 14 former, 14 never smokers, Table 6). Potential key regulatory hub, mir-424 and its transcription target FOXP2 were validated by qRT-PCR in these independent tumor and adjacent normal samples and in the original sample set, analyzed together with pval < 0.05 (Fig. 15) when compared between never and ever smokers. qPCR showed that mir-424 is expressed at a higher level in the small airways near the adenocarcinoma tumor in never smokers than ever smokers (Fig. 16). Together, computational and experimental testing demonstrate that our signature validates both as a whole computationally with GSEA and for specific genes via qPCR in independent groups of samples. This suggest that our signature is robust and will replicate in any future studies with additional independent samples collected, whether the testing is done with RNAseq, microarrays, or qPCR.

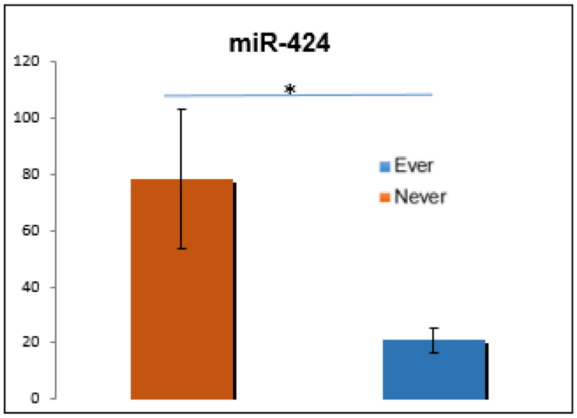


Fig. 16: miR-424 is Upregulated in the Small Airways Near the Tumor in Never Smokers in an Independent Cohort. Expression of mir-424 was queried in the small airways near the tumor of ever and never smokers with lung adenocarcinoma. X-axis is relative expression.

Identification of Never Smoker Adenocarcinoma Therapeutics via the Connectivity Map

Although developing drugs against a certain target is a feasible approach for developing new therapeutics for lung cancer, identifying FDA approved compounds that could be repurposed may have a shorter path to the clinic. The Connectivity map (Cmap) helps investigators identify already existing FDA approved bioactive compounds which may reverse a molecular phenotype of interest. Specifically, the Cmap is a large publicly available compendium of microarray data reflecting gene-expression responses to drug therapy. I leveraged the Cmap to identify compounds that reverse the 120 gene never smoker tumor specific expression pattern. One of the most significant hits is a drug called

Altretamine (p=0.002). This chemotherapeutic is currently in use in the clinic in ovarian cancer[150]. Interestingly, never smoker lung cancer is much more common in females[31], hinting that gender may play a role in never smoker lung cancer development and could influence efficacy of chemotherapeutics. Althgouth Altretamine works as an alkylating agent, through a mechanism similar to mustard gas[151], its exact mechanism of action remains to be characterized thoroughly.

Discussion

Our study represents the most comprehensive profiling of the regulation of never smoker lung cancer to date. Here, we leverage high-throughput high-coverage large and small RNA sequencing of tumor versus paired adjacent noncancerous lung tissues resected from adenocarcinoma patients with varied smoking histories. Building on previous work showing clinical and genomic differences between ever and never smoker lung adenocarcinoma, we have revealed unique changes in the transcriptomic landscape of never smoker lung adenocarcinoma compared to ever smoker lung adenocarcinoma.

Differential expression analysis of large and small RNA from clinical specimens as stratified by smoking status has enabled us to gain unprecedented insight into the regulatory networks underpinning lung carcinogenesis. The large RNA changes discovered in our dataset were significantly related to results generated in two independent gene expression profiling experiments from different laboratories. Interestingly, famous cancer genes such as RBP1 and MDM2 were uncovered by this analysis, suggesting that the set of genes is likely related to oncogenesis in never smokers. In a literature review of all upregulated never smoker tumor specific genes, it can be observed that many of these genes have already been implicated in ovarian and breast cancer. Although this observation is casual rather than statistical, it is interesting to note that never smoker lung cancer occurs at much higher frequency in women, suggesting that further investigation is needed to determine if there are any shared mechanisms. Adding to this observation, the connectivity map suggested ovarian cancer drug Altretamine as a drug to reverse the never smoker tumor specific signature.

Although finding a protein coding drug target can be a desirable outcome, many have suggested that miRNAs may be better drug targets as they regulate entire pathways[152,153]. To explore this possibility, network analysis was performed by integrating select mRNA and miRNA expression data with a network of prior putative targets. This analysis uncovered a mRNA-miRNA regulatory network which regulates gene expression changes unique to the never smoker tumor. Together the large RNA analysis and network analysis support the hypothesis that never smoker lung adenocarcinoma is a disparate disease from ever smoker lung cancers and triggers the need to identify therapeutic targets specific for never smokers.

To address this clinical need, miR-424 is identified as a key hub in the never smoker specific lung adenocarcinoma regulatory network because of its high degree of regulatory connections in the network. miR-424 has mRNA targets that are important in many pathways, specifically in those pathways that would be affected by cancer. This miRNA has been reported to play an oncogenic role in colorectal and pancreatic cancers, and is highly connected in our never smoker tumor regulatory network. Our network predicted that mir-424 suppressed FOXP2, a transcription factor important in lung development. Unique never smoker tumor perturbation of mir-424 and FOXP2 was confirmed using qRT-PCR in an independent set of samples. Preliminary data suggests that this pattern of miR-424 expression extends into the small airways of an independent cohort of never smokers with lung adenocarcinoma, raising the possibility of a field of injury specific to never smokers. At this time, further functional characterization of mir-424 and FOXP2 is needed to understand their potential as therapeutic targets and their roles in the airway epithelium.

The oncogenic, never smoker specific molecular derangements detailed in this study will ultimately contribute to the development and clinical deployment of new therapies for lung adenocarcinoma in never smokers.

Materials and Methods

Patient Demographics

Current, former, and never smokers who had no history of other exposures underwent surgery as part of their treatment, at which point samples were collected. At the time of collection, current smokers had a higher pack year burden than former smokers ($p<0.05$). The three groups were well balanced for age with a mean of 55 years for each group. There was an insignificant but higher ratio of females in the never smoker group. RIN (RNA Integrity Number) differed between all three groups ($p<0.05$), but was not found to be associated with results and thus not confounding.

Sample Collection and Processing

Through a collaboration with Lovelace Respiratory Research Institute, (LRRI), samples were collected from patients undergoing tumor resection at the Mayo clinic. Collected samples were sent to LRRI, where RNA was isolated and then shipped to Boston University.

Library preparation was done using Illumina's TruSeq (RNAseq) sample preparation kit starting with 200-500 ng of total RNA from each sample. The large RNA was isolated using poly-A selection and fragmented to get a range of fragment lengths centered around 200 nucleotides. Fragments were randomly primed for reverse transcription followed by first and second-strand synthesis to create double-stranded cDNA fragments. cDNA ends were repaired, ligated to a unique barcoded index paired end adapter. These fragments were then PCR amplified, purified, and subjected to cluster generation on a cBot machine using Illumina TruSeq Paired-End Cluster Generation Kits. Next, the samples were sequenced four per lane on a HiSeq machine. Sequencing generated approximately 40 million 99 nucleotide paired end reads with an average inner distance of -25 nucleotides.

The small RNA fraction (fewer than 200 nucleotides) was isolated and then 200ng was processed using the Illumina TruSeq Small RNA Sample Prep Kit. Samples were multiplexed and sequenced on the Illumina HiSeq 2000 generating 35-bp reads. Up to 10 samples were pooled per lane obtaining an average of approximately seven million reads per sample.

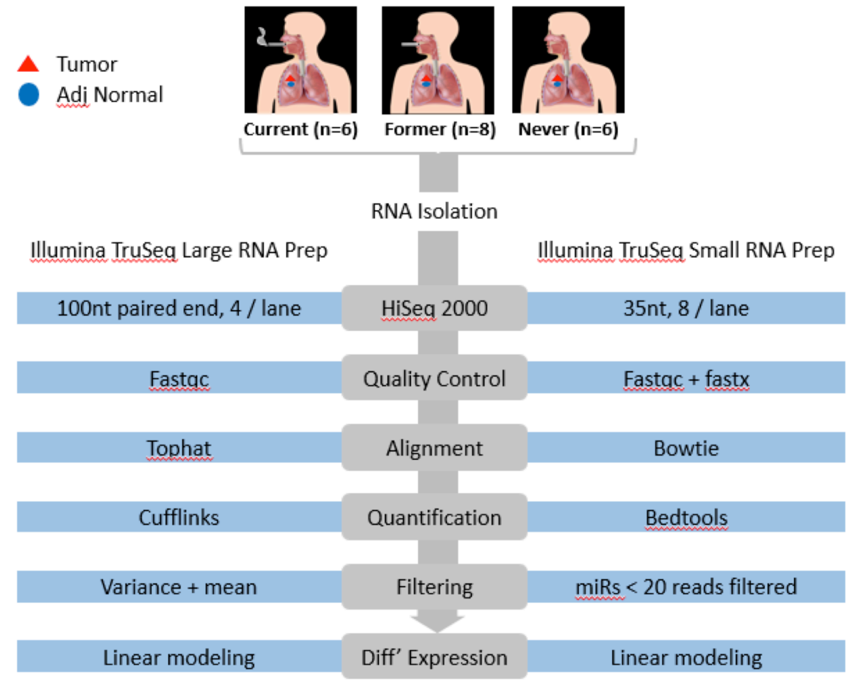


Fig. 17: Processing of RNAseq Data.

Processing of Large RNA and miRNAseq Data

mRNA and miRNA fastq files were initially filtered for quality using fastqc. mRNA reads were aligned to hg19 using Tophat[101], and quantified to Ensembl using Cufflinks[102]. After adapter trimming with the fastx toolkit, miRNA reads were aligned with Bowtie[101], counted with bedtools[154], and RPM normalized. QC metrics were reviewed, such as alignment statistics and PCA. One sample and its pair were excluded for low QC metrics and PCA outlier status. The bottom 30% of genes by mean FPKM and the bottom 30% by variance were removed from further analysis. miRNAs with an average count below 20 were removed. These filtering methods were employed to avoid testing genes ineligible for linear modeling. An overview of sequencing processing methods can be seen in Fig. 17.

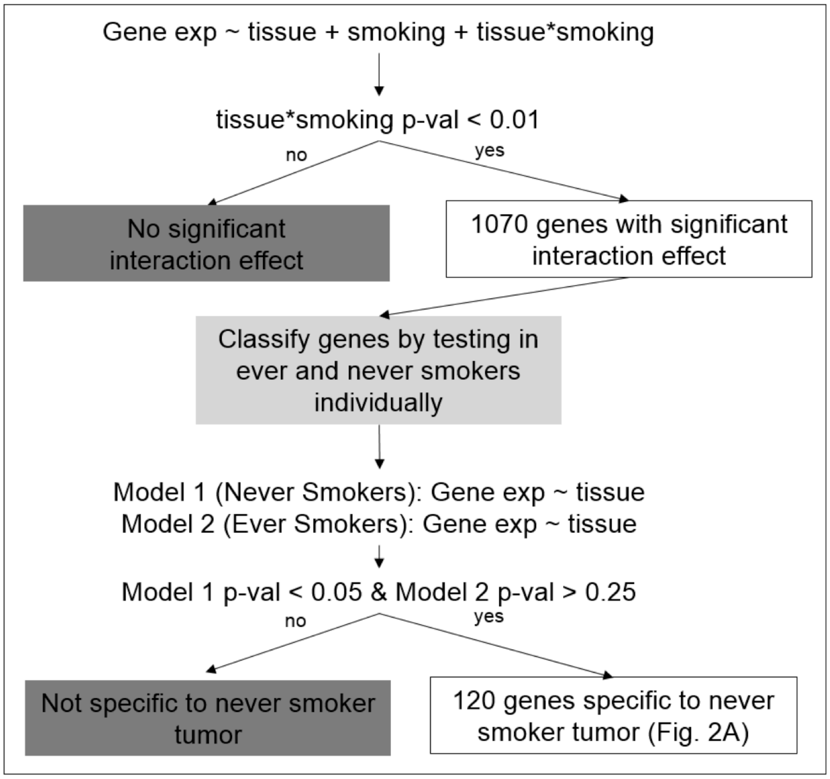

Gene exp ~ tissue + smoking + tissue*smoking
tissue*smoking p-val < 0.01
no
yes
No significant interaction effect
1070 genes with significant interaction effect
Classify genes by testing in ever and never smokers individually
Model 1 (Never Smokers): Gene exp ~ tissue
Model 2 (Ever Smokers): Gene exp ~ tissue
Model 1 p-val < 0.05 & Model 2 p-val > 0.25
no
yes
Not specific to never smoker tumor
120 genes specific to never smoker tumor (Fig. 2A)

Fig. 18. Linear Modeling to Find Never Smoker Tumor Specific Gene Expression Changes.

Linear Modeling to Find Unique Tumor Genes and miRNAs by Smoking Status

Model 1 (RNA = tissue + smoking, random=~1|patient) accounts for tissue, smoking status, and patient. ‘Tissue’ is a fixed effect controlling for the histology and site of the sample (tumor or adjacent-normal), and ‘Smoking’ is a fixed effect controlling for the smoking status of the patient (ever or never). ‘Patient’ is a random effect controlling for patient specific effects. Model 2 (RNA = tissue + smoking + tissue:smoking, random=~1|patient) contains an additional interaction effect between histological status of the tissue sample and smoking status of the patient. The two models were compared by a likelihood ratio test and those genes with a p value of less than 0.01 were determined to be associated with the interaction between tissue type and smoking status. To determine the direction of change of these genes, a post-hoc paired t-test was done separately in ever and never smokers between tumor and adjacent normal tissue, and identified never smoker tumor specific changes in gene expression. To be considered a never smoker tumor specific gene, the never smoker t-test had to have a p-value of less than 0.05 and the ever smoker t-test had to have a p-value of greater than 0.25. This would indicate that the gene was significantly differentially expressed between tumor and adjacent normal in never smokers but not significantly differentially expressed in ever smokers. An overview of this approach can be seen in Fig. 18.

The resulting signature was validated using GSVA. Microarray expression data was downloaded from GEO, and RPM normalized RNAseq data was downloaded from TCGA portal. Patients with the smoking status "reformed smokers quit >= 15 years" were excluded from TCGA dataset. The never smoker tumor specific gene expression signature was split into groups by direction. These two gene sets were projected into Landi et al. and TCGA expression data. The same linear models as above were then run on these independent datasets and tested for significance.

Pathway Identification

EnrichR was used to determine pathway enrichment. First, the signature was split by up-in-the-never-smoker-tumor or down-in-the-never-smoker-tumor. These two gene lists were uploaded separately to enrichR. Scores for each gene were not included.

mRNA-miRNA Construction with miRconnX

MirconnX was used to build an integrated anticorrelation prior information network, leveraging expression values from large RNA and miRNAseq as well as miRNA target prediction information. The network was constructed by submitting the expression data for only the samples which had both mRNA and miRNAseq data. Gene expression for all protein coding genes that passed the filter and the 15 significant miRNA were submitted to mirconnX, and then a second network was created by submitting only significant protein coding genes and all miRNA.

qRT-pcr Validation of mir-424

Under IRB approval, tumor, adjacent normal, and small airway brushings were obtained from ever and never smokers undergoing tumor resection surgery at MD Anderson. RNA was isolated and shipped to Boston University. MiR-424 expression was measured by qRT-PCR in 32 ever smoker and 20 never smoker paired lung adenocarcinoma and adjacent normal tissue. Samples were analyzed with qRT-PCR using Taqman assays and RNU44 as a control. FOXP2 was measured in the same samples with qRT-PCR using Qiagen RT^2 Primer Assay and UBC as a control. Fold change for both miR-424 and FOXP2 was measured by dividing the relative expression of the tumor with the relative expression of the adjacent normal in each matching pair. Additionally, MiR-424 expression was measured by qRT-PCR in 19 ever smoker and 14 never smoker small airway brushing samples (collected near the tumor). Samples were analyzed with qRT-PCR using Taqman assays and RNU44 as a control.

Conclusions and Future Directions

Impact of Shared Emphysema-IPF Gene Expression and Regulatory Network

Emphysema and IPF are both progressive diseases with a dearth of therapeutic options to address the underlying disease mechanism. Funding has been historically low for these diseases, especially when contrasted with incidence and mortality rate. Clinicians lack not only effective treatment options, but also adequate tools to determine the rate at which patients will decline. This is ultimately because of a lack of understanding of molecular mechanisms of the diseases.

Although other publications have speculated about the existence of shared pathways, this study, by profiling COPD and IPF together, conclusively demonstrated that these two diseases have common pathways. To an outside observer this may be surprising as COPD and IPF have different clinical presentations. COPD appears to be more of a disease of wasting whereas IPF is characterized by deleterious fibrosis. However, both diseases have an overlapping risk associated with them: cigarette smoking. Based on this, one possible hypothesis is that gene expression changes in common are more likely to be causative than reactionary.

The major impact of this discovery of shared gene expression on the field is the presentation of a pathway and a miRNA as potential therapeutic targets. The hypoxia component of the p53 pathway was found to be upregulated in both emphysema and IPF lung tissue. Fortuitously, this is a pathway that is already well characterized due to its prominent role in tumorigenesis and oncogenesis in malignant disease. Drugs that can perturb members of this pathway already exist, and a subset is FDA approved. Hopefully this means that existing drugs can be pivoted rather than having to develop drugs de novo. Interestingly, a patient diagnosed with COPD is more likely to develop lung cancer, and vise-versa. Since COPD is a risk factor for lung cancer and the p53 pathway has been implicated in both diseases, one can infer that targeting this pathway could be used in a chemoprevention and/or COPD-prevention setting.

Secondly, although miRNA targeting is a much newer field, several clinical trials demonstrate feasibility in a human disease setting[155]. MiR-96 shows promise for both diseases, but especially for IPF. By miRNA arrays, miR-96 is observed as upregulated in IPF. When miR-96 is overexpressed in cell lines, gene expression is perturbed in a way that resembles IPF gene expression. Thus, targeting miR-96 for destruction may help to reverse the IPF phenotype. Further studies are needed to determine if miR-96 could also be a suitable target for emphysema

Forward progress in the COPD and IPF research space has also been limited by disease heterogeneity. Specifically, part of what makes emphysema difficult to treat is variability of severity within a patient and/or between patients. This idea of heterogeneity is supported by the current study, which shows that even between highly differentially expressed genes within emphysema vs. control lungs, there are some minor gene expression differences between at least two clusters of emphysema lungs (Fig. 3). Analysis of associated clinical variables shows that one group may trend toward being "less healthy" than the other by select variables but this trend is not significant.

Impact of Never Smoker Tumor Specific Gene Expression Signature

Although lung cancer has for a long time been and is currently the leading cause of cancer mortality in the United States, funding in this area has been chronically limited. Treatment options as well as a fundamental understanding of the molecular progression of this disease are limited to date. Moreover, some genomic evidence exists to suggest that there may be major molecular differences between ever and never smokers with lung cancer. Despite the potential for these differences, ever and never smokers currently receive similar clinical treatment. Results from this thesis add to the hypothesis that never smokers with lung adenocarcinoma could be considered as a distinct disease group instead of being included with ever smokers. First, this work uncovered unique changes in gene expression in the never smoker tumor. These transcriptomic modifications represent not only differences in gene expression, but also changes in pathway usage that may be driving the development of the tumor. Moreover, there were unique changes in miRNA expression in the never smoker tumor. This is particularly relevant in terms of target discovery, because miRNA have the potential to regulate entire pathways. Using directed integrative networking techniques, my analysis uncovered the potential never smoker tumor miRNA-mRNA regulatory network. While many arms of this network could possibly be targeted for chemotherapy or chemoprevention, we initially focused on miR-424 because it has been implicated in many other cancers. Initial qPCR validation confirmed that miR-424 and its predicted target FOXP2 are perturbed in an independent sample set, setting the stage for further validation of the mir-424-FOXP2 regulatory link as a potentially drugable interaction. Additionally, qPCR characterization of miR-424 expression unveiled higher expression in the small airways near the adenocarcinoma tumor in never smoker as compared to ever smokers. Although preliminary, oncomiR expression in the small airways near the tumor implies that never smokers with lung cancer may have a field of injury like their ever smoking counterparts. If this hypothesis is true, airway gene expression could hold the potential to diagnose lung cancer in this cohort. While additional functional genomic studies are needed to further characterize miR-424 and its interaction with FOXP2, this regulatory event could be part of the next wave of chemotherapeutics or diagnostics for never smoker lung cancer.

In addition, chemotherapeutic options in the never smoker lung cancer space are still quite limited. The process of moving from an identified target to a clinically useful drug is very long and costly. A potentially more rapid approach is to pinpoint already FDA approved bioactive drugs and reposition them in the context of never smoker lung cancer. Following this logic, I identified Altretamine as a compound which is able to reverse the never smoker tumor specific signature. Altretamine is a chemotherapeutic already in use for the treatment of ovarian cancer. Interestingly, a casual survey of the literature on the never smoker tumor specific signature genes revealed that many have been previously implicated in ovarian and breast cancer. Although further functional genomic studies are needed, Altretamine may represent a chemotherapeutic that could be repurposed to treat never smokers with lung cancer.

Limitations

While this study has provided a thorough investigation of the COPD, IPF, and lung cancer transcriptome, there are a number of limitations. Firstly, analysis of changes in gene expression associated with IPF and COPD only included seventy-five samples, a subset of all samples sequenced in total. It is well known that IPF and especially COPD are highly heterogeneous diseases, and the number of samples represented in this study may be insufficient to fully capture this disease diversity within patient groups. Some of this heterogeneity can even be observed in the heatmap (Fig. 3) of significant emphysema vs control genes. Here, it is clear that there are two subgroups within emphysema. Although not statistically significant, it does appear that one group appears to be more "ill" than the other, with shorter 6 minute walk distances and worse scores on questionnaires. By analyzing more samples, it would be possible to test if this trend is real or a mirage. Moreover, there are major cell type differences between emphysema, IPF, and healthy lung tissue. Although the expression of several genes was localized to the airway epithelium using IHC, the cell type of expression of the rest of the signature remains in question. This problem could be addressed using laser microdissection of whole tissue samples to zoom in on airway epithelium or single-cell next generation sequencing.

One other challenge in my study of emphysema in particular is that the variable used to measure emphysema severity may not correlate perfectly with severity in the exact location sampled and profiled. In this study "percent emphysema" was used, which is a representation of the function of the whole lung, not only the section of the tissue sampled. In a patient with emphysema, alveolar destruction is variable and thus some samples may have more apoptotic cells than others. This could be solved in future studies by collecting better annotation on the degree of alveolar destruction severity in the area from where the sample is being collected.

Lastly, the major limitation preventing the findings of my work in emphysema and IPF from having greater impact is the lack of functional validation. While the overexpression of miR-96 was profiled with gene expression in cell lines, the effect on cell phenotype was never tested.   Moreover, there exist a number of potential drugs to target the p53/hypoxia pathway, none of which were tested out in the relevant cell lines in this study.

Like the work in emphysema and IPF, the study of never smoker lung cancer is also limited by low sample numbers. Discovery of the never smoker tumor specific genes in this study relied on ever smokers as a negative control. In order for a gene to be included in the “never smoker tumor specific” category, it had to not be statistically significant in a comparison of the tumor and adjacent normal in ever smokers. There are a number of reasons why a gene can have an insignificant p-value in a statistical test. My choices of statistical model and cutoff assume that a gene is not detected as changing because it is truly not different between the ever smoker tumor and adjacent normal. However, it is possible that there are not enough patients in the study to detect the change, or my choice of statistical test is not robust enough to detect a change. While boosting the sample number would partially address this issue, there is no perfect solution. The issue of how to prove a lack of connection between a phenotype and an effect is an unsolved challenge not only in gene expression studies, but also in other fields such as epidemiology.

The never smoker tumor signature provides a tantalizing peek into the world of large noncoding RNA. Although the original intention of the presented analysis was not focused on this breed of RNA, about half of the signal is noncoding. When performing RNA sequencing, it is necessary to perform at least one step to avoid sequencing ribosomal RNA, which dominates the cellular RNA pool and varies very little between cells, conditions, and individuals. In order to avoid sequencing ribosomal RNA, we employed a poly-A selection. However, not all large regulatory RNAs are polyadenylated. Although the signature contains a large percentage of noncoding RNA, it may be providing only a glimpse into critical regulatory circuitry of the never smoker tumor. Furthermore, many long noncoding RNAs are antisense to annotated genes. The RNA-sequencing methods used in these studies were not strand-specific so there are likely antisense transcripts that could not be identified unless a stranded RNA-seq protocol is utilized. The present study also did not probe the relationship between never smoker tumor lncRNAs and protein coding genes, which could be used to draw hypothesis about the functional role of lncRNAs.

Next Steps

One of the overall impacts of this study as a whole is to underscore the importance of noncoding RNA. In COPD, IPF, and never smoker lung cancer a miRNA-mRNA regulatory network was revealed. In order to better characterize the full transcriptional potential and regulatory networks of these diseases, whole transcriptome sequencing should be employed in the future. Also known as “total RNA” sequencing, this methodology enables detection not only of large RNAs such as protein coding genes, but also profiles lncRNAs, pseudogenes, and RNAs which use up miRNAs (sponge RNAs)[99]. In addition, strand specific RNAseq could provide better resolution for the alignment of certain antisense transcripts, which sometimes can hide in a number of genomic locations, such as in an intron of a protein coding gene, and will be lost without strand information for alignment. Moreover, a number of very recent studies lately have suggested that RNAs may form circles after transcription which play a regulatory "sponge RNA" like role[156–158]. These circular RNAs do not have poly-A tails, and will thus not be included in poly-A RNA sequencing. Although further investigations are needed to support this claim, circular RNAs may in the future also prove to be targetable for the treatment of diseases such as emphysema, IPF, and lung cancer. It is technically possible to sequence circular RNAs with total RNA sequencing, but the best approach is to utilize special enzymes such as RNAseH in library preparation to nick the circularized RNA[159].

Furthermore, sequencing of additional sample types would yield additional information about disease biology. COPD, IPF, lung cancer, and healthy lung all have very different cell types. Thus, it would be helpful to use laser capture microdissection to select for only one cell type. Gene expression analysis would be much less likely to be confounded when comparing only one cell type, although heat generated during the laser capture can selectively degrade some RNAs. Single cell sequencing would also be useful in avoiding cell type heterogeneity.

In the lung cancer space, studies have revealed that tumors are rarely a subclonal population, and that there can be heterogeneity within a tumor in terms of presence or absence of a mutation. It is possible that while collecting a sample from a tumor, the area sampled may not be an accurate representation of the tumor as a whole. In the future, this issue would be best addressed with multiple fine needle biopsies of the tumor, which should all undergo transcriptional profiling. By comparing gene expression within the tissue, it would be possible to assess the effects of tumor heterogeneity and correct for it before comparing tumors or comparing to other non-cancerous tissues. Lastly, sequencing bronchial airway epithelium from patients with and without emphysema, IPF, or never smoker lung cancer would provide a tantalizing view into the reaction of the airway epithelium to the presence of disease.

RNA seq data can be analyzed from many different angles, and in this case analysis as a whole of the emphysema, IPF, and lung cancer RNAseq data is not “complete”. Since analysis was done on seventy-five IPF, COPD, and control samples, hundreds more samples from this group have been sequenced. The analysis described in this document has yet to be extended into this much larger cohort of samples. The larger sample pool is much more powered for analyses such as disease subclass discovery. In addition, including more samples empowers us to ask more questions regarding the clinical data, such as searching for gene expression changes associated with percent emphysema, FEV1/FVC ratio, or other clinical variables. The total pool of lung cancer samples sequenced also includes samples from other lung cancer subtypes, including squameous cell carcinoma and broncheoalveolar carcinoma. With this larger subset many analyses are possible, such as comparing the adjacent normal between smokers with adenocarcinoma and smokers with squameous cell carcinoma to determine if the local tumor microenvironment is affected by cancer subtype. Additionally, the lung cancer data was only mined for changes in gene expression. To date, no analysis has been done to test for never smoker tumor specific changes in splicing. Since RNA sequencing is an unbiased platform, it is also possible to test the data for mutations, such as gene fusions and indels. In summary, many future directions remain to be explored both within existing data and in potential future yet to be generated data.

BIBLIOGRAPHY

1. American Cancer Society. Cancer Facts & Figures 2014. (2014).
2. Murphy SL, Xu J & Kochanek KD. Deaths: Preliminary Data for 2012. *National Vital Statistics Reports* **60,** (2012).
3. Mannino, D. M., Homa, D. M., Akinbami, L. J., Ford, E. S. & Redd, S. C. Chronic Obstructive Pulmonary Disease Surveillance --- United States, 1971--2000. *Respiratory Care* **76,** 1184–1199 (2002).
4. Murphy, S., Xu, J. & Kochanek, K. Deaths: Preliminary data for 2010. *National Vital Statistics Reports* **60,** (2012).
5. Olson, A. L. *et al.* Mortality from Pulmonary Fibrosis Increased in the United States from 1992 to 2003. *American Journal of Respiratory and Critical Care Medicine* **176,** 277–284 (2007).
6. Howlader N, Noone AM, Krapcho M, Garshell J, Neyman N, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Cho H, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). PDF Version - SEER Cancer Statistics Review (CSR), 1975-2010. *SEER Cancer Statistics Review, 1975-2010, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2010/, based on November 2012 SEER data submission, posted to the SEER web site, April 2013.* at <http://seer.cancer.gov/csr/1975_2010/sections.html>
7. Oh, C. K., Murray, L. A. & Molfino, N. A. Smoking and Idiopathic Pulmonary Fibrosis. *Pulmonary Medicine* **2012,** e808260 (2012).

8. Perez-Padilla, R., Schilmann, A. & Riojas-Rodriguez, H. Respiratory health effects of indoor air pollution [Review article]. *International Journal of Tuberculosis and Lung Disease* **14,** 1079–1086 (2010).

9. Sethi, J. M. & Rochester, C. L. Smoking and chronic obstructive pulmonary disease. *Clinics in Chest Medicine* **21,** 67–86 (2000).

10. Brawley, O. W., Glynn, T. J., Khuri, F. R., Wender, R. C. & Seffrin, J. R. The first surgeon general's report on smoking and health: The 50th anniversary. *CA: A Cancer Journal for Clinicians* **64,** 5–8 (2014).

11. Dept. of Health and Human Services, Centers for Disease Control and P. and H. P. *The Health Consequences of Smoking: a report of the Surgeon General*. (2004). at <http://www.cdc.gov/tobacco/sgr/sgr_2004/chapters.htm>

12. Pauwels, R. A., Buist, A. S., Calverley, P. M. A., Jenkins, C. R. & Hurd, S. S. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop Summary. *American Journal of Respiratory and Critical Care Medicine* **163,** 1256–1276 (2001).

13. Vestbo, J., Hurd, S. S. & Rodriguez-Roisin, R. The 2011 revision of the global strategy for the diagnosis, management and prevention of COPD (GOLD) – why and what? *Clinical Respiratory Journal* **6,** 208–214 (2012).

14. Lee, S.-H. *et al.* Antielastin autoimmunity in tobacco smoking–induced emphysema. *Nature Medicine* **13,** 567–569 (2007).

15. Merchant, R. K., Schwartz, D. A., Helmers, R. A., Dayton, C. S. & Hunninghake, G. W. Bronchoalveolar Lavage Cellularity: The Distribution in Normal Volunteers. *American Review of Respiratory Disease* **146,** 448–453 (1992).

16. Niewoehner, D. E., Kleinerman, J. & Rice, D. B. Pathologic Changes in the Peripheral Airways of Young Cigarette Smokers. *New England Journal of Medicine* **291,** 755–758 (1974).

17. Grumelli, S. *et al.* An Immune Basis for Lung Parenchymal Destruction in Chronic Obstructive Pulmonary Disease and Emphysema. *PLoS Medicine* **1,** e8 (2004).

18. Hautamaki, R. D., Kobayashi, D. K., Senior, R. M. & Shapiro, S. D. Requirement for Macrophage Elastase for Cigarette Smoke-Induced Emphysema in Mice. *Science* **277,** 2002–2004 (1997).

19. Maeno, T. *et al.* CD8+ T Cells Are Required for Inflammation and Destruction in Cigarette Smoke-Induced Emphysema in Mice. *Journal of Immunology* **178,** 8090–8096 (2007).

20. Kasahara, Y. *et al.* Inhibition of VEGF receptors causes lung cell apoptosis and emphysema. *Journal of Clinical Investigation* **106,** 1311–1319 (2000).

21. Chilosi, M., Poletti, V. & Rossi, A. The pathogenesis of COPD and IPF: Distinct horns of the same devil? *Respiratory Research* **13,** 1–9 (2012).

22. Faner, R., Rojas, M., MacNee, W. & Agustí, A. Abnormal Lung Aging in Chronic Obstructive Pulmonary Disease and Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* **186,** 306–313 (2012).

23. Garantziotis, S. & Schwartz, D. Host-Environment Interactions in Pulmonary Fibrosis. *Seminars in Respiratory and Critical Care Medicine* **27,** 574–580 (2006).

24. King Jr, T. E., Pardo, A. & Selman, M. Idiopathic pulmonary fibrosis. *Lancet* **378,** 1949–1961 (2011).

25. Macneal, K. & Schwartz, D. A. The Genetic and Environmental Causes of Pulmonary Fibrosis. *Proceedings of the American Thoracic Society* **9,** 120–125 (2012).

26. Geiser, T. Idiopathic Pulmonary Fibrosis - a disorder of alveolar wound repair? *Swiss Medical Weekly* **133,** 405–411 (2003).

27. Selman, M., King, J., Talmadge E. & Pardo, A. Idiopathic Pulmonary Fibrosis: Prevailing and Evolving Hypotheses about Its Pathogenesis and Implications for Therapy. *Annals of Internal Medicine* **134,** 136–151 (2001).

28. Sisson, T. H. *et al.* Targeted Injury of Type II Alveolar Epithelial Cells Induces Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* **181,** 254–263 (2010).

29. Raghu, G. *et al.* 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,** 788–824 (2011).

30. Cottin, V. *et al.* Pulmonary hypertension in patients with combined pulmonary fibrosis and emphysema syndrome. *European Respiratory Journal* **35,** 105–111 (2010).

31. Couraud, S., Zalcman, G., Milleron, B., Morin, F. & Souquet, P.-J. Lung cancer in never smokers – A review. *European Journal of Cancer* **48,** 1299–1311 (2012).

32. Devesa, S. S., Bray, F., Vizcaino, A. P. & Parkin, D. M. International lung cancer trends by histologic type: Male:Female differences diminishing and adenocarcinoma rates rising. *International Journal of Cancer* **117,** 294–299 (2005).

33. Trédaniel, J., Boffetta, P., Saracci, R. & Hirsch, A. Non-smoker lung cancer deaths attributable to exposure to spouse's environmental tobacco smoke. *International Journal of Epidemiology* **26,** 939–944 (1997).

34. Vineis, P. *et al.* Tobacco and Cancer: Recent Epidemiological Evidence. *JNCI Journal of the National Cancer Institute* **96,** 99–106 (2004).

35. Vineis, P. *et al.* Environmental tobacco smoke and risk of respiratory cancer and chronic obstructive pulmonary disease in former smokers and never smokers in the EPIC prospective study. *BMJ: British Medical Journal* **330,** 277 (2005).

36. Wald, N. J., Nanchahal, K., Thompson, S. G. & Cuckle, H. S. Does breathing other people's tobacco smoke cause lung cancer? *British Medical Journal (Clinical research ed.)* **293,** 1217 (1986).

37. Yu, I. T. S., Chiu, Y., Au, J. S. K., Wong, T. & Tang, J. Dose-Response Relationship between Cooking Fumes Exposures and Lung Cancer among Chinese Nonsmoking Women. *Cancer Research* **66,** 4961–4967 (2006).

38. Ko, Y. C. *et al.* Risk factors for primary lung cancer among non-smoking women in Taiwan. *International Journal of Epidemiology* **26,** 24–31 (1997).

39. Koo, L. C. & Ho, J. H.-C. Diet as a confounder of the association between air pollution and female lung cancer: Hong Kong studies on exposures to environmental tobacco smoke, incense, and cooking fumes as examples. *Lung Cancer* **14, Supplement 1,** S47–S61 (1996).

40. Selikoff IJ, Churg J & Hammond E. ASbestos exposure and neoplasia. *JAMA: The Journal of the American Medical Association* **188,** 22–26 (1964).

41. Carbone, M., Kratzke, R. A. & Testa, J. R. The pathogenesis of mesothelioma. *Seminars in Oncology* **29,** 2–17 (2002).

42. Mossman, B. T., Bignon, J., Corn, M., Seaton, A. & Gee, J. B. Asbestos: scientific developments and implications for public policy. *Science* **247,** 294–301 (1990).

43. Loon, A. J. van *et al.* Occupational exposure to carcinogens and risk of lung cancer: results from The Netherlands cohort study. *Occupational and Environmental Medicine* **54,** 817–824 (1997).

44. Samet, J. M. Radon and Lung Cancer. *JNCI: Journal of the National Cancer Institute* **81,** 745–758 (1989).

45. Belinsky, S. A. *et al.* Plutonium targets the p16 gene for inactivation by promoter hypermethylation in human lung adenocarcinoma. *Carcinogenesis* **25,** 1063–1067 (2004).

46. Canver, C.C., Memoli, V.A., Vanderveer, P.L., Dingivan, C.A. & Mentzer Jr, R.M. Sex hormone receptors in non-small-cell lung cancer in human beings. *Journal of Thoracic and Cardiovascular Surgery* **108,** 153–157 (1994).

47. Beattie, C. W., Hansen, N. W. & Thomas, P. A. Steroid Receptors in Human Lung Cancer. *Cancer Research* **45,** 4206–4214 (1985).

48. Omoto, Y. *et al.* Expression, Function, and Clinical Implications of the Estrogen Receptor β in Human Lung Cancers. *Biochemical and Biophysical Research Communications* **285,** 340–347 (2001).

49. Fasco, M. J., Hurteau, G. J. & Spivack, S. D. Gender-dependent expression of alpha and beta estrogen receptors in human nontumor and tumor lung tissue. *Molecular and Cellular Endocrinology* **188,** 125–140 (2002).

50. Hershberger, P. A. *et al.* Regulation of Endogenous Gene Expression in Human Non–Small Cell Lung Cancer Cells by Estrogen Receptor Ligands. *Cancer Research* **65,** 1598–1605 (2005).

51. Mizushima, Y. & Kobayashi, M. Clinical characteristics of synchronous multiple lung cancer associated with idiopathic pulmonary fibrosis : A review of Japanese cases. *Chest* **108,** 1272–1277 (1995).

52. Turner-Warwick, M., Lebowitz, M., Burrows, B. & Johnson, A. Cryptogenic fibrosing alveolitis and lung cancer. *Thorax* **35,** 496–499 (1980).

53. Kawai, T., Yakumaru, K., Suzuki, M. & Kageyama, K. Diffuse Interstitial Pulmonary Fibrosis and Lung Cancer. *Pathology International* **37,** 11–19 (1987).

54. Wu, A. H. *et al.* Family History of Cancer and Risk of Lung Cancer among Lifetime Nonsmoking Women in the United States. *American Journal of Epidemiology* **143,** 535–542 (1996).

55. Schwartz, A. G., Yang, P. & Swanson, G. M. Familial Risk of Lung Cancer among Nonsmokers and Their Relatives. *American Journal of Epidemiology* **144,** 554–562 (1996).

56. Subramanian, J. & Govindan, R. Molecular genetics of lung cancer in people who have never smoked. *Lancet Oncology* **9,** 676–682 (2008).

57. Sun, S., Schiller, J. H. & Gazdar, A. F. Lung cancer in never smokers--a different disease. *Nature Reviews. Cancer* **7,** 778–790 (2007).

58. Powell, C. A. *et al.* Gene Expression in Lung Adenocarcinomas of Smokers and Nonsmokers. *American Journal of Respiratory Cell and Molecular Biology* **29,** 157–162 (2003).

59. Huncharek, M., Muscat, J. & Geschwind, J.-F. K-ras oncogene mutation as a prognostic marker in non-small cell lung cancer: a combined analysis of 881 cases. *Carcinogenesis* **20,** 1507–1510 (1999).

60. Rodenhuis, S. ras and human tumors. *Seminars in Cancer Biology* **3,** 241–247 (1992).

61. Ahrendt, S. A. *et al.* Cigarette smoking is strongly associated with mutation of the K-ras gene in patients with primary adenocarcinoma of the lung. *Cancer* **92,** 1525–1530 (2001).

62. Wong, D. W.-S. *et al.* The EML4-ALK fusion gene is involved in various histologic types of lung cancers from nonsmokers with wild-type EGFR and KRAS. *Cancer* **115,** 1723–1733 (2009).

63. Lim, S.-T. *et al.* Gefitinib is more effective in never-smokers with non-small-cell lung cancer: experience among Asian patients. *British Journal of Cancer* **93,** 23–28 (2005).

64. Mok, T. S. *et al.* Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *New England Journal of Medicine* **361,** 947–957 (2009).

65. Lynch, T. J. *et al.* Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib. *New England Journal of Medicine* **350,** 2129–2139 (2004).

66. Lipson, D. *et al.* Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. *Nature Medicine* **18,** 382–384 (2012).

67. Soda, M. *et al.* Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. *Nature* **448,** 561–566 (2007).

68. Wong, M. P. *et al.* Chromosomal aberrations of primary lung adenocarcinomas in nonsmokers. *Cancer* **97,** 1263–1270 (2003).

69. Wang, Y.-C. *et al.* Inactivation of hMLH1 and hMSH2 by promoter methylation in primary non-small cell lung tumors and matched sputum samples. *Journal of Clinical Investigation* **111,** 887–895 (2003).

70. Golpon, H. A. *et al.* Emphysema Lung Tissue Gene Expression Profiling. *American Journal of Respiratory Cell and Molecular Biology* **31,** 595–600 (2004).

71. Spira, A. *et al.* Gene Expression Profiling of Human Lung Tissue from Smokers with Severe Emphysema. *American Journal of Respiratory Cell and Molecular Biology* **31,** 601–610 (2004).

72. Wang, I.-M. *et al.* Gene Expression Profiling in Patients with Chronic Obstructive Pulmonary Disease and Lung Cancer. *American Journal of Respiratory and Critical Care Medicine* **177,** 402–411 (2008).

73. Ning, W. *et al.* Comprehensive gene expression profiles reveal pathways related to the pathogenesis of chronic obstructive pulmonary disease. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 14895–14900 (2004).

74. Pierrou, S. *et al.* Expression of Genes Involved in Oxidative Stress Responses in Airway Epithelial Cells of Smokers with Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine* **175,** 577–586 (2007).

75. Bhattacharya, S. *et al.* Molecular Biomarkers for Quantitative and Discrete COPD Phenotypes. *American Journal of Respiratory Cell and Molecular Biology* **40,** 359–367 (2009).

76. Zeskind, J. E., Lenburg, M. E. & Spira, A. Translating the COPD Transcriptome. *Proceedings of the American Thoracic Society* **5,** 834–841 (2008).

77. Subramanian, A. *et al.* 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,** 15545–15550 (2005).

78. Campbell, J. D. *et al.* A gene expression signature of emphysema-related lung destruction and its reversal by the tripeptide GHK. *Genome Medicine* **4,** 1–16 (2012).

79. Ezzie, M. E. *et al.* Gene expression networks in COPD: microRNA and mRNA regulation. *Thorax* **67,** 122–131 (2012).

80. Steiling, K. *et al.* A Dynamic Bronchial Airway Gene Expression Signature of COPD and Lung Function Impairment. *American Journal of Respiratory and Critical Care Medicine* (2013). doi:10.1164/rccm.201208-1449OC

81. Selman, M. *et al.* Gene Expression Profiles Distinguish Idiopathic Pulmonary Fibrosis from Hypersensitivity Pneumonitis. *American Journal of Respiratory and Critical Care Medicine* **173,** 188–198 (2006).

82. Pardo, A. *et al.* Up-Regulation and Profibrotic Role of Osteopontin in Human Idiopathic Pulmonary Fibrosis. *PLoS Medicine* **2,** e251 (2005).

83. Beer, D. G. *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* **8,** 816–824 (2002).

84. Landi, M. T. *et al.* Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival. *PLoS ONE* **3,** e1651 (2008).

85. Landi, M. T. *et al.* MicroRNA Expression Differentiates Histology and Predicts Survival of Lung Cancer. *Clinical Cancer Research* **16,** 430–441 (2010).

86. Deng, N., Sanchez, C. G., Lasky, J. A. & Zhu, D. Detecting Splicing Variants in Idiopathic Pulmonary Fibrosis from Non-Differentially Expressed Genes. *PLoS ONE* **8,** e68352 (2013).

87. Kim, S. C. *et al.* A High-Dimensional, Deep-Sequencing Study of Lung Adenocarcinoma in Female Never-Smokers. *PLoS ONE* **8,** e55596 (2013).

88. Govindan, R. *et al.* Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. *Cell* **150,** 1121–1134 (2012).

89. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185–193 (2003).

90. Wright, M. W. & Bruford, E. A. Naming 'junk': Human non-protein coding RNA (ncRNA) gene nomenclature. *Human Genomics* **5,** 90 (2011).

91. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379,** 131–137 (1996).

92. Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464,** 1071–1076 (2010).

93. Rinn, J. L. & Chang, H. Y. Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry* **81,** 145–166 (2012).

94. Gibb, E. A. *et al.* Human Cancer Long Non-Coding RNA Transcriptomes. *PLoS ONE* **6,** e25915 (2011).

95. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell* **120,** 15–20 (2005).

96. Eulalio, A. *et al.* Deadenylation is a widespread effect of miRNA regulation. *RNA* **15,** 21–32 (2009).

97. Plath, K. *et al.* Role of Histone H3 Lysine 27 Methylation in X Inactivation. *Science* **300,** 131–135 (2003).

98. Kwek, K. Y. *et al.* U1 snRNA associates with TFIIH and regulates transcriptional initiation. *Nature Structural & Molecular Biology* **9,** 800–805 (2002).

99. Ebert, M. S. & Sharp, P. A. Emerging Roles for Natural MicroRNA Sponges. *Current Biology* **20,** R858–R861 (2010).

100. Tay, Y., Rinn, J. & Pandolfi, P. P. The multilayered complexity of ceRNA crosstalk and competition. *Nature* **505,** 344–352 (2014).

101. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

102. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7,** 562–578 (2012).

103. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

104. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).

105. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nature Genetics* **39,** 1278–1284 (2007).

106. Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* **19,** 92–105 (2009).

107. Krüger, J. & Rehmsmeier, M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Research* **34,** W451–W454 (2006).

108. Krek, A. *et al.* Combinatorial microRNA target predictions. *Nature Genetics* **37,** 495–500 (2005).

109. Kass, D. J. & Kaminski, N. Evolving Genomic Approaches to Idiopathic Pulmonary Fibrosis: Moving Beyond Genes. *Clinical and Translational Science* **4,** 372–379 (2011).

110. Zuo, F. *et al.* Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 6292–6297 (2002).

111. Pottelberge, G. R. V. *et al.* MicroRNA Expression in Induced Sputum of Smokers and Patients with Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine* **183,** 898–906 (2011).

112. Milosevic, J. *et al.* Profibrotic Role of miR-154 in Pulmonary Fibrosis. *American Journal of Respiratory Cell and Molecular Biology* **47,** 879–887 (2012).

113. Pandit, K. V. *et al.* Inhibition and Role of let-7d in Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* **182,** 220–229 (2010).

114. Huang, G. T., Athanassiou, C. & Benos, P. V. mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic Acids Research* **39,** W416–W423 (2011).

115. Thannickal, V. J. Evolving Concepts of Apoptosis in Idiopathic Pulmonary Fibrosis. *Proceedings of the American Thoracic Society* **3,** 350–356 (2006).

116. Nakashima, N. *et al.* The p53–Mdm2 association in epithelial cells in idiopathic pulmonary fibrosis and non-specific interstitial pneumonia. *Journal of Clinical Pathology* **58,** 583–589 (2005).

117. Morissette, M. C., Vachon-Beaudoin, G., Parent, J., Chakir, J. & Milot, J. Increased p53 Level, Bax/Bcl-xL Ratio, and TRAIL Receptor Expression in Human Emphysema. *American Journal of Respiratory and Critical Care Medicine* **178,** 240–247 (2008).

118. Denko, N. C. Hypoxia, HIF1 and glucose metabolism in the solid tumour. *Nature Reviews. Cancer* **8,** 705–713 (2008).

119. Wade, M., Li, Y.-C. & Wahl, G. M. MDM2, MDMX and p53 in oncogenesis and cancer therapy. *Nature Reviews. Cancer* **13,** 83–96 (2013).

120. Budanov, A. V., Sablina, A. A., Feinstein, E., Koonin, E. V. & Chumakov, P. M. Regeneration of Peroxiredoxins by p53-Regulated Sestrins, Homologs of Bacterial AhpD. *Science* **304,** 596–600 (2004).

121. Bianchi, M. G., Bardelli, D., Chiu, M. & Bussolati, O. Changes in the expression of the glutamate transporter EAAT3/EAAC1 in health and disease. *Cellular and Molecular Life Sciences* **71,** 2001–2015 (2014).

122. Fujita, H., Sato, K., Wen, T.-C., Peng, Y. & Sakanaka, M. Differential Expressions of Glycine Transporter 1 and Three Glutamate Transporter mRNA in the Hippocampus of Gerbils With Transient Forebrain Ischemia. *Journal of Cerebral Blood Flow and Metabolism* **19,** 604–615 (1999).

123. Ortutay, C., Nore, B. F., Vihinen, M. & Smith, C. I. E. in *Advances in Genetics* (ed. Jeffrey C. Hall, J. C. D. and T. F.) **Volume 64,** 51–80 (Academic Press, 2008).

124. Su, Q., Zhou, Y. & Johns, R. A. Bruton's tyrosine kinase (BTK) is a binding partner for hypoxia induced mitogenic factor (HIMF/FIZZ1) and mediates myeloid cell chemotaxis. *FASEB Journal* **21,** 1376–1382 (2007).

125. Matthews, J. M. & Visvader, J. E. LIM-domain-binding protein 1: a multifunctional cofactor that interacts with diverse proteins. *EMBO Reports* **4,** 1132–1137 (2003).

126. Lin, J. *et al.* FHL family members suppress vascular endothelial growth factor expression through blockade of dimerization of HIF1α and HIF1β. *IUBMB Life* **64,** 921–930 (2012).

127. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10,** R25 (2009).

128. Piccolo, S. R., Withers, M. R., Francis, O. E., Bild, A. H. & Johnson, W. E. Multiplatform single-sample estimates of transcriptional activation. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 17778–17783 (2013).

129. Kass, D. J. *et al.* Cytokine-Like Factor 1 Gene Expression Is Enriched in Idiopathic Pulmonary Fibrosis and Drives the Accumulation of CD4+ T Cells in Murine Lungs: Evidence for an Antifibrotic Role in Bleomycin Injury. *American Journal of Pathology* **180,** 1963–1978 (2012).

130. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative CT method. *Nature Protocols* **3,** 1101–1108 (2008).

131. Subramanian, J. & Govindan, R. Molecular genetics of lung cancer in people who have never smoked. *Lancet Oncology* **9,** 676–682 (2008).

132. Jemal, A. *et al.* Cancer statistics, 2008. *CA: A Cancer Journal for Clinicians* **58,** 71–96 (2008).

133. Jang, J. S. *et al.* Increased miR-708 Expression in NSCLC and Its Association with Poor Survival in Lung Adenocarcinoma from Never Smokers. *Clinical Cancer Research* **18,** 3658–3667 (2012).

134. Spira, A. *et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine* **13,** 361–366 (2007).

135. Shu, W. *et al.* Foxp2 and Foxp1 cooperatively regulate lung and esophagus. *Development* **134,** 1991–2000 (2007).

136. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14,** 7 (2013).

137. Chène, P. Inhibiting the p53–MDM2 interaction: an important target for cancer therapy. *Nature Reviews. Cancer* **3,** 102–109 (2003).

138. Jang, H., Choi, S.-Y., Cho, E.-J. & Youn, H.-D. Cabin1 restrains p53 activity on chromatin. *Nature Structural & Molecular Biology* **16,** 910–915 (2009).

139. Tovar, C. *et al.* MDM2 Small-Molecule Antagonist RG7112 Activates p53 Signaling and Regresses Human Tumors in Preclinical Cancer Models. *Cancer Research* **73,** 2587–2597 (2013).

140. Mir, R. *et al.* Mdm2 antagonists induce apoptosis and synergize with cisplatin overcoming chemoresistance in TP53 wild-type ovarian cancer cells. *International Journal of Cancer* **132,** 1525–1536 (2013).

141. Zelcer, N., Saeki, T., Reid, G., Beijnen, J. H. & Borst, P. Characterization of Drug Transport by the Human Multidrug Resistance Protein 3 (ABCC3). *Journal of Biological Chemistry* **276,** 46400–46407 (2001).

142. O'Brien, C. *et al.* Functional Genomics Identifies ABCC3 as a Mediator of Taxane Resistance in HER2-Amplified Breast Cancer. *Cancer Research* **68,** 5380–5389 (2008).

143. Cipriano, R. *et al.* Conserved Oncogenic Behavior of the FAM83 Family Regulates MAPK Signaling in Human Cancer. *Molecular Cancer Research* molcanres.0289.2013 (2014). doi:10.1158/1541-7786.MCR-13-0289

144. Trivedi, M. V. *et al.* Abstract P6-04-05: GPR110 overexpression increases tumorigenic potential of HER2+ breast cancer cells. *Cancer Research* **73,** P6–04–05–P6–04–05 (2013).

145. Kim, J. *et al.* Detection of ovarian cancer-specific gene by differentially expressed gene polymerase chain reaction prescreening and direct DNA sequencing. *ASCO Meeting Abstracts* **25,** 21106 (2007).

146. Staff, A. C. *et al.* Growth differentiation factor-15 as a prognostic biomarker in ovarian cancer. *Gynecologic Oncology* **118,** 237–243 (2010).

147. Ghosh, G. *et al.* Hypoxia-induced microRNA-424 expression in human endothelial cells regulates HIF-α isoforms and promotes angiogenesis. *Journal of Clinical Investigation* **120,** 4141–4154 (2010).

148. Liu, Z. Tumor suppressive microRNA-424 inhibits osteosarcoma cell migration and invasion via targeting fatty acid synthase. *Experimental and Therapeutic Medicine* (2013). doi:10.3892/etm.2013.959

149. Wu, K. *et al.* MicroRNA-424-5p Suppresses the Expression of SOCS6 in Pancreatic Cancer. *Pathology Oncology Research* **19,** 739–748 (2013).

150. Hall, M. & Rustin, G. Recurrent Ovarian Cancer: When and How to Treat. *Current Oncology Reports* **13,** 459–471 (2011).

151. Puyo, S., Montaudon, D. & Pourquier, P. From old alkylating agents to new minor groove binders. *Critical Reviews in Oncology/Hematology* **89,** 43–61 (2014).

152. Garzon, R., Marcucci, G. & Croce, C. M. Targeting microRNAs in cancer: rationale, strategies and challenges. *Nature Reviews. Drug Discovery* **9,** 775–789 (2010).

153. Yang, N. *et al.* MicroRNA Microarray Identifies Let-7i as a Novel Biomarker and Therapeutic Target in Human Epithelial Ovarian Cancer. *Cancer Research* **68,** 10307–10314 (2008).

154. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

155. Hydbring, P. & Badalian-Very, G. Clinical applications of microRNAs. *F1000Research* (2013). doi:10.12688/f1000research.2-136.v1

156. Du Toit, A. RNA: Circular RNAs as miRNA sponges. *Nature Reviews. Molecular Cell Biology* **14,** 195–195 (2013).

157. Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L. & Brown, P. O. Cell-Type Specific Features of Circular RNA Expression. *PLoS Genetics* **9,** e1003777 (2013).

158. Stower, H. Regulatory RNA: Circular sponges. *Nature Reviews. Genetics* **14,** 238–238 (2013).

159. Jeck, W. R. & Sharpless, N. E. Detecting and characterizing circular RNAs. *Nature Biotechnology* **32,** 453–461 (2014).

CURRICULUM VITAE

REBECCA KUSKO
69 Concord Ave Unit 3, Somerville MA 02143
617-871-9362; rkusko@bu.edu; 08/25/1987

SUMMARY OF QUALIFICATIONS

- **Computational Genomicist** experienced with Illumina and Ion Torrent RNAseq of 500+ clinical samples representing lung cancer, COPD, and IPF including analysis focused on disease associated transcripts and ncRNA-mRNA networks
- **Clinical Translational** know-how, works directly with clinicians on study design as well as with functional genomics benchwork scientists to plan molecular biology and validation experiments

PROFESSIONAL EXPERIENCE

Boston University School of Medicine, Boston MA January 2010 - present
Spira-Lenburg Lab, Computational Biomedicine
Doctoral Researcher

- Using RNAseq data, identified potentially actionable gene expression changes, pathways, and miRNA regulatory networks associated with both COPD and IPF, and validated results with gene expression arrays
- Discovered a unique gene expression signature and ncRNA network in never smoker adenocarcinoma in RNAseq and miRNAseq data, validated results in independent cohort, and identified candidate FDA approved compounds to reverse this oncogenic signature
- Demonstrated connections between lung cancer associated perturbations in gene expression in the small airways near a lung cancer tumor as compared to larger airways, paving the way for minimally invasive tools to diagnose and monitor lung cancer progression
- Participated in preparing and writing of grant applications

Immuneering, Cambridge, MA April 2014-present
Consultant

- Assisted with GWAS of a large clinical cohort and post-GWAS functional analysis

Boston University School of Medicine, Boston MA April 2010 - June 2010
Shared between Montano Lab, Infectious Diseases and Sebastiani Lab, Biostatistics
Rotation Student

- Led a multi-institutional project on premature expression of an age-associated pathway in HIV Infection
- Developed a genetic signature of aging by comparing gene expression in older subjects to younger subjects, observed that HIV accelerates the signature, which has implications for the management of HIV associated wasting

Massachusetts Institute of Technology, Cambridge, MA June 2006 - August 2009
Thilly lab, Samson Lab and Lauffenburger Lab, Biological Engineering
Undergraduate Researcher, Undergraduate Research Opportunity Program (UROP)

- Updated epidemiological mortality database, enabling both testing of cancer modeling hypothesis and the publication of a manuscript
- Studied the DNA damage network by making improvements to existing Pulsed Field Gel Electrophoresis protocols and studied network perturbations through cell culture and western blotting
- Mentored and taught cell culture and western blotting four female freshmen through IAP Research Mentor Program (IRMP), all four students went on to successfully join MIT laboratories as undergraduate researchers

EDUCATION

Boston University School of Medicine (BUSM), Boston, MA
Expected January 2015
PhD, Genetics and Genomics
Massachusetts Institute of Technology (MIT), Cambridge, MA
June 2009
SB, Biological Engineering

PUBLICATIONS

- K. Krysan, **R. Kusko**, T. Grogan, J. D. O'Hearn, K. L. Reckamp, T. C. Walser, E. B. Garon, M. E. Lenburg, S. Sharma, A. E. Spira, D. Elashoff, S. M. Dubinett. "PGE2-driven Expression of c-Myc and OncomiR-17-92 Contributes to Apoptosis Resistance in NSCLC" Molecular Cancer Research (2014).
- L. Kini, P. Herrero-Jimenez, T. Kamath, J. Sanghvi, E. Gutierrez Jr., D. Hensle, J. Kogel, **R. Kusko**, K. Rexer, R. Kurzweil, P. Refinetti, S. Morgenthaler,V. Koledova, E. V. Gostjeva and W. G. Thilly. "Mutator/Hypermutable Fetal/Juvenile Metakaryotic Stem Cells and Human Colorectal Carcinogenesis" *Frontiers in Oncology: Cancer Genetics* (2013). 3:267.
- **R. Kusko**, C. Banerjee, K. K. Long, , A. Carcy, J. Otis, P. Sebastiani, S. Melov, M. Tarnopolsky, S. Bhasin, M. Montano. "Premature Expression of an Age-Associated Muscle Senescence Pathway in HIV Infection." *Skeletal Muscle* (2012).

SELECTED ORAL PRESENTATIONS

- **Lung SPORE Workshop.** "Mapping the airway-wide molecular field of injury in smokers with lung cancer" NCI, Washington DC, July 2014
- **Advances in Genome Biology and Technology (AGBT)**. "SEQing the Shared and Distinct Transcriptional Events Underlying Lung Adenocarcinoma in Ever Smokers and Never Smokers". Marco Island, FL. February 2014
- **Flemish Training Network Life Sciences Next Gen Seq Conference** *(Invited Keynote Lecture)*. "Redefining the Lung Disease Transcriptome with Next Generation RNA-Sequencing". Leuven, Belgium. September 2012
- **Federation of American Societies for Experimental Biology (FASEB) Lung Epithelium in Health and Disease**. "Redefining the Lung Disease Transcriptome with Next Generation RNA-sequencing". Saxtons River, VT. July 2012

- **Advances in Genome Biology and Technology (AGBT)**. “Redefining the Lung Disease Transcriptome with Next Generation RNA-Sequencing”. Marco Island, FL. February 2012
- **Illumina User Group Symposium**. “Redefining the Lung Disease Transcriptome with Next Generation RNA-Sequencing”. Cambridge, MA, September 2011

HONORS/AWARDS:

- Graduate Medical Sciences (GMS) Travel Grant (2014, 2013, 2012, 2011, 2010)
- Graduate Program in Genetics and Genomics Travel Grant (2013)
- Pulmonary NIH T32 Training Grant Appointment (2013, 2012, 2011)
- Russek Student Achievement Day First Place (2012)
- Genome Science Institute (GSI) Poster Award (2012, 2011)
- Russek Student Achievement Day Genetics and Genomics Travel Grant (2012, 2011, 2010)
- Evan’s Day First Place Computational Project (2011)
- Genome Science Institute (GSI) Oral Distinction (2010)
- Integrated Cancer Biology Program (ICBP, MIT) UROP Grant (2008)
- IBM Thomas J. Watson Memorial Scholarship (2005-2009)

TEACHING/LEADERSHIP

Rotation Student Mentor (Boston University School of Medicine) Spring 2012-Summer 2013

- Trained graduate students and postdocs in analysis of microarray and RNAseq datasets, mentored three Bioinformatics PhD students and one MD/PhD rotation student, all stayed on as full time students

Teaching Assistant (Boston University School of Medicine) Spring 2013, Fall 2012, Fall 2010

- Lead discussion sections, graded homework, midterm papers and finals, and orchestrated review sessions
- Courses: Translational Genomics, Foundations in Biomedical Sciences, Principles of Genetics and Genomics

Associate Advisor (Massachusetts Institute of Technology) Fall 2007 - Spring 2009

- Co-mentored freshmen students with Professor Thilly, coached students through a seminar