2023

Airway gene expression alterations in association with radiographic abnormalities of the lung

BOSTON UNIVERSITY

SCHOOL OF MEDICINE

Dissertation

AIRWAY GENE EXPRESSION ALTERATIONS IN ASSOCIATION WITH RADIOGRAPHIC ABNORMALITIES OF THE LUNG

by

KE XU

B.A., Hamilton College, 2011

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2021

Approved by

First Reader

Marc E. Lenburg, Ph.D.
Professor of Medicine

Second Reader

Ehab Billatos, M.D.
Assistant Professor of Medicine

Third Reader

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

“A recluse knows not what year it is,

But sees the entire world with a falling leaf.”

- Anonymous, Chinese Poem.

山僧不解数甲子，一叶落知天下秋。

佚名，唐诗

DEDICATION

I would like to dedicate this work to my dear parents Keqi Xu and Cailian Wang. The three of us have lived separately in different counties for 16 years, but our love endures and grows. And I would also like to dedicate this work to Yumeng Guo. You are the most intelligent person, the kindest companion, and the dearest friend.

ACKNOWLEDGMENTS

Coming from a background in molecular biology and advanced microscopy, I did not consider the possibility of completing a thesis that primarily involves computational work. I remember vividly at one of my medical school interviews when I responded to an interviewer's question on whether I would be interested in learning computational biology with an unapologetic no. That response perhaps, was what got me rejected from that school and led me to Boston University, where I embarked on my journey in computational medicine, serendipitously.

A rotation in the Spira-Lenburg lab quickly changed my perception of computational research. It opened my eyes to technologies so powerful that yielded knowledge that could not only provide insight into disease pathophysiology but be directly translated into clinical applications. Thus, I was over exhilarated to have the opportunity to join the lab at the end of my rotation, which I believe is one of the best choices I have made. To me, Drs. Avrum Spira and Marc Lenburg complemented each other perfectly, representing the finest mentorship, providing both general guidance of the research direction and detailed advice on the scientific approach while allowing me the greatest flexibility in fulfilling my scientific curiosities. I truly appreciate Dr. Lenburg's vast knowledge of statistical methods and the effectiveness in sharing them that led to significant discoveries. I also enjoyed meetings with Dr. Spira during the late hours to go over results, and to shape my analysis in a more impactful direction. Both Drs. Spira and Lenburg offered the most generous help when progress is slow, the most insightful guidance when results are inconclusive, and the kindest encouragement when challenges

are faced. The members of my thesis committee, Drs. Ehab Billatos, Stefano Monti, and Joshua Campbell have also provided great insights throughout the last four years and helped advance my research both in its scientific significance and clinical implications. Dr. Billatos provide insights into the clinical implications of the analysis. Dr. Monti has always asked critical questions that helped to improve the quality of the research. Dr. Campbell has allowed me the opportunity to develop the single cell RNA-sequencing protocol and to learn the analysis of single cell transcriptomics, both of which became a significant component of this work.

One of my biggest fears in joining the Bioinformatics Program, as the first M.D./Ph.D. student at Boston University School of Medicine was whether a lack of prior experience would hinder the ability to learn and apply computational algorithms to my research effectively. This initial hesitation quickly dissipated, thanks to my colleagues, many of whom became my best friends. Rui (Ray) Hong, Xingyi Shi, and Yusuke Koga helped me with my course works, patiently explaining to me concepts in Bioinformatics and correcting countless errors in my code. Yue (Jason) Zhao and Junming Hu enriched my scientific endeavors by accompanying me through academic challenges outside of BU, many led to both challenging and fun learning experiences. I also thank my lab mates Reid McMurry, Jiarui Zhang, Elizabeth Becker, Carter Merenstein, Eric Reed, and all others for stimulating discussions and great entertainment, a part I miss since the start of the COVID19 epidemic.

Markov property has taught us that the future is independent of the past, given today. While it remains one of my favorite scientific properties, it is shockingly inadequate

when applied in the real life. If we are what we eat, our existence is the direct product of the combined interactions of the past. I want to thank Drs. Robert (Bob) Simon, Daniel Chambliss, Larry Knop, and Wei-Jen Chang from Hamilton College, as well as Drs. Katia Manova from Memorial Sloan Kettering Cancer Center, Robert Weinberg from Whitehead Institute for their continued guidance and phone conversation that could last all night when needed.

Moreover, I would like to thank the directors of both the Bioinformatics Program, Dr. Thomas Tullius, and the M.D./Ph.D Program, Drs. John Schwarz, Vickery Trinkaus-Randall, and Steven Borkan for their help along the way. And I sincerely appreciate the administrative support from David King, Caroline Lyman, Johanna Vasquez, Mary Ellen Gipson-Fitzpatrick, Mildred Agosto, Katie Harper, and Donna Gibson.

Lastly, I would like to acknowledge all my collaborators. They include Drs. George Washko and Alejandro Diaz of Brigham and Women's Hospital; Travis Sullivan and Dr. Kimberly Rieger-Christ of Lahey Hospital; Drs. Gang Liu, Yuriy Alekseyev, and Robert Smyth of Boston University. Much of this work relies on their diligent work and mentorship.

My idol physician-scientist Dr. Judah Folkman represented a way of living that balances scientific curiosity, clinical practice, and when time permits, family. Still seeking and striving to live the most meaningful life to me, I feel grateful to have worked in the Spira-Lenburg Lab, where curiosity is cherished, nurtured, and encouraged to bring scientific discoveries and clinical applications. And I cannot think of a more meaningful way to live than marrying medicine with science to help those in need, advancing

knowledge to bring equality to all, and bridging conflicting ideologies by solving common challenges we face as a species. This, in the context of a global pandemic, internal unrest, and international conflicts becomes even more relevant.

Onward to the next chapter! Excited and anxious, yet more confident, I will cherish this experience forever.

AIRWAY GENE EXPRESSION ALTERATIONS IN

ASSOCIATION WITH RADIOGRAPHIC ABNORMALITIES OF THE LUNG

KE XU

Boston University School of Medicine, 2021
Ph.D. degree requirements completed in 2021
Dual M.D./Ph.D. degrees expected in 2023

Major Professor: Marc E. Lenburg, Ph.D. Professor of Medicine

High-resolution computed tomography (HRCT) of the chest is commonly used in the diagnosis of a variety of lung diseases. Structural changes associated with clinical characteristics of disease may also define specific disease-associated physiologic states that may provide insights into disease pathophysiology. Gene expression profiling is potentially a useful adjunct to HRCT to identify molecular correlates of the observed structural changes. However, it is difficult to directly access diseased distal airway or lung parenchyma routinely for profiling studies.

Previously, we have profiled bronchial airway in normal-appearing epithelial cells at the mainstem bronchus, detecting distinct gene expression alterations related to the clinical diagnosis of chronic obstructive pulmonary disease (COPD) and lung cancer. These gene expression alterations offer insights into the molecular events related to diseased tissue at more distal airways and in the parenchyma, which we hypothesize are due to a field-of-injury effect. Here, we expand this prior work by correlating airway gene expression to COPD and bronchiectasis phenotypes defined by HRCT to better understand

the pathophysiology of these diseases. Additionally, we classified pulmonary nodules as malignant or benign by combining HRCT nodule imaging characteristics with gene expression profiling of the nasal airway.

First, we collected brushing samples from the main-stem bronchus and assessed gene expression alterations associated with COPD phenotypes defined by K-means clustering of HRCT-based imaging features. We found three imaging clusters, which correlated with incremental severity of COPD: preserved, interstitial predominant, and emphysema predominant. 357 genes were differentially expressed between the normal and the emphysema predominant clusters. Functional analysis of the differentially expressed genes suggests a possible induction of inflammatory processes and repression of T-cell related biologic pathways, in the emphysema predominant cluster.

We then discovered gene expression alterations associated with radiographic evidence of bronchiectasis (BE), an underdiagnosed obstructive pulmonary disease with unclear pathophysiology. We found 655 genes were differentially expressed in bronchial epithelium from individuals with radiographic evidence of BE despite none of the study participants having a clinical BE diagnosis. In addition to biological pathways that had been previously associated with BE, novel pathways that may play important roles in BE initiation were also discovered. Furthermore, we leveraged an independent single-cell RNA-sequencing dataset of the bronchial epithelium to explore whether the observed gene expression alterations might be cell-type dependent. We computationally detected an increased presence of ciliated and deuterosomal cells, as well as a decreased presence of

basal cells in subjects with widespread radiographic BE, which may reflect a shift in the cellular landscape of the airway during BE initiation.

Finally, we identified gene expression alterations within the nasal epithelium associated with the presence of malignant pulmonary nodules. A computational model was constructed for determining whether a nodule is malignant or benign that combines gene expression and imaging features extracted from HRCT. Leveraging data from single-cell RNA sequencing, we found genes increased in patients with lung cancer are expressed at higher levels within a novel cluster of nasal epithelial cells, termed keratinizing epithelial cells.

In summary, we leveraged gene expression profiling of the proximal airway and discovered novel biological pathways that potentially drive the structural changes representative of physiologic states defined by chest HRCT in COPD and BE. This approach may also be combined with chest HRCT to detect weak signals related to malignant pulmonary nodules.

TABLE OF CONTENTS

DEDICATION v

ACKNOWLEDGMENTS vi

ABSTRACT x

TABLE OF CONTENTS xiii

LIST OF TABLES xviii

LIST OF FIGURES xx

LIST OF ILLUSTRATIONS xxii

LIST OF ABBREVIATIONS xxiii

CHAPTER ONE: 1

INTRODUCTION 1

1.1 High Resolution Computed Tomography 2

1.2 Application of HRCT in chronic obstructive pulmonary disease, bronchiectasis, and lung cancer 2

1.2.1 Role of HRCT in COPD 2

1.2.2 Role of HRCT in bronchiectasis 2

1.2.3 Application of HRCT in lung cancer 3

1.3 The Transcriptome 4

1.4 Field-of-injury effect in airway epithelium 5

1.5 XGBoost 6

1.6 Dissertation Aims 6

Aim 1: Examine Gene expression alterations associated with HRCT-derived imaging phenotypes related to COPD 7
Aim 2: Examine Gene expression alterations associated with HRCT-derived radiographic bronchiectasis 8
Aim 3: Examine Gene expression alterations associated with malignant pulmonary nodules and create a computational model to differentiate indeterminate nodules 8
CHAPTER TWO: Gene expression alterations associated with 10
CT-derived imaging phenotypes related to COPD 10
2.1 Abstract 11
2.2 Introduction 12
2.3 Material and Methods 14
2.3.1 Study Participants and Sample Analysis 14
2.3.2 HRCT acquisition and processing 15
2.3.3 COPD imaging cluster derivation 16
2.3.4 Biospecimen collection in DECAMP 17
2.3.5 Data preprocessing and quality control 17
2.3.6 Analytic strategies of differential gene expression 18
2.3.7 Determining cell type abundance in bulk tissue 19
2.3.8 Statistical analysis 19
2.4 Results 19
2.4.1 Replication of three COPD imaging clusters in DECAMP 19
2.4.2 Correlation between COPD imaging clusters and clinical characteristics 20

2.4.3 Genes differentially expressed between the preserved and the emphysema predominant clusters 21
2.4.4 Gene expression profile of the individuals of the interstitial cluster 22
2.4.5 Correlation between the differentially expressed genes and interferon-beta ... 23
2.4.6 Compare and contrast of previous genes differentially expressed in patients with clinical COPD 23
2.4.7 Estimation of immune cell type abundance from bulk tissues 24
2.5 Discussion 34
2.6 Conclusion 40
CHAPTER THREE: 41
Gene expression alterations associated with 41
HRCT-derived radiographic bronchiectasis 41
3.1 Abstract 42
3.2 Introduction 43
3.3 Material and Methods 44
3.3.1 Study Participants and Sample Analysis 44
3.3.2 HRCT acquisition and characterization of radiographic bronchiectasis 45
3.3.3 Analytic strategies of differential gene expression 46
3.3.4 Single cell RNA-sequencing workflow 47
3.3.5 Deconvolution of bulk RNA-sequencing samples 48
3.3.6 Statistical analysis 48
3.4 Results 48

3.4.1 Participant demographics, pulmonary function, and imaging measurement ... 48
3.4.2 Identification of three distinct clusters of participants based on gene expression profiles ... 49
3.4.3 Functional analysis of differentially expressed genes by radiographic bronchiectasis ... 51
3.4.4 Widespread radiographic BE correlates with increased proportions of ciliated and deuterosomal cells, and decreased proportions of basal cells ... 53
3.5 Discussion ... 67
3.6 Conclusion ... 71
CHAPTER FOUR: Gene expression alterations associated with malignant pulmonary nodules and creation of a computational model to differentiate indeterminate nodules ... 72
4.1 Abstract ... 73
4.2 Introduction ... 74
4.3 Material and Methods ... 76
4.3.1 Study Participants ... 76
4.3.2 HRCT acquisition and characterization of pulmonary nodules ... 77
4.3.3 Analytic strategies of differential gene expression ... 77
4.3.4 Single cell RNA-sequencing protocol ... 78
4.3.5 Single cell RNA-sequencing workflow ... 78
4.3.5 Compare and contrast of identified cell types to cell types inferred by cellassign ... 79
4.3.6 Feature selection and Modelling ... 79

4.3.7 Statistical analysis 81
4.4 Results 84
4.4.1 Participant demographics and pulmonary function 84
4.4.2 Pre-filtering of samples of potentially lower quality. 84
4.4.3 Genes differentially expressed between subjects with/without a malignant pulmonary nodule 86
4.4.4 Examination of nasal epithelium at a single-cell resolution 86
4.4.5 Comparison and contrast with cell type identification by different methods .. 87
4.4.6 Genes up-regulated in participants with a malignant nodule were enriched in the keratinizing epithelial cells 87
4.4.7 Model with combined features achieved highest classifier performance 88
4.5 Discussion 102
4.6 Conclusion 108
CHAPTER FIVE: 109
General Conclusions and Future Directions 109
LIST OF JOURNAL ABBREVIATIONS 114
BIBLIOGRAPHY 118
CURRICULUM VITAE 136

LIST OF TABLES

Table 2.1. Clinical characteristics of COPDGene subjects in the three imaging clusters…… 25

Table 2.2. Clinical characteristics of DECAMP subjects in the three imaging clusters…26

Table 2.3. Functional analysis of differentially expressed genes between individuals of the preserved and the emphysema predominant imaging clusters. 27

Table 2.4. Gene set enrichment analysis showed enrichment of Hallmark genesets in participants within the emphysema predominant and the preserved clusters 27

Table 3.1. Clinical characteristics of subjects with and without radiographic BE 54

Table 3.2. Distribution of radiographic BE by lobe..... 55

Table 3.3. Clinical characteristics of subjects with and without widespread radiographic BE…….. 56

Table 3.4. Clinical characteristics of the participants of the three clusters based on gene expression profiles 57

Table 3.5. Clinical characteristics of the participants with and without widespread BE in the bronchiectatic cluster 58

Table 3.6. Clinical characteristics of the participants without widespread BE in subgroups……. 59

Table 3.7. Functional analysis of differentially expressed genes of the five gene clusters……..… 60

Table 3.8. Gene set enrichment analysis showed enrichment of Hallmark genes in participants with and without radiographic BE 60

Table 3.9. Genes associated with ciliogenesis 61

Table 4.1. Clinical characteristics of the participants with and without a malignant pulmonary nodule 89

Table 4.2. Comparison of samples filter by TIN and by combined metrics……..90

Table 4.3. Subject demographics for nasal single cell RNA-sequencing profiling ….....90

Table 4.4. Proportions of different types of cells in the nasal compartment ……. ..91

Table 4.5. Compare and contrast of identified cell types to cell types inferred by cellassign.…….. 92

Table 4.6. List of features included for model construction ……..……..…….....93

Table 4.7. List of the top 50 features in the final model ……..……..……...94

LIST OF FIGURES

Figure 2.1. Representative CT images from each of the three imaging clusters 20

Figure 2.2. Comparison of Clinical Characteristics and Mortality of the Clusters in the COPDGene Cohort 28

Figure 2.3. Unsupervised heatmap of the 179 genes differentially expressed between Individuals of the emphysema and preserved clusters. .. ……...29

Figure 2.4. Gene set variation analysis was used to summarize the expression of each gene cluster in each sample …….……...30

Figure 2.5. Correlation between the differentially expressed genes and interferon-beta . 31

Figure 2.6. GSEA results assessing the enrichment of the 98 genes with relation to clinical COPD in participants within the emphysema predominant cluster on *t* statistics…….. 32

Figure 2.7. Estimated enrichment of T cells and neutrophils in individuals from the three imaging clusters 33

Figure 3.1. Schematic representation of participant clustering based on imaging and gene expressions. ……45

Figure 3.2. Unsupervised heatmap of the 655 genes associated with widespread radiographic bronchiectasis (presence of radiographic BE in at least 3 lobes) 62

Figure 3.3. Modular gene expressions in genomic clusters ……..……...63

Figure 3.4. Modular gene expressions associated with the number of lobes with radiographic BE …….…….……..….. 63

Figure 3.5. GSEA results assessing the enrichment of the 310 genes with relation to

ciliogenesis in participants with widespread radiographic BE based on *t* statistics…64

Figure 3.6. Single-cell RNA-seq analysis of genes differentially expressed in participants with widespread radiographic BE ……...……...65

Figure 3.7. Boxplots of all cell type proportions estimated by AutoGeneS. ……...66

Figure 4.1. Protocol for single cell dissociation ……..…….....82

Figure 4.2. Schematic representation of creating a model to predict malignancy of an indeterminant pulmonary nodule …….……..83

Figure 4.3. Distribution of TIN and library sizes …….…….....95

Figure 4.4. Unsupervised heatmap of the QC metrics …….....96

Figure 4.5. Semi-supervised heatmap of the 75 genes associated with malignant pulmonary nodules ……...……...……...97

Figure 4.6. Single-cell RNA-seq analysis of the nasal epithelium from 15 ever-smokers …….…..98

Figure 4.7. UMAP projections showing the expression pattern of genes highly enriched among the keratinizing epithelial cells. ……..……...99

Figure 4.8. Enrichment of lung cancer associated genes in the keratinizing epithelial cells.……..…….100

Figure 4.9. Classifier performance of models leveraging clinical, radiomic, genomic, and combined features ……..…..101

LIST OF ILLUSTRATIONS

Illustration 1.1. Simplified representation of boosting algorithm for binary classification…….7

Illustration 3.1. Proposed mechanism of early BE development……... ...70

Illustration 4.1. Schematic representation of radiomic features.……80

LIST OF ABBREVIATIONS

ANOVA	Analysis of variance
BE	Bronchiectasis
BU	Boston University
CARET	Classification and regression training
COPD	Chronic obstructive pulmonary disease
COPDGene	Genetic epidemiology of chronic obstructive pulmonary disease
DECAMP	Detection of early lung cancer among military personnel
FEV1	Forced expiratory volume in one second
FVC	Forced vital capacity
FDR	False discovery rate
GOLD	Global initiative for chronic obstructive lung disease
GSEA	Gene set enrichment analysis
GSVA	Gene set variation analysis
KEGG	Kyoto encyclopedia of genes and genomes
LFC	Log fold change
LIMMA	Linear Models for Microarray Data
MSigDB	Molecular signatures database
PFTs	Pulmonary functional tests
RIN	RNA integrity number
SGRQ	St. George's Respiratory Questionnaire
STAR	Spliced Transcripts Alignment to a Reference

SU2C	Stand Up to Cancer
TIN	Transcript integrity number
UMAP	Uniform manifold approximation and projection
XGBoost	eXtreme gradient boosting

CHAPTER ONE:

INTRODUCTION

1.1 High Resolution Computed Tomography

Chest radiograph is the initial imaging tool for the lung parenchyma due to its low cost, minimal radiation, wide availability, and ease of performance[1]. However, it is normal in 10-15 percent of symptomatic patients with infiltrative lung disease[2], in up to 12 percent of those with bronchiectasis[3], and about 60 percent in patients with emphysema[4]. Chest radiograph has also been shown to have an overall sensitivity of 80 percent for detection of diffuse lung disease[5]. Thus, high resolution computed tomography (HRCT, also called thin-section CT scanning), is frequently applied for more information on the lung[6,7].

1.2 Application of HRCT in chronic obstructive pulmonary disease, bronchiectasis, and lung cancer

1.2.1 Role of HRCT in COPD

Chronic obstructive pulmonary disease (COPD) is clinically defined in functional terms as a slowly progressive disorder with airway obstruction that does not change markedly over several months[8]. While not used for diagnosing COPD, HRCT could detect morphological abnormalities that cause airway obstruction, which is divided into emphysema and chronic bronchitis. HRCT is more sensitive than other imaging modalities in detecting emphysema[9]. The role of HRCT in the assessment of chronic bronchitis is less defined[10].

1.2.2 Role of HRCT in bronchiectasis

Imaging is essential in diagnosing bronchiectasis[11]. HRCT the most sensitive and specific non-invasive method for diagnosing bronchiectasis. Additionally, the pattern of

disease on HRCT may be related to disease etiology[12] and disease severity[13]. Some of the bronchiectasis-related imaging signs on HRCT include bronchial dilation, lack of bronchial tapering, and visualization of bronchi within 1 cm of the pleura, and vascular abnormalities.

1.2.3 Application of HRCT in lung cancer

The utility of imaging and lung cancer screening has been evolving. Three large American screening programs and another in Czechoslovakia have previously demonstrated that a combination of chest radiography and sputum analysis was able to increase detection of early-stage lung cancer, which led to improved 5-yr survival rates in the screened versus control groups[14 - 18]. However, none of these cohorts showed a statistically significant reduction in overall mortality. Low-dose computed tomography (CT) detects many more lung nodules than chest radiography. However, only a small percentage of these nodules turn out to be lung cancer, ranging from 13 to 27% [15, 19, 20, 21]. The high discovery rate and a low cancer prevalence in the screened patients called for a method to help to differentiate benign from malignant nodules. By assessing the patterns of calcification at both low-dose and high-resolution CT (HRCT), and by repeated scanning, one group demonstrated high specificity in differentiating malignant nodules[19]. The most recent guideline for lung cancer screening recommends annual screening for lung cancer with low-dose computed tomography (LDCT) in adults aged 50-80 years who have a 20 pack-year smoking history and currently smoke or have quit within the past 15 years[22].

1.3 The Transcriptome

This body of work will focus on the application of transcriptomics, the study of total RNA in a tissue or cell. Total transcript abundance directly measures the levels of gene expression and positively correlates with protein levels. Two methods were used to generate the datasets in the following analyses: RNA-sequencing of the bulk tissue and single cell RNA sequencing.

RNA-sequencing was developed more than a decade ago and has since seen a universal adoption in molecular biology[23]. By directly sampling transcript abundance, it allows us to detect transcriptomic changes associated with a physiological or pathological process. The captured transcriptomic changes have also been successfully leveraged to build disease biomarkers that are clinically relevant in providing advanced care.

While RNA-seq from bulk tissue and/or cultured cells has advanced our understanding of biology, it lacks the power to differentiate transcriptomic changes driven by either a change in cellular landscape, or an altered transcriptomic control of gene activation, or a mixture of both. Sequencing of bulk tissues also masks the expression from rare cell populations. Thus, single cell sequencing technology has been developed to enable us to move beyond bulk RNA sequencing[24]. Large-scale cell atlas projects were initiated to determine the various cell types in an organism or tissue[25, 26], which provided us with an unprecedented resolution of the cellular composition of the human body and brain. Some early and exciting discoveries revealed by single cell RNA-sequencing included the discovery of ionocyte, a rare cell type that may play important roles in cystic fibrosis [27, 28].

1.4 Field-of-injury effect in airway epithelium

The concept of field-of-injury stems from "field effect in cancer" that was originally proposed in 1953[30]. The field effect is represented by molecular abnormalities in tissues that appear histological normal, adjacent to the diseased tissue. Such effect has been reported in several sites and organs, ranging from head and neck[31], breast[32], colon and rectum[33], and lung[34,35]. The field effect may be caused by two mechanisms that are not mutually exclusive: genetic alteration in a stepwise fashion and epigenetic perturbation resulting from a shared tissue microenvironment. In the stepwise genetic alteration model, some or all cells within a field harvest critical genetic alteration (initiation) that lead to more genetic alterations (promotion), cells that fail to balance the expression of oncogenes and tumor suppressor genes will start proliferating in an uncontrolled fashion that leads to abnormal histology while the rest of tissue remain normal histologically. A classic example of this sequential development of colorectal cancer with a sequential accumulation of mutations in specific cancer-related genes, APC, KRAS, and p53[36,37]. The field effect by epigenetic perturbation is achieved by hypermethylation of the DNA promoter of tumor suppressor genes, caused by shared exposure to chemical compounds, microbial invasion, or environmental pollutants.

Leveraging the "field of injury" concept, our lab has previously demonstrated gene expression differences in the normal airway epithelium of smokers can serve as a sensitive and specific indicator of lung cancer risk, both in the bronchial[35] and nasal[38] airways. We have also found gene expression alterations in normal appearing mainstem bronchus correlated with spirometrically defined Chronic Obstructive Pulmonary Disease (COPD)[39].

1.5 XGBoost

In modeling, boosting is an ensemble technique in which new models are added to correct the errors made by existing models. Gradient boosting is a modified boosting approach in which new models are created to predict the residuals of prior models. Some of the advantage of gradient boosting include ease of implementation, fast computation, and application of an ensemble learning algorithm, which combines the predictions of multiple base learners (Illustration 1.1). The flexibility of the ensemble model allows for learning more complex relationships between features and labels in the training set.

XGBoost stands for eXtreme Gradient Boosting and is an implementation of gradient boosted decision trees designed for speed and performance[40]. In recent competitions held by Kaggle (www.kaggle.com), a popular online community of data scientists and machine learning practitioners, XGBoost was the winning solution for over 20 major competitions[41].

1.6 Dissertation Aims

The following aims seek to extend the field of injury hypothesis by using gene-expression leveraging both bulk and single cell RNA sequencing to capture transcriptomic alterations associated with 1) subgroups of patients defined by imaging features associated with chronic pulmonary obstructive disease, 2) radiographic bronchiectasis, and 3) malignant pulmonary nodules. Collectively, these investigations will provide insight into molecular workings associated with the radiographic changes, and may be clinically useful in creating a more personalized medicine approach.

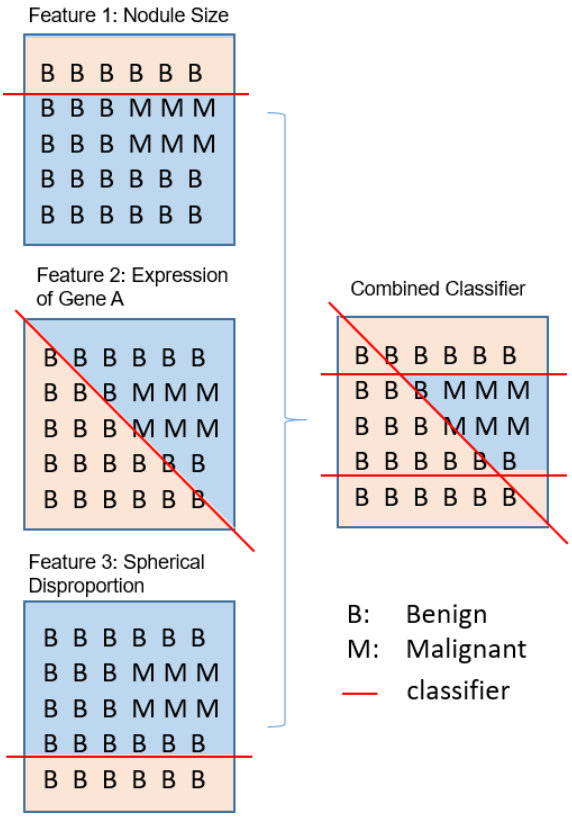

Illustration 1.1. Simplified representation of boosting algorithm for binary classification. A classifier differentiating benign and malignant nodules was created for each of the three representative features (clinical (nodule size), genomic (Gene A), radiomic (spherical disproportion)). The combined classifier represents a single strong learner by iteratively adding the three learners.

Aim 1: Examine Gene expression alterations associated with HRCT-derived imaging phenotypes related to COPD

While the current COPD evaluation guidelines have moved from a spirometry based grading of severity to a composite gradation based on airflow obstruction, symptoms, and risk of exacerbation using various thresholds[42,43], lack of objective measurements of symptoms and evidence-based threshold has made the new guidelines impractical to

follow[44]. HRCT has the potential to objectively detect COPD subgroups, map the disease path to monitor progression and extrapolate its origin, and help to disentangle COPD pathophysiology. Here, we aimed to derive patient subgroups that share similar imaging features and examine whether the subgroups correlated with disease severity. Next, we captured the molecular workings associated with these subgroups using transcriptomic analysis.

Aim 2: Examine Gene expression alterations associated with HRCT-derived radiographic bronchiectasis

Bronchiectasis (BE) is an underdiagnosed obstructive pulmonary disease characterized by permanent dilation of the airway[45]. The pathophysiology of bronchiectasis remains unclear. Here, we aim to examine CT imaging scans and capture radiographic evidence for BE from subjects without a diagnosis for the disease. Next, we correlate gene expression alterations at the mainstem bronchus to the presence of radiographic BE to identify pathways that may play important roles in early BE initiation.

Aim 3: Examine Gene expression alterations associated with malignant pulmonary nodules and create a computational model to differentiate indeterminate nodules

The latest guideline for lung cancer screening recommends annual screening for lung cancer with low-dose computed tomography (LDCT) in adults aged 50-80 years who have a 20 pack-year smoking history and currently smoke or have quit within the past 15

years[22]. With the advent of more imaging studies, the number of detected nodules is rising when only a small percentage of these nodules are cancerous. Thus, there is an urgent need for making a sensitive and specific classifier distinguishing indeterminate pulmonary nodules. Here, we aim to derive gene expression alterations within the nasal epithelium that are associated with malignant nodules. Following on from this, we will integrate these genomic features with radiomic features with clinical factors to build a classifier leveraging XGBoost algorithm.

CHAPTER TWO: Gene expression alterations associated with CT-derived imaging phenotypes related to COPD

Disclaimer: The figures and text in this chapter were originally published as:

Billatos E, Ash SY, Duan F, Xu K, et al. Distinguishing smoking related lung disease phenotypes via imaging and molecular features. Chest. Published online September 15, 2020. doi:10.1016/j.chest.2020.08.2115

2.1 Abstract

There is a growing interest in utilizing multiple disease-related parameters to identify novel patient phenotypes and understand the pathophysiology of Chronic Obstructive Pulmonary Disease (COPD). High-Resolution Computed Tomography (HRCT) of the chest and airway gene expression have previously been used independently to characterize lung function impairment and understand COPD subtypes. In the present study, we sought to combine these two modalities by identifying bronchial airway gene expression correlates of HRCT phenotypes to explore their biological basis.

Using K-means clustering, we clustered participants from the COPDGene study (n=5273) based on CT imaging characteristics and then evaluated their clinical phenotypes. These clusters were replicated in the Detection of Early Lung Cancer Among Military Personnel (DECAMP) cohort (n=360), from which 146 samples were further characterized using bronchial epithelial gene expression.

CT identified three patient clusters that are phenotypically distinct. The preserved cluster is enriched for individuals with normal lung function, the emphysema cluster is enriched for individuals with obstructive lung disease, and the interstitial cluster is enriched for individuals with an intermediate severity of COPD. In longitudinal follow-up,

individuals from the emphysema group had greater declines in exercise capacity and lung function, more emphysema, more exacerbations, and higher mortality.

We found that 179 genes are differentially expressed between the emphysema and the preserved clusters (DEG; 99 genes down-regulated and 80 genes up-regulated). The 99 genes down-regulated in the emphysema cluster were enriched for genes in T-cell-related pathways, while the 80 genes up-regulated in the emphysema cluster were enriched for genes in pathways related to inflammation and TNF-alpha signaling. While we did not detect significant expression differences between the preserved and interstitial clusters, GSVA showed that in the interstitial cluster, the expression levels of the DEGs were intermediate between the preserved and emphysema clusters. In parallel, we found an enrichment of T cells in individuals of the preserved cluster and a reduction of neutrophils via computational deconvolution of the bulk tissues.

Taken together, the gene expression alterations associated with structural changes of the emphysema predominant cluster suggest potential roles of immune cells in subgroups of individuals with COPD. A well-balanced orchestration of innate and adaptive immunity may be crucial for guarding airway epithelium against changes that lead to emphysematous changes.

2.2 Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a major cause of morbidity and mortality, both in the USA and in the world[42, 46]. It is currently the 3rd leading cause of death – over six million people died of COPD in 2019, responsible for approximately 6%

of total death globally[47]. Despite advancements in diagnostics, therapeutics, and care guidelines, the problem of COPD in the USA is undeniably significant, and the health burden of patients, their families, and society, in general, is still growing[48]. Globally, this increasing burden is projected to increase in coming decades due to aging, continued exposure to risk factors[49], underdiagnoses[50], overdiagnosis[51], and mistreatment[52, 53].

COPD care is further complicated by ever-evolving society guidelines, confusing criteria thresholds, and conflicting therapeutic strategies[10]. Recent COPD evaluation guidelines, including the Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria[42, 43] have moved from a spirometry based grading of severity to a composite gradation based on airflow obstruction, symptoms, and risk of exacerbation using various thresholds. However, the utility of the new gradation system is yet to be validated. Lack of confidence from clinicians, and arguably rightfully so, stems from the fact that none of the guideline-recommended thresholds for COPD severity is based on evidence from randomized trials[55]. In a recent study of 445 patients with spirometry proven COPD, the multidimentional system had caused massive reclassification of COPD for up to half of patients. In one report, the new staging system classified patients into COPD stages that had little correlation with either the physicians or their patients' impressions about the COPD severity[56].

Sensitive mathematical models or precise scoring systems are key to more accurate and practical clinical classifications of COPD that guide diagnoses and treatment options[57]. Multiple investigations have utilized CT to characterize COPD-related lung diseases and most have focused on a specific feature such as an objective measure of emphysema for

correlative investigation, prognostication, and an intermediate endpoint for therapeutic trials. Several groups are now leveraging CT as well as clinical data for more general subgrouping efforts through supervised and unsupervised learning algorithms[58–62]. While these approaches identify clinically relevant subgroups, they have generally lacked organ-specific molecular correlates.

Here, we leveraged clinical and imaging data from a large research cohort, the COPDGene Study, combined with clinical and bronchial epithelial gene expression data from the Detection of Early Lung Cancer Among Military Personnel (DECAMP) Study to identify COPD-related subgroups. It remains to be seen, however, whether the transcriptomic of the bronchial airway epithelium can capture the molecular workings related to the disease subtypes defined by CT. The following study, therefore, explores how the field of injury phenomenon can be extended to begin to explore the molecular working that could drive the structural changes on CT, and whether these alterations are similar or dissimilar to previously described gene expression alterations related to clinical COPD[39].

2.3 Material and Methods

2.3.1 Study Participants and Sample Analysis

Participants were part of two large cohorts, COPDGene[63] and the Detection of Early Lung Cancer Among Military Personnel (DECAMP)[64]. The COPDGene Study (NCT00608764) cohort is a multicenter longitudinal observational investigation of

smokers focused on the epidemiologic and genetic factors associated with COPD. 10,306 COPDGene participants were initially recruited between October 2006 and January 2011. All participants were invited to return for five- and ten-year follow-up visits. Our analyses were limited to those individuals who had completed both baseline and five-year follow-up visits.

DECAMP is a multi-center consortium comprised of 15 military treatment facilities, Veterans Affairs (VA) hospitals, and academic centers across the United States. Participants were recruited into one of two study protocols, designated as DECAMP-1 (NCT01785342) and DECAMP-2 (NCT02504697) to develop an integrated panel of biomarkers that discriminate benign and malignant indeterminate nodules detected on CT scans. Briefly, participants of DECAMP-1 were adults aged 45 and older with indeterminate pulmonary nodules and heavy smoking history. Study participants of DECAMP-2 were aged 50-79 with a heavy smoking history and a family history of lung cancer or a personal history of COPD.

This study was approved by the Human Research Protection Office (HRPO) for the Department of Defense, and the individual site IRBs for every participating site. All subjects were approached for written informed consent to participate in the study per IRB regulations.

2.3.2 HRCT acquisition and processing

For COPDGene participants, volumetric CT scans of the chest were performed at both maximal inflation and relaxed exhalation. Images were acquired with the following

CT protocol: for General Electric (GE) LightSpeed-16, GE VCT-64, Siemens Sensation-16 and -64, and Philips 40- and 60-slice scanners with 120kVp, 200mAs, and 0.5s rotation time. Images were reconstructed using a standard algorithm at 0.625mm slice thickness and 0.625mm intervals for GE scanners; using a B31f algorithm at 0.625 (Sensation-16) or 0.75mm slice thickness and 0.5mm intervals for Siemens scanners; and using a B algorithm at 0.9mm slice thickness and 0.45mm intervals for Philips scanners.

DECAMP-1 utilized CT scans collected as part of routine clinical care while DECAMP-2 utilized a standardized protocol for image acquisition and reconstruction. DECAMP-2 scans were collected using low dose helical computed tomography on a minimum 16-slice scanner. The scans were acquired at 2.5 to 5 mm but reconstructed into 1 mm slice thickness using the soft tissue and lung algorithms.

2.3.3 COPD imaging cluster derivation

The extraction of imaging features has been previously described in detail[65]. Cluster analysis was performed by our collaborators using a parsimonious set of variables selected to represent the breadth of airway, lung parenchyma and extrapulmonary processes. The imaging features were log-transformed and standardized as needed to address distribution skewness and range. K-means clustering was then applied to these variables to group the subjects into clusters. The optimum number of clusters was determined using the Silhouette. Because these methods suggested differing numbers of clusters (2 and 4 respectively), the average of the number of clusters suggested (3) was used.

2.3.4 Biospecimen collection in DECAMP

All individuals in the DECAMP study underwent bronchoscopy. Bronchial airway epithelial cells were obtained from brushings of the right mainstem bronchus collected during fiberoptic bronchoscopy with an endoscopic cytobrush (Cellebrity Endoscopic Cytology Brush, Boston Scientific, Boston). The brushes were immediately placed in 1mL of RNAprotect Cell Reagent (Qiagen, Valencia, CA) and kept at -80 Degree Celsius, until RNA isolation was performed.

2.3.5 Data preprocessing and quality control

Total RNA was isolated using the miRNasey Mini Kit (Qiagen, Valencia, CA). RNA integrity was assessed by Agilent BioAnalyzer, and RNA purity was confirmed using a NanoDrop spectrophotometer. Libraries were generated using the Illumina TruSeq Stranded Total RNA kit and sequenced on Illumina NextSeq 500 and Illumina HiSeq2500 instruments with 75 base-pair paired-end reads (Illumina, San Diego, CA). We developed an automatic pipeline (https://github.com/compbiomed/RNA_Seq) with standard setups based on the Nextflow framework to obtain the expression levels for each gene[66]. Reads were aligned to the Genome Reference Consortium human build 37 (GRCh37) using STAR[67]. Gene and transcript level counts were calculated using RSEM[68] using Ensembl v75 annotation.

All samples included for the analysis had sex annotation correlated with the expression of the constitutively expressed Y-linked genes CYorf15A, DDX3Y, KDM5D,

RPS4Y1, USP9Y, and UTY. No additional filtering of samples was carried out for calculating clinical correlations.

2.3.6 Analytic strategies of differential gene expression

The LIMMA package (version 3.10) in R (version 3.6.0) was used to assess the differential gene expression[69]. The counts were then filtered based on counts per million (CPM) such that a gene could only be included if its CPM was greater than 1 in 10% of the total number of patients. A false discovery rate (FDR) of 0.25 was used to select significantly differentially expressed genes. The functional analysis of the differentially expressed genes was performed using the STRING database[70]. Heatmaps were used to visualize the data and identify unsupervised participant clusters using the "Ward.D2" algorithm.

To identify genes differentially expressed between subjects within each imaging cluster:

(1) Gene ~ imaging cluster + smoking + error

Gene set enrichment analysis (GSEA) [71] was performed on pre-ranked gene lists created by pairwise comparisons between participants' clusters using Hallmark gene sets[72]. Gene set variation analysis (GSVA) was performed using gene sets of interest using standard workflow[73].

2.3.7 Determining cell type abundance in bulk tissue

CIBERSORTx[74], a machine learning method to infer cell-type abundance was used to detect the presence of immune cells within each bulk tissue. LM22 signature genes file was downloaded from the CIBERSORT resource and the deconvolution was performed using standard setting (version Jar Version 1.05). [https://cibersort.stanford.edu/manual.php]

2.3.8 Statistical analysis

Data are presented as means and standard deviations for continuous measurements and number and percentage for categorical features. P values were calculated using a Student's T test, Fisher's exact test, Kruskal test, or ANOVA F test.

2.4 Results

2.4.1 Replication of three COPD imaging clusters in DECAMP

Three distinct clusters of COPDGene participants were identified using quantitative imaging features: preserved, interstitial predominant, and emphysema predominant (Table 2.1). The details were published[65]. Briefly, the individuals in the preserved cluster generally had preserved airway wall thickness and the fewest number of parenchymal abnormalities (emphysema and interstitial features). The individuals in the emphysema cluster had the highest emphysema scores and mildly thickened airway walls, whereas those in the interstitial predominant cluster had the highest number of interstitial changes

and highest airway wall thickness. Representative CT images for each of the three clusters were shown in Figure 2.1.

When the same imaging clustering technique was applied to the DECAMP cohort to derive three imaging clusters, the findings mirrored those found in COPDGene. In the DECAMP cohort, those in the preserved cluster and those in the emphysema cluster had the least severe and most severe clinical phenotypes, respectively. The interstitial cluster had an intermediate clinical phenotype (Table 2.2).

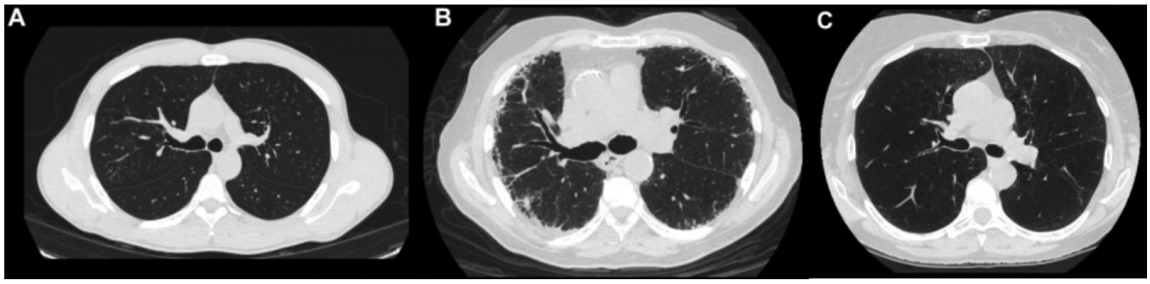

Figure 2.1 Representative CT images from each of the three imaging clusters: A, Preserved. B, Interstitial Predominant. C, Emphysema Predominant.

2.4.2 Correlation between COPD imaging clusters and clinical characteristics

In the COPDGene cohort, the individuals in the preserved cluster tended to have normal spirometry, normal six-minute walk distance (6MWD) and preserved respiratory health status as assessed by St. George's Respiratory Questionnaire (SGRQ) (Figure 2.2 A). Individuals in the emphysema cluster have expiratory airflow obstruction, reduced 6MWD, and the lowest respiratory health quality of life, while the individuals in the interstitial cluster were generally intermediate between the preserved and emphysema clusters in these characteristics. With longitudinal follow-ups, individuals of the emphysema cluster had the highest mortality followed by the interstitial and then the

preserved cluster (Figure 2.2 B). Though DECAMP cohort does not include 6MWD and SGRQ as part of the measurement, it also showed there was a significant reduction of FEV1% in individuals from the preserved, to the interstitial predominant, and the emphysema predominant (Table 2.2).

2.4.3 Genes differentially expressed between the preserved and the emphysema predominant clusters

We analyzed the bronchial epithelial gene expression associated with the imaging cluster using a subset of individuals from DECAMP (N = 224) and identified 179 genes that were differentially expressed between the preserved and emphysema clusters (FDR<0.25) (Figure 2.3). 99 genes were expressed at lower levels in the airway epithelial cells from individuals in the emphysema cluster relative to the preserved cluster and 80 genes were expressed at higher levels in the emphysema cluster relative to the preserved cluster. Genes expressed at lower levels in individuals from the emphysema cluster were genes involved in pathways that play regulatory roles in the regulation of T-cells and ribosomal small subunit assembly (Table 2.3); whereas genes expressed at higher levels in individuals from the emphysema cluster were genes related to inflammation and cell adhesion pathways.

To more fully characterize the biology of the gene expression differences associated with the imaging clusters, GSEA was performed on pre-ranked gene lists created by the pair-wise comparison between individuals of the emphysema predominant cluster and the

preserved cluster to identify enrichment of Hallmark genesets from the MsigDB databases. Gene sets with significant enrichment (Family-wise error rate < 0.05) from individuals in the emphysema cluster relative to those in the preserved cluster included up-regulated pathways involved in the inflammatory response, androgen response, and TNF-α signaling via NF-κB pathways (Table 2.4). While interferon alpha response, allograft rejection, and pancreas beta cells pathways related genes were shown to be enriched in individuals in the preserved cluster.

2.4.4 Gene expression profile of the individuals of the interstitial cluster

There are no differentially expressed genes between the interstitial cluster and either the preserved or the emphysema cluster. To further characterize the interstitial cluster, gene set variation analysis (GSVA) was performed, using the genes differentially expressed between the emphysema and preserved clusters, to summarize the expression of the emphysema-increased and emphysema-decreased genes in each of the three groups (Figure 2.4). The 80 genes with increased expression in the emphysema cluster relative to the preserved cluster were expressed at an intermediate level in the interstitial cluster patients, significantly different from either the preserved or the emphysema cluster ($P < 0.001$). Similarly, the 99 genes with decreased expression in the emphysema cluster relative to the preserved cluster were also expressed at an intermediate level in the interstitial cluster patients ($P < 0.001$).

2.4.5 Correlation between the differentially expressed genes and interferon-beta

Interestingly, the gene expression patterns of these 179 genes seem to correlate with that of peripheral blood mononuclear cells (PBMC) treated by Interferon-beta, in both directions (GSE26104)[75]. The 80 genes highly expressed in the emphysema cluster are expressed at higher levels in the PBMCs of patients with multiple sclerosis (MS) after interferon-beta treatment, while the 99 genes highly expressed in the preserved cluster are expressed at lower levels in the same patients after Interferon-beta treatment (Figure 2.5). The same trend was also observed in at least 4 other studies on the effect of Interferon-beta on a variety of cell types including hepatocytes, fibroblasts, and bronchial epithelial cells (GSE48400[76], GSE125066[77], GSE19392[78]).

2.4.6 Compare and contrast of previous genes differentially expressed in patients with clinical COPD

A previous study from our lab used gene expression profiling of bronchial brushings obtained from 238 individuals with and without COPD to examine genes differentially expressed in patients with COPD[39]. This analysis identified 98 genes, which were associated with spirometrically defined COPD status, FEV1%, and FEV1/FVC. 44 and 54 genes were up- and down-regulated in patients with COPD, respectively. Comparing these genes to the 179 gens we discovered, we found very limit overlaps between them – only three genes up-regulated in patients with COPD from Steiling et al were also up-regulated in individuals of the emphysema predominant cluster (PTGS2, LCN2, IRAK3).

While most of the genes previously detected to be associated with clinical COPD in the bronchial epithelium were not differentially expressed in individuals with emphysema predominant cluster, they did show strong correlations to the gene expression profiles of the emphysema predominant imaging cluster ($p < 0.001$, Figure 2.4). Here, we ranked genes based on the comparison between the emphysema cluster relative to the preserved cluster. We then used GSEA to determine if the genes increased or decreased in association with spirometrically defined COPD are enriched among the genes at the top and bottom of this ranked list. The 44 genes up-regulated in patients with COPD were in general expressed at higher levels among individuals within the emphysema predominant cluster. Conversely, the 54 genes down-regulated in patients with COPD were in general expressed at lower levels among the same individuals

2.4.7 Estimation of immune cell type abundance from bulk tissues

Functional analysis of the gene expression profiling associated with the emphysema predominant cluster suggests a potential involvement of immune cells. We further examined this correction by determining cell type abundance and expression from bulk tissues with digital cytometry using CIBERSORTx. We estimated abundance scores of 22 different types of immune cells using the LM22 signature gene file provided by the CIBERSORT online resource and focused on the abundance of T-cells and neutrophils (Figure 2.7). We found an increase of T-cell abundance from the emphysema predominant to the interstitial and the preserved cluster, as well as a decrease presence of neutrophil abundance in the preserved clusters.

Table 2.1 Clinical characteristics of COPDGene subjects in the three imaging clusters. The mean and standard deviation are shown for continuous variables. The count and proportion are shown for categorical variables. *p values calculated using an Anova test or Fisher's exact test.

	Preserved (N = 2623)	Interstitial Predominant (N = 1910)	Emphysema Predominant (N = 740)	*P* value*
Age (mean (SD))	60.22 (9)	59.97 (9)	65.57 (8)	< 0.001
Sex				
Male (%)	987 (80)	1364 (83)	431 (58)	<0.001
Race				
Black (%)	987 (37.6)	1364 (71.4)	431 (58.2)	<0.001
Smoking Status				
Current (%)	1144 (43.6)	1087 (56.9)	153 (20.7)	<0.001
Pack-years (mean (SD))	39.11 (21.09)	48.35 (28.14)	55.57 (26.81)	<0.001
Body Mass Index (kg/m2) (mean (SD))	27.92 (5.56)	30.97 (6.22)	25.47 (5.00)	<0.001
FEV1 % predicted (mean (SD))	86.58 (20.07)	73.14 (22.08)	40.62 (19.91)	<0.001
Radiologic Measures				
Interstitial Features (Percent Lung) (mean (SD))	4.55 (2.66)	7.89 (5.37)	5.34 (3.27)	<0.001
Emphysema (Percent Lung) (mean (SD))	3.66 (5.42)	4.97 (6.63)	48.65 (15.22)	<0.001
Pectoralis Muscle Area (cm2) (mean (SD))	36.26 (12.38)	48.64 (16.66)	30.94 (10.63)	<0.001
Airway Wall Thickness (mm) (mean (SD))	0.91 (0.13)	1.25 (0.19)	1.08 (0.22)	<0.001

Table 2.2 Clinical characteristics of DECAMP subjects in the three imaging clusters. The mean and standard deviation are shown for continuous variables. The count and proportion are shown for categorical variables. *p values calculated using an Anova test or Fisher's exact test.

	Preserved (N = 141)	Interstitial Predominant (N = 153)	Emphysema Predominant (N = 66)	*P* value*
Age (mean (SD))	63.91 (8)	66.14 (8)	68.11 (6)	0.001
Sex				
Male (%)	97 (68.8)	131 (85.6)	58 (87.9)	<0.001
Race				
Black (%)	13 (10.7)	32 (22.4)	12 (19.7)	0.041
Smoking Status				
Current (%)	68 (51.5)	64 (45.1)	24 (38.7)	0.227
Pack-years (mean (SD))	47.01 (26)	49.08 (26)	52.49 (27)	0.381
Body Mass Index (kg/m2) (mean (SD))	27.49 (6.01)	28.52 (6.13)	24.40 (5.40)	<0.001
FEV1 % predicted (mean (SD))	80.13 (17.27)	73.49 (18.12)	54.61 (19.61)	<0.001
Radiologic Measures				
Interstitial Features (Percent Lung) (mean (SD))	7.06 (4.50)	12.62 (9.31)	6.81 (8.18)	<0.001
Emphysema (Percent Lung) (mean (SD))	2.81 (2.77)	13.08 (8.32)	52.18 (16.60)	<0.001
Pectoralis Muscle Area (cm2) (mean (SD))	43.04 (13.62)	47.60 (12.87)	39.37 (10.08)	<0.001
Airway Wall Thickness (mm) (mean (SD))	2.10 (0.35)	2.22 (0.35)	2.01 (0.36)	<0.001

Table 2.3 Functional analysis of differentially expressed genes between individuals of the preserved and the emphysema predominant imaging clusters. Genes involved in listed pathways were listed. All functional enrichments listed here had a false discovery rate < 0.05.

Gene Cluster (number of genes)	GO-term	Description	Genes
1 (N=45)	GO:0035725	Sodium ion transmembrane transport	ACTN1, EPDR1, FZD8, GRIK2, SCN1B,SCNN1B, PCDHA10 PCDHB9, PCDHB14, PCDHGC5, SLC5A5, TMEM108,
	GO:0007156	Homophilic cell adhesion	
	GO:0050808	Synapse organization	
	GO:0007155	Cell adhesion	
	GO:0007267	Cell-cell signaling	
2 (N=35)	GO:0090023	Positive regulation of neutrophil chemotaxis	CCL20, CXCL2, CXCL3, CXCL5, CXCL8
	GO:0045236	CXCR chemokine receptor binding	
	GO:0008009	Chemokine activity	
3 (N=58)	GO:1902715	Positive regulation of interferon-gamma secretion	CCL5, CCR2, CD2, CD244, CD3E, CD3G, LCK
	GO:0042608	T cell receptor binding	
	GO:0010820	Positive regulation of T cell chemotaxis	
4 (N=22)	NA		
5 (N=19)	GO:0000028	Ribosomal small subunit assembly	RPL12, RPL13A, RPL29, RPL31, RPLP1, PRS3, RPS4X, RPS110, RPS19
	GO:0006614	SRP-dependent co-translational protein	
	GO:0000184	Nuclear-transcribed mRNA catabolic process	

Table 2.4 Gene set enrichment analysis showed enrichment of Hallmark genesets in participants within the emphysema predominant and the preserved clusters. All functional enrichments listed here had a family-wise error rate < 0.05.

Pathways Enriched in Participants within the Emphysema Predominant Cluster			
	Enrichment Scores	Normalized Enrichment Scores	FWER p-val
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.69	2.48	<0.001
HALLMARK_ANDROGEN_RESPONSE	0.56	1.84	0.003
HALLMARK_INFLAMMATORY_RESPONSE	0.50	1.82	0.005
HALLMARK_MTORC1_SIGNALING	0.48	1.76	0.018
HALLMARK_HYPOXIA	0.49	1.74	0.023
Pathways Enriched in Participants within the Preserved Cluster			
	Enrichment Scores	Normalized Enrichment Scores	FWER p-val
HALLMARK_INTERFERON_ALPHA_RESPONSE	-0.63	-2.16	<0.001
HALLMARK_ALLOGRAFT_REJECTION	-0.56	-2.05	<0.001
HALLMARK_PANCREAS_BETA_CELLS	-0.75	-1.91	0.002
HALLMARK_INTERFERON_GAMMA_RESPONSE	-0.48	-1.75	0.015

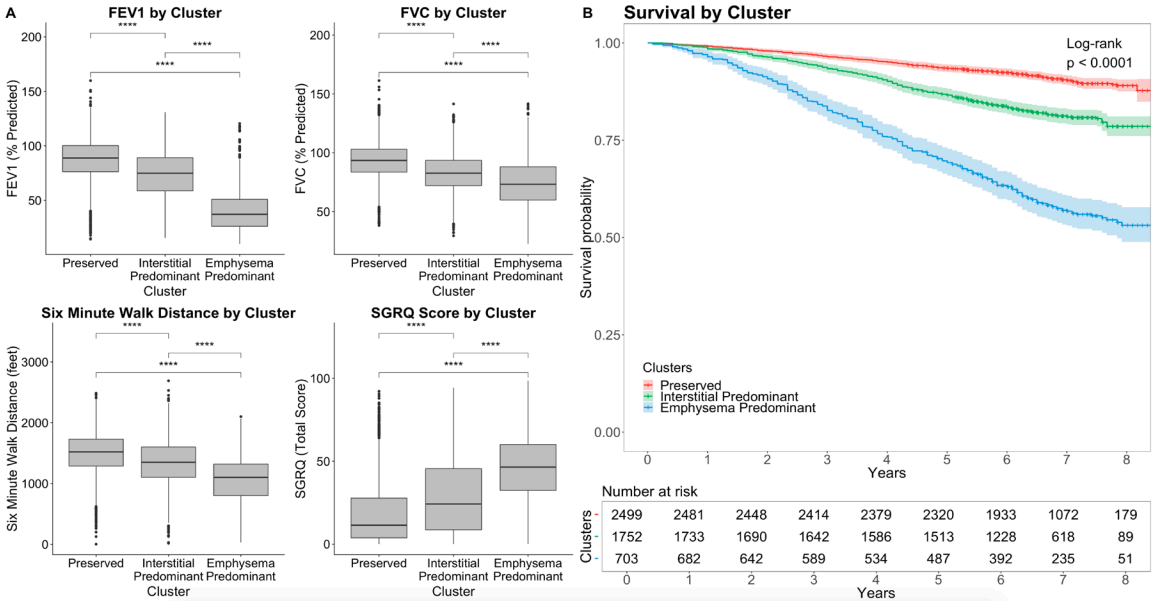

Figure 2.2 Comparison of Clinical Characteristics and Mortality of the Clusters in the COPDGene Cohort. A. The clinical characteristics identified in COPDGene were compared between the three clusters. B. The survival of the three clusters identified in COPDGene is demonstrated in this Kaplan-Meier curve. Individuals in the emphysema-predominant cluster had the lowest 5-year survival while individuals in the preserved cluster had the highest 5-year survival. Global differences for each clinical characteristic among the three clusters were assessed using ANOVA and found to be statistically significantly different (ANOVA $P < 0.001$). Pairwise differences were assessed using t-tests. mm = millimeter; cm2 = square centimeter; ANOVA = analysis of variance. Symbols for pairwise comparisons: ns = $P > 0.05$; * = $P \leq 0.05$; ** = $P \leq 0.01$; *** = $P \leq 0.001$; **** = $P \leq 0.0001$.

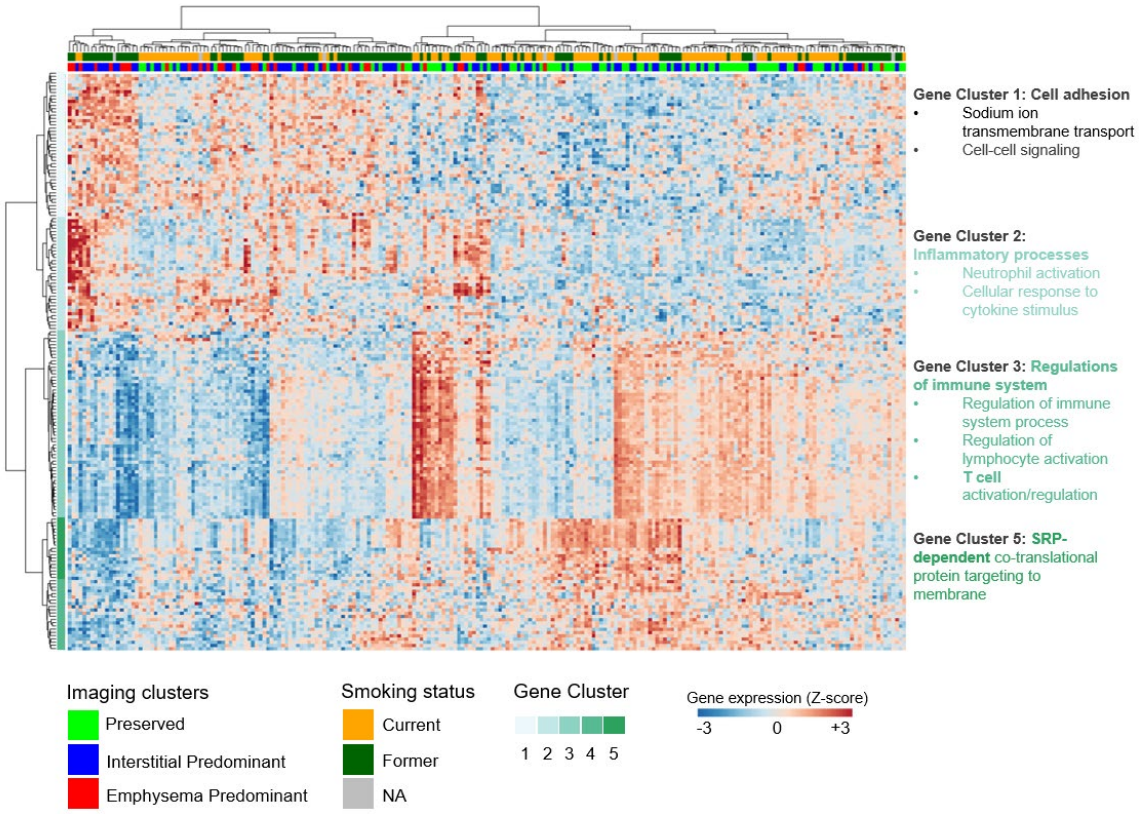

Figure 2.3 Unsupervised heatmap of the 179 genes differentially expressed between individuals of the emphysema and preserved clusters. Based on hierarchical clustering, genes were grouped into five gene clusters (1-5). Biological pathways in which these clusters of genes were enriched were shown on the side. False discovery rate < 0.1; Fold change > 0.25.

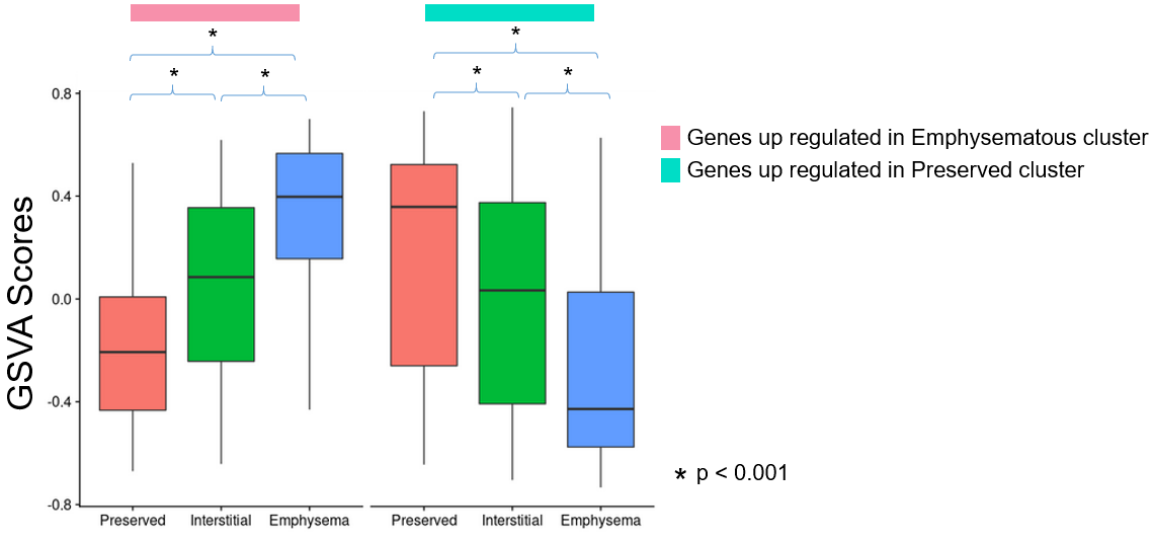

Figure 2.4 Gene set variation analysis was used to summarize the expression of each gene cluster in each sample. Variation in these summary scores was then examined as a function of imaging clusters. For both gene clusters, there was a significant difference between the imaging clusters by analysis of variance (each, $P < .001$). Post-hoc Tukey's honestly significant difference test was applied to examine the pairwise differences between groups. $*P \leq .05$; $**P \leq .01$.

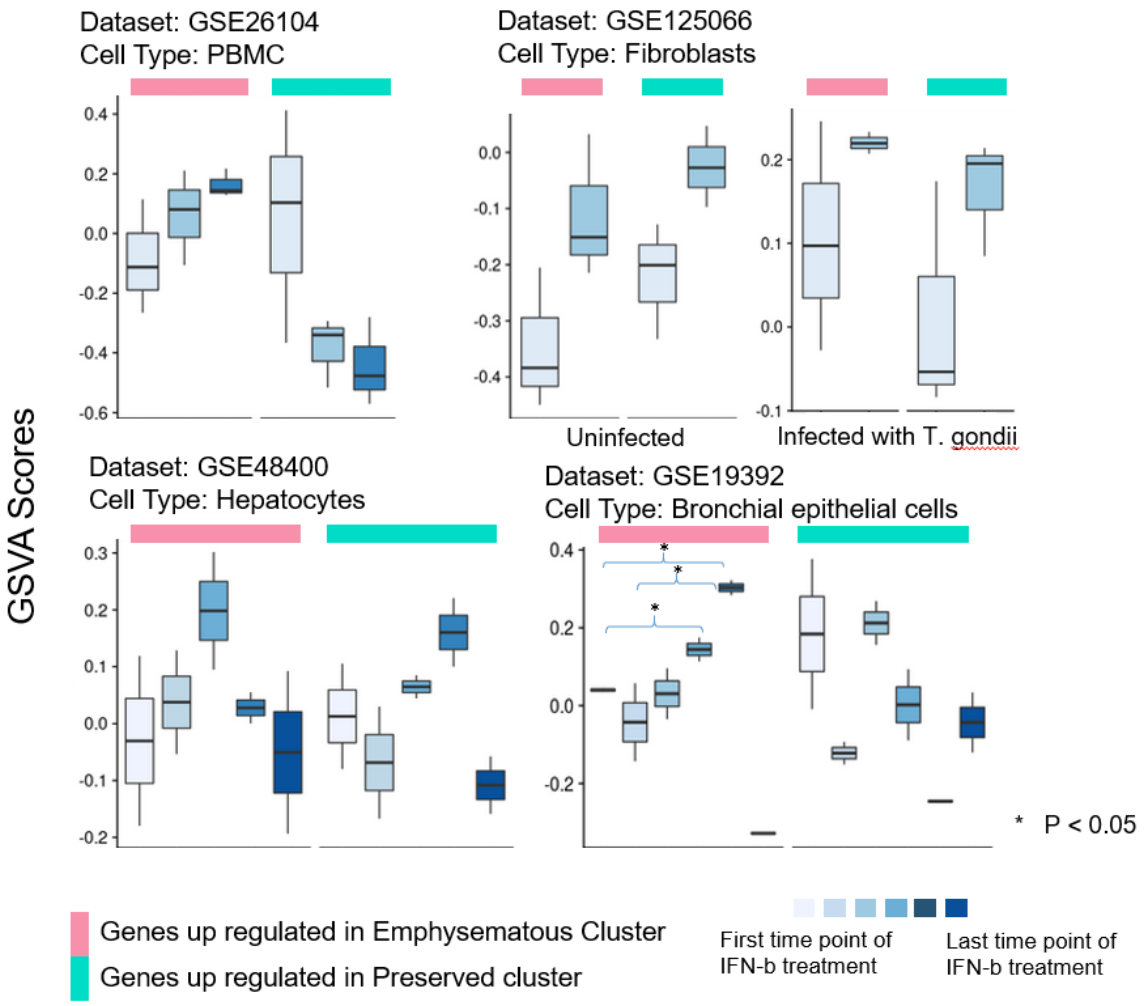

Figure 2.5 Correlation between the differentially expressed genes and interferon-beta. Gene set variation analysis was performed in each sample treated by interferon-beta, measured at different time points. In general, interferon-beta treatment of a variety of cell types induced an increased expression of genes up-regulated in emphysematous clusters, and a decreased expression of genes up-regulated in the preserved cluster.

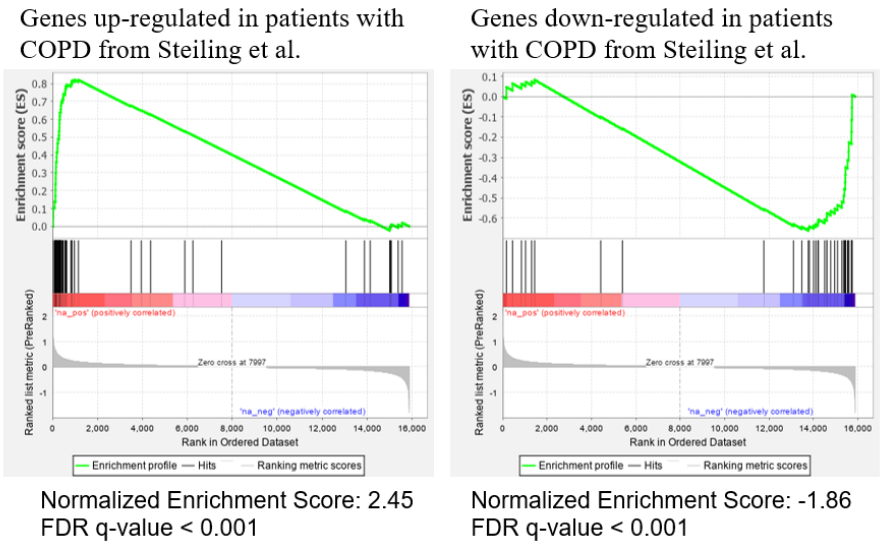

Figure 2.6 GSEA results assessing the enrichment of the 98 genes with relation to clinical COPD in participants within the emphysema predominant cluster on *t* statistics. Each vertical bar represents a single gene within a gene set and its occurrence among the rank like.

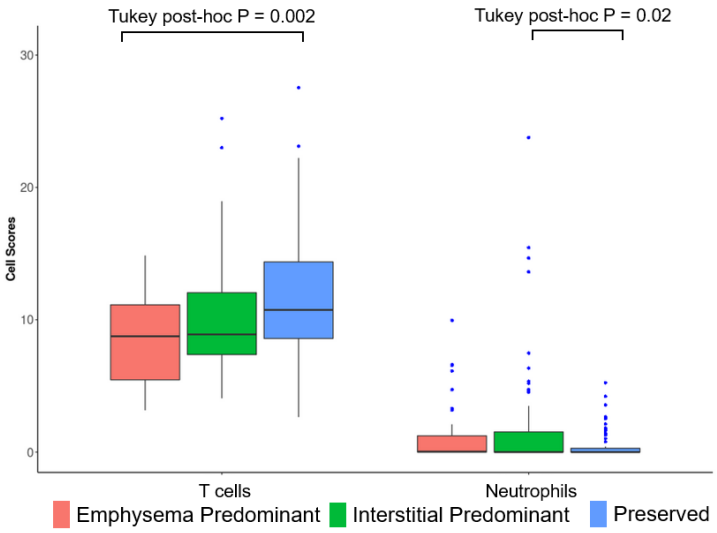

Figure 2.7 Estimated enrichment of T cells and neutrophils in individuals from the three imaging clusters. CIBERSORTx was used to derive the enrichment of 22 different kinds of immune cells present among bulk RNA-sequencing samples. Post-hoc Tukey's honestly significant difference test was applied to examine the pairwise differences between groups.

2.5 Discussion

Using quantitative CT imaging from the COPDGene cohort, we identified three subgroups of individuals associated with specific clinical characteristics. Individuals classified in the emphysema cluster have worse physiologic measures, reduced exercise capacity, worse respiratory health-related quality of life, and lower survival. Those who are classified in the preserved cluster have little-to-no emphysema, relatively normal physiologic measures, a normal exercise capacity and symptoms assessment, and a higher survival when compared to the emphysema cluster; and the interstitial cluster is composed of individuals whose level of disease is intermediate in severity. We identified the same imaging clusters in an independent cohort (DECAMP), corresponding to a similar pattern of clinical characteristics. Moreover, by leveraging bronchial epithelial gene expression from patients in the DECAMP cohort, we identified differential gene expression patterns when comparing the emphysema cluster to the preserved cluster that replicates previously published signatures generated from spirometrically derived COPD[39].

We found 179 genes were differentially expressed between individuals of the preserved and the emphysema predominant clusters. Functional analysis revealed that the genes up-regulated in individuals of the emphysema predominant cluster were enriched in sodium ion transmembrane transport, cell adhesion, and cell-cell signaling pathways. And genes up-regulated in individuals of the preserved cluster were enriched in positive regulation of interferon-gamma secretion, T cell regulation, and ribosomal small subunit assembly activities. Gene set enrichment analysis (GSEA) showed that genes of the TNF-α signaling via NF-κB pathway were strongly enriched in those within the emphysema

predominant cluster, whereas genes of the interferon alpha and gamma response pathway were strongly enriched in those within the preserved predominant cluster.

The TNF-α signaling via NF-κB pathway has been implicated in many different processes in the body and has been specifically shown to play a central role in airway inflammation in asthma and COPD. Many NF-κB-mediated processes are insensitive to the actions of steroids and some have proposed targeting NF-κB signaling as a potential intervention for steroid-refractory airway disease[79–81]. Our current work suggests that such an intervention may be of particular importance in those with emphysema predominant disease.

Type I and II interferons (interferon alpha and gamma) intrinsically promote an antimicrobial state in cells that are infected as well as neighboring cells to help limit the spread of infection[82]. Furthermore, they stimulate the adaptive immune system via antigen presentation and natural killer cell activation in response to infection[83]. But they also play a role in non-infected cells as they are constitutively secreted in low amounts by many tissues to keep the cell primed for future responses[84,85]. Here, we found genes enriched in the interferon alpha and gamma response pathways were lowly expressed among individuals within the emphysema cluster. Similar to our findings, previous reports have shown a relative decrease in the interferon response in patients with COPD from BAL fluid[86] and resected lung tissue[87]. We hypothesize that the down-regulation of the interferon alpha and gamma pathways plays a role in the pathogenesis of COPD and emphysema, both directly and indirectly. First, the down-regulation of constitutive baseline expression of interferon alpha and gamma leads to an enhanced susceptibility to infection which may

explain why patients with emphysema are more prone to infection[88,89]. Secondly, a down-regulation of the interferon alpha and gamma response may directly result in tissue injury, destruction, and remodeling and may explain the documented correlation between exacerbation frequency and emphysema progression both in our study and in prior work[90–92].

It is worth pointing out, though interferon alpha and beta are both subtypes of Type I interferon, their activity seemed to be inversely related in our analysis. Specifically, while genes enriched in the interferon alpha pathway were lowly expressed among individuals of the emphysema cluster, genes highly expressed among these individuals seemed to correlate with those that were up-regulated as the result of interferon beta treatment. Gene expression signatures from four published datasets that involved the response to interferon beta were identified as concordant (GSE26104, GSE19392, GSE125066, and GSE48400); we found that the gene set variation analysis scores from the signature of genes increased in the emphysema cluster is significantly increased in bronchial epithelial cells after interferon beta treatment (GSE19392). We also found a similar increase in the gene set variation analysis scores from the emphysema-increased signature in datasets that examined the response of peripheral blood mononuclear cells, hepatocytes, and fibroblasts to interferon beta. Previously, interferon alpha and beta have been reported to exhibit key differences in several biological properties. Interferon beta, but not interferon alpha induces the association of tyrosine-phosphorylated receptor components ifnar1 and ifnar2, and has activity in cells lacking the interferon receptor-associated, Janus kinase tyk2[93]. It has also been reported that interferon alpha and beta differ by their capability of clearing chronic

hepatitis C virus (HCV) infection[94] with interferon beta being capable of clearing HCV at a faster rate. Thus, we hypothesize the increased expression related to interferon beta among individuals with emphysema may indicate an active infection associated with COPD that requires a more potent inflammatory response.

The activation of TNF-α signaling and interferon beta pathways has implications in a possible shift in immune cell composition within the bronchial airways of individuals of the emphysema predominant cluster. TNF-α activation directly leads to the release of multiple chemokines that induce neutrophil influx to inflammatory foci[95, 96, 97]. Meanwhile, while the immunomodulatory mechanism of action is not fully understood, interferon beta activation reduces CD4 and CD8 T cell reactivity[98], inhibits T cell activation and proliferation[99,100], and has been shown to prevent transmigration of autoreactive T cells into the central nervous system[101]. To test whether an increase in neutrophils and a decrease in T cells are present in individuals within the emphysema predominant cluster, we derived cell enrichment scores using CIBERSORTx, a tool that allows for computational deconvolution of bulk tissues into an abundance of cell types. The result does provide some evidence that T cells are indeed less abundant in individuals within the emphysema predominant cluster, and that neutrophils are less abundant in individuals within the preserved cluster.

Previously, Steiling et al discovered a set of 98 genes that correlated with spirometrically defined COPD status. Between these 98 genes and the 179 genes we discovered, only three genes were overlapped. The overall trend of expression of the 98 genes, however, seems to be coherent using GSEA where the 44 genes up-regulated in

patients with spirometrically defined COPD are highly expressed in individuals within the emphysema predominant cluster, and vice versa. Overall, the comparison between the gene lists suggests the patient subgroups defined by pulmonary function test and CT imaging are not completely identical, but there exist share molecular mechanisms between the two. Spiromertrically defined COPD may have resulted from a plethora of pathologic structural changes of the lung parenchyma, including the emphysematous processes that characterize the emphysema predominant imaging cluster in this current analysis. An alternative cause of few genes overlap between the two analyses lies upon the difference between the methods of data collection – RNA-sequencing samples included in Steiling et al were profiled using Affymetrix Human Gene 1.0 ST Arrays, whereas the samples in DECAMP were profiled using Illumina Sequencing Platform.

A central question raised by this work and other COPD phenotyping efforts is how to interpret the groups of disease identified. One possibility is that the radiographic subtypes represent temporal disease stages: i.e. individuals from the preserved group with progressive disease move into the interstitial group and then to the emphysema group. In this model, the interstitial features identified on CT likely represent inflammation and edema that precedes the development of emphysema in some cases[102,103,104]. If this hypothesis is correct, the identification of interstitial features may enable earlier disease detection and prognostication. Both the GSVA of the differentially expressed genes and the computationally derived cell proportions for T cells and neutrophils show individuals of the interstitial predominant cluster represent an intermediary group between the preserved and the emphysema predominant clusters, lending support to this model.

Rather than seeing the interstitial group as an intermediate or early disease phenotype between the preserved and emphysema groups, it is also possible that these groups represent three distinct phenotypes of response to cigarette smoke: a group relatively resistant to smoking, another group that has evidence of inflammation and experiences the development of airway and interstitial disease, and a third group that experiences the development of progressive emphysema. This mirrors the clinical observation that various lung-function trajectories lead to COPD[104] and suggests that these different radiographic phenotypes may reflect distinct pathogenic mechanisms[105] that all result in a common physiologic abnormality (ie, airflow limitation). In this schema, a subset of the second group, those with interstitial predominance, may have early or subtle pulmonary fibrosis and may go on to experience more advanced fibrotic disease, as has been suggested for patients with visually defined interstitial lung abnormalities[103].

One of the strengths of our study is the ability to identify imaging clusters that replicate across two separate cohorts, despite differences in patient population and image processing. Another strength is the ability to associate imaging cluster membership with gene expression differences in the bronchial airway epithelium because these gene expression associations argue for an underlying biologic cause for the imaging-based subgroups and begin to suggest the molecular processes that differentiate the groups.

Our study did have several limitations as well. The two cohorts are not identical and have differences in inclusion and exclusion criteria. For instance, COPDGene excluded patients with interstitial lung disease, which potentially could limit the ability to draw accurate conclusions regarding interstitial features. No such exclusion criterion was

included in the DECAMP study. Also, and notably, the demographics of the two cohorts are quite different, as are several of the imaging variables, with the latter differences likely related both to demographics and differences in image acquisition techniques. The replication of the clusters in the DECAMP study suggests that they may be robust to these differences; however, additional work is needed in other cohorts to determine whether this is the case. Finally, we selected imaging variables based on prior knowledge and experience. An unsupervised learning algorithm may be used to generate imaging clusters that do not depend on the domain knowledge associated with COPD.

2.6 Conclusion

In conclusion, clustering smokers using quantitative CT imaging-based measures in two distinct cohorts enabled the identification of three subgroups of disease which have organ specific molecular correlates. These subgroups differ from those defined by the pulmonary function test. Gene expression differences in individuals within the emphysema predominant group reflect biological processes that have been related to COPD and suggest prominent roles of the immune cells in emphysema. Further work is needed to better understand the significance of these clusters, the pathophysiologic and molecular differences between them, and their potential utility for defining disease prognosis and management. Additional understanding of the similarity and differences in interferon pathways may also provide more insight into the interplay between the chemokines and cytokines that may be important in COPD treatment.

CHAPTER THREE:

Gene expression alterations associated with HRCT-derived radiographic bronchiectasis

Disclaimer: Part of the figures and text in this chapter is being prepared for publication.

3.1 Abstract

Bronchiectasis (BE) is an increasingly recognized disease characterized by pathologic dilation of airways, yet its pathophysiology is poorly understood. Identifying airway gene expression alterations associated with radiographic BE may provide insights into the molecular changes associated with BE initiation.

173 subjects without a prior clinical diagnosis for BE, with or without radiographic BE, were examined. We detected widespread radiographic BE (in 3 or more lobes) in 20 of 173 participants, who presented with more cardinal bronchiectatic symptoms such as cough and mucus production. Transcriptomic assessment of bronchial epithelial cells revealed 298 genes with roles in cilium organization and endopeptidase activity pathways were up-regulated in participants with widespread radiographic BE; while 357 genes involved in cell adhesion and Wnt signaling pathways were down-regulated in them. Leveraging an independent single-cell RNA-seq dataset of bronchial epithelial cells, we found genes up-regulated in participants with radiographic BE were expressed at higher levels in ciliated and deuterosomal cells; and genes down-regulated were expressed at higher levels in basal cells. Deconvolution of the bulk RNA-seq samples into proportions of various cell types also showed an increased presence of ciliated and deuterosomal cells, as well as a decreased presence of basal cells in subjects with widespread radiographic BE, respectively.

The regulatory pattern suggests a compensatory response of producing more ciliated cells in an inflammatory environment, accompanied by a loss of surface integrity in the early stage of BE. While requiring independent confirmation and longitudinal studies, our findings provide the first pathophysiological concept for radiographic BE to be better understood.

3.2 Introduction

Bronchiectasis (BE) is a pathologic dilation of bronchi[1]. Once considered an orphan disease that affects less than 200,000 people in the United States (US), BE has become increasingly common in the US, with over 70,000 new diagnoses in 2013 and between 340,000 and 522,000 prevalent cases requiring treatment[2]. Computed tomography (CT) is not only the gold standard for diagnosing BE in patients with symptomatic disease, it can also detect bronchial abnormalities in asymptomatic or mildly symptomatic individuals that are scanned for unrelated reasons. A term used for this situation is radiographic BE[1,3]. Molecular profiling of the airway epithelium in patients with radiographic BE may provide insights into the pathogenesis of BE and the possible biological pathways not yet targeted by current treatments[4–7].

Previous studies of bronchoalveolar lavage fluid and sputum have detected gene expression alterations or protein level changes related to bronchiectasis[8–10]. Yet, these studies were limited by patient numbers and a small number of genes or proteins as compared to transcriptomic profiling through RNA sequencing (RNA-seq), which allows

for a comprehensive analysis of gene expression. The utility of gene expression in BE has not been sufficiently explored, possibly due to a paucity of pertinent samples at the site of the abnormality such as endobronchial tissue. Previously, our group has used the bronchial tissue at the mainstem bronchus, from which gene expression alterations in normal-appearing airway epithelial cells could be detected in association with smoking[11], lung cancer[12], and chronic obstructive pulmonary disease[13,14]. Importantly, these gene expression alterations not only served as diagnostic biomarkers but have provided insights into the molecular events related to these lung pathologies. We hypothesize that mainstem bronchus epithelial gene expression would be especially well suited to the study of radiographic BE due to the proximity between the profiled biospecimen and the site of disease.

Here, we aimed to characterize gene expression alterations in radiographic BE using bronchial epithelium cells and explore whether the genes show cell-type dependent expression leveraging a single-cell RNA-seq dataset of airway epithelium at the mainstem bronchus. Moreover, we examined the relationship between gene expression and clinical characteristics.

3.3 Material and Methods

3.3.1 Study Participants and Sample Analysis

The participants of this analysis were recruited as part of the two DECAMP cohorts, described in detail in Section 2.3.1. Pertinent to this analysis is that none of the participants

of the DECAMP cohorts had a clinical diagnosis of BE (participants with chronic lung disease other than COPD were excluded).

From the large cohort of DECAMP of 360 subjects (169 from DECAMP-1 and 191 from DECAMP-2), 129 participants from DECAMP-1, and 44 participants from DECAMP-2 with matching RNA-seq from right mainstem bronchus and CT scans were available for analysis as of February 2021 (**Figure 3.1**).

3.3.2 HRCT acquisition and characterization of radiographic bronchiectasis

The acquisition of HRCT was described in detail in Section 2.3.2. Pertinent to this study is that the images were adequate to assess bronchiectasis. The detection of bronchiectasis was visually performed by a single reader, a pulmonologist (AD) with over 10 years of experience in lung imaging. In brief, bronchiectasis on CT was coded as yes, no, or equivocal.

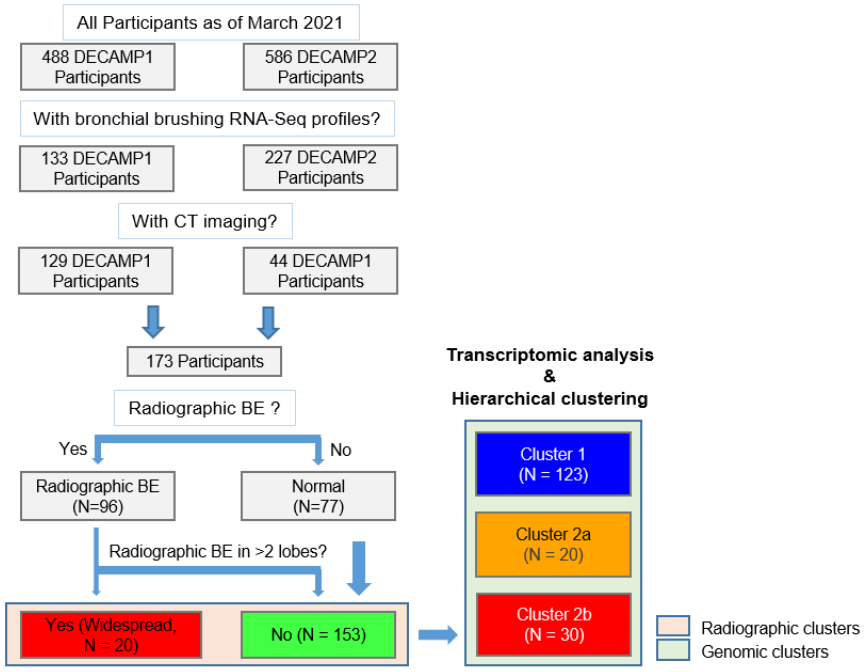

Figure 3.1 Schematic representation of participant clustering based on imaging and gene expressions.

Radiographic bronchiectasis was defined with one or more of the following criteria: **a)** airway dilation (airway lumen diameter greater than adjacent pulmonary vessel diameter)**; b)** abnormal airway tapering of any extent (no decrease in or increase in lumen moving from proximal to distal airways); and **c)** visualization of a bronchus within 1 cm of the pleura. A CT case was defined with at least one lobe meeting the above BE criteria. The lingula was considered a separate lobe. We further defined widespread radiographic BE when 3 or more lobes were involved.

3.3.3 Analytic strategies of differential gene expression

The preprocessing steps of the analysis were described in detail from Sections 2.3.4 to 2.3.5. The corrected counts were then filtered based on counts per million (CPM) such that a gene could only be included if its CPM was greater than 1 in 10% of the total number of patients. False discovery rate (FDR) of 0.1 and log fold change (logFC) of 0.25 were used to filter significantly differentially expressed genes. Heatmaps were used to visualize the data and identify unsupervised participant clusters using the "Ward.D2" algorithm.

To identify genes differentially expressed between subjects with and without radiographic BE, I first modeled gene expression on the presence of any lobe of the lung having radiographic BE, correcting for sex and smoking status:

(1) Gene ~ radiographic BE + sex + smoking + error

I then examined gene expression as a function of the presence of having widespread radiographic BE, also correcting for sex and smoking status:

(2) Gene ~ widespread radiographic BE + sex + smoking + error

3.3.4 Single cell RNA-sequencing workflow

The Seurat R package (version 3.1)[116] was used for downstream analyses including normalization, scaling, clustering of cells, and identifying cluster marker genes on a dataset from Deprez et al. [https://www.genomique.eu/cellbrowser/HCA/][117]. Cells collected from the tracheal epithelium (9 healthy volunteers, N = 20519) were selected for this analysis for matching anatomical locations. Cells were filtered out if they met any of the following criteria: 1) bottom quantile for a total number of genes detected, 2) bottom quantile for total library size and 3) 30% of counts mapped to the mitochondrial genome. Overall, 7343 cells were filtered and 13176 cells were kept for further analysis. UMAP dimensionality reduction was performed using the first 15 principal components with a resolution setting of 0.8. Cell types were previously assigned by Deprez et al and individually validated with previously reported cell markers. Gene set variation analysis was performed at a single-cell level.

3.3.5 Deconvolution of bulk RNA-sequencing samples

To dissect cell population proportions from bulk RNA samples, reference gene expression profiles (GEPs) were derived using the bronchial scRNA-seq data. Marker genes of each cell type were identified using the FindMarkers function within the Seurat package using the MAST method of modeling (FDR < 0.05, logFC > 0.25). After generating the GEPs, we applied an optimized function within the AutoGeneS package[118] to further identify the top 1000 most informative genes from Seurat selected genes. We then applied the deconvolve function within the AutoGeneS to predict cell proportion from the bulk RNA-seq data based on the 1000 genes with model parameter set to Nu Support Vector Regression (nusvr).

3.3.6 Statistical analysis

Data are presented as means and standard deviations for continuous measurements and number and percentage for categorical features. P values were calculated using a Student's T test, Fisher's exact test, Kruskal test, or ANOVA F test.

3.4 Results

3.4.1 Participant demographics, pulmonary function, and imaging measurement

Of the 173 evaluated participants, 96 participants had radiographic BE. Participants with and without radiographic BE showed similar clinical characteristics (**Table 3.1**). Radiographic BE was predominant in the lower lobes (47%), whereas the lingula was the

least affected (5.5%) (**Table 3.2**). There were 20 participants with widespread radiographic BE (3 females and 17 males; 66 $\pm$ 5 yr), and 153 participants with limited or no radiographic BE (29 females and 124 males; 67 $\pm$ 8 yr). We found participants with and without widespread radiographic BE showed similar clinical features except those with widespread radiographic BE were more likely to report shortness of breath (p-value = 0.01, **Table 3.3**).

3.4.2 Identification of three distinct clusters of participants based on gene expression profiles

We first performed differential gene expression analysis comparing bronchial epithelium of participants with and without radiographic BE and found no genes were differentially expressed (data not shown). We then compared participants with and without widespread radiographic BE and discovered 655 genes were significantly (false discovery rate q value < 0.1, log fold change > 0.25) differentially expressed when controlling for sex and smoking status (**Figure 3.2**). Unsupervised clustering using the 655 genes first separated the 173 participants into two genomic clusters. The predominant cluster (N = 123, light blue, left branch in **Figure 3.2**) was primarily composed of participants without widespread radiographic BE. The smaller cluster on the right branch of the dendrogram contained two subgroups of participants, one that included most of the participants with widespread BE (N = 30, red), and another which was composed of participants without

widespread BE, yet demonstrated gene expression patterns similar to those with widespread BE (N = 20, orange).

In addition to a different number of lobes with radiographic BE (p-value < 0.0001), these three clusters of participants differ by the likelihood of having cardinal symptoms associated with BE – cough and phlegm production. The red cluster of participants with the highest average of number lobes with radiographic BE had the highest proportion of participants complaining about both cough and phlegm (p-value = 0.002). Interestingly, the orange cluster had a higher proportion of current smokers (p-value = 0.006) but otherwise consists of individuals with similar clinical characteristics to the participants in the blue cluster with the lowest average number of lobes with radiographic BE (**Figure 3.2 and Table 3.4**). Based on differences both in the gene expression and clinical characteristics that correlated with an increasing presentation of symptoms related to bronchiectasis, we named these three participants' clusters normal (light blue), intermediate (orange), and bronchiectatic (red).

Further examination of the participants in the bronchiectatic cluster who do not have widespread radiographic BE, failed to identify significant differences in clinical characteristics compared to the participants with widespread BE within the bronchiectatic cluster (**Table 3.5**). However, when compared to the participants without widespread BE in the normal and intermediate clusters, non-BE participants in the bronchiectatic gene expression cluster exhibit an increased likelihood of cough and phlegm production (p-value = 0.02) (**Table 3.6**).

3.4.3 Functional analysis of differentially expressed genes by radiographic bronchiectasis

To better understand the differences in gene expression among the three patient clusters, especially that between the intermediate cluster and either the normal or bronchiectatic cluster, we first divided the 655 genes into five co-expression clusters (A-E) based on hierarchical clustering. A composite expression score for each of the five gene clusters for each participant was then calculated using gene set variation analysis (GSVA) (**Figure 3.3**). The normal and the bronchiectatic clusters showed clear distinctions in all gene clusters, gene clusters A and B were expressed at higher levels among participants of the normal cluster, and gene clusters C, D, and E were expressed at higher levels among participants of the bronchiectatic cluster. Participants in the intermediate cluster, however, were displaying two patterns of gene expression: one in which the intermediate and bronchiectatic clusters exhibited similar levels of gene expression relative to the normal cluster (gene cluster A, D, and E); and another pattern in which intermediate cluster exhibited gene expression intermediate between the normal and the bronchiectatic clusters (Gene clusters B and C).

Functional enrichment analysis showed that gene clusters A and B were significantly enriched for genes involved in cell adhesion and Wnt signaling, respectively (**Table 3.7**), whereas gene clusters C and E were enriched for genes involved in endopeptidase activity, and genes in cluster D were enriched for genes involved in cilium organization. When examining gene expression as a function of the number of affected

lobes, we also observed a dramatic shift in expression between 2 and 3 affected lobes (**Figure 3.4**).

To further explore the possible biological processes contributing to the BE-associated gene expression differences, we performed gene set enrichment analysis (GSEA) using a catalog of curated Hallmark gene sets[22]. We found genes up-regulated among participants with widespread radiographic BE were enriched in interferon-gamma, oxidative phosphorylation, and interferon-alpha pathways; while genes up-regulated in participants without widespread radiographic BE were enriched in pathway down-regulated by KRAS activation, epithelial-mesenchymal transition, and pancreas beta cells pathways (**Table 3.8**).

Moreover, we compared the differentially expressed genes to a signature of ciliogenesis, which contained a list of 310 genes up-regulated with cilia organization and associated with primary ciliary dyskinesia (PCD) [119,120], a significant risk factor for BE. 42 of the 310 genes were up-regulated among patients with widespread radiographic BE (**Table 3.9**). Using GSEA, we found significant enrichment of ciliogenesis-associated genes among the genes expressed at higher levels in participants with widespread radiographic BE (**Figure 3.5**).

3.4.4 Widespread radiographic BE correlates with increased proportions of ciliated and deuterosomal cells, and decreased proportions of basal cells

To explore whether the observed gene expression alterations might be specific to specific cell types in the bronchial epithelium, we leveraged a single cell RNA-sequencing dataset consisting of bronchial epithelial cells collected by bronchoscopic biopsy of 9 healthy volunteers[121] and calculated a per cell GSVA enrichment scores for the Cluster A & B genes as well as the Cluster C-E genes. We found the genes increased in individuals with widespread radiographic BE (Clusters A & B) were expressed almost exclusively among the deuterosomal cells and the multiciliated cells, whereas the genes up-regulated in individuals without widespread radiographic BE (Clusters C-E) were expressed at the highest levels among basal cells (**Figure 3.6 A-C**). To further explore the possibility that the gene expression alterations observed at the bulk level are consistent with altered epithelial cell prevalence in the bronchial airway of individuals with widespread radiographic BE, we computationally deconvolved cell population proportions in the bulk RNA-seq data using gene markers identified from the single-cell data. The predicted proportions of both the multiciliated and deuterosomal cells increased from the normal to the intermediate and the bronchiectatic cluster (**Figure 3.6 D**); with the shift being most pronounced for the immature deuterosomal cells. In contrast, the proportion of basal cells incrementally decreased from the normal to the intermediate and the bronchiectatic cluster. Of note, the proportion of goblet cells, previously shown to be correlated with cigarette smoking[121], was increased in the smoker-predominant intermediate cluster (**Figure 3.7**).

Table 3.1 Clinical characteristics of subjects with and without radiographic BE. The mean and standard deviation are shown for continuous variables. The count and proportion are shown for categorical variables. *p values calculated using a Student's t-test or Fisher's exact test. **Missing pack-years for 2 subjects with radiographic BE.

	Participants with Radiographic BE (N = 96)	Participants without Radiographic BE (N = 77)	*P* value*
Age (mean (SD))	67 (7)	67 (8)	0.89
Sex			0.70
Male (%)	77 (80)	64 (83)	
Female (%)	19 (20)	13 (17)	
Race			0.88
White (%)	71 (74)	58 (75)	
Black (%)	12 (13)	9 (12)	
Asian (%)	2 (2)	3 (4)	
Others/Unknown (%)	11 (12)	7 (9)	
Smoking Status			0.25
Current (%)	38 (40)	37 (48)	
Former (%)	55 (57)	35 (46)	
Unknown (%)	3 (3)	5 (7)	
Pack-years (mean (SD))	50 (25) **	50 (27)	0.97
FEV1 % predicted (mean (SD))	72 (20)	76 (21)	0.18
FEV1/FEV (mean (SD))	0.6 (0.1)	0.6 (0.1)	0.62
Cough			0.87
Yes (%)	42 (44)	35 (46)	
No (%)	46 (48)	35 (46)	
Unknown (%)	8 (8)	7 (9)	
Phlegm			0.87
Yes (%)	42 (44)	34 (44)	
No (%)	47 (49)	35 (46)	
Unknown (%)	7 (7)	8 (10)	
Shortness of Breath			0.74
Yes (%)	56 (34)	42 (55)	
No (%)	33 (58)	28 (36)	
Unknown (%)	7 (7)	7 (9)	

Definition of abbreviation: BE = bronchiectasis.
The mean and standard deviation are shown for continuous variables.
* *P* values calculated using a Student's t-test or Fisher's exact test.
**Missing pack-years for 2 subjects with Radiographic BE.

Table 3.2 Distribution of radiographic BE by lobe.

	No.	%
Left lung	64	35.2
Left upper lobe	24	13.2
Lingula	10	5.5
Left lower lobe	30	16.5
Right lung	118	64.8
Right upper lobe	39	21.4
Right middle lobe	24	13.2
Right lower lobe	55	30.2

Table 3.3 Clinical characteristics of subjects with and without widespread radiographic BE. The mean and standard deviation are shown for continuous variables. The count and proportion are shown for categorical variables. *p values calculated using a Student's t-test or Fisher's exact test. **Missing pack-years for 2 subjects without widespread radiographic BE.

	Participants with Widespread Radiographic BE (N = 20)	Participants without Widespread Radiographic BE (N = 153)	*P* value*
Age (mean (SD))	66 (5)	67 (8)	0.62
Sex			1
Male (%)	17 (85)	124 (81)	
Female (%)	3 (15)	29 (19)	
Race			0.63
White (%)	15 (75)	114 (74.5)	
Black (%)	3 (15)	18 (11.8)	
Asian (%)	1 (5)	4 (2.6)	
Others/Unknown (%)	1 (5)	17 (11.1)	
Smoking Status			1
Current (%)	9 (45)	66 (43.1)	
Former (%)	10 (50)	80 (52.3)	
Unknown (%)	1 (5)	7 (4.6)	
Pack-years (mean (SD))	45 (19)	51 (26)**	0.37
FEV1 % predicted (mean (SD))	69 (22)	75 (20)	0.23
FEV1/FEV (mean (SD))	0.6 (0.2)	0.6 (0.1)	1
Number of lobes with Radiographic BE (mean (SD))	3.4 (0.6)	0.8 (0.8)	**<2e-16**
Cough			1
Yes (%)	9 (45)	68 (44)	
No (%)	9 (45)	72 (47)	
Unknown (%)	2 (10)	13 (9)	
Phlegm			0.46
Yes (%)	11 (55)	65 (43)	
No (%)	8 (40)	74 (48)	
Unknown (%)	1 (5)	14 (9)	
Shortness of Breath			**0.01**
Yes (%)	17 (85)	81 (53)	
No (%)	2 (10)	59 (39)	
Unknown (%)	1 (5)	13 (9)	

Table 3.4 Clinical characteristics of the participants of the three clusters based on gene expression profiles. The mean and standard deviation are shown for continuous variables. The count and proportion are shown for categorical variables. *p values calculated using either an ANOVA F test or Fisher's exact test. **Missing pack-years for 1 subject within the normal cluster and 1 subject within the intermediate cluster.

Patient Clusters	Normal (N = 123)	Intermediate (N = 20)	Bronchiectatic (N = 30)	*P* value*
Age (mean (SD))	67 (8)	65 (6)	67 (6)	0.29
Sex				0.14
Male (%)	98 (80)	15 (75)	28 (93)	
Female (%)	25 (20)	5 (25)	2 (7)	
Race				0.22
White (%)	88 (72)	15 (75)	26 (87)	
Black (%)	17 (14)	2 (10)	2 (7)	
Asian (%)	2 (2)	2 (10)	1 (3)	
Others/Unknown (%)	16 (13)	1 (5)	1 (3)	
Smoking Status				**0.006**
Current (%)	46 (37)	15 (75)	14 (47)	
Former (%)	70 (57)	4 (20)	16 (53)	
Unknown (%)	7 (6)	1 (5)	0 (0)	
Pack-years (mean (SD))	50 (26)	61 (31)**	44 (20)**	0.09
FEV1 % predicted (mean (SD))	75 (20)	78 (20)	68 (21)	0.16
FEV1/FEV (mean (SD))	0.6 (0.1)	0.7 (0.1)	0.6 (0.2)	0.07
Widespread Radiographic BE				**1.18E-07**
Yes (%)	4 (3)	3 (15)	13 (43)	
No (%)	119 (97)	17 (85)	17 (57)	
Number of lobes with Radiographic BE (mean (SD))	0.9 (0.9)	1.1 (1.3)	1.8 (1.6)	**0.0001**
Cough				**0.0003**
Yes (%)	48 (39)	6 (30)	23 (77)	
No (%)	66 (54)	10 (50)	5 (17)	
Unknown (%)	9 (7)	4 (30)	2 (7)	
Phlegm				**0.02**
Yes (%)	49 (40)	7 (35)	20 (67)	
No (%)	63 (51)	11 (55)	8 (27)	
Unknown (%)	11 (9)	2 (10)	2 (7)	
Both cough and phlegm				**0.002**
Yes (%)	33 (27)	4 (20)	18 (60)	
No (%)	79 (64)	13 (65)	10 (33)	
Unknown (%)	11 (9)	3 (15)	2 (7)	
Shortness of Breath				0.15
Yes (%)	64 (52)	13 (65)	21 (70)	
No (%)	49 (40)	5 (25)	7 (23)	
Unknown (%)	10 (8)	2 (10)	2 (7)	

Table 3.5 Clinical characteristics of the participants with and without widespread BE in the bronchiectatic cluster. The mean and standard deviation are shown for continuous variables. The count and proportion are shown for categorical variables. *p values calculated using either an ANOVA F test or Fisher's exact test. **Missing pack-years for 1 subject without widespread radiographic BE.

Patient Clusters	Participants with Radiographic BE (N = 13)	Participants without Radiographic BE (N = 17)	*P* value*
Age (mean (SD))	68 (5)	67 (7)	0.71
Sex			
Male (%)	11 (85)	17 (100)	0.18
Female (%)	2 (15)	0 (0)	
Race			0.70
White (%)	10 (77)	16 (94)	
Black (%)	1 (8)	1 (6)	
Asian (%)	1 (8)	0 (0)	
Others/Unknown (%)	1 (8)	0 (0)	
Smoking Status			1
Current (%)	6 (46)	8 (47)	
Former (%)	7 (54)	9 (53)	
Unknown (%)	0 (0)	0 (0)	
Pack-years (mean (SD))	40 (14)	48 (23)**	0.30
FEV1 % predicted (mean (SD))	66 (21)	69 (21)	0.68
FEV1/FEV (mean (SD))	0.6 (0.1)	0.6 (0.2)	0.82
Cough			0.62
Yes (%)	9 (69)	14 (82)	
No (%)	3 (23)	2 (12)	
Unknown (%)	1 (8)	1 (6)	
Phlegm			1
Yes (%)	9 (69)	11 (65)	
No (%)	4 (31)	4 (24)	
Unknown (%)	0 (0)	2 (12)	
Both cough and phlegm			1
Yes (%)	8 (62)	10 (59)	
No (%)	5 (39)	5 (29)	
Unknown (%)	0 (0)	2 (12)	
Shortness of Breath			0.18
Yes (%)	11 (85)	10 (59)	
No (%)	1 (8)	6 (35)	
Unknown (%)	1 (8)	1 (6)	

Table 3.6 Clinical characteristics of the participants without widespread BE in the subgroups. The mean and standard deviation are shown for continuous variables. The count and proportion are shown for categorical variables. *p values calculated using either an ANOVA F test or Fisher's exact test. **Missing pack-years for 1 subject without widespread radiographic BE.

Patient Clusters	Normal (N = 119)	Intermediate (N = 17)	Bronchiectatic (N = 17)	*P* value*
Age (mean (SD))	68 (8)	65 (7)	67 (7)	0.37
Sex				
Male (%)	95 (80)	12 (71)	17 (100)	**0.04**
Female (%)	24 (20)	5 (29)	0 (0)	
Race				0.15
White (%)	85 (71)	13 (77)	16 (94)	
Black (%)	16 (13)	1 (6)	1 (6)	
Asian (%)	2 (2)	2 (12)	0 (0)	
Others/Unknown (%)	16 (14)	1 (6)	0 (0)	
Smoking Status				**0.02**
Current (%)	45 (38)	13 (77)	8 (47)	
Former (%)	68 (57)	3 (18)	9 (53)	
Unknown (%)	6 (5)	1 (6)	0 (0)	
Pack-years (mean (SD))	50 (26)	58 (31)**	48 (23)**	0.49
FEV1 % predicted (mean (SD))	75 (20)	78 (21)	69 (21)	0.42
FEV1/FEV (mean (SD))	0.6 (0.1)	0.6 (0.1)	0.6 (0.2)	0.37
Radiographic BE				0.78
Yes (%)	61 (49)	8 (47)	7 (41)	
No (%)	58 (51)	9 (53)	10 (59)	
Cough				**0.003**
Yes (%)	48 (40)	6 (35)	14 (82)	
No (%)	62 (52)	8 (47)	2 (12)	
Unknown (%)	9 (8)	3 (18)	1 (6)	
Phlegm				0.1
Yes (%)	48 (40)	6 (35)	11 (65)	
No (%)	61 (51)	9 (53)	4 (24)	
Unknown (%)	10 (8)	2 (12)	2 (12)	
Both cough and phlegm				**0.02**
Yes (%)	33 (28)	4 (24)	10 (59)	
No (%)	75 (63)	10 (59)	5 (29)	
Unknown (%)	11 (9)	3 (18)	2 (12)	
Shortness of Breath				0.76
Yes (%)	61 (51)	10 (59)	10 (59)	
No (%)	48 (40)	5 (29)	6 (35)	
Unknown (%)	10 (8)	2 (12)	1 (6)	

Table 3.7 Functional analysis of differentially expressed genes of the five gene clusters. All functional enrichments listed here had a false discovery rate < 0.05.

Gene Cluster (number of genes)	GO-term	Description	Genes
A (N=105)	GO:0007156	Hemophilic cell adhesion	ACKR3, CD81, PCDHGA4, PCDHGA7, PCDHGA9, PCDHGA11, PCDHGA12, PCDHGB4, PCDHGB6, PCDHGB7, PCDHGC5, PPAP2B, PTPRS, SMAD6, STRC
	GO:0022610	Biological adhesion	
	GO:0095609	Cell-cell adhesion	
B (N=136)	GO:0060828	Regulation of canonical Wnt signaling pathway	DACT2, EGFR, FGF9, FGFR2, FZD7, IGFBP4, KANK1, LGR5, LGR6, LRP4, MCC, SEMA5A, SNAI2, SULF2, TLE2, WNT2B, WNT3A, WNT5A
	GO:0030111	Regulation of Wnt signaling pathway	
	GO:0090263	Positive regulation of canonical Wnt signaling pathway	
C (N=58)	GO:0004298	Threonine-type endopeptidase activity	PSMB8, PSMB9, PSMB10
D (N=85)	GO:0044782	Cilium organization	ARL3, B9D2, C1orf192, C6orf165, C21orf59, CC2D2A, CCDC65, CCDC176, CEP97, DNAL1, DYNC2LI1, DYX1C1, IFT22, IFT43, MAP9, SPAG1, TCTEX1D2, TMEM17, TUBA1A
	GO:0060271	Cilium assembly	
	GO:0070925	Organelle assembly	
E (N=104)	GO:0004298	Threonine-type endopeptidase activity	PSMA3, PSMA5, PSMA6, PSMB5

Table 3.8 Gene set enrichment analysis showed enrichment of Hallmark genes in participants with and without radiographic BE. All functional enrichments listed here had a family-wise error rate (FWER) < 0.05.

Pathways Enriched in Participants with Widespread Radiographic BE			
	Enrichment Scores	Normalized Enrichment Scores	FWER p-val
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.67	3.15	0
HALLMARK_OXIDATIVE_PHOSPHORYLATION	0.59	2.83	0
HALLMARK_INTERFERON_ALPHA_RESPONSE	0.65	2.69	0
HALLMARK_MYC_TARGETS_V1	0.51	2.41	0
HALLMARK_ALLOGRAFT_REJECTION	0.51	2.34	0
HALLMARK_COMPLEMENT	0.47	2.18	0
HALLMARK_MTORC1_SIGNALING	0.43	2.06	0
HALLMARK_PROTEIN_SECRETION	0.45	1.92	0.001
HALLMARK_INFLAMMATORY_RESPONSE	0.37	1.76	0.009
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	0.48	1.74	0.013
HALLMARK_APOPTOSIS	0.38	1.74	0.015
HALLMARK_SPERMATOGENESIS	0.41	1.70	0.025
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	0.40	1.69	0.026
HALLMARK_DNA_REPAIR	0.35	1.64	0.044
Pathways Enriched in Participants without Widespread Radiographic BE			
	Enrichment Scores	Normalized Enrichment Scores	FWER p-val
HALLMARK_KRAS_SIGNALING_DN	-0.53	-2.05	0.002
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	-0.42	-1.75	0.023
HALLMARK_PANCREAS_BETA_CELLS	-0.61	-1.72	0.027

Table 3.9 **Genes associated with ciliogenesis.** 42 genes (bold) up-regulated among participants with widespread radiographic BE were previously recognized in a panel of genes (N = 310) important in ciliogenesis.

ABHD12B	C1orf158	CCDC146	DIXDC1	DYNLRB2	HAGHL	KIF3A	LRRC79	NEK11	RPGRIP1L	STX2	TTC8
AGR3	**C1orf189**	CCDC147	DNAAF1	**DYRK3**	HEATR2	KIF3B	LRRC80	NME5	RSPH1	TAF1B	**TUBA1A**
AK8	**C1orf192**	CCDC164	DNAAF2	**DYX1C1**	HOOK1	KIF6	LRRC81	NME7	RSPH10B	TBPL1	TUBB4B
AKAP14	C1orf87	CCDC17	DNAAF3	DZIP1	HYDIN	KIF9	LRRC82	NME8	RSPH4A	TCTEX1D1	TUBD1
ANKMY1	C20orf26	CCDC170	DNAH10	DZIP3	IFT122	KIFAP3	LRRC83	NPHP1	RSPH9	**TCTEX1D2**	TUBE1
APOBEC4	**C20orf85**	**CCDC176**	DNAH11	EFCAB1	IFT140	KIFAP4	LRRC84	NPHP4	RTDR1	TCTN1	TUSC3
ARL3	C21orf58	CCDC33	DNAH12	EFCAB6	IFT172	KLHDC9	LRRC85	NQO1	**RUVBL1**	TCTN1	UCHL1
ARL6	**C21orf59**	CCDC37	DNAH2	EFHB	IFT46	KTN1	LRTOMT	NSUN7	RUVBL2	TEKT1	VWA3B
ARMC2	C22orf23	CCDC39	DNAH3	EFHC1	**IFT57**	**LCA5L**	LRWD1	NUP62CL	SLC22A16	TEKT2	WDPCP
ARMC4	C4orf22	CCDC40	DNAH5	EFHC2	IFT74	LRGUK	LZTFL1	PACRG	SLC22A4	TEX26	WDR16
B9D1	C6orf165	CCDC41	DNAH6	ELL3	IFT81	LRRC18	MAATS1	PDE6B	SLC4A8	TEX9	WDR19
B9D2	**C9orf116**	CCDC60	DNAH7	ENKD1	IFT88	LRRC23	MAK	PFN2	SMYD2	THNSL1	WDR38
BBS10	C9orf117	**CCDC65**	DNAH9	ENKUR	IL20RA	**LRRC34**	MAP6	**PHTF1**	**SOD1**	TMEM107	WDR54
BBS2	**C9orf135**	CCDC78	DNAI1	FABP6	INTU	LRRC43	MAPRE3	PIFO	**SPA17**	TMEM254	WDR60
BBS4	C9orf24	CCDC81	DNAI2	FAM154B	IQCD	LRRC46	MDH1B	PIH1D2	SPAG16	TMEM67	WDR78
BBS5	**CALML4**	CCDC89	**DNAJA1**	FAM206A	IQCG	LRRC48	MEIG1	PIH1D3	SPAG17	TNFAIP8L1	WDR96
BEST4	CAPS	**CCT6B**	DNAJB13	FAM216B	IQCH	LRRC49	MKS1	PLEKHB1	SPAG6	TPPP3	WRAP53
BPHL	**CAPSL**	CETN2	**DNAL1**	FAM81A	IQUB	LRRC6	MLF1	**PPIL6**	SPAG8	TSGA10	WRB
C10orf107	CASC1	CFTR	**DNAL4**	FBXO15	KATNAL2	LRRC71	MLH1	PPOX	SPATA17	TSNAXIP1	XRN2
C10orf67	CATSPERB	CLGN	DNALI1	FOXJ1	**KBTBD4**	LRRC72	MNS1	PPP1R32	SPATA18	TSPAN6	ZBBX
C11orf49	CBY1	CREB3L4	**DUSP14**	FSIP1	KCNE1	LRRC73	**MORN2**	PROM1	**SPATA4**	TTC18	ZCWPW1
C11orf63	CCDC103	CSPP1	DYDC1	GLB1L	KIF19	LRRC74	MORN3	RBKS	SPATA6	TTC21A	ZMYND10
C11orf65	**CCDC104**	CYB5D1	DYNC2H1	GPR162	KIF21A	LRRC75	MPDZ	RFX3	SPATA8	TTC26	ZMYND12
C11orf70	CCDC11	CYB5D2	**DYNC2LI1**	GPX4	KIF23	LRRC76	MROH9	RGS22	SPEF1	TTC29	ZNF474
C11orf74	CCDC114	DAW1	DYNLL1	GSTA1	KIF24	LRRC77	MSMB	**ROPN1L**	STOML3	TTC30A	
C15orf26	CCDC135	**DHX40**	DYNLRB1	GSTA3	KIF27	LRRC78	**MYCBP**	RPGR	STRBP	TTC30B	

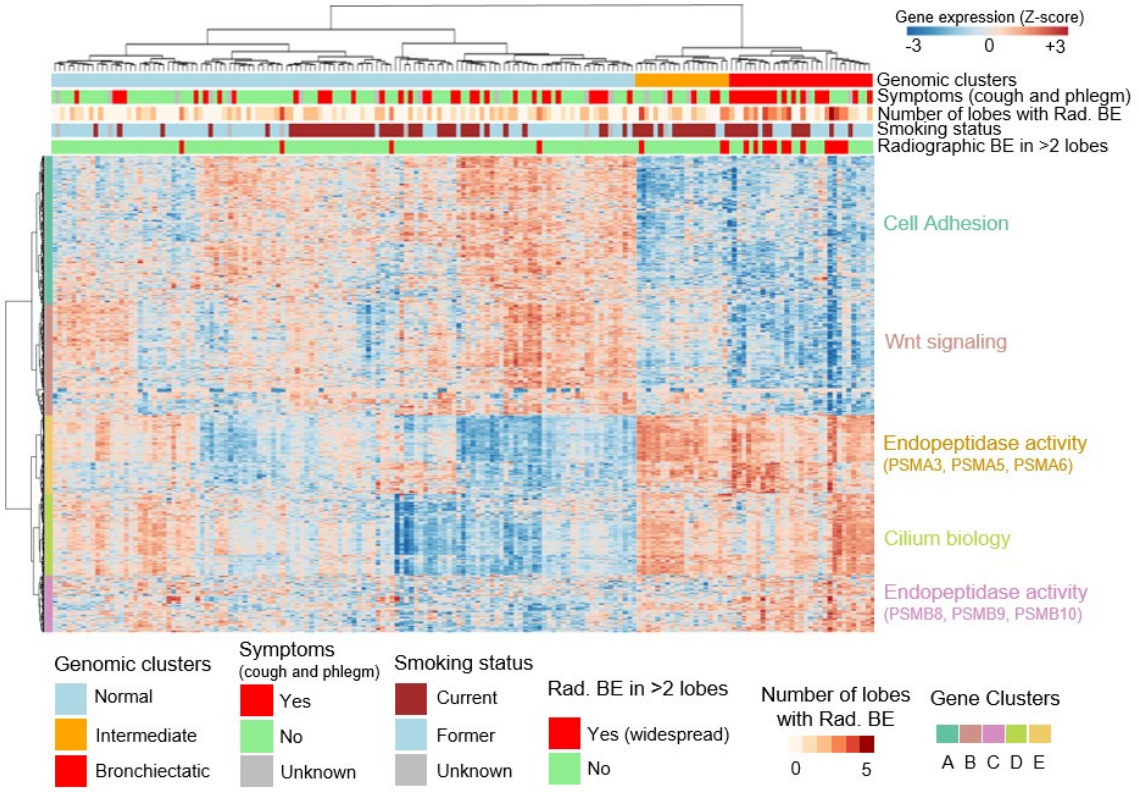

Figure 3.2 Unsupervised heatmap of the 655 genes associated with widespread radiographic bronchiectasis (presence of radiographic BE in at least 3 lobes). Based on hierarchical clustering, participants were grouped into three genomic clusters (normal, intermediate, and bronchiectatic), while genes were grouped into five gene clusters (A-E). Biological pathways in which these clusters of genes were enriched were shown on the side. False discovery rate < 0.1; Fold change > 0.25.

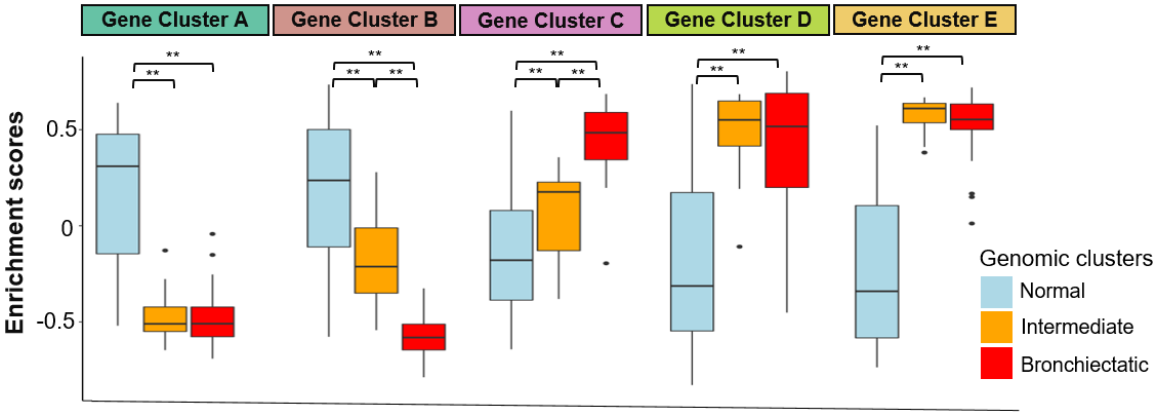

Figure 3.3 Modular gene expressions in genomic clusters. Gene set variation analysis was performed within each sample using genes extracted from Gene Clusters A-E. Participants of the normal and bronchiectatic showed significant differences of expression in all five gene clusters. Participants from the intermediate cluster showed an intermediate level of expression in Gene Clusters B and C, but in general, were more similar to those of the bronchiectatic cluster. ** Tukey adjusted p values < 0.01.

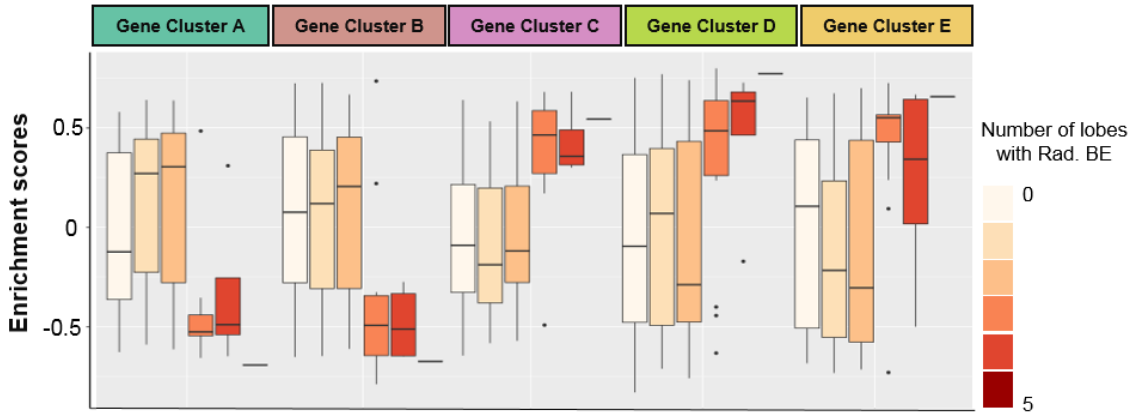

Figure 3.4 Modular gene expressions associated with the number of lobes with radiographic BE. Gene set variation analysis was performed within each sample using genes extracted from Gene Clusters A-E. There is a dramatic shift in expression between 2 and 3 affected lobes. ** Tukey adjusted p values < 0.01.

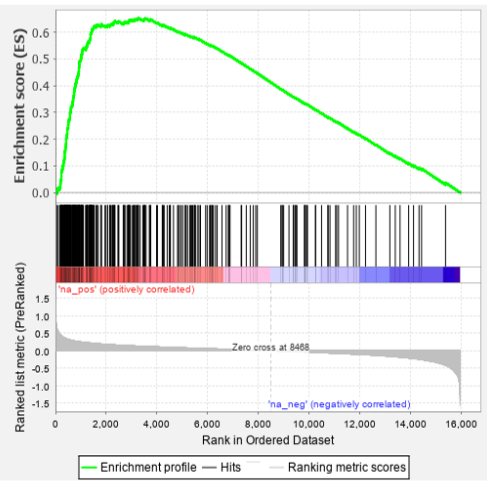

Figure 3.5 GSEA results assessing the enrichment of the 310 genes with relation to ciliogenesis in participants with widespread radiographic BE based on *t* statistics. Each vertical bar represents a single gene within a gene set and its occurrence among the rank like.

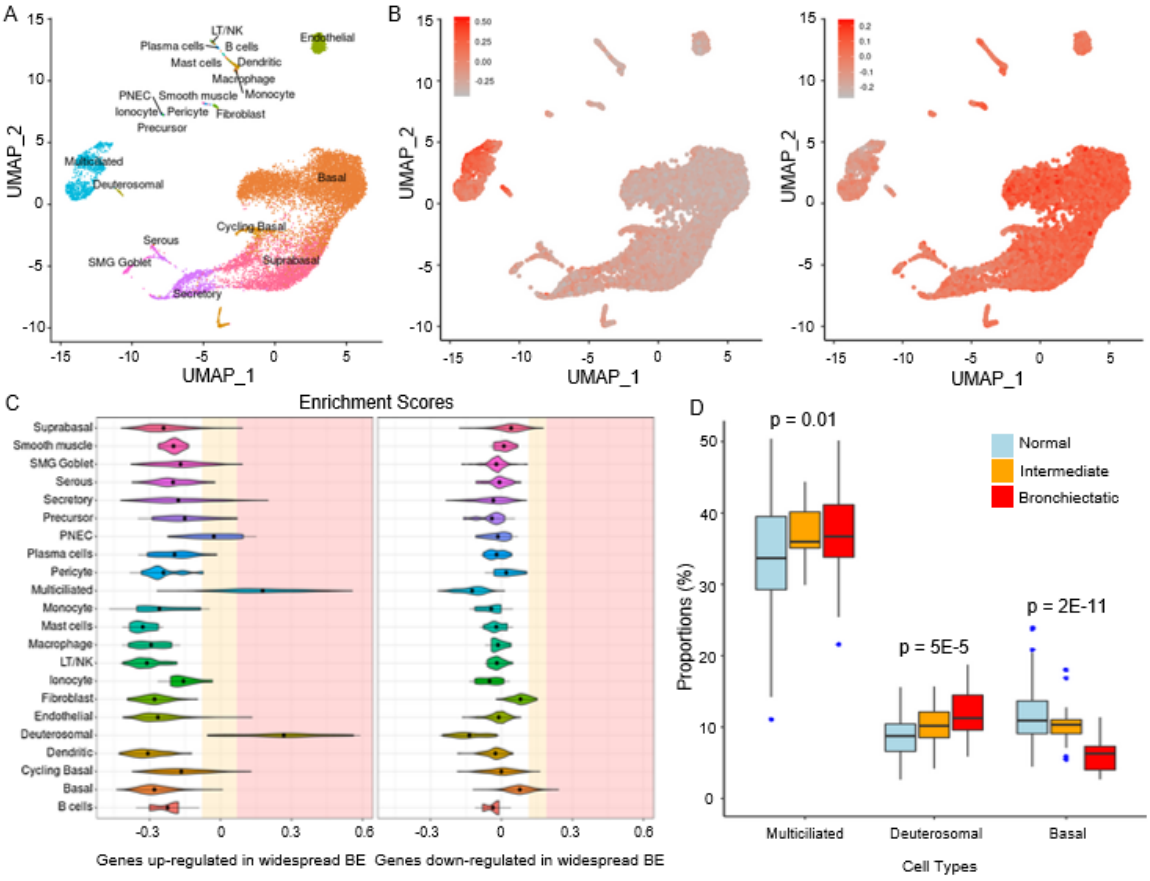

Figure 3.6 Single-cell RNA-seq analysis of genes differentially expressed in participants with widespread radiographic BE. (A) Single-cell RNA-seq of bronchial brushings from 9 subjects (n=13,176 cells) were clustered. Cell types were previously assigned and reported by Deprez et al[23]. (B) UMAP projections showing the expression pattern of genes up- (left) and down-regulated (right) in widespread radiographic BE across different cell types. The cells are colored gray for low expression and red for high expression of metagene scores of each set of genes. (C) Violin plot showing the metagene score for each set of gene modules across the cell types. For each violin plot, metagene expression is designated as elevated (light yellow) or highly elevated (pink) if it is greater than one or two standard deviations above the mean metagene score, respectively. (D) Boxplots of multiciliated, deuterosomal, and basal cell proportions estimated by

AutoGeneS in bulk RNA-seq data from the bronchial brushings (N = 173) obtained from the DECAMP cohort. Significant cell proportion differences among the normal, intermediate, and bronchiectatic clusters were determined by the Kruskal test.

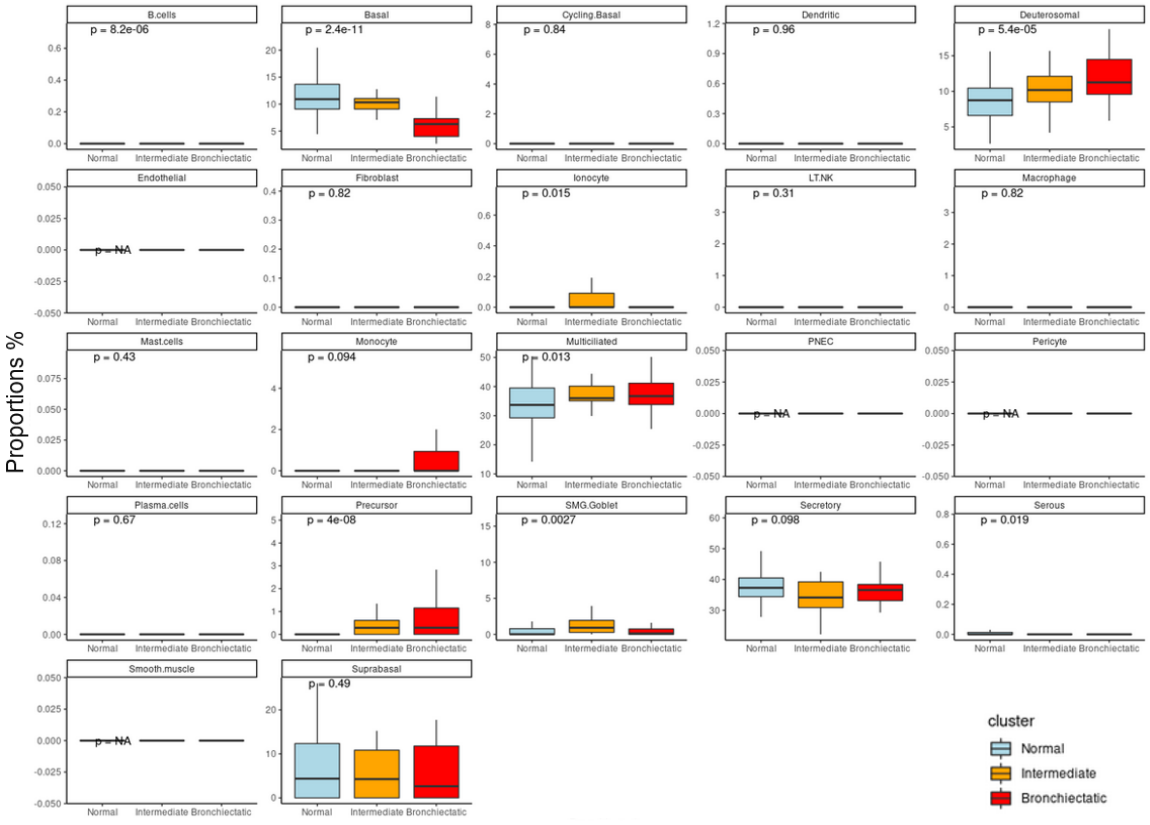

Figure 3.7 Boxplots of all cell type proportions estimated by AutoGeneS. Bulk RNA-seq data from the bronchial brushings (N = 173) obtained from the DECAMP cohort were used as input. Significant cell proportion differences among the normal, intermediate, and bronchiectatic clusters were determined by the Kruskal test.

3.5 Discussion

Current knowledge about the pathogenesis of bronchiectasis (BE) has been summarized as a "vicious cycle" model[122] in which epithelial dysfunction, chronic infection, recurring inflammation, and structural damage are involved in a cycle of events that promote the enlargement of bronchi. Our transcriptomic analysis in individuals without a previous clinical diagnosis of BE but who show signs of BE in multiple lobes on CT ("widespread radiographic BE") potentially offers insights into early molecular changes associated with BE development. Consistent with the proposed vicious cycle model, a loss of gene expression related to cell adhesion and an increased level of gene expression related to inflammation is detected in participants with radiographic BE. Our analysis also reveals two biologic pathways that may play important roles in the initiation of BE that have not been previously discussed – a decreased expression of genes in the Wnt signaling pathway and an increased expression of genes in the cilium biology pathway.

While participants with or without radiographic BE do not show clear clinical differences, the distribution of radiographic BE is consistent with existing findings that BE most commonly occurs in the lower lobes[123–124]. The decreased expression of genes involved in cell adhesion and Wnt signaling with widespread radiographic BE are consistent with the bronchial dilation observed in patients with clinical BE[123]. Activation of the Wnt signaling pathway is important in maintaining the epithelial niche via the balance of epithelial/mesenchymal pairing[126-128]. LGR5, the gene most down-regulated in

participants with widespread radiographic BE, has previously been reported to be required for maintaining the epithelial progenitor niche[127,129]. Single-cell RNA-sequencing of the bronchial epithelium also suggests that genes decreased among participants with widespread BE are highly expressed among the basal cells of the bronchial airway, suggesting that these cells may be less prevalent or less active in individuals with widespread radiographic BE which may alter the structure of the airway or the ability to repair damaged airway epithelium.

Because BE is often accompanied by loss of cilia[130], we were intrigued to observe increased expression of cilia-related genes in participants with widespread radiographic BE. scRNA-seq of the bronchial epithelium also shows that these genes to be exclusively expressed among the ciliated cells. Previously, 310 genes were found to be up-regulated in ciliogenesis, the biologic process for the production of new cilium[120]. Of these 310 genes, 42 were up-regulated among participants with widespread BE. We hypothesize that the increased expression in genes important for ciliogenesis could be a response to cilium damage. A compensatory increase in ciliated cells is also supported by the deconvolution result, which predicts significant increases in the proportion of both the multiciliated cells and deuterosomal cells in samples from the bronchiectatic cluster. The predicted increase in the proportion of deuterosomal cells in samples in the bronchiectatic cluster is the most dramatic. These cells are marked by high expression of DEUP1, FOXN4, and CDC20B, which have been reported as a precursor of multiciliated cells[131]. An alternative hypothesis

is that the overproduction of certain cilium-related proteins contributes to faulty cilia assembly, reduced clearing capacity of the lung[132,] and BE pathogenesis.

Inflammatory pathways up-regulated in participants with widespread radiographic BE (as evidenced both via GSEA and enrichment of genes in the immune-related endopeptidase activity pathways[133]) could reflect increased immune infiltration related to radiographic BE. Large immunohistological studies observed the presence of CD8+ T cells[134,135], CD4+ T cells, macrophages, neutrophils, and interleukin 8 positive cells in the airways of patients with BE[136].

Taken together, the regulatory pattern suggests a compensatory response of producing more cilia or ciliated cells in an inflammatory environment, accompanied by a loss of surface integrity in the early stage of BE (Illustration 3.1). The participants of the "bronchiectatic" gene expression cluster are characterized by having more symptoms such as cough and phlegm. Though 13 of the 30 participants in the bronchiectatic gene expression cluster have widespread radiographic BE, the other 17 do not. Interestingly, these 17 participants differ from the other participants without widespread BE in the other two gene expression clusters in that they are significantly more likely to have cough and phlegm production. Thus, the gene expression profile that defines the bronchiectatic cluster may be the consequence of two separate mechanisms – one that is associated with BE and one that is associated with cough and phlegm production that could be BE-dependent or independent. This discovery is also interesting in that it suggests not all who complain about cough and phlegm are the same, some may have transcriptomic changes in the airway that reflect radiographic BE.

While the participants in the intermediate gene expression cluster demonstrate a gene expression profile that is intermediate between the normal to the bronchiectatic clusters, they differ from those of the other clusters by being predominantly current smokers (75% compared to 37% in the normal cluster and 47% in the bronchiectatic cluster). This intermediate cluster could suggest a previously unappreciated risk of BE among smokers. Alternatively, those in the intermediate cluster may progress to develop more symptoms such as cough and phlegm production, and become more similar to those within the bronchiectatic group but without widespread BE. Longitudinal follow-up of participants of the intermediate cluster may validate whether smoking indeed leads to an increased risk for developing radiographic evidence of BE.

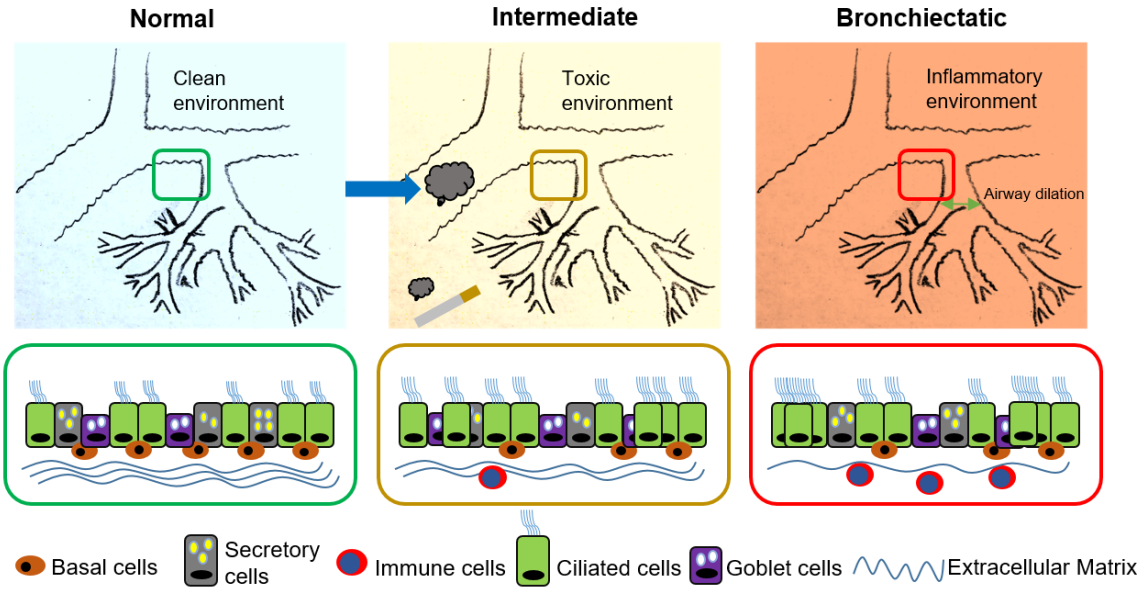

Illustration 3.1. Proposed mechanism of early BE development.

3.6 Conclusion

In conclusion, gene expression differences in individuals with radiographic signs of BE in multiple lobes reflect biologic processes that have previously been related to BE, as well as novel processes that may be associated with BE initiation. The gene expression alterations were also detected in a subpopulation of participants who present with cough and phlegm production but did not have widespread radiographic BE, and an intermediate pattern of gene expression enriched for current smokers. Longitudinal clinical follow-up and molecular profiling of the participants will provide an opportunity to explore the potential for progression and the molecular risk factors for developing radiographic BE.

CHAPTER FOUR: Gene expression alterations associated with malignant pulmonary nodules and creation of a computational model to differentiate indeterminate nodules

Disclaimer: Part of the figures and text in this chapter is being prepared for publication.

4.1 Abstract

There has been a dramatic increase in recent years in the detection of indeterminate pulmonary nodules (IPN), both incidentally and through improved uptake in lung cancer CT screening programs targeting high-risk individuals. Distinguishing benign from malignant nodules remains a challenge. Here, we sought to determine the probability of malignancy in IPNs using clinical factors, CT findings, and genomic alterations within the nasal epithelium.

IPNs on CT of the chest were adjudicated for 221 subjects with paired bulk RNA-sequencing profiling of nasal epithelium. Subjects with a malignant pulmonary nodule were older and had worse airflow obstruction compared to subjects with benign nodules. Malignant nodules were larger but did not show correlation with location (upper versus lower lobe). Transcriptomic assessment of nasal epithelial cells revealed 44 genes with roles in cornification and keratinocyte differentiation pathways that were up-regulated in participants with a malignant nodule; while 31 genes involved in extracellular matrix receptor interaction and focal adhesion pathways were down-regulated.

Concomitantly, we profiled 52,936 cells from 17 samples, 15 of which had IPNs. We identified known cell types such as basal, club, secretory, ciliated, ionocytes, and immune cells. We also discovered a novel cluster of cells termed keratinizing epithelial cells found to be a group of secretory cells characterized by high expression of genes enriched in cornification and keratinization activities. Interestingly, we found that genes

up-regulated in subjects with malignancy were expressed at higher levels in this novel cluster.

Modeling with an extreme gradient boosting (XGBoosting) algorithm, we demonstrated that a model leveraging combined clinical, radiomic, and genomic features achieved the highest classifier performance in determining malignant pulmonary nodules. While the utility of this model should be further validated with an independent cohort, we demonstrated that gene expression profiling of the nasal epithelium may be useful in combination with chest CT, to detect weak signals related to cancerous pulmonary nodules.

4.2 Introduction

Incidental pulmonary nodules are increasingly common sequelae of routine medical care, with an incidence that is much greater than previously recognized[137]. Between 2006 and 2012 the frequency of nodule identification in chest CT imaging increased from 24 to 31% for all scans performed. It is extrapolated that more than 4.8 million Americans underwent at least one chest CT scan and 1.57 million had a nodule identified within that period. The increase in indeterminate pulmonary nodules (IPN), found incidentally and through screening programs targeting high-risk individuals for lung cancer poses a clinical challenge as the vast majority of IPNs are benign[138]. Only approximately 5% of those patients found to have an IPN received a new lung cancer diagnosis within a following 2-year period. Therefore, as lung cancer screening becomes more prevalent[139, 140] and more patients become eligible[141], the need for predictive tools that differentiate benign from malignant nodules is becoming more urgent.

Distinct radiomic features depicted on a CT of the chest can be used to facilitate the prediction of malignancy, especially large nodule size, part-solid appearance, and/or spiculation. The Brock predictive model is one of the more widely used tools leveraging CT scan findings. First developed from participants enrolled in the Pan-Canadian Early Detection of Lung Cancer Study[142], it has been validated in several studies[143, 144, 145]. However, it has been reported that the Brock model tended to overestimate the malignancy risk for lung nodules, especially in Asian countries[146, 147]. Its utility is also limited as it requires six nodule-level and three participant-level inputs, which may not always be readily available. Moreover, the inputs in the Brock model suffer from a lack of objective assessment and could be affected by the condition of the CT scans.

Gene expression profiling is potentially a useful adjunct to CT to identify molecular alterations associated with malignancy. However, it is difficult to directly access diseased distal airway or lung parenchyma routinely for profiling studies. Previously, we have profiled bronchial airway in normal-appearing epithelial cells at the mainstem bronchus, detecting distinct gene expression alterations related to the clinical diagnosis of chronic obstructive pulmonary disease (COPD) and lung cancer[39, 148]. These gene expression alterations offer insights into the molecular events related to diseased tissue at more distal airways and in the parenchyma, which we hypothesize are due to a field-of-injury effect. We have also leveraged gene expression of the nasal epithelium and detected gene expression patterns associated with COPD and lung cancer[38, 149]. Sampling the nasal epithelium provides multiple advantages, with ease of access and safety being the most

obvious. However, it remains to be determined, whether the field-of-injury effect exists for lung cancer in nasal epithelium.

Here, we expand this prior work by correlating nasal gene expression to the status of small IPNs from the DECAMP cohort, combined with single cell RNA-sequencing from the SU2C cohort to identify possible changes in cell landscape within the nasal epithelium that correlate with lung cancer. We then classified pulmonary nodules as malignant or benign by combining HRCT nodule radiomic characteristics with gene expression profiling of the nasal airway using a machine learning algorithm. The following study, therefore, explores how the field-of-injury phenomenon can be extended in the nasal epithelium to create a predictive tool that differentiates benign from malignant nodules, and whether genomic features could improve the model accuracy compared to radiomic features alone.

4.3 Material and Methods

4.3.1 Study Participants

For the bulk RNA-sequencing profile of the nasal epithelium, participants were recruited as part of the DECAMP-1 cohort, described in detail in Section 2.3.1. 221 participants with matching RNA-seq of the nasal epithelium from the inferior turbinate and CT scans were available for analysis as of March 2021.

For the single cell RNA-sequencing profile of the nasal epithelium, participants were recruited as part of the Stand Up To Cancer (SU2C) cohort. SU2C is a multi-center consortium that aims to “address critical unmet needs for risk and response in the lung

cancer interception setting". Patients with an indeterminant pulmonary nodule on CT were recruited for RNA profiling of nasal samples. 17 nasal swabs from 15 participants and volunteers were included for the analysis.

4.3.2 HRCT acquisition and characterization of pulmonary nodules

The acquisition of HRCT was described in detail in Section 2.3.2. Pertinent to this study is that each nodule was assessed at 7 perimeters: interior, centroid (at 15, 20, and 25 mm from the center of the nodule), and boundary (10, 15, and 20 mm from the periphery of the nodule). 66 radiomic features were measured in each perimeter.

4.3.3 Analytic strategies of differential gene expression

The preprocessing steps of the analysis were described in detail from Sections 2.3.4 to 2.3.5, and in Section 3.3.3. To identify genes differentially expressed between subjects with and without malignant nodules, I modeled gene expression on malignant nodule, correcting for sex, smoking status, batch, and median TIN value:

(1) Gene ~ nodule + sex + smoking + batch + TIN + error

A gene with an unadjusted p-value < 0.05 and log fold change >0.25 was considered as differentially expressed.

4.3.4 Single cell RNA-sequencing protocol

A protocol was developed to maximize the quantity and viability of cells collected from the nasal epithelium (Figure 4.2). Cells were initially dissociated from two nasal swabs [CytoSoft, Camarillo, CA] obtained from the inferior turbinate. The cells were then washed with Phosphate-buffered saline (PBS, Sigma Aldrich, Burlington, MA) and dissociated to single-cell suspensions using 0.25% Trypsin/EDTA (Thermo Fisher, Waltham, MA). Red blood cells were removed after treatment with 1X RBC Lysis Buffer [Thermo Fisher, Waltham, MA] for 2 minutes. Trypan blue exclusion [STEMCELL, Vancouver, BC, Canada] was used for measuring cell viability. The final concentration of cells was measured using a hemocytometer under a light microscope before library preparation using the 10X Genomics Platform [Pleasanton, CA].

The nasal samples were sequenced on an Illumina NextSeq 500 with 75 base-pair single-end reads. Reads were demultiplexed using Cell Ranger (v3.1.0) and aligned to human genome build hg38 (v1.2.0) and tabulated according to a unique combination of a Universal Molecular Identifier (UMI) and alignment position.

4.3.5 Single cell RNA-sequencing workflow

The workflow has been described in detail in Section 3.3.4. Additionally, batch correction of the single cell dataset was performed using the *IntegrateData* function from the Seurat package. Overall, 10469 cells were filtered, and 52936 cells were kept for further analysis. UMAP dimensionality reduction was performed using the first 30 principal

components with a resolution setting of 1.2. The cell types were identified with known cell markers or results of the functional analysis performed on highly enriched genes.

4.3.5 Compare and contrast of identified cell types to cell types inferred by cellassign

Cellassign[150] is a new machine learning based computational algorithm that assigns cells measured using single cell RNA sequencing to known cell types based on marker gene information from a different dataset. Unlike other methods for assigning cell types from single cell RNA-seq data, cellassign does not require labeled single cell or purified bulk expression data. Therefore, cellassign provides an opportunity to identify cell clusters without clustering of a new single cell RNA-seq dataset, given a list of marker genes from an existing dataset. Previously, Deprez et al have profiled over 10,000 cells of the nasal epithelium via brushing and have identified cell types using marker gene expression[117]. To compare cell type assignments deduced from an existing dataset and clustering with marker genes, we applied cellassign (version 0.99.21) R package with marker gene information calculated from the Deprez dataset.

4.3.6 Feature selection and Modelling

The workflow of the feature selection and modeling using XGBoost[40] within the caret R package (version 6.0) is shown below (Figure 4.2). A preliminary model leveraging each set of features - clinical, radiomic, and genomic - was initially built to differentiate malignant and benign nodules. Briefly, a missing value in a categorical variable was named

"Unknown". Any numeric feature with missing values in more than 10% of the sample size was discarded; otherwise, the missing value was replaced by the sample median. Highly correlated features were identified using *findCorrelation* function, and only one of two features greater than 0.7 correlation was kept in the model.

The clinical features were: age, sex, smoking status, pack-years, FEV1, FEV/FVC, nodule size, nodule location, and longest axis of a nodule. The radiomic features combine radiomic features collected from the interior of each nodule and engineered features by taking the difference between measurements collected from the different boundary and centroid peripheries (Illustration 4.1). For example, feature "Information Dimension" was measured at both Centroid 15mm and Centroid 20mm perimeters, the difference of the two measurements was calculated to create a new feature named "Centroid_20_15_DIFF_Infromation Dimension". The genomic features were genes differentially expressed in the nasal epithelium between individuals with and without a malignant pulmonary nodule.

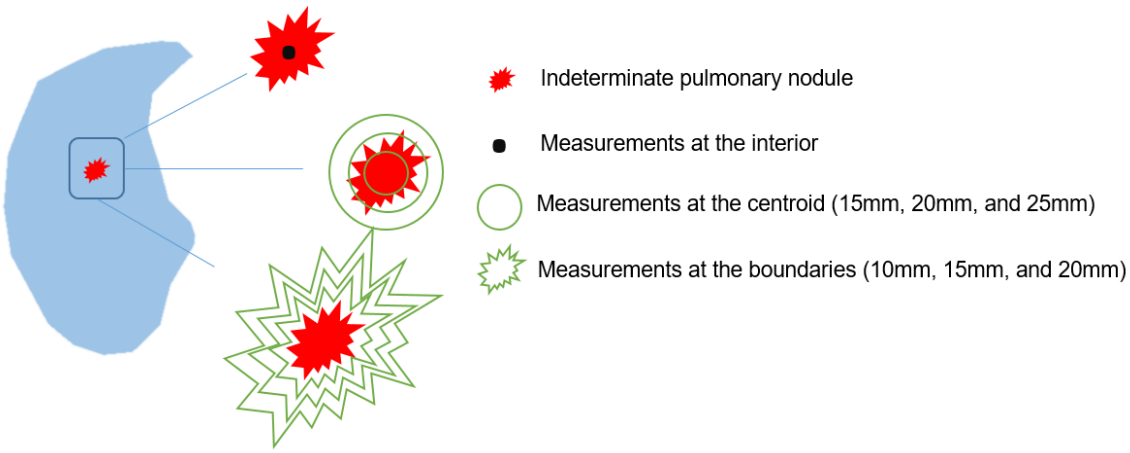

Illustration 4.1 Schematic representation of radiomic features.

The model for XGBoost was optimized on the "error" with its objective set to "binary: logistic". Five-fold cross-validation was applied within the training process. Hyperparameters tuned for the model included: nround (50, 100, 200, 300, 500, and 1000), max_depth (2, 4, 6, and 8), eta (0.001, 0.01, 0.1, 0.2), min_child_weight (2, 4, 6, and 8), and lambda (0.1 and 0.2).

4.3.7 Statistical analysis

The statistical analysis has been described in detail in Chapter 2.3.8.

4.3.8 Identification of samples of lower qualities

Part of the challenge to transcriptomic analysis of the nasal samples was that many started with very low sample qualities indicated by low Transcript Integrity Numbers (TIN). Here, using a similar model set up and hyperparameters as described in 4.3.6, I created a model that uses relevant features to differentiate samples of high qualities from the others. These features include library size, library variance, expression of top 1 most abundant gene, and probability of biological sex prediction using gene XIST and RPS4Y1.

Revised Protocol for Single Cell Suspension from Nasal Brushings (2020.01.01)

Sort Buffer = 1% FBS in PBS **Wash Buffer** = **PBS**

Nasal Brush Prep: (in 1mL Sort Buffer)

- Rinse each brush (after clipping the tip from the entire brush) w/15mL **Sort Buffer** into 100mm plate (15X using 1mL pipette)
- **Please note:** You should use the pipettor to wash up and down the brush tip to get maximal cell recovery.
- Rinse out the tube holding brushes 3X with 1 mL **Sort Buffer**
- Transfer contents of plate to a 50mL tube
- Rinse off plate 3X with 5 mL **Sort Buffer** => transfer to tube
- Check the plate using a light microscope to make sure you've transferred most cells.
- Centrifuge 2000 rpm, 5min
- Wash with 20mL **PBS**, don't have to disturb the cell pellet
- Centrifuge 2000 rpm, 5min. Aspirate to cell pellet.

Trypsinization

- Add 0.75mL 0.25% Trypsin:EDTA => transfer to 6 well plate
- Add another 0.75 mL 0.25% Trypsin:EDTA => transfer remainder to plate
- **Incubate for 5 min** => mix cells with 1mL pipet (7X) followed by 200uL (7X) followed by 20uL (7X) pipet to mechanically dissociate cells (repeat up to 2 times = 10 minutes max total incubation time)
- **Please note:** stop the trypsinization step when you see most of the cells are in single cell suspension. Over-trypsinization may damage the cells for down-stream processing and affect sequencing quality.
- Add 4 mL **Sort Buffer** and mix well to inactivate Trypsin
- Pipet cells through 40um strainer into 50mL tube
- Wash well with 1mL **Sort Buffer** and pipet through strainer until about 15mL total
- Transfer everything to a 15mL conical tube (sharper tip at the bottom for more obvious cell pellet)
- Centrifuge **for** 5 min, 2000 rpm. Aspirate.

RBC Lysis

- Prepare 1X **RBC Lysis Buffer for 1 sample (5mL)**
  - 0.5mL 10X RBC Lysis + 4.5mL UP dH2O
- Resuspend Cells in 5mL 1X **RBC Lysis Buffer**
  - **Tube = 2 min on ice, shake gently at 1min**
  - **Please note:** If you started with a very bloody sample, you may want to incubate for an extra minute.
  - Add 10mL **PBS** to Stop reaction
  - Centrifuge 5 min, 2000 rpm (re-centrifuge if needed)
    - Resuspend cells in 50uL **Sort Buffer**

Cell count

Figure 4.1 Protocol for single cell dissociation.

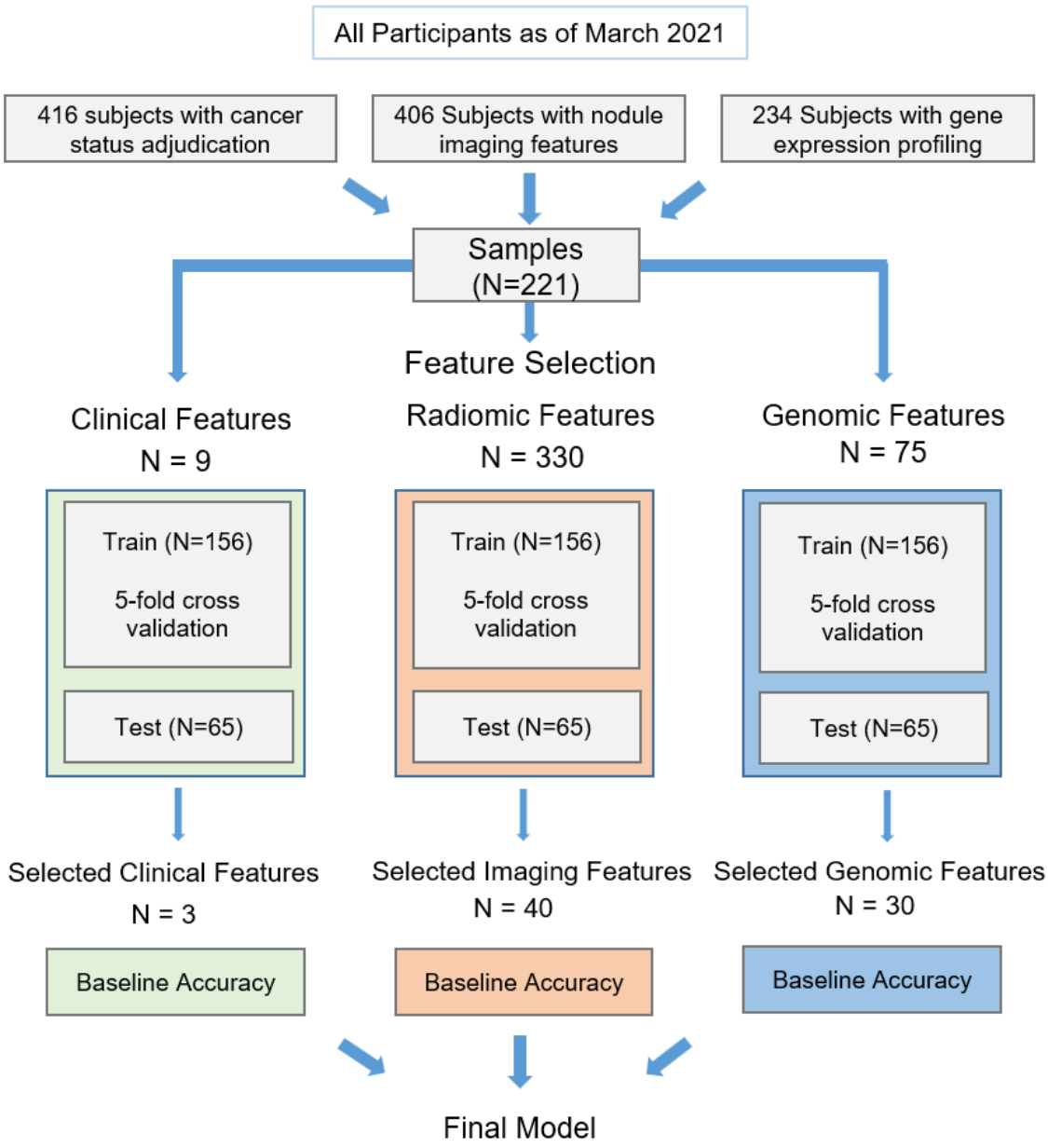

Figure 4.2 Schematic representation of creating a model to predict malignancy of an indeterminant pulmonary nodule. 221 samples with complete clinical, radiomic, and genomic features were included for this analysis. Samples were separated into training and testing sets (7:3) for separate feature selections. Selected features were then combined for model building.

4.4 Results

4.4.1 Participant demographics and pulmonary function

Individuals with a malignant nodule were older and had worse FEV1% predicted. Consistent with the previous findings [142, 143, 144], malignant nodules were significantly larger and had increased length. Sex, smoking status, pack-years, as well as the location of the nodule, however, did not correlate with nodule status (Table 4.1).

4.4.2 Pre-filtering of samples of potentially lower quality.

One of the challenges we faced in discovering malignancy-associated gene expression alterations in the nasal epithelium was that the samples suffer from significant batch effects and low RNA quality (Figure 4.3). Initial analysis including samples of lower quality did not generate satisfactory results (data not shown) with even batch correction applied. Moreover, our routine quality control procedure using sex-linked gene expression alone failed to exclude these samples.

To identify and exclude samples of lower quality from the analysis, two methods were tested: one that relies heavily on the Transcript integrity number (TIN)[145] and one that relies on a combination of various quality control metrics. We first used a TIN value less than 60 and identified 61 samples, which included samples thought to have negatively impacted the analysis. However, this cut-off also discarded samples without obvious quality issues (Figure 4.4). We then applied a novel method to identify samples with lower quality by leveraging not one QC metric (TIN), but a set of QC metrics including TIN,

RIN, and spectrometric readouts. Unsupervised clustering of QC metrics showed that in general, samples with lower TIN were clustered together (left branch of the dendrogram). However, some samples with relatively high TIN had very low RIN[146] and had been flagged by the spectrometric readout for potential contamination and/or fragmented RNA.

We then examined a variety of features embedded in the count matrix that could reflect the quality of the expression profile. In early data exploratory analysis, I defined sample quality using the unsupervised clustering of samples in Figure 4.3 and found that sample library size and probability of biological sex prediction were positively correlated with sample quality, whereas library variance and expression of top 1 most abundant gene were negatively correlated with sample quality. Using these four features, an XGBoost model was applied and predicted 40 samples should be excluded from the analysis (Table 4.2).

I then applied this model to published RNA sequencing samples previously labeled with sample quality (keep or discard) and was able to achieve 95% accuracy determining whether a sample was of good quality (Table 4.2). While needing further refinement, this method may be further developed to address challenges associated with samples of lower qualities in general. Moreover, this challenge coincides with another common challenge in bioinformatic analysis, which is to perform quality control procedures on samples published on Gene Expression Omnibus. An online resource for data curation for most published studies. However, the original sequencing files may not always be available, and only the count matrix is published. Additionally, the pre-processing the original read file creates delays in the workflow.

4.4.3 Genes differentially expressed between subjects with/without a malignant pulmonary nodule

We performed differential gene expression analysis comparing participants with and without a malignant pulmonary nodule and discovered 75 genes were significantly (p-value < 0.05, log fold change > 0.25) differentially expressed when controlling for sex, smoking status, batch, and TIN (Figure 4.5). Functional enrichment analysis showed that the 44 genes up-regulated in participants with a malignant pulmonary nodule were enriched for genes involved in cell cornification and keratinocyte differentiation, whereas the 31 genes down-regulated in participants with a malignant pulmonary nodule were enriched for genes involved in extracellular matrix receptor interaction and focal adhesion pathways.

4.4.4 Examination of the nasal epithelium at a single-cell resolution

Single cell RNA-sequencing was completed for a total number of 17 samples with 52,936 high-quality cells (Table 4.3). Leveraging single cell gene expression profiles and known cell markers, we were able to identify basal cells, club cells, secretory cells, ciliated cells, ionocytes, keratinizing epithelial cells, and immune cells (Figure 4.6).

Similar to the previously reported result by Deprez et al, we observed that ciliated cells, deuterosomal cells, immune cells, ionocytes were clustered as isolated clusters, contrasting with the other epithelial cells grouped in a large cluster of cells. One cluster of cells, dark green on the right consists of cells primarily from one individual. Unlike Deprez et al, we further sub-divided the secretory cell population into clusters defined either by

the top gene expressed (i.e. C15orf48 and STATH), or the molecular process enriched by the top expressed genes (keratinizing epithelial cells) (Figure 4.7).

Club cells and ionocytes represent the most and least abundant cell type, at 15.69% and 0.45%, respectively (Table 4.4). It is worth pointing out a large proportion of cells (~25%) was determined to be at a transitional state, from one cell type to another.

4.4.5 Comparison and contrast with cell type identification by different methods

Cell types of the single cell RNA-sequencing of the nasal epithelium were assigned either via the traditional workflow, where cells were grouped into clusters that share similar gene expression profiles before identification using known marker genes, or by using a set of marker genes derived from the single cell RNA-sequencing profiles from Deprez et al. Here, we saw good agreement between the cell types identified via either approach. In cell types with distinct gene expression profiles such as the cycling basal cells, macrophages, deuterosomal cells, ionocytes, and multiciliated cells, the overlap was more complete. The secretory cell population defined by Deprez et al was further divided into club cells, secretory cell (STATH+), secretory (C15orf48+), goblet cells, and the keratinizing epithelial cells.

4.4.6 Genes up-regulated in participants with a malignant nodule were enriched in the keratinizing epithelial cells

Using GSVA, we calculated metagene scores for both genes up- and down-regulated in participants with a malignant nodule in each cell to understand how biological

processes are associated with the up- and down-regulated genes in the bulk RNA-seq data are distributed across nasal cell populations (**Figure 4.8**). In the nose, the genes up-regulated in participants with a malignant nodule were moderately expressed across many cell types but were most highly expressed in the keratinizing epithelial cells. The genes down-regulated in participants with a malignant nodule were expressed across most cell types and slightly more highly expressed by macrophages, followed by T cells.

4.4.7 Model with combined features achieved highest classifier performance

Preliminary models using only the 9 clinical, 330 radiomoic, and 75 genomic variables selected 3 clinical, 40 radiomic, and 30 genomic features that have an importance score greater than 0.001 (Table 4.6). The most important features from the clinical, genomic, and radiomic variables were nodule size, ITGB3, and difference in standard deviation between centroid at 20mm and 15mm periphery. Using these selected features, we determined the baseline accuracies of each model to be 0.66, 0.65, and 0.65, respectively (Figure 4.9). The area under the ROC curve (AUC) for the model with only clinical, radiomic, and genomic features were 0.64, 0.63, and 0.7, respectively. These features were then combined to build a model that achieved the highest classifier performance with an AUC of 0.74, a significant improvement compared to the model using only the radiomic features (p value < 0.05). The model with combined features also had improvement in accuracy, sensitivity, specificity, and precision.

TABLE 4.1 Clinical characteristics of the participants with and without a malignant pulmonary nodule. The mean and standard deviation are shown for continuous variables. The count and proportion are shown for categorical variables. *p values calculated using either a student t test or Fisher's exact test. **Missing FEV1% predicted and FVC for 7 subjects with benign nodules and 16 subjects with malignant nodules. ***Missing FEV1/FVC for 7 subjects with benign nodules and 17 subjects with malignant nodules.

	Benign (N = 94)	Malignant (N = 165)	*P* value*
Age (mean (SD))	67 (8)	69 (8)	**0.03**
Sex			0.86
Male (%)	74 (79)	127 (77)	
Female (%)	20 (21)	38 (23)	
Race			0.90
White (%)	66 (70)	124 (75)	
Black (%)	13 (14)	29 (18)	
Asian (%)	2 (2)	4 (2)	
Others/Unknown (%)	13 (14)	8 (5)	
Smoking Status			0.43
Current (%)	38 (40)	78 (47)	
Former (%)	51 (54)	79 (48)	
Never (%)	0	1 (1)	
Unknown (%)	5 (5)	7 (4)	
Pack-years (mean (SD))	48 (26)	53 (27)	0.16
FEV1 % predicted (mean (SD)) **	80 (19)	75 (19)	**0.04**
FEV1/FEV (mean (SD)) ***	0.7 (0.1)	0.6 (0.1)	0.10
FVC (mean (SD)) **	93 (16)	91 (17)	0.31
Size (mm^3) (mean (SD))	1.2 (0.6)	1.6 (0.6)	**1.049e-07**
Location			0.60
Lower lobes (%)	35 (37.2)	55 (33.3)	
Upper lobes (%)	58 (61.7)	109 (66.1)	
Unknown (%)	1 (1.1)	1 (0.6)	
Longest axis (mm) (mean (SD))	11.6 (6.0)	13.7 (7.2)	**0.01**

TABLE 4.2 Comparison of samples filter by TIN and by combined metrics. 443 RNA-sequencing samples were determined as of sufficient quality by using a combination of metrics, whereas 422 samples were determined as having sufficient quality using TIN > 60 as a cutoff. 3 out of 4 samples previously determined as lower quality were determined by modeling with combined metrics.

Determined by TIN > 60	Determined by combined metrics: Discard	Determined by combined metrics: Keep
Discard	25	36
Keep	15	407

Determined by TIN > 60 (GSE70285)	Determined by combined metrics: Discard	Determined by combined metrics: Keep
Discard	3	0
Keep	1	16

TABLE 4.3 Subject demographics for nasal single cell RNA-sequencing profiling.

Sample #	Sample ID	Number of cells	Sex	Age	Smoking Status
1	BU_L1	13117	M	30	Never
2	BU_R1	12663	M	30	Never
3	KA_left	1576	F	34	Unknown
4	KA_right	1652	F	34	Unknown
5	LH_0012	3044	M	63	Former
6	LH_0034	6726	M	64	Former
7	LH_0951	107	M	66	Current
8	LH_1169	7298	F	75	Former
9	LH_1925	123	M	59	Current
10	LH_2077	572	F	62	Former
11	LH_2130	304	M	56	Current
12	LH_2158	157	F	65	Current
13	LH_0138	724	M	61	Current
14	LH_0287	739	F	64	Current
15	LH_0516	287	M	74	Former
16	LH_1979	2612	M	68	Current
17	LH_2085	1235	F	70	Former

TABLE 4.4 Proportions of different types of cells in the nasal compartment. A total of 52936 cells were collected from 17 samples. Cells were clustered and identified using known cell markers or result of functional analysis performed on highly enriched genes.

	Number of cells	Proportions (%)
Club cells	8308	15.69
Transitional Cells_Basal_Club	6711	12.68
Ciliated cells	6525	12.33
Transitional Cells_Club_Secretory	5750	10.86
C15orf48 secretory	5128	9.69
Goblet cells	4082	7.71
STATH secretory	3653	6.90
Basal cells	3323	6.28
Keratinizing epithelial cells	3261	6.16
1169 predominant cells	1525	2.88
KRT15 17 Basal cells	1112	2.10
Transitional Cells_Basal_keratinizing epithelial cells	935	1.77
MUC4 NEAT1 cells	771	1.46
Proliferative basal cells	417	0.79
T cells	384	0.73
Deuterosomal cells	383	0.72
Macrophages	338	0.64
Ionocytes	238	0.45
Unidentified	92	0.17

TABLE 4.5 Compare and contrast of identified cell types to cell types inferred by cellassign. Cell types assigned by either method showed good agreement, especially for basal cells, cycling basal cells, macrophages, deuterosomal cells, ionocytes, and multiciliated cells.

	Ke Xu's Assignment													
Barbry's Assignment	**Basal cells**	**KRT15.17.Basal.cells**	**Cycling basal cells**	**Macrophages**	**Deuterosomal**	**Ionocytes**	**T cells**	**Multiciliated cells**	**Club cells**	**STATH secretory**	**C15orf48 secretory**	**Goblet cells**	**Keratinizing epithelial cells**	**MUC4 NEAT1 cells**
Basal	446	184	0	0	0	0	1	1	1	0	0	0	3	2
Cycling Basal	265	35	140	0	0	0	0	3	2	0	1	2	2	0
Dendritic	0	0	0	70	0	0	9	0	0	0	0	0	0	1
Deuterosomal	0	1	0	0	48	0	0	0	0	0	0	0	0	0
Fibroblast	387	147	0	3	1	0	3	5	14	22	8	1	29	11
Ionocyte	0	0	0	0	0	82	0	0	0	1	0	0	3	2
LT/NK	0	0	0	0	0	0	98	0	0	0	0	0	1	0
Monocyte	0	0	0	40	0	0	20	1	0	0	0	0	1	0
Multiciliated	4	2	0	2	36	1	0	2270	16	2	5	5	4	0
Precursor	2	3	11	0	0	0	0	0	0	0	0	0	0	0
Secretory	49	30	3	13	49	6	14	158	3088	1164	1939	1499	1080	279
Serous	2	10	0	0	0	0	0	1	3	103	0	0	1	0
SMG Goblet	0	0	0	0	0	0	0	0	0	2	0	37	0	0
Suprabasal	138	2	0	0	0	0	0	1	0	0	0	0	103	0

TABLE 4.6 List of features included for model construction. 3 clinical variables, 30 genes, and 40 radiomic features that have an importance score of >0.001 were selected.

Clinical Variables		Radiomic Variables	
Variable	Importance	Variable	Importance
Size	0.87	img_Centroid_20_15_DIFF_Standard.Deviation	0.1590
Pack Years	0.12	Spherical Disproportion	0.1343
Age	0.01	img_Centroid_20_15_DIFF_Information.Dimension	0.0901
		img_Centroid_20_15_DIFF_IMC1	0.0766
		img_Boudary_15_10_DIFF_Information.Dimension	0.0518
		img_Centroid_20_15_DIFF_Box.Counting.Dimension	0.0492
		img_Boudary_15_10_DIFF_Cluster.Prominence	0.0363
		img_Centroid_25_20_DIFF_Box.Counting.Dimension	0.0329
Genomic Variables		img_Boudary_15_10_DIFF_Compactness.2	0.0280
Variable	Importance	img_Boudary_15_20_DIFF_IMC1	0.0274
ITGB3	0.1534	img_Boudary_15_10_DIFF_SRE	0.0230
ETV4	0.0975	Energy	0.0198
PNMA6C	0.0758	Skewness	0.0196
CERS4	0.0738	img_Centroid_20_15_DIFF_Ventilation.Heterogeneity	0.0196
ANO4	0.0599	img_Boudary_15_20_DIFF_Cluster.Tendency	0.0193
FLG2	0.0545	img_Boudary_15_20_DIFF_LAA910Perc	0.0187
FAM74A3	0.0463	img_Centroid_25_20_DIFF_LRLGLE	0.0170
RP11-25H12.1	0.0445	Sum Variance	0.0160
CDH2	0.0444	Range	0.0154
RP11-156K13.1	0.0367	img_Boudary_15_20_DIFF_LAA856Perc	0.0153
VIPR1	0.0364	img_Centroid_20_15_DIFF_Minimum.Intensity	0.0152
DTNA	0.0343	img_Centroid_25_20_DIFF_Minimum.Intensity	0.0148
CCL2	0.0327	img_Boudary_15_20_DIFF_Information.Dimension	0.0136
AACSP1	0.0277	img_Centroid_20_15_DIFF_Maximum.Probability	0.0126
SH3GL1P1	0.0175	img_Boudary_15_10_DIFF_Extruded.Surface.Volume.Ratio	0.0097
SPP1	0.0166	img_Boudary_15_10_DIFF_Maximum.3D.Diameter	0.0090
RP13-616I3.1	0.0152	img_Centroid_20_15_DIFF_SRE	0.0065
IVL	0.0149	img_Boudary_15_20_DIFF_Correlation	0.0056
AC104532.2	0.0142	img_Centroid_25_20_DIFF_IMC1	0.0050
RP11-1348G14.4	0.0137	LRLGLE	0.0050
PRSS30P	0.0123	RP	0.0039
PNLIPRP3	0.0122	Ventilation Heterogeneity	0.0038
CTD-2562J15.6	0.0113	img_Boudary_15_20_DIFF_SRE	0.0038
KCTD16	0.0107	img_Centroid_25_20_DIFF_Standard.Deviation	0.0036
SCARNA10	0.0107	img_Centroid_25_20_DIFF_Information.Dimension	0.0033
C21orf88	0.0092	img_Centroid_25_20_DIFF_Maximum.3D.Diameter	0.0029
CTSW	0.0092	img_Centroid_20_15_DIFF_Sum.Variance	0.0026
VN1R82P	0.0069	img_Centroid_25_20_DIFF_SRLGLE	0.0021
AC005532.5	0.0028	img_Boudary_15_20_DIFF_LRLGLE	0.0017
CLSTN2	0.0021	img_Boudary_15_10_DIFF_Kurtosis	0.0015

TABLE 4.7 List of the top 50 features in the final model. Genomic, radiomic, and clinical features all contributed to the final model.

Features	Importance	Type
ITGB3	8.77E-02	Genomic
img_Centroid_20_15_DIFF_Standard.Deviation	7.47E-02	Radiomic
Spherical Disproportion	6.19E-02	Radiomic
img_Centroid_20_15_DIFF_Information.Dimension	3.98E-02	Radiomic
img_Boudary_15_10_DIFF_SRE	3.86E-02	Radiomic
FAM74A3	3.55E-02	Genomic
img_Boudary_15_10_DIFF_Compactness.2	3.04E-02	Radiomic
img_Boudary_15_20_DIFF_LAA910Perc	2.94E-02	Radiomic
img_Boudary_15_20_DIFF_Cluster.Tendency	2.92E-02	Radiomic
RP13-616I3.1	2.79E-02	Genomic
CERS4	2.56E-02	Genomic
PNMA6C	2.44E-02	Genomic
img_Centroid_20_15_DIFF_IMC1	2.42E-02	Radiomic
img_Boudary_15_10_DIFF_Extruded.Surface.Volume.Ratio	2.35E-02	Radiomic
img_Boudary_15_20_DIFF_LAA856Perc	2.25E-02	Radiomic
RP11-25H12.1	2.06E-02	Genomic
img_Centroid_20_15_DIFF_SRE	2.05E-02	Radiomic
age	2.00E-02	Clinical
FLG2	1.83E-02	Genomic
img_Centroid_25_20_DIFF_Maximum.3D.Diameter	1.65E-02	Radiomic
size	1.63E-02	Clinical
img_Boudary_15_10_DIFF_Maximum.3D.Diameter	1.62E-02	Radiomic
ANO4	1.53E-02	Genomic
img_Centroid_20_15_DIFF_Box.Counting.Dimension	1.50E-02	Radiomic
KCTD16	1.47E-02	Genomic
img_Boudary_15_20_DIFF_SRE	1.45E-02	Radiomic
DTNA	1.44E-02	Genomic
img_Centroid_25_20_DIFF_Information.Dimension	1.25E-02	Radiomic
IVL	1.24E-02	Genomic
PNLIPRP3	1.19E-02	Genomic
Skewness	1.09E-02	Radiomic
CCL2	1.09E-02	Genomic
VIPR1	1.08E-02	Genomic
VN1R82P	1.06E-02	Genomic
Range	1.05E-02	Radiomic
ETV4	9.32E-03	Genomic
img_Boudary_15_20_DIFF_Correlation	9.31E-03	Radiomic
SPP1	8.97E-03	Genomic
RP11-156K13.1	8.40E-03	Genomic
C21orf88	7.94E-03	Genomic
img_Boudary_15_10_DIFF_Cluster.Prominence	7.79E-03	Radiomic
img_Boudary_15_10_DIFF_Information.Dimension	6.95E-03	Radiomic
img_Centroid_25_20_DIFF_Minimum.Intensity	6.46E-03	Radiomic
LRLGLE	6.22E-03	Radiomic
img_Centroid_25_20_DIFF_IMC1	6.05E-03	Radiomic
Energy	5.76E-03	Radiomic
packyears	5.58E-03	Clinical
img_Boudary_15_20_DIFF_Information.Dimension	4.81E-03	Radiomic
img_Boudary_15_20_DIFF_IMC1	4.46E-03	Radiomic

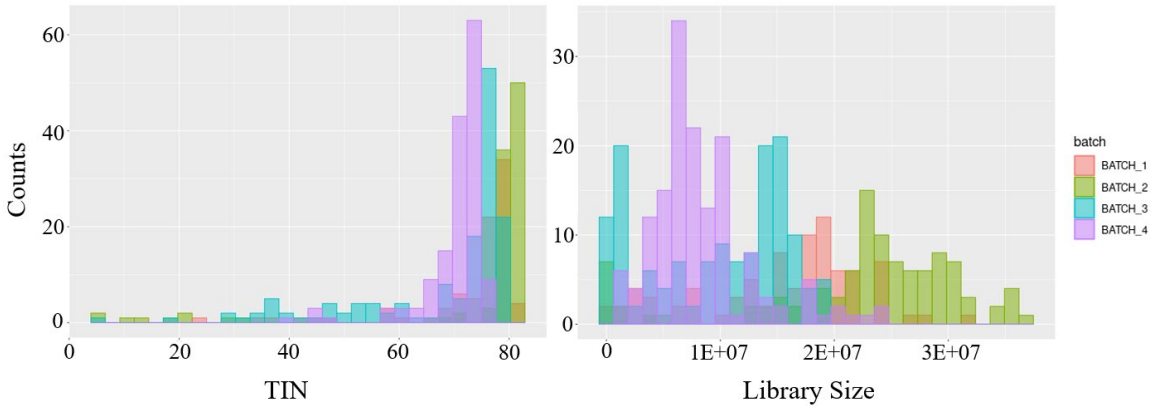

Figure 4.3 Distribution of TIN and library sizes. Samples included for this analysis were collected and sequenced in four batches, representing a varied distribution of TINs and library sizes.

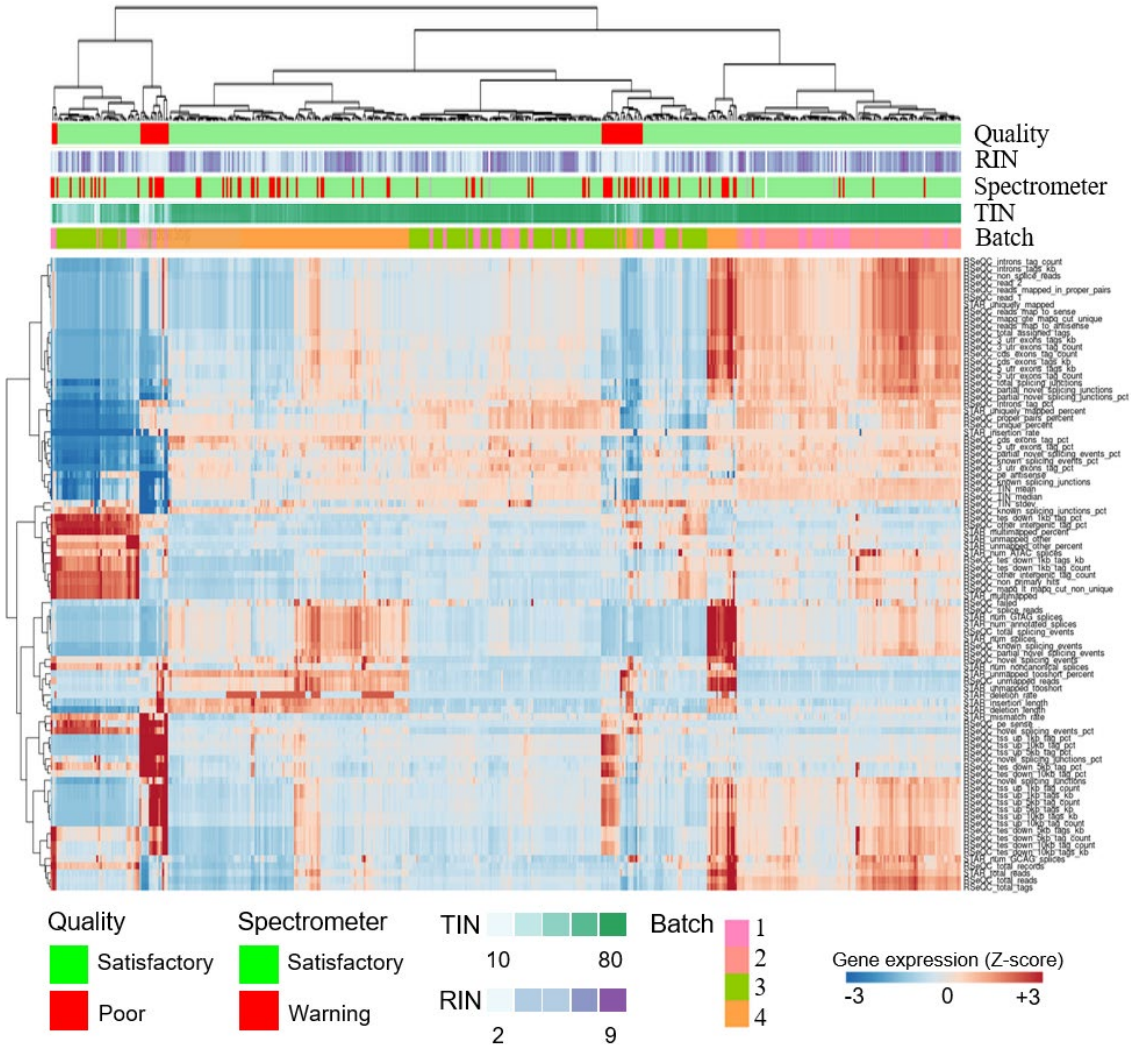

Figure 4.4 Unsupervised heatmap of the QC metrics. Based on hierarchical clustering, samples were grouped by their QC metrics. TIN, RIN, and RNA quality assessed by spectrometer were taken into consideration for the final determination ("Quality") whether a sample will be included for the analysis.

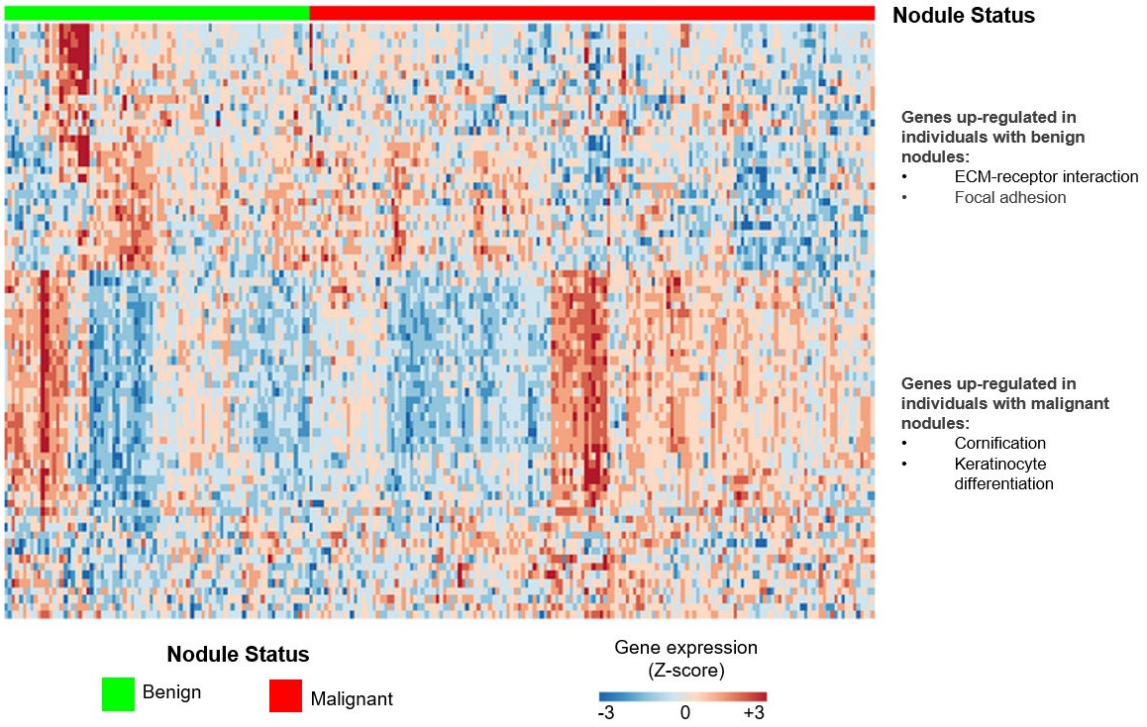

Figure 4.5 Semi-supervised heatmap of the 75 genes associated with malignant pulmonary nodules. Participants were grouped into two clusters by the status of their pulmonary nodule, while genes were grouped via hierarchical clustering. Biological pathways in which these clusters of genes were enriched were shown on the side. p value < 0.05; Fold change > 0.25.

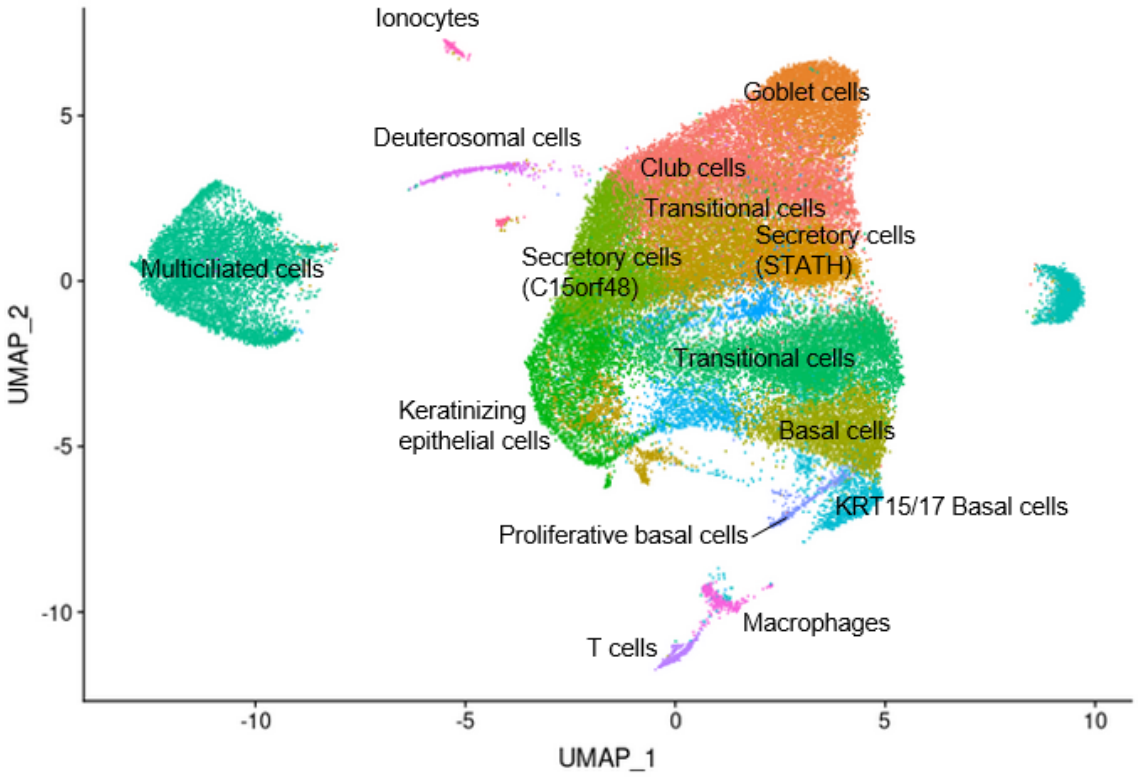

Figure 4.6 Single-cell RNA-seq analysis of the nasal epithelium from 15 ever-smokers. Single-cell RNA-seq of nasal brushings from 15 subjects (n=52,936 cells) were clustered.

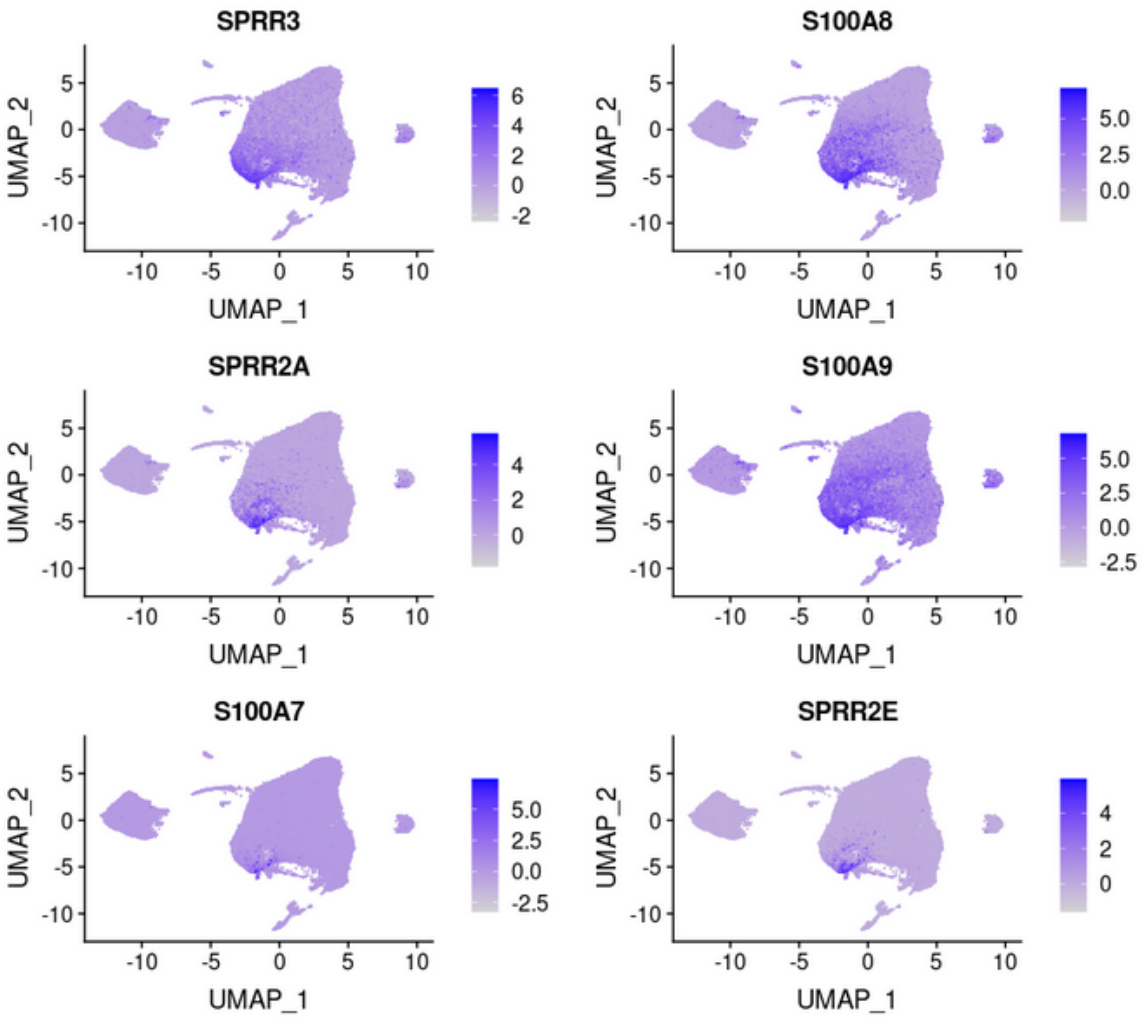

Figure 4.7 UMAP projections showing the expression pattern of genes highly enriched among the keratinizing epithelial cells. The top six differentially expressed genes within the keratinizing epithelial cells, along with other genes highly expressed among the keratinizing epithelial cells, were enriched in cornification and keratinization processes.

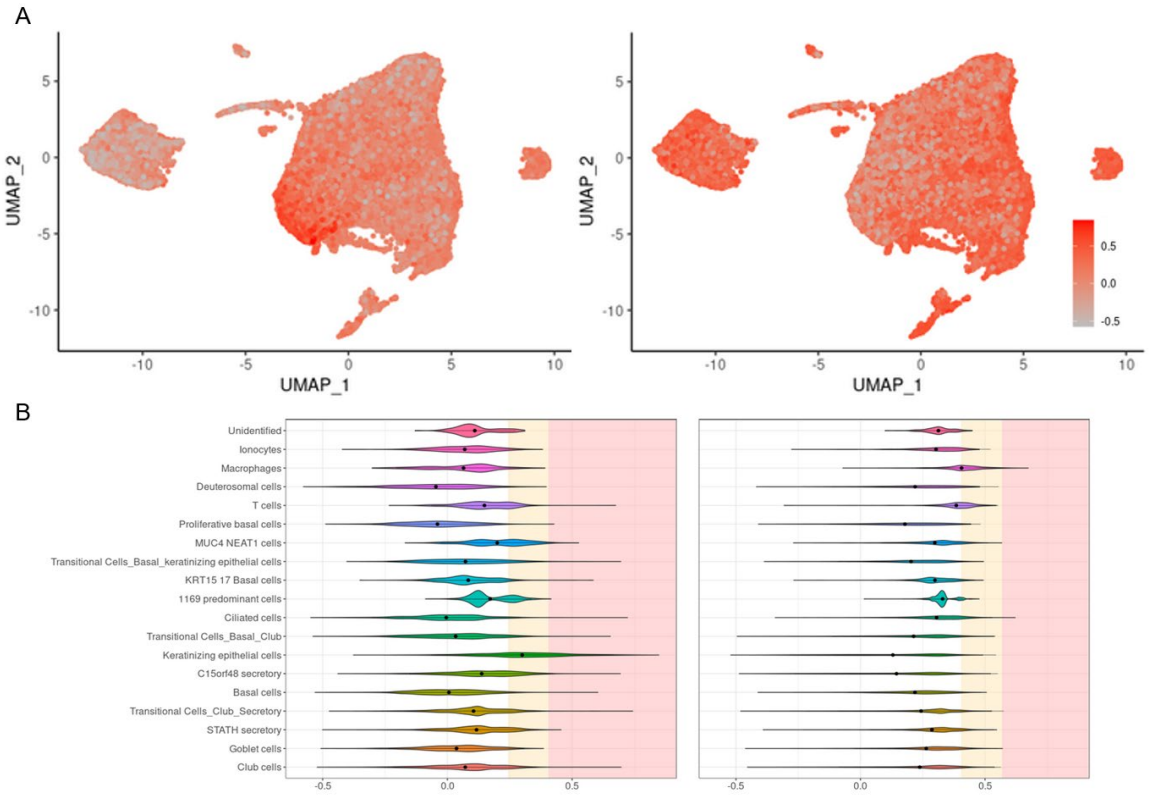

Figure 4.8 Enrichment of lung cancer associated genes in the keratinizing epithelial cells. (A) UMAP projections showing the expression pattern of nasal genes positively- (left) and negatively associated (right) with malignant pulmonary nodules across different cell types. The cells are colored grey for low expression and pink for high expression of metagene scores of each set of genes. (B) Violin plot showing the metagene score for each set of gene module (left: positive association with malignant nodules; right-negative association with malignant nodules) across the cell types. For each violin plot, metagene expression is designated as elevated (light yellow) or highly elevated (pink) if it is greater than one or two standard deviation above the mean metagene score, respectively.

A

	Clinical features	Radiomic features	Genomic features	Combined features
Number of Features	3	40	30	73
Accuracy	0.66	0.65	0.65	**0.71**
Sensitivity	0.83	0.76	0.81	**0.86**
Specificity	0.35	0.43	0.35	**0.43**
Precision	0.70	0.71	0.69	**0.73**
AUC	0.64	0.63	0.70	**0.74**

B

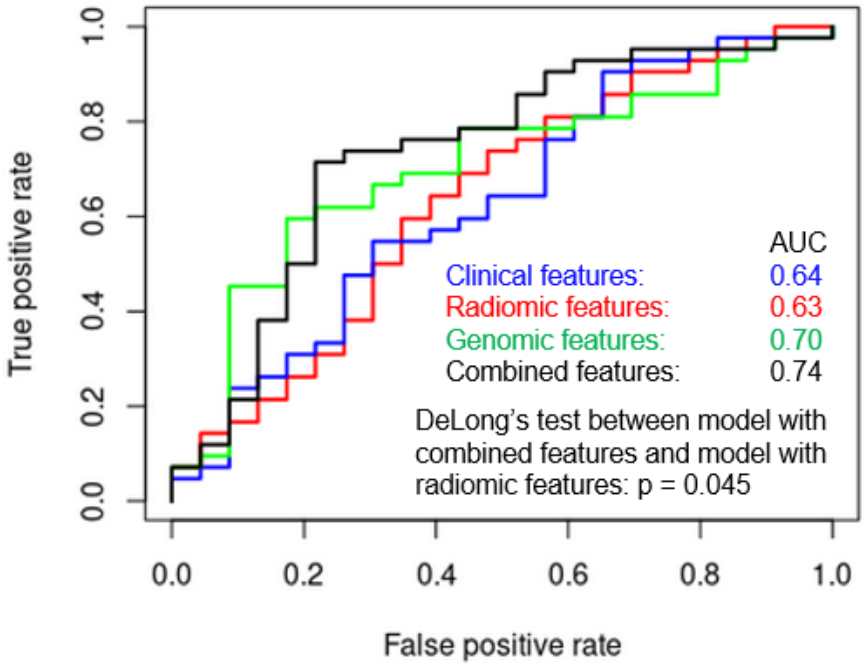

Figure 4.9 Classifier performance of models leveraging clinical, radiomic, genomic, and combined features. (A) Model leveraging a combination of clinical, radiomic, and genomic features showed superior accuracy, sensitivity, specificity, precision, and AUC. (B) Shown are receiver-operating-characteristic curves for the subset of patients with an indeterminant pulmonary nodule. The area under the curve (AUC) was 0.74 (95% CI, 0.60

to 0.87) for the model leveraging a combination of features and 0.63 (95% CI, 0.48 to 0.78) the model only leveraging radiomic features (P = 0.045).

4.5 Discussion

In this study, we detected gene expression alterations in normal-appearing nasal epithelial cells that were associated with malignant pulmonary nodules. To better understand the relationship between the differentially expressed genes and the microenvironment of nasal epithelium, we profiled over 50,000 cells with single cell RNA-sequencing from 15 individuals, 13 of whom had an indeterminate pulmonary nodule (IPN). This single cell RNA-sequencing dataset containing a large number of cells provided us with an unprecedented resolution of the nasal epithelium. We examined the expression of the altered genes across various cell types and saw genes positively associated with malignancy were expressed at high levels in a group of cells that we termed "keratinizing epithelial cells". We further demonstrated the feasibility of integrating the differentially expressed genes into a machine learning model that incorporates radiomic and clinical features to improve model accuracy in determining whether an IPN is malignant. Overall, our findings strengthen the field of injury hypothesis in which there are gene expression alterations in normal-appearing epithelial cells throughout the entire airway of individuals with lung cancer. Furthermore, that these genomic signals could be useful in combination with radiomic features to predict the nature of IPNs.

One unique challenge associated with this analysis was the overall reduced RNA-sequencing quality of the nasal samples compared to the bronchial samples. Both the RNA integrity (RIN) and transcript integrity numbers (TIN) for the nasal samples were

significantly lower. Multiple factors might have contributed to this reduced quality. First, nasal samples may contain significant amounts of mucus, making RNA extraction more difficult. Second, nasal swabs may collect less tissue compared to bronchial brushings as subjects were not sedated. The discomfort associated with nasal sampling is not insignificant, many react to the swab by extending their head, resulting in fewer cells collected. Moreover, nasal samples were often contaminated by the presence of microbes. Here, using a combination of metrics derived from alignment, TIN, RNA, as well as assessment from spectrometric analysis of the RNA content, we were able to detect samples potentially with lower quality to be excluded from downstream analysis.

Single-cell RNA-sequencing is a fast-developing method that has seen great advances in the last five years. A key component of the analysis relies on clustering sequenced cells into groups that share similar gene expression profiles for cell type assignment. However, such assignment often relies on only a handful of genes that have been previously associated with a certain cell type, either through immunohistochemistry or flow cytometry, which limits our capability to identify novel cell clusters. Moreover, in situations where investigators wish to identify and quantify specific cell types of interest across multiple samples or experiments, this method may be cumbersome with differences in clustering strategies resulting in different cell classification. We first derived cell clusters using the traditional workflow and compared the assignment to cell types derived from a novel method by Zhang et al[152]. The new method claims to automate the process of assigning cells in a highly scalable manner across large datasets while controlling for batch and sample effects, leveraging prior knowledge of gene sets to annotate cell types de novo.

We decided to apply gene sets derived from the Deprez dataset onto our single cell RNA-sequencing datasets.

Overall, the cell types assigned by these two methods were comparable. Cell types with distinct markers, genes that are exclusively expressed or expressed at much higher levels in a particular cell type, overlapped the most. For example, ionocytes[27,28], a rare cell type that co-expresses FOXI1, multiple subunits of the vacuolar-type H^+-ATPase (V-ATPase) and CFTR, the gene that is mutated in cystic fibrosis, showed 92% overlap. Similarly, multiciliated cells[151], a common cell type lining the airways and responsible for removing particles from the airways, showed 93% overlap. Because our dataset included a larger number of cells, we were able to sub-divide the secretory cells into STATH+ secretory cells, C15orf48+ secretory cells, goblet cells, and keratinizing epithelial cells. However, it is noteworthy that gene expression alterations within the secretory cells represented gradual shifts.

We identified three subgroups of basal cells in our datasets: KRT5+ basal cells, KRT15/KRT17+ basal cells, and MKI67 proliferating basal cells. The basal cells are relatively undifferentiated epithelial cells, which are located attached to the basal lamina of the stratified and pseudostratified airway epithelium[152]. Basal cells separate the underlying basement membrane from the airway and serve as the progenitor cell population. Basal cells also maintain epithelial integrity by repairing and regenerating the respiratory system[153, 154]. Details about basal cell subpopulations in the nasal epithelium were unclear. Basal cell subsets expressing distinct keratin (KRT) isoforms have been described. Previously, Smirnova et al[155] had noticed that some $KRT5^+$ cells are $p63^-$, which

was attributed to cell renewal in the conducting airways. In contrast to Nakajima et al[156], KRT15 or KRT17 expression was not limited to distal cells in the distal airways. The MKI67+ could represent a group of basal cells undergoing activating cell division, which may be related to self-renewal of the stem cell populations[157].

Similar to the finding of deuterosomal cells in the bronchial epithelium, discussed in detail in Chapter 3, we found deuterosomal cells representing less than 1% of the cell population in the nasal epithelium. Based on the similar gene expression profile to their counterpart in the bronchial epithelium, we hypothesize that they are the precursors of mature multiciliated cells. In particular, the deutersomal cells are characterized by high expression of genes CDC20B, PLK4, and MYB, genes that code for proteins in deuterosome-mediated centriole production in multiciliated cells[158].

Previously, Deprez et al reported a group of secretory cells that express SCEL, SPRR1A, and SPRR1B at higher levels. This group of cells is most similar to the keratinizing epithelial cells found in our datasets. SPRR1A and SPRR1B, coding for Small Proline Rich Protein (Cornifin) 1A and 1B, serve as cross-linked envelop protein of keratinocytes[161]. In addition to SPRR1A and SPRR1B, we also observed high expression of SPRR2A, SPRR2E, and SPRR3 in the keratinizing epithelial cell. Interestingly, S100A7, S100A8, S100A9 were also expressed at higher levels in the keratinizing epithelial cells. Previous studies have shown that S100A8 and S100A9 are highly expressed by neutrophils and monocytes, and are found at high levels in inflammatory conditions[160, 161]. Thus, it could be possible that the keratinizing processes within these cells were the result of a compensatory response of the nasal epithelium to an inflammatory

environment, driven by the activation of S100A7, S100A, and S100A9 (Figure 4.7). Another possible explanation for the presence of these keratinizing epithelial cells was that they represent real keratinocytes present in the outer nostril epithelium, which were harvested during nasal brushing. While this explanation is likely, these cells do express gene markers that are related to the secretory cell population. Therefore, we hypothesize the keratinizing epithelial cells are secretory cells that undergo metaplasia, from a stratified non-keratinized epithelium to stratified keratinized epithelium, under an inflammatory environment and possibly driven by pathologic processes. Metaplasia requires reprogramming of the progenitor cell populations[162]. Our UMAP plot of the cell clusters showed that there was a group of transitional cells between the basal cells and keratinizing epithelial cells (Figure 4.6). Further examination of genes, especially genes of transcription factors, would provide more insight into the possible molecular drivers behind the observed metaplasia. Continued examination of the transition between the basal cells and keratinizing epithelial cells may also contribute to a further understanding of the role of inflammation and stem cell reprogramming[163].

The enrichment of nasal genes positively associated with malignancy in the keratinizing epithelial could be explained by 1) an increase in the proportion of keratinizing epithelial cells in the nasal microenvironment or 2) increased expression of genes that are characteristic of the keratinizing epithelial cells, as part of metaplasia process. Though the two mechanisms are not mutually exclusive. Squamous cell metaplasia is a well-known epithelial alteration of the human tracheobronchial mucosa[164]. Squamous metaplasia in bronchial epithelium has previously been found to be associated with COPD[165], and it is a

precursor to low-grade dysplasia[166], which can culminate in high-grade dysplasia and carcinoma [167, 168, 169]. The detection of biologic processes in the nasal epithelium strengthens the field of injury hypothesis in which there are gene expression alterations in normal-appearing epithelial cells throughout the entire airway of individuals with lung cancer.

The application of XGBoost algorithm to build a classifier differentiating malignant from benign nodules marks a departure from our previous attempts in building prediction tools that utilize logistic regression models[38, 170]. XGBoost algorithm has won multiple classifier building competitions[171] and is especially suited for non-linear models. It has also been applied to identify cancer Tissue-of-Origin based on copy number variations[172], detect cancer-related long non-coding RNAs[173], and predict 1-year survival in non-small-cell lung cancer patients[174].

While we did not have the features that are part of the model built by McWilliams et al[142], we did have information on age, sex, FEV1%, nodule size, and whether a nodule was located in the upper lung. We also integrated pack years into our model using the clinical variables and found that nodule size, pack-years, and age were the three most important features. However, we did not find the location of the nodule or sex to be as important. Leveraging radiomic features derived at different peripheries of a nodule, our model achieved a modest predictive power with an AUC of 0.63, slightly lower than that achieved by using the clinical features. Future work could examine how to utilize these imaging features to achieve better predictive power. With the combination of genomic features, genes that were differentially expressed in the nasal epithelium from subjects with

lung cancer, the predictive power of the model increased significantly to achieve an AUC of 0.74 ($p < 0.05$). However, an independent cohort is needed to validate the utility of this model.

4.6 Conclusion

In conclusion, nasal gene expression differences in individuals with malignancy reflect squamous metaplasia that has previously been related to lung cancer, strengthening the field of injury hypothesis observed in smoking, COPD, and lung cancer. A large number of nasal epithelial cells have been profiled for an unprecedented resolution of the cellular landscape of the proximal airway, while also providing us the opportunity to examine possible shifts in cellular composition that may be related to lung cancer distally. Finally, a classifier combined clinical factors (age, nodule size, pack years), radiomic features (40 measurements) and nasal gene expression (30 genes) had statistically significantly higher area under the curve (0.74; $p < 0.04$) and sensitivity (0.71; $p < 0.03$) than a radiomic feature only model in the test set of samples. While an independent cohort is needed to validate the model utility and more samples are needed for improving the model accuracy, we have demonstrated the feasibility of combining features acquired from different methods to detect weak signals associated with malignant pulmonary nodules.

CHAPTER FIVE:

General Conclusions and Future Directions

The studies featured in this dissertation collectively used high-throughput transcriptomic profiling of bronchial and nasal airway epithelium to assess the molecular alterations associated with a variety of radiographic abnormalities. Taken together, the result from these chapters contend that:

- Our analyses strengthen the field of injury hypothesis in that gene expression alterations at the proximal airway correlated with pathologic processes in the distal airway, and that these gene expression alterations provided insights into the molecular events related to the lung pathologies. We found 1) evidence in an altered inflammatory environment in the bronchial epithelium in the analysis of gene expression alterations associated with patient subgroups related to chronic obstructive pulmonary disease (COPD), 2) signals of reduced cell adhesion pathway in the bronchial epithelium from those with widespread radiographic bronchiectasis (BE), and 3) markers of squamous metaplasia in the nasal epithelium to be expressed at high levels among those with malignant pulmonary nodules. The inflammation, cell adhesion, and squamous metaplasia all have been previously associated with COPD, BE, and lung cancer pathophysiology.
- Radiomic features from High Resolution Computed Tomography (HRCT) may be useful in clustering patients into subgroups that correlated with similar clinical characteristics, disease progression, and gene expression profiles of the airway. These subgroups may improve our understanding of disease pathophysiology. Longitudinal follow-ups would allow the examination of whether these subgroups warrant differential care.

- Gene expression alterations may be caused by a shift of cellular composition within the tissue microenvironment. Previous transcriptomic analyses have the limitation of separating transcriptional control from a change in the cellular composition of a tissue. Using advanced computational algorithms leveraging marker genes derived from single cell RNA-sequencing, we demonstrate that gene expression alterations related to COPD subgroups and radiographic BE could be a result of the change tissue microenvironment.
- Single cell RNA-sequencing is a powerful tool in the examination of cell transcriptomics at a high resolution. In addition to detecting previously unknown, rare cell populations that may be disease-related (ionocytes and deuterosomal cells), differentiating subgroups of a cell type (KRT5+ basal cells, KRT15/17+ basal cells, and proliferating basal cells), it can also reveal various cell states of a cell type that could reflect the influence of chemokine and cytokines (keratinizing epithelial cells). Knowing the expression profile of individual single cell also provide us with the opportunity to "project" gene set enrichment scores derived from a traditional method like Gene Set Variation Analysis (GSVA), which could be useful to further examination of transcriptomic alterations detected at the bulk level.

There are several limitations throughout these analyses, some are consistent across all three studies:

- The derivation of patient subgroups using radiographic features is prone to individual subjectivity. This limitation range from detecting signs for radiographic bronchiectasis (BE) from one expert, to selecting pre-determined features associated with chronic pulmonary obstructive disease (COPD) by a group of experts to discover imaging clusters, and imaging analysis using predefined measurement in the case of assessing indeterminate pulmonary nodules (IPN). While we tried our best at each step to our best intention to keep the assessment as objective as possible, it is hard to achieve complete objectivity. Possible solutions may include clustering patients using features in an unsupervised manner. Imaging analysis through deep learning may be another solution[175,176], although this computer vision may require a significantly larger number of images than our current datasets. Computer vision may also be less useful in detecting new clusters de novo.
- Perhaps one reason for our semi-supervised or supervised derivation of subgroups of patients with similar imaging features was that our datasets consist HRCT scans captured at different clinical sites using different instruments, leading to artifacts in images that require experts to intervene to obtain meaningful measurements. The batch effect, whether across clinical sites, or within the same sites may always propose a challenge in studies like this. However, an imaging standardization tool may be useful in reducing the batch effect[177].
- Lack of longitudinal follow-ups of the participants in each analysis has limited us from more insights into what each patient subgroup represents. A central question

in the work related to COPD imaging clustering is how to interpret the groups of disease identified – whether they represent temporal disease stages or groups with differential disease tolerance may lead to different clinical implications. In the analysis of radiographic BE, a longitudinal follow-up would allow us to assess the BE risks associated with smoking and further interrogate whether cough and phlegm production were indeed signs of radiographic BE in a subset of patients. Furthermore, longitudinal follow-up of the subjects that have been profiled by single cell RNA-sequencing would allow us the analysis of cell microenvironment changes that may be related to malignancy[178,179].

To our knowledge, this body of work marks the novel application of human airway transcriptomics in assessing molecular workings associated with radiographic abnormality. The combination of transcriptomics and imaging together may not only lead to a better understanding of disease pathophysiology but produce useful tools in personalized medicine to benefit patients directly.

LIST OF JOURNAL ABBREVIATIONS

ACM	Asscoiation for Computing Machinery
Acad. Radiol.	Academic Radiology
Am. J. Physiol. Lung. Cell. Mol. Physiol.	American Journal of Physiology-Lung Cellular and Molecular Physiology
Am. J. Pathol.	American Journal of Pathology
Am. J. Respir. Cell. Mol. Biol.	American Journal of Respiratory Cell and Molecular Biology
Am. J. Respir. Crit. Care Med.	American Journal of Respiratory and Critical Care Medicine
Am. J. Roentgenol.	American Journal of Roentgenology
Am. Rev. Respir. Dis.	American review of respiratory disease
Ann. Am. Thorac. Soc.	Annals of the. American Thoracic Society
Ann. Intern. Med.	Annals of Internal Medicine
Annu. Rev. Immunol.	Annual Review of Immunology
Arch. Dis. Child.	Archives of Disease in Childhood
Arthritis. Rheum.	Arthritis & Rheumatology
Bioinforma. Oxf. Engl.	Bioinformatics Oxford
Biomed. Res. Int.	BioMed Research International
BMC. Molecular. Biol.	BMC Molecular Biology
BMC. Pulm. Med.	BMC Pulmonary Medicine
Br. J. Pharmacol.	British Journal of Pharmacology

Cancer. Prev. Res.	Cancer Prevention Research
Chron. Respir. Dis.	Chronic Respiratory Disease
COPD	Journal of Chronic Obstructive Pulmonary Disease
Dig. Dis. Sci.	Digestive Diseases and Sciences
Digit. Med.	Digital Medicine
Dis. Model. Mech.	Disease Models & Mechanisms
Eur. Radiol.	European Radiology
Eur. Respir. J.	European Respiratory Journal
Front. Genet.	Frontiers in Genetics
IEEE. Trans. Med. Imaging.	IEEE Transactions on Medical Imaging
Hum. Mutat.	Human Mutation
Int. J. Chron. Obstruct. Pulmon. Dis.	International Journal of Chronic Obstructive Pulmonary Disease
JAMA	Journal of the American Medical Association
JAMA. Oncol.	JAMA Oncology
J. Biol. Chem.	Journal of Biological Chemistry
J. Clin. Virol. Off. Publ. Pan. Am. Soc. Clin. Virol.	Journal of the Pan American Society for Clinical Virology
J. Immunol.	Journal of Immunology
J. Leukoc. Biol.	Journal of Leukocyte Biology
J. Natl. Cancer. Inst.	Journal of the National Cancer Institute
J. Pediatr.	Journal of Pediatrics
J. Thorac. Dis.	Journal of Thoracic Disease

J. Thorac. Imaging.	Journal of Thoracic Imaging
Lancet. Respir. Med.	Lancet Respiratory Medicine
Nat. Biotechnol.	Nature Biotechnology
Nat. Commun.	Nature Communication
Nat. Method.	Nature Method
Nat. Rev. Cancer.	Nature Reviews Cancer
Nat. Rev. Dis. Primer.	Nature Reviews Disease Primers
Nat. Rev. Immunol.	Nature Reviews Immunology
NAR. Genomics. Bioinforma.	NAR Genomics and Bioinformatics
N. Engl. J. Med.	New England Journal of Medicine
Nucleic. Acids. Res.	Nucleic Acids Research
Pathol. Int.	Pathology International
Pediatr. Pulmonol.	Pediatric Pulmonology
Pharmacol. Ther.	Pharmacology & Therapeutics
PLoS. Med.	PLOS Medicine
PLoS. Pathog.	PLOS Pathogens
Prim. Care. Respir. J.	Primary Care Respiratory Journal
Proc. Am. Thorac. Soc.	Proceedings of the American Thoracic Society
Proc. Natl. Acad. Sci. U S A.	Proceedings of the National Academy of Sciences of the United States of America
Pulm. Pharmacol. Ther.	Pulmonary Pharmacology and Therapeutics
Respirol Carlton Vic	Respirology

Respir. Med.	Respiratory Medicine
Respir. Res.	Respiratory Research
Sci. Adv.	Science Advances
Sci. Rep.	Scientific Reports
Sci. Signal.	Science Signaling
Trends. Mol. Med.	Trends in Molecular Medicine
Virchows. Archiv. B Cell Pathol.	Virchows Archiv B Cell Pathology

BIBLIOGRAPHY

1. Stark P. High resolution computed tomography of the lungs In: Post T, ed. UpToDate. *UpToDate*; 2021. Accessed March 1st, 2021. www.uptodate.com

2. Epler GR, McLoud TC, Gaensler EA, Mikus JP, Carrington CB. Normal chest roentgenograms in chronic diffuse infiltrative lung disease. *N Engl J Med.* 1978 Apr 27;298(17):934-9. doi: 10.1056/NEJM197804272981703. PMID: 642974.

3. van der Bruggen-Bogaarts BA, van der Bruggen HM, van Waes PF, Lammers JW. Screening for bronchiectasis. A comparative study between chest radiography and high-resolution CT. *Chest.* 1996 Mar;109(3):608-11. doi: 10.1378/chest.109.3.608. PMID: 8617064.

4. Thurlbeck WM, Simon G. Radiographic appearance of the chest in emphysema. *AJR Am J Roentgenol*. 1978 Mar;130(3):429-40. doi: 10.2214/ajr.130.3.429. PMID: 415543.

5. Mathieson JR, Mayo JR, Staples CA, Müller NL. Chronic diffuse infiltrative lung disease: comparison of diagnostic accuracy of CT and chest radiography. *Radiology*. 1989 Apr;171(1):111-6. doi: 10.1148/radiology.171.1.2928513. PMID: 2928513.

6. Müller NL, Miller RR. Computed tomography of chronic diffuse infiltrative lung disease. Part 1. *Am Rev Respir Dis*. 1990 Nov;142(5):1206-15. doi: 10.1164/ajrccm/142.5.1206. PMID: 2240845.

7. Müller NL, Miller RR. Computed tomography of chronic diffuse infiltrative lung disease. Part 2. *Am Rev Respir Dis*. 1990 Dec;142(6 Pt 1):1440-8. doi: 10.1164/ajrccm/142.6_Pt_1.1440. PMID: 2252265.

8. Current best practice for nebuliser treatment. The Nebulizer Project Group of the British Thoracic Society Standards of Care Committee. *Thorax*. 1997 Apr;52 Suppl 2:S1-3. Erratum in: Thorax 1997 Sep;52(9):838. PMID: 9155846.

9. Miller RR, Müller NL, Vedal S, Morrison NJ, Staples CA. Limitations of computed tomography in the assessment of emphysema. *Am Rev Respir Dis*. 1989 Apr;139(4):980-3. doi: 10.1164/ajrccm/139.4.980. PMID: 2930075.

10. Müller NL, Coxson H. Chronic obstructive pulmonary disease. 4: imaging the lungs in patients with chronic obstructive pulmonary disease. *Thorax*. 2002;57(11):982-985. doi:10.1136/thorax.57.11.982

11. Pasteur MC, Bilton D, Hill AT; British Thoracic Society Bronchiectasis non-CF Guideline Group. British Thoracic Society guideline for non-CF bronchiectasis. Thorax. 2010 Jul;65 Suppl 1:i1-58. doi: 10.1136/thx.2010.136119. PMID: 20627931.
12. Edwards EA, Metcalfe R, Milne DG, Thompson J, Byrnes CA. Retrospective review of children presenting with non cystic fibrosis bronchiectasis: HRCT features and clinical relationships. *Pediatr Pulmonol*. 2003 Aug;36(2):87-93. doi: 10.1002/ppul.10339. PMID: 12833486.

13. Alzeer AH. HRCT score in bronchiectasis: correlation with pulmonary function tests and pulmonary artery pressure. *Ann Thorac Med*. 2008;3(3):82-86. doi:10.4103/1817-1737.39675

14. Tockman M. Survival and mortality from lung cancer in a screening population: the Johns Hopkins study. *Chest 1986*;*89*:*324S*–326S.

15. Fontana RS, Sanderson DR, Woolner LB, Taylor WF, Miller WE, Muhm JR. Lung cancer screening: the Mayo program. *J Occup Med* 1986*;*28*:*746–*750.*

16. Frost JK, Ball WC Jr, Levin ML, et al. Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Johns Hopkins study. *Amer Rev Respir Dis* 1984*;*130*:*549–554.

17. Melamed MR, Flehinger BJ, Zaman MB, Heelan RT, Perchick WA, Martini N. Screening for early lung cancer. Results of the Memorial Sloan-Kettering study in New York. *Chest* 1984;86:44–53.

18. Kubik A, Parkin DM, Khlat M, Erban J, Polak J, Adamec M. Lack of benefit from semi-annual screening for cancer of the lung: follow-up report of a randomized controlled trial on a population of high-risk males in Czechoslovakia. *Int J Cancer* 1990;45:26–33.

19. Henschke CI, McCauley DI, Yankelevitz DF, et al. Early Lung Cancer Action Project: overall design and findings from baseline screening. *Lancet* 1999*;*354:99–105.

20. Kaneko M, Eguchi K, Ohmatsu H, et al. Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography. *Radiology 1996*;*201*:*798*–802.

21. Sone S, Takashima S, Li F, et al. Mass screening for lung cancer with mobile spiral computed tomography scanner. *Lancet* 1998*;*351*:*1242*–1245.*

22. US Preventive Services Task Force, Krist AH, Davidson KW, Mangione CM, Barry MJ, Cabana M, Caughey AB, Davis EM, Donahue KE, Doubeni CA, Kubik M, Landefeld CS, Li L, Ogedegbe G, Owens DK, Pbert L, Silverstein M, Stevermer J, Tseng CW, Wong JB. Screening for Lung Cancer: US Preventive Services Task Force Recommendation

Statement. *JAMA*. 2021 Mar 9;325(10):962-970. doi: 10.1001/jama.2021.1117. PMID: 33687470.

23. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat Rev Genet* **20,** 631–656 (2019).

24. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).

25. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Göttgens B, Hacohen N, Haniffa M, Hemberg M, Kim S, Klenerman P, Kriegstein A, Lein E, Linnarsson S, Lundberg E, Lundeberg J, Majumder P, Marioni JC, Merad M, Mhlanga M, Nawijn M, Netea M, Nolan G, Pe'er D, Phillipakis A, Ponting CP, Quake S, Reik W, Rozenblatt-Rosen O, Sanes J, Satija R, Schumacher TN, Shalek A, Shapiro E, Sharma P, Shin JW, Stegle O, Stratton M, Stubbington MJT, Theis FJ, Uhlen M, van Oudenaarden A, Wagner A, Watt F, Weissman J, Wold B, Xavier R, Yosef N; Human Cell Atlas Meeting Participants. The Human Cell Atlas. *Elife*. 2017 Dec 5;6:e27041. doi: 10.7554/eLife.27041. PMID: 29206104; PMCID: PMC5762154.

26. Insel TR, Landis SC, Collins FS. Research priorities. The NIH BRAIN Initiative. *Science*. 2013 May 10;340(6133):687-8. doi: 10.1126/science.1239276. PMID: 23661744; PMCID: PMC5101945.

27. Montoro, D.T., Haber, A.L., Biton, M. *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560,** 319–324 (2018).

28. Plasschaert, L.W., Žilionis, R., Choo-Wing, R. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560,** 377–381 (2018).
29. Chai H, Brown RE. Field effect in cancer-an update. *Ann Clin Lab Sci*. 2009 Fall;39(4):331-7. PMID: 19880759.

30. Slaughter DP, Southwick HW, Smejkal W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer* 1953*;6:*963–968.

31. Braakhuis BJ, Tabor MP, Kummer JA, Leemans CR, Brakenhoff RH. A genetic explanation of Slaughter's concept of field cancerization: evidence and clinical implications. *Cancer Res* 2003;63:1727–1730.

32. Yan PS, Venkataramu C, Ibrahim A, Liu JC, Shen RZ, Diaz NM, Centeno B, Weber F, Leu YW, Shapiro CL, Eng C, Yeatman TJ, Huang TH. Mapping geographic zones of

cancer risk with epigenetic biomarkers in normal breast tissue. *Clin Cancer Res* 2006;12:626–636.

33. Shen L, Kondo Y, Rosner GL, Xiao L, Hernandez NS, Vilaythong J, Houlihan PS, Krouse RS, Prasad AR, Einspahr JG, Buckmeier J, Alberts DS, Hamilton SR, Issa JP. MGMT promoter methylation and field defect in sporadic colorectal cancer. *J Natl Cancer Inst* 2005*;*97*:*1330–1338.

34. Franklin WA, Gazdar AF, Haney J, Wistuba, II, La Rosa FG, Kennedy T, Ritchey DM, Miller YE. Widely dispersed p53 mutation in respiratory epithelium. A novel mechanism for field carcinogenesis. *J Clin Invest* 1997*;*100*:*2133–2137.

35. Spira A, Beane JE, Shah V, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med*. 2007;13(3):361-366. doi:10.1038/nm1556

36. Suzuki H, Watkins DN, Jair KW, Schuebel KE, Markowitz SD, Chen WD, Pretlow TP, Yang B, Akiyama Y, Van Engeland M, Toyota M, Tokino T, Hinoda Y, Imai K, Herman JG, Baylin SB. Epigenetic inactivation of SFRP genes allows constitutive WNT signaling in colorectal cancer. *Nat Genet*. 2004 Apr;36(4):417-22. doi: 10.1038/ng1330. Epub 2004 Mar 14. PMID: 15034581.

37. Baker SJ, Fearon ER, Nigro JM, Hamilton SR, Preisinger AC, Jessup JM, vanTuinen P, Ledbetter DH, Barker DF, Nakamura Y, White R, Vogelstein B. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science*. 1989 Apr 14;244(4901):217-21. doi: 10.1126/science.2649981. PMID: 2649981.

38. AEGIS Study Team. Shared Gene Expression Alterations in Nasal and Bronchial Epithelium for Lung Cancer Detection. *J Natl Cancer Inst*. 2017 Jul 1;109(7):djw327. doi: 10.1093/jnci/djw327. PMID: 28376173; PMCID: PMC6059169.

39. Steiling K, van den Berge M, Hijazi K, et al. A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. *Am J Respir Crit Care Med*. 2013;187(9):933-942. doi:10.1164/rccm.201208-1449OC

40. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International. Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.

41. Machine learning challenge winning solutions. Github; 2021. Accessed March 1st, https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions

42. Vestbo J, Hurd SS, Agustí AG, Jones PW, Vogelmeier C, Anzueto A, Barnes PJ, Fabbri LM, Martinez FJ, Nishimura M, Stockley RA, Sin DD, Rodriguez-Roisin R. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med*. 2013 Feb 15;187(4):347-65. doi: 10.1164/rccm.201204-0596PP. Epub 2012 Aug 9. PMID: 22878278.

43. Qaseem A (WAS 24), Wilt TJ, Weinberger SE, et al, and the American College of Physicians, and the American College of Chest Physicians, and the American Thoracic Society, and the European Respiratory Society. Diagnosis and management of stable chronic obstructive pulmonary disease: a clinical practice guideline update from the American College of Physicians, American College of Chest Physicians, American Thoracic Society, and European Respiratory Society. *Ann Intern Med* 2011; 155: 179-91.

44. Mapel DW, Dalal AA, Johnson PT, et al. Application of the new GOLD COPD staging system to a USA primary care cohort, with comparison to physician and patients impressions of severity. *Int J Chron Obstruct Pulmon Dis* 2015; 10: 1477-86

45. Chalmers JD, Chang AB, Chotirmall SH, Dhar R, McShane PJ. Bronchiectasis. *Nat Rev Dis Primer*. 2018;4(1):45. doi:10.1038/s41572-018-0042-3

46. Chronic respiratory diseases: Burden of COPD. *World Health Organization*. https://www.who.int/respiratory/copd/burden/en/

47. The Top 10 Causes of Death. *World Health Organization*, World Health Organization, www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

48. Han MK, Martinez CH, Au DH, Bourbeau J, et al. Meeting the challenge of COPD care delivery in the USA: a multiprovider perspective. *Lancet Respir Med*. 2016 Jun;4(6):473-526

49. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med* 2006; 3(11): e442.

50. Haroon S, Jordan RE, Fitzmaurice DA, Adab P. Case finding for COPD in primary care: a qualitative study of the views of health professionals. *Int J Chron Obstruct Pulmon Dis* 2015; 10: 1711-18

51. Walters JA, Walters EH, Nelson M et al. Factors associated with misdiagnosis of COPD in primary care. *Prim Care Respir J* 2011; 20: 396-402.

52. Davis KJ, Landis SH, Oh YM, et al. Continuing to CONFRONT COPD International Physician Survey: physician knowledge and application of COPD management guidelines in 12 countries. *Int J Chron Obstruct Pulmon Dis* 2015; 10: 39-55.

53. Perez X, Wisnivesky JP, Lurslurchachai L, et al. Barriers to adherence to COPD guidelines among primary care providers. *Respir Med* 2012; 106: 374-81.
54. Rennard S, Thomashow B, Crapo J, et al. Introducing the COPD Foundation Guide for Diagnosis and Management of COPD, recommendations of the COPD Foundation. COPD 2013; 10: 378-89.

55. Rodrigo GJ, Soler-Cataluna JJ, Solanes I, et al, Assessment of the internal structure of GOLD 2011 system. *Pulm Pharmacol Ther* 2015; 30: 87-92.

56. Mapel DW, Dalal AA, Johnson PT, et al. Application of the new GOLD COPD staging system to a USA primary care cohort, with comparison to physician and patients impressions of severity. *Int J Chron Obstruct Pulmon Dis* 2015; 10: 1477-86

57. Overington JD, Huang YC, Abramsom MJ, el al. Implementing clinical guidelines for chronic obstructive pulmonary disease: barriers and solutions. *J Thorac Dis* 2014; 6: 1586-96.

58. Castaldi PJ, Dy J, Ross J, et al. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax* 2014;69(5):415–422.

59. Garcia-Aymerich J, Gómez FP, Benet M, et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax* 2011;66(5):430–437.

60. Rennard SI. Chronic obstructive pulmonary disease: Linking outcomes and pathobiology of disease modification. *Proc Am Thorac Soc* 2006;3(3):276–280.

61. Rennard SI, Locantore N, Delafont B, et al. Identification of five chronic obstructive pulmonary disease subgroups with different prognoses in the ECLIPSE cohort using cluster analysis. *Ann Am Thorac Soc* 2015;12(3):303–312.

62. Sieren JP, Newell JD, Barr RG, et al. SPIROMICS Protocol for Multicenter Quantitative Computed Tomography to Phenotype the Lungs. *Am J Respir Crit Care Med* 2016;194(7):794–806.

63. Regan EA, Hokanson JE, Murphy JR, et al. Genetic Epidemiology of COPD (COPDGene) Study Design. *COPD* 2010;7(1):32–43.

64. Billatos, E., Duan, F., Moses, E. *et al.* Detection of early lung cancer among military personnel (DECAMP) consortium: study protocols. *BMC Pulm Med* **19,** 59 (2019).

65. Billatos E, Ash SY, Duan F, et al. Distinguishing smoking related lung disease phenotypes via imaging and molecular features. *Chest*. Published online September 15, 2020. doi:10.1016/j.chest.2020.08.2115

66. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. 2017;35(4):316-319. doi:10.1038/nbt.3820

67. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinforma Oxf Engl*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

68. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12(1):323. doi:10.1186/1471-2105-12-323

69. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/nar/gkv007

70. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607-D613. doi:10.1093/nar/gky1131

71. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102

72. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinforma Oxf Engl*. 2011;27(12):1739-1740. doi:10.1093/bioinformatics/btr260

73. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:7. doi:10.1186/1471-2105-14-7

74. Newman, A.M., Steen, C.B., Liu, C.L. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37,** 773–782 (2019).

75. Malhotra S, Bustamante MF, Pérez-Miralles F, Rio J, Ruiz de Villa MC, Vegas E, Nonell L, Deisenhammer F, Fissolo N, Nurtdinov RN, Montalban X, Comabella M. Search for specific biomarkers of IFNβ bioactivity in patients with multiple sclerosis. *PLoS One*. 2011;6(8):e23634. doi: 10.1371/journal.pone.0023634. Epub 2011 Aug 23. PMID: 21886806; PMCID: PMC3160307.

76. Bolen CR, Ding S, Robek MD, Kleinstein SH. Dynamic expression profiling of type I and type III interferon-stimulated hepatocytes reveals a stable hierarchy of gene expression. *Hepatology*. 2014 Apr;59(4):1262-72. doi: 10.1002/hep.26657. Epub 2014 Feb 18. PMID: 23929627; PMCID: PMC3938553.

77. Matta SK, Olias P, Huang Z, Wang Q, Park E, Yokoyama WM, Sibley LD. *Toxoplasma gondii* effector TgIST blocks type I interferon signaling to promote infection. *Proc Natl Acad Sci U S A*. 2019 Aug 27;116(35):17480-17491. doi: 10.1073/pnas.1904637116. Epub 2019 Aug 14. PMID: 31413201; PMCID: PMC6717281.

78. Shapira SD, Gat-Viks I, Shum BO, Dricot A, de Grace MM, Wu L, Gupta PB, Hao T, Silver SJ, Root DE, Hill DE, Regev A, Hacohen N. A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell*. 2009 Dec 24;139(7):1255-67. doi: 10.1016/j.cell.2009.12.018. PMID: 20064372; PMCID: PMC2892837.

79. Schuliga M. NF-kappaB Signaling in Chronic Inflammatory Airway Disease. *Biomolecules* 2015;5(3):1266–1283.

80. Edwards MR, Bartlett NW, Clarke D, Birrell M, Belvisi M, Johnston SL. Targeting the NF-kappaB pathway in asthma and chronic obstructive pulmonary disease. *Pharmacol Ther* 2009;121(1):1–13.

81. Zaynagetdinov R, Sherrill TP, Gleaves LA, et al. Chronic NF-κB activation links COPD and lung cancer through generation of an immunosuppressive microenvironment in the lungs. *Oncotarget* 2015;7(5):5470–5482.

82. Ivashkiv LB, Donlin LT. Regulation of type I interferon responses. *Nat Rev Immunol* 2014;14(1):36–49.

83. Singanayagam A, Loo S-L, Calderazzo MA, et al. Anti-viral immunity is impaired in COPD patients with frequent exacerbations. *Am J Physiol Lung Cell Mol Physiol* 2019.

84. Gough DJ, Messina NL, Clarke CJP, Johnstone RW, Levy DE. Constitutive type I interferon modulates homeostatic balance through tonic signaling. *Immunity* 2012;36(2):166–174.

85. Crotta S, Davidson S, Mahlakoiv T, et al. Type I and type III interferons drive redundant amplification loops to induce a transcriptional signature in influenza-infected airway epithelia. *PLoS Pathog* 2013;9(11):e1003773.

86. Mallia P, Message SD, Gielen V, et al. Experimental rhinovirus infection as a human model of chronic obstructive pulmonary disease exacerbation. *Am J Respir Crit Care Med* 2011;183(6):734–742.

87. García-Valero J, Olloquequi J, Montes JF, et al. Deficient pulmonary IFN-β expression in COPD patients. *PLoS ONE* [Internet] 2019 [cited 2019 Oct 9];14(6).

88. Leung JM, Tiew PY, Mac Aogáin M, et al. The role of acute and chronic respiratory colonization and infections in the pathogenesis of COPD. *Respirol Carlton Vic* 2017;22(4):634–650.

89. Zwaans W a. R, Mallia P, Winden MEC van, Rohde GGU. The relevance of respiratory viral infections in the exacerbations of chronic obstructive pulmonary disease—a systematic review. *J Clin Virol Off Publ Pan Am Soc Clin Virol* 2014;61(2):181–188.

90. George SN, Garcha DS, Mackay AJ, et al. Human rhinovirus infection during naturally occurring COPD exacerbations. *Eur Respir J* 2014;44(1):87–96.

91. Seemungal T, Harper-Owen R, Bhowmik A, et al. Respiratory viruses, symptoms, and inflammatory markers in acute exacerbations and stable chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2001;164(9):1618–1623.

92. Parr DG. Quantifying the Lung at Risk in Chronic Obstructive Pulmonary Disease. Does Emphysema Beget Emphysema? *Am J Respir Crit Care Med* 2017;196(5):535–536.

93. Runkel L, Pfeffer L, Lewerenz M, Monneron D, Yang CH, Murti A, Pellegrini S, Goelz S, Uzé G, Mogensen K. Differences in activity between alpha and beta type I interferons explored by mutational analysis. *J Biol Chem*. 1998 Apr 3;273(14):8003-8. doi: 10.1074/jbc.273.14.8003. PMID: 9525899.

94. Furusyo N, Hayashi J, Ohmiya M, Sawayama Y, Kawakami Y, Ariyama I, Kinukawa N, Kashiwagi S. Differences between interferon-alpha and -beta treatment for patients with chronic hepatitis C virus infection. *Dig Dis Sci*. 1999 Mar;44(3):608-17. doi: 10.1023/a:1026625928117. PMID: 10080158.

95. Vieira SM, Lemos HP, Grespan R, et al. A crucial role for TNF-alpha in mediating neutrophil influx induced by endogenously generated or exogenous chemokines,

KC/CXCL1 and LIX/CXCL5. *Br J Pharmacol*. 2009;158(3):779-789. doi:10.1111/j.1476-5381.2009.00367.x

96. Tessier PA, Naccache PH, Clark-Lewis I, Gladue RP, Neote KS, McColl SR. Chemokine networks in vivo: involvement of C-X-C and C-C chemokines in neutrophil extravasation in vivo in response to TNF-alpha. *J Immunol*. 1997 Oct 1;159(7):3595-602. PMID: 9317159.

97. Chen K, Wei Y, Alter A, Sharp GC, Braley-Mullen H. Chemokine expression during development of fibrosis versus resolution in a murine model of granulomatous experimental autoimmune thyroiditis. *J Leukoc Biol*. 2005 Sep;78(3):716-24. doi: 10.1189/jlb.0205102. Epub 2005 Jun 16. PMID: 15961577.

98. Zafranskaya M, Oschmann P, Engel R, et al. Interferon-beta therapy reduces CD4+ and CD8+ T-cell reactivity in multiple sclerosis. *Immunology*. 2007;121(1):29-39. doi:10.1111/j.1365-2567.2006.02518.x

99. Rudick RA, Carpenter CS, Cookfair DL, Tuohy VK, Ransohoff RM. In vitro and in vivo inhibition of mitogen-driven T-cell activation by recombinant interferon beta. *Neurology*. 1993 Oct;43(10):2080-7. doi: 10.1212/wnl.43.10.2080. PMID: 8105424.

100. Pette M, Pette DF, Muraro PA, Farnon E, Martin R, McFarland HF. Interferon-beta interferes with the proliferation but not with the cytokine secretion of myelin basic protein-specific, T-helper type 1 lymphocytes. *Neurology*. 1997 Aug;49(2):385-92. doi: 10.1212/wnl.49.2.385. PMID: 9270566.

101. Theofilopoulos AN, Baccala R, Beutler B, Kono DH. Type I interferons (alpha/beta) in immunity and autoimmunity. *Annu Rev Immunol*. 2005;23:307-36. doi: 10.1146/annurev.immunol.23.021704.115843. PMID: 15771573.

102. Putman RK, Hatabu H, Araki T, Gudmundsson G, Gao W, Nishino M, Okajima Y, Dupuis J, Latourelle JC, Cho MH, El-Chemaly S, Coxson HO, Celli BR, Fernandez IE, Zazueta OE, Ross JC, Harmouche R, Estépar RS, Diaz AA, Sigurdsson S, Gudmundsson EF, Eiríksdottír G, Aspelund T, Budoff MJ, Kinney GL, Hokanson JE, Williams MC, Murchison JT, MacNee W, Hoffmann U, O'Donnell CJ, Launer LJ, Harrris TB, Gudnason V, Silverman EK, O'Connor GT, Washko GR, Rosas IO, Hunninghake GM; Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) Investigators; COPDGene Investigators. Association Between Interstitial Lung Abnormalities and All-Cause Mortality. *JAMA*. 2016 Feb 16;315(7):672-81. doi: 10.1001/jama.2016.0518. PMID: 26881370; PMCID: PMC4828973.

103. Ash SY, Washko GR. Interstitial lung abnormalities: risk and opportunity. *Lancet Respir Med*. 2017 Feb;5(2):95-96. doi: 10.1016/S2213-2600(17)30006-1. PMID: 28145231; PMCID: PMC5797991.

104. Agusti A, Faner R. Lung function trajectories in health and disease. *Lancet Respir Med*. 2019 Apr;7(4):358-364. doi: 10.1016/S2213-2600(18)30529-0. Epub 2019 Feb 11. PMID: 30765254.

105. Agustí A, Hogg JC. Update on the Pathogenesis of Chronic Obstructive Pulmonary Disease. *N Engl J Med*. 2019 Sep 26;381(13):1248-1256. doi: 10.1056/NEJMra1900475. PMID: 31553836.

106. Weycker D, Hansen GL, Seifer FD. Prevalence and incidence of noncystic fibrosis bronchiectasis among US adults in 2013. *Chron Respir Dis*. 2017;14(4):377-384. doi:10.1177/1479972317709649

107. Martinez CH, Okajima Y, Yen A, Maselli DJ, Nardelli P, Rahaghi F, Young K, Kinney G, Hatt C, Galban C, Washko GR, Han M, Estépar RSJ, Diaz AA. Paired CT Measures of Emphysema and Small Airways Disease and Lung Function and Exercise Capacity in Smokers with Radiographic Bronchiectasis. *Acad Radiol*. 2020 Mar 23:S1076-6332(20)30099-4. doi: 10.1016/j.acra.2020.02.013. Epub ahead of print. PMID: 32217055; PMCID: PMC7508820.

108. Metersky M, Chalmers J. Bronchiectasis insanity: Doing the same thing over and over again and expecting different results? *F1000Research*. 2019;8. doi:10.12688/f1000research.17295.1

109. Barker AF, O'Donnell AE, Flume P, et al. Aztreonam for inhalation solution in patients with non-cystic fibrosis bronchiectasis (AIR-BX1 and AIR-BX2): two randomised double-blind, placebo-controlled phase 3 trials. *Lancet Respir Med*. 2014;2(9):738-749. doi:10.1016/S2213-2600(14)70165-1

110. De Soyza A, Aksamit T, Bandel T-J, et al. RESPIRE 1: a phase III placebo-controlled randomised trial of ciprofloxacin dry powder for inhalation in non-cystic fibrosis bronchiectasis. *Eur Respir J*. 2018;51(1). doi:10.1183/13993003.02052-2017

111. Bilton D, Tino G, Barker AF, et al. Inhaled mannitol for non-cystic fibrosis bronchiectasis: a randomised, controlled trial. *Thorax*. 2014;69(12):1073-1079. doi:10.1136/thoraxjnl-2014-205587

112. Chen AC-H, Pena OM, Nel HJ, et al. Airway cells from protracted bacterial bronchitis and bronchiectasis share similar gene expression profiles. *Pediatr Pulmonol*. 2018;53(5):575-582. doi:10.1002/ppul.23984

113. Chalmers JD, Moffitt KL, Suarez-Cuartin G, et al. Neutrophil Elastase Activity Is Associated with Exacerbations and Lung Function Decline in Bronchiectasis. *Am J Respir Crit Care Med*. 2017;195(10):1384-1393. doi:10.1164/rccm.201605-1027OC

114. Guan W, Gao Y, Xu G, et al. Sputum matrix metalloproteinase-8 and -9 and tissue inhibitor of metalloproteinase-1 in bronchiectasis: Clinical correlates and prognostic implications. *Respirology*. 2015;20(7):1073-1081. doi:https://doi.org/10.1111/resp.12582

115. Sridhar S, Schembri F, Zeskind J, et al. Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium. *BMC Genomics*. 2008;9:259. doi:10.1186/1471-2164-9-259

116. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36(5):411-420. doi:10.1038/nbt.4096

117. Deprez M, Zaragosi LE, Truchi M, Becavin C, Ruiz García S, Arguel MJ, Plaisant M, Magnone V, Lebrigand K, Abelanet S, Brau F, Paquet A, Pe'er D, Marquette CH, Leroy S, Barbry P. A Single-Cell Atlas of the Human Healthy Airways. *Am J Respir Crit Care Med*. 2020 Dec 15;202(12):1636-1645. doi: 10.1164/rccm.201911-2199OC. PMID: 32726565.

118. Aliee H, Theis F. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. bioRxiv. Published online February 23, 2020:2020.02.21.940650. doi:10.1101/2020.02.21.940650

119. Horani A, Ferkol TW. Understanding Primary Ciliary Dyskinesia and Other Ciliopathies. *J Pediatr*. 2020 Nov 23:S0022-3476(20)31452-9. doi: 10.1016/j.jpeds.2020.11.040.

120. Paff T, Kooi IE, Moutaouakil Y, et al. Diagnostic yield of a targeted gene panel in primary ciliary dyskinesia patients. *Hum Mutat*. 2018;39(5):653-665. doi:https://doi.org/10.1002/humu.23403

121. Duclos GE, Teixeira VH, Autissier P, et al. Characterizing smoking-induced transcriptional heterogeneity in the human bronchial epithelium at single-cell resolution. *Sci Adv*. 2019;5(12):eaaw3413. doi:10.1126/sciadv.aaw3413

122. Flume PA, Chalmers JD, Olivier KN. Advances in bronchiectasis: endotyping, genetics, microbiome, and disease heterogeneity. *Lancet*. 2018;392(10150):880-890. doi:10.1016/S0140-6736(18)31767-7

123. Whitwell F. A Study of the Pathology and Pathogenesis of Bronchiectasis *. *Thorax*. 1952;7(3):213-239.

124. King PT, Holdsworth SR, Freezer NJ, Villanueva E, Holmes PW. Characterisation of the onset and presenting clinical features of adult bronchiectasis. *Respir Med*. 2006;100(12):2183-2189. doi:10.1016/j.rmed.2006.03.012

125. Field CE. Bronchiectasis. Third report on a follow-up study of medical and surgical cases from childhood. *Arch Dis Child*. 1969;44(237):551-561.

126. Nabhan AN, Brownfield DG, Harbury PB, Krasnow MA, Desai TJ. Single-cell Wnt signaling niches maintain stemness of alveolar type 2 cells. *Science*. 2018;359(6380):1118-1123. doi:10.1126/science.aam6603

127. Yan KS, Janda CY, Chang J, et al. Non-equivalence of Wnt and R-spondin ligands during Lgr5+ intestinal stem-cell self-renewal. *Nature*. 2017;545(7653):238-242. doi:10.1038/nature22313

128. Zepp JA, Zacharias WJ, Frank DB, et al. Distinct mesenchymal lineages and niches promote epithelial self-renewal and myofibrogenesis in the lung. *Cell*. 2017;170(6):1134-1148.e10. doi:10.1016/j.cell.2017.07.034

129. Barker N, van Es JH, Kuipers J, et al. Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature*. 2007;449(7165):1003-1007. doi:10.1038/nature06196

130. Goddard M. Chapter 3 Histopathology of bronchiectasis. Published 2011. Accessed November 12, 2020. /paper/Chapter-3-Histopathology-of-bronchiectasis-Goddard/f1bbc368c49d13cbc222d22e12dfd57143676768

131. Ruiz García S, Deprez M, Lebrigand K, Cavard A, Paquet A, Arguel MJ, Magnone V, Truchi M, Caballero I, Leroy S, Marquette CH, Marcet B, Barbry P, Zaragosi LE. Novel dynamics of human mucociliary differentiation revealed by single-cell RNA sequencing of nasal epithelial cultures. *Development*. 2019 Oct 23;146(20):dev177428. doi: 10.1242/dev.177428. PMID: 31558434; PMCID: PMC6826037.

132. Guo Z, Chen W, Wang L, Qian L. Clinical and Genetic Spectrum of Children with Primary Ciliary Dyskinesia in China. *J Pediatr*. 2020 Oct;225:157-165.e5. doi: 10.1016/j.jpeds.2020.05.052. Epub 2020 Jun 2. PMID: 32502479.

133. Rouette A, Trofimov A, Haberl D, et al. Expression of immunoproteasome genes is regulated by cell-intrinsic and -extrinsic factors in human cancers. *Sci Rep*. 2016;6(1):34019. doi:10.1038/srep34019

134. Silva JR, Jones JA, Cole PJ, Poulter LW. The immunological component of the cellular inflammatory infiltrate in bronchiectasis. *Thorax*. 1989;44(8):668-673.

135. Lapa e Silva JR, Guerreiro D, Noble B, Poulter LW, Cole PJ. Immunopathology of experimental bronchiectasis. *Am J Respir Cell Mol Biol*. 1989;1(4):297-304. doi:10.1165/ajrcmb/1.4.297

136. Gaga M, Bentley A, Humbert M, et al. Increases in CD4+ T lymphocytes, macrophages, neutrophils and interleukin 8 positive cells in the airways of patients with bronchiectasis. *Thorax*. 1998;53(8):685-691.

137. Gould MK, Tang T, Liu IL, Lee J, Zheng C, Danforth KN, Kosco AE, Di Fiore JL, Suh DE. Recent Trends in the Identification of Incidental Pulmonary Nodules. *Am J Respir Crit Care Med*. 2015 Nov 15;192(10):1208-14. doi: 10.1164/rccm.201505-0990OC. PMID: 26214244.

138. Massion PP, Walker RC. Indeterminate pulmonary nodules: risk for having or for developing lung cancer?. *Cancer Prev Res (Phila)*. 2014;7(12):1173-1178. doi:10.1158/1940-6207.CAPR-14-0364

139. Boiselle PM, Chiles C, Patz E, Tammemägi M, Wood DE. Expert opinion: United States Preventive Services Task Force recommendation on screening for lung cancer. *J Thorac Imaging*. 2014 Jul;29(4):197. doi: 10.1097/RTI.0000000000000094. PMID: 24905632.

140. de Koning HJ, Meza R, Plevritis SK, ten Haaf K, Munshi VN, Jeon J, Erdogan SA, Kong CY, Han SS, van Rosmalen J, Choi SE, Pinsky PF, Berrington de Gonzalez A, Berg CD, Black WC, Tammemägi MC, Hazelton WD, Feuer EJ, McMahon PM. Benefits and harms of computed tomography lung cancer screening strategies: a comparative modeling study for the U.S. Preventive Services Task Force. *Ann Intern Med*. 2014 Mar 4;160(5):311-20. doi: 10.7326/M13-2316. PMID: 24379002; PMCID: PMC4116741.

141. Ito Fukunaga M, Wiener RS, Slatore CG. The 2021 US Preventive Services Task Force Recommendation on Lung Cancer Screening: The More Things Stay the Same…. *JAMA Oncol*. 2021 Mar 9. doi: 10.1001/jamaoncol.2020.8376. Epub ahead of print. PMID: 33687430.

142. McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of cancer in pulmonary nodules detected on first screening CT. *N Engl J Med*. 2013 Sep 5;369(10):910.

143. Winkler Wille MM, van Riel SJ, Saghir Z, Dirksen A, Pedersen JH, Jacobs C, Thomsen LH, Scholten ET, Skovgaard LT, van Ginneken B. Predictive Accuracy of the

PanCan Lung Cancer Risk Prediction Model -External Validation based on CT from the Danish Lung Cancer Screening Trial. *European radiology*. 25 (10): 3093-9.

144. Al-Ameri A, Malhotra P, Thygesen H, Plant PK, Vaidyanathan S, Karthik S, Scarsbrook A, Callister ME. Risk of malignancy in pulmonary nodules: A validation study of four prediction models. *Lung cancer* (Amsterdam, Netherlands). 89 (1): 27-30.

145. Callister ME, Baldwin DR, Akram AR, Barnard S, Cane P, Draffan J, Franks K, Gleeson F, Graham R, Malhotra P, Prokop M, Rodger K, Subesinghe M, Waller D, Woolhouse I. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax*. 70 Suppl 2: ii1-ii54.

146. Kim, H., Kim, H.Y., Goo, J.M. *et al.* External validation and comparison of the Brock model and Lung-RADS for the baseline lung cancer CT screening using data from the Korean Lung Cancer Screening Project. *Eur Radiol* (2020).

147. Winter A, Aberle DR, Hsu W (2019) External validation and recalibration of the Brock model to predict probability of cancer in pulmonary nodules using NLST data. *Thorax* 74:551–563

148. Silvestri GA, Vachani A, Whitney D, Elashoff M, Porta Smith K, Ferguson JS, Parsons E, Mitra N, Brody J, Lenburg ME, Spira A; AEGIS Study Team. A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. *N Engl J Med*. 2015 Jul 16;373(3):243-51. doi: 10.1056/NEJMoa1504601. Epub 2015 May 17. PMID: 25981554; PMCID: PMC4838273.

149. Faiz A, Imkamp K, van der Wiel E, et al. Identifying a nasal gene expression signature associated with hyperinflation and treatment response in severe COPD. *Sci Rep*. 2020;10(1):17415. Published 2020 Oct 15. doi:10.1038/s41598-020-72551-0

150. Zhang, A.W., O'Flanagan, C., Chavez, E.A. *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* **16,** 1007–1015 (2019).

151. Yaghi A, Dolovich MB. Airway Epithelial Cell Cilia and Obstructive Lung Disease. *Cells*. 2016;5(4):40. Published 2016 Nov 11. doi:10.3390/cells5040040

152. Rock JR, Randell SH, Hogan BL. Airway basal stem cells: a perspective on their roles in epithelial homeostasis and remodeling. *Dis Model Mech*. 2010;3(9-10):545–56. doi: 10.1242/dmm.006031.

153. Staudt MR, et al. Airway Basal stem/progenitor cells have diminished capacity to regenerate airway epithelium in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2014;190(8):955–8. doi: 10.1164/rccm.201406-1167LE.

154. Hogan BL, et al. Repair and regeneration of the respiratory system: complexity, plasticity, and mechanisms of lung stem cell function. *Cell Stem Cell*. 2014;15(2):123–38. doi: 10.1016/j.stem.2014.07.012.

155. Smirnova NF, Schamberger AC, Nayakanti S, Hatz R, Behr J, Eickelberg O. Detection and quantification of epithelial progenitor cell populations in human healthy and IPF lungs. *Respir Res*. 2016;17(1):83. Published 2016 Jul 16. doi:10.1186/s12931-016-0404-x

156. Nakajima M, et al. Immunohistochemical and ultrastructural studies of basal cells, Clara cells and bronchiolar cuboidal cells in normal human airways. *Pathol Int.* 1998;48(12):944–53. doi: 10.1111/j.1440-1827.1998.tb03865.x.

157. Cheshier SH, Morrison SJ, Liao X, Weissman IL. In vivo proliferation and cell cycle kinetics of long-term self-renewing hematopoietic stem cells. *Proc Natl Acad Sci U S A*. 1999 Mar 16;96(6):3120-5. doi: 10.1073/pnas.96.6.3120. PMID: 10077647; PMCID: PMC15905.

158. Revinski DR, Zaragosi LE, Boutin C, Ruiz-Garcia S, Deprez M, Thomé V, Rosnet O, Gay AS, Mercey O, Paquet A, Pons N, Ponzio G, Marcet B, Kodjabachian L, Barbry P. CDC20B is required for deuterosome-mediated centriole production in multiciliated cells. *Nat Commun*. 2018 Nov 7;9(1):4668. doi: 10.1038/s41467-018-06768-z. PMID: 30405130; PMCID: PMC6220262.

159. Gibbs S, Fijneman R, Wiegant J, van Kessel AG, van De Putte P, Backendorf C. Molecular characterization and evolution of the SPRR family of keratinocyte differentiation markers encoding small proline-rich proteins. *Genomics*. 1993 Jun;16(3):630-7. doi: 10.1006/geno.1993.1240. PMID: 8325635.

160. Ryckman C, Vandal K, Rouleau P, Talbot M, Tessier PA. Proinflammatory activities of S100: proteins S100A8, S100A9, and S100A8/A9 induce neutrophil chemotaxis and adhesion. *J Immunol*. 2003 Mar 15;170(6):3233-42. doi: 10.4049/jimmunol.170.6.3233. PMID: 12626582.

161. Ryckman C, McColl SR, Vandal K, de Médicis R, Lussier A, Poubelle PE, Tessier PA. Role of S100A8 and S100A9 in neutrophil recruitment in response to monosodium urate monohydrate crystals in the air-pouch model of acute gouty arthritis. *Arthritis Rheum*. 2003 Aug;48(8):2310-20. doi: 10.1002/art.11079. PMID: 12905486.

162. Mills JC, Sansom OJ. Reserve stem cells: Differentiated cells reprogram to fuel repair, metaplasia, and neoplasia in the adult gastrointestinal tract. *Sci Signal*. 2015;8(385):re8. Published 2015 Jul 14. doi:10.1126/scisignal.aaa7540

163. Herfs M, Hubert P, Delvenne P. Epithelial metaplasia: adult stem cell reprogramming and (pre)neoplastic transformation mediated by inflammation? *Trends Mol Med*. 2009 Jun;15(6):245-53. doi: 10.1016/j.molmed.2009.04.002. Epub 2009 May 18. PMID: 19457719.

164. Leube, R.E., Rustad, T.J. Squamous cell metaplasia in the human lung: molecular characteristics of epithelial stratification. *Virchows Archiv B Cell Pathol* **61,** 227–253 (1992).

165. Rigden HM, Alias A, Havelock T, et al. Squamous Metaplasia Is Increased in the Bronchial Epithelium of Smokers with Chronic Obstructive Pulmonary Disease. *PLoS One*. 2016;11(5):e0156009. Published 2016 May 26. doi:10.1371/journal.pone.0156009

166. Giroux V, Rustgi AK. Metaplasia: tissue injury adaptation and a precursor to the dysplasia-cancer sequence. *Nat Rev Cancer*. 2017;17(10):594-604. doi:10.1038/nrc.2017.68

167. Hirano T, Franzen B, Kato H, *et al*. Genesis of squamous cell lung carcinoma. Sequential changes of proliferation, DNA ploidy, and p53 expression. *Am J Pathol* 1994;**144**:296–302.

168. Wistuba II, Behrens C, Milchgrub S, *et al*. Sequential molecular abnormalities are involved in the multistage development of squamous cell lung carcinoma. *Oncogene* 1999;**18**:643–650.

169. Beane JE, Mazzilli SA, Campbell JD, et al. Molecular subtyping reveals immune alterations associated with progression of bronchial premalignant lesions. *Nat Commun*. 2019;10(1):1856. Published 2019 Apr 23. doi:10.1038/s41467-019-09834-2

170. Beane J, Sebastiani P, Whitfield TH, et al. A prediction model for lung cancer diagnosis that integrates genomic and clinical features. *Cancer Prev Res (Phila)*. 2008;1(1):56-64. doi:10.1158/1940-6207.CAPR-08-0011

171. https://www.quora.com/Why-is-XGBoost-among-most-used-machine-learning-method-on-Kaggle#JKatL

172. Zhang Y, Feng T, Wang S, et al. A Novel XGBoost Method to Identify Cancer Tissue-of-Origin Based on Copy Number Variations. *Front Genet*. 2020;11:585029. Published 2020 Nov 20. doi:10.3389/fgene.2020.585029

173. Zhang X, Li T, Wang J, Li J, Chen L, Liu C. Identification of Cancer-Related Long Non-Coding RNAs Using XGBoost With High Accuracy. *Front Genet*. 2019;10:735. Published 2019 Aug 9. doi:10.3389/fgene.2019.00735

174. Huang Z, Hu C, Chi C, Jiang Z, Tong Y, Zhao C. An Artificial Intelligence Model for Predicting 1-Year Survival of Bone Metastases in Non-Small-Cell Lung Cancer Patients Based on XGBoost Algorithm. *Biomed Res Int*. 2020;2020:3462363. Published 2020 Jun 27. doi:10.1155/2020/3462363

175. Esteva, A., Chou, K., Yeung, S. *et al.* Deep learning-enabled medical computer vision. *npj Digit. Med.* **4,** 5 (2021).

176. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542,** 115–118 (2017).

177. Madabhushi A, Udupa JK. Interplay between intensity standardization and inhomogeneity correction in MR image processing. *IEEE Trans Med Imaging*. 2005 May;24(5):561-76. doi: 10.1109/TMI.2004.843256. PMID: 15889544.

178. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, Deschler DG, Varvares MA, Mylvaganam R, Rozenblatt-Rosen O, Rocco JW, Faquin WC, Lin DT, Regev A, Bernstein BE. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell*. 2017 Dec 14;171(7):1611-1624.e24. doi: 10.1016/j.cell.2017.10.044. Epub 2017 Nov 30. PMID: 29198524; PMCID: PMC5878932.

179. Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, Fallahi-Sichani M, Dutton-Regester K, Lin JR, Cohen O, Shah P, Lu D, Genshaft AS, Hughes TK, Ziegler CG, Kazer SW, Gaillard A, Kolb KE, Villani AC, Johannessen CM, Andreev AY, Van Allen EM, Bertagnolli M, Sorger PK, Sullivan RJ, Flaherty KT, Frederick DT, Jané-Valbuena J, Yoon CH, Rozenblatt-Rosen O, Shalek AK, Regev A, Garraway LA. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016 Apr 8;352(6282):189-96. doi: 10.1126/science.aad0501. PMID: 27124452; PMCID: PMC4944528.

CURRICULUM VITAE