

2023

# Leveraging transcriptomic regulation to understand, diagnose and intercept early lung cancer pathogenesis

---

<https://hdl.handle.net/2144/47470>

*"Downloaded from OpenBU. Boston University's institutional repository."*

BOSTON UNIVERSITY  
GRADUATE SCHOOL OF ARTS AND SCIENCES  
AND  
COLLEGE OF ENGINEERING

Dissertation

**LEVERAGING TRANSCRIPTOMIC REGULATION TO UNDERSTAND,  
DIAGNOSE AND INTERCEPT EARLY LUNG CANCER PATHOGENESIS**

by

**BOTING NING**

B.A., Vanderbilt University, 2013  
MPH, Boston University, 2016

Submitted in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

2023



Approved by

First Reader

---

Marc E. Lenburg, Ph.D.  
Professor of Medicine  
Professor of Pathology and Laboratory Medicine

Second Reader

---

Jennifer E. Beane-Ebel, Ph.D.  
Associate Professor of Medicine

## **DEDICATION**

I would like to dedicate this work to my family, who inspired and supported me.

## ACKNOWLEDGMENTS

To my advisors Dr. Avrum Spira, Dr. Marc Lenburg and Dr. Jennifer Beane: Thank you for trusting and accepting me to the lab, and for the opportunities, challenges and guidance that help me grow as a computational biologist.

To my thesis advisory committee members Dr. Juan Fuxman Bass, Dr. Eric Kolaczyk, Dr. Xaralabos Varelas and Dr. Stefano Monti: Thank you for the support and suggestions to my thesis, and the encouragement to pursue better science.

To experimental collaborators Dr. Sarah Mazzilli, Dr. Roxy Pfefferkorn, Dr. Darren Chiu: Thank you for your generous help to my various projects.

To experimental collaborators Dr. Xaralabos Varelas, Dr. Andrew Tilston-Lunel, Dr. Julia Hicks-Berthet and Joseph Kern: Thank you for involving me in your exciting research, and always providing me inspirations from different perspectives.

To Dr. Adam Gower: Thank you for helping with all the heavy-lifting computational tasks.

To the BUSM sequencing and microarray core member Dr. Gang Liu, Dr. Yuriy Alekseyev, Hanqiao and Sherry: Thank you for generating the data that support my thesis work.

To the CBM scientific program manager Liz, Erin, Ipsita and Maria: Thank you for organizing the funding, conferences and all the paperwork that I am always bad at.

To the CBM administration: Brianna, Donna, Jess and Katie: Thank you for your managing CBM activities.

To lab mates of the Spira & Lenburg lab Carter, Chris, Conor, Dylan, Eddy, Grant, Jiarui, Ke, Kelley, Minyi, Regan, Robert, Sean and Xingyi: Thank you for listening to my work and pushing me to do more.

To current and former CBM members Eric, David, Shiyi, Tyler, Yue, Yuqing, Yusuke and Zhe: Thank you for making CBM fun.

To my PhD student cohort: Aaron, Anthony, Jamie, Nick, Rui. Thank you for the unforgettable memories during the journey.

To the Boston University Bioinformatics Program administration team: Thank you for the guidance and support.

To my wife Xiangyi Ren, my son Zhixian Ning and my dog Cash: thank you for being you.

To my parents Cheng Ning and Jie Ding: Thank you for your unconditional love and always believing in me.

**LEVERAGING TRANSCRIPTOMIC REGULATION TO UNDERSTAND,  
DIAGNOSE AND INTERCEPT EARLY LUNG CANCER PATHOGENESIS**

**BOTING NING**

Boston University Graduate School of Arts and Sciences

and College of Engineering, 2023

Major Professor: Marc E. Lenburg, Professor of Medicine, Professor of Pathology and  
Laboratory Medicine

**ABSTRACT**

Lung cancer is the leading cause of cancer death in the U.S., largely due to the lack of treatment options to intercept the progression of early lung cancers and methods to diagnose lung cancer at early stages. Prior studies indicated that the lack of immune surveillance is associated with the progression of bronchial premalignant lesions (PMLs) and the gene alterations in the nasal epithelium can be leveraged for the early detection of lung cancer. Yet, the regulatory mechanism of these gene expression alterations is still less understood. Thus, there are unmet needs to study the gene expression regulation for better disease management of early lung cancer, including further understanding the biology of early lung cancer development, identifying potential interception strategies, and improving the lung cancer diagnosis.

My dissertation addresses these challenges by investigating the transcriptional and post-transcriptional gene expression regulators, including transcription factors and microRNAs (miRNAs), to facilitate the understanding, interception, and diagnosis of early lung cancer. First, I explored the miRNA regulatory landscape to identify miRNA-gene regulatory relationships associated with bronchial PML progression and molecular

subtypes. Using matched gene and microRNA expression profiles from patients with bronchial premalignant lesions, I identified epithelial miR-149-5p to be a key regulator of gene expression contributing to PML progression. By suppressing NLRC5, miR-149-5p inhibits MHC-I gene expression of epithelial cells, promoting early immune depletion and lesion progression. I also developed a novel statistical framework, Differential Regulation Analysis of miRNA (DReAmiR), that characterizes miRNA-mediated gene regulatory network rewiring across multiple groups from transcriptomic profiles, and identified regulatory network differences across PML molecular subtypes. Secondly, I investigated the alterations in the Hippo pathway to identify potential drug targets to intercept the progression of bronchial PMLs. I found that Hippo pathway effectors YAP/TAZ, together with transcription factors TEAD and TP63, cooperatively promote basal cell proliferation and repress signals associated with interferon responses and immune cell communication. Further *in silico* drug screening with external datasets identified small compounds that can reverse the direct regulated gene signature to potentially intercept bronchial PML progression. Lastly, I integrated miRNA and gene expression profiles in the nasal epithelium to distinguish malignant from benign indeterminate pulmonary nodules. I built an ensemble classifier consisting of nasal epithelial miRNA expression features, miRNA-gene top scoring pairs, and clinical features. The performance of the ensemble classifier exceeded that of the classifier built with clinical features alone.

Collectively, my thesis investigated the gene expression regulation mechanisms to facilitate the understanding, interception, and diagnosis of early lung cancer pathogenesis.

## TABLE OF CONTENTS

DEDICATION.....	iv
ACKNOWLEDGMENTS .....	v
ABSTRACT.....	vii
TABLE OF CONTENTS.....	x
LIST OF TABLES.....	xvi
LIST OF FIGURES .....	xvii
LIST OF ABBREVIATIONS.....	xx
CHAPTER 1 INTRODUCTION .....	1
1.1 Lung Cancer.....	1
1.2 Bronchial premalignant lesion .....	2
1.3 miRNA-mediated gene regulation .....	5
1.4 Hippo pathway in cancer .....	8
1.5 Field of injury and field cancerization.....	11
1.6 Non-invasive biomarker for the early detection of lung cancer .....	13
1.7 Dissertation Aims .....	16
CHAPTER 2 THE ROLE OF EPITHELIAL MIR-149 IN IMMUNE MODULATION AND PROGRESSION OF BRONCHIAL PREMALIGNANT LESIONS .....	19
2.1 INTRODUCTION .....	19
2.2 METHODS .....	22
2.2.1 Sample Collection.....	22

2.2.2 miRNA-Seq Library Preparation, Sequence Data Processing, and Sample Filtering .....	24
2.2.3 Construct miRNA-Module Network to Identify miRNAs Associating with Gene Modules.....	26
2.2.4 Identification of miRNAs and Genes Associating with PML Progression Status .....	27
2.2.5 Examination of the Cell-Type Specific Expression of miR-149-5p .....	28
2.2.6 Analysis of NLRC5 ChIP-seq data .....	29
2.2.7 Analysis of FANTOM5 gene expression data .....	29
2.2.8 Cell-type deconvolution analysis .....	29
2.2.9 Data availability .....	30
2.3 RESULTS .....	30
2.3.1 Patient population .....	30
2.3.2 Identification of miRNAs regulating the immune-related module using miRNA-mRNA Network .....	32
2.3.3 hsa-miR-149-5p is Associated with Immune-Related Gene Module and Lesion Progression .....	36
2.3.4 miR-149-5p Regulates MHC Class I Genes Through Suppressing NLRC5 Expression .....	39
2.3.5 has-miR-149-5p Expression is Enriched within Epithelial Cells .....	43
2.3.6 Interaction between miR-149-5p and the NLRC5 can be observed in epithelial cell .....	45

2.3.7 miRNA-gene module network in brushing samples collected from the normal- appearing airways.....	46
DISCUSSION.....	49
CHAPTER 3 DIFFERENTIAL REGULATION ANALYSIS QUANTIFIES MIRNA	
REGULATORY ROLES AND CONTEXT-SPECIFIC TARGETS.....	57
3.1 INTRODUCTION.....	57
3.2 METHODS.....	60
3.2.1 DReAmiR.....	60
3.2.2 Other Functions.....	63
3.2.3 Data Simulation.....	64
3.2.4 Cell-type-specific mmu-miR-155 KO RNA-seq data.....	66
3.2.5 Breast cancer molecular subtype analysis.....	67
3.2.6 Analysis of biopsies and brushings from patients with PMLs across molecular subtypes.....	68
3.2.7 Data and Code Availability.....	68
3.3 RESULTS.....	68
3.3.1 DReAmiR method overview.....	68
3.3.2 Benchmark on simulation data and comparison vs. other methods.....	71
3.3.3 DReAmiR identified mmu-miR-155 cell-type-specific functional pathways .	74
3.3.4 DReAmiR identified BRCA subtype-specific targets from bulk RNA-seq data .....	76

3.3.5 DReAmiR identifies miRNA with context-specific target genes between the Proliferative and Inflammatory PMLs .....	80
3.4 Discussion.....	84
CHAPTER 4 CONVERGENCE OF YAP/TAZ, TEAD AND P63 ACTIVITY DIRECTS	
PREMALIGNANT LUNG GENE EXPRESSION .....	91
4.1 INTRODUCTION .....	91
4.2 METHODS .....	93
4.2.1 Primary human bronchial epithelial cell culture.....	93
4.2.2 Immunoprecipitation and Immunoblotting .....	94
4.2.3 TP63 and isoform expression data analysis in TCGA LUSC and PML data ...	94
4.2.4 HBEC RNA-seq experiments .....	95
4.2.5 ChIP-seq experiments .....	96
4.2.6 Derivation of gene expression signature from RNA-seq siRNA experiments	98
4.2.7 Derivation of direct target genes of TEAD and TP63 from ChIP-seq experiments .....	98
4.2.8 Computational analyses of TEAD-TP63 direct target genes in human patient data .....	99
4.2.9 Single-cell RNA-seq data analysis.....	101
4.2.10 CP1 and chemical compound analysis.....	102
4.2.11 Datasets used and Code Availability .....	102
4.3 RESULTS .....	102
4.3.1 TP63 and TEAD expression is elevated in PML histological progression ..	102

4.3.2 YAP, TEAD and TP63 bind to the same genomic sites in basal bronchial epithelial cells.....	105
4.3.3 TP63 and TEAD co-regulate gene expression in the basal bronchial epithelial cells.....	108
4.3.4 The TP63/TEAD repressed gene program is associated with early immune evasion in the bronchial premalignant lesions .....	111
4.3.5 YAP/TAZ-TEAD-p63 down-regulate Major Histocompatibility Complex factors transactivator CIITA in bronchial epithelial cells .....	115
4.3.6 Palmitoylation inhibitor blocks TEAD DNA-binding and may reverse the TEAD-TP63 directly regulated gene program .....	119
4.4 DISCUSSION.....	121
CHAPTER 5 MICRORNA EXPRESSION DIFFERENCES IN NASAL EPITHELIUM FOR IDENTIFYING MALIGNANT INDETERMINATE PULMONARY NODULES .....	
5.1 INTRODUCTION .....	128
5.2 METHODS .....	130
5.2.1 Study enrollment and sample collection .....	130
5.2.2 miRNA sequencing and processing .....	130
5.2.3 Derivation of miRNA signatures and TSPs associated with cancer status ....	131
5.2.4 Ensemble Learning Pipeline .....	133
5.3 RESULTS .....	134
5.3.1 Study Population .....	134

5.3.2 miRNA expression signature and miRNA-gene TSPs associated with IPN	
status .....	134
5.3.3 Development of an integrated ensemble learning model for the IPN diagnosis	
.....	141
5.4 DISCUSSION.....	144
CHAPTER 6 GENERAL CONCLUSIONS AND FUTURE DIRECTIONS.....	148
APPENDIX A SUPPLEMENTARY FIGURES .....	157
APPENDIX B SUPPLEMENTARY TABLES.....	171
BIBLIOGRAPHY.....	182
CURRICULUM VITAE.....	213

## LIST OF TABLES

Table 2.1. Biopsy sample clinical annotation across four PML molecular subtypes. ....	31
Table 2.2. Brushing sample clinical annotation across four PML molecular subtypes. Statistical tests for categorical clinical variables (dysplasia grade, smoking status, progression status, and batch) were conducted using Chi-square tests. Statistical tests for continuous variables (TIN) were compared using two-sided Student’s t-tests. Percentages are reported for categorical variables and mean/standard deviations are reported for the continuous variable. ....	47
Table 5.1. DEAMP I nasal miRNA sample clinical annotation by cancer status.....	135
Table B.1. Connections in the miRNA-Gene Module Network.....	172
Table B.2. Test statistics for miRNA connected to the immune-related modules in biopsy samples.....	173
Table B.3. Differential Expression for miRNA connected to the immune-related module. .....	174
Table B.4. Differential Expression for miR-149-5p predicted target genes in Beane <i>et al.</i> Validation cohort and in Merrick <i>et al.</i> .....	175
Table B.5. Differential expression for NLRC5 targets in Beane <i>et al.</i> Validation cohort and in Merrick <i>et al.</i> .....	176
Table B.6. Comparison of model performance metrics in simulated datasets.....	178
Table B.7. Functional pathway enrichment results of genes in the basal or luminal A cluster.....	179
Table B.8. Differential expression results for the MHC II genes by progression status in PML datasets.....	180
Table B.9. Clinical annotation for DEAMP I nasal epithelial brushings with both miRNA and gene expression profiles between the discovery and validation cohort. ....	181

## LIST OF FIGURES

Figure 2.1. Analysis design diagram.....	34
Figure 2.2. miRNA-Gene module network captures miRNA regulatory roles. ....	35
Figure 2.3. miR-149-5p targets genes in the immune-related gene module and is upregulated in the progressive PML samples within the Proliferative subtype.....	38
Figure 2.4, The predicted target genes of miR-149-5p were associated with PML progression within the Proliferative subtype. ....	40
Figure 2.5. NLRC5 regulates MHC Class I gene expression associated with PML progression. ....	42
Figure 2.6. miR-149-5p is highly expressed in and regulates NLRC5 within epithelial cells. ....	44
Figure 2.7. miRNA-gene module network consensus analysis between the biopsy and brushing sample. ....	48
Figure 2.8. Summary diagram. ....	51
Figure 3.1. DReAmiR method overview. ....	69
Figure 3.2. DReAmiR performance in simulated data. ....	72
Figure 3.3. DReAmiR identifies miR-155 target genes involved in functional pathways with cell-type specificity.....	75
Figure 3.4. The miR-23b prioritized targets from TCGA BRCA are uniquely altered by perturbation in the cell line of the same subtype. ....	79
Figure 3.5. DReAmiR identified miRNAs with differential regulatory roles between the proliferative and inflammatory PMLs. ....	81
Figure 3.6. miRNA differential regulatory patterns across molecular subtypes were different between endobronchial biopsy and mainstem airway brushing samples...	83
Figure 4.1. TP63 is associated with human LUSC carcinogenesis and early lung cancer progression.....	104
Figure 4.2. TP63 interacts and co-binds to chromatin with YAP/TEAD in HBECs.....	107
Figure 4.3. YAP/TEAD/TP63 together regulate target genes associated with carcinogenesis pathways in HBECs.....	110

Figure 4.4. TEAD-TP63 direct regulated genes are associated with human bronchial PML progressive pathology and early immune evasion. ....	114
Figure 4.5. CIITA associates with bronchial PML progressive pathology and tracks with suppressing MHC Class II gene expression and the presence of Th1 T cells. ....	117
Figure 4.6. Blocking the DNA-binding ability of TEADs via CP1 may reverse the TEAD-TP63 directly regulated gene signature.....	120
Figure 4.7. Summary diagram. ....	123
Figure 5.1. Smoking-associated miRNAs in the nasal epithelium of DECAMP I samples were enriched within the bronchial smoking-associated miRNA signature. ....	136
Figure 5.2. Predicted target genes of the cancer-associated miRNAs in the DECAMP I nasal samples were dysregulated in previously derived nasal cancer-associated gene signatures. ....	138
Figure 5.3. miRNA-target gene top-scoring pairs reflected cancer-associated gene alterations.....	140
Figure 5.4. Diagram for the ensemble learning model training pipeline and cross-validation schema.....	142
Figure 5.5. The performance of ensemble classifier integrating clinical and miRNA features in differentiating malignant and benign IPNs. ....	143
Figure 6. The schematic diagram for overall conclusions and future directions. ....	151
Figure A.1. GSVA scores of miRNAs associated with a gene module were negatively correlated with the gene module metagene scores.....	157
Figure A.2. Genes of the immune-related gene module were enriched among the genes negatively correlated with miR-149-5p. ....	158
Figure A.3. The expression level of miR-149-5p was significantly negatively correlated with that of NLRC5 across datasets.....	159
Figure A.4. NLRC5 regulates the expression of MHC Class I genes.....	160
Figure A.5. Correlation between miR-149-5p and NLRC5 expression level within the samples from the FANTOM5 project. ....	161
Figure A.6. DReAmiR performance in simulated data.....	162

Figure A.7. DReAmiR performance in simulated data with different sample sizes across groups.....	163
Figure A.8. The expression level of miR-155 prioritized target genes across four mice immune cell-types of the control group.....	164
Figure A.9. TP63 isoform expression levels in TCGA-LUSC and in bronchial PML biopsy data. ....	165
Figure A.10. ChIP-seq analysis of YAP/TEAD/TP63 chromatin binding profiles.....	166
Figure A.11. Transcriptomic analysis of TEAD-TP63 direct regulated target genes.....	167
Figure A.12. Transcriptomic analysis of TEAD-TP63 direct regulated target genes in human bronchial PML data and lung scRNA-seq data.....	168
Figure A.13. Analysis CIITA in human bronchial PML data and lung scRNA-seq data. ....	170

## LIST OF ABBREVIATIONS

AD	Anderson-Darling
Adaboost	Adaptive Boosting
AEGIS	Airway Epithelial Gene Expression in the Diagnosis of Lung Cancer
BRCA	Breast Invasive Carcinoma
ChIP-seq	Chromatin immunoprecipitation sequencing
CIS	Carcinoma <i>in situ</i>
CLASH	Cross-linking ligation and sequencing of hybrids
CLIP-seq	Crosslinking immunoprecipitation sequencing
CMAP	Connectivity mapping
COPD	Chronic obstructive pulmonary disease
CPM	Counts per million
CT	Computed Tomography
DCIS	Ductal carcinoma <i>in situ</i>
DECAMP	Detection of Early Lung Cancer Among Military Personnel
DReAmiR	Differential regulatory analysis of microRNA
FANTOM5	Functional Annotation of the Mouse/Mammalian Genome
FDR	False discovery rate
FEV	Forced expiratory volume
FPR	False positive rate
GBM	Gradient boosting machine

GRL	Graph representation learning
GSEA	Gene set enrichment analysis
GSVA	Gene set variation analysis
HBEC	Primary Human Bronchial Epithelial Cell
HNSCC	Head and Neck squamous cell carcinoma
IPF	Idiopathic pulmonary fibrosis
IPN	Indeterminate pulmonary nodule
IQR	Interquartile range
KS	Kolmogorov–Smirnov
LDCT	Low-dose Computed Tomography
LINE	Large information network embedding
logFC	log Fold-change
LR	Logistic regression
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MACS2	Model-based Analysis of ChIP-Seq 2
Max-dES	Maximum difference in enrichment scores
MHC	Major histocompatibility complex
miRISC	miRNA-induced silencing complex
miRNA	microRNA
MSigDB	The Molecular Signatures Database
NB	Naïve Bayes

NCCN	National Comprehensive Cancer Network
NLR	NOD-like receptor
NLST	National Lung Screening Trial
NSCLC	Non-small cell lung cancer
PCA	Principal component analysis
PML	Premalignant lesion
Pre-miRNA	Precursor microRNA
Pri-miRNA	Primary microRNA
RF	Random forest
RIN	RNA integrity number
RPM	Reads per million
SCLC	Small cell lung cancer
SMOTE	Synthetic Minority Oversampling Technique
SpQN	Spatial quantile normalization
SVM	Supported vector machine
TCGA	The Cancer Genome Atlas
TF	Transcription factor
TIN	Transcript integrity number
TMM	Trimmed mean of M-values
TPM	Transcripts per million
TPR	True positive rate
tSNE	t-distributed stochastic neighbor embedding

TSP	Top-scoring pair
UMAP	Uniform Manifold Approximation and Projection
UTR	Untranslated region
WGCNA	Weighted gene correlation network analysis

## CHAPTER 1 INTRODUCTION

### *1.1 Lung Cancer*

Lung cancer is one of the most frequently diagnosed cancer type and is the leading cause of cancer mortality. Based on the latest statistics, there are over 200,000 lung cancer cases in the United States per year, leading to more than 25% of all cancer deaths<sup>1</sup>. Meanwhile, the overall lung cancer 5-year survival rate is 21% for all stages combined, which is much lower than 98% for prostate cancer or 90% for breast cancer. The high mortality of lung cancer is associated with its complex molecular profiles and the lack of early detection and management strategies.

Various factors could contribute to the development of lung cancer, depending on the specific subtype. There are two major subtypes of lung cancers based on its biology and histology: small-cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), where the latter accounts for over 80% of all lung cancer cases worldwide<sup>2</sup>. NSCLC can be further classified into adenocarcinoma (LUAD), squamous cell carcinoma (LUSC) and large cell carcinoma<sup>3</sup>. Cigarette smoking, including second-hand smoking, is the leading risk factor for developing lung cancer, particularly for LUSC which arises from the airway epithelium<sup>4</sup>. Environmental and occupational exposures to carcinogens, including radon, asbestos, arsenic compounds and air pollution, can also increase the risk of lung cancers<sup>5-8</sup>. Furthermore, germline and somatic genetic alterations can increase risk of lung cancer. Gene mutations in EGFR are found in about 15% of LUAD cases, causing uncontrolled cell growth and proliferations<sup>9</sup>. G12 mutations of KRAS are found in LUAD patients who do not have smoking history<sup>10</sup>. ALK and BRAF mutations are

also identified in less than 10% of NSCLC patients<sup>11,12</sup>. Meanwhile, less is known about the driving mutation for SCLC beyond TP53 or RB1<sup>13</sup>. These factors together reflect the complicated mechanism behind the carcinogenesis of lung cancer and the difficulty to develop preventative, diagnostic and therapeutic strategies for better disease management.

Although the National Lung Screening Trial demonstrated low-dose CT screening among patients with high-risk for lung cancer can reduce the mortality by 20%<sup>14</sup>, the increase in the 5-year survival rate of all lung cancer cases over the past decades is low comparing to the other cancer types<sup>15</sup>. The high mortality rate of lung cancer can be largely attributed to the inability to detect lung cancer at early stages when it is still curable. In fact, only less than 20% of lung cancer cases were diagnosed early on when the tumor was still localized to its primary site and the 5-year survival rate is around 60%, compared to 6% 5-year survival rate when diagnosed late at distal sites<sup>1</sup>.

Thus, there is a critically unmet need to develop tools for the early detection of lung cancers, to identify the cellular and molecular alterations associated with early lung cancer development, and to explore potential interception strategies with the ultimate goal of reducing overall lung cancer cases and mortality.

### ***1.2 Bronchial premalignant lesion***

Commonly found in smokers and chronic obstructive pulmonary disease (COPD) patients, bronchial premalignant lesions (PMLs) are the presumed precursors of LUSC<sup>16,17</sup>. They are histological abnormalities observed in the bronchial airways,

characterized by various degree of morphological change and proliferation of basement membrane<sup>18</sup>. Depending on cell morphology and histologic appearances, the stepwise histological progression of PMLs can be graded as normal epithelium, hyperplasia, squamous, metaplasia, dysplasia (mild, moderate, and severe), carcinoma in situ (CIS), and all the way to invasive carcinoma<sup>19</sup>. Hyperplasia and metaplasia are considered to be normal physiological response to tissue damage, while dysplasia is thought to be the beginning of the oncogenic process<sup>20</sup>. The presence of is high-grade dysplasia is significantly associated with higher risk for lung cancer both at the site of the lesion and other regions in the lung<sup>21</sup>. Meanwhile, the sequential progression of PML is accompanied by various genetic abnormalities. Loss of heterozygosity on chromosome 3p and 9p in the normal-appearing airway epithelium were documented as the early driver during the initial of PML progression<sup>22</sup>. Also, increased expression levels of p53, Bcl-2 and cyclin-D1, and miR-241/301 have been observed during the transition to higher-grade dysplasia and CIS<sup>23-25</sup>. The well documented histologic and genetic features and the stepwise progression nature of bronchial PML make it a good model for early lung cancer study.

Intuitively, it may be easier to prevent LUSC by intercepting PML progression before it becomes invasive than to cure high-grade LUSC. While many therapeutic options exist for managing LUSC<sup>26</sup>, including targeted therapies and immune checkpoint blockade, few treatments are available for intercepting PMLs due to the lack of understanding in molecular drivers of PML progression. Most lung cancer prevention trials focused on removing residual chemicals from cigarette smoking or inflammation in the airway

epithelium<sup>27,28</sup>. Also, there are on-going clinical trials that evaluate the efficacy of phytochemical antioxidants or electrocautery ablation in preventing the progression of high-grade bronchial PMLs<sup>29,30</sup>. However, limited success has been made potentially due to the lack of knowledge about the molecular alterations that drive the PML progression. One intriguing observation is that not all PMLs will progress to invasive cancer and most actually regress without any intervention<sup>31,32</sup>. Understanding the dynamic of bronchial PML histological progression and developing biomarkers to identify those PML that will progress are crucial for lung cancer prevention. Studies comparing molecular profiles between precancerous bronchial lesions with various outcomes have indicated that innate and adaptive immune alterations, chromosomal instability and copy number variation in epithelial cells, and stroma changes are associated with risk for progression to invasive LUSC<sup>33-37</sup>. Also, transcriptomic alterations in the normal-appearing airway can be leveraged to identify PML that may progress<sup>38</sup>. Notably, work by Beane *et al.*<sup>39</sup> utilized longitudinally collected endobronchial biopsy and mainstem bronchial brushing samples from patients with PMLs to investigate the molecular mechanism that contribute to the progressive phenotype of PMLs. Based on the pattern of co-expressed gene modules, the PMLs can be classified into four transcriptionally distinct molecular subtypes with various activities of epithelial and immune pathways. The proliferative subtype is enriched PML with higher grade dysplasia and had high basal cell signal and proliferative gene expression. Particularly, a gene module consisting of genes in the interferon response pathway and antigen processing/presentation pathway were strongly suppressed in proliferative PMLs that progress to higher grades, comparing to the ones that regress to

normal, suggesting modulating of immune microenvironment may help intercept early lung cancers. Collectively, these findings may ultimately pave the way for better disease management for patients with PMLs, or other patients at risk for lung cancer.

Yet, the gene regulatory mechanism related to early immune evasion is not known. Thus, there is need to probe transcriptional and post-transcriptional gene expression regulation that contributes to the bronchial PML histological progression.

### ***1.3 miRNA-mediated gene regulation***

microRNAs (miRNAs) are a class of small non-coding RNA that mediate post-transcriptional gene regulation<sup>40</sup>. In mammals, the miRNA genes, located either in the introns of protein-coding genes or in separated miRNA loci, are transcribed mostly by Pol II, and less frequently by Pol III<sup>41,42</sup>. The transcripts, called primary miRNAs (pri-miRNAs), are generally longer than 1kb and contain a stem-loop structure, 5' cap and 3' polyadenylated tail. The pri-miRNAs are recognized and cleaved by Microprocessor, a complex consisting of double-stranded RNase DROSHA and double-stranded RNA-binding protein DGCR8, within the nucleus<sup>43</sup>. The products are around 60-70-nucleotide with hairpin structure called precursor miRNAs (pre-miRNAs)<sup>44,45</sup>. Then, the pre-miRNAs are exported to the cytoplasm by XPO5 and are further processed by RNase III DICER I to cleave the stem-loop, leaving a 22-nucleotide long mature miRNA duplex<sup>46-48</sup>. Finally, the miRNA-induced silencing complex (miRISC) is formed by AGO proteins binding to a single strand of the mature miRNA that induces RNA silencing by recognizing the 3' untranslated region of mRNA transcripts via the miRNA seeding

region (the 2-7 nucleotide on the 5' end of the mature miRNAs)<sup>49-51</sup>. Among the many different mechanisms observed, including transcript deadenylation and translational repression, the majority of RNA silencing is through mRNA transcript destabilization<sup>52</sup>. The ability to predict target based on their sequence and unidirectional regulation (expression suppression) makes statistical network model an ideal approach to study miRNA-mediated gene regulation. There are several unique features to the miRNA regulatory network, comparing to other biological networks (e.g. protein-protein network). First, a single miRNA can potentially regulate multiple, even up to a hundreds of target genes<sup>53</sup>. The target genes are often within the similar functional pathways, which allows the miRNA to have stronger regulation in short timeframe and more specific control over the network<sup>54,55</sup>. Second, miRNAs locus often localized to the genome regions as polycistronic clusters, are transcribed together and tend to regulate genes in the same functional pathways<sup>56-58</sup>. The co-expressed miRNAs allow cellular signal to be amplified more rapidly and more robustly to gene expression changes. Furthermore, given the smaller number of miRNAs comparing to the number of genes in the genome, and the many-to-many regulatory relationship, miRNAs tend to be the hubs in the network, giving them stronger influence in the network behavior and potential translational impact. Typically, the miRNA regulatory network is modeled as a bipartite weighted directional network. A node can be either the miRNA or the target gene. The edge points from the miRNA node to the gene node, representing direct regulation with the weight as the association strength. Many computational tools have been developed to characterize the miRNA-mediated gene regulation networks from sequencing data, based

on miRNA-gene expression correlation, target gene enrichment in functional pathways or differentially expression gene sets, or experimental evidence alone<sup>59-63</sup>.

An important feature of biological networks, including transcriptional regulatory networks, is that they are not static. The context-specific transcriptional regulatory network has been described for transcription factors, where the chromosomal binding pattern of transcription factor or the co-binding factors changes between conditions<sup>64-66</sup>. Similarly, the miRNA-mediated regulatory network is dynamic between different biological and cellular conditions<sup>67-70</sup>, although the underlying mechanism is not yet fully understood. The reprogramming process provides further flexibility in the miRNA-mediated gene regulation, but no computational tools have been built to specifically address this question.

miRNA-mediated gene regulation is involved in all biological processes and the dysregulation of miRNA can lead to diseases, including cancer<sup>71,72</sup>. Based on the functions of the target genes, a miRNA can either be oncomiR that promotes tumorigenesis, or tumor suppressor miRNA that inhibits cancer development. A famous example of an oncomiR is the miR-17-92 cluster, consisting of miR-17, miR-18a, miR-19a/b, miR-20a and miR-92, which promotes various cancer types by suppressing BIM, TSP1 and CTGF<sup>73,74</sup>. The down-regulation of tumor suppressor miRNA miR-34 has been widely reported in lung, breast, prostate and colorectal cancer<sup>75-78</sup>, which induces the epithelial-to-mesenchymal transition, cell proliferation and tumor metastasis.

Functionally interacting with p53, miR-34 may also suppress the expression of PDL1 and facilitate the immune evasion of acute myeloid leukemia<sup>79</sup> and NSCLC<sup>80</sup>. Defects in the

miRNA processing machinery, including various DROSHA/DGCR8 single-point mutation<sup>81</sup>, XPO5 inactivating mutations<sup>82</sup>, truncation of the C-terminal catalytic domain of DICER<sup>83</sup>, or simply the expression alterations of these genes, may also lead to cancer development.

In addition, our lab previously has shown that airway epithelial miRNAs are responsible for gene expression alterations in response to cigarette smoking and lung carcinogenesis<sup>84,85</sup>. Here, I hypothesize that exploring the miRNA regulatory landscape and miRNA-mediated regulatory network rewiring, using a novel network approach, can provide mechanistic understating of the altered gene expression regulation associated with bronchial PML molecular subtypes and phenotypes.

#### ***1.4 Hippo pathway in cancer***

First discovered in the *Drosophila* as a potential tumor suppressive pathway<sup>86</sup>, the Hippo pathway is an evolutionary conserved signal cascade that control cell proliferation and tissue size<sup>87</sup>. Various upstream signal may activate the Hippo pathway, including cell-cell contact altering ECM stiffness<sup>88</sup>, the G-protein coupled receptor signal regulated by estrogen<sup>89,90</sup>, energy stress or hypoxia signal<sup>91,92</sup> and cell polarity<sup>93</sup>. The core components of the Hippo pathway are the MST1/2 and LATS1/2 kinases. Upon receiving upstream signal, MST1/2 are phosphorylated by the TAO kinases and are activated<sup>94,95</sup>. Then, MST1/2 undergo dimerization<sup>96</sup>, which leads to the recruitment and phosphorylation of LATS1/2<sup>97,98</sup>. Alternatively, LATS1/2 can be phosphorylated and activated directly by MAP4Ks without MST1/2<sup>99</sup>. The activated LATS1/2 can then undergo

autophosphorylation to phosphorylate and inactivate YAP and TAZ<sup>100</sup>. This leads to YAP and TAZ binding to 14-3-3, which sequester YAP and TAZ in the cytoplasm for degradation<sup>100</sup>. The cytoplasmic and nuclear localization of YAP and TAZ determine the transcriptional output of the Hippo pathway. YAP and TAZ are transcriptional coactivators, meaning they do not have DNA-binding ability to regulate gene expression. Instead, when the Hippo pathway is inactivated, YAP and TAZ enter the nucleus and bind to transcription factors, the most well-known is TEAD1-4<sup>101</sup>, to mediate the transcriptional regulation related to cell survival and proliferation<sup>102-104</sup>. Interestingly, YAP and TAZ can also bind to other transcription factors to regulate gene expression, including p73<sup>105</sup>, SMADS<sup>106</sup>, ErbBs<sup>107</sup>, EGR1<sup>108</sup> and RUNX<sup>109</sup>.

Dysregulation of the Hippo pathway and the association with various aspects of cancer development, including cancer initiation, invasion, tumor metastasis and drug resistance, have been well documented. The regulated target genes of Hippo pathway has been thoroughly studied and the elevated YAP/TAZ target gene activity have been suggested in different cancer types<sup>110</sup>, highlighting its importance for carcinogenesis of particular the squamous carcinoma. YAP and TAZ can induce the gene expression involved in DNA synthesis and cell cycle progression to promote abnormal cellular growth<sup>111,112</sup>. Gene programs activated by YAP and TAZ also maintain the cancer stem cell identity or reprogram differentiated cells to a stem cell-like state<sup>113-115</sup>. Cancer cell also increase resistance to cell death by YAP/TAZ upregulating Bcl2 genes<sup>116</sup>. Lower levels of YAP/TAZ also weaken the tumor metastasis ability in different cancer types<sup>113,117,118</sup>. In lung cancer, the YAP/TAZ expression level and nuclear localization are associated with

higher histological grade and worse patient outcome<sup>119-121</sup>. The acquired resistance to EGFR inhibitor are also correlated with an elevated YAP expression in NSCLC patients<sup>122</sup>. Furthermore, YAP nuclear localization driven by Crb3 knockdown and loss of cell polarity in the mice model results in cell morphological changes and transcriptional signature that highly resemble human precancerous airway<sup>123,124</sup>. Given the importance of the Hippo pathway in the carcinogenesis, there has been extensive effort to design therapeutic that targets the Hippo pathway. The main focus of the field is to identify inhibitors that block the transcriptional regulation by TEAD. Verteporfin is among the first small compound identified. Disrupting the binding between YAP and TEAD, verteporfin showed good tumor suppressing ability in mice model<sup>125</sup> and in human colon and endometrial cell lines<sup>126,127</sup>. Small molecule inhibitors MYF-01-37<sup>128</sup> and C19<sup>129</sup> were designed for similar purpose and exhibited efficacy *in vivo*. However, the target specificity was not ideal. An alternative strategy to target Hippo pathway or YAP/TAZ transcriptional regulation activity was to disrupt the post-translation autopalmitoylation of TEADs and ablate the DNA-binding activity, which theoretically would not affect its cellular localization or binding to YAP or TAZ<sup>130,131</sup>. MGH-CP1 was identified by small-molecule library screen followed by *in vitro* autopalmitoylation assay to inhibit the TEAD2/4 autopalmitoylation<sup>132</sup>. Additionally, its ability to block downstream transcription was successfully demonstrated in human embryonic kidney, fibroblast and breast cancer cell lines. However, whether the Hippo pathway activity is associated with bronchial PML progressions and whether it can serve a potential therapeutic target to intercept early lung cancer are less understood.

### ***1.5 Field of injury and field cancerization***

*Paragraphs of chapter 1.5-1.6 were adapted from sections written for the review:*

Paez R, Kammer MK, Tanner NT, Heidman BE, Peikert T, Babach M, Iams WT, Ning B, Lenburg ME, Mallow C, Yarmus L, Fong KM, Deppen S, Grogan EL and Maldonado F. State-of-the-Art review on biomarkers for the early detection of lung cancer. Submitted to American Journal of Respiratory and Critical Care Medicine.

Initially proposed by Slaughter *et al.*<sup>133</sup>, the term “field cancerization” described the molecular abnormalities related to tumorigenesis can be detected in the normal-appearing tissue adjacent to the oral squamous premalignant lesions. Similar observations have been made in other cancer types since then, including various genetic and molecular alterations. Genomic instability and epigenetic alterations have been characterized in the histologically normal epithelium adjacent to breast tumor and in pre-invasive ductal carcinoma *in situ* (DCIS)<sup>134–136</sup>. Abnormal gene expression associated with head and neck squamous cell carcinoma (HNSCC) carcinogenesis process were found in the normal mucosa among HNSCC patients<sup>137,138</sup>. In the adjacent normal prostate tissues, oncogenic Somatic mitochondrial DNA mutations and cancer-associated gene signature can also be detected<sup>139–141</sup>. The translational utility of the field of cancerization theory opened new door for the development of minimally invasive cancer risk evaluation tools, including using DNA methylation panels measured in normal mucosa to quantify risk of esophageal squamous cell carcinoma<sup>142</sup>, and leveraging the spectroscopic microscopy on normal rectal colonocytes for colorectal cancer risk stratification<sup>143</sup>.

In lung cancer, field of injury and field cancerization, describe molecular alterations in the normal-appearing respiratory tract associated with lung cancer that may reflect exposure to carcinogens and/or the process of carcinogenesis<sup>144</sup>. Injury due to exposure to

carcinogens such as cigarette smoking, induces physiological responses and genetic alterations in the entire respiratory tract<sup>145,146</sup>. In contrast, field cancerization describes cancer-associated alterations in the tumor-adjacent normal tissues, including gene mutations of KRAS, p53, and EGFR detected in histologically normal airway epithelium<sup>147-149</sup>. A study from Kadara *et al.* also revealed a lung cancer-associated gene signature in the adjacent normal bronchial epithelium, including the upregulation of LPTM4B whose expression levels were dependent on the distance from tumors<sup>150</sup>. The field of cancerization effect has also been observed for bronchial PML during the early lung cancer phase. Gene expression alterations in the mainstem airway, collected by bronchial brushings, were showed to be involved in the LUSC carcinogenesis and were used to developed a biomarker to predict the presence of PMLs<sup>38</sup>. Demonstrated in the same study, the changes in the biomarker scores between sequential brushing samples collected longitudinally was able to predict whether the worst PML histology will regress in the future. The transcriptionally distinct bronchial PML molecular subtypes were also identified in the normal appearing uninvolved bronchial airways with high specificity (91%) but low sensitivity (31-38%), reflecting a heterogeneous impact of bronchial PML on the entire airway where a subset of PML may impose widespread damage while the others are more localized<sup>39</sup>.

In my thesis, I will mainly utilize two data type collected from the airway fields. The bronchial brushing samples collected from patients with bronchial PMLs will be leveraged to examined the conservation of miRNA-mediated gene regulatory network between the PML site and the field. The nasal epithelial miRNA expression will be used

to construct an integrated biomarker for the early detection of lung cancer.

### ***1.6 Non-invasive biomarker for the early detection of lung cancer***

Pulmonary nodules are found in about 24-30% of patients undergoing CT incidentally or as part of screening<sup>151</sup>. Despite an overall reduction in lung cancer mortality due to screening, the results of NLST showed high false-positive rate, and the high cost and the stringent enrollment criteria prohibits it from adapting to larger population<sup>152</sup>. Thus, there is an unmet need to develop a low-cost and minimally invasive biomarker for the early detection of lung cancer and for better cancer risk stratification that can be applied to broader population of low lung cancer risk. The ability to detect cancer-associated molecular differences in relatively accessible normal-appearing airway epithelium has served as the foundation for building several non-invasive biomarkers that aid in the diagnosis of lung cancer and might eventually serve as tool for identifying patients at elevated risk of lung cancer.

Bronchoscopic sampling of the suspicious lesion with molecular analysis of brushings from the mainstem bronchus is an attractive alternative to tissue collection via bronchoscopy which have low sensitivity, especially in smaller, or more peripheral nodules. Blomquist *et al.* reported the levels of 14 genes of the antioxidant and DNA repair pathways, measured via RT-PCR, in the normal bronchial airway epithelium were associated with lung cancer status, and resulted in a biomarker currently being tested in the Lung Cancer Risk Test trial (NCT01130285)<sup>153,154</sup>. Similarly, using microarray-based transcriptomic profiling, Spira *et al.* developed and validated an 80-gene signature for

differentiating ever smokers with or without lung cancer using endobronchial brushings from normal-appearing main-stem airway<sup>155</sup>. This signature was later refined into a 23-gene lung cancer biomarker for suspected lung cancer patients undergoing bronchoscopy<sup>156</sup>. This biomarker, marketed as Percepta™ by Veracyte, Inc, was extensively validated in independent cohorts consisting of more than 600 samples<sup>157</sup>. The high sensitivity either alone (89%) or in combination with bronchoscopy (97%) and high NPV of this biomarker allows physicians to choose surveillance by imaging for intermediate-risk patients with an inconclusive bronchoscopy and negative biomarker results.

While biomarkers based on bronchial airway gene expression have shown great clinical utility, the collection of the required biospecimen via bronchoscopy potentially limits the intended use population due to the potential complications of bronchoscopy, and the need to balance benefit against risk and costs<sup>158</sup>. As an alternative, researchers have explored lung cancer-associated gene expression alterations in the nasal epithelium and aimed to develop a potentially less invasive diagnostic method. Studies demonstrated that bronchial and nasal epithelium share similar transcriptomic alteration associated with smoking status<sup>159</sup>, and lung diseases such as COPD<sup>160</sup> and idiopathic pulmonary fibrosis (IPF)<sup>161</sup>, suggesting the involvement of nasal epithelium in the field of injury and field cancerization. Using nasal brushing samples and microarray expression profiling from participants in the AEGIS cohorts, Perez-Rogers *et al.* developed a clinical-genomic classifier with 30 genes measured in the nasal epithelium that can identify lung cancer among ever-smokers with high sensitivity of 91%<sup>162</sup>. During the ASCO 2021, Veracyte,

Inc. announced the latest genomic classifier for lung cancer risk stratification, profiled from nasal swab with total RNA sequencing, with 95% sensitivity for classification as low risk and 90% specificity for classification as the high risk<sup>163</sup>. Similar high performance was observed among patients who met or did not meet the screening criteria<sup>164</sup>. These studies highlight clinical utility of nasal epithelium gene expression for early lung cancer detection, particularly among patients whose pretest risk are low for bronchoscopy.

Meanwhile, there is evidence that an integrated biomarker combining information from different data modalities may aid the early detection or diagnosis of lung cancer.

Combining gene expression profile in the normal appearing airway epithelium collected via bronchial brushings and clinical risk factors, including age, mass size, and lymphadenopathy, the clinicogenomic model achieved significantly improved performance for lung cancer classification than the biomarker with clinical risk factors alone<sup>165</sup>. Pavel *et al.* demonstrated that adding miR-146a-5p expression to the bronchial gene biomarker can significantly improve the performance of the lung cancer diagnosis within the samples of AEGIS cohorts<sup>166</sup>. Furthermore, the expression of 3 miRNAs and the promoter methylation status of 3 genes measured in the sputum samples improves the early detection of NSCLC<sup>167</sup>. These studies demonstrate the feasibility and importance of integrating data from different platforms to build multi-modal biomarkers for early lung cancer diagnosis.

The recent effort also extends the usage of early lung cancer biomarkers to its intended use populations, since the screening criteria from NLST excludes the majority of lung

cancer cases being screened<sup>168</sup>. Instead, the Detection of Early Lung Cancer Among Military Personnel (DECAMP) prospective observational trials (NCT01785342 and NCT02504697) uses the National Comprehensive Cancer Network (NCCN) relaxed eligibility criteria to include individuals from a broader risk spectrum<sup>169</sup>. The inclusion criteria and risk background of DECAMP make nasal swab collection a reasonable strategy given its lower chance of complications and higher accessibility. With the miRNA and gene expression profiles from the nasal swabs collected from DECAMP, I will address the unmet need to differentiate malignant from the majority of benign IPNs by building an integrated biomarker.

### ***1.7 Dissertation Aims***

In the following aims, I investigate the gene expression regulation, both transcriptional and post-transcriptional, to facilitate the understanding, interception and the early detection of lung cancer.

#### **Aim 1: Identify microRNA-mRNA regulatory interactions associated with bronchial premalignant lesion molecular subtypes and progression**

Previous study indicated immune evasions take place early during the PML development process and contribute the progression pathologies. Yet, it is unclear how miRNA-mediated gene regulations are associated with PML. Thus, I utilized matched miRNA and gene expression profiles from longitudinally collected bronchial PML biopsy samples to construct miRNA-gene regulatory network and to identify miRNAs that regulate the immune-related gene module. Furthermore, I developed novel computational framework

and R package *DReAmiR* to investigate the miRNA-mediated regulatory network rewiring associated with PML molecular subtypes. The results of Aim 1 are discussed in Chapter 2 and 3.

**Aim 2: Identify potential drug targets to intercept the progression of bronchial premalignant lesions by identifying regulatory alterations in the Hippo pathway**

Evidence indicated both Hippo and TP63 pathway activities are crucial for lung cancer development, but their function in bronchial PML is much less investigated. Integrating chromatin binding profiles and gene signatures of YAP/TAZ, TEAD and TP63, I showed that they localize to overlapped chromatin regions and have concordant gene signatures. I further explored the functions of the direct regulated genes and revealed their contribution to early immune evasions associated with bronchial PML progression. Additional, I examined the small compound treatment gene expression signatures and showed the feasibility of targeting Hippo pathway to intercept early lung cancer. The results from Aim 2 are described in Chapter 4.

**Aim 3: Identify microRNA expression profiles in the nasal epithelium to distinguish malignant vs. benign pulmonary nodules**

There is a clinical unmet need for a non-invasive biomarker to differentiate the malign IPNs from the majority of benign IPNs. With the matched miRNA and gene expression data from nasal epithelial brushings from the DECAMP I cohort, I derived miRNA expression signatures and top-scoring pairs associated with cancer status. Additionally, I constructed an ensemble classifier integrating miRNA-related and clinical features, and compared its performance to a classifier with only clinical variables. The results are

discussed in Chapter 5.

## CHAPTER 2 THE ROLE OF EPITHELIAL MIR-149 IN IMMUNE MODULATION AND PROGRESSION OF BRONCHIAL PREMALIGNANT LESIONS

*Adapted from the following manuscript:*

B Ning, RM Pfefferkorn, G Liu, S Zhang, H Liu, C Stevenson, ME Reid, SA Mazzilli, AE Spira, ME Lenburg, and JE Beane. The role of epithelial miR-149 in immune modulation and progression of bronchial premalignant lesions. *In preparation*.

### **2.1 INTRODUCTION**

The National Lung Screening Trial (NLST) and its extended follow-up analysis have shown that screening among the individuals with high risk for lung cancer with the use of low-dose computed tomography (CT) can reduce mortality from lung cancer<sup>14,170</sup>.

Despite the advance in screening, lung cancer still has the largest number of estimated deaths per year in both females and males in the United States as of 2020<sup>171</sup>. This is in part due to our incomplete understanding of the molecular events associated with early lung carcinogenesis and the inability to intercept the disease progression at pre-invasive stages. Bronchial premalignant lesions (PMLs), the histological abnormalities in the bronchial airway epithelium, are the precursors for lung squamous cell carcinoma (LUSC). The pathological grade progression of PMLs has been well characterized: from normal, hyperplasia, metaplasia, dysplasia (mild, moderate, and severe) to carcinoma in situ (CIS) and eventually to invasive LUSC<sup>172</sup>. Higher-grade persistent or progressive PMLs are associated with a higher risk for lung cancers, but not all PMLs will progress to invasive cancers. More than half of the higher-grade PMLs regress to lower-grade

spontaneously without any intervention and only a small fraction of PMLs eventually progress to CIS or invasive cancer<sup>16,31,173</sup>. Although large consortium efforts have tremendously improved our understanding of gene expression alterations in LUSC<sup>174,175</sup>, The cross-sectional profiling of advanced tumor tissues of these studies limits their ability to identify the gene expression alterations associated with early lung cancer progression. Thus, there is an unmet need to characterize the temporal dynamics of gene expression in bronchial PMLs in order to better understand the mechanism that drives the progression of PMLs<sup>176,177</sup>.

Previous work from our group utilized 302 endobronchial biopsies and 160 normal-appearing bronchial brushings longitudinally collected from 49 patients and investigated gene expression alterations associated with PML progression<sup>39</sup>. We identified nine co-expressed gene modules and classified the PMLs into four molecular subtypes based on the gene module expression patterns. Among these, the PMLs of the Proliferative subtype are enriched with dysplasia, have high expression of genes in proliferation pathways, and have high basal cell and low ciliated cell composition. We have also shown that an immune-related gene module, consisting of genes involved in interferon response and antigen processing and presentation pathways, is down-regulated among the progressive/persistent PMLs comparing to regressive ones in the Proliferative subtype and suggest immune evasion may contribute to the PML progression. The association between lack of immune surveillance and progressive/persistent higher grade PMLs has been supported by recent work from other groups as well<sup>33,35,178</sup>. However, still very little is known about what drives the immune-related gene regulation that leads to early

immune evasion and PML progression.

We hypothesize that microRNA (miRNA) may be regulating the expression of genes in the interferon signaling and antigen processing and presentation module. miRNAs are short non-coding RNAs that suppress gene expression, and facilitate transcript degradation through complementary base-pair binding to the 3' untranslated region of gene transcripts<sup>50</sup>. Depending on the target genes that miRNA interacts with, miRNAs can be oncogenic or tumor suppressive and regulate the hallmarks of cancer in a wide-range of cancer types<sup>179,180</sup>. Studies stemming from The Cancer Genome Atlas (TCGA) suggest miRNA mediated gene regulation contributes to cancer subtype expression patterns, epithelial-to-mesenchymal transition, metastasis, and patient survival in breast<sup>181,182</sup>, ovarian<sup>183,184</sup>, colon and rectal<sup>185</sup> and lung cancers<sup>186</sup>. In the early cancer setting, miRNA biomarkers have been developed for early diagnosis and outcome prediction<sup>187-189</sup>. Particularly in the early lung cancer setting, our lab has previously identified four miRNAs whose expression in main-stem bronchial brushings can improve the gene-based biomarker for better early lung cancer diagnostics<sup>166</sup>. However, there has not been a study specifically on how miRNA expression alterations may contribute to bronchial PML progression.

We hypothesized that miRNAs might contribute to the progressive lesion pathology by suppressing genes in the immune-related gene module. In this project, we utilized sample matched mRNA and miRNA sequencing data from longitudinally collected endobronchial lesion biopsies to probe miRNA-mediated gene regulation that leads to PML progression. We identified miR-149-5p as a potential regulator of the immune-

related gene module associated with higher-grade lesion progression. This miRNA is highly enriched within the epithelial cell population and its expression is positively correlated with basal cell markers within the PMLs. We also found that the miR-149-5p target gene NLRC5, which regulates MHC Class I and antigen processing genes, is down-regulated in progressing/persistent PMLs. This leads us to hypothesize that miR-149-5p contributes to early immune suppression via modulation of epithelial-associated genes involved in antigen presentation and recruitment of immune cells. These data also suggest the potential of miR-149-5p as a therapeutic target for preventing PML progression to lung cancer.

## **2.2 METHODS**

### *2.2.1 Sample Collection*

Patient enrollment and sample collection procedures were the same as described in our previous publication<sup>39</sup>. Briefly, we collected endobronchial biopsies and brushing samples from high-risk individuals undergoing cancer screenings at about 1-year intervals at Roswell Park Comprehensive Cancer Center between 2010 and 2015. The patient inclusion criteria include: patient needs to have a previous history of aerodigestive cancer and no disease at the time of enrollment or age greater than 50; patient need to have a smoking history (either current or former) of at least 20 peck-years and at least one other risk factor including moderate COPD (forced expiratory volume (FEV1) < 70%), confirmed asbestos-related lung disease or a strong family history of lung cancer (at least 1–2 first-degree relatives).

Two biopsy samples were obtained at each abnormal and suspicious region. One was used for histological evaluation and the other used for mRNA and miRNA sequencing profiling. Brushing samples were collected in the normal-looking regions within the mainstem bronchus. The progression status for each biopsy sample was defined by comparing the histology of this biopsy to the worst histological biopsy sampled at the same anatomic location in the future. A lesion that changes from normal/hyperplasia/metaplasia to dysplasia or stays stably in dysplasia grades was annotated as “progressive/persistent”. A lesion that moves back within dysplasia grades or from dysplasia grade to normal/hyperplasia/metaplasia was annotated as “regressive”. A lesion that changes only between normal/hyperplasia/metaplasia at all future timepoints was annotated as “normal/stable”. Finally, those lesions without subsequent samples were annotated as “unknown”. The analysis of this project was performed on the samples from the Discovery cohort, collected between 2010 and 2012 as mentioned in our previous publication, and 197 biopsies and 91 brushings from 30 subjects were included.

This project was approved by the Institutional Review Boards at Boston University Medical Center and Roswell Park Comprehensive Cancer Center. Written informed consent was obtained from each subject and the participations were voluntary during sample collection.

### 2.2.2 miRNA-Seq Library Preparation, Sequence Data Processing, and Sample Filtering

In total, miRNA library preparation and sequencing were performed for 167 biopsies and 91 brushing samples with matched gene expression profiles. miRNA was extracted from biopsies or brushing samples using the miRNeasy Mini kit (Qiagen) according to the manufacturer's instructions. Sequencing libraries were prepared using NEBNext Multiplex Small RNA Library Prep Set for Illumina (NEBioLabs). Sequencing adapters that target the 3' hydroxyl group of small RNAs were ligated and the transcripts were reverse transcribed and PCR-amplified into single-stranded cDNA libraries. The libraries were pooled and size selected for small RNA fragments on PAGE gel in groups of 6-10. The samples were then sequenced on Illumina® HiSeq 2500 to generate more than 10 million single-read 36-bp reads per sample<sup>190</sup>.

De-multiplexing and generation of FASTQ files were performed using Illumina CASAVA v1.8.2. FastQC v0.11.7 was used to examine the quality of raw reads and cutadapt v1.18 was used for trimming the sequencing adaptors (5'-AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3')<sup>191,192</sup>. Reads with lengths shorter than 16 nt or longer than 25 nt were discarded, which indicated a read is not a properly sequenced mature miRNA transcript. Then, the alignment and quantification of mature miRNAs based on miRBase v22 were performed with miRDeep2 v0.1.0<sup>193,194</sup>. We then performed miRNA level and sample level quality control separately in biopsy and brushing data. We limited the analysis to the miRNA samples with matched mRNA expression profiles previously analyzed (N = 249 samples)<sup>39</sup>. The raw counts were converted into log<sub>2</sub> counts per million (logCPM) and quantile normalized with voom<sup>195</sup>.

miRNAs with logCPM less than 1 in more than half of the samples were removed. Principal component analysis (PCA) and between-sample Pearson correlation were derived from the resulting expression matrix. Next, sample level filtering was conducted among biopsy and brushing samples separately. We excluded a sample if it had more than one sequencing quality metric being outside 2 standard deviations from the mean. The quality metrics we used included the first and the second principal components from the PCA, the mean Pearson correlation with all other samples, and the transcript integrity number (TIN) of the matched mRNA sequencing sample (calculated using RSeQC v3.0.0<sup>196</sup>). Finally, we kept only those samples that also had good quality matched mRNA sequencing data that were previously used to derive gene modules and molecular subtypes. The same gene filter and normalization described above were performed on the remaining samples. The residual miRNA expressions adjusting for sequencing batch were computed with limma<sup>197</sup> for downstream analysis. The expression profiles for 153 biopsy samples with 525 miRNAs and 87 brushing samples with 488 miRNAs from 30 patients were eventually used.

Sample matched miRNA expression data for the TCGA LUSC used in our previous publication (N=446) were also used to validate the miRNA-gene correlation. miRNA expression data were obtained using TCGAbiolink<sup>198</sup>. Counts associated with the same mature miRNA transcript were aggregated. miRNA counts normalization was performed the same way as in our datasets. The Pearson correlation coefficients were then calculated between miRNA and mRNA data on expression residuals adjusting for the plate as previously described.

### *2.2.3 Construct miRNA-Module Network to Identify miRNAs Associating with Gene Modules*

To capture the association of miRNAs with 9 co-expressed gene modules within the biopsy samples, we first constructed a miRNA-mRNA regulatory network based on the genes that were used in the weighted correlation network analysis (WGCNA) within the biopsy samples in the discovery cohort (N=11852; **Supplementary Table B.1**)<sup>199</sup>. For this network, we utilized both the predicted gene targets for a miRNA and the correlation between a miRNA and its targets (**Figure 2.1a**). The predicted gene targets for each miRNA were defined as those identified by any of the three miRNA target databases, including TargetScan v7.2 (conserved targets from default predictions), starBase v2.0 (strict stringency of CLIP data), and miRTarBase v7.0<sup>71,200,201</sup>. Between each miRNA and its predicted target gene, the Pearson correlation coefficient was calculated using the residual expression values adjusting for sequencing batch, and the significance level was adjusted using false discovery rate (FDR). Only the edges with significant and negative Pearson correlation coefficients (FDR  $\leq$  0.05) between miRNA and predicted target genes were retained in the network.

After constructing the miRNA-mRNA regulatory network, we sought to filter for the connections from miRNAs whose regulatory relationship with a gene module was stronger and more specific than with other gene modules (**Figure 2.1b**). For this purpose, we filtered the miRNAs connecting to the genes of a gene module based on two statistical tests. More specifically, for a gene module, we first performed Fisher-Z transformation on the Pearson correlation coefficient density of a miRNA with its targets within this

module and compared to the density with its targets in other modules using a two-sample or one-sample (if only one target is within a group) t-test to select for miRNAs whose relationship is stronger with this gene module (ie. more negative). Next, we used Fisher-exact tests to examine whether or not a miRNA had more significantly negative-correlated target genes in the module of interest than in other modules (odds ratio > 1). The results of both tests were adjusted using FDR, and miRNAs with significant results for the same gene module from both tests (FDR ≤ 0.1) were considered specifically regulating this particular gene module. Gene-set variation analysis (GSVA) was used to summarize miRNA set expression scores for the remaining miRNAs connecting to each gene module<sup>202</sup>.

#### *2.2.4 Identification of miRNAs and Genes Associating with PML Progression Status*

The association between miRNAs that regulate immune-related modules and the PML progression status was determined based on whether the expression level of a miRNA is significantly different between the progressive/persistent and the regressive PMLs of the Proliferative subtype (p-value < 0.05). Within each model, miRNA residual expressions adjusted for sequencing batch were used as the dependent variable, PML progression status was the main independent variable, and patient was included as a random effect using ‘duplicateCorrelation()’ function from limma v3.44.3<sup>197</sup>.

To identify the genes that are associated with PML progression status, we used the full gene expression data of the discovery and validation cohorts from our previous publication, including those samples without matched miRNA sequencing data<sup>39</sup>. Similar

linear mixed-effect models were performed with gene residual expressions adjusted for sequencing batch and TIN as the dependent variable, PML progression status as the independent variable while adjusting patient as a random effect. We further validated gene associations with PML progression status using an external microarray dataset, in which linear models were used with probe intensities as the dependent variable and progression status, as defined in our study, as the main independent variable.

#### *2.2.5 Examination of the Cell-Type Specific Expression of miR-149-5p*

To examine whether a miRNA is potentially specifically expressed within a certain cell type, we examined the correlation between the expression level of this miRNA and cell-type marker genes. The method for calculating Pearson correlation coefficients has been described in the previous section. Cell-type marker genes we used were: CD3g for T cells, CD19 for B cells, CD68 for macrophages, KRT5 for basal cells, FOXJ1 for ciliated cells, MUC5ac for goblet epithelial cells, , and SCGB1A1 for club cells.

Also, miRNA expression data from human primary cell-types were used to evaluate whether or not the expression of a miRNA is enriched within a certain group of samples<sup>203</sup>. The data contains miRNA expression profiles for 2595 mature miRNAs across 399 samples. The library size normalized (CPM) miRNA expression matrix was directly downloaded from <https://fantom.gsc.riken.jp/data/> and was log<sub>2</sub>-transformed. Classification of FANTOM5 samples to cell types were based on the sample annotations. Samples were ranked by their expression levels of miR-149-5p. Then, gene set enrichment analysis (GSEA<sup>204</sup>) was used to test whether or not samples with higher miR-

149-5p expression were enriched for samples from a specific group of cell-types.

#### *2.2.6 Analysis of NLRC5 ChIP-seq data*

NLRC5 peak locations and genome coverages from previous NLRC5 ChIP-seq data in mice were obtained from GSE59092<sup>205</sup>. Normalized read coverage in reads per million (RPM) around the NLRC5 regulated genes were averaged across biological duplicates. Genome track visualizations were made with karyoploteR v1.16.0<sup>206</sup>.

#### *2.2.7 Analysis of FANTOM5 gene expression data*

Matched gene expression profiles in transcripts per million (TPM; N=394) from the functional annotation of the mammalian genome 5 (FANTOM5) project were used to examine the miRNA-mRNA correlation by cell type<sup>203</sup>. The count table obtained from the FANTOM5 atlas of miRNAs website (<https://fantom.gsc.riken.jp/data/>) and were log2-transformed for correlation analysis.

#### *2.2.8 Cell-type deconvolution analysis*

Cell-type deconvolution was conducted using AutogeneS<sup>207</sup>. Based on the package recommendation, gene expression data from biopsy and brushing samples was first normalized to transcript per million (TPM), and batch correction was performed using combat<sup>208</sup>. Cell type reference was generated using single-cell RNA sequencing data containing 2,075 cells from 17 bronchial brushings from patients undergoing bronchoscopy for suspicion of lung cancer<sup>209</sup>.

### *2.2.9 Data availability*

Gene expression data from GSE109743 was used to construct miRNA-gene correlation network<sup>39</sup>. miRNA expression data from GSE93284 and gene expression from GSE66499 and GSE114489 were used to validate the miRNA-gene correlation and gene association with lesion outcome<sup>33,156,166</sup>.

## **2.3 RESULTS**

### *2.3.1 Patient population*

153 miRNA expression data from PML biopsy samples with matching gene expression profiles were obtained from 28 individual patients after quality filtering. Since quite a few samples were removed for low quality compared our previous analysis using mRNA expression data, we compared the clinical characteristics for the biopsy samples between the four molecular subtypes and examined whether the correlation with phenotypes was preserved in this smaller dataset (**Table 2.1**). Similar to our previous observation, significant differences were found between the molecular subtypes for the lesion-related characteristics (chi-square test p-value < 0.05): the higher-grade dysplasia lesions are enriched in the Proliferative subtypes; current smokers were enriched in both the Proliferative and the Secretory subtypes. Sequencing batch and TIN for matched RNA samples are not significantly different between the molecular subtypes. These observations indicated the patients with miRNA expression data have similar associations between molecular subtypes and phenotypes as we described in our original publication.

	<b>Proliferative (n=45)</b>	<b>Inflammatory (n=27)</b>	<b>Secretory (n=44)</b>	<b>Normal (n=37)</b>	
<b>Dysplasia Grade</b>					
<b>Normal</b>	3 (6.7)	4 (14.8)	12 (27.3)	10 (27.0)	<b>chi=37.22, p=0.0049</b>
<b>Hyperplasia</b>	4 (8.9)	8 (29.6)	7 (15.9)	6 (16.2)	
<b>Metaplasia</b>	7 (15.6)	8 (29.6)	9 (20.5)	12 (32.3)	
<b>Mild Dysplasia</b>	5 (11.1)	1 (3.7)	6 (13.6)	2 (5.4)	
<b>Moderate Dysplasia</b>	19 (42.2)	4 (14.8)	7 (15.9)	5 (13.5)	
<b>Severe Dysplasia</b>	7 (15.6)	1 (3.7)	1 (2.3)	1 (2.7)	
<b>Smoking Status (genomic prediction)</b>					
	38 (84.44)	11 (40.7)	32 (72.7)	12 (32.4)	<b>chi=30.23, p &lt; 0.001</b>
<b>Progression Status</b>					
<b>Progression/persistent</b>	14 (31.1)	4 (14.8)	13 (29.6)	5 (13.6)	<b>chi=25.69, p=0.0023</b>
<b>Regressing</b>	14 (31.1)	3 (11.1)	4 (9.1)	3 (8.1)	
<b>Normal/Stable</b>	7 (15.6)	12 (44.4)	8 (18.2)	14 (37.8)	
<b>UNK</b>	10 (22.2)	8 (20.6)	19 (43.2)	15 (40.5)	
<b>Batch</b>					
<b>1</b>	11 (24.4)	9 (33.3)	9 (20.5)	8 (21.6)	<b>chi=20.51, p=0.058</b>
<b>2</b>	13 (28.9)	5 (18.5)	17 (38.6)	8 (21.6)	
<b>3</b>	8 (17.8)	7 (25.3)	7 (15.9)	16 (43.2)	
<b>4</b>	11 (24.4)	2 (7.4)	9 (20.5)	4 (10.8)	
<b>5</b>	2 (4.4)	4 (14.8)	2 (4.6)	1 (2.7)	
<b>TIN (Matched mRNA Sample)</b>					
	78.3 (1.5)	78.3 (1.5)	78.1 (1.4)	77.8 (2.9)	<b>F=0.46, p=0.71</b>

**Table 2.1. Biopsy sample clinical annotation across four PML molecular subtypes.**

Statistical tests for categorical clinical variables (dysplasia grade, smoking status, progression status, and batch) were conducted using Chi-square tests. Statistical tests for continuous variables (TIN) were compared using two-sided Student's t-tests. Percentages are reported for categorical variables and mean/standard deviations are reported for the continuous variable.

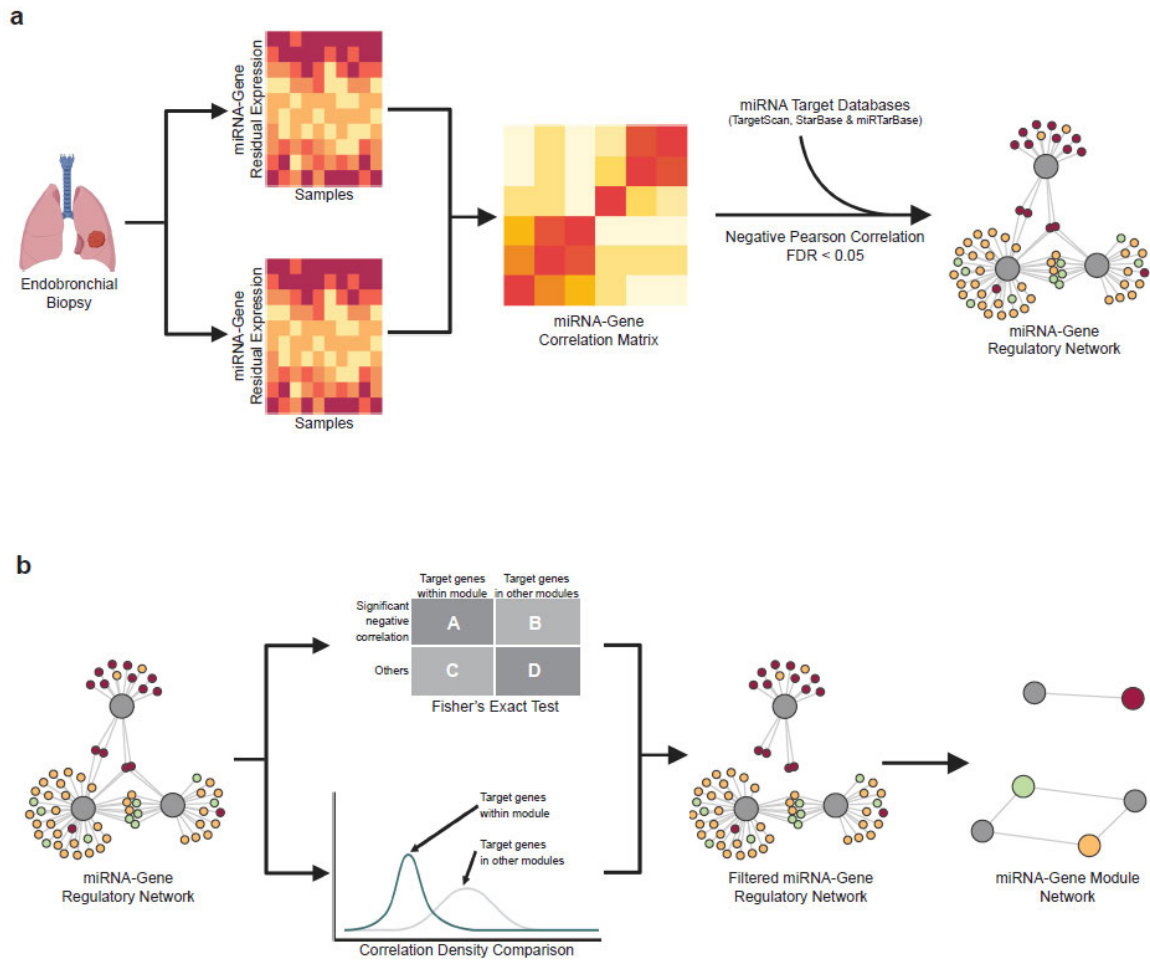
### *2.3.2 Identification of miRNAs regulating the immune-related module using miRNA-mRNA Network*

Constructing miRNA-mRNA regulatory networks has been frequently performed to elucidate how miRNAs, through negatively regulating target genes, may be functionally associated with cancer<sup>210,211</sup>. In these networks, the nodes are miRNAs and mRNAs, and the edges or connections represent negative correlations between miRNA and predicted target genes. For our case, we wanted to further classify the miRNAs based on whether they have central gene-expression regulatory roles within the miRNA-gene network in each of the nine gene modules we previously described<sup>39</sup>. This could be accomplished by testing whether a miRNA is a “hub” node in each gene module, in which a miRNA is assigned to the gene module it has the largest number of connections or is most negatively correlated with<sup>199</sup>. However, such approach has several disadvantages. First, our gene modules have various sizes. Classification based on miRNA connection degree would bias towards larger gene modules by chance, and smaller gene modules may end up with no connected miRNAs. Also, such method allows each miRNA to be associated with only one gene module yet it has been shown that miRNAs can regulate multiple different functional pathways. For example, the miR-34 family has been shown to suppress neoplastic proliferation while also being a central regulator for motile ciliogenesis in multiciliated epithelia<sup>212,213</sup>. miRNAs of the miR-200 family can target genes in both cytoskeleton processes related pathway and CD8<sup>+</sup> T cell fate specification<sup>55,214</sup>.

To overcome these issues, we constructed a miRNA-gene module network by applying

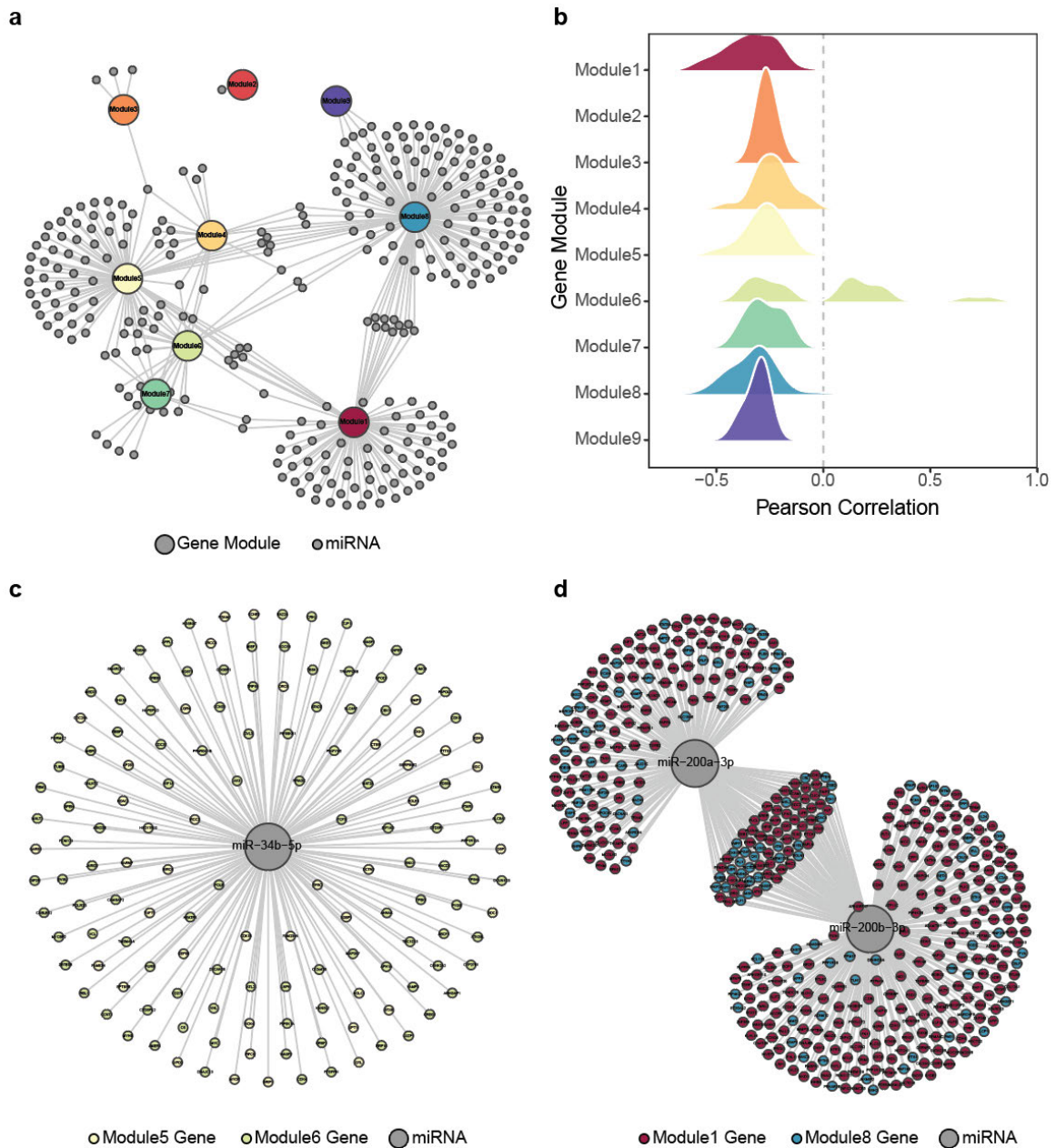
an edge-filtering strategy (**Figure 2.1b**; see Methods). To achieve this, we first built a miRNA-gene network based on the significantly negative correlations between miRNAs and their predicted target genes using 153 lesion biopsy samples with both miRNA and gene expression profiles. Then, for each co-expressed gene module, we identified and retained only the connections from miRNAs whose associations with the genes in this gene module are stronger (more negative) and the negatively correlated target genes are enriched. The edges between miRNA and target genes in the other gene modules were removed. After excluding nodes without any connections and summarizing miRNA-gene connections to the gene module level, we resulted in a bipartite, directed network consisted of 287 miRNA nodes targeting genes in nine gene modules (**Figure 2.2a**; **Table B.2**).

The connection pattern of the miRNA-gene module network agrees with our assumptions. miRNA connected to all nine gene modules were identified, and the gene modules with more connected miRNAs do not necessarily contain more genes. Also, as expected, we observed 58 miRNAs that have connections to more than one gene module. To further validate our network captures biologically relevant miRNA regulations, we then examined whether the expression levels of miRNAs are negative correlated with that of the gene modules they are connecting to. The correlation densities between the expression levels of miRNAs associating with a gene module and the module score for that module were all left-shifted, with the exception of module 6 which contained both positive- and negative-correlated genes (**Figure 2.2b**). We also observed that the GSVA scores for miRNAs connected to each gene module are significantly negatively correlated



**Figure 2.1. Analysis design diagram.**

**a.** 148 longitudinally collected endobronchial biopsies from 30 patients were collected and both mRNA and miRNA were sequenced. A miRNA-gene regulatory network was derived based on both significant negative correlation and evidence that the miRNA targets the mRNA. Pearson correlation was calculated from expression residuals for each miRNA-mRNA pair, and those with significant negative Pearson correlation coefficients ( $FDR \leq 0.05$ ) were selected. Target information was obtained by combining both sequence-based prediction databases and experimentally validated databases. **b.** The Fisher-Z transformed Pearson Correlation coefficient densities of one miRNA with its targets within the module of interest were compared to that with its targets in other modules using a t-test to select for miRNAs that strongly regulate the module of interest ( $t < 0$ ). Fisher-exact tests were performed to select for miRNA with significantly more connected target genes in the module of interest than in other modules ( $OR > 1$ ). miRNAs with significant results from both tests ( $FDR \leq 0.1$ ) were selected to construct a miRNA-Gene module network.



**Figure 2.2. miRNA-Gene module network captures miRNA regulatory roles.**

**a.** miRNA-Gene module network after connection filtering. miRNAs were shown by small grey circles and the gene modules were shown as large colored circles. Edges showed the connection after filtering (see Method) between the miRNAs and gene modules. **b.** Pearson correlation densities between miRNA connecting to each gene module in the miRNA-gene module network and the predicted target genes of that gene. **c-d.** Network plots of miRNAs connecting to two gene modules and the predicted target genes within those gene modules: miRNA-34c-5p and miR-34b-5p were connected to the cell cycle, DNA replication gene module, and the cilia biogenesis and function gene module (Module 5 and 6; miR-200a-3p and miR-200b-3p were connected to the extracellular matrix, cell adhesion gene module and the immune activation, inflammatory response gene module (Module 1 and 8).

with the corresponding module GSVA scores (**Figure A.1**; FDR < 0.05, Pearson correlation). Moreover, the connections between miRNA and gene module in the network reflect multifaceted functions of miRNAs demonstrated by previous studies. miR-34b-5p are connected to gene modules with genes associated with cell proliferation and cilia biogenesis<sup>212,213</sup> (module 5 and 6), while miR-200a/b-3p are connected to gene modules with genes associated with ECM and inflammatory response<sup>55,214</sup> (module 1 and 8; **Figure 2.2c-d**). These results suggest that our miRNA-gene module network reveals biological meaningful regulatory relationships between miRNAs and the predicted target genes.

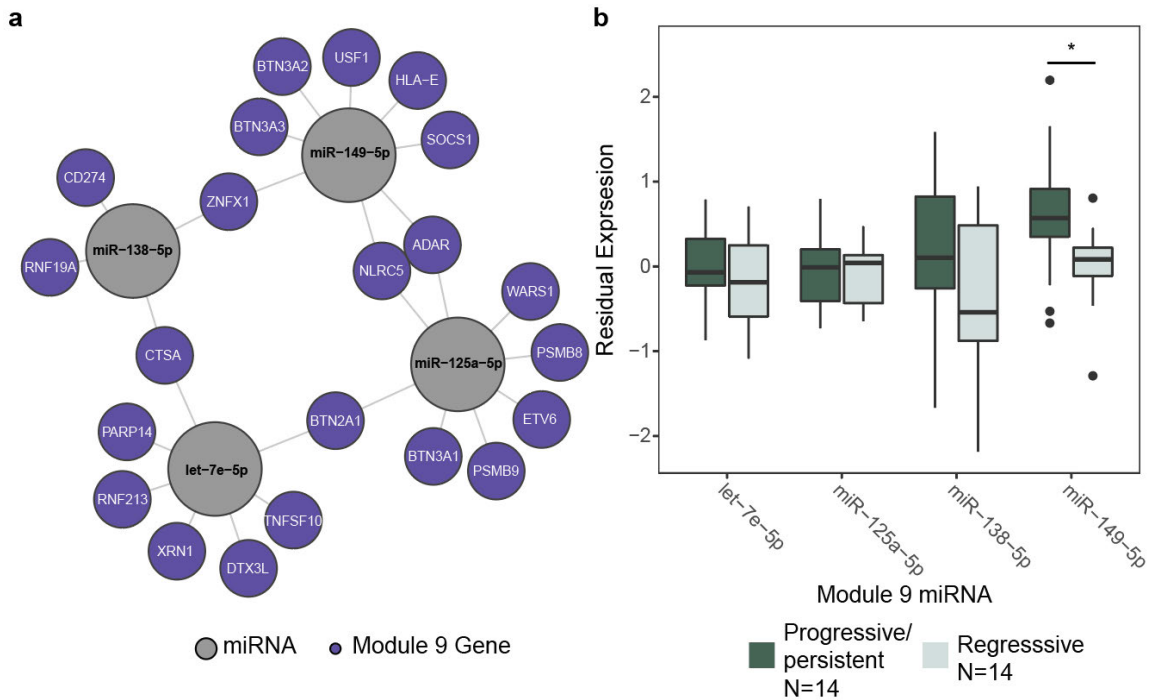
### *2.3.3 hsa-miR-149-5p is Associated with Immune-Related Gene Module and Lesion Progression*

The immune-related gene module, enriched with genes associated with interferon response and antigen presentation and processing pathways, has previously been shown to be down-regulated among the progressive/persistent comparing to the regressive PMLs of the Proliferative subtype. In order to identify miRNAs that potentially contribute to the PML progressive phenotype, we first identified the miRNAs that are connected to the immune-related gene module. Based on the miRNA-gene module network, four miRNAs (hsa-let-7e-5p, hsa-miR-125a-5p, hsa-miR-138-5p and hsa-miR-149-5p) are found to be significantly negatively correlated with 21 genes of the immune-related gene module (**Figure 2.3a**). For each of these four miRNAs, as described above, the correlation density with predicted target genes in the immune-related gene module is significantly

more negative than that with predicted target genes in the other gene modules, and the significant negatively correlated target genes are enriched within the immune-related gene module (FDR < 0.15, t-test and Fisher-exact test; **Table B.3**).

Next, we examined whether or not the expression levels of these miRNAs are associated with the progressive status of PMLs of the Proliferative subtype in our biopsy dataset. hsa-miR-149-5p was significantly upregulated within the progressive/persistent (N=14) comparing the regressive (N=14) PMLs of the Proliferative subtype (p-value < 0.05, linear model; **Figure 2.3b and Table B.4**), suggesting its higher expression is associated with worse lesion prognosis. As oncogenic role for hsa-miR-149-5p has been previously identified in different cancer settings<sup>215,216</sup>.

We further validated the association between hsa-miR-149-5p and the immune-related gene module in three datasets with sample matched miRNA and gene expression profiles: airway brushings samples from our project, TCGA LUSC tumor samples and the mainstem bronchus brushing from the Airway Epithelial Gene Expression in the Diagnosis of Lung Cancer (AEGIS) trials. Genes of the immune-related gene module were enriched among the genes that were negatively correlated with hsa-miR-149-5p in all datasets (p-value < 0.001, GSEA; **Figure A.2a**). Also, the genes in each leading edge were strongly overlapped (**Figure A.2b**). These observations suggest hsa-miR-149-5p may promote the progressive PML phenotype by specifically suppressing the target genes the immune-related gene modules.



**Figure 2.3. miR-149-5p targets genes in the immune-related gene module and is upregulated in the progressive PML samples within the Proliferative subtype.**

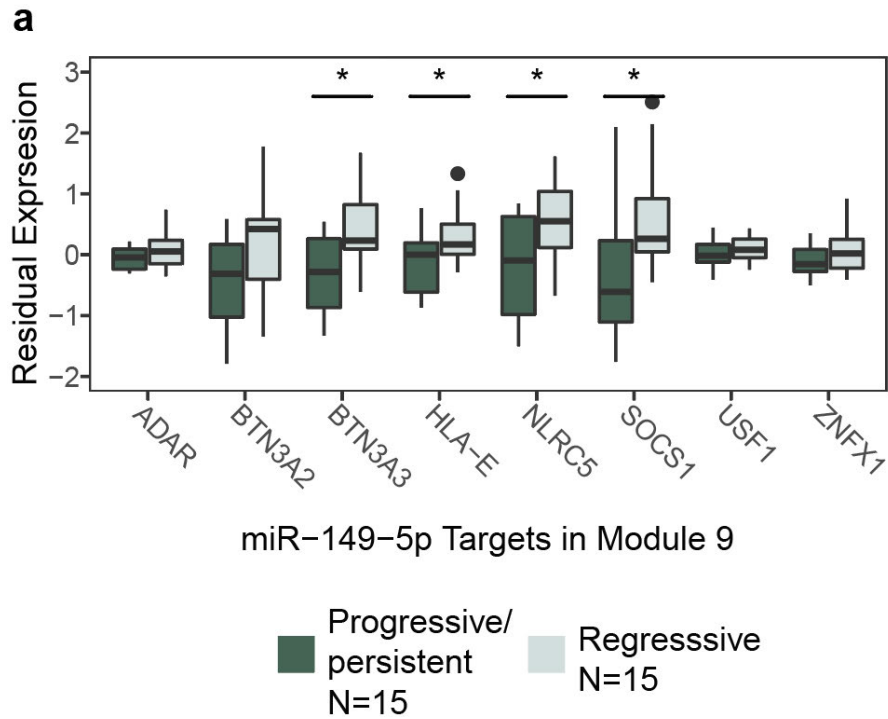
**a.** Network plot of miRNAs connecting to the immune-related gene module (Module 9). miRNAs were shown by large grey circles and the significantly negatively correlated target genes of the immune-related gene module were shown as purple circles. Edges showed significantly negative correlations between the miRNAs and the predicted target genes. **b.** Boxplots of residual expression levels of the four miRNAs specifically connected to the immune-related gene module.

\*p-value < 0.05.

#### *2.3.4 miR-149-5p Regulates MHC Class I Genes Through Suppressing NLRC5*

##### *Expression*

Next, we sought to better elucidate the mechanism of how hsa-miR-149-5p up-regulation contributes to lesion progression by investigating the targets it regulates. Eight significantly negatively correlated target genes of hsa-miR-149-5p were found in the immune-related gene module. We compared the expression levels of these genes between the progressive/persistent (N=15) versus the regressive (N=15) PMLs of the Proliferative subtype in the full gene expression data. The expression level of four negatively correlated predicted target genes, BTN3A3, HLA-E, NLRC5 and SOCS1, are significantly lower in the progressive/persistent proliferative PMLs than in the regressive ones (p-value < 0.05, linear model), while all of them are down-regulated in the progressive/persistent PML samples (**Figure 2.4 and Table B.5**). The lower expression of these genes among PMLs that progress is also observed in the proliferative PML samples from our validation cohort (N=7 and 12) and in an independent dataset from GSE114489 (N=32 and 15) (**Table B.5**). Since hsa-miR-149-5p is elevated in the progressive/persistent PMLs of the Proliferative subtype, these results suggest that it may be responsible for some of the decreased gene expression associated with lesion progression.

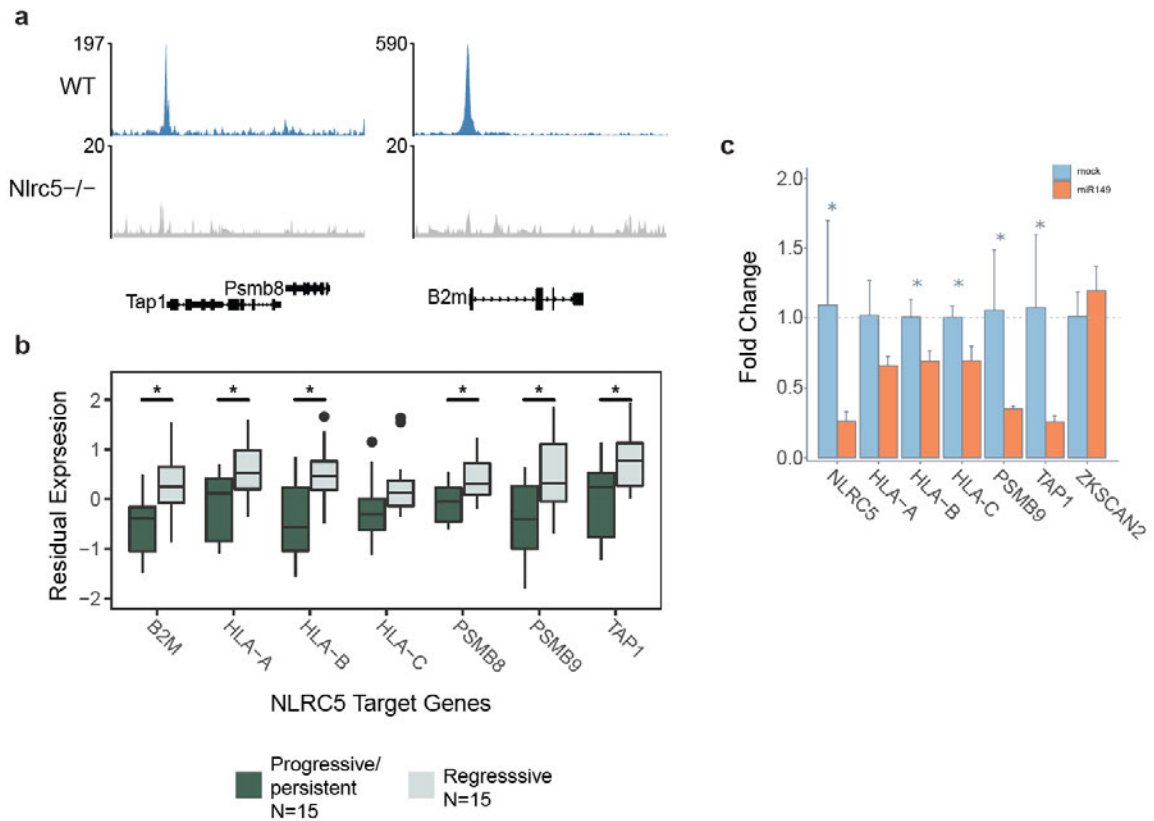


**Figure 2.4, The predicted target genes of miR-149-5p were associated with PML progression within the Proliferative subtype.**

Boxplots of residual expression levels of the miR-149-5p significantly negatively correlated target genes of the immune-related gene modules between the progressive/persistent PML samples and the regressive samples within the Proliferative subtype. \*p-value < 0.05.

Among these genes, we found NLRC5 to be particularly interesting. NLRC5 is a predicted target for hsa-miR-149-5p in two out of the three miRNA target databases we used. The negative correlations are statistically significant in the sample matched miRNA and mRNA expression profiles in tumor samples from TCGA LUSC (N=446), the bronchial brushing samples (N=82) from our dataset, and the mainstem bronchus brushing samples (N=341) from AEGIS clinical trials ( $p \leq 0.05$ , Pearson correlation; **Figure A.3a-c**).

Previous studies have showed that NLRC5 is a member of the NOD-like receptor (NLR) family and is a transcriptional activator of the MHC class I genes<sup>217</sup>. Since previously we have shown that the depletion of CD8 T cells is associated with PML progression<sup>39</sup>, we hypothesized this could partially be due to hsa-miR-149-5p regulating the expression of the MHC Class I gene by down-regulating NLRC5 expression. Seven genes (B2M, HLA-A/B/C, PSMB8/9 and TAP1) were identified to be directly regulated by NLRC5 based on literature reviews<sup>218,219</sup>. NLRC5 promoter-binding at the mouse orthologues of these genes can be observed via ChIP-seq experiments from mice T cells<sup>205</sup> (**Figure 2.5a and Figure A.4**). All these genes are within the immune-related gene module, and the expression levels for all, except HLA-C, were significantly down-regulated within the progressive/persistent proliferative PML (**Figure 2.5b and Table B.6**). Further validation was done in data from our validation cohort from GSE114489 revealed the same dysregulation (**Table B.6**). qPCR experiments in SW900 lung cancer cell lines transfected with miR-149 or mock sequence demonstrated that NLRC5, as well as its target genes HLA-B/C, PSMB9 and TAP1, were



**Figure 2.5. NLRC5 regulates MHC Class I gene expression associated with PML progression.**

**a.** WT and Nlrc5<sup>-/-</sup> ChIP-seq tracks from (GSE59092) shows NLRC5 binding at the promoter regions of B2m and Tap1. **b.** Boxplots of residual expression levels of the NLRC5 target genes in the progressive/persistent and regressive PML samples within the Proliferative subtype. **c.** qPCR of NLRC5 and NLRC5 regulated genes expression levels in SW900 in miR-149 transfection and mock control samples (n=3). ZKSCAN2 was used as negative control. \*p-value < 0.05.

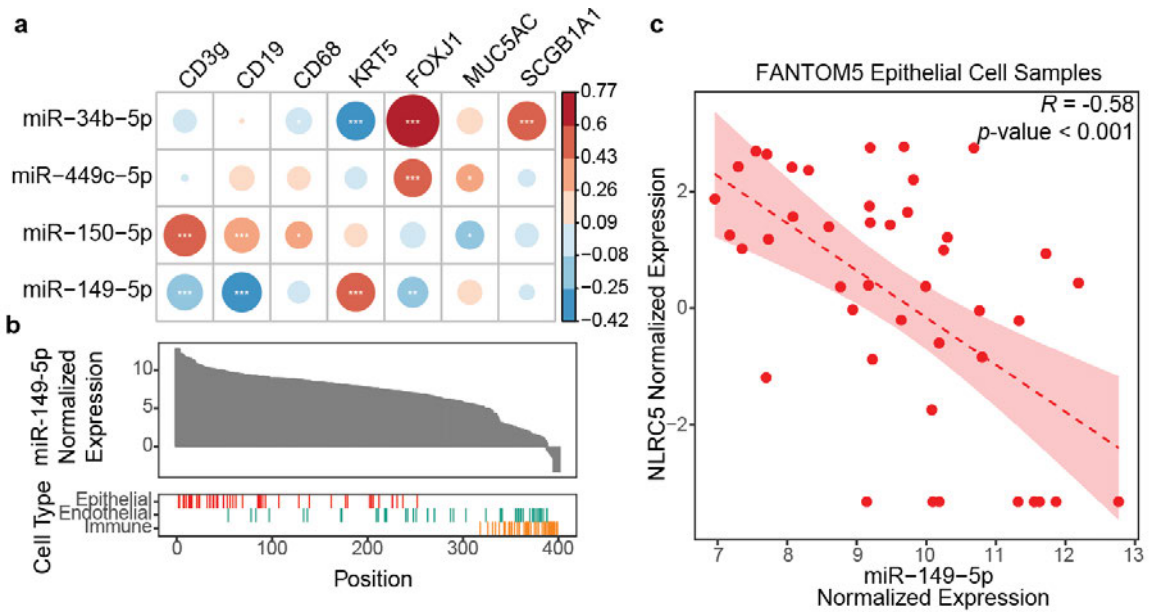
significantly decreased upon transfection with miR-149 (**Figure 2.3c**; t-test p-value < 0.05). These data suggest that hsa-miR-149-5p potentially promotes the decrease of CD8 T cells and lesion progression by suppressing NLRC5 and MHC Class I gene expression.

### *2.3.5 has-miR-149-5p Expression is Enriched within Epithelial Cells*

Cell-type-specific expression of genes or miRNAs can potentially affect interpretation of data from tissue sequencing experiments. To further explore our hypothesis, we aimed to determine whether the expression of hsa-miR-149-5p is specific to certain cell-types.

First, we examined the correlation pattern between hsa-miR-149-5p and canonical cell type markers within our lesion biopsy samples. As shown in **Figure 2.6a**, significantly positive correlations were only observed between has-miR-149-5p and the basal cell markers KRT5 (FDR < 0.001, Pearson correlation) but no other markers. In comparison, hsa-miR-34b-5p and hsa-miR-449c-5p, which are expressed highly in mucociliary epithelia, were positively correlated with FOXP1, and hematopoiesis essential hsa-miR-150-5p were positively correlated with immune cell markers<sup>213,220</sup>.

Also, we utilized cell-type-specific transcriptomic data from the FANTOM5 project<sup>203</sup>. FANTOM5 samples were ranked based on the normalized expression level of hsa-miR-149-5p. We then selected samples of epithelial, lymphoid, or myeloid cell-types, and examined whether any cell-type group was enriched among those with high hsa-miR-149-5p expression. Based on GSEA analysis, samples belonging to the epithelial cell group (N=44) were positively enriched among samples with high hsa-miR-149-5p expression (p < 0.001, GSEA), while the endothelial (N=42) and the immune cell (N=36)



**Figure 2.6. miR-149-5p is highly expressed in and regulates NLRC5 within epithelial cells.**

**a.** Bubble plot of the correlation between the expression level of miR-149-5p and cell-type marker genes, including CD3g (T cells), CD19 (B cells), CD68 (macrophages), KRT5 (basal cells), FOXJ1 (ciliated cells), MUC5AC (club cells) and SCGB1A1 (goblet cells). Correlation results for miR-34b-5p, miR-449c-5p and miR-150-5p were also shown as controls. \*FDR  $\leq 0.05$ ; \*\*FDR  $\leq 0.01$ ; \*\*\*FDR  $\leq 0.001$ . **b.** Enrichment of miR-149-5p across cell-type compartments in the FANTOM5 project. (Top) Barchart showed the rank list in which FANTOM5 samples (N=393) were ranked by the normalized expression levels of miR-149-5p. (Bottom) Vertical bars indicated the position of FANTOM5 samples belonging to epithelial, endothelial, or immune cell compartments. **c.** Scatter plot of the correlation between the normalized expression levels of miR-149-5p and NLRC5 within the FANTOM5 samples belongs to the epithelial cell compartment (N=41). Linear fitness was shown as a dashed line and 95% CIs were shown as shade.

groups were negatively enriched ( $p = 0.05$  and  $p < 0.01$ , GSEA; **Figure 2.6b**). Our findings indicate that the expression of hsa-miR-149-5p is highly enriched within epithelial cell-types, particularly in basal cells.

### *2.3.6 Interaction between miR-149-5p and the NLRC5 can be observed in epithelial cell*

The observation that hsa-miR-149-5p is highly expressed among epithelial cells is interesting given its implication in immune-regulation. There are two potential hypotheses to explain the apparent conflict: hsa-miR-149-5p functions within epithelial cells to indirectly regulate immune response, or its differential expression level reflects cell-type composition changes associated with lesion progression status. In fact, the significant correlation between the expression level of hsa-miR-149-5p and CD3g and CD19 may indicate an inverse relationship between the abundance of immune and epithelial cells (**Figure 6A**) and potentially favors the latter.

To test these hypotheses, we examined the correlation between hsa-miR-149-5p and NLRC5 among samples belonging to specific cell-types within the FANTOM5 data. A significant negative correlation between the expression levels of hsa-miR-149-5p and NLRC5 was observed among the samples of the epithelial cell group ( $p < 0.01$ , Pearson Correlation; **Figure 2.6c**), but not in those of the endothelial or immune cell groups (**Figure A.4a-b**), suggesting that the suppression of NLRC5 by hsa-miR-149-5p may be specific to epithelial cells. Together, the negative regulation relationship between hsa-miR-149-5p and NLRC5 within the epithelial cells indicated that hsa-miR-149-5p may be a key driver for immune suppression in early lung cancer that exerts its function within

epithelial cells.

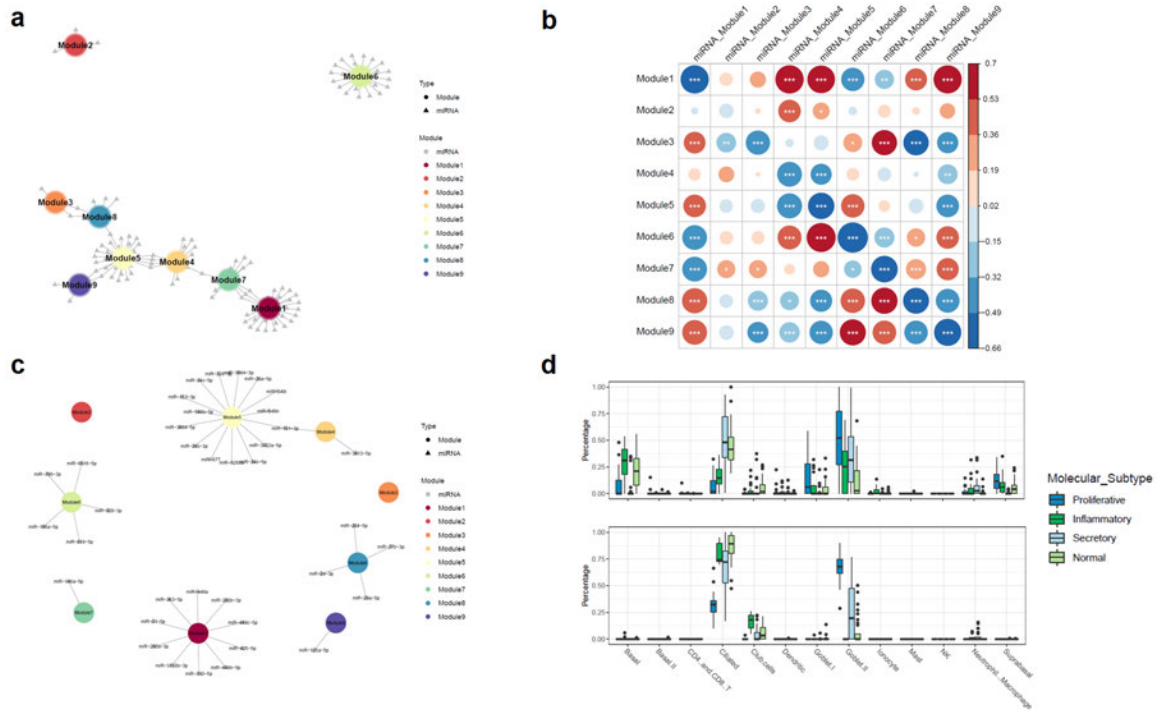
### *2.3.7 miRNA-gene module network in brushing samples collected from the normal-appearing airways*

We further sought to explore the miRNA regulated gene expression in the bronchial brushing samples, and explore the field cancerization effect associated with PML. Using the 82 samples with match gene and miRNA expression profiles (**Table 2.2**), we constructed a miRNA-gene module regulatory network as in the biopsy samples. This network contained 3093 genes and 468 miRNAs with 8738 edges. Using the filters described earlier, we removed the potentially non-specific regulatory connections to identify a miRNA-gene module network of 104 miRNAs connecting to 1107 genes from the co-expressed gene modules with 1851 edges (**Figure 2.7a**). Of these miRNAs, 13 were connected to more than gene modules. The metagene scores calculated from the miRNAs connected to each gene module were significant negatively correlated with the gene module scores, suggesting module-specific regulatory relationships (**Figure 2.7b**). Next, we compared the miRNA regulating each gene module in both biopsy and in brushing samples. Out of the 114 miRNAs in the brushing miRNA-gene module regulatory network, only 13 were also observed in the network from the biopsy samples (**Figure 2.7c**). Most of these miRNAs were connected to the modules associated with ECM pathways or cell proliferation pathways. These differences in miRNA-gene module connection patterns suggested the miRNA-regulation landscape might be different between the airway epithelium at the lesions and the normal-appearing airways.

	<b>Proliferative (n=15)</b>	<b>Inflammatory (n=11)</b>	<b>Secretory (n=25)</b>	<b>Normal (n=36)</b>	
<b>Dysplasia Grade</b>					
<b>Normal</b>	0 (0)	1 (9.1)	3 (12.0)	2 (5.6)	chi=20.08, p=0.17
<b>Hyperplasia</b>	1 (6.7)	4 (36.4)	2 (8.0)	4 (11.1)	
<b>Metaplasia</b>	0 (0)	2 (18.2)	3 (12.0)	10 (27.8)	
<b>Mild Dysplasia</b>	3 (20.0)	0 (0)	2 (8.0)	4 (11.1)	
<b>Moderate Dysplasia</b>	8 (53.3)	2 (18.2)	8 (32.0)	11 (30.6)	
<b>Severe Dysplasia</b>	3 (20.0)	2 (18.2)	7 (28.0)	5 (13.9)	
<b>Smoking Status (genomic prediction)</b>	11 (73.3)	2 (18.2)	13 (52.0)	16 (44.4)	<b>chi=8.11, p=0.043</b>
<b>Batch</b>					
<b>1</b>	1 (6.7)	2 (18.2)	6 (24.0)	5 (13.9)	chi=114.02, p=0.30
<b>2</b>	3 (20.0)	3 (27.3)	2 (8.0)	10 (27.8)	
<b>3</b>	5 (33.3)	1 (9.1)	7 (28.0)	7 (19.4)	
<b>4</b>	6 (40.0)	4 (36.4)	10 (4.0)	14 (38.9)	
<b>5</b>	0 (0)	1 (9.1)	0 (0)	0 (0)	
<b>TIN (Matched mRNA Sample)</b>	73.7 (3.2)	72.1 (4.6)	72.5 (3.6)	72.7 (3.0)	F=0.29, p=0.83

**Table 2.2. Brushing sample clinical annotation across four PML molecular subtypes.**

Statistical tests for categorical clinical variables (dysplasia grade, smoking status, progression status, and batch) were conducted using Chi-square tests. Statistical tests for continuous variables (TIN) were compared using two-sided Student's t-tests. Percentages are reported for categorical variables and mean/standard deviations are reported for the continuous variable.



**Figure 2.7. miRNA-gene module network consensus analysis between the biopsy and brushing sample.**

**a.** miRNA-Gene module network after connection filtering. miRNAs were shown by small grey circles and the gene modules were shown as large colored circles. Edges showed the connection after filtering (see Method) between the miRNAs and gene modules. **b.** Pearson correlation between metagene scores calculated from miRNA connecting to each gene module in the miRNA-gene module network and the gene module scores. **c.** Consensus miRNA-gene module network between the PML biopsy and brushing samples. Each edge showed the miRNA-gene module connected in both networks constructed in the PML biopsy and the brushing datasets. **d.** Boxplots showing the AutoGeneS cell-type deconvolution results across the four molecular subtypes for the PML biopsy (top) and the brushing (bottom) datasets. \* FDR  $\leq$  0.05; \*\* FDR  $\leq$  0.01; \*\*\* FDR  $\leq$  0.001.

## ***DISCUSSION***

Lack of understanding of the earliest molecular event associated with bronchial PML and effective strategies to intercept PML progression contributes to the high prevalence and mortality of LUSC. We previously demonstrated bronchial PMLs can be classified into four distinct molecular subtypes based on gene expression patterns. Decreased gene expression involved in antigen processing and presentation, as well as decreased abundance of anti-tumor immune cell infiltration, is associated with the persistent/progressive Proliferative PMLs<sup>39</sup>. However, the transcriptional regulators contributing to the immune evasion and progressive pathology of bronchial PMLs are still poorly understood. In this study, we measured miRNA expression profiles from longitudinally collected endobronchial biopsies and examined the miRNA-mediated gene expression network of patients undergoing lung cancer screening. With miRNA and mRNA expression data from the same samples, we constructed a miRNA-gene module network and identified four miRNAs regulating the immune-related gene modules. We further showed that miR-149-5p, a miRNA highly expressed in the airway basal cells, is up-regulated among the persistent/progressive Proliferative PMLs. Through negatively regulating the MHC Class I transactivator NLRC5, miR-149-5p suppresses MHC Class I gene expression which may promote immune evasion, particularly the lack of CD8<sup>+</sup> T cells recruitment, and PML progression (**Figure 2.8**). Together, these data suggest the molecular mechanism by which miRNA dysregulation leads to alterations in gene expression and an altered immune microenvironment.

Our miRNA-gene module network analysis across multiple datasets suggested that miR-149-5p is a regulator of the immune-related gene module in the lung. The expression level of miR-149-5p and its target genes are inversely associated with the progression status of bronchial PMLs, suggesting it to be a driver of immune evasion and early lung cancer progression. Intriguingly, previous studies have largely characterized miR-149 being a tumor-suppressive miRNA across different cancer settings. In NSCLC, Yang *et al.* showed the expression of miR-149-5p is up-regulated in the tumor compared with adjacent tissues<sup>221</sup>, while others suggested miR-149-5p inhibits the epithelial-to-mesenchymal process through targeting FOXM1 in lung cancer cell line<sup>222</sup>. Higher miR-149-5p expression has been associated with reduced drug sensitivity and cancer progression in breast<sup>223,224</sup>, gastric<sup>225</sup>, and oral cancer<sup>226</sup> through down-regulating My99, IL-6, AKT, and CDK6. In contrast, the oncogenic role of miR-149-5p has been reported in prostate cancer<sup>227</sup> and melanoma<sup>228</sup>. Similarly, our data suggest miR-149-5p acts as an oncomiR in the bronchial PMLs. Notably, Srivastava *et al.* showed that IFN- $\gamma$  activity may suppress miR-149 levels and drive inflammatory response in keratinocyte<sup>229</sup>. A similar association is observed among progressive/persistent Proliferative PMLs in our data, which are characterized by low expression of genes related to the interferon response pathways and high levels of miR-149-5p. Given these observations, miR-149-5p activity may be tissue or cancer-type specific, and the underlying mechanism requires more study to elucidate.

Here, through exploring the miRNA-mediated regulatory landscape within bronchial PMLs, we revealed a novel regulatory mechanism of NLRC5 and MHC Class I by miR-

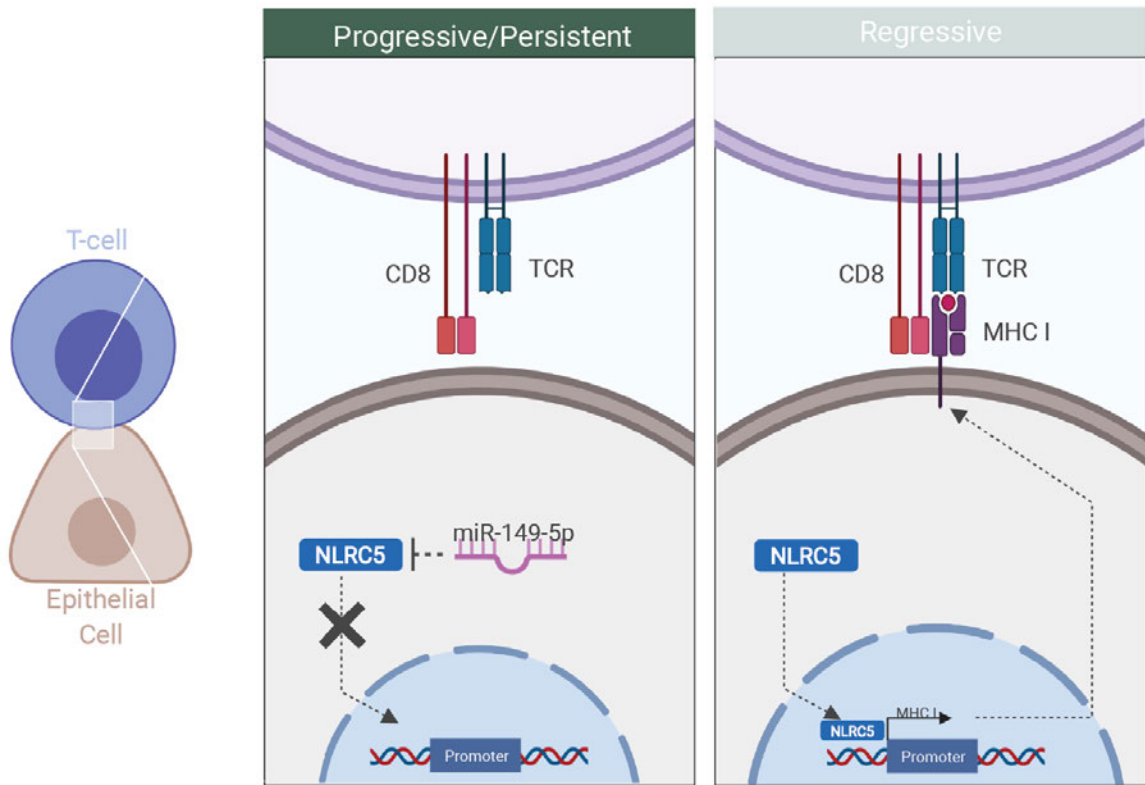


Figure 2.8. Summary diagram.

149-5p, and their association with early immune evasion and bronchial PML progression. Our *in vitro* experiments suggested that NLRC5 and MHC Class I genes are suppressed by miR-149-5p in bronchial epithelial cells, and are correlated with anti-tumor immune cell infiltration, similar to observations in the metastatic melanoma context<sup>230</sup>. In bronchial PMLs, a previous study reported loss of heterozygosity and hypermethylation in the HLA region led to dysfunctional antigen processing and presentation<sup>35</sup>. Thus, our work expands the current understanding and reveals an additional mechanism regulating antigen processing and presentation.

The association between NLRC5 regulated MHC Class I gene expression and tumor immune evasion have been extensively studied<sup>218,219,231</sup>. Reduced NLRC5 expression can lead to decreased T cell cytotoxicity in mice<sup>232,233</sup>. Using the solid tumor data from TCGA, Yoshihama *et al.* demonstrated that lower NLRC5 expression is associated with deficient CD8<sup>+</sup> T cell activation and poor prognosis<sup>234</sup>. Furthermore, recent work demonstrated that impaired MHC Class I processing and presentation pathway confers lung cancer resistance to immune therapy<sup>235</sup>, and increasing MHC gene expression can improve immune checkpoint blockade response<sup>236</sup>, suggesting their critical roles and therapeutic potential in the context of lung cancer management. Interestingly, Ayukawa *et al.* recently demonstrated that MHC class I in normal epithelium may facilitate the removal of premalignant cells through interacting with LILRB3 independently from NK or CD8<sup>+</sup> T cell cytotoxic activities<sup>237</sup>. Collectively, these data suggest the critical and multi-faceted roles of miR-149-5p-mediated NLRC5 and MHC Class I gene expression associated with the immune microenvironment and the pathological progression of

bronchial PMLs.

Additionally, we showed that hsa-miR-149-5p is highly enriched among the bronchial airway epithelial cell populations, suggesting interesting crosstalk between the epithelium and the immune microenvironment. More specifically, the cell-type marker gene correlation analysis demonstrated that miR-149-5p is likely expressed within the basal cell, the progenitor stem cell, and the cell of origin of LUSC in the proximal airway<sup>238,239</sup>. Meanwhile, the proportion of basal cell is not associated with PML progression status, and adjusting for basal cell abundance does not affect the association between expression and progression status. Laughney *et al.* also demonstrated an inverse association between expression levels of the transcription factors specifying lung progenitor cells, SOX2, and MHC Class I genes in primary and metastatic lung tumors<sup>240</sup>. Similarly, the stem cell program in colon cancer has been associated with decreased levels of antigen presentation and elevated immune evasions<sup>241,242</sup>. These observations suggest that the cell state of basal cells in the bronchial airway may alter the immune microenvironment and drives the progression of bronchial PMLs. Therapeutics that direct cell fate among the airway cell populations might help control the lesions. Future investigation will be needed to further unravel the causality between epithelial cell state and the lack of antigen presentation, and determine whether immune surveillance may direct epithelial cell fates.

Our results also highlight the possibility of miR-149-5p being a potential therapeutic target for preventing PML progression. The miRNAs' ability to regulate the expression levels of multiple target genes makes them ideal candidates for therapeutics or as drug

targets<sup>243,244</sup>. Previous studies explored the therapeutic potential by modulating cancer-associated miRNA levels, either increasing the levels of tumor suppressive miRNAs or decreasing the levels of oncogenic miRNAs. For example, delivery of let-7 or miR-34a mimic with various strategies has shown to inhibit tumor initiation and growth in pancreatic<sup>245</sup> and lung cancer mice models<sup>246,247</sup>. In contrast, suppressing the functions of oncogenic miRNAs, such as miR-10b, using anti-miR (antisense oligomer) has demonstrated efficacy in reducing breast cancer metastasis<sup>248</sup> and glioma growth<sup>249</sup>. Despite these successful examples, only a handful of miRNA therapeutics for cancer have moved to the clinical trial stage. One challenge during miRNA therapy development was the accurate identification of miRNAs that specifically regulate a gene or pathway of interest, given that miRNAs can potentially regulate hundreds of target genes. Combining the network approach and in vitro assay, we have found that miR-149-5p targets and suppresses the NLRC5 expression in the airway epithelial cells. Another challenge of miRNA therapeutics is the delivery methods that ensure both miRNA stability and targeted delivery while minimizing toxicity. Recent advances in lipid nanoparticle platforms<sup>250</sup> may help to overcome these issues and to develop miR-149-5p anti-miRs as a bronchial lesion interception strategy.

Our results also reveal strikingly different miRNA-gene regulatory landscapes between the endobronchial biopsy and the mainstem bronchial brushing samples, despite the previously observed similar gene co-expression pattern. While the previous work demonstrated that the co-expression patterns of gene modules and the transcriptionally distinct molecular subtypes can be found in both lesion sites and normal airways<sup>39</sup>, the

miRNA-gene module networks indicate the lack of consensus connections between tissue or sample types. We hypothesized there are two main reasons for the discrepancy. First, the sample size of the biopsy dataset is almost twice as that of the brushing dataset, and the estimated miRNA-gene correlation might be more robust and stronger in the biopsy data. Therefore, some weaker regulatory relationships might not reach statistical significant levels and would be filtered out in the brushing miRNA-gene network. Second, the sample collection procedure and the tissue location differences may result in very different cellular composition between datasets. Biopsy of PMLs may be composed mainly of epithelial basal cell, while the mainstem bronchial airway brushings would have more ciliated cells (**Figure 2.7d**). The difference in cellular composition has been associated with miRNA differential expression and regulatory roles<sup>251</sup>. Notably, we found miR-34/449, two cilia cell specific miRNAs<sup>213</sup>, were not connected to the cilia biogenesis gene module in the brushing miRNA-gene module network as in the biopsy network. These results highlight the potential that a similar co-expressed gene modules or gene programs might be associated with different cell types between sample types. Our data reveal a cell-cell communication model where the upregulation of miR-149-5p in basal cells represses the anti-tumor immune response associated with the progression of bronchial PMLs. The results not only suggest a novel mechanism that drives the immune-evasive microenvironment but also provide a potential biomarker measured from biopsy samples for the detection of progressive PMLs and candidate therapeutic targets. To our knowledge, this is the first study analyzing miRNA-mediated gene expression regulation associated with premalignant lesion progression. Future studies are

needed to comprehensively characterize the alterations in both the epithelial and immune cell populations, including single-cell sequencing to address cellular composition changes associated with PML histological progression, targeted TCR/BCR sequencing to address how neoantigens shape the adaptive immune cell populations, and spatial transcriptomic to quantify how the cellular organizations and basal membrane structure are changed during early lung cancer progressions. Putting everything together, our data reveal miRNA-mediated gene regulation that contributes to early immune-evasion and PML progression and supports the potential of intercepting the progression of PMLs through modulating the miR-149-5p levels.

## CHAPTER 3 DIFFERENTIAL REGULATION ANALYSIS QUANTIFIES MIRNA REGULATORY ROLES AND CONTEXT-SPECIFIC TARGETS

*Adapted from the following manuscript:*

B Ning, Spira T, JE Beane and ME Lenburg. Differential regulation analysis quantifies miRNA regulatory roles and context-specific targets. *bioRxiv*.

### **3.1 INTRODUCTION**

Gene expression regulation within cells can be modeled via networks, where the transcriptional regulators, such as transcription factors (TFs), and their downstream target genes are represented as nodes and regulatory relationships are represented as edges<sup>252,253</sup>. Through tight control of the transcriptional network, cells can accurately coordinate gene expression and establish correct cellular functions under normal conditions. In the meantime, the gain or loss of connectivity in the transcriptional network, or “rewiring” events, can result in altered gene expression profiles observed among differentiating or treated/perturbed cells or between cancer-subtypes<sup>65,254–256</sup>. Computational methods have been developed to detect and statistically quantify transcriptional network rewiring events between two groups from bulk gene expression profiles. These methods utilize gene-gene expression correlation coefficients, graphical models, or Latent Dirichlet allocation with TF chromatin-binding profiles to infer the transcriptional regulator (typically TFs) whose connectivity with downstream targets is significantly rewired<sup>257–260</sup>.

MicroRNA (miRNA) is a class of short, non-coding RNAs which utilizes complementary sequence pairing between its seed sequence and the 3' untranslated region (UTR) of gene

transcripts to repress target gene expression levels<sup>261–263</sup>. Through acting as post-transcriptional regulators, miRNAs participate in a wide range of biological and cellular processes, including cell differentiation, development, and carcinogenesis processes<sup>72,251,264,265</sup>. It has been shown that miRNA may undergo network rewiring and have different functions between cell-types or cancer molecular subtypes<sup>67,69,70,266</sup>.

Particularly shown by the study from Hsin *et al.*, the rewiring of miRNA regulatory networks between cell-types is predominantly due to miRNA target binding switching, rather than 3'UTR isoform or expression levels<sup>267</sup>. While these studies provided valuable knowledge on miRNA regulatory networks, they often rely on single-cell RNA sequencing or complicated functional profiling such as CLIP-seq with Halo-enhanced Ago2 pull-down<sup>268</sup> or CLASH<sup>269</sup>, which is expensive and not always feasible. Thus, it is advantageous to develop methods capable of identifying miRNAs with differential regulatory roles across either multiple cell-types or cancer subtypes from bulk miRNA and gene expression profiles.

While Chapter 1 explored the miRNAs that regulate specific gene modules and drive the progressiveness of proliferative lesions, less is known on how miRNA-mediated regulation may contribute to the molecular subtypes. Given that the molecular subtypes were originally defined based on gene co-expression<sup>39</sup>, miRNA differential expression analysis between molecular subtypes would not yield further insights since the target genes of the resulted miRNAs will simply be the module defining genes. However, evidence showed the molecular subtypes were enriched with different epithelial cell populations, which suggests differential miRNA-mediated gene regulatory network or

miRNA network rewiring may exist across the molecular subtype. Furthermore, the regressive proliferative lesions were shown to be transcriptionally similar to the inflammatory samples, much less high-grade dysplasia lesions were classified as the inflammatory subtype. Thus, identifying miRNAs with subtype-specific target genes may provide further biological mechanism on the development of the molecular subtypes. Most current computational methods for detecting rewiring events, are not optimized for miRNA-mediated gene regulation and cannot be directly applied to scenarios involving miRNAs for several reasons. 1) Previous methods were developed to build gene-gene networks (all nodes connected and associations between nodes could be either positive or negative), but did not include the miRNA predicted target information nor the miRNA gene repression activity in the computational model; 2) Previous methods outputted either single miRNA-gene connection or single module consisted of multiple interconnected miRNAs/genes rather than potential miRNAs; 3) Previous methods focus on comparisons between two groups, whereas when comparing between cell-types or cancer molecular subtypes, researchers are often dealing with more than two groups. Here, we present a novel computational framework and R package *Differential Regulation Analysis of miRNA* (DReAmiR; <https://github.com/ningb/DReAmiR>) to address the aforementioned challenges. Integrating mRNA and miRNA expression profiles with the predicted target information, DReAmiR first estimates miRNA-mRNA correlation matrices per group while removing the bias from the mean-correlation relationship<sup>270</sup>. Then, it identifies miRNA with significantly context-specific targets across multiple experimental groups based on differential enrichment of predicted targets

between groups using genes ranked by mRNA / miRNA correlation coefficients. Finally, DReAmiR can prioritize target genes uniquely regulated in each group using either a graph embedding-based approach or iteratively maximizing the difference in enrichment score (dES).

## **3.2 METHODS**

### *3.2.1 DReAmiR*

The DReAmiR workflow is divided into three main steps (**Figure 1**): (1) Remove mean-correlation relationship using spatial quantile normalization (SpQN), (2) Identify miRNAs with significant context-specific targets with differential regulation analysis, and (3) Prioritize the target genes specifically regulated by a miRNA in each group.

*SpQN*. To remove the bias from gene/miRNA expression levels on correlation estimations (the mean-correlation relationship) in the miRNA-gene correlation matrix, we modified the SpQN algorithm developed by Wang Y. *et al.* (2020) to accommodate an asymmetrical matrix<sup>270</sup>. Briefly, a Pearson correlation matrix is first constructed for each experimental group between miRNAs and genes and sorted by miRNA and gene expression levels. Then, the correlation matrix is separated into non-overlapped submatrices based on gene and miRNA expression levels. Next, for each non-overlapped sub-matrix, a larger overlapping matrix is used for estimating the empirical correlation coefficient distribution. The number of bins for genes and miRNAs and the overlap size can be specified by the user. By default, DReAmiR separates the correlation matrix into 20x20 submatrices and constructs overlapping matrices with 1000 genes and 150

miRNAs each, which are suitable for typical RNA and small-RNA sequencing experiments in which about 10-20 thousand genes and around one thousand miRNAs can be detected. These parameters are chosen to balance the smoothness in normalization and the total running time. Using the sub-matrix close to the top right corner as reference (by default, the sub-matrix corresponding the second-highest miRNA and gene expression), 1-dimensional quantile normalization is applied to match the correlation density distribution of all other bins. Further information on parameter tunings for SpQN can be found in Wang Y. *et al.* (2020)<sup>270</sup>.

*Differential regulation analysis.* After correcting for the mean-correlation relationship, differential regulation analysis can be performed to identify miRNAs with significant context-specific target genes. This step requires two inputs: the correlation matrices for each group and the predicted target genes of miRNAs. For each miRNA, the genes are first ranked by their normalized correlation coefficients to the miRNA, from negative to positive, to generate a gene rank list for each group. Then, the enrichment pattern of predicted target genes in each rank list is compared using the Anderson-Darling (AD) test. Permutation of group label is performed and the observed test statistic is compared to the distribution from permutations to generate empirical p-values. When only two groups are present, Kolmogorov-Smirnov (KS) test is used instead. Post-hoc analysis using the KS test can be performed afterward to examine pair-wise significance.

*Target Prioritization.* For the miRNAs whose regulated genes are significantly different between groups, DReAmiR can further prioritize their target genes in the leading edge of each group based on association strength, assuming that being in the leading edge

suggests a gene is strongly regulated by the miRNA in at least one group. DReAmiR provides two methods for this purpose.

(1) The graph embedding and node clustering method aims to classify and label the target gene clusters as either group-specific or shared between groups. First, a network is constructed with nodes representing genes and group-specific miRNAs. An edge between a gene node and a group-specific miRNA node represents the gene being in the leading-edge of that group based on differential regulation analysis, with the association strength as the edge weight. Then, the large information network embedding (LINE) algorithm<sup>271</sup> is used to identify embedded features. We chose LINE over other network embedding methods due to its scalability and ability to preserve both local and global network structures. Fuzzy k-means clustering is performed within the embedding space to cluster the target genes. Clusters containing one or more group-specific miRNA nodes are labeled accordingly. For each of the remaining clusters, a one-tail KS test is performed to compare whether the node membership density associated with each of the labeled clusters is significantly higher than the other labeled clusters, and the cluster labels are then assigned based on the number of significant results. A cluster with membership density significantly higher for one labeled group is labeled as a uniquely regulated target cluster for that group. A cluster with more than one significant result is labeled as a shared target cluster. Cluster without any significant result is a common target gene cluster regulated equally across all groups.

(2) Difference in enrichment score (*dES*) maximization method aims to find the set of target genes that are most strongly regulated in the group of interests compared to the

reference group. The reference group can either be another group that the user wants to compare to, or all other groups combined. Starting with all the genes in the leading edge in the group of interests (notated as  $\mathcal{X}$ ) as the baseline target set, DReAmiR first calculates the  $ES_{normalized}$ , which is the observed enrichment score divided by the mean of permutation  $ES$  values from shuffling the rank list 500 times. Then,  $dES$  is derived as  $ES_{normalized}^{\mathcal{X}} - ES_{normalized}^{Reference}$ . Then, each predicted target gene in the baseline target set is removed one at a time and the changes in  $dES$  are calculated. The gene that yields the largest increase in  $dES$  is then removed to form the new target set. The process is repeated until no gene remains in the target set or the minimum target size specified by the user is reached (by default 20). The  $dES$  calculated at each step is recorded and back-traced. The set of genes giving the largest  $dES$  is selected as the final target set for the group of interest, which yields the largest separation in enrichment score compared to the reference group. Of note, the direction of enrichment to test, either positive or negative, can be selected by the user and the calculation in  $ES$  calculation is changed accordingly.

### 3.2.2 Other Functions

To facilitate the workflow, DReAmiR also contains several utility functions:

*Correlation matrix and predicted target gene matrix construction.* To simplify the workflow and ensure the reproducibility of the analysis, DReAmiR only requires three simple inputs: gene expression matrix, miRNA expression matrix, and the group label. DReAmiR examines the format of these inputs and returns group-specific correlation matrices for the samples with complete records. In addition, with the help of multiMiR

Bioconductor package<sup>272</sup>, users can specify which and the number of miRNA target databases to use. The intersection of databases will be used for miRNA-gene target information. For example, if a gene is recorded as the target of a miRNA by 3 out of the 5 miRNA target databases, it will be retained for down-stream analysis. DReAmiR directly generates a binary target matrix for miRNA-gene pair with the same size as the miRNA-gene correlation matrix.

*Visualization.* The output from the differential regulation analysis can be visualized as a multi-group enrichment plot (**Fig 2a** and **Fig 3b**). The plot contains two panels: the top panel depicts the enrichment patterns and the enrichment score per group, and the bottom panel depicts the position of predicted target genes in each rank list.

*Parallelization.* DReAmiR contains several computationally intensive tasks, including the group-wise correlation matrix calculation, the SpQN, and max-dES with iterative searches. To reduce the analysis time, we implemented parallelization for these steps when users have access to multi-core computation.

### 3.2.3 Data Simulation

To evaluate the performance of AD test used by DReAmiR in detecting miRNAs with different regulatory roles across multiple groups, we generated sample matched gene and miRNA expression where the predicted target genes of a miRNA have different patterns of negative enrichment along the gene rank list by miRNA-gene correlation strength. The simulation dataset contained three groups of samples with expression profiles of 1000 genes and 20 miRNAs. 10 miRNAs were assigned to have context-specific target genes

between groups (true positive) and 10 to be not different (true negative). For each miRNA, we first randomly picked  $N_{target}$  genes to be the predicted targets, mimicking the information from miRNA predicted databases. The covariance between the predicted target genes and miRNAs was simulated by a normal distribution with negative mean  $\mathcal{N}(-0.5, 1)$  representing globally suppressive effects on gene expression by miRNAs. Then, for those miRNAs assigned to be true positive, we randomly picked  $P_{variable\%}$  percent of the predicted targets per group to be the group-specific target and should be negatively regulated more strongly within that group. The association between group-specific target genes and miRNA was simulated by shifting the covariance towards the negative direction proportionally from the mean within two groups with a left-skewed beta distribution  $\text{Beta}(\alpha, \beta)$  where  $\alpha = 20$  and  $\beta \sim \mathcal{U}(1, 5)$ . Next, the nearest positive-definite of the covariance matrix is calculated using the ‘make.positive.definite’ function from `lqmm` R packages<sup>273</sup>. The gene and miRNA expression matrices were then simulated for  $N_{sample}$  per group by multivariate Gaussian distribution  $\mathcal{N}(\mu = 0, \Sigma)$  using ‘`mvrnorm`’ function from the `MASS` R package<sup>274</sup>. Finally, the miRNA-gene correlation matrix was calculated and exported along with the predicted targets per miRNA as the input for performance evaluation.

Evaluation of DReAmiR performance was conducted by simulating datasets with the following parameters:

$$N_{target} = \{40, 60, 80, \text{and } 100\}$$

$$P_{variable\%} = \{60, 70, 80, \text{and } 90\}$$

$$N_{sample} = \{30, 50, \text{and } 70\}$$

With each combination of simulation parameters, we ran the simulation 20 times to obtain confidence intervals. Differential regulation analysis was run using default parameters. As a comparison, we also evaluated the performance of two other methods using the same simulated dataset: (1) comparing the correlation density between groups using ANOVA, and (2) summarizing the p-values with Edgington's method<sup>275</sup> implemented in the *metap* R package<sup>276</sup> from two-group KS-test comparisons. All tests were performed with a significance value of  $p\text{-value} < 0.05$ . The performance metrics between methods were compared using two-group Student's T-tests.

#### *3.2.4 Cell-type-specific mmu-miR-155 KO RNA-seq data*

The gene count tables for primary dendritic cells, B cells, CD4<sup>+</sup> T cells, and macrophages from C56BL/6J wild-type and miR-155 KO mice were obtained from GSE116348<sup>267</sup>. The raw gene count data was normalized first using the Trimmed Mean of the M-values and transformed into log<sub>10</sub> counts per million (logCPM)<sup>197,277</sup>. Then, genes with mean expression across samples equal to or less than 1 and interquartile range (IQR) equal to 0 were removed. The filtered count table with 10843 genes was TMM normalized again before conducting differential expression analysis.

Differential expression analysis was conducted within each cell-type separately. Within samples from each cell-type, the normalized expression data were first voom transformed<sup>195</sup>. Gene association with KO treatment was calculated by comparing the samples from the mmu-miR-155 KO group to the WT sample group using *limma* R package<sup>197</sup>. Genes were ranked by their association with mmu-miR-155 KO using t-

statistics to generate cell-type specific rank lists. Signature genes were selected using log fold-change  $< 0$  and FDR  $\leq 0.05$ . TargetScan Mouse v7.1<sup>263</sup> was used to predict the target genes for mmu-miR-155-5p (the major mature miRNA from pre-miR-155). To perform the max-dES target prioritization for each cell-type, gene rank lists generated from the other three cell-types together were used as the reference. Functional enrichment analysis of the mmu-miR-155 target genes was conducted using the hypeR R package<sup>278</sup>.

### *3.2.5 Breast cancer molecular subtype analysis*

miRNA and gene expression data and breast cancer molecular subtypes for TCGA BRCA samples were obtained using TCGAbiolinks<sup>198</sup>. We kept samples with both the miRNA and gene expression data from tumor samples and with PAM50 molecular subtypes (N=1064). miRNA and gene filtering was conducted using the same method described above, yielding 13846 genes and 584 miRNAs. Additionally, residual expression levels were calculated adjusting for the plate number using edgeR R package<sup>277</sup>, which were used for constructing correlation matrices per molecular subtypes. miRNA predicted target genes were queried from TargetScan v7.2<sup>263</sup> and only the conserved target sites were used.

RNA-seq profiles of hsa-miR-23b perturbation assays (either hsa-miR-23b over-expression or hsa-miR-23b sponge vector transfection) and controls in MCF-7 and MDA-MB-231 cell lines were obtained from GSE37918<sup>279</sup>. The rank lists were generated using fold-change in each cell-line by comparing to the corresponding controls and signs were set that the negative association indicated a gene being regulated by miR-23b.

### *3.2.6 Analysis of biopsies and brushings from patients with PMLs across molecular subtypes*

The matched gene and miRNA sequencing data of endobronchial biopsies (N=148) and mainstem bronchial airway brushings (N=82) from the previous chapter was used for the analysis. The residual expression values were used as input for the SpQN before running the differential regulation analysis. miRNA predicted target genes were queried from TargetScan v7.2<sup>280</sup> and only the conserved target sites were used. Finally, we prioritized the target genes for the Proliferative or the Inflammatory subtypes using the other subtype as reference group, respectively.

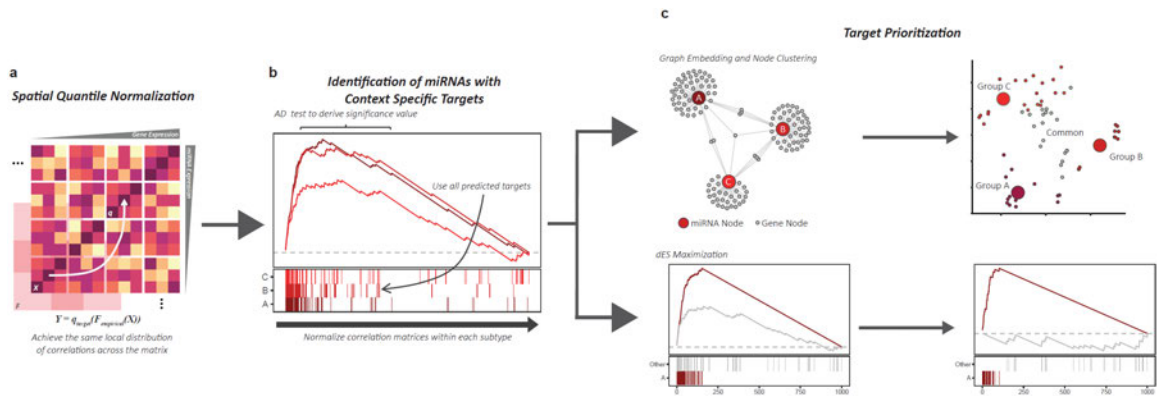
### *3.2.7 Data and Code Availability*

All data used in this project were publicly available. DReAmiR is an R package and can be downloaded from <https://github.com/ningb/DReAmiR>.

## **3.3 RESULTS**

### *3.3.1 DReAmiR method overview*

A toy example is described in **Figure 3.1** to outline the basic workflow of the DReAmiR package. With the user-provided miRNA and gene expression matrix and the group label, DReAmiR computes the miRNA-gene correlation matrices for each group. DReAmiR also generates the corresponding miRNA-target gene matrix for the miRNAs and genes based on the intersection of miRNA target databases that the user specifies. SpQN can be



**Figure 3.1. DReAmiR method overview.**

DReAmiR consists of three major steps. **a.** SpQN removes the mean-correlation relationship in the miRNA-gene correlation matrices. **b.** The differential regulation analysis identifies miRNAs with context-specific target genes by comparing the ranking of miRNA-gene correlation coefficients within each group. **c.** Two methods were developed to prioritize experimental group-specific target genes.

performed as an optional step to adjust the correlation estimations and remove the mean-correlation relationship (**Figure 3.1a**). Next, DReAmiR performs differential regulation analysis for each miRNA using AD test (**Figure 3.1b**). By default, the genes are ranked by their correlation coefficients with miRNA within each group based on the correlation matrices generated in the previous step. Alternatively, the user can construct the gene rank list manually using other metrics such t-statistic from perturbation experiments. By comparing the negative enrichment pattern of target genes along the gene rank list by their correlation with the miRNA across groups rather than simply comparing the correlation densities, DReAmiR better captures the miRNA rewiring events and is less affected by the sample sizes. Parallel computation is implemented to speed up the process when multiple cores are available. An enrichment plot can be plotted to help visualize the difference in the enrichment patterns of target genes between groups. In the example shown in **Figure 3.1b**, the target genes of this miRNA were strongly negatively enriched in group A and B, but less in group C.

After getting a list of miRNA with significant context-specific targets, it might be important to identify group specific target genes for each miRNA. DReAmiR provides two independent methods to address this target prioritization step (**Figure 3.1c**). The graph embedding and node clustering method focuses on all predicted target genes in the leading-edge and aims to label each target as specifically regulated in one group, shared by multiple groups, or common across all. In the given example, four clusters of target genes were identified for the miRNA, where three of them were group-specific and one was shared across all groups. The max-dES method focuses on one particular group and

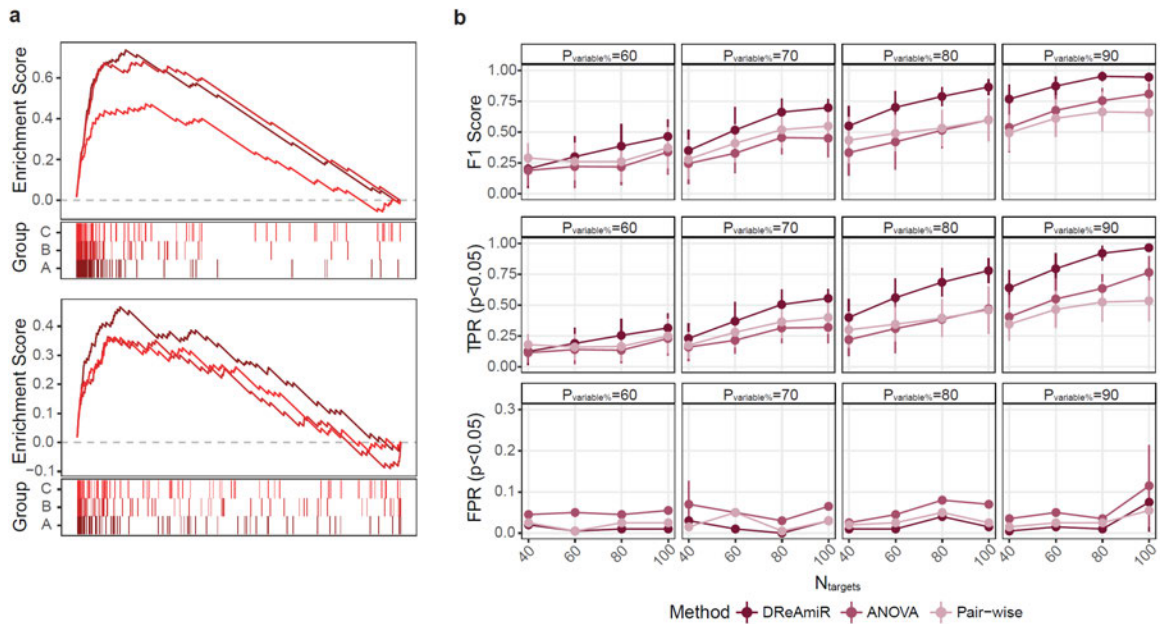
aims to find the set of predicted target genes that generate the largest difference between this and the reference group. Assuming the group A was the group of interests, we combined the group B and C to generate the background rank list. While the resulted group-specific target genes from two methods could be overlapped, the user can choose which method to use depending on the specific biological question and study design.

### 3.3.2 Benchmark on simulation data and comparison vs. other methods

Biological network algorithms are typically tested using ground truth data where the edges between individual nodes are known and experimentally validated. Yet, no such dataset is available for miRNA differential regulation and a method for similar purposes has not been developed to the best of our knowledge. Thus, we aimed to evaluate DReAmiR based on its robustness and the biological plausibility of its results.

To evaluate the ability of AD test used by DReAmiR to identify miRNAs with context-specific target genes across multiple groups, we simulated gene and miRNA expression data with both true positive and true negative miRNAs and predicted target gene information 20 times for each parameter combination (**Figure 3.2a**). Then, we performed differential regulation analysis using AD test with default settings. AD test demonstrated good performance under most scenarios, and the F1 score and TPR increased with  $N_{sample}$  and  $N_{target}$  (**Figure 3.2b** and **Fig A.6**). Among these three parameters, the TPR and F1 score of AD test performance were most strongly affected by the  $P_{variable\%}$ .

Based on observations from Hsin JP *et al.*<sup>267</sup>, the majority of predicted targets are group-specific, suggesting AD test should perform well in realistic settings, especially when



**Figure 3.2. DReAmiR performance in simulated data.**

**a.** Enrichment plots for simulated true positive (top) and true negative (bottom) differentially regulating miRNAs. **b.** Performance comparisons between DReAmiR, ANOVA and pair-wise comparison in the simulated data with sample size per group equalled 70. The simulation was performed across different simulation parameters, including the number of predicted target genes per miRNA, and the percentage of variable targets per group. F1 score (F1), true positive rates (TPR), and false positive rates (FPR) across 20 iterations per combination of simulated parameters were evaluated by the mean and standard deviation.

$N_{sample}$  and  $N_{target}$  are large enough. In the meantime, the FPR remains low (most less than 5%) regardless of the simulation parameters tested.

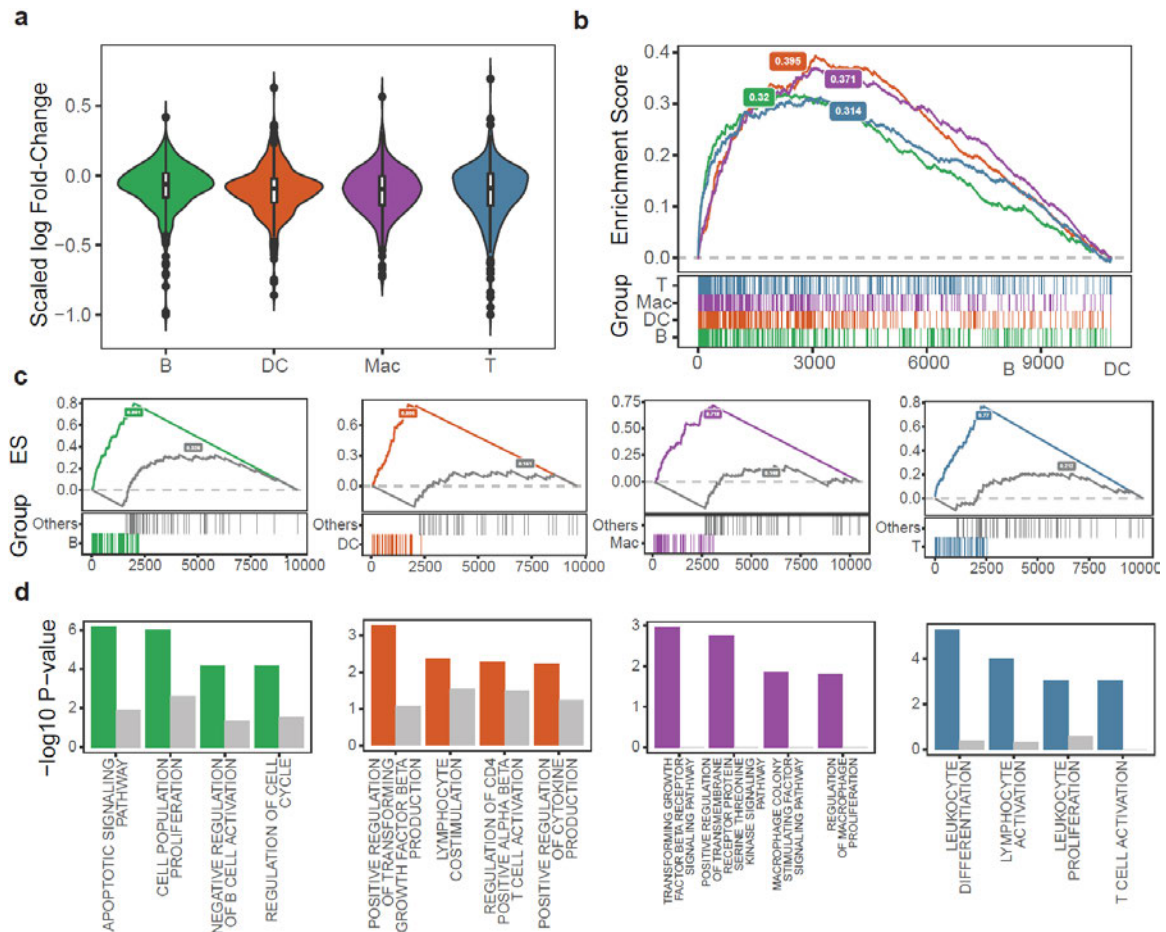
In addition, we also evaluated the performance of two alternative strategies in detecting miRNA regulatory rewiring across multiple groups, using the same simulated dataset.

First, we used ANOVA to compare the correlation density distribution of predicted target genes for each miRNA, representing the underlying model of general gene-gene network-based algorithms, such as DGCA<sup>257</sup>. Second, we used the pair-wise group comparison function in DReAmiR based on KS-test and summarized the p-values using Edgington's method, mimicking methods that only allow two-group comparison. Under all simulated parameters, AD test outperformed these two methods (**Figure 3.2b** and **Fig A.6**). More specifically, the F1 score and TPR from AD test were significantly higher than those from ANOVA or p-value summation when  $P_{variable\%}$  or  $N_{target}$  is large (**Table B.6**). Notably, the FPR from the ANOVA method is generally significantly higher than from AD test. In contrast, while the p-value summation method can achieve relatively low FPR, the F1 score and TPR are much lower than AD test. Similar difference in performance between the three methods was observed, where the sample size was different across groups (**Figure A.7**), suggesting that AD test is not strongly affected by the sample size bias in correlation estimation. The simulation results suggested that DReAmiR is better at identifying miRNA with context-specific target genes than other methods and can achieve good performance under reasonable conditions (sample size, target number, and effect size) while maintaining high specificity.

### 3.3.3 DReAmiR identified mmu-miR-155 cell-type-specific functional pathways

Having benchmarked DReAmiR in simulated data, we next sought to evaluate whether DReAmiR can identify miRNA known to regulate different targets in different contexts. We picked mmu-miR-155 as an example since the rewiring of the mmu-miR-155 regulatory network across immune cell types has been clearly demonstrated, and its functions in immune cells have been extensively studied<sup>267,281,282</sup>. Using all mmu-miR-155 predicted target genes as the target set (N=430) and the average logFC from differential expression analysis (comparing samples in the WT to the KO group) for generating the rank list per cell-type, differential regulation analysis was performed and the rank list from each cell-type was shuffled 500 times to calculate the permutation p-value. mmu-miR-155 regulated different target genes across four immune cell-types (**Figure 3.3b**; Permutation p-value = 0.01). In contrast, the average logFC distribution of mmu-miR-155 predicted target genes were not significantly different between four immune cell-types (**Figure 3.3a**; ANOVA p-value=0.08). These observations suggest DReAmiR may discover miRNAs with context-specific targets that are not captured by comparing average association strength.

We then used the max-dES method to prioritize group-specific target genes and examine whether the prioritized target genes were associated with cell-type-specific functions. The max-dES yielded 55, 40, 43, and 58 predicted target genes for B-cell, dendritic cell, macrophage, and CD4 T-cell, respectively, when each cell-type was compared to the other three as reference (**Figure 3.3c**). Target gene expression per cell-type was not strongly associated with whether a target gene was prioritized for a cell-type (**Fig A.8**) as



**Figure 3.3. DReAmiR identifies miR-155 target genes involved in functional pathways with cell-type specificity.**

**a.** Enrichment plot of miR-155 predicted target genes in the gene rank list sorted by logFC in miR-155 KO samples comparing to the controls within each immune cell-types (Permutation p-value = 0.01). **b.** logFC densities of miR-155 predicted target genes in miR-155 KO samples compared to the controls within each immune cell-types (ANOVA p-value = 0.08). **c.** Enrichment plots for the miR-155 prioritized target genes per immune cell-type, using the max-dES method. For each cell-type, the gene list ranked by logFC associated with miR-155 in the other three cell-types compared to the were used as the reference group. **d.** Functional pathway enrichment results for the prioritized miR-155 target genes per cell-type (colored), and the differentially expressed miR-155 target genes following miR-155 KO (grey).

previously suggested<sup>267</sup>, indicating the cell-type specific regulatory behavior of miR-155 was not solely determined by the target gene expression level. Notably, the rankings of the prioritized target genes overlapped between the group of interests and the reference group, meaning max-dES did not simply prioritize the top-ranking target genes.

Functional pathway enrichment analysis was then performed on the prioritized target genes and the known pathways related to mmu-miR-155 cell-type-specific functions were among the top significantly enriched pathways (**Figure 3.3d**). For example, proliferative and apoptotic signaling pathways in B cells<sup>283,284</sup>, cytokine and co-stimulation related pathways in dendritic cells<sup>285,286</sup>, and activation and differentiation pathways in CD4+ T cells<sup>287,288</sup>. However, such enrichment was much weaker for the differentially expressed ( $\log_{2}FC < 0$  and  $FDR \leq 0.05$ ) miR-155 target genes (N=36, 110, 6 and 13). It is also worth noting that while it may be possible to alter the threshold for differential expression analysis to achieve similar functional enrichment results, no parameter tuning was needed for DReAmiR to generate such results. Taken together, this evidence suggested that DReAmiR can better identify miRNA rewiring events and the targets prioritized from DReAmiR are biologically informative.

#### 3.3.4 DReAmiR identified BRCA subtype-specific targets from bulk RNA-seq data

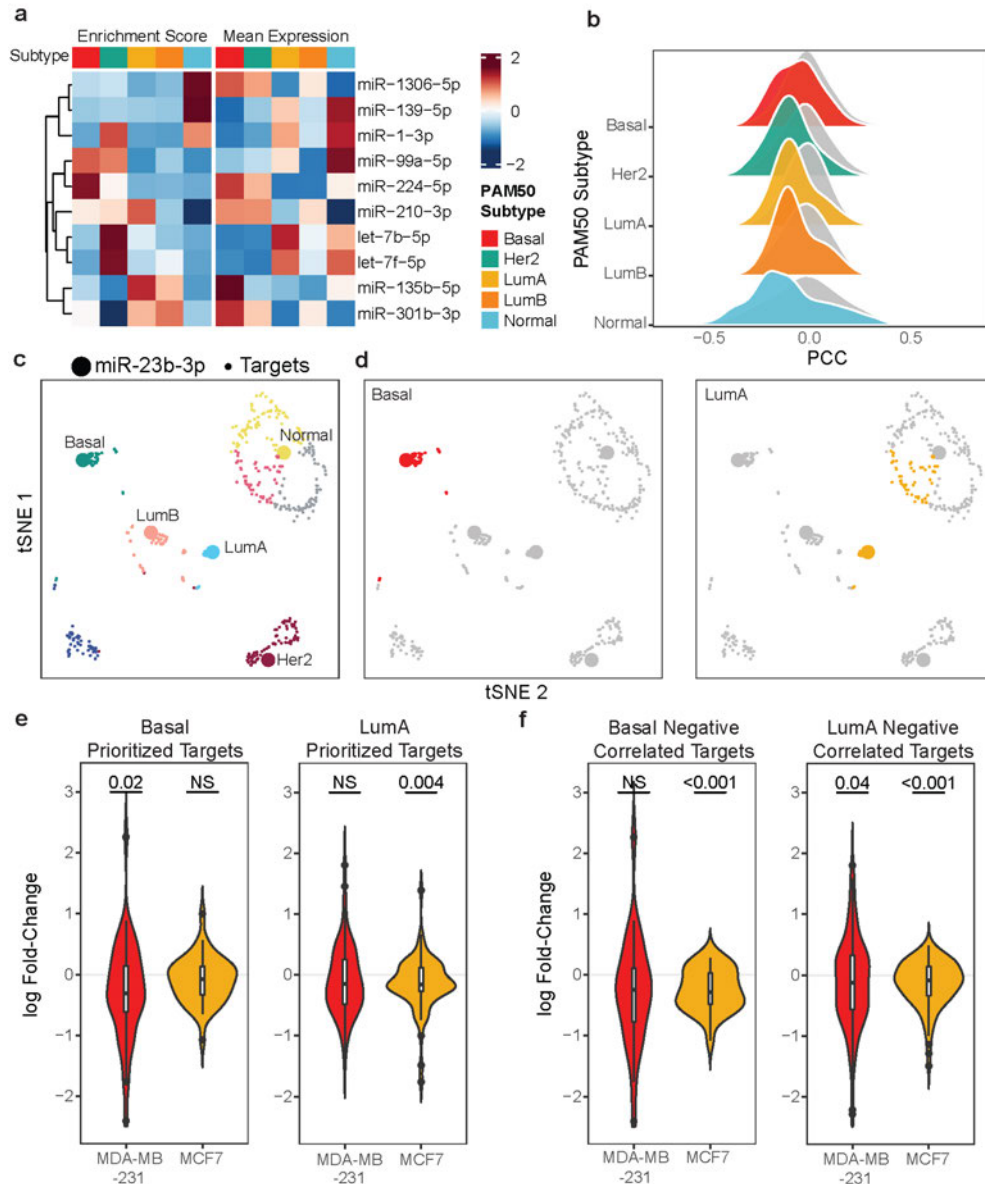
We next sought to demonstrate DReAmiR in a realistic use case, where researchers want to identify miRNAs with significant context-specific targets from sample-matched bulk miRNA and mRNA expression profiles between cancer subtypes, and validate the candidates through *in vitro* experiments. We performed the differential regulation

analysis in TCGA BRCA data between five PAM50 breast cancer subtypes<sup>289,290</sup>, which yielded 10 miRNAs with significant context-specific target genes (**Figure 3.4a**; Permutation FDR  $\leq 0.1$ ). The mean expression levels of these miRNAs within each subtype showed very different patterns compared to their ES values, suggesting context-specific target gene regulation is not directly driven by differential expression levels. Then, we prioritized subtype-specific target genes for each molecular subtype using the graph embedding and node embedding method. The expression correlation densities between miRNA and the prioritized target genes for each subtype were significantly lower among the samples of each subtype compared to all other subtypes (**Figure 3.4b**; one-tail KS test, p-values  $< 0.01$ ), suggesting the target genes prioritized by DReAmiR were specific to each subtype.

To support that the prioritized target genes are indeed altered in a subtype-specific fashion, we examined whether the target genes prioritized for a breast cancer subtype were specifically altered with miRNA expression manipulation in the cell line of the same subtype. hsa-miR-23b was taken as an example since it is the only miRNA that we found perturbation assay with transcriptomic profiles<sup>279</sup> in cell lines of different breast cancer subtypes: MCF-7 for the luminal A subtype and MDA-MB-231 for the basal subtype<sup>291</sup>, even though hsa-miR-23b does not appear to regulate different targets in the different breast cancer subtypes in the TCGA BRCA data (permutation FDR = 0.48). Fuzzy k-means clustering on the graph embedding features derived from the TCGA BRCA data revealed eight different clusters. hsa-miR-23b in the five breast cancer subtypes were in separate clusters (**Figure 3.4c**). One cluster (42 predicted target genes)

was labeled as unique to the basal subtype, and two clusters (72 predicted target genes) were labeled to be associated with Luminal A subtype with one shared with the normal subtype (**Figure 3.4d**). Genes in the basal subtype cluster were enriched in oxidative phosphorylation and Kit signaling pathways, and those in luminal A subtype clusters were enriched in citrate cycle and interleukin signaling pathways (**Table B2**; hypergeometric test p-value < 0.01).

We next tested whether the targets of hsa-miR-23b that are differentially expressed in Basal and LumA breast cancer in the TCGA BRCA data show cell-line subtype-specific response to hsa-miR-23b overexpression *in vitro*. In the *in vitro* hsa-miR-23b overexpression experiment, the logFC of the basal-specific cluster genes were significantly lower than zero in MDA-MB-231, but not MCF7, suggesting these genes were specifically suppressed by hsa-miR-23b among the samples of the basal subtype (**Figure 3.4e**; one-tail t-test p-value < 0.05). A similar observation was seen for the luminal A specific gene cluster which were repressed by hsa-miR-23b overexpression in MCF7, but not MDA-MB31 (**Figure 3.4e**; one-tail t-test p-value < 0.05). In contrast, the logFC densities of the predicted target genes that were significantly negatively correlated with hsa-miR-23b (N=27) in the BRCA luminal A subtype samples were significantly lower than zero in both MDA-MB-231 and MCF-7 (**Figure 3.4f**; one-tail t-test, p-value < 0.05), while significant change was seen for the negatively correlated hsa-miR-23b predicted target genes in the BRCA basal subtypes only in MCF7 (one-tail t-test, p-value < 0.01) but not in MDA-MB-231. These observations highlighted the utility of DReAmiR for identifying cancer subtype-specific miRNA target genes and selecting



**Figure 3.4. The miR-23b prioritized targets from TCGA BRCA are uniquely altered by perturbation in the cell line of the same subtype.**

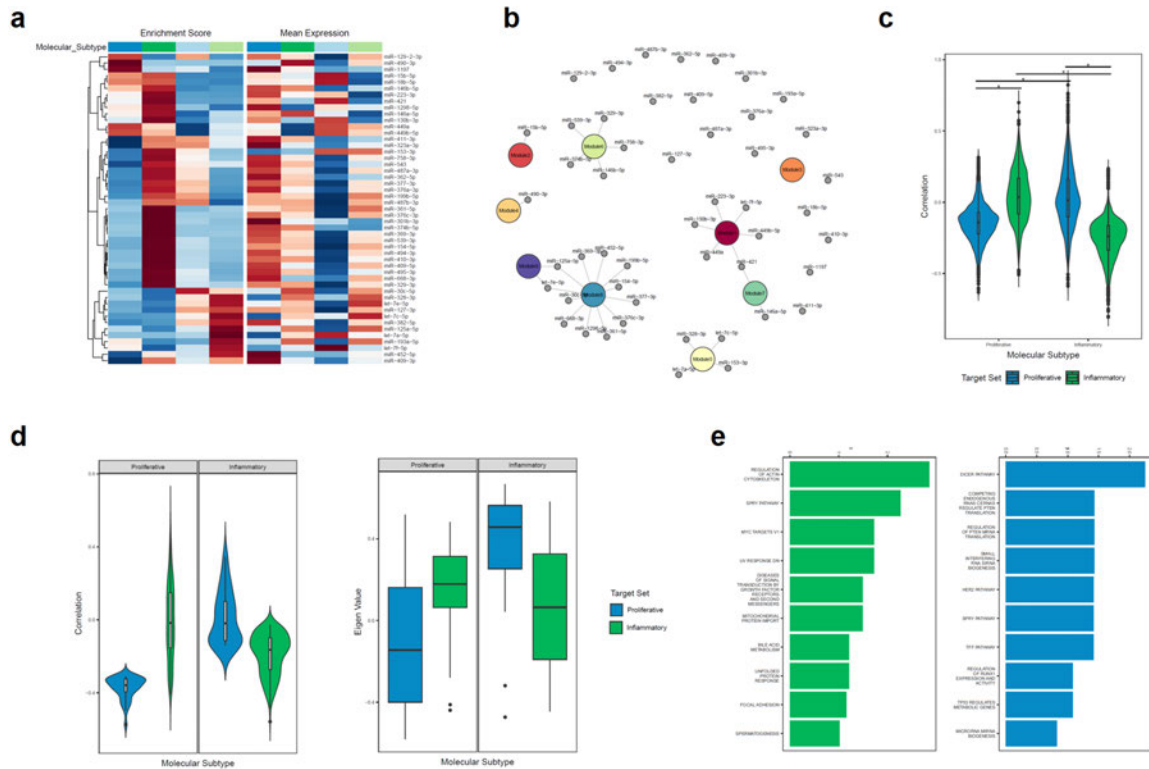
**a.** Heatmap of the ES and mean expression levels of differentially regulating miRNAs in TCGA BRCA data by the PAM50 subtypes. The rows were scaled and clustered by the ES values. **b.** Correlation densities between significantly differentially regulating miRNAs and their prioritized target genes for each subtype (colored) or the prioritized target genes for the other four subtypes (grey). Comparisons within all subtypes were significant (one-tail KS-test p-value < 0.05). **c-d.** tSNE plots were generated using 10 embedded features for the leading-edge genes from hsa-miR-23b differential regulation analysis. Points were colored by fuzzy k-means clusters (**c**) and subtype assignment to either basal or luminal A subtype (**d**). The number of genes assigned to Basal subtype=42, Her2=128, LumA=72, LumB=64, Normal=261. The size of the points depicted the node type. **e.** LogFC densities of hsa-miR-23b prioritized target genes in MCF7 and MDA-MB-231 perturbation assays. **f.** LogFC densities of predicted target genes that were significantly negatively correlated with hsa-miR-23b in TCGA BRCA RNA-seq data in MCF7 and MDA-MB-231 perturbation assays.

uniquely regulated miRNA target genes from bulk expression data for *in vitro* studies.

### *3.3.5 DReAmiR identifies miRNA with context-specific target genes between the Proliferative and Inflammatory PMLs*

To examine the miRNA whose context-specific regulatory pattern was associated with the PML molecular subtypes, we first performed differential regulation analysis in the endobronchial biopsy data. We identified 49 miRNAs with significant context-specific target genes (**Figure 3.5a**; AD-test, p-value < 0.05). The ES per subtype showed strong difference between proliferative and inflammatory samples and very different patterns comparing to the mean expression levels of miRNAs. Based on the miRNA-gene module regulatory network we derived in the previous chapter, many of these miRNAs were not connected to any gene modules, and the connections were not strongly enriched for any gene modules (**Figure 3.5b**), suggesting context-specific target regulation of miRNAs were not dependent miRNA expression values, and may yield additional insights than miRNA-gene module network analysis.

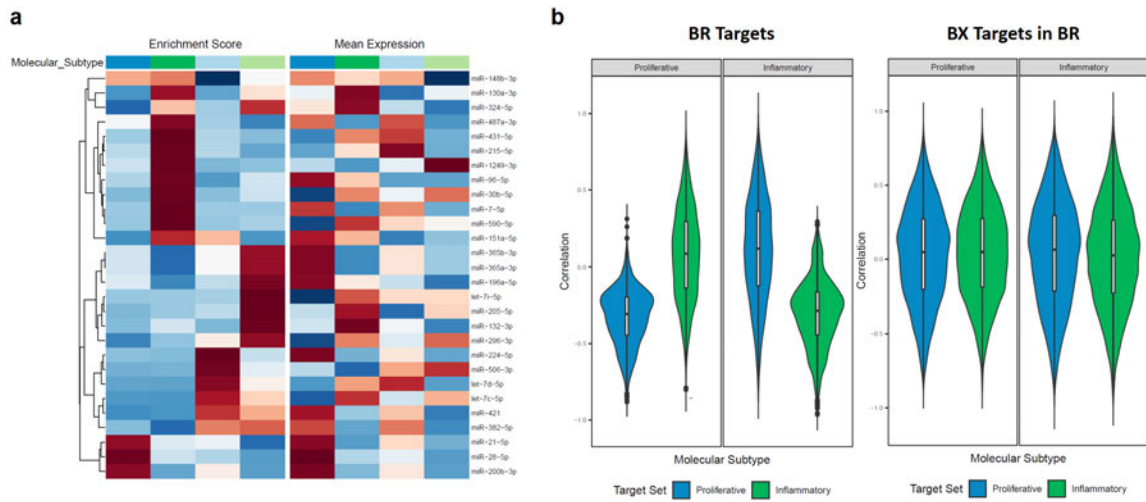
Next, we prioritized the target genes of these miRNAs using the max-dES method for the proliferative and the inflammatory subtypes separately, with the other subtype as the reference group. The correlation density between miRNA and the proliferative subtype-specific target genes were significantly lower than with the inflammatory subtype-specific target genes in the proliferative PML samples, and lower than with the same genes in the inflammatory PML samples (**Figure 3.5c**; one-tail Wilcoxon test, p-value < 0.001). Same differences were observed for the inflammatory subtype-specific target



**Figure 3.5. DReAmiR identified miRNAs with differential regulatory roles between the proliferative and inflammatory PMLs.**

**a.** Heatmap of the ES and mean expression levels of differentially regulating miRNAs in PML biopsy data by the molecular subtypes. The rows were scaled and clustered by the ES values. The top color bar indicated the molecular subtypes. **b.** The miRNAs with significant context-specific target genes in the previous derived miRNA-gene module network. **c.** Violin plot showing the correlation density between miRNAs with significant context-specific target genes and the prioritized target genes for proliferative or inflammatory subtype within the PML samples of each subtype. \*Wilcoxon sum rank test p-value < 0.01. **d.** The correlation density between miR-421 and the proliferative or inflammatory subtype prioritized genes (left) and the prioritized gene set metagene expression levels within samples of each molecular subtype. The violin and box plots were colored by the molecular subtype that the target genes were prioritized. The panels reflected the sample subtype among which the correlation or metagene expression levels were calculated. **e.** Top enriched functional pathways associated with the miR-421 target genes that were prioritized for the proliferative (right) and the inflammatory (left) subtypes.

genes, suggesting DReAmiR detect subtype-specific correlations and regulatory relationships between the PML molecular subtypes. Of these miRNAs, we highlighted the prioritized target genes of miR-421 to be particularly interesting. Not only did they show subtype-specific correlation patterns with the prioritized target genes, the metagene scores of the target also exhibited similar trend (**Figure 3.5d**). The target genes prioritized for the proliferative subtype were associated with RNA processing pathways, while genes prioritized for the inflammatory subtype were associated with various cell growth signaling pathways (**Figure 3.5e**), suggesting miR-421 may potentially suppress target genes of different functions that contribute to the distinct molecular subtypes. Next, we performed the DReAmiR workflow in the brushing samples collected from PML patients, and compared the results to that from analysis in the biopsy data. Differential regulation analysis showed 28 miRNAs with significant context-specific target genes (**Figure 3.6a**; AD-test,  $p$ -value  $< 0.05$ ), and 4 were also significant in the biopsy analysis. While the target genes prioritized in the brushing samples exhibited subtype-specific regulatory patterns in the brushing dataset as expected, the target genes prioritized in the biopsy samples showed similar correlation densities across subtypes in the brushing dataset (**Figure 3.6b**). These results suggested the context-specific target gene correlation patterns of miRNAs are not well conserved between bronchial data types, potentially due to the cellular compositions.



**Figure 3.6. miRNA differential regulatory patterns across molecular subtypes were different between endobronchial biopsy and mainstem airway brushing samples.**

**a.** Heatmap of the ES and mean expression levels of differentially regulating miRNAs in PML brushing data by the molecular subtypes. The rows were scaled and clustered by the ES values. The top color bar indicated the molecular subtypes. **b.** Violin plot showing the correlation density between miRNAs with significant context-specific target genes and the prioritized target genes for proliferative or inflammatory subtype within the PML samples of each subtype in brushing samples (left) or in biopsy samples (right).

### **3.4 Discussion**

Understanding miRNA regulatory network differences between cell types or disease states is essential to investigate the activities of miRNAs and their regulated gene programs. Existing methods were mostly developed for gene-gene networks, and fail to incorporate the predicted target information and the gene expression suppression nature of miRNAs. Furthermore, these methods often end at the rewiring detection step and do not quantitatively identify the group-specific targets of the regulators. In order to address these shortcomings, we created DReAmiR, a novel computational tool for characterizing miRNA-mediated regulatory network rewiring. First, DReAmiR calculates the miRNA-gene correlation matrix for each group and corrects the bias from the mean-correlation relationship using SpQN. Then, utilizing a GSEA-like model, DReAmiR detects miRNA with group-specific target regulation across multiple groups by comparing the negative enrichment patterns of the predicted target genes within gene rank list based on miRNA-gene correlation coefficients per group. Finally, DReAmiR identifies group-specific target genes through target prioritization.

Correlation estimation between miRNA and gene bulk expression profiles are affected by various technical factors, including expression levels and sample size. The accuracy of the differential regulation analysis depends on the ability to estimate miRNA-gene expression correlation free from such potential bias. Mean-correlation relationship was first described by Wang Y *et al.*<sup>270</sup>, and described a bias from the noise introduced during the high-throughput sequencing process that makes the absolute correlation coefficients estimated between highly expressed genes appear to be larger than those more lowly

expressed. DReAmiR adopts and modifies the Wang Y *et al.* method to remove such bias before running the differential regulation analysis, such that the expression level of miRNA and gene may have less effect on downstream analysis.

Through testing our method in simulated data, we showed using AD test, DReAmiR can achieve high performance under various conditions when correlation is used for ranking genes, particularly with a large number of predicted target genes and larger sample size, while maintaining low FPR. Compared with other strategies for modeling miRNA rewiring events, such as comparing correlation densities or summing p-values from two-group comparisons, the differential regulation analysis achieved better performance across different parameters settings, including various effect strength and sample sizes. Using expression and mRNA-binding profiles from mmu-miR-155 KO mice, Hsin P *et al.* showed that the regulatory network rewiring of mmu-miR-155 between immune cell-types is a result of switching of binding targets<sup>267</sup>. Furthermore, while all the predicted target genes are being suppressed to some degree, the cell-type-specific targets of mmu-miR-155 have a stronger association with mmu-miR-155 KO in that cell-type compared to all predicted targets. Hence, simply comparing the association strength, either based on fold-change in perturbation assays or correlation coefficient, between a miRNA and all of its predicted targets lack the sensitivity for detecting miRNA-related rewiring events. To better identify miRNAs with context-specific target genes, DReAmiR uses a framework similar to gene-set enrichment analysis (GSEA)<sup>204</sup>. One major task for GSEA is to examine whether genes of interest (the gene set) are concordantly positively or negatively enriched along a gene rank list, where the genes are ranked by their association with

certain phenotypes. In DReAmiR, we use an Anderson Darling's test and ask whether the predicted target genes of a miRNA (a target set) are similarly negatively enriched along multiple rank lists where genes are ranked by their association to a miRNA in each group. Applying DReAmiR to the RNA-seq data from Hsin P *et al*, we identified mmu-miR-155 as an miRNA with significant context-specific targets. Yet, simply comparing logFC densities of predicted target genes failed to detect the rewiring event. These results demonstrate that DReAmiR, by modeling the negative enrichment pattern of miRNA's predicted target genes in the correlation rank list across multiple groups, is usable and suitable for detecting miRNA regulatory network rewiring.

A typical task for studying miRNA is to identify top candidate target genes *in silico* and validate the findings through *in vitro* or *in vivo* assays. We demonstrated the utility of DReAmiR for such real-world scenario with the TCGA BRCA subtype analysis.

Conventionally, putative target genes are identified by setting a significance threshold to filter a set of genes through subtype-specific negative correlation analysis in bulk expression profiles. However, such a task may often fail because it does not inform whether the targets identified are specific to a group, and setting a hard threshold introduces unnecessary bias. Indeed, the target genes identified using this strategy did not show subtype-specific expression changes in cell lines with miR-23b perturbations. In contrast, DReAmiR does not set a specific hard threshold on the correlation coefficient nor does it significance level to filter the target genes. More importantly, the miRNA target genes identified through target prioritization for each cancer subtype were altered in a cancer subtype-specific fashion in cell line perturbation assays and were enriched for

different functional pathways. Notably, hsa-miR-23b regulation over several of the prioritized target genes, including PIL3R3 and PAK2, have been suggested in breast cancer settings<sup>279,292</sup>. Yet, the BRCA subtype specific regulation was not extensively studied and may inform future investigations. These observations, which were not found by traditional methods, highlighted the unique advantage of DReAmiR.

DReAmiR also revealed the differential miRNA regulatory roles between the Proliferative and the Inflammatory PMLs. In the previous chapter, we characterized miRNAs associated with the progression the histological grades of PMLs through correlation network across all PML samples. Here, we explored the differential regulation of miRNAs between molecular subtypes through examining the subtype-specific miRNA-gene correlation patterns. Particularly, we highlighted miR-421 whose molecular subtype prioritized target genes exhibited subtype-specific patterns of both correlation densities and expression levels. Furthermore, similar to the miRNA-gene module network analysis, we observed strong differences between the miRNA with context-specific target in biopsy and brushing samples, reflecting the potential effects of cellular composition between tissue types. These results may be utilized to understand the fundamental differences between PML molecular subtypes, and to potentially “treat” a subset of high-grade PMLs by directing them into lesions of low-risk subtypes.

DReAmiR offers two target prioritization methods to identify putative group-specific target genes for miRNAs: graph embedding and dES maximization. These two methods aim to answer similar questions but from different perspectives. Graph embedding aims to characterize the distance between the gene nodes and group-specific miRNA nodes in

the embedded space, and assign target genes to one or multiple groups based on cluster membership probabilities. Thus, it globally examines all target genes and all groups at once. This is useful when users are interested in all experimental groups or disease subtypes, or targets shared between groups, as we showed in the TCGA BRCA subtype analysis. dES maximization, on the other hand, tries to iteratively search for a set of target whose normalized enrichment score along the rank list of the group of interests most different from the background. The background can either be another group that users want to make contrast with, or all other samples. This can be used when samples from multiple groups are similar based on prior belief and the analysis is logical as a two-group comparison. For example, in the mice immune cell types analysis, we identified the target genes strongly associated with mmu-miR-155 KO in one cell type of interests, with samples from the remaining three closely related cell types combined as the background. This should be used when there is one group that is of particular interest to the users, such as a cancer-subtype with significantly worse prognosis comparing to others. While two methods often yield very similar results, we advise users to choose the method better suited to their experimental design and biological questions for the best interpretability. Meanwhile, on-going analysis will try to evaluate how much additional benefits the target prioritization will provide comparing to using the leading-edge genes from each group alone.

Although DReAmiR was primarily developed for miRNA analysis, it can be extended for other types of transcriptomic regulatory networks and groups other than cancer subtypes or cell types, as long as a target set and group-specific rank lists could be defined. Also,

we included a user-defined option to choose the direction of interest for enrichment such that either suppressive or activating regulators could be modeled during the target prioritization step. For example, DReAmiR can be used to examine the group-specific activities of RNA-binding proteins, long non-coding RNAs, or even TFs between multiple cancer subtypes. It can also be applied to study chromatin regulator behavior or histone modifications, where the chromatin-binding peak intensities are used for generating rank lists and peaks associated with certain target genes or functions as the target set. As an extension to gene set variation analysis<sup>202</sup>, DReAmiR can also answer a question like whether genes in a predefined gene set are similarly correlated with a phenotype across groups, similar to what we did for the mmu-miR-155 KO dataset.

There are several limitations for DReAmiR worth mentioning. First, DReAmiR does not provide a mechanism for why miRNA regulatory rewiring happens between conditions, and additional biochemical studies will need to be performed for this purpose. Also, even though we extensively validate DReAmiR results using known biological observations, there are limited examples of known context-specific miRNA regulation to serve as ground-truth for validating DReAmiR. With sequencing technology development, we hope methods such as cross-linking ligation and sequencing of hybrids<sup>293</sup> (CLASH) can be improved such that the exact binding between all miRNAs and target genes could be examined for large sample size and single-cell level for DReAmiR validation.

In conclusion, DReAmiR is a new approach to characterize miRNA-mediated gene regulatory network rewiring across multiple groups from transcriptomic profiles. The method may offer novel insights into cell-type and cancer subtype-specific miRNA

regulatory roles. In the future, we plan to apply the method to the premalignant bronchial lesion biopsy data and examine whether miRNA with differential regulatory roles may be associated with the molecular subtypes.

## CHAPTER 4 CONVERGENCE OF YAP/TAZ, TEAD AND TP63 ACTIVITY DIRECTS PREMALIGNANT LUNG GENE EXPRESSION

*Adapted from the following manuscript:*

Ning B, Tilston-Lunel A, Simonetti J, Hicks-Berthet J, Matschulat A, Pfefferkorn R, Spira AE, Mazzili SA, Lenburg ME, Beane JE and Varelas X. Convergence of YAP/TAZ, TEAD and P63 activity directs premalignant lung gene expression. *In submission*.

### **4.1 INTRODUCTION**

Lung cancer accounts for the largest number of deaths in the United States among all cancer types, making up over 20% of cancer-related deaths in 2020<sup>171</sup>. The development of lung squamous cell carcinoma (LUSC), one of the most common subtypes of lung cancer, is preceded by the formation of bronchial premalignant lesions (PMLs), which are characterized by the abnormal expansion and morphological alteration of airway basal cells<sup>294</sup> that progress through a series of histological grades, from normal, hyperplasia, metaplasia to dysplasia. Our poor understanding of the early molecular events associated with these precancer states makes it difficult to develop potential interception strategies for LUSC<sup>176,177</sup>. Previous studies profiling gene expression in bronchial PML samples have suggested that progressive higher grade lesions show immune evasion profiles, including impaired antigen presentation and decreased lymphoid and myeloid populations<sup>33–35,39,178</sup>. Although immune evasion is a feature of LUSC<sup>295</sup>, the mechanisms contributing to similar phenotypes in bronchial PML progression is poorly understood.

A transcription factor important for controlling bronchial basal cell identity is the p53 family member TP63 (also known as p63)<sup>296–298</sup>, which is encoded by the *TP63* gene. Amplification and overexpression of *TP63* are frequently observed in squamous cell carcinoma, including LUSC<sup>174,299</sup>, and ectopic expression of the  $\Delta$ Np63 isoform ( $\Delta$ Np63) has been shown to drive to the development of squamous metaplasia in the mouse lung and promote proliferative phenotypes in skin epithelial basal cells<sup>300,301</sup>. The activity of TP63 is regulated by a number of transcriptional co-factors, including the Hippo signaling pathway effectors YAP and TAZ<sup>302–304</sup>, which in the lung associate with TP63 to regulate airway basal epithelial cell growth<sup>305</sup>.

Recent evidence has implicated the aberrant activity of YAP/TAZ in bronchial PML development. For example, YAP/TAZ regulated transcription is associated with the progression of human PMLs, and the aberrant activation of YAP/TAZ in the bronchial epithelium of mice drives epithelial growth and PML-like pathology<sup>124</sup>. YAP and TAZ encoding genes are frequently amplified in squamous carcinoma<sup>306</sup> and ample evidence has suggested the oncogenic role of YAP/TAZ across multiple cancer types<sup>118,307–310</sup>. YAP/TAZ functions rely on their ability to associate with the TEAD family of transcription factors<sup>101,311</sup>, including the essential roles for YAP/TAZ in the development and homeostasis of the lung<sup>101,312</sup>.

The relationship between YAP/TAZ, TEAD and TP63 has not been explored. Given the implication of these factors in PML development we set out to investigate the binding

pattern and gene expression program of these factors in human bronchial epithelial cells (HBECs) proliferating in a basal state. We found that YAP, TEADs, and TP63 associate with shared chromatin regions and co-regulate a gene expression program in proliferating HBECs. Integrating genomic binding and gene expression profiling, we demonstrate that gene targets directly induced by TEAD and TP63 are enriched for pro-proliferative genes and those directly repressed are enriched for genes involved in interferon responses and immune regulation. Genes repressed by YAP/TAZ-TEAD-TP63 are notably enriched among the genes down-regulated in progressive/persistent PMLs and includes *CIITA* (also known as MHC2TA), known as a “master” transcriptional co-activator of major histocompatibility complex (MHC) class II gene transcription. Our data suggest that a YAP/TAZ-TEAD-p63 regulated network contributes to a bronchial basal cell proliferative state and the immune-evasive microenvironment observed in lung PMLs. Taken together, our results provide insight into the functions of transcriptional complexes that contribute to the early stages of lung carcinogenesis and offer potential new avenues to develop lung cancer interception strategies.

## **4.2 METHODS**

### *4.2.1 Primary human bronchial epithelial cell culture*

HBECs (Lonza Lot# 269120 and 451973) were cultured in Pneumacult EX Plus media (StemCell Technologies). siRNA transfection was carried out with Lipofectamine RNAimax (Invitrogen, 13778150) on low density proliferating cells and were maintained in submerged culture for 48 hours before lysis.

#### *4.2.2 Immunoprecipitation and Immunoblotting*

HBECS were cultured in Pneumacult EX plus (StemCell Technologies) and proliferating cells were lysed in Tris-buffered saline with 0.1% Tween (TBS-T) detergent. Lysates were subjected to immunoprecipitation using an anti-pan-TEAD antibody to isolate endogenous TEAD proteins and then analyzed by immunoblotting using an anti-TP63 antibody.

#### *4.2.3 TP63 and isoform expression data analysis in TCGA LUSC and PML data*

Copy number data for TCGA LUSC samples (PanCancer Atlas; N=487), TP63 amplification frequency and expression level z-score relative to normal samples across TCGA cancer types were downloaded from cBioPortal<sup>186,313,314</sup>. A list of transcription factor genes in LUSC was obtained from aracne.networks R package<sup>315,316</sup>. TCGA LUSC gene and transcript level expression data of the TCGA LUSC samples were downloaded using TCGAbiolinks<sup>198</sup> (legacy data) for primary tumor (N=502) and normal tissue (N=51) samples. The count data were normalized using the trimmed mean of M-values (TMM) from edgeR R package<sup>277</sup> and transformed into log<sub>2</sub> counts per million. For analysis on Tap63 and  $\Delta$ Np63, raw counts related to each isoform were summed before normalization based on annotation from UCSC genome browser. Gene and TP63 isoform over-expressions were examined with a linear model comparing tumor to normal samples adjusting for the plate. To assess the association between TP63 isoform expression level and the histological grades in Beane et al., same normalization was performed and a linear mixed-effect model was fitted with the lesion grade as the main independent

variable, adjusting for sequencing batch and median TIN and the patient was adjusted as a random effect.

#### *4.2.4 HBEC RNA-seq experiments*

For generating the YAP/TAZ, TEAD and TP63 regulated gene expression signature in human airway cells, HBECs (Lot# 619261 and 18TL386664) were cultured in Pneumacult EX plus and transfected with control and dual Yap/Taz targeting, pan-TEAD targeting and TP63 targeting siRNAs and RNA was extracted for quality assessment and library prep for RNA-sequencing. 3 unique siRNA control sequences were used, and all siRNA transfected samples were collected in triplicate, 48 hours after transfection. RNA quality for all samples was assessed by BioAnalyzer before proceeding with library preparation for sequencing. Sequencing libraries were prepared from total RNA samples using Illumina TruSeq RNA Sample Preparation Kit v2. The libraries from individual samples were pooled sequencing. HBEC samples were sequenced on the Illumina HiSeq 2500 platform to generate single end 50bp reads.

FASTQ files were demultiplexed and created by Illumina BaseSpace. The quality of the FASTQ files was examined with FastQC<sup>191</sup>. The samples were aligned to the build version hg19 of the human genome using STAR 2-pass alignment<sup>317</sup>. RSEM<sup>318</sup> was then used to quantify the gene and transcript counts using Ensembl v75 annotation, and RSeQC<sup>196</sup> was used to calculate the quality metrics. The count data were normalized by the library sizes using the TMM and transformed into log<sub>2</sub> counts per million using

edgeR R package<sup>277</sup>.

#### *4.2.5 ChIP-seq experiments*

HBECs (Lot# 619261 and 18TL386664) for ChIP were cultured in Pneumocult EX Plus media and cross-linked in 1mM EGS in PBS for 30min followed by a 1% formaldehyde treatment for 10 min. Fixation was subsequently neutralized with 0.125M glycine in PBS. Harvested chromatin was isolated as single samples from each patient line, sonicated using the Bioruptor UCD- 200 and the incubated with the following antibodies at 4 °C overnight: Rabbit anti-Yap (Abcam, Cat# ab52771, 3ug), Rabbit anti-TEAD (AvivaSysBio, Cat# ARP38276, 1ug), and Mouse anti-P63 (Biocare # CM163, 5ug). Immunoprecipitated complexes were collected by Protein A/G Magnetic beads (Pierce, 8802). Samples were washed with low salt buffer (20mM Tris, 140mM NaCl, 1mM EDTA, 0.1% NaDeoxycholate, 0.1% SDS, 1% Triton X-100), followed by a high salt buffer (20mM Tris, 500mM NaCl, 1mM EDTA, 0.5% NaDeoxycholate, 1% Triton X-100), and a LiCl buffer (20mM Tris, 1mM EDTA, 0.1% NaDeoxycholate, 1% Triton X-100, 250mM LiCl). Chromatin was de-crosslinked overnight at 65C and purified using the Qiaquick PCR purification kit (Qiagen, 28104). For ChIP-seq, the purified DNA was ligated to specific adaptors and sequenced using DNB-seq, performed by BGI, to a depth of 40 million reads.

The ChIP-seq fastq files were aligned to the build version hg19 of the human genome using Bowtie2<sup>319</sup> with the default parameters. Reads that were unmapped, not primary

alignment or with MAPQ score lower than 30 were removed. Duplicated reads were marked by Picard<sup>320</sup> and were discarded from the alignments and the resulting SAM files were converted to BAM format with samtools<sup>321</sup>. Peak-calling was performed for each individual replicate against the IgG control ChIP-seq consistently using the narrow-peak mode from Model-based Analysis for ChIP-Seq (MACS2)<sup>322</sup> at a p-value cutoff of 0.05 with nomodel option and extsize of 150. Additionally, peaks were filter by summit fold-change  $> 2$ . Peaks within the blacklisted regions ([hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/](http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/)) were removed. Overlapped peaks between replicates were then identified using the `findOverlapsOfPeaks` function from the `ChIPpeakAnno` R package<sup>323</sup> and were used for downstream analysis.

Overlapped peaks between ChIP-seq experiments were found using the `findOverlapsOfPeaks` function from the same package.

The normalized read density for each factor was calculated using `callpeak` function of MACS2 (`-B -SPMR -nomodel -extsize 150`) from pooled replicates for genome track visualization using `karyoploter` R package<sup>206</sup> and read coverage visualization within up- and down-stream 2kb window around the peak center. Significance of peak overlap was calculated with the `enrichPeakOverlap` function from `ChIPseeker` R package<sup>324</sup>. Motif enrichment analysis was done within the YAP peak, YAP-TEAD overlapped peak, and the YAP-TP63 overlapped peak regions using `findMotifsGenome.pl` function from HOMER software suite<sup>325</sup> with the default parameters. The distances between peak locations and TSS for each factor were calculated using the `annotatePeakInBatch` function from the `ChIPpeakAnno` R package<sup>323</sup>.

#### *4.2.6 Derivation of gene expression signature from RNA-seq siRNA experiments*

Gene expression signature was generated comparing each siRNA knockdown with the control experiments in HBECs separately. First, we excluded genes from the count table if the interquartile range was equal to zero or the sum of counts was less or equal to 1 across samples. This yielded 13976, 13918, 13938, and 13859 genes for the siYT, siTEAD, siTP63, and siLATS experiments, respectively. The remaining genes were TMM normalized again. Then, the data was voom-transformed and the differentially expressed genes associated with siRNA treatment were identified using a linear model in limma R package with treatment as the main independent variable, adjusting for cell line<sup>195,197</sup>. Genes significantly associated with siRNA treatment were filtered at FDR < 0.05 and absolute log fold change greater than 0.5. Genes were ranked by the t-statistic for their association with treatment effect to generate the rank list for each siRNA. The enrichment of differentially expressed genes on rank lists from another siRNA KO experiment was examined by GSEA<sup>204</sup> using the fgsea R package<sup>326</sup>. Expression residual values adjusting for cell line were used for heatmap visualization using the ComplexHeatmap R package<sup>327</sup>.

#### *4.2.7 Derivation of direct target genes of TEAD and TP63 from ChIP-seq experiments*

Genes with a transcriptional start site (TSS) within 50kb from TEAD-TP63 overlapped peaks were assigned as direct target genes. Next, to account for the potential long-range interaction, we utilized promoter capture Hi-C interaction data of lung tissue from 3div.kr<sup>328</sup>. P-value cutoff of 0.05 was used to filter the promoter-promoter and promoter-

other interactions, which resulted in 15545 and 52254 pairs of chromatin interactions respectively. Genes with TSS overlapping with a promoter-containing fragment that had interacting fragment overlapped with a TEAD-TP63 overlapped peak were assigned as direct target gene. Then, the target gene sets were filtered based on their association with siYT, siTEAD and siTP63 treatments. Genes significantly up-regulated (FDR < 0.05 and log fold-change > 0.5) in all siRNA treatments compared to the controls were assigned as “repressed targets”, whereas genes significantly down-regulated (FDR < 0.05 and log fold-change < -0.5) were assigned as “induced targets.” Functional pathway enrichment analysis for the TEAD-TP63 induced and repressed target genes were performed using the hypergeometric test implemented in the R package hyper<sup>278</sup> and the Molecular Signatures Database (MSigDB) from the Broad Institute. The enrichment of TEAD-TP63 induced and repressed target genes within PML co-expressed gene modules were examined with Fisher’s exact test.

#### *4.2.8 Computational analyses of TEAD-TP63 direct target genes in human patient data*

We obtained bulk gene expression profiles of endobronchial biopsies including various PML histological grades and progression status from two studies: the discovery and validation cohort from Beane et al. (GSE109743; discovery cohort with 190 biopsies from 29 subjects; validation cohort with 105 biopsies from 20 subjects)<sup>39</sup> and Merrick et al. (GSE114489; 63 biopsies from 42 subjects)<sup>33</sup>. For the samples from Beane et al., the residual expression values adjusting for batch and RNA quality measured by the transcript integrity number (TIN)<sup>196</sup> were first calculated as in the original study and were

used for further analysis.

A metagene score for TEAD-TP63 direct induced and repressed target genes was calculated using GSVA<sup>202</sup> for each sample within each dataset separately. Correlation between TF levels (YAP, TAZ TP63, and TEAD1-4) and metagene scores were calculated with Pearson correlation.

To assess the association between TEAD-TP63 induced and repressed metagene scores and the histological grades in Beane et al., a linear mixed-effect model was used with the histological grade as the main independent variable (coded as a continuous variable from normal to severe dysplasia/carcinoma in situ), and the patient was adjusted as a random effect using nlme<sup>329</sup>. For Merrick et al., a linear model was used with the histological grade as the main independent variable (coded as a continuous variable from normal to severe dysplasia/carcinoma in situ).

To study the association between TEAD-TP63 induced and repressed target genes and the lesion progression status, a gene rank list was first calculated for each dataset. For Beane et al., genes were ranked by the t-statistic for their association with progression status from a linear mixed effect model, comparing progressive/persistent lesions to the regressive ones among samples of the Proliferative subtype, adjusting for the patient as a random variable using duplicateCorrelation function from limma<sup>197</sup>. For the Merrick et al., genes were ranked by the t-statistic of a linear model comparing all progressive/persistent samples (including the persistent bronchial dysplasia and progressive non-dysplasia groups in the original annotation) to the regressive ones (regressive bronchial dysplasia group). Then, GSEA<sup>204</sup> was used to test whether the

TEAD-TP63 direct induced and repress target genes were enriched within the rank lists. Immune infiltration scores of 24 immune cell-types within the bulk RNA-seq samples were calculated using GSVA<sup>202</sup> based on the immune cell-type-specific signature genes from Bindea et al.<sup>330</sup>. The association between metagene scores of TEAD-TP63 induced and repressed target and the immune cell-type scores were calculated with Pearson correlation. The association between the immune cell-type scores and lesion progression status was examined using the same model as described above.

#### *4.2.9 Single-cell RNA-seq data analysis*

10X Chromium single-cell RNA-seq datasets of the human healthy airway and normal lung tissue, including normalized count data and annotations, were obtained from Travaglini et al. (EGAS00001004344)<sup>331</sup> and Deprez et al (EGAS00001004082)<sup>332</sup>. Cell clustering and cell-type annotation from the original studies were used. For Travaglini, the tSNE coordinates across all cells was calculated using the top 20 principle components with 2K highly-variable genes for visualization using the RunTSNE function from Seurat<sup>333</sup>. For Deprez et al. healthy airway dataset, only the proximal and intermediate airway biopsy samples were used for our analysis to match the cellular composition of bronchial premalignant lesion bulk RNA-seq data. The metagene scores of TEAD-TP63 direct induced/repressed target gene sets and MHC Class II genes were calculated using AUCell R package<sup>254</sup> based on normalized count data and were compared between cell-types using one-tail Wilcoxon test. Ligand-receptor analysis was performed in Travaglini et al. normal lung dataset using CellChat<sup>334</sup>.

#### *4.2.10 CP1 and chemical compound analysis*

The RNA sequencing data from MGH-CP1 treatment and control experiments in MDA-MB-231 cells was obtained from GSE140396<sup>132</sup>. Data processing and gene filtering was performed as described above. MGH-CP1 associated gene signature were derived by comparing the samples from MGH-CP1 treatment group to the DMSO control group. CMAP signature connectivity analysis was performed using the top 150 TEAD-TP63 direct induced and repressed target genes, ranking by log fold-change comparing siRNA treatment group (siYAP/TAZ, siTEAD and siTP63 combined) to the control group. CMAP results were filtered by: “perturbation\_type” = “trt\_cp”, cell line derived from lung, and qc\_pass = “1”.

#### *4.2.11 Datasets used and Code Availability*

RNA-seq and ChIP Seq datasets have been deposited to NCBI GEO GSE213656 and GSE158307.

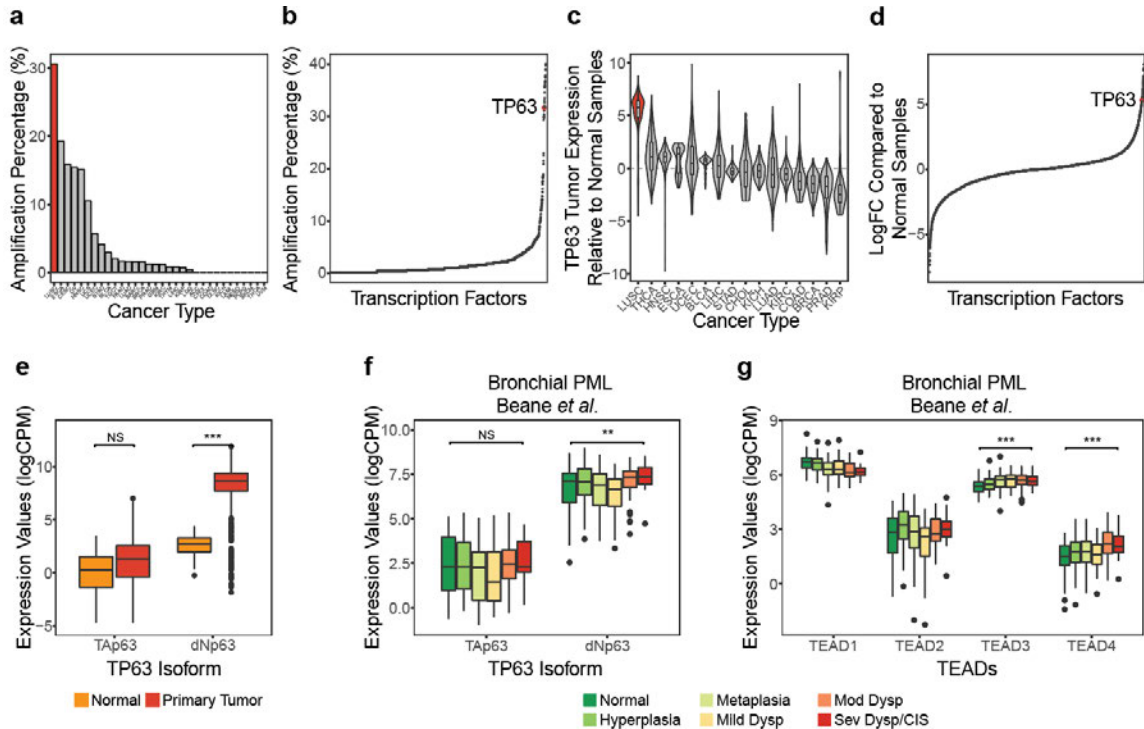
### **4.3 RESULTS**

#### *4.3.1 TP63 and TEAD expression is elevated in PML histological progression*

In prior work, we demonstrated that activation of the transcriptional effectors YAP and TAZ stimulates lung epithelial basal cell growth and induces gene expression associated with progressive bronchial PML<sup>124</sup>. Given the reported association of TP63 with YAP and TAZ<sup>305</sup> and the critical functions of TP63 in maintaining airway basal stem cell

identity, we hypothesized cooperation between YAP/TAZ and TP63 in airway epithelial cells contributes to the progression of premalignant human airway disease and LUSC. We first sought to examine the association of *TP63* with lung squamous tumor samples. Analysis of LUSC data available from The Cancer Genome Atlas (TCGA) showed that over 30% of tumors have an amplification of *TP63*, which is more frequent compared to other cancers profiled in TCGA (**Figure 4.1a**) and is among the most frequently amplified transcription factors in LUSC (**Figure 4.1b**). We found that *TP63* is also significantly overexpressed in primary tumor samples compared to normal tissues in TCGA LUSC data compared to other cancer types (**Figure 4.1c**; linear regression model p-value  $\leq 0.001$ ), ranking high among transcription factors expressed in this cancer subtype (**Figure 4.1d**). Notably,  $\Delta$ Np63, the major TP63 isoform with oncogenic functions<sup>335</sup>, is more strongly over-expressed in primary LUSC tumor samples than the full-length Tap63 (**Figure 4.1e** and **Figure A.9a**; linear model p-value  $< 0.001$  and p-value  $< 0.05$ ). Similar high expression of  $\Delta$ Np63 was observed in high-grade PML samples<sup>39</sup> compared with low-grade PMLs (**Figure 4.1f** and **Figure A.9b**; mixed effect model p-value  $< 0.01$ ), and  $\Delta$ Np63 was the dominant isoform across all stages of PML samples, suggesting that TP63 activity may be important for high-grade PML development.

To further explore the association between *TP63* and increasing PML histologic grade, we performed TF enrichment among genes (N=822) in a previously defined co-expressed gene module that is significantly increased in higher grade PML and is enriched with genes involved in cell cycle and DNA replication pathways<sup>39</sup>. Results from both Binding



**Figure 4.1. TP63 is associated with human LUSC carcinogenesis and early lung cancer progression.** **a.** TP63 amplification frequency in TCGA samples by cancer types. **b.** Transcription factors ranked by amplification frequencies in the TCGA LUSC samples. **c.** TP63 expression z-scores in TCGA primary tumor samples relative to normal samples by cancer types. **d.** Transcription factors ranked by logFC comparing TCGA LUSC tumor to normal samples. **e-f.** Boxplots show the TP63 isoform (TAp63 and dNp63) expression levels between normal and primary tumor samples in TCGA LUSC (**e**), and across bronchial PML histological grades in Beane *et al.* (**f**). \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ . **g.** Boxplots show the TEAD3/4 expression levels across bronchial PML histological grades in Beane *et al.* \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

Analysis for Regulation of Transcription (BART)<sup>336</sup> and ChIP-X Enrichment Analysis 3 (ChEA3)<sup>337</sup> indicated TP63 as a highly significant TF regulating the genes in this histologic-grade associated gene module (p-value < 0.001). Notably, TP63 was also listed as the top TF regulating genes that are up-regulated in LUSC compared to normal tissues in BART-Cancer<sup>338</sup> (p-value < 0.001).

We also found that the expression levels of *TEAD3* and *TEAD4*, which encode transcription factors of the TEAD family that are regulated by YAP/TAZ binding, were significantly increased with higher histologic grades in PML samples<sup>39</sup> (**Figure 4.1g**; mixed effect model p-value < 0.001). The *TEAD1* and *TEAD2* family members did not exhibit similar increases (**Figure A.9c**). These observations suggested that TP63 and TEAD transcription factors, both of which are linked to YAP/TAZ function, may be associated with bronchial PML progression.

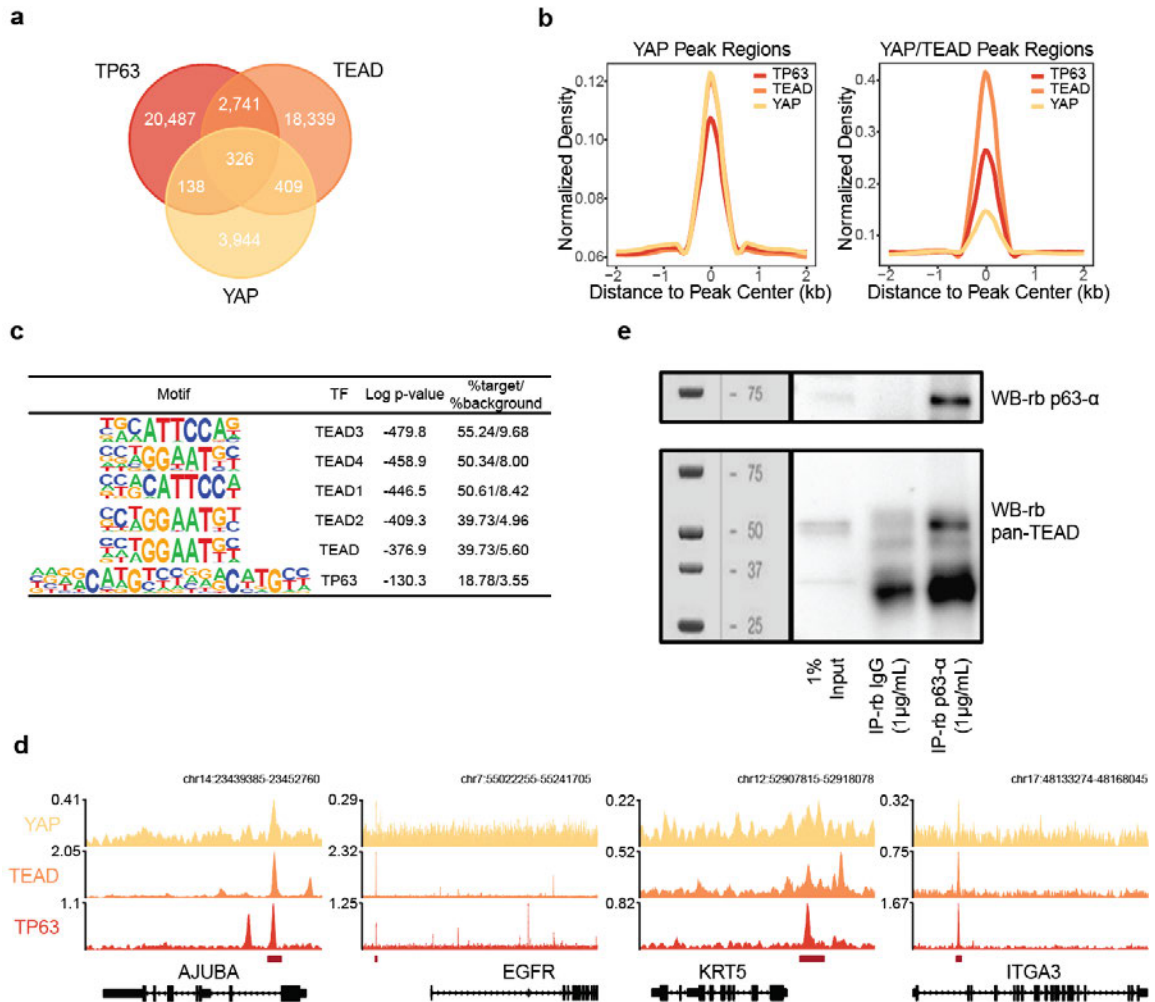
#### *4.3.2 YAP, TEAD and TP63 bind to the same genomic sites in basal bronchial epithelial cells*

To characterize the genes directly regulated by YAP/TAZ, TEAD and TP63, we performed chromatin immunoprecipitation sequencing (ChIP-seq) from proliferating HBECs using antibodies targeting YAP, TEADs (pan-TEAD antibody) and TP63. In total, 4817, 21925, and 23692 consensus peaks (overlapped between replicates) were identified for YAP, TEAD, and TP63, respectively. While about 25% of the peaks from each ChIP-seq experiment were located within 2.5 kb from the gene TSSs, many peaks were located further away from gene promoter regions (**Figure A.10a**), indicating

potential long-range gene regulation for YAP, TEAD, and TP63, as previously suggested<sup>112,339</sup>.

We next compared the chromatin binding patterns of YAP, TEAD, and TP63 and found significant peak overlaps between the three factors: 735 peaks were overlapped between YAP and TEAD, 464 were overlapped between YAP and TP63, and 326 were overlapped between all three (**Figure 4.2a**; hypergeometric test p-value < 0.001).

Intriguingly, coverage density analysis not only revealed strong TEAD coverage at the YAP binding regions but also TP63 coverage at both the YAP binding and YAP-TEAD co-binding regions (**Figure 4.2b**). HOMER motif enrichment analysis on the YAP-TEAD co-binding peaks identified TEAD and TP63 motifs as the two most significantly enriched TF binding motifs (**Figure 4.2c**; p-value < 0.001). Similarly, both TEAD and TP63 motifs were significantly enriched at the YAP binding regions and YAP-TP63 co-binding regions, suggesting these are the primary DNA binding factors mediating YAP function in HBECs (**Figure A.10b**). Regions bound by YAP, TEAD and TP63 included the promoters/enhancers of target genes identified in other contexts, including *AJUBA* for YAP and *EGFR* for TP63, as well as genes associated with basal cell identity, such as *KRT5* and *ITGA3* (**Figure 4.2d**)<sup>110,340</sup>. These analyses showed highly overlapped chromatin-binding profiles between YAP, TEAD, and TP63, prompting us to test for physical association between these factors. YAP is documented to interact with TEADs<sup>101</sup> and TP63<sup>302,305,341,342</sup>, so we tested whether TEAD associates with TP63. Co-immunoprecipitation experiments from primary HBECs showed a strong interaction between TEAD and TP63 (**Figure 4.2e**). Taken together these data suggest that YAP,



**Figure 4.2. TP63 interacts and co-binds to chromatin with YAP/TEAD in HBECS.**

**a.** Venn diagram shows peak overlaps between YAP, TEAD and TP63 chromatin binding domains in HBECS. **b.** Distribution of YAP/TEAD/TP63 ChIP-seq signal around  $\pm 2$  kb of YAP and YAP/TEAD overlapped peak regions (N=4817 and 735). **c.** Top transcription factor binding motifs enriched in the YAP/TEAD overlapped peak regions in HBECS. P-values were calculated by HOMER. **d.** YAP, TEAD and TP63 ChIP-seq tracks shows the co-binding at the promoter regions of Hippo or TP63 canonical target genes. Overlapped peak regions are shown in red strips. **e.** Western blot showing TEAD and TP63 co-immunoprecipitated together in HBECS.

TEAD and p63 form a transcriptional complex in proliferating basal bronchial epithelial cells.

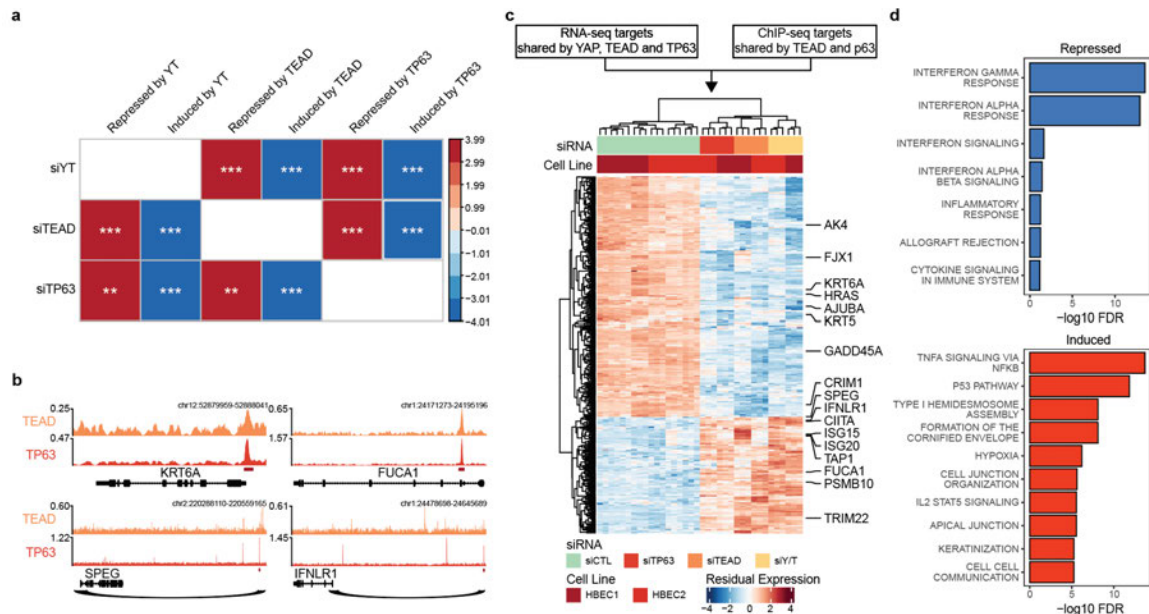
#### *4.3.3 TP63 and TEAD co-regulate gene expression in the basal bronchial epithelial cells*

To gain insight into the transcriptional relationship between YAP, TEAD, and TP63, we performed bulk RNA sequencing on proliferating HBECs treated with siRNA targeting YAP/TAZ, TEADs, and TP63. Differential expression analysis comparing siRNA treated samples to the controls identified 2581, 2120, and 1566 genes down-regulated in expression following siRNA-mediated knockdown of YAP/TAZ, TEAD, and TP63, respectively (i.e., genes normally induced by these factors). This analysis also identified 2510, 2096, and 1391 genes, that were up-regulated in expression following siRNA-mediated knockdown YAP/TAZ, TEAD, and TP63, respectively (i.e., genes normally repressed by these factors). YAP/TAZ, TEAD, and TP63 induced genes (i.e., genes down-regulated with siRNA-mediated knockdown) were significantly enriched within each other's respective gene sets, and a similar pattern was observed for YAP/TAZ, TEAD, and TP63 repressed genes (i.e., genes up-regulated with siRNA-mediated knockdown) (**Figure 4.3a**; GSEA FDR < 0.01), suggesting that YAP, TEAD and TP63 regulate a shared gene expression program in HBECs.

Next, we sought to identify genes directly co-regulated by YAP, TEAD, and TP63 by integrating the chromatin binding profiles from ChIP-seq experiments with the gene expression profiles from the RNA-sequencing of the siRNA experiments. Since only TEAD and p63 directly bind DNA, and due to higher quality data obtained from our ChIP-

seq analysis of TEAD and TP63, we combined our TEAD-p63 overlapped peaks (N=3067) with our RNA-seq analysis to identify potential direct targets. Genes with TSS within 50kb from the TEAD-TP63 overlapped binding regions or potentially regulated by TEAD and TP63 through long range interactions at distal regions (based on pcHi-C data from Jung et al.<sup>328</sup>) were labeled as direct targets (**Figure 4.3b**). This analysis identified 260 TEAD-TP63 directly induced (i.e., genes with binding peaks that were down-regulated following siRNA-mediated knockdown) and 126 directly repressed (i.e., genes with binding peaks that were up-regulated following siRNA-mediated knockdown) target genes (**Figure 4.3c**). Among the TEAD-TP63 direct induced targets, we found several canonical targets for both the Hippo pathway (AJUBA, GADD45A, FJX1, and CRIM1)<sup>110</sup> and TP63 pathway (AK4, KRT5/6A, and HRAS)<sup>340</sup>. We also observed several interferon response and antigen processing related genes among the TEAD-TP63 direct repressed genes.

To further validate the regulation of target genes via endogenous YAP/TAZ activation, and to test the effects of Hippo pathway regulation of these targets, we depleted the LATS1/2 kinases in HBECs using siRNA and found that TEAD-TP63 directly induced and repressed target genes were among the genes most down- and up-regulated in the siLATS treatment samples compared to the controls (**Figure A.11**; GSEA p-value  $\leq 0.005$ ). Functional enrichment analysis revealed that the TEAD-TP63 induced genes are associated with cell proliferation and extracellular matrix-associated pathways, and the repressed genes are strongly enriched for interferon alpha and gamma responses (**Figure 4.3d** hypergeometric FDR < 0.001).



**Figure 4.3. YAP/TEAD/TP63 together regulate target genes associated with carcinogenesis pathways in HBECs.**

**a.** Correlation plot summarizes the GSEA results of genes associated with YAP/TAZ, TEAD and TP63 siRNA treatments in HBECs. Rank lists were generated by ranking genes by t-statistics for their association with siRNA treatment comparing to the controls in HBECs. Genes significantly up or down-regulated with the siRNA treatments were used as gene sets (absolute logFC > 0.5 and FDR < 0.05). \*\*p-value < 0.01, \*\*\*p-value < 0.001. **b.** TEAD and TP63 ChIP-seq tracks shows the representative co-binding associated TEAD-TP63 direct target genes. Overlapped peak regions are shown in red strips. Only the direct target genes are plotted. **c.** Heatmap of gene expression significantly altered in siYAP/TAZ, siTEAD and siTP63 treatment (absolute logFC > 0.5 and FDR ≤ 0.05). Genes annotated on the right are associated with interferon response pathways or shown to be canonical target genes of Hippo or TP63 pathways. **d.** Top enriched functional pathways associated with the TEAD-TP63 direct repressed (top) and induced (bottom) target genes.

#### *4.3.4 The TP63/TEAD repressed gene program is associated with early immune evasion in the bronchial premalignant lesions*

To explore a potential the relationship between YAP/TAZ, TEAD and p63 transcriptional regulation and the gene expression changes associated with bronchial carcinogenesis, we measured metagene scores of directly induced and repressed target genes shared by these factors in human PML patient endobronchial biopsy samples factors. Metagene scores were calculated for the induced and repressed targets separately in three gene expression datasets which examined progressive PML pathology, which included RNA sequencing data from Beane et al., (GSE109743) which defined both a discovery cohort and an independent validation cohort<sup>39</sup> and Affymetrix Gene 1.0 ST microarray data from Merrick et al., (GSE114489)<sup>33</sup>. First, we validated that the expression of TEAD-TP63 direct targets were correlated with the expression levels of *YAP*, *TAZ/WWTR1*, *TEAD*, and *TP63*. A strong positive correlation was observed between the metagene score for the directly induced targets of TEAD-TP63 and *YAP*, *TAZ/WWTR1*, *TP63*, and *TEAD2/3/4* (*TEAD1* did not show a similar correlation) in Beane et al. discovery cohort<sup>39</sup>, and conversely a negative correlation was observed for the directly repressed targets of TEAD-TP63 (**Figure 4.4a**). Similar correlation patterns were also observed in the Beane et al. validation cohort and the Merrick et al. dataset (**Figure A.12a**).

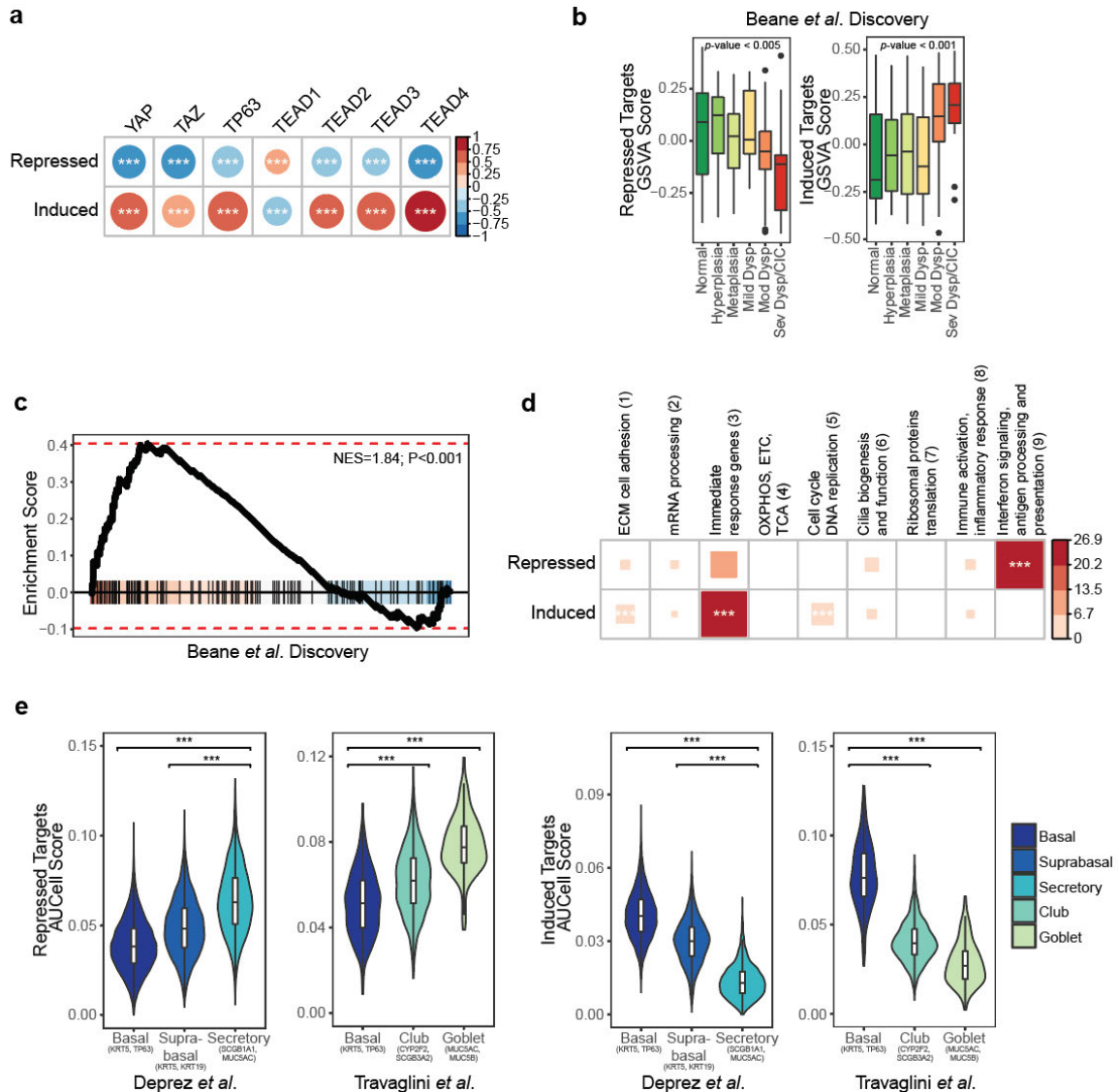
Notably, TEAD-TP63 direct target genes were significantly associated with increased PML histologic severity; the metagene score of the induced targets were significantly increased in higher grade PML samples (linear model p-value < 0.001 in all three datasets) and the metagene score of the repressed targets were decreased, although less

significantly (**Figure 4.4b** and **Figure A.12b**). TEAD-TP63 directly repressed target genes were significantly enriched among the genes down-regulated in progressive/persistent compared with regressive PMLs among the samples of Proliferative subtype described in the Beane et al. discovery (GSEA p-value <0.001) and validation cohort (GSEA p-value <0.05)<sup>39</sup>, and among all samples in Merrick et al. (GSEA p-value <0.001)<sup>33</sup> (**Figure 4.4c** and **Figure A.12c**). Directly induced genes were also strongly enriched among the genes up-regulated in progressive/persistent PMLs in the Merrick et al. cohort<sup>33</sup> (**Figure A.12c**; GSEA p-value <0.001), although this enrichment was not as clear in the Beane et al. data. Collectively, these observations suggest that shared TEAD and p63 activities are associated with precancerous airway disease progression.

To gain functional insight into TEAD-TP63-regulated genes, we explored potential associations with gene modules identified from network analyses of PML data from prior work, which revealed significant overlap between TEAD-TP63 direct induced target genes and three co-expressed gene modules (Modules 1, 3, and 5) described in Beane et al<sup>39</sup>. TEAD-TP63 targets in these modules were enriched for genes associated with extracellular matrix/cell adhesion, immediate response, and cell-cycle/DNA-replication pathways, respectively (**Figure 4.4d**; Fisher's exact test p-value < 0.001), suggesting these gene networks are induced by TEAD-P63 in bronchial PMLs. We also observed a significant overlap between the TEAD-TP63 direct repressed target genes and co-expressed gene module (Module 9) from Beane et al., which is enriched for genes encoding antigen presentation and interferon response pathways factors, strongly

associated with PML progressive pathology and is correlated with the level of immune cell infiltration, including cytotoxic cells, CD8+ T cells, NK cells, Th1 CD4+ T cells, and activated dendritic cells<sup>39</sup>.

Previous studies have suggested immune regulatory functions reside in distinct subsets of airway epithelial cells, with airway secretory cells playing key roles in promoting lymphocytic infiltration<sup>343,344</sup>. We therefore examined the cell-type expression of the genes directly induced or repressed by TEAD-TP63 by calculating the metagene score of TEAD-TP63 direct induced and repressed target genes with AUCell<sup>254</sup> in two normal human airway/lung single-cell RNA-seq datasets, from Deprez et al.,<sup>332</sup> (N=41134) and Travaglini et al.,<sup>331</sup> (N=65662). High expression of genes directly induced by TEAD-TP63 was observed within the basal and suprabasal epithelial cell subsets, while repressed target genes were expressed at lower levels in these same subsets (**Figure 4.4e** and **Figure A.12d** Wilcoxon one-tail test p-value < 0.001). These observations suggest cooperation between YAP/TAZ, TEAD, and TP63 in basal and suprabasal cells.



**Figure 4.4. TEAD-TP63 direct regulated genes are associated with human bronchial PML progressive pathology and early immune evasion.**

**a.** Correlation plot shows the correlation between the expression levels of transcription factors and metagene scores of TEAD-TP63 direct induced and repressed target genes (calculated with GSEA) in Beane *et al.* Discovery cohort. The color and the size of the circles indicate the Pearson correlation coefficients. \*\*\*Pearson correlation, p-value < 0.005. **b.** The metagene scores of TEAD-TP63 direct repressed (left) and induced (right) target gene sets across human bronchial PML data by histological grades in Beane *et al.* Discovery cohort. **c.** Enrichment plot for TEAD-TP63 direct repressed target genes among genes ranked by t-statistics comparing the regressive PML samples to the progressive/persistent ones of the Proliferative subtypes in the Beane *et al.* Discovery cohort (GSEA; p-value < 0.001). **d.** Bubble plot shows the enrichment of TEAD-TP63 direct repressed and induced target gene sets among human bronchial PML co-expressed gene modules. The color and the size of the squares indicate the odds ratio. \*\*\*Fisher's exact test p-value < 0.001. **e.** Violin plots show the summarized expression of TEAD-TP63 direct repressed (left) and induced (right) target genes (calculated using AUCell) in the healthy human airway scRNA-seq data from Deprez *et al.* and human lung scRNA-seq data from Travaglini *et al.* (\*\*\*)one-tail Wilcoxon test, p-value < 0.001).

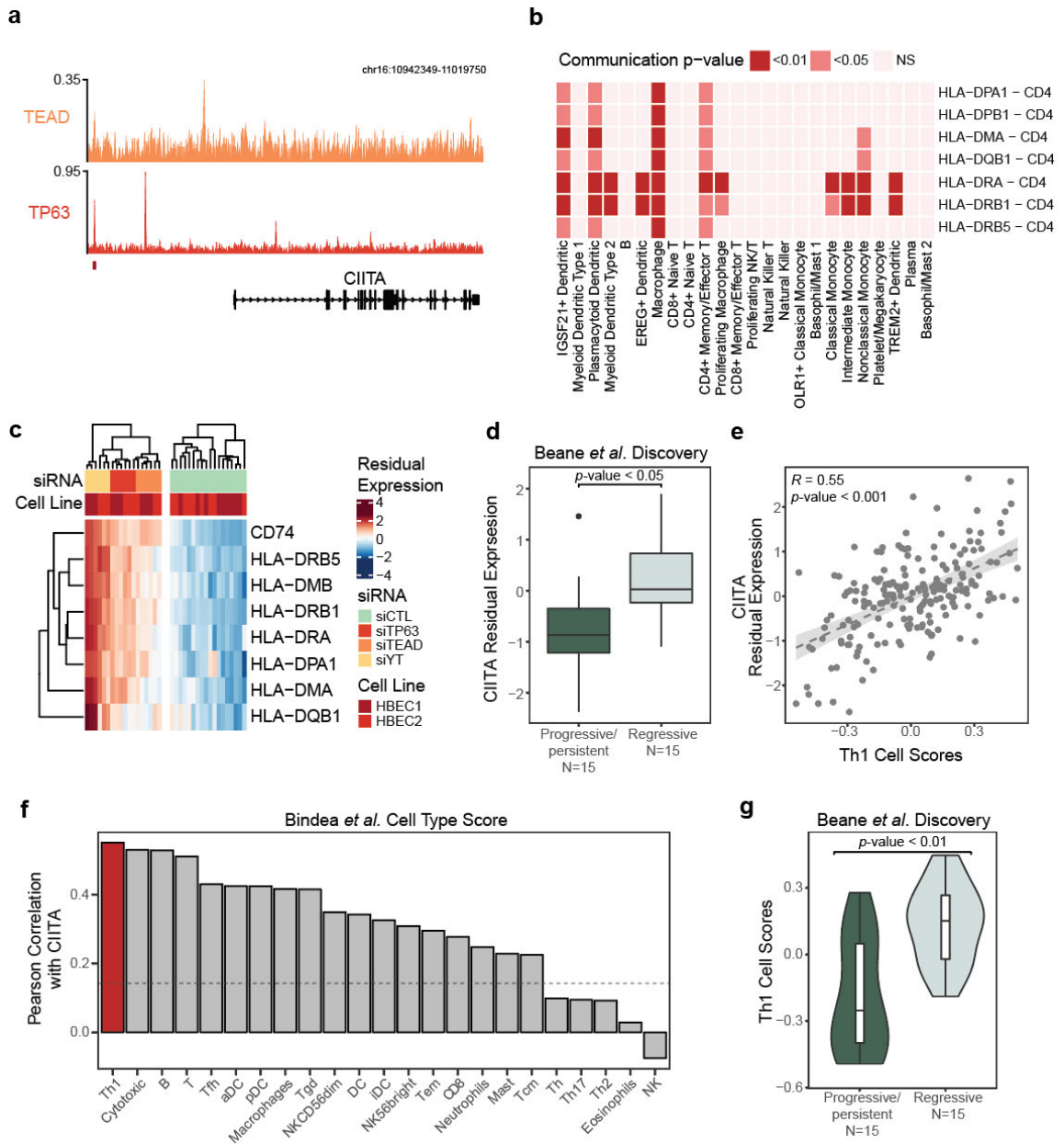
*4.3.5 YAP/TAZ-TEAD-TP63 down-regulate Major Histocompatibility Complex factors transactivator CIITA in bronchial epithelial cells.*

To explore potential basal cell-immune crosstalk downstream of TEAD-TP63 activity we further examined the target genes, and found CIITA, a MHC Class II transactivator that plays critical functions in inducing the expression of MHC-II related genes<sup>345,346</sup>, as a TEAD-TP63 direct repressed target gene (**Figure 4.5a**). We hypothesized that YAP/TEAD/TP63 may repress the MHC Class II gene expressions by down-regulating CIITA to suppress the immune infiltration. To test this, we used CellChat<sup>334</sup> to investigate mediators of ligand-receptor signaling within lung single-cell RNA-seq data from Travaglini et al.<sup>331</sup>, which is a dataset with detailed annotation of immune cell subsets. 61246 significant ligand-receptor interactions between 57 cell-types were identified (p-value < 0.05), and 1210 ligand-receptor interaction pairs involving ligands expressed on basal cells. Among these were 44 interactions between MHC class II genes expressed in basal cells and CD4 in various immune cells (**Figure 4.5b**), including mature CD4<sup>+</sup> T cells, dendritic cells and macrophages. More focused analyses confirmed that most of the genes encoding MHC II family factors were induced following YAP/TAZ, TEAD, or p63 knockdown, indicating repression of MHC family gene expression by YAP/TAZ, TEAD and TP63 (**Figure 4.5c**; linear model FDR < 0.05), including various HLA class II histocompatibility antigens, and CD74, the HLA-DR antigens-associated invariant chain that plays essential roles in the formation and transport of the MHC class II complex<sup>347</sup>.

CIITA belongs to the antigen presentation/interferon response co-expressed gene module

(Module 9) previously identified in Beane et al.<sup>39</sup> as being down-regulated amongst progressive/persistent Proliferative subtype PMLs (**Figure 4.5d**; linear model p-value < 0.05). Similar association between lower CIITA expression and PML progression was observed in the Beane et al. validation (**Figure A.13a**; p-value 0.45) and in Merrick et al. datasets (p-value < 0.05)<sup>33,39</sup>. Concordantly, most of the MHC Class II genes were strongly down-regulated among the progressive/persistent PMLs across three datasets (**Table B.8**). These data therefore suggest that repression of CIITA mediated MHC Class II expression by YAP/TAZ-TEAD-TP63 is associated with early immune-evasion and PML progression.

In previous work by Merrick et al.<sup>33</sup>, increased epithelial MHC Class II molecule HLA-DRA expression had been associated with a regressive PML phenotype and associated with elevated expression of Th1 marker genes. Similarly, ligand-receptor analysis showed potential communication between bronchial basal population and CD4+ T cells utilizing MHC Class II and CD4 interactions. Hence, we sought to further quantify the association between CIITA expression and markers of Th1 cells in PML human patient data. Our analysis showed a strong correlation between the expression level of CIITA and Th1 cell-type score, calculated using signature genes from Bindea et al.<sup>330</sup> (**Figure 4.5e** and **Figure A.13b**; Pearson correlation p-value < 0.001) with the correlation between CIITA and Th1 being strongest compared to other immune cell-types in data from Beane et al. discovery cohort<sup>39</sup> (**Figure 4.5f** and **Figure A.13c**). Consistently, we observed that the Th1 score is decreased in the progressive/persistent PMLs among the samples of the Proliferative subtype PMLs in Beane et al.<sup>39</sup> discovery (**Figure 4.5g** and **Figure A.13d**;



**Figure 4.5. CIITA associates with bronchial PML progressive pathology and tracks with suppressing MHC Class II gene expression and the presence of Th1 T cells.**

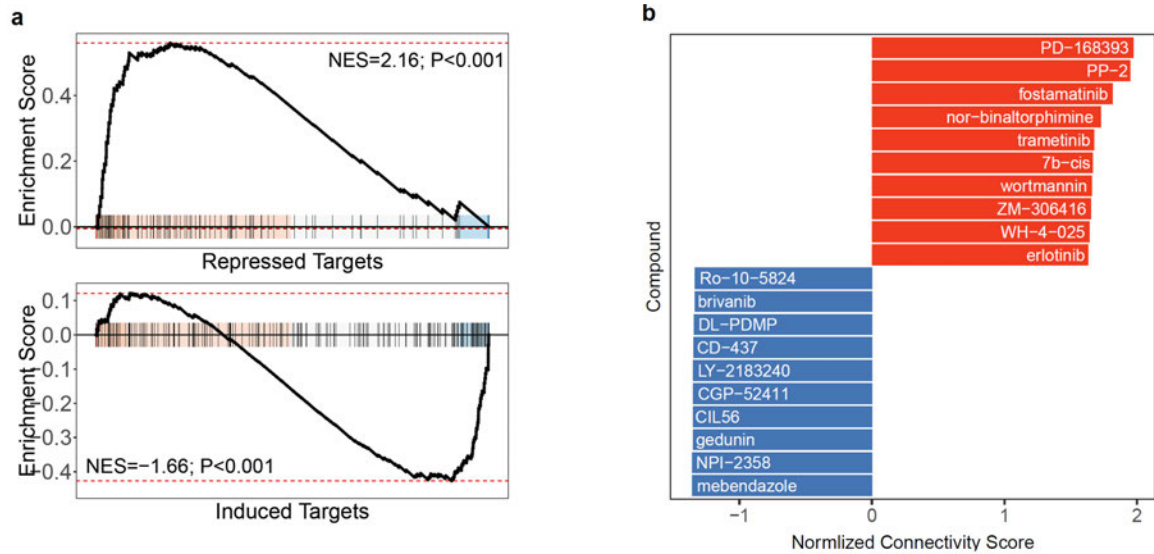
**a.** TEAD and TP63 ChIP-seq tracks shows the co-binding peaks associated with CIITA. Overlapped peak regions are shown in red strips. **b.** Heatmap of communication significance levels between MHC Class II genes in basal cell and binding partners in immune cells in the human lung scRNA-seq data from Travaglini *et al.* The ligand-receptor pairs that involve MHC Class II gene expression in the basal cells are plotted as the row, and the immune cell types that the basal cells are communicating to are plotted as the columns. The color of heatmap reflect the significance levels of the cell-cell communication based on CellChat. **c.** Heatmap of significantly repressed MHC Class II gene expression (FDR < 0.01) in siYAP/TAZ, siTEAD and siTP63 treatment in HBEcs. **d.** Expression level of CIITA in progressive/persistent and regressive PML samples of the Proliferative subtype in Beane *et al.* Discovery cohort. **e.** Scatter plots show the Pearson correlation between the expression level of CIITA and Th1 scores

(calculated using GSVA based on genes from Bindea *et al.*) in Beane *et al.* Discovery cohort. **f.** Immune cell-type ranked by their Pearson correlation coefficients with CIITA expression level in Beane *et al.* Discovery cohort. The dashed line indicates the Pearson correlation coefficient that reaches p-value = 0.05. **g.** Th1 cell scores in progressive/persistent and regressive PML samples of the Proliferative subtype in Beane *et al.* Discovery cohort.

linear model, p-value < 0.01) and validation cohorts (p-value = 0.08), and among all samples in Merrick et al.<sup>9</sup> (p-value = 0.25), suggesting that decreased Th1 infiltration is predictive of PML progression. We observed that MHC Class II genes are strongly expressed among the secretory epithelial cells in two lung scRNA-seq datasets (**Figure A.13e**), highlighting the role of secretory cells in antigen presentation and suggesting that epithelial expression of MHC II genes is associated with lower Th1 infiltration. Taken together, our observations suggest that YAP/TAZ-TEAD-TP63 activity in bronchial epithelial cells repress MHC Class II genes, potentially through down-regulating CIITA, which contributes to PML progression in part by suppressing the local presence of Th1 cells.

#### *4.3.6 Palmitoylation inhibitor blocks TEAD DNA-binding and may reverse the TEAD-TP63 directly regulated gene program*

Finally, we sought to investigate whether chemical compounds can be identified to reverse the YAP-TEAD-TP63 directly regulated gene signatures. Previous study suggested small molecule MGH-CP1 targeting the TEAD auto-palmitoylation pocket may inhibit TEAD-mediated transcriptional regulation *in vivo*<sup>132</sup>. We obtained gene expression data of MDA-MB-231 treated with MGH-CP1 or DMSO (GSE140396<sup>132</sup>) and compared the gene expression changes associated with MGH-CP1 treatment with the TEAD-TP63 direct regulated gene signature. GSEA showed the TEAD-TP63 directly repressed target genes in HBECs were significantly positive-enriched among the genes up-regulated with MGH-CP1 treatment in MDA-MB-231, and the direct induced genes



**Figure 4.6. Blocking the DNA-binding ability of TEADs via CP1 may reverse the TEAD-TP63 directly regulated gene signature.**

**a.** Enrichment plot for TEAD-TP63 direct repressed (top) or induced (bottom) target genes among genes ranked by t-statistics comparing the MDA-MB-231 cell line samples in the CP1 treatment group to the ones in the DMSO control group from GSE140396 (GSEA; p-value < 0.001). **b.** Bar chart for the top small compound perturbation in lung cancer cell lines that generate positive (red) or negative (blue) connectivity to the TEAD-Tp63 directly regulated target genes.

were negatively enriched (**Figure 4.6a**; GSEA p-value < 0.001). This evidence demonstrated inhibiting TEAD-mediated transcription using MGH-CP1 may reverse the TEAD-TP63 direct regulated gene signature in bronchial epithelium and could be a feasible bronchial PML progression interception strategy.

To identify other novel compounds with potential treatment effects, we performed *in silico* chemical compound screening using the Connectivity Map (CMAP)<sup>348,349</sup>. We identified 250 chemical compound that are significantly positively connected to the TEAD-TP63 direct regulated target genes (**Figure 4.6b**; FDR < 0.01). 178 compounds of these were annotated compounds with names not starting with “BRD”. These compounds are potential drug candidates to reverse the Hippo and TP63 regulated gene programs in the human bronchial epithelial cells and may be repurposed to intercept the early lung cancer progression. Notably, trametinib, a MAPK inhibitor<sup>350</sup> were among the top compounds that were positively associated with TEAD-TP63 direct regulated genes. Combination treatment of MGH-CP1 and trametinib could synergistically decreased viability of mice metastatic melanoma cell lines<sup>351</sup>. These observations highlighted the potential of combined inhibition of TEAD and MAPKL signaling pathways to intercept the pathological progression of bronchial PMLs.

#### **4.4 DISCUSSION**

Our study demonstrates that the activity of a gene expression program that is cooperatively regulated by the TEAD and TP63 transcription factors is increased in progressive bronchial PMLs, and that these factors are modulated by the transcriptional

effectors YAP and TAZ (biological hypothesis depicted in **Figure 4.7**). Our observations strongly suggest that these factors assemble as a transcriptional complex in bronchial basal cells, as YAP, TEAD and TP63 physically interact and occupy similar chromatin binding sites and control a conserved gene expression program that strongly associates with PML progression. We mapped genes directly regulated by TEAD and TP63 by ChIP-seq and examined the gene expression consequences of siRNA-mediated gene silencing using RNA-seq. Directly induced genes of TP63 and TEAD encode factors involved in cell proliferation and extracellular matrix production, while directly repressed genes include genes associated with interferon downstream signaling and antigen presentation pathways. Our analysis of directly repressed TEAD-TP63 targets showed a particularly strong association between these immune modulating target genes and genes downregulated in the progressive PMLs, suggesting that TEAD-TP63 activity modulates immune function in the lung. Notable genes directly regulated by TEAD-TP63 included *CIITA*, which encodes a transcriptional transactivator that functions as a key regulator of MHC Class II genes. Our analyses across several datasets demonstrated that low *CIITA* expression is associated with progressive PML pathology and is negatively correlated with genes associated with Th1 cell activity, suggesting that epithelial control of MHC presentation by YAP-TEAD-TP63 modifies immune cell responses in early cancer development.

YAP and TP63 have been reported to associate in several contexts, including airway epithelial cells<sup>302,305,341,342</sup>, and the oncogenic functions of YAP/TAZ rely on their association with the TEAD family of transcription factors<sup>101,311</sup>. The physical and

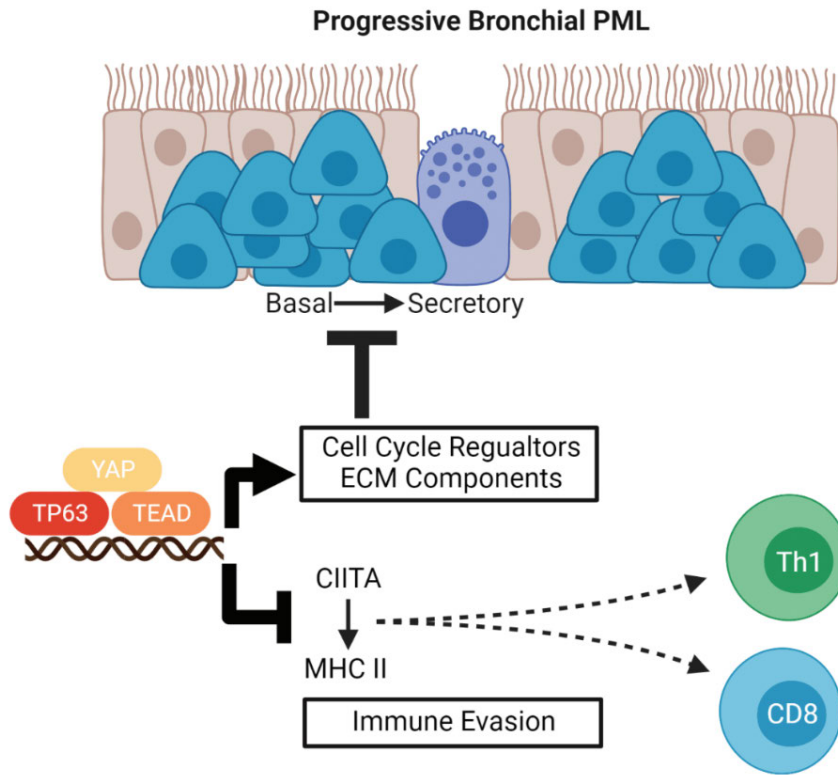


Figure 4.7. Summary diagram.

functional association of TEAD and TP63 suggest that the activity of YAP/TAZ-TEAD in basal cells is mediated by TP63, potentially directing the context of this complex to control lineage specific events. Our observations suggest that the cooperative activity of YAP/TAZ, TEAD, and TP63 induce the expression of genes that promote basal cell proliferation, as well as extracellular matrix components that may be supportive of basal cell self-renewal. Consistent with a pro-proliferative role, genes co-regulated by YAP/TAZ, TEAD and TP63 were strongly enriched among the genes in a proliferation-related co-expression module that was previously shown to be elevated in higher grade PML samples<sup>39</sup>. The regulation of such genes is consistent with observations that induced nuclear YAP/TAZ activity in mouse basal cells promotes basal cell expansion in vivo<sup>124</sup>; and observations that deletion of TP63 results in a loss of basal cells from the airways of mice<sup>297,298</sup>. Further, TP63 is frequently amplified in squamous cell carcinoma, and increased TP63 has been linked to YAP activation and alteration of TEAD binding<sup>352,353</sup>. Thus, an interconnected relationship exists between these factors that appears linked to the development of squamous carcinomas.

Our data suggest poorly explored functions for YAP/TAZ-TEAD-TP63 contribute to the progression of PMLs, including repression of immune modulating factors that may in turn lead to immune evasion. Previous studies have suggested that decreased levels of immune surveillance, particularly decreased interferon responses and antigen processing/presentation, is associated with progressive/persistent PMLs<sup>33,35,39</sup>. We found YAP/TAZ, TEAD, and TP63 directly repress many of the genes linked to immune surveillance in bronchial PMLs, including repression of genes involved in interferon

response and antigen presentation pathways. Interestingly, in multiple PML datasets we observed that decreased levels of anti-tumor immune cells, including cytotoxic CD8+ T cells, Th1 cells, NK cells, and activated dendritic cells was associated with decreased expression of genes that are repressed by YAP/TAZ-TEAD-TP63. Collectively these observations suggest that the repression mechanisms controlled by YAP/TAZ-TEAD-TP63 are central for PML progression.

The identification of the MHC II transactivator CIITA as a YAP/TAZ-TEAD-TP63 target gene was notable, particularly given the reported low expression of MHC class II genes in progressive bronchial PML<sup>33,35</sup> and the similar decreases observed in CIITA and MHC class II gene expression with poor immunotherapy responses in melanoma patients and in a rat model of breast cancer<sup>354,355</sup> and with promoting intestinal tumorigenesis<sup>356</sup>. Moreover, higher MHC Class II gene expression has been suggested as prognosis marker for colorectal carcinoma and triple-negative breast cancer survival<sup>357,358</sup>. Thus, the ability for YAP/TAZ-TEAD-TP63 to repress the expression of CIITA and MHC class II molecules in expanding basal epithelial cells may be a key mechanism of how early PMLs evade immune clearance.

Interestingly, many of the genes repressed by YAP/TAZ, TEAD, and TP63, including the MHC Class II genes, were highly expressed in airway secretory cells. This raises interesting questions about how the composition of the bronchial epithelium might influence immune-surveillance. Lung club cells have been shown to be crucial for the efficacy of radiation and immune checkpoint inhibitor combined therapy for non-small cell lung cancer<sup>344</sup>, and MHC class II expressing lung epithelial cells act as antigen-

presenting cells to direct CD4<sup>+</sup> T helper cell functions<sup>343</sup>. Thus, increased YAP/TAZ-TEAD-TP63 activity that favors the basal cell state would be associated with less immune infiltration and a worse prognosis in PMLs. Interestingly, similar stem-cell-like populations with high developmental plasticity and proliferation potential have been observed in adenocarcinoma and metastatic lung cancers<sup>240,359</sup>, suggesting possibly similar mechanism that couple cell fate with immune control.

Our data also proved the feasibility of using TEAD auto-palmitoylation inhibitor to reverse the TEAD-TP63 direct regulated gene expression associated with bronchial PML progression. We have also performed small compound screening testing the transcriptomic alterations associated with 10 different TEAD auto-palmitoylation inhibitors and measure the gene expression changes in HBECs. Three of these compounds were able to reverse the TEAD-TP63 direct regulated gene signatures similar to the MGH-CP1 experiments we showed above, including increasing the expression levels of HLA-DRA and CD74. Yet, due to the patent restrictions associated with these compounds, the results cannot be included in this thesis. Furthermore, inhibition of ERBB receptors has been shown to treat precancer airway lesion development<sup>124</sup>. Our connectivity analysis confirmed this observation and suggested combined MGH-CP1 and trametinib may provide additional treatment effect in intercepting early lung cancer progression. Meanwhile, other strategies may be effective in intercepting early lung cancer progression, including targeting TP63-mediated transcription or disrupting the binding between YAP, TEAD and TP63, which requires further investigation.

Collectively our results identify important roles for YAP/TAZ-TEAD-TP63 in the early

development of lung cancer, which notably includes immune-suppressive roles, and suggest that an assessment of the activity of this transcriptional complex may offer a means to identify immune evasive bronchial PMLs. Finally, targeting the YAP/TAZ-TEAD-TP63 complex may provide a therapeutic opportunity for intercepting early lung carcinogenesis, which is something that may be feasible given recent efforts that have been devoted towards developing YAP/TAZ-TEAD inhibitors<sup>87</sup>.

**CHAPTER 5 MICRORNA EXPRESSION DIFFERENCES IN NASAL  
EPITHELIUM FOR IDENTIFYING MALIGNANT INDETERMINATE  
PULMONARY NODULES**

***5.1 INTRODUCTION***

Lung cancer is the top cause of cancer prevalence of mortality in the US, causing over 230,000 new cases and nearly 132,000 deaths in 2021<sup>1</sup>. This can be largely attributed to the lack of detection of early-stage lung cancer when the survival rate is high. Results from the National Lung Screening Trial (NLST) showed that early detection through a low-dose computerized tomography (LDCT) can significantly reduce mortality<sup>14,170</sup>. However, the clinical management of indeterminate pulmonary nodules (IPNs), non-calcified nodules with 7–30mm size in diameter identified during CT screenings, is still challenging. While high-risk patients are followed-up with invasive procedures, such as tissue collection using bronchoscopy, and the low-risk ones are followed-up with repeat CT, those classified as intermediate-risk represent a diagnostic dilemma<sup>360</sup>. With the majority of IPNs being benign<sup>151,361</sup>, unnecessary invasive follow-up procedures increase the risk of complications, cost, and anxiety<sup>158</sup>. Thus, there is a unmet need for a diagnostic test that can accurately identify those patients with malignant IPNs from the majority of benign cases.

The field of injury describes the molecular alterations associated with carcinogen exposure or lung cancer development and can be observed in the normal-appearing airway throughout the respiratory tract<sup>144</sup>. Based on this principle, our group has previously built a genomic classifier for lung cancer diagnosis based on gene expression

profiles in the bronchial epithelium<sup>155-157</sup>, suggesting the clinical utility of transcriptional profiles from normal-appearing airways. Similar field effects can be observed in the nasal epithelium as well. Our group previously demonstrated that the smoking-related gene expression alterations in the nasal epithelium are similar to those observed in the bronchial epithelium<sup>159</sup>. Consistent gene signatures between nasal and bronchial epithelium have also been reported for COPD<sup>160</sup> and IPF<sup>161</sup>. Moreover, Perez-Roger *et al.* reported a nasal lung cancer classifier for lung cancer detection among ever smokers and highlighted the utility of the more readily accessible nasal epithelium for cancer risk stratification<sup>162</sup>. These results suggested that the airway cancer field extends to the nasal epithelium and the nasal gene expression measured through nasal brushing could serve as a non-invasive alternative to bronchoscopy for lung cancer diagnosis.

miRNA are a class of small non-coding RNA that suppresses gene expression through base-pairing between the seed region and the 3' untranslated regions. miRNA-mediated gene expression regulation plays important role in almost all biological processes, including cancer<sup>72,362</sup>. The field of injury concept can be extended to miRNA expression. Previous studies from our group showed that the bronchial miRNA expressions and the regulated gene signatures reflect the effects of cigarette smoking and are involved in lung cancer development<sup>84,85</sup>. Notably, the study by Pavel *et al.* demonstrated that miRNA expression alterations in the bronchial epithelium can be used to improve the gene-based biomarker for the early detection of lung cancer<sup>166</sup>.

Given this prior evidence, we hypothesize that nasal miRNA expression can be utilized for the early detection of lung cancer and the classification of IPNs. In this study, we used

matched gene and miRNA expression profiles from the nasal airway brushings collected from DECAMP I subjects to investigate the miRNA expression alterations associated with IPN malignant status. Furthermore, by combining top scoring pairs between miRNAs and the predicted target genes and building an ensemble learning model, we evaluated the performance of an integrated biomarker.

## **5.2 METHODS**

### *5.2.1 Study enrollment and sample collection*

Nasal airway epithelial brushings were collected from 235 ever smokers who had incidental pulmonary nodules between 7-30mm on chest CT from the DECAMP I cohort (NCT01785342). Subjects were followed up for up to two years after the nasal brushing collection until a final diagnosis of lung cancer status was made.

### *5.2.2 miRNA sequencing and processing*

The miRNA processing procedures were the same as described in Aim 1. Briefly, we used NEBNext Multiplex Small RNA Sample Prep Kit (Illumina) to construct the sequencing library. The samples were then sequenced using Illumina HiSeq 2000. The quality of the sequencing reads was examined with fastqc<sup>191</sup>. Reads shorter than 15 nt were excluded. Then, the reads were aligned to human reference miRNAs in miRBase v22 and quantified using miRDeep2 with default settings<sup>193,194</sup>. All quality metrics were summarized with multiqc<sup>363</sup>. A two-step quality control strategy was used to filter out

samples with poor quality. First, given that the majority of reads in the library should be from miRNAs that are 21-23 bps long, we filtered by sequencing quality removing samples with a read mapping percentage < 70% or an average read length lower than 20bp. Samples were then TMM normalized using edgeR<sup>277</sup> and transformed into log CPM. Lowly expressed miRNAs were removed if the sum of counts across all samples were lower or equal to 1. Next, we excluded a sample if it had at least two outlier measurements (2 standard deviations from the mean of all samples) among PC1, PC2, RIN, and between-sample correlation. The counts for the remaining samples were normalized again as described above. 189 samples remained after the two-step filtering (65 benign and 124 cancer) with 681 miRNAs. 154 samples had matched nasal gene expression profiles after quality filtering<sup>209</sup>.

### *5.2.3 Derivation of miRNA signatures and TSPs associated with cancer status*

Voom<sup>195</sup> was used to transform miRNA counts and Limma<sup>197</sup> was used to identify miRNA expression alterations associated with IPN malignant status, adjusting for smoking status, batch and RIN. Similarly, the smoking-associated miRNAs were identified using linear regression models adjusting for cancer status, batch, and RIN. To identify miRNA that is actively regulating gene expression, we also compared the density of correlations between a miRNA's expression profile and the expression profiles of all of its predicted targets with the density of correlations between that miRNA and all of the genes that are not predicted to be targeted by that miRNA. Three databases were used for target gene identification (TargetScan (v7.2; conserved targets), miRDB (v5.0),

and miRTarBase (v7.0)<sup>71,201,364</sup>), and target gene or a miRNA was defined as those identified by at least two databases. The Pearson correlation between miRNAs and genes was calculated using the residual expression value adjusted for technical variables (batch and RIN). The distribution of correlation coefficients of all predicted target genes of a miRNA was compared to the distribution of all the genes that were not predicted targets using the KS-test. miRNAs with significant KS test results (FDR < 0.05) and a negative shift in the correlation distribution with predicted target genes compared with genes that were not target were then selected for differential expression analysis.

TSPs were derived in two formats based on miRNA and the predicted target gene expression values. Both the miRNA and gene expression data were normalized by log CPM and batch-corrected before TSP calculation. Binary TSPs were calculated as whether the expression level of the predicted target gene was higher than the miRNA (true = 1 and false = 0). Continuous TSPs were calculated as the difference between the batch-corrected log CPM target gene and miRNA expression values. Fisher exact test was used to identify the cancer-associated binary TSP and two-group t-test was used for the continuous TSP.

The association between the predicted target genes of cancer-associated miRNAs or the genes involved in cancer-associated TSPs and the activities of previously derived gene signatures were examined using GSEA<sup>204</sup>.

#### 5.3.4 Ensemble Learning Pipeline

A novel ensemble learning model framework was proposed to integrate metrics from different feature types for better classification of lung cancer status. The DECAMP-I samples with both miRNA and gene expression profiles after QC were randomly split 4-fold into discovery and validation cohorts (N=116 and 38; **Table B.9**). Next, we performed the feature and model selection within the discovery cohort with 100-time repeated 5-fold cross-validation. Within each iteration, we first performed feature selection in the inner training dataset with each miRNA-related feature type, including cancer-associated differential expression miRNAs, and binary and continuous TSPs. The redundant and highly correlated significant features were filtered out using Boruta<sup>365</sup>. Then, with the selected features in each feature type, we trained weak learners using various machine learning models, including naïve-Bayesian (NB), random forest (RF), linear discriminative analysis (LDA), supported vector machine with different kernels (SVM), logistic regression (LR) and gradient boosting machine (GBM). A clinical classifier, including smoking status, pack-year, age and nodule size, was also trained separately at this step. The prediction probabilities from each model were extracted and were used to train the ensemble model, including RF, tree bagging, Adaboost, GBM, and LR. `findCorrelation` function from caret R package<sup>366</sup> was used before ensemble model training to remove highly correlated weak learners. We run the whole CV pipeline using the Pearson correlation cutoff from 0.5 to 0.95 at 0.05 increments. Synthetic Minority Oversampling Technique (SMOTE) was performed to up-sample the benign samples during the model training steps to alleviate the class imbalance. Finally, the final model

was chosen based on the highest average AUC in the inner-testing dataset across CVs. The top features and the weak learners were selected based on their selected frequencies during CV and the average number of selected features/learners for the final ensemble model. The entire training data set was used to re-train the final model, and the validation cohort was used for independent performance evaluation.

### **5.3 RESULTS**

#### *5.3.1 Study Population*

We removed those samples with poor sequencing quality (average sequencing length less than or miRNA mapping percentage less than 70%) or samples that were significantly different from the rest (based on the first two principal components, RIN, and between-sample correlation). The baseline characteristics of the remaining 189 samples (65 benign and 124 cancer), including 138 samples from last year and 51 samples, were summarized in **Table 5.1**. Technical variables such as batch or RIN were not significantly different between the benign and the cancer groups.

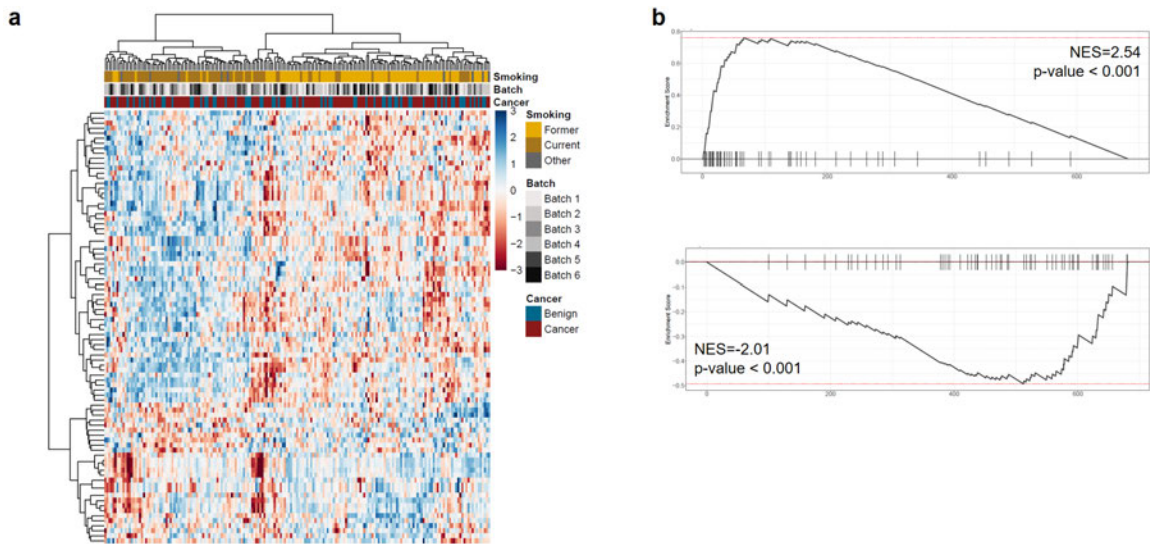
#### *5.3.2 miRNA expression signature and miRNA-gene TSPs associated with IPN status*

We first explored the smoking-associated miRNAs and validate previous findings in the DECAMP-I nasal miRNA expression profiles to examine the quality of our data. Differential expression analysis showed 86 miRNAs whose expression levels were significantly different between current and former smokers, adjusted for cancer status,

Variables	Benign (N=65)	Cancer (N=124)	Statistics	p-value
<b>Batch</b>				
1	12 (20.0%)	27 (21.8%)	4.67	0.46
2	16 (24.6%)	23 (18.6%)		
3	16 (24.6%)	26 (21.0%)		
4	4 (6.15%)	13 (10.5%)		
5	13 (21.5%)	23 (18.6%)		
6	2 (3.1%)	12 (9.7%)		
<b>RIN</b>	5.6 (1.75)	5.09 (1.73)	1.89	0.062
<b>Gender (=Male)</b>	51 (78.5%)	94 (75.8%)	0.86	0.72
<b>Race</b>				
African American	10 (15.4%)	29 (23.4%)	2.97	0.23
White	45 (69.2%)	84 (67.7%)		
Others	10 (15.4%)	11 (8.8%)		
<b>Age (year)</b>	66.1 (7.79)	69.8 (8.38)	-3.05	<b>0.003</b>
<b>Nodule Size (cm)</b>	1.21 (0.59)	1.66 (0.56)	-4.98	<b>&lt; 0.001</b>
<b>Pack-year</b>	50.42 (27.43)	51.23 (24.9)	-0.20	0.84
<b>Smoking Status (=Current)</b>	27 (41.5%)	55 (50.82%)	0.83	0.84

**Table 5.1. DEAMP I nasal miRNA sample clinical annotation by cancer status.**

Statistical tests for categorical clinical variables (batch, gender, race and smoking status) were conducted using Chi-square tests. Statistical tests for continuous variables (RIN, age, nodule size and pack-year) were compared using two-sided Student's t-tests. Percentages are reported for categorical variables and mean/standard deviations are reported for the continuous variable.



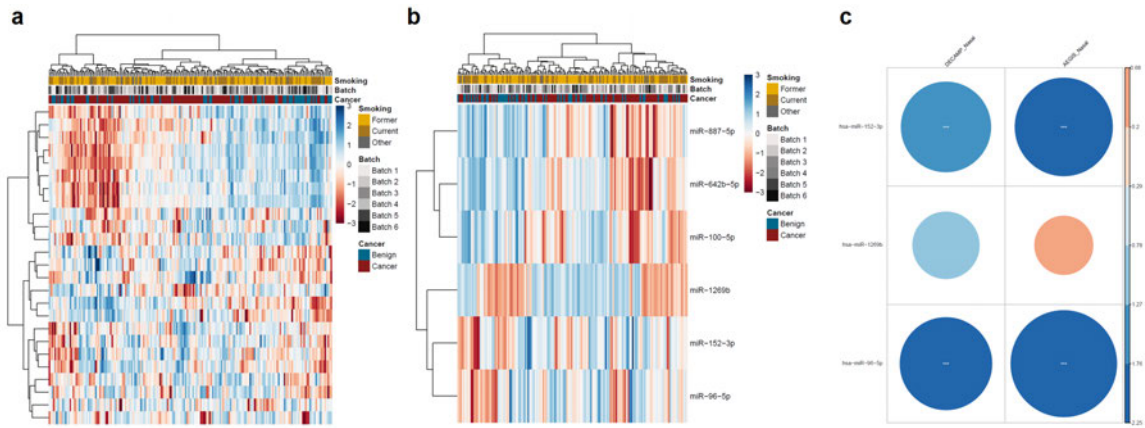
**Figure 5.1. Smoking-associated miRNAs in the nasal epithelium of DECAMP I samples were enriched within the bronchial smoking-associated miRNA signature.**

**a.** Heatmap of smoking-associated miRNAs residual expression values adjusted for cancer status, batch and RIN in the DECAMP I nasal miRNA expression profiles (FDR < 0.05). The miRNA expression values were scaled by row. The top color bars indicate the sample smoking status, batch and cancer status. **b.** Enrichment plot of miRNAs upregulated (top) or downregulated (bottom) in the current compared to former smokers in the AEGIS bronchial brushing dataset among all miRNAs ranked by their t-statistics associated with smoking status in the DECAMP I nasal dataset.

batch, and RIN (**Figure 5.1a**; linear model, FDR < 0.05). miRNA up- and down-regulated with current smoking in the AEGIS bronchial miRNA sequencing dataset were significantly enriched with the miRNAs rank list in the DECAMP-1 dataset (**Fig 5.1b**; GSEA p-value < 0.0001). Among these, we found miRNAs whose expression levels were previously indicated to be associated with smoking, such as miR-218-5p, miR-365a-3p/5p, and miR-181<sup>84,166</sup>. These results suggested the data quality is fine and the smoking-related field of injury could be observed in the distal nasal epithelium.

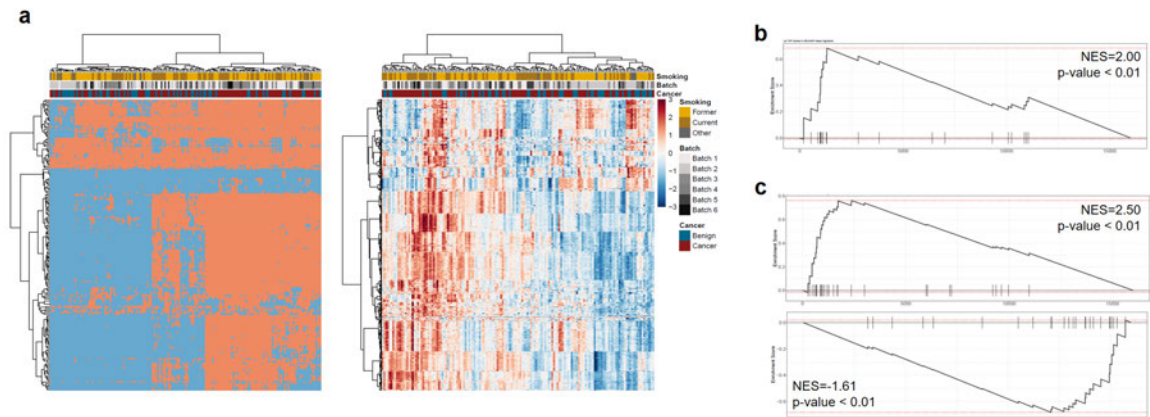
Next, we aimed to identify cancer-associated miRNA in the nasal epithelium. Differential expression analysis identified only 25 miRNAs whose expression levels were significantly altered between the malignant and the benign IPN subjects, suggesting a weak cancer signal in the nasal epithelium (**Figure 5.2a**; linear model, p-value < 0.05).

To overcome the signal issue, we filtered for miRNAs whose expression levels were more negatively correlated with the predicted target genes and may be actively regulated gene expressions based on KS-test using samples with both miRNA and gene expression data. Of the 114 miRNAs that passed the filter (KS-test, FDR < 0.05), 6 were significantly altered in cancer compared to the benign samples (**Figure 5.2b**; linear model, p-value < 0.05). Notably, the significantly negative correlated predicted target genes of miR-152-3p and miR-96-5p, both up-regulated in the cancer samples, were significantly enriched among the genes that were decreased in the nasal epithelium of cancer subjects in both DECAMP-I and AEGIS cohort (**Figure 5.2c**; GSEA, p-value < 0.01). These observations suggested the nasal miRNA expressions may reflect cancer-associated alterations.



**Figure 5.2. Predicted target genes of the cancer-associated miRNAs in the DECAMP I nasal samples were dysregulated in previously derived nasal cancer-associated gene signatures.**  
**a-b.** Heatmap of cancer-associated miRNAs residual expression values, before (a) or after (b) the selection of miRNAs more negatively correlated with predicted target genes than non-target genes, adjusted for smoking status, batch and RIN in the DECAMP I nasal miRNA expression profiles ( $p$ -value  $< 0.05$ ). The miRNA expression values were scaled by row. The top color bars indicate the sample smoking status, batch and cancer status. **c.** Bubble plot of the predicted target genes significantly negative correlated with miRNAs that were down-regulated in cancer compared with benign samples in DECAMP I or AEIGS bronchial genes ranked by t-statistics of association with cancer status.

Given the relatively weak association between miRNA expression levels and cancer status in the nasal epithelium, we have also identified cancer-associated top-scoring-pairs (TSPs) in miRNA expression profiles. The method aims to identify a pair of features whose relative relationship can be used for binary classification<sup>367</sup>. Compared with standard expression level-based classifiers, a TSP is more robust to data normalization and can potentially generate classification features beyond the miRNA or gene expression alone. We constructed 7414 non-constant binary and 22090 continuous TSPs between miRNA and the predicted target genes among samples with both miRNA and mRNA sequencing data. 189 binary TSPs and 875 continuous TSPs were significantly associated with the IPN cancer status (**Figure 5.3a**; Fisher Exact test and two-group t-test  $p$ -value  $< 0.05$ ). The binary TSP scores were all higher in the cancer sample, and the involved genes were significantly enriched amongst the genes whose expression increased in DECAMP-I nasal gene expression profiles of the subjects with malignant IPNs (**Figure 5.3b**; GSEA,  $p$ -value  $< 0.01$ ). Similarly, the genes involved in the continuous TSPs that were elevated in the cancer subjects were also enriched among the up-regulated genes in the cancer samples, whereas genes from the continuous TSPs that decreased in cancer samples showed opposite enrichment (**Figure 5.3c**; GSEA  $p$ -value  $< 0.01$ ). Notably, the miRNAs involved in these TSP were not strongly overlapped with those cancer-associated miRNAs. These results suggested TSP may carry independent cancer-associated signals and highlighted the potential of utilizing feature combinations for better cancer diagnosis.



**Figure 5.3. miRNA-target gene top-scoring pairs reflected cancer-associated gene alterations.**

**a.** Heatmap of cancer-associated miRNA-target gene binary (left) and continuous (right) TSPs in the DECAMP I nasal miRNA expression profiles ( $p$ -value < 0.05). The miRNA expression values were scaled by row. The top color bars indicate the sample smoking status, batch and cancer status. **b-c.** Enrichment plot of predicted target genes involved in binary (b) or continuous (c) TSPs among all genes ranked by their  $t$ -statistics associated with cancer status in the DECAMP I bronchial dataset.

### *5.3.3 Development of an integrated ensemble learning model for the IPN diagnosis*

Finally, we aimed to combine clinical and miRNA features to build an integrated classifier for the early detection of lung cancer and compare its performance with the classifier containing clinical variables only (**Figure 5.4**). The ensemble learning pipeline allowed combining information from different feature types and models while keeping the training process separate. Thus, the most suitable model could be trained for each type of feature to utilize independent information and achieve overall better performance. Based on the AUC from cross-validation, we selected the random forest model as the final model with the Pearson correlation cutoff of 0.95. This model contained 12 weak learners constructed separately on four feature types: 3 from clinical variables, 3 from miRNA expressions (N=7), 2 from binary TSPs (N=34), and 4 from continuous TSPs (N=16) (**Figure 5.5a**). No strong correlation was observed between or within feature types within the discovery cohort, suggesting the redundant features were removed during the cross-validation (**Figure 5.5b**). Meanwhile, the Pearson correlation of the cancer status prediction probabilities between the weak learners constructed on clinical variables and miRNA features were lower than 0.5, suggesting the miRNA-based learners carried additional information for the classification.

The ensemble model was used to predict cancer status for the samples in the validation cohort (**Figure 5.5c**). The model performance was compared to a GBM classifier built with clinical variable only, which achieved the best performance during CV among all weak learners. The clinical variable classifier had an AUC of 0.64 (95% CI: 0.46-0.78), with a low sensitivity of 0.43. In contrast, the ensemble classifier achieved an AUC

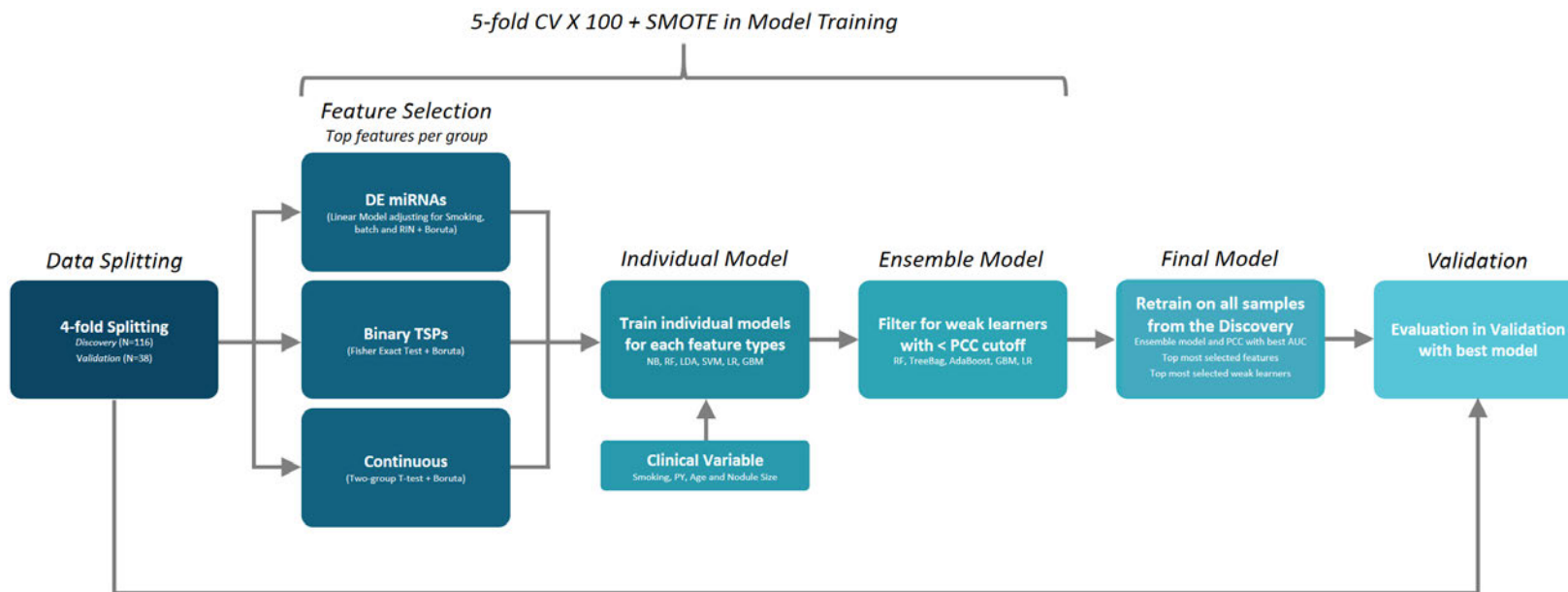
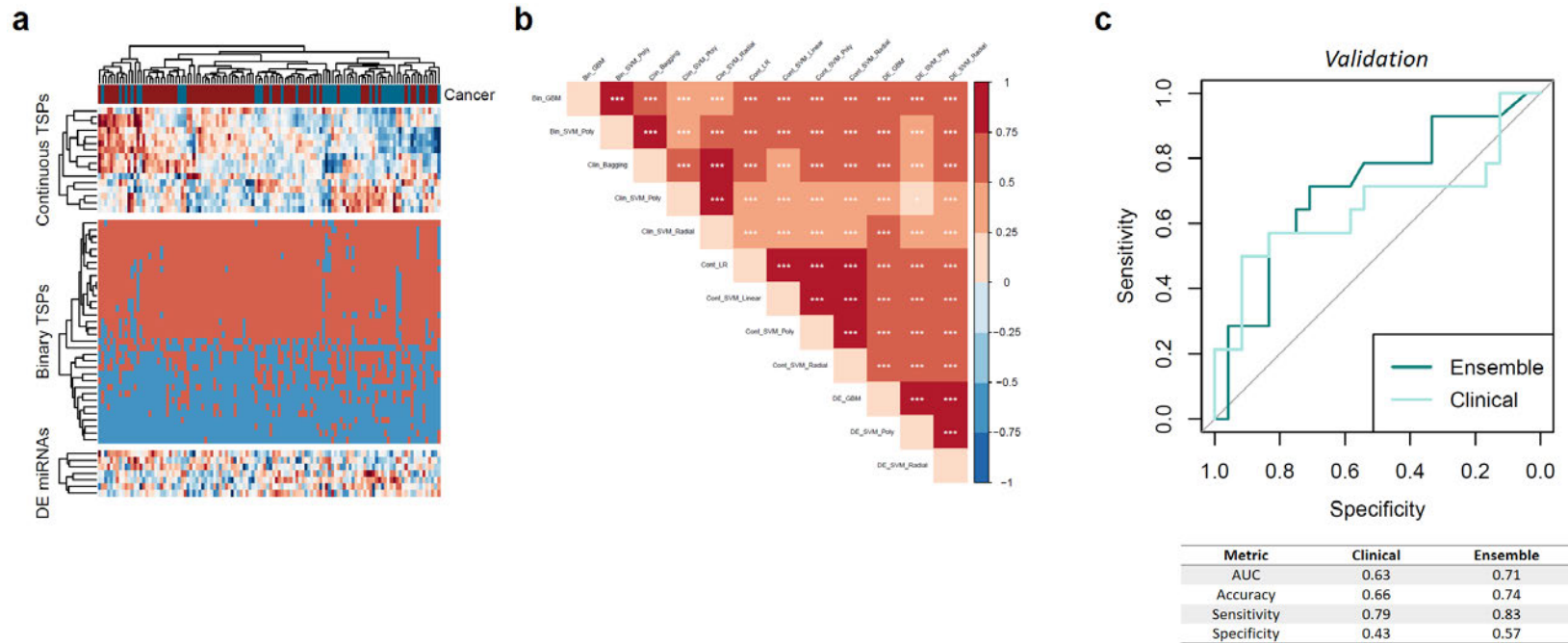


Figure 5.4. Diagram for the ensemble learning model training pipeline and cross-validation schema.



**Figure 5.5. The performance of ensemble classifier integrating clinical and miRNA features in differentiating malignant and benign IPNs.**  
**a.** Heatmap of levels cancer-associated miRNAs, binary and continuous TSPs selected for the final ensemble classifier in the samples of the DECAMP I discovery cohort. The feature values were scaled by row. The top color bars indicate the sample cancer status. **b.** Correlation plot showing for the Pearson correlation coefficients between the cancer status prediction probability of all weak learners of the ensemble classifier in the DECAMP I discovery cohort. **c.** ROC curve and prediction performance for the ensemble classifier and the classifier based on clinical features only in the DECAMP I validation cohort.

higher than the clinical classifier at 0.71 and an accuracy of 0.74 (95% CI: 0.57-0.87), although no significant difference based on DeLong's test ( $p$ -value = 0.2). Notably, the ensemble classifier was able to achieve much better specificity at 0.57 while maintaining the sensitivity at 0.83. In summary, by utilizing nasal miRNA expression alterations and combining weak learners built on different feature types, we were able to slightly increase the predictive ability to facilitate the detection of malignant IPNs.

#### ***5.4 DISCUSSION***

To solve the unmet clinical need to discriminate the malignant IPNs from the majority of benign ones, we examined the matched nasal gene and miRNA expression profiles of high-risk smokers with IPNs from the DECAMP I cohort. Our study extends from the previous project in two important aspects. First, the AEGIS trial used in the previous nasal genomic classifier study<sup>162</sup> enrolled high-risk smokers who were undergoing bronchoscopy for suspicion of lung cancer, more than of whom had nodules greater than 30mm. In contrast, the DECAMP I cohort included patients from broader risk backgrounds and with nodule sizes between 7-30mm. Given the lower risk of lung cancer in the population that DECAMP represents, reducing the potential complications through the less invasive nasal swab sampling procedure became particularly critical. Secondly, previous work for the diagnosis of lung cancer was conducted utilizing miRNA expression data in the bronchial epithelium. We demonstrated, that with the integrated model, the miRNA alteration in the distal field could be utilized for early cancer detection, which further supports the field of cancerization hypothesis and serves as a

proof-of-concept for building a minimally invasive integrated biomarker.

While we initially observed relatively weak cancer-associated signal in the nasal miRNA expression profiles, two strategies were leveraged to improve the signature derivation: filtering miRNAs that are more negatively correlated with the predicted target genes comparing the non-target ones and constructing top-scoring-pairs using miRNAs and the predicted target genes. These strategies identified stronger cancer-associated signals that were likely biological meaningful. Notably, the target genes of the cancer-associated miRNAs and the genes involved in the TSPs that were up-regulated in samples with malignant IPNs were enriched among the genes whose expression were reduced in bronchial epithelium of cancer patients in both DECAMP I and AEGIS trial, an alteration that was associated with the immune system signaling<sup>162</sup>. Thus, cancer-associated miRNA signature and TSPs may suggest that immune alterations can be observed in the nasal epithelium as well. The underlying mechanism and function of such change could be an interesting question for future investigation.

Furthermore, we were able to build an integrated predictive model with both miRNA-related and clinical features for discriminating malignant and benign IPNs by integrating the nasal epithelial miRNA-related and clinical features. The high sensitivity of this classifier in the validation dataset highlights its potential clinical utility as a rule-out test. Particularly within the screening setting where the majority of IPNs are benign, our classifier could be used to select those patients who are at higher risk for lung cancer and should be follow-up with bronchoscopy. A similar model training framework was previously utilized to build a clinical-genomic classifier for lung cancer risk

stratification<sup>368</sup>. Here, we proposed a novel ensemble model training framework in which various weak learners can be constructed with features from different data modalities (differentially expressed miRNAs, TSPs, and clinical features) separately the orthogonal information from different features could be better utilized. The modular nature of this framework could be potentially useful when incorporating additional data and features in our future work.

Two issues emerged during the classifier development process. First, the cancer signal captured by nasal miRNA expression was weak. We alleviated this issue by leveraging gene expression data and constructing TSP to boost the signal. Second, there were almost twice as many malignant IPNs as the benign samples in the DECAMP-I dataset, causing the severe class imbalance issue. As a result, the classifier tended to have high sensitivity but very low specificity. By implementing the up-sampling procedure in the ensemble learning pipeline, we were able to achieve reasonable specificity in the validation while maintaining high sensitivity. Notably, the miRNAs are usually the weakest predictive features, compared to either gene expressions or imaging features. We believe the performance could be further improved by adding in results from gene expression and imaging analysis.

In summary, our work demonstrated the clinical utility of nasal miRNA for the early diagnosis of malignant IPNs and risk stratification for lung cancer. By integrating miRNA and clinical features with a novel ensemble learning classifier, the ensemble classifier achieved better performance than the clinical features-only biomarker. While this study is limited by the relatively small sample size and the nasal gene-based classifier

and imaging feature classifier are still work-in-progress, we believe the framework proposed could be potentially helpful for future projects.

## CHAPTER 6 GENERAL CONCLUSIONS AND FUTURE DIRECTIONS

Collectively, the work in this dissertation thesis investigated the transcriptional regulation in the respiratory tract epithelium to facilitate the understanding, interception, and diagnosis of early lung cancers.

PMLs are the presumed precursors of LUSC. Previous publications from our group suggested the lack of anti-tumor immune cell infiltration is associated with the progression of PMLs. In Chapters 2 and 3, we examined the miRNA regulatory landscape associated with various PML phenotypes to better understand the early molecular events that drive PML progression. Notably, we identified an airway basal-specific miRNA, miR-149-5p, that may suppress the cytotoxic T cell activity by repressing expression of the MHC Class I regulator NLRC5. This finding reveals an interesting cell-cell communication between the epithelial basal cells and immune cells, whose alteration may drive the progression of PML. Also, we developed a novel computational framework and R package DReAmiR to identify miRNA-mediated regulatory network rewiring events and to better understand the molecular differences between PML molecular subtypes. Current work leveraging human tissue samples and multiplexed imaging analysis is being conducted to validate the computational findings and explore the spatial pattern of the miRNA and gene expressions.

In Chapter 4, we investigated the transcriptional crosstalk between the Hippo and TP63 pathways, two important oncogenic pathways, in the context of bronchial PML.

Integrating ChIP-seq and RNA-seq data, we identified highly overlapped chromatin-binding profiles and concordant direct regulated gene signatures between the Hippo

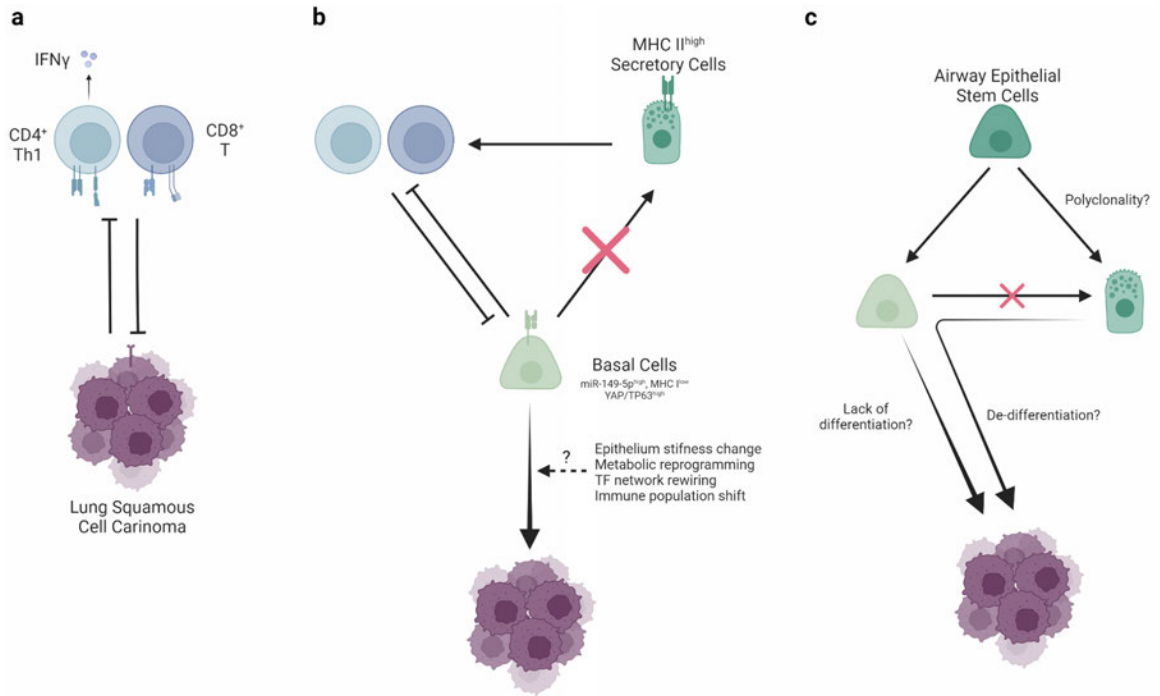
pathway effector and TF, YAP and TEAD, and the TP63. Together, YAP-TEAD-TP63 direct regulated genes contribute to early lung carcinogenesis by promoting bronchial airway basal cell proliferation and suppressing cell differentiation and immune surveillance. Notably, YAP-TEAD-TP63 together represses the expression of HLA-DRA and CD74, genes encoding the components of the MHC II complex, which are highly expressed in the epithelial secretory cell population under normal conditions.

Furthermore, small compound screening showed that the overlapped signatures can be reversed by blocking the DNA binding ability of TEADs through an auto-palmitoylation inhibitor. These findings offer potential insight into the transcriptional regulatory mechanism of early lung carcinogenesis and suggest disruption of the DNA binding ability of TEADs as an early lung cancer progression interception strategy.

Finally, there is a currently unmet need to discriminant the malignant IPNs found during chest CT screening for better clinical management. In Chapter 5, we studied the miRNA expression profiles in the nasal epithelium collected from patients with IPNs. We derived cancer-associated nasal miRNA signatures and miRNA-target gene TSPs and built an ensemble classifier to predict the cancer status of the IPN patients. The ensemble classifier achieved better performance than the classifier built with clinical features alone. This work provided a useful framework for future research to include predictive features from other data modalities and to build an integrative biomarker for the early detection of lung cancer.

Theodosius Dobzhansky stated in his 1973 essay that “nothing in biology makes sense except in the light of evolution”<sup>369</sup>. Similarly for early lung cancer development, not only

do we see an evolution of predator-prey interaction between immune and tumor-initiating basal cells, but also a potential competition between basal and other airway epithelial cells. More generally, as an extension of the field cancerization theory, the results from this thesis highlighted the critical role of normal epithelial cells within the early lung cancer microenvironment that can affect the outcomes of PMLs. Immune evasion in the tumor microenvironment has been proved to be an important mechanism through which tumor cells grow and acquire therapeutic resistance<sup>370,371</sup>. Traditionally, the mechanism can be summarized as the mutual competition between the malignant tumor cells and immune cells: immune cells recognize and eliminate tumor cells while tumor cells utilize immune checkpoints to inhibit immune cell activities (**Figure 6a**). However, the work from this thesis showed a dysregulated multi-cellular program consisting of not only the tumor-initiating cells (airway basal cell in lung squamous carcinoma) and immune cells, but also the non-malignant epithelial cells, particularly the secretory cells (**Figure 6b**). In conjunction with the lack of antigen presentation on the surface of the premalignant basal cells (Chapter 2), the absence of the MHC II expressing secretory cell population (Chapter 4) is contributing to the immune evasions of bronchial PMLs as well. Furthermore, this model indicates that the intact immune surveillance in bronchial premalignancy might be compromised as the epithelium acquires a more stem-cell-like state. Thus, rather than a simple shift of cellular composition (i.e. lack of immune infiltration) during early lung carcinogenesis, the balance of different epithelial cell subsets may be the root cause that drives early lung carcinogenesis and could be a novel therapeutic research angle.



**Figure 6. The schematic diagram for overall conclusions and future directions.**

Given that some of the aforementioned observations can be explained by the relative abundance of multiple cell states, rather than the simple presence or absence of a particular cell type, it is crucial to develop a computational algorithm to facilitate the discovery and quantification of the multi-cellular programs. Single-cell data offered the ability to probe dysregulated gene programs and cell states in diseases at high resolution and recent advances in the ligand-receptor analysis provided methods to characterize cell-cell interaction<sup>334,372,373</sup>. Yet, these methods rely strongly on known ligand-receptor gene expression without modeling the underlying genetic program of each cell group. The results were often affected by biological and technical variabilities and were limited between a pair of cell types. Meanwhile, tools like EcotypeR<sup>374</sup> and DIALOGUE<sup>375</sup> started to give insight into the interaction between cell types but lack mechanistic explanations. Therefore, we propose to develop a method that directly models multi-cellular gene programs (such as the malignant-normal-immune cell interaction) and examines their alterations in cancer settings. While extensive research is still needed, an initial conceptual design is described here. The key component of the tool is to construct meta-cell clusters of different cell states that are linked by cellular differentiation trajectories and/or cell-cell communication factors (ligand-receptor pair, cytokines, etc). Affinity matrix with adaptive Gaussian kernel can be used to then scale and harmonize the cell-cell distance measured by expression similarities or cell-cell communication strength. Graph-based clustering algorithm or archetypal analysis can then be performed to identify meta-cell clusters. Then, a Bayesian hierarchical model can be used to identify the latent gene programs within each cluster incorporating sample-level information, such

as disease status and sequencing batch<sup>376,377</sup>. Alternatively, canonical correlation analysis or similar methods could be used to identify co-varying gene programs across cell types. Combining this approach with cell-type deconvolution tools that provide cell-type-specific expression profiles, we can further project a multi-cellular program and calculate a gene-set score to capture both cell-state variance and gene expression changes from the bulk sequencing samples. Furthermore, recent studies showed some cell programs, including the levels of epithelial MHC I/II and the immune cell activations, can form spatially distinct structures that are associated with cancer phenotypes<sup>378,379</sup>. Under the Bayesian framework, the physical locations of cells and the spatial distance between cell clusters can be incorporated intuitively as priors to better characterize the multi-cellular programs.

Another question that remains is to elucidate the altered transcriptional regulatory networks that accompany the transition from premalignant lesions to carcinoma *in situ*. While the effective immune responses have been proved essential in lung cancer treatment<sup>380,381</sup>, several discrepancies exist comparing our findings in bronchial PMLs to those observed from invasive lung cancers. First, we observed higher expression levels of immune checkpoint genes, including PD-1/PD-L1, CD80/CTLA4, and LAG3, among the regressive proliferative lesions, among which immune infiltration is high. Second, while NLRC5 expression was associated with better bronchial PML outcome, it was associated with a lower 5-year survival rate in TCGA lung cancer samples<sup>234</sup>. Furthermore, the correlation between immune-related gene module (Module 9) levels and PML progression status was almost reversed in the CIS samples<sup>374</sup>. These discrepancies

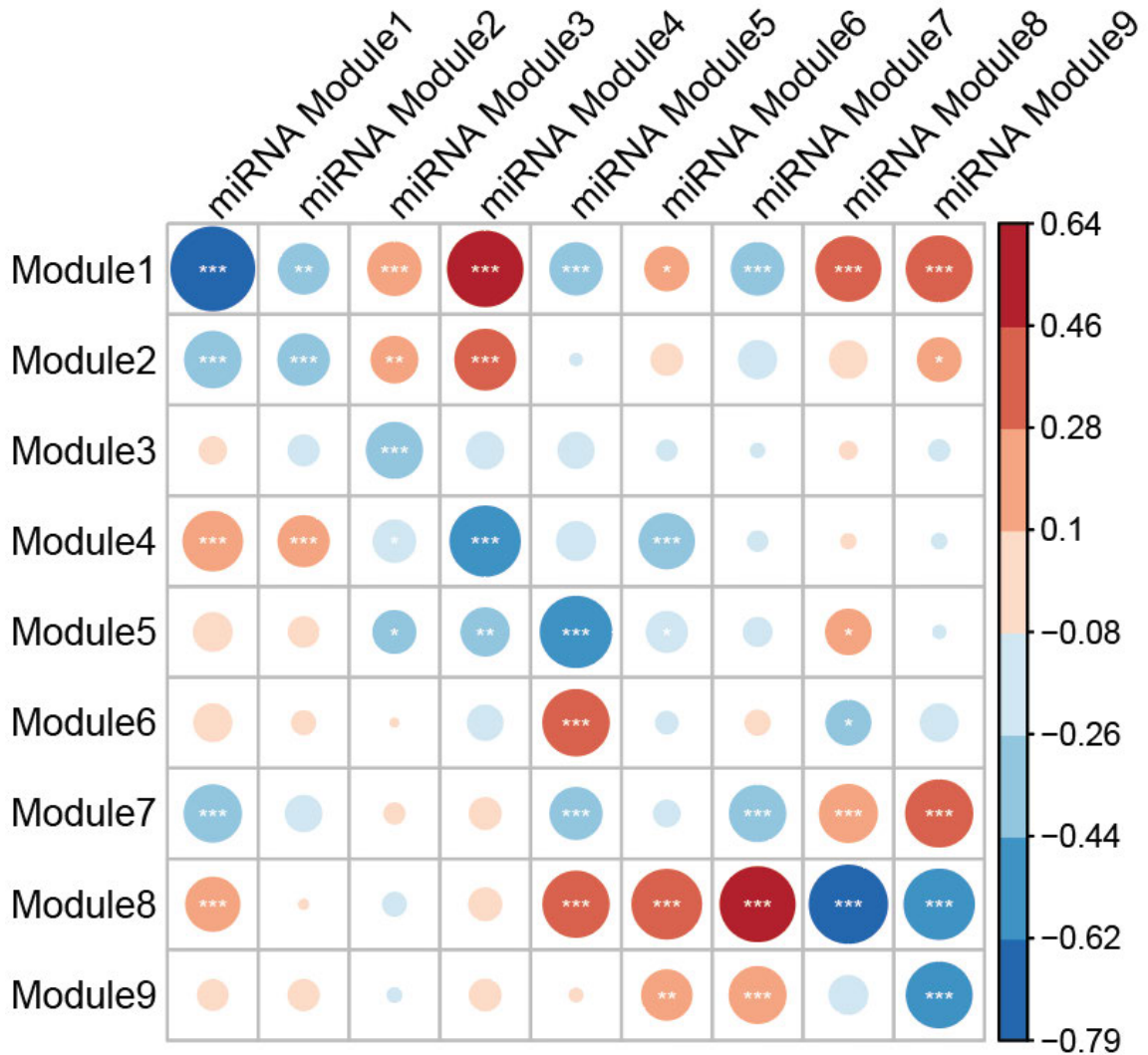
indicate that a mechanistic switch may take place during the transition from PML to CIS and alter the association between gene expression and phenotype. One explanation is that the immune response against bronchial PMLs is dominated by MHC I/II-mediated immune recognition and activation, rather than the immune checkpoint-mediated cytotoxic activity repression. Thus, a high checkpoint level may simply suggest the presence of immune cells, rather than their function or exhaustion states. Ongoing multiplexed imaging-based immune characterization could potentially help to disentangle this dilemma. Meanwhile, several other potential mechanisms exist and require further investigation, including changes in cell stiffness<sup>382,383</sup> and metabolic reprogramming<sup>384</sup>. The key obstacle to fully understanding the full spectrum of PML progression is the lack of an adequate animal model. Our collaborators showed the disruption of the Hippo pathway in mice bronchial airway basal cells may lead to cancer, but similar genetic modification in secretory cells would only result in high-grade dysplasia (data not published). Combining the genetically engineered mice model with NTCU treatment<sup>385</sup>, and *in vivo* CRISPR-Cas9 screen could be performed to evaluate how each gene contributes to histological progression, particularly their dependence on the Hippo pathway.

Finally, it is critical to explore how the normal epithelial cell population affects early cancer progression with the lens of epithelial development. Work from this thesis suggests the ability of bronchial PMLs to progress to high-grade dysplasia is tightly associated with the loss of epithelial lineage fate specification, supporting the observation that cancer cells acquire high epigenetic and phenotypic plasticity<sup>240,359,386</sup>. Yet, it is

unclear whether the tumors arise solely from a basal cell population that has differentiation capability during early cancer development, or it could be also from a differentiated cell population that undergoes a de-differentiation process (**Figure 6c**). Also, given the importance of enhancer rewiring in cell fate specification, functional genomics characterization, including accessible genomic region or histone modification characterizations, in the model system, combined with computational methods such as DReAmiR, could reveal further details. Another question that remains is what is the clonal origin of the pro-immune secretory cells in relation to the malignant cells. Recent studies suggested somatic mutations and even cancer driver mutations can be widely observed within normal tissues<sup>387</sup>, which form distinct normal clones around the premalignant epithelial cells<sup>388,389</sup>. In the esophagus, these normal cells acquire mutations, including NOTCH mutations, to increase their fitness against premalignant cells, provide anti-tumor effects, and might even indicate better patient outcomes<sup>390</sup>. Similarly, work from our lab has also revealed high prevalence of NOTCH mutations among high-grade dysplasia samples which were not significantly associated with the progression status (data not published). Given these observations, there might exist a normal airway epithelial clone that harbors cancer-driving mutations and proliferates to compete with and eliminate the malignant clones in bronchial airways. To answer this question, the ability to combine cell lineage barcoding and spatial single-cell sequencing techniques to trace clonal evolution is needed. Furthermore, if the normal and premalignant epithelial cells indeed evolve from separate clones, it is unclear whether the neoantigens presented by MHCs are from the malignant clones and can trigger an

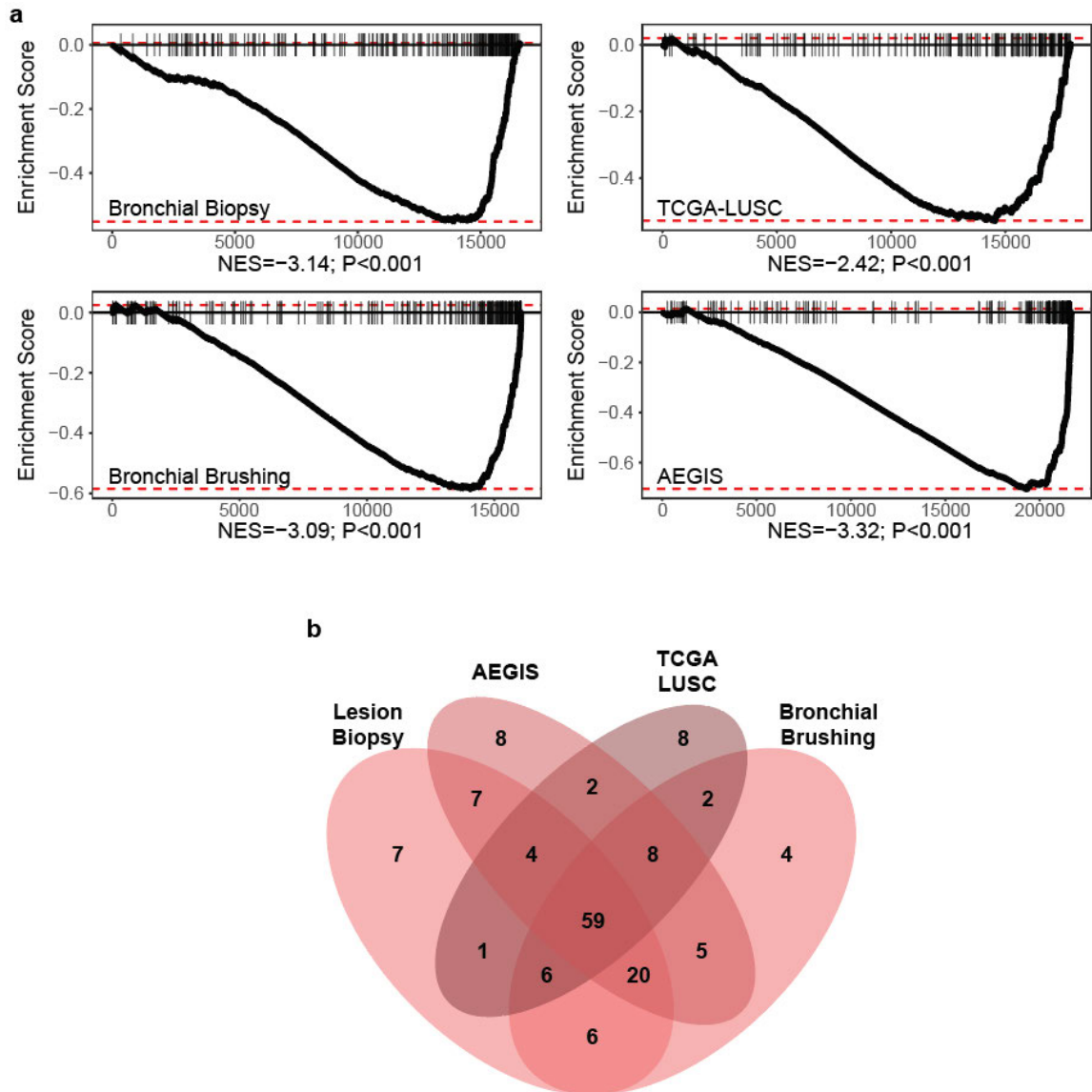
effective cytotoxic response. Thus, it would be interesting to characterize the immune repertoire via TCR-seq and probe the antigen bound by MHC I/II complex via immunopeptidomics<sup>391</sup> within the cell lineage resolved models.

## APPENDIX A SUPPLEMENTARY FIGURES



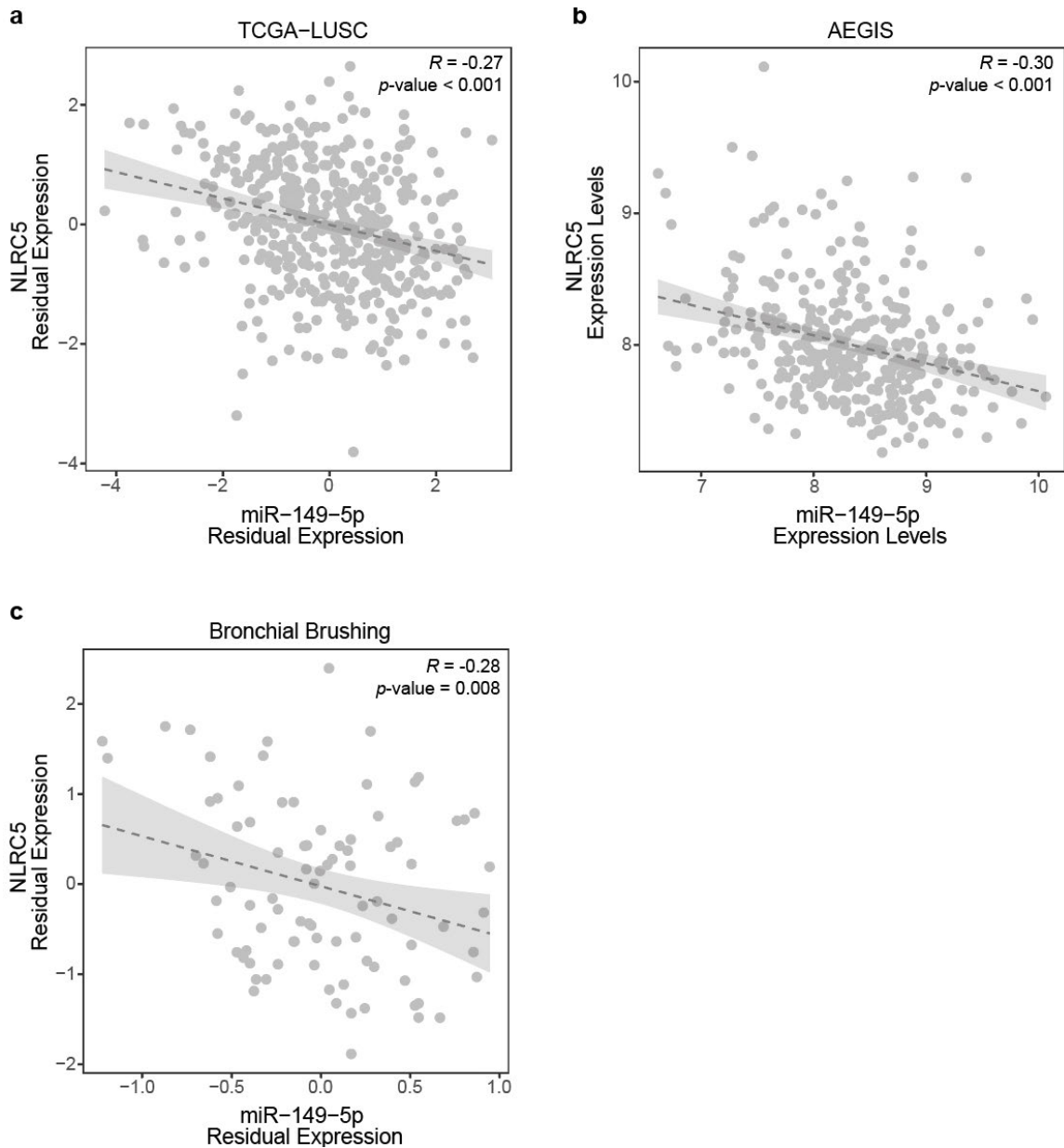
**Figure A.1. GSVAs scores of miRNAs associated with a gene module were negatively correlated with the gene module metagene scores.**

Bubble plots of the correlation between GSVAs scores of miRNA associated with a gene module, from the miRNA-gene module network, and the gene module metagene scores. \* FDR  $\leq 0.05$ ; \*\* FDR  $\leq 0.01$ ; \*\*\* FDR  $\leq 0.001$ .



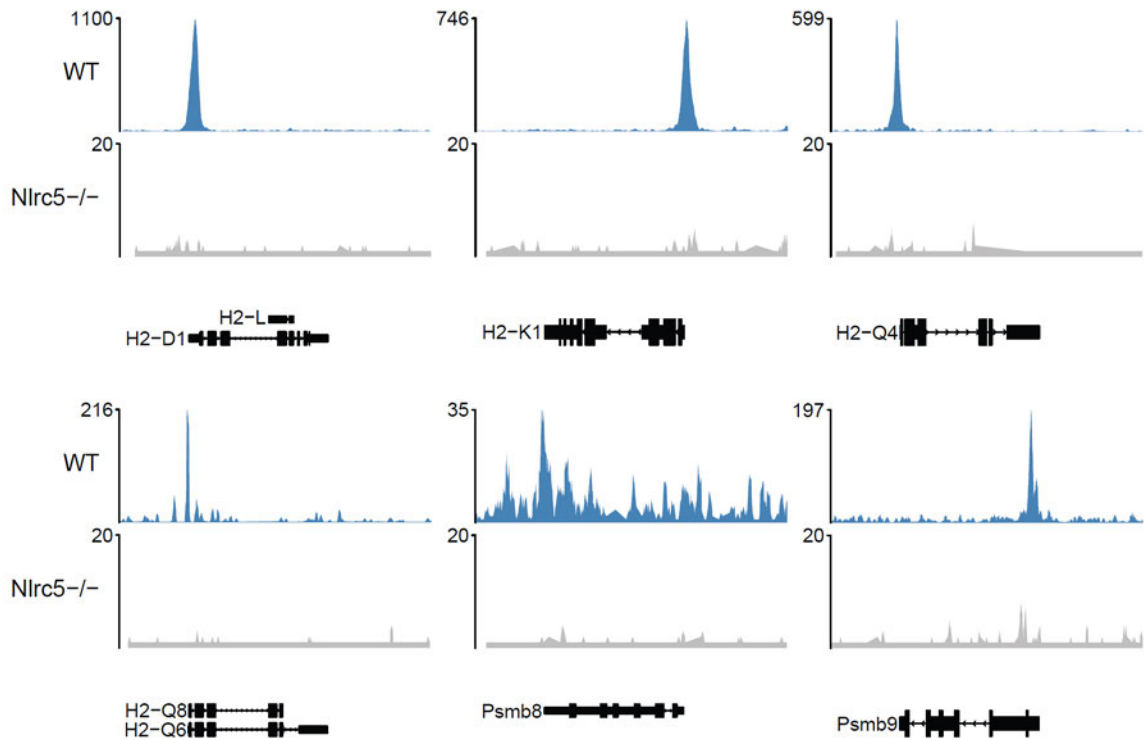
**Figure A.2. Genes of the immune-related gene module were enriched among the genes negatively correlated with miR-149-5p.**

**a.** Enrichment plot of genes of the immune-related gene module (N=224) among all genes ranked by their expression level correlation with miR-149-5p across four datasets: (from top to bottom) lesion biopsy samples, bronchial brushing samples, TCGA-LUSC primary tumor samples, AEGIS bronchial brushing samples. **b.** Overlap of leading-edge genes from **a** between four datasets.



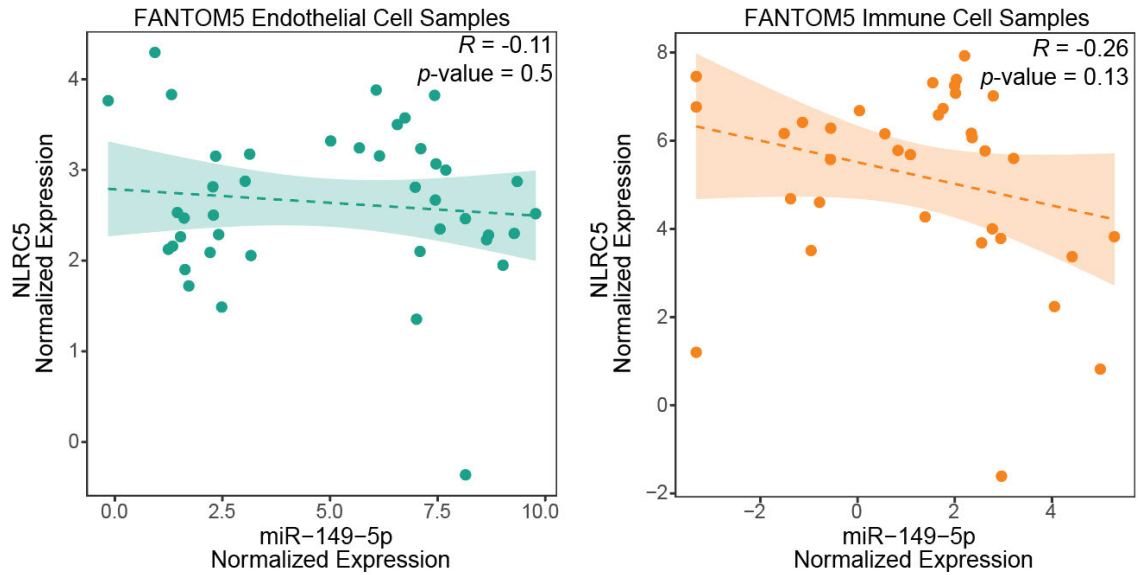
**Figure A.3. The expression level of miR-149-5p was significantly negatively correlated with that of NLRC5 across datasets.**

Scatter plots show the expression level correlation between miR-149-5p and NLRC5 across three validation datasets. Linear fitness was shown as a dashed line and 95% CIs were shown as shade. There was a significantly negative correlation, calculated using Pearson correlation, between miR-149-5p and NLRC5 in TCGA-LUSC and AEGIS bronchial brushing datasets.



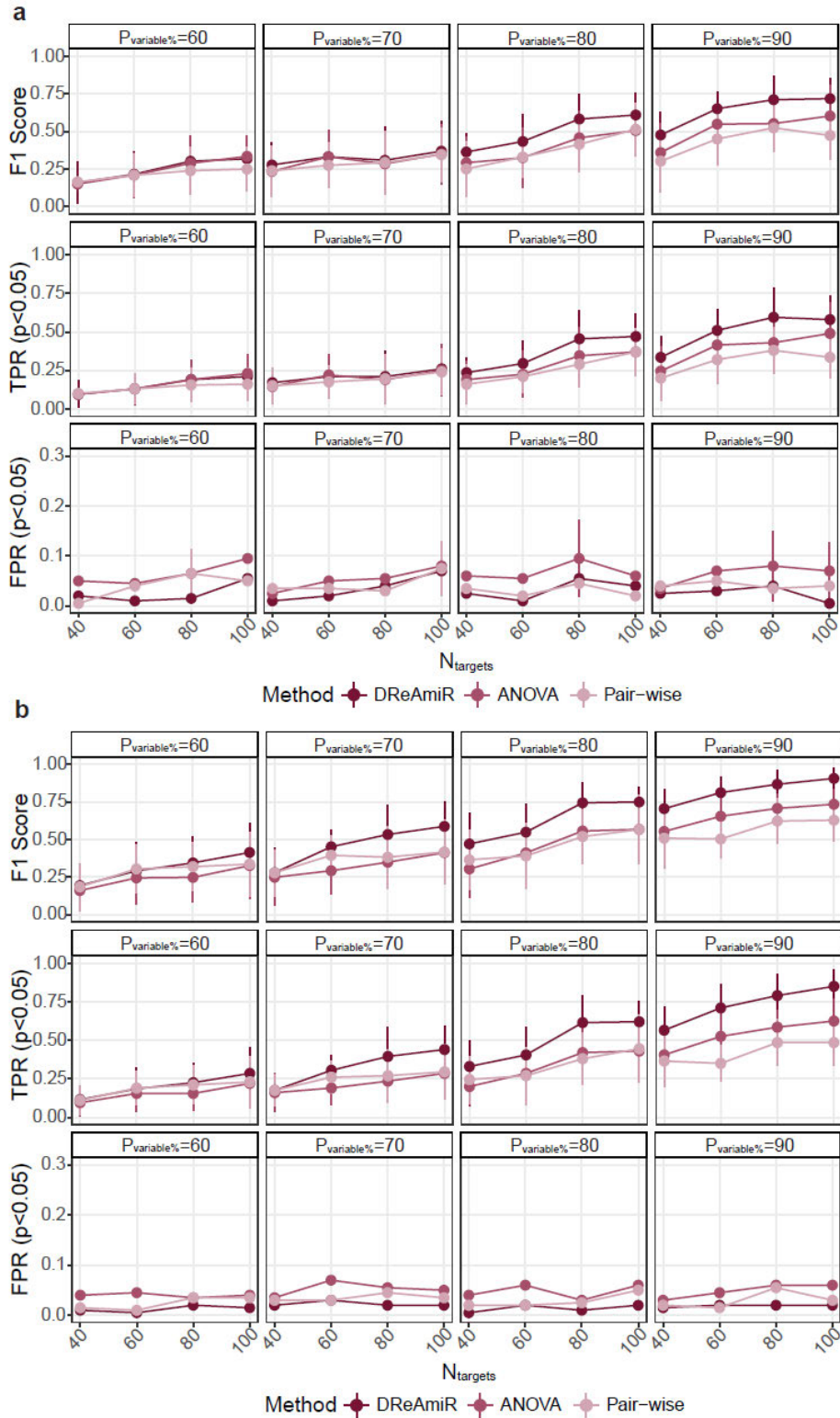
**Figure A.4. NLRC5 regulates the expression of MHC Class I genes.**

WT and Nlrc5<sup>-/-</sup> ChIP-seq tracks from (GSE59092) show NLRC5 binding at the promoter regions of H2-D1, H2-K1, H2-Q4, Psmb8, Psmb9, and H2-Q8.



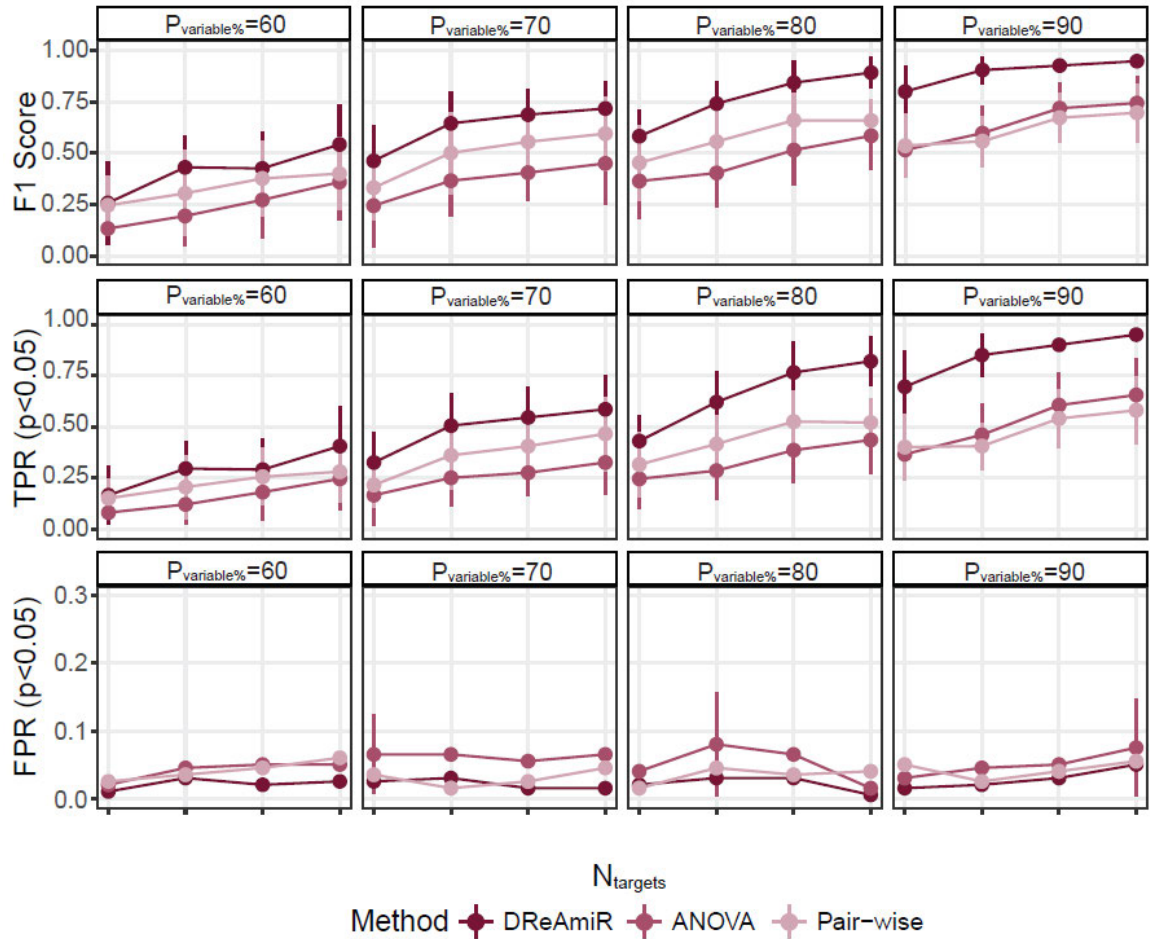
**Figure A.5. Correlation between miR-149-5p and NLRC5 expression level within the samples from the FANTOM5 project.**

**a-b.** Scatter plot of the correlation between the normalized expression levels of miR-149-5p and NLRC5 within the FANTOM5 samples belong to endothelial (**a**; N=39) and immune (**b**; N=29) cell compartments. Linear fitness was shown as dashed lines and 95% CIs were shown as shade.

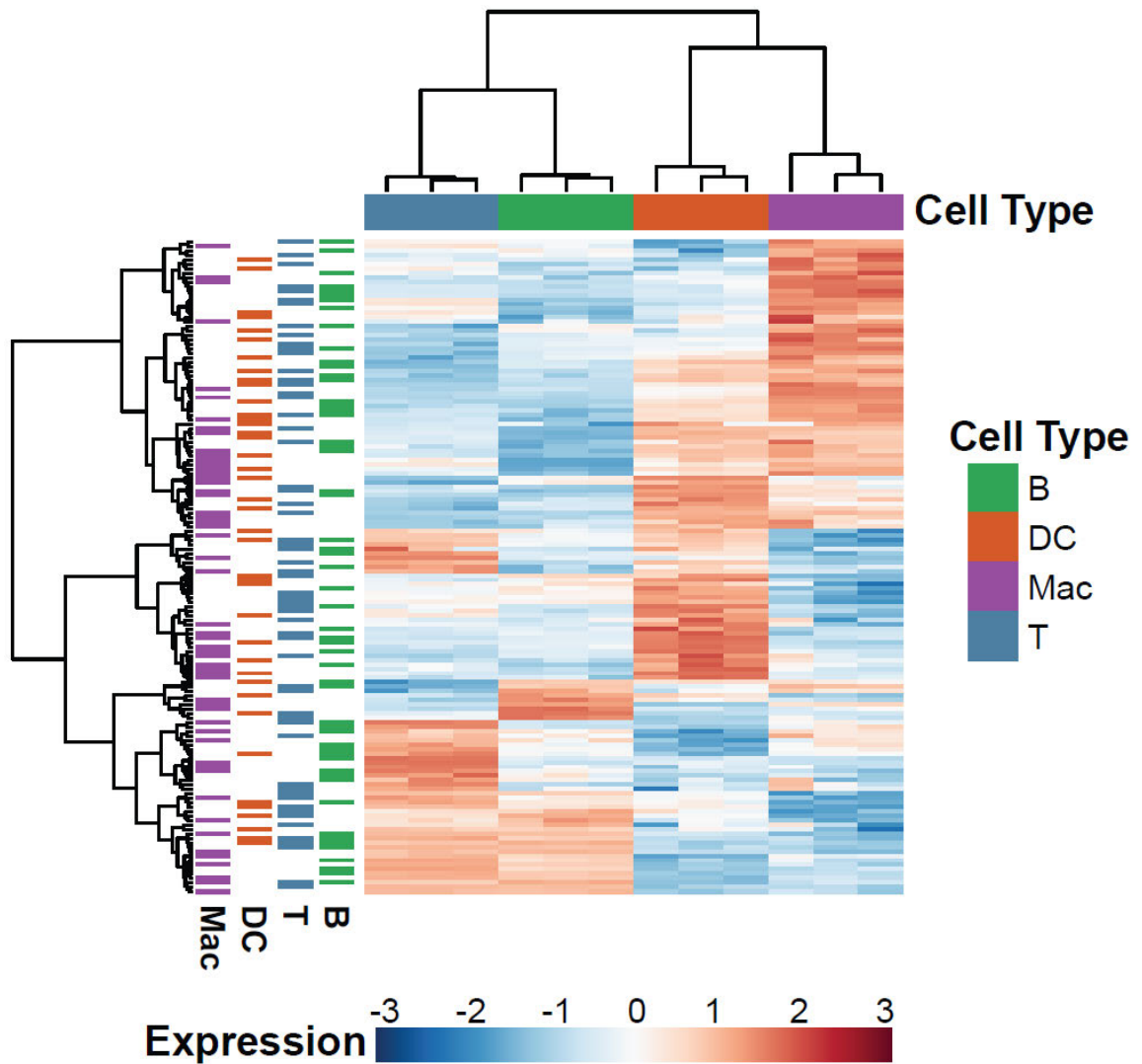


**Figure A.6. DReAmiR performance in simulated data.**

DReAmiR performance in simulated data with  $N_{sample}$  equaled 30 (a) and 50 (b).

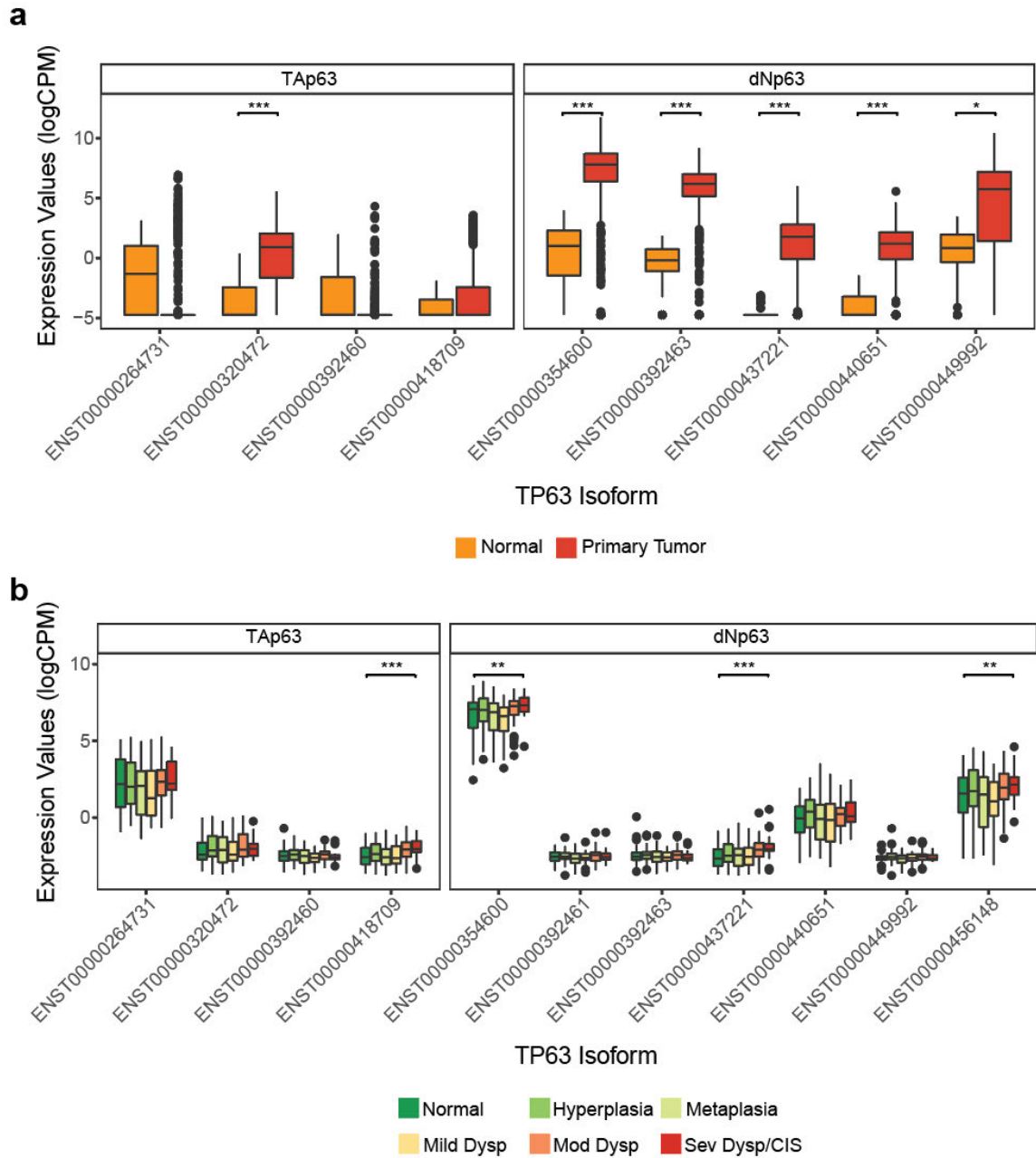


**Figure A.7. DReAmiR performance in simulated data with different sample sizes across groups.** DReAmiR performance in simulated data with different sample sizes across groups ( $N_{sample} = 50, 60$  and  $70$ ).



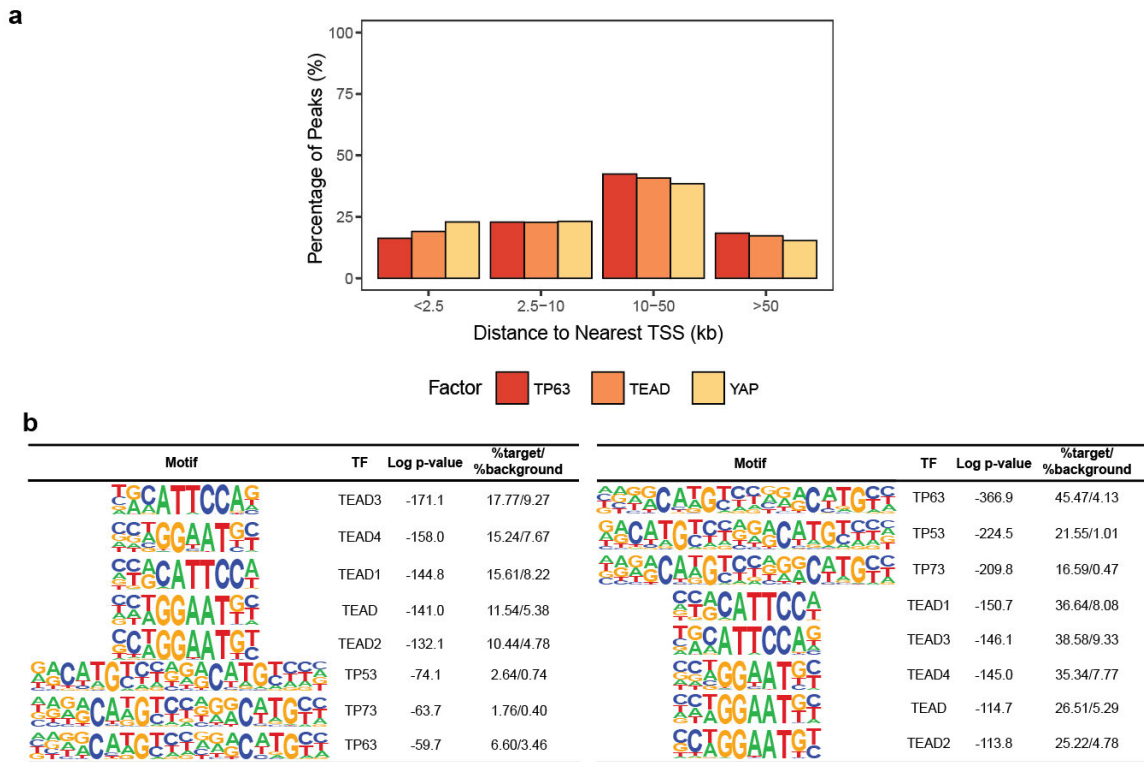
**Figure A.8. The expression level of miR-155 prioritized target genes across four mice immune cell-types of the control group.**

The color bar on the top showed the cell-types from the mice of the control group, and the color bar on the left of the heatmap showed which cell-type the prioritized target gene belonged to. The expression values were scaled per gene.



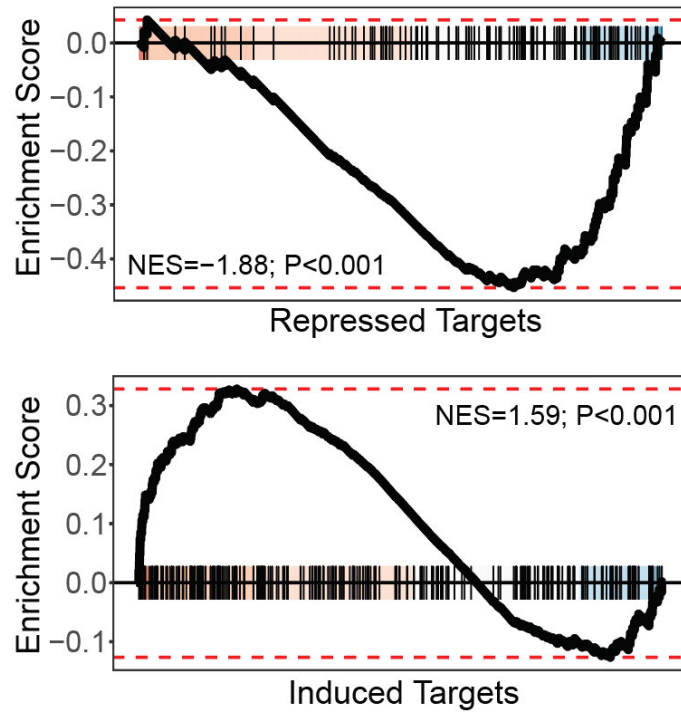
**Figure A.9. TP63 isoform expression levels in TCGA-LUSC and in bronchial PML biopsy data.**

**a.** Boxplots show all TP63 isoforms expression levels between normal and primary tumor samples in TCGA LUSC. Only those with significant increase in primary tumors are marked. \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ . **b.** Boxplots show all TP63 isoforms expression levels across bronchial PML histological grades in Beane *et al.* \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ . **c.** Boxplots show the TEAD1/2 expression levels across bronchial PML histological grades in Beane *et al.* \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

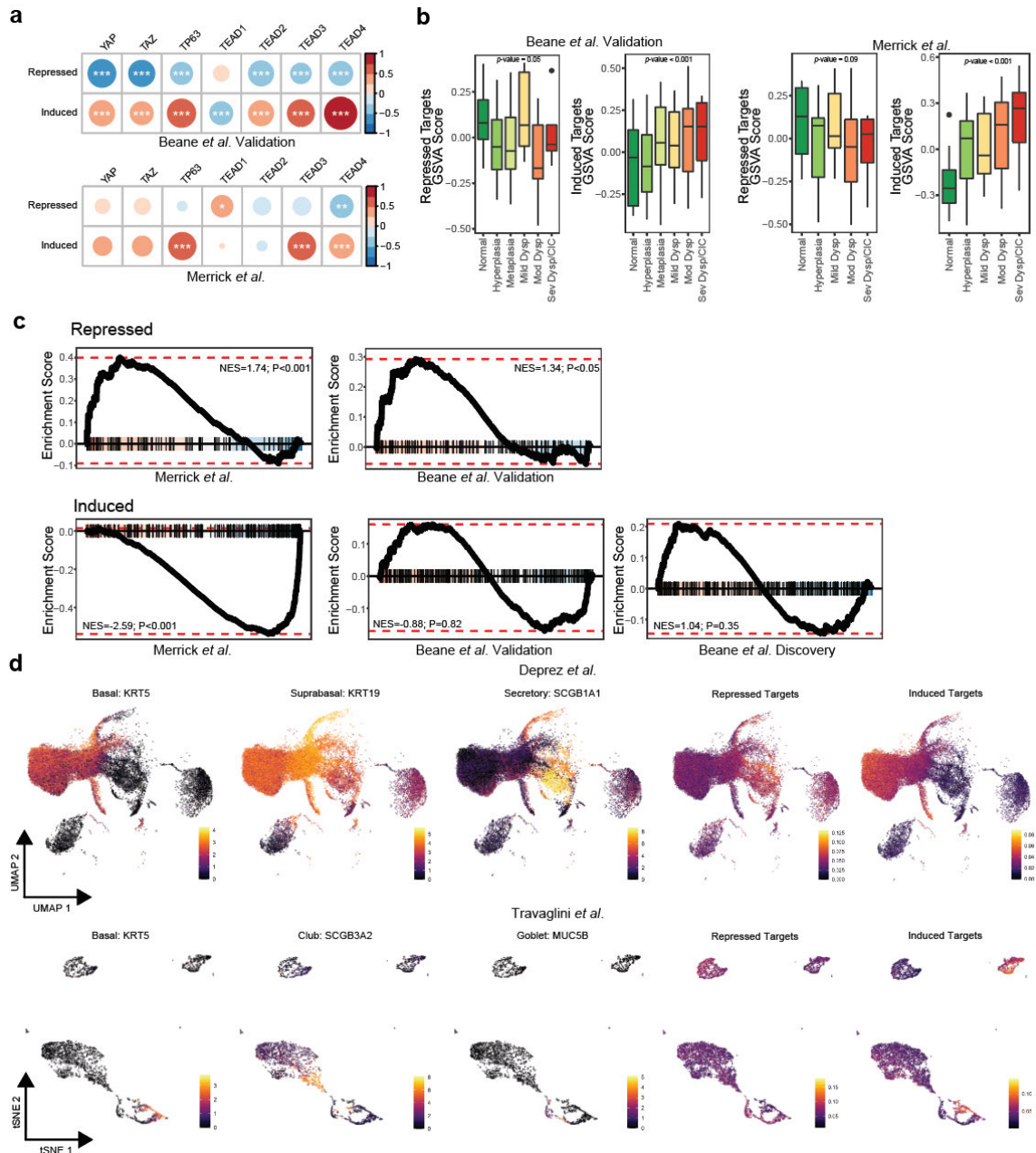


**Figure A.10. ChIP-seq analysis of YAP/TEAD/TP63 chromatin binding profiles.**

**a.** Distribution of YAP/TEAD/TP63 peaks by distance between peak locations and nearest TSS. **b.** Top transcription factor binding motifs enriched in the YAP (left) and YAP/TP63 (right) co-binding sites in HBECs. Only unique motifs are shown. P-values were calculated by HOMER.



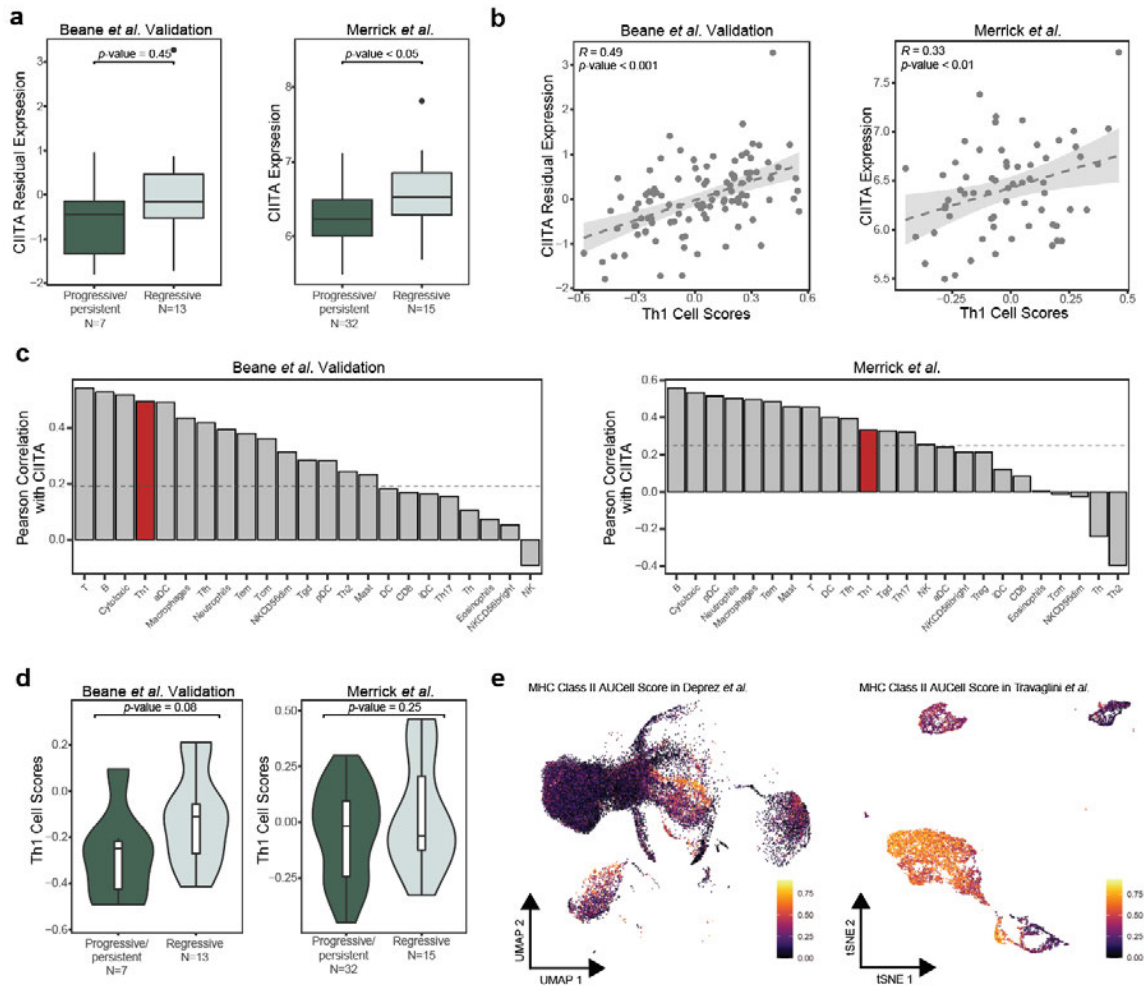
**Figure A.11. Transcriptomic analysis of TEAD-TP63 direct regulated target genes.** Enrichment plots for TEAD-TP63 repressed (top) and induced (bottom) target genes among genes ranked by t-statistic for their association with siLATS treatment in HBECs (GSEA; p-value <0.005).



**Figure A.12. Transcriptomic analysis of TEAD-TP63 direct regulated target genes in human bronchial PML data and lung scRNA-seq data.**

**a.** Correlation plot shows the correlation between the expression levels of transcription factors and metagene scores of TEAD-TP63 direct induced and repressed target genes (calculated with GSEA) in Beane *et al.* Validation cohort (top) and Merrick *et al.* (bottom). The color and the size of the circles indicate the Pearson correlation coefficients. \*\*\*Pearson correlation,  $p$ -value  $< 0.005$ . **b.** The metagene scores of TEAD-TP63 direct repressed and induced target gene sets across human bronchial PML data by histological grades in Beane *et al.* Validation cohort (left) and Merrick *et al.* (right). **c.** Enrichment plot for TEAD-TP63 direct repressed (top) and induced (bottom) target genes among genes ranked by  $t$ -statistics comparing the regressive PML samples to the progressive/persistent ones of the Proliferative subtypes in the Beane *et al.* Discovery/Validation cohort or comparing regressive to progressive/persistent bronchial

PML samples in Merrick *et al.* **d.** (Top) UMAP plots show the cell-type marker genes and AUCell scores for TEAD-TP63 direct induced/repressed target gene sets in the healthy human airway scRNA-seq data from Depez *et al.* (Bottom) tSNE plots show the cell-type marker genes and AUCell scores for TEAD-TP63 direct induced/repressed target gene sets in and human lung scRNA-seq data from Travaglini *et al.* Only the epithelial cells are shown.



**Figure A.13. Analysis CIITA in human bronchial PML data and lung scRNA-seq data.**

**a.** (Left) Expression level of CIITA in progressive/persistent and regressive PML samples of the Proliferative subtype in Beane *et al.* Validation cohort. (Right) Expression level of CIITA in progressive/persistent and regressive PML samples in Merrick *et al.* **b.** Scatter plots show the Pearson correlation between the expression level of CIITA and Th1 scores (calculated using GSVA based on genes from Bindea *et al.*) in Beane *et al.* Validation cohort (left) and Merrick *et al.* (right). **c.** Immune cell-type ranked by their Pearson correlation coefficients with CIITA expression levels in Beane *et al.* Validation cohort (Left) and in Merrick *et al.* (Right). The dashed line indicates the Pearson correlation coefficient that reaches  $p$ -value = 0.05. **d.** (Left) Th1 cell scores in progressive/persistent and regressive PML samples of the Proliferative subtype in Beane *et al.* Validation cohort. (Right) Th1 cell scores in progressive/persistent and regressive PML samples in Merrick *et al.*. **e.** (Left) UMAP plots show the MHC Class II gene metagene scores calculated with AUCcell in the healthy human airway scRNA-seq data from Deprez *et al.* (Right) tSNE plots show the MHC Class II gene metagene scores calculated with AUCcell in and human lung scRNA-seq data from Travaglini *et al.* Only the epithelial cells are shown.

## APPENDIX B SUPPLEMENTARY TABLES

Gene Module Number	miRNA
Module 1	hsa-let-7f-5p, hsa-miR-103a-3p, hsa-miR-1299, hsa-miR-1301-3p, hsa-miR-1307-3p, hsa-miR-130b-3p, hsa-miR-130b-5p, hsa-miR-144-3p, hsa-miR-15a-5p, hsa-miR-15b-3p, hsa-miR-17-5p, hsa-miR-182-5p, hsa-miR-183-5p, hsa-miR-184, hsa-miR-18a-5p, hsa-miR-191-5p, hsa-miR-19a-3p, hsa-miR-200a-3p, hsa-miR-200b-3p, hsa-miR-200c-3p, hsa-miR-203a-3p, hsa-miR-20a-5p, hsa-miR-21-5p, hsa-miR-210-3p, hsa-miR-223-3p, hsa-miR-223-5p, hsa-miR-224-5p, hsa-miR-25-3p, hsa-miR-25-5p, hsa-miR-3065-5p, hsa-miR-30d-5p, hsa-miR-30e-5p, hsa-miR-31-5p, hsa-miR-32-5p, hsa-miR-320b, hsa-miR-320c, hsa-miR-330-3p, hsa-miR-330-5p, hsa-miR-342-5p, hsa-miR-378a-3p, hsa-miR-378a-5p, hsa-miR-378c, hsa-miR-378e, hsa-miR-378f, hsa-miR-378i, hsa-miR-421, hsa-miR-422a, hsa-miR-423-5p, hsa-miR-425-5p, hsa-miR-429, hsa-miR-449a, hsa-miR-449b-5p, hsa-miR-449c-5p, hsa-miR-454-3p, hsa-miR-4659a-3p, hsa-miR-4685-3p, hsa-miR-5001-3p, hsa-miR-501-3p, hsa-miR-548am-3p, hsa-miR-548ap-5p, hsa-miR-548av-5p, hsa-miR-548j-5p, hsa-miR-548k, hsa-miR-548o-3p, hsa-miR-561-5p, hsa-miR-574-5p, hsa-miR-625-5p, hsa-miR-629-5p, hsa-miR-641, hsa-miR-671-5p, hsa-miR-6842-3p, hsa-miR-7-5p, hsa-miR-744-5p, hsa-miR-760, hsa-miR-769-3p, hsa-miR-877-5p, hsa-miR-92a-3p, hsa-miR-93-5p, hsa-miR-941, hsa-miR-942-5p, hsa-miR-96-5p, hsa-miR-98-5p
Module 2	hsa-miR-1303
Module 3	hsa-let-7b-5p, hsa-miR-1-3p, hsa-miR-133a-3p, hsa-miR-133b
Module 4	hsa-let-7c-5p, hsa-let-7i-5p, hsa-miR-1-3p, hsa-miR-100-5p, hsa-miR-101-3p, hsa-miR-130a-3p, hsa-miR-140-3p, hsa-miR-145-5p, hsa-miR-181c-5p, hsa-miR-186-5p, hsa-miR-195-5p, hsa-miR-218-5p, hsa-miR-26b-5p, hsa-miR-326, hsa-miR-363-3p, hsa-miR-497-5p, hsa-miR-505-3p, hsa-miR-9-5p
Module 5	hsa-let-7c-5p, hsa-let-7d-5p, hsa-miR-1-3p, hsa-miR-100-5p, hsa-miR-101-3p, hsa-miR-106a-5p, hsa-miR-106b-5p, hsa-miR-107, hsa-miR-1294, hsa-miR-1295a, hsa-miR-1306-5p, hsa-miR-140-3p, hsa-miR-142-5p, hsa-miR-144-3p, hsa-miR-148a-5p, hsa-miR-151a-5p, hsa-miR-151b, hsa-miR-153-3p, hsa-miR-15a-5p, hsa-miR-15b-5p, hsa-miR-1827, hsa-miR-185-5p, hsa-miR-186-5p, hsa-miR-187-3p, hsa-miR-18b-5p, hsa-miR-190a-5p, hsa-miR-192-5p, hsa-miR-193a-5p, hsa-miR-194-5p, hsa-miR-196b-5p, hsa-miR-20b-5p, hsa-miR-215-5p, hsa-miR-25-3p, hsa-miR-29b-2-5p, hsa-miR-29c-3p, hsa-miR-30c-2-3p, hsa-miR-3143, hsa-miR-3158-3p, hsa-miR-3200-3p, hsa-miR-324-3p, hsa-miR-328-3p, hsa-miR-339-3p, hsa-miR-340-3p, hsa-miR-34b-3p, hsa-miR-34b-5p, hsa-miR-34c-5p, hsa-miR-363-3p, hsa-miR-3688-3p, hsa-miR-3909, hsa-miR-3913-5p, hsa-miR-4326, hsa-miR-4510, hsa-miR-4662a-5p, hsa-miR-484, hsa-miR-486-5p, hsa-miR-499a-5p, hsa-miR-500a-3p, hsa-miR-501-3p, hsa-miR-502-3p, hsa-miR-503-5p, hsa-miR-505-3p, hsa-miR-505-5p, hsa-miR-548ad-5p, hsa-miR-548ae-5p, hsa-miR-548ay-5p, hsa-miR-548b-5p, hsa-miR-548i, hsa-miR-576-5p, hsa-miR-584-5p, hsa-miR-589-5p, hsa-miR-6130, hsa-miR-625-5p, hsa-miR-628-5p, hsa-miR-642a-5p, hsa-miR-6513-3p, hsa-miR-660-5p, hsa-miR-664a-5p, hsa-miR-92b-3p, hsa-miR-942-5p
Module 6	hsa-miR-106b-5p, hsa-miR-107, hsa-miR-1271-5p, hsa-miR-1278, hsa-miR-1295a, hsa-miR-150-5p, hsa-miR-15b-5p, hsa-miR-190a-5p, hsa-miR-194-5p, hsa-miR-19b-3p, hsa-miR-328-3p, hsa-miR-329-3p, hsa-miR-335-5p, hsa-miR-34b-5p, hsa-miR-363-3p, hsa-miR-374a-5p, hsa-miR-374b-5p, hsa-miR-449c-5p, hsa-miR-497-5p, hsa-miR-502-3p, hsa-miR-503-5p, hsa-miR-539-3p, hsa-miR-653-5p, hsa-miR-766-3p, hsa-miR-9-5p
Module 7	hsa-let-7i-5p, hsa-miR-106b-5p, hsa-miR-142-5p, hsa-miR-146a-5p, hsa-miR-26b-5p, hsa-miR-324-3p, hsa-miR-324-5p, hsa-miR-3613-3p, hsa-miR-421, hsa-miR-454-3p, hsa-miR-503-5p, hsa-miR-652-3p

Module 8	<p>hsa-let-7c-3p, hsa-let-7e-5p, hsa-let-7f-2-3p, hsa-miR-10a-5p, hsa-miR-10b-5p, hsa-miR-125a-3p, hsa-miR-125a-5p, hsa-miR-125b-2-3p, hsa-miR-125b-5p, hsa-miR-1260a, hsa-miR-1260b, hsa-miR-1271-5p, hsa-miR-1275, hsa-miR-1277-5p, hsa-miR-1287-5p, hsa-miR-1296-5p, hsa-miR-132-5p, hsa-miR-135b-3p, hsa-miR-135b-5p, hsa-miR-136-5p, hsa-miR-138-5p, hsa-miR-141-3p, hsa-miR-141-5p, hsa-miR-143-3p, hsa-miR-149-5p, hsa-miR-152-3p, hsa-miR-154-5p, hsa-miR-181a-5p, hsa-miR-181b-5p, hsa-miR-181c-5p, hsa-miR-181d-5p, hsa-miR-188-5p, hsa-miR-193b-3p, hsa-miR-193b-5p, hsa-miR-195-5p, hsa-miR-199b-5p, hsa-miR-200a-3p, hsa-miR-200b-3p, hsa-miR-203a-3p, hsa-miR-204-5p, hsa-miR-205-5p, hsa-miR-212-5p, hsa-miR-214-3p, hsa-miR-214-5p, hsa-miR-218-5p, hsa-miR-221-3p, hsa-miR-221-5p, hsa-miR-222-3p, hsa-miR-224-3p, hsa-miR-23a-3p, hsa-miR-23b-3p, hsa-miR-23c, hsa-miR-24-3p, hsa-miR-26a-2-3p, hsa-miR-26a-5p, hsa-miR-27a-3p, hsa-miR-27b-3p, hsa-miR-27b-5p, hsa-miR-28-5p, hsa-miR-296-3p, hsa-miR-299-3p, hsa-miR-299-5p, hsa-miR-29a-3p, hsa-miR-29b-3p, hsa-miR-29c-3p, hsa-miR-30a-3p, hsa-miR-30a-5p, hsa-miR-30b-3p, hsa-miR-30b-5p, hsa-miR-30c-1-3p, hsa-miR-30c-2-3p, hsa-miR-30c-5p, hsa-miR-31-5p, hsa-miR-32-3p, hsa-miR-320d, hsa-miR-331-3p, hsa-miR-335-5p, hsa-miR-33a-5p, hsa-miR-340-5p, hsa-miR-342-3p, hsa-miR-345-5p, hsa-miR-34a-5p, hsa-miR-361-3p, hsa-miR-361-5p, hsa-miR-365a-3p, hsa-miR-365b-3p, hsa-miR-369-3p, hsa-miR-376c-3p, hsa-miR-377-3p, hsa-miR-378a-3p, hsa-miR-378a-5p, hsa-miR-378e, hsa-miR-378f, hsa-miR-378g, hsa-miR-382-3p, hsa-miR-3910, hsa-miR-422a, hsa-miR-429, hsa-miR-452-5p, hsa-miR-455-3p, hsa-miR-455-5p, hsa-miR-4662a-5p, hsa-miR-497-5p, hsa-miR-499a-5p, hsa-miR-504-5p, hsa-miR-511-3p, hsa-miR-511-5p, hsa-miR-582-5p, hsa-miR-628-5p, hsa-miR-655-3p, hsa-miR-656-3p, hsa-miR-664b-3p, hsa-miR-671-5p, hsa-miR-708-3p, hsa-miR-708-5p, hsa-miR-769-5p, hsa-miR-7977, hsa-miR-874-3p, hsa-miR-887-3p, hsa-miR-92a-1-5p, hsa-miR-944, hsa-miR-99a-3p, hsa-miR-99a-5p, hsa-miR-99b-3p, hsa-miR-99b-5p</p>
Module 9	hsa-let-7e-5p, hsa-miR-125a-5p, hsa-miR-138-5p, hsa-miR-149-5p

**Table B.1. Connections in the miRNA-Gene Module Network.**

	<b>t-statistics</b>	<b>t test p-value</b>	<b>t test FDR</b>	<b>Odds Ratio</b>	<b>Fisher-exact test p-value</b>	<b>Fisher-exact test FDR</b>
hsa-let-7e-5p	-3.142	0.006	0.033	6.052	0.001	0.025
hsa-miR-125a-5p	-3.604	0.003	0.022	4.902	0.004	0.081
hsa-miR-138-5p	-2.756	0.020	0.084	7.724	0.006	0.092
hsa-miR-149-5p	-4.937	0.001	0.007	14.645	0.000	0.002

**Table B.2.** Test statistics for miRNA connected to the immune-related modules in biopsy samples.

	<b>t-statistics</b>	<b>p-value</b>
hsa-let-7e-5p	-0.757	0.455
hsa-miR-125a-5p	0.153	0.880
hsa-miR-138-5p	-1.171	0.251
hsa-miR-149-5p	-2.177	<b>0.037</b>

**Table B.3. Differential Expression for miRNA connected to the immune-related module.**

	Discovery cohort t-statistics	Discovery cohort p-value	Validation cohort t-statistics	Validation cohort p-value	GSE114489 t-statistics	GSE114489 p-value
ADAR	0.755	0.456	1.449	0.163	0.024	0.981
BTN3A2	1.743	0.092	1.894	0.073	1.536	0.131
BTN3A3	2.308	<b>0.028</b>	1.210	0.240	2.039	<b>0.047</b>
HLA-E	2.048	<b>0.049</b>	1.635	0.118	0.325	0.746
NLRC5	2.236	<b>0.033</b>	1.239	0.230	1.397	0.169
SOCS1	2.297	<b>0.029</b>	1.388	0.180	0.789	0.434
ZNFX1	0.911	0.369	0.626	0.538	-0.225	0.823
USF1	0.714	0.481	0.204	0.840	0.301	0.765

**Table B.4. Differential Expression for miR-149-5p predicted target genes in Beane *et al.* Validation cohort and in Merrick *et al.***

	<b>Discovery cohort t-statistics</b>	<b>Discovery cohort p-value</b>	<b>Validation cohort t-statistics</b>	<b>Validation cohort p-value</b>	<b>GSE114489 t-statistics</b>	<b>GSE114489 p-value</b>
HLA-A	2.636	<b>0.013</b>	2.007	0.058	1.400	0.168
HLA-B	2.896	<b>0.007</b>	1.841	0.080	1.512	0.137
HLA-C	1.849	0.074	1.290	0.212	0.785	0.436
B2M	2.944	<b>0.006</b>	2.081	<b>0.050</b>	1.842	0.071
PSMB8	2.488	<b>0.019</b>	1.412	0.173	0.245	0.808
PSMB9	2.405	<b>0.023</b>	1.568	0.133	1.783	0.081
TAP1	2.645	<b>0.013</b>	1.296	0.210	0.601	0.551

**Table B.5. Differential expression for NLRC5 targets in Beane *et al.* Validation cohort and in Merrick *et al.***

<b>NSample=30</b>													
NTargets	Nvar. percent	VS. ANOVA F1		VS. ANOVA TPR		VS. ANOVA FPR		VS. Pair-wise F1		VS. Pair-wise TPR		VS. Pair-wise FPR	
		T	P-value	T	P-value	T	P-value	T	P-value	T	P-value	T	P-value
40	60%	0.17	0.87	0.00	1.00	-1.55	0.13	-0.12	0.91	-0.16	0.88	1.18	0.25
60	60%	0.12	0.90	0.00	1.00	-2.08	0.05	0.14	0.89	0.00	1.00	-1.99	0.06
80	60%	0.28	0.78	0.00	1.00	-2.69	0.01	1.32	0.19	1.04	0.30	-3.66	0.00
100	60%	0.38	0.71	0.53	0.60	-1.30	0.20	1.50	0.14	1.40	0.17	0.20	0.84
40	70%	0.88	0.38	0.78	0.44	-1.24	0.22	0.77	0.45	0.59	0.56	-1.69	0.10
60	70%	0.00	1.00	0.27	0.79	-1.83	0.08	1.29	0.20	1.08	0.28	-0.74	0.47
80	70%	0.36	0.72	0.42	0.68	-0.79	0.44	0.22	0.83	0.29	0.77	0.65	0.52
100	70%	0.31	0.76	0.19	0.85	-0.35	0.73	0.40	0.69	0.41	0.69	-0.21	0.84
40	80%	1.51	0.14	1.29	0.21	-1.93	0.06	2.31	0.03	2.08	0.04	-0.56	0.58
60	80%	1.81	0.08	1.48	0.15	-2.97	0.01	2.06	0.05	2.09	0.04	-0.87	0.39
80	80%	2.30	0.03	1.88	0.07	-1.75	0.09	3.05	0.00	3.08	0.00	0.49	0.63
100	80%	2.32	0.03	2.36	0.02	-0.88	0.38	1.87	0.07	2.12	0.04	1.23	0.23
40	90%	2.10	0.04	1.94	0.06	-0.56	0.58	3.05	0.00	3.01	0.00	-0.77	0.45
60	90%	2.88	0.01	2.27	0.03	-1.82	0.08	4.33	0.00	4.12	0.00	-1.00	0.32
80	90%	3.04	0.00	2.92	0.01	-1.84	0.07	3.73	0.00	4.03	0.00	0.23	0.82
100	90%	2.04	0.05	1.58	0.12	-4.74	0.00	5.14	0.00	5.35	0.00	-2.45	0.02
<b>NSample=50</b>													
NTargets	Nvar. percent	VS. ANOVA F1		VS. ANOVA TPR		VS. ANOVA FPR		VS. Pair-wise F1		VS. Pair-wise TPR		VS. Pair-wise FPR	
		T	P-value	T	P-value	T	P-value	T	P-value	T	P-value	T	P-value
40	60%	0.74	0.47	0.69	0.49	-1.80	0.08	0.17	0.86	0.17	0.87	-0.47	0.64
60	60%	0.82	0.42	0.75	0.46	-3.21	0.00	-0.24	0.81	-0.13	0.90	-0.59	0.56
80	60%	1.81	0.08	1.88	0.07	-0.85	0.40	0.48	0.63	0.38	0.71	-0.85	0.40
100	60%	1.34	0.19	1.26	0.22	-1.59	0.12	1.21	0.24	1.04	0.31	-1.46	0.15
40	70%	0.54	0.59	0.40	0.69	-1.05	0.30	0.01	0.99	0.00	1.00	-0.72	0.48
60	70%	3.71	0.00	3.60	0.00	-1.93	0.06	1.45	0.16	1.43	0.16	0.00	1.00
80	70%	3.18	0.00	3.13	0.00	-1.96	0.06	2.37	0.02	2.20	0.03	-1.40	0.17
100	70%	3.32	0.00	3.35	0.00	-1.45	0.16	2.85	0.01	2.75	0.01	-0.94	0.36
40	80%	2.68	0.01	2.73	0.01	-2.45	0.02	1.66	0.11	1.66	0.10	-1.18	0.25
60	80%	2.41	0.02	2.35	0.02	-1.95	0.06	2.44	0.02	2.33	0.03	0.00	1.00
80	80%	3.74	0.00	3.34	0.00	-1.38	0.18	4.53	0.00	4.33	0.00	-1.24	0.22
100	80%	4.96	0.00	4.51	0.00	-2.25	0.03	3.31	0.00	3.06	0.00	-1.67	0.10
40	90%	3.48	0.00	3.59	0.00	-1.13	0.27	3.69	0.00	4.00	0.00	-0.41	0.69
60	90%	3.71	0.00	3.65	0.00	-1.71	0.10	8.31	0.00	8.22	0.00	0.41	0.69
80	90%	5.39	0.00	5.06	0.00	-1.95	0.06	6.31	0.00	6.63	0.00	-1.33	0.19
100	90%	6.25	0.00	5.54	0.00	-1.74	0.09	8.06	0.00	8.93	0.00	-0.64	0.53
<b>NSample=70</b>													
NTargets	Nvar. percent	VS. ANOVA F1		VS. ANOVA TPR		VS. ANOVA FPR		VS. Pair-wise F1		VS. Pair-wise TPR		VS. Pair-wise FPR	
		T	P-value	T	P-value	T	P-value	T	P-value	T	P-value	T	P-value
40	60%	0.28	0.78	0.32	0.75	-1.71	0.10	-2.01	0.05	-1.76	0.09	-0.33	0.75
60	60%	1.46	0.15	1.30	0.20	-2.78	0.01	0.77	0.44	0.82	0.42	0.00	1.00
80	60%	3.26	0.00	3.14	0.00	-1.91	0.07	2.34	0.02	2.24	0.03	-1.24	0.22
100	60%	2.44	0.02	2.09	0.04	-2.68	0.01	1.89	0.07	1.59	0.12	-1.24	0.22

40	70%	1.98	0.06	1.87	0.07	-2.05	0.05	1.36	0.18	1.52	0.14	0.89	0.38
60	70%	3.38	0.00	3.64	0.00	-2.63	0.01	1.98	0.06	2.05	0.05	-2.99	0.01
80	70%	5.26	0.00	4.88	0.00	-2.85	0.01	3.48	0.00	3.48	0.00	-1.00	0.33
100	70%	6.57	0.00	7.06	0.00	-1.67	0.10	4.36	0.00	4.61	0.00	0.00	1.00
40	80%	3.94	0.00	3.99	0.00	-1.06	0.30	2.09	0.04	2.13	0.04	-0.87	0.39
60	80%	4.80	0.00	4.44	0.00	-2.31	0.03	4.83	0.00	4.69	0.00	-1.06	0.30
80	80%	7.49	0.00	7.74	0.00	-1.66	0.11	6.94	0.00	6.88	0.00	-0.52	0.60
100	80%	7.56	0.00	7.84	0.00	-2.35	0.03	6.43	0.00	6.65	0.00	-0.56	0.58
40	90%	4.41	0.00	4.59	0.00	-2.14	0.04	6.57	0.00	6.79	0.00	-1.04	0.31
60	90%	6.18	0.00	5.28	0.00	-1.54	0.14	6.84	0.00	7.50	0.00	-0.56	0.58
80	90%	8.21	0.00	9.56	0.00	-1.93	0.06	7.89	0.00	10.20	0.00	-1.06	0.30
100	90%	6.41	0.00	6.23	0.00	-1.47	0.15	7.91	0.00	10.90	0.00	0.86	0.40
<b>Different Nsample across groups</b>													
NTargets	Nvar. percent	VS. ANOVA F1		VS. ANOVA TPR		VS. ANOVA FPR		VS. Pair-wise F1		VS. Pair-wise TPR		VS. Pair-wise FPR	
		T	P-value	T	P-value	T	P-value	T	P-value	T	P-value	T	P-value
40	60%	2.23	0.03	2.22	0.03	-0.87	0.39	0.21	0.83	0.39	0.70	-1.24	0.22
60	60%	4.98	0.00	4.82	0.00	-0.75	0.46	2.18	0.04	1.99	0.05	-0.27	0.79
80	60%	2.66	0.01	2.44	0.02	-1.55	0.13	0.87	0.39	0.77	0.45	-1.40	0.17
100	60%	3.07	0.00	2.90	0.01	-1.49	0.15	2.43	0.02	2.28	0.03	-1.68	0.10
40	70%	3.65	0.00	3.40	0.00	-2.43	0.02	2.48	0.02	2.67	0.01	-0.61	0.55
60	70%	5.48	0.00	5.49	0.00	-1.37	0.18	2.57	0.01	2.83	0.01	0.89	0.38
80	70%	6.90	0.00	6.54	0.00	-2.12	0.04	3.16	0.00	3.22	0.00	-0.78	0.44
100	70%	5.06	0.00	5.06	0.00	-2.93	0.01	2.47	0.02	2.18	0.04	-1.72	0.10
40	80%	4.41	0.00	4.34	0.00	-1.23	0.23	2.63	0.01	2.56	0.01	0.41	0.69
60	80%	7.68	0.00	7.41	0.00	-2.48	0.02	4.57	0.00	4.47	0.00	-0.97	0.34
80	80%	7.26	0.00	7.76	0.00	-1.58	0.13	4.82	0.00	5.10	0.00	-0.27	0.79
100	80%	7.61	0.00	8.51	0.00	-1.04	0.31	8.30	0.00	8.08	0.00	-2.85	0.01
40	90%	7.40	0.00	7.21	0.00	-1.13	0.27	5.89	0.00	5.56	0.00	-2.21	0.03
60	90%	9.40	0.00	9.52	0.00	-1.53	0.14	11.17	0.00	12.80	0.00	-0.37	0.71
80	90%	6.31	0.00	6.47	0.00	-1.16	0.25	7.92	0.00	8.46	0.00	-0.65	0.52
100	90%	6.30	0.00	6.88	0.00	-1.07	0.29	7.37	0.00	9.15	0.00	-0.22	0.83

Table B.6. Comparison of model performance metrics in simulated datasets.

<b>Basal Subtype</b>			
<b>Label</b>	<b>p-value</b>	<b>FDR</b>	<b>hits</b>
KEGG OXIDATIVE PHOSPHORYLATION	0.0017	0.29	COX15,NDUFA3,NDUFC1
REACTOME SIGNALING BY KIT IN DISEASE	0.00058	0.7	PIK3R3,STAT5A
REACTOME_INACTIVATION_OF_CSF3_G_CSF_SIGNALING	0.00091	0.7	STAT5A,UBE2D1
<b>Luminal A Subtype</b>			
<b>Label</b>	<b>p-value</b>	<b>FDR</b>	<b>hits</b>
KEGG CITRATE CYCLE TCA CYCLE	0.00012	0.022	CS,PCK2,SDHD
KEGG JAK STAT SIGNALING PATHWAY	0.012	0.48	IL21R,JAK1,STAT1
REACTOME INTERLEUKIN 21 SIGNALING	3.30E-06	0.0052	IL21R,JAK1,STAT1
REACTOME INTERLEUKIN 9 SIGNALING	0.00033	0.13	JAK1,STAT1
REACTOME INTERLEUKIN 2 FAMILY SIGNALING	0.00034	0.13	IL21R,JAK1,STAT1
BIOCARTA FAS PATHWAY	0.00011	0.02	CASP7,FAS,PAK2

**Table B.7. Functional pathway enrichment results of genes in the basal or luminal A cluster.**

<b>Beane <i>et al.</i> Discovery</b>			
<b>HGNC Symbol</b>	<b>t</b>	<b>Log Fold-Change</b>	<b>p-value</b>
HLA-DQB1	1.7113435	1.1135827	0.09737
HLA-DRB1	3.2263701	1.1658441	<b>0.00303</b>
HLA-DRB5	0.9287384	0.5031092	0.36046
HLA-DRA	3.3775499	1.1424434	<b>0.00205</b>
HLA-DPA1	3.1243929	1.1164404	<b>0.00394</b>
HLA-DMB	2.8866788	1.008559	<b>0.00716</b>
CD74	3.0278447	0.9760697	<b>0.00503</b>
HLA-DMA	3.1985615	0.878681	<b>0.00326</b>
<b>Beane <i>et al.</i> Validation</b>			
<b>HGNC Symbol</b>	<b>t</b>	<b>Log Fold-Change</b>	<b>p-value</b>
HLA-DQB1	2.6047404	1.7482128	<b>0.01691</b>
HLA-DRB1	2.1361416	1.1705276	<b>0.04515</b>
HLA-DRB5	1.0516345	1.2410555	0.30544
HLA-DRA	1.6845378	0.7584402	0.10754
HLA-DPA1	1.6868599	0.7547033	0.10709
HLA-DMB	1.664055	0.6853659	0.11161
CD74	1.8993898	0.6822461	0.07196
HLA-DMA	1.9385173	0.6313718	0.06673
<b>Merrick <i>et al.</i></b>			
<b>HGNC Symbol</b>	<b>t</b>	<b>Log Fold-Change</b>	<b>p-value</b>
HLA-DQB1	1.7787442	0.5326371	0.08148
HLA-DRB1	1.3549279	0.4843854	0.18165
HLA-DRB5	1.6509376	0.4228367	0.10514
HLA-DRA	3.1550894	0.7706171	<b>0.00274</b>
HLA-DPA1	2.7434545	0.8531956	<b>0.00847</b>
HLA-DMB	2.7516354	0.4806581	<b>0.00829</b>
CD74	2.8076756	0.575176	<b>0.00714</b>

**Table B.8. Differential expression results for the MHC II genes by progression status in PML datasets**

Variables	Discovery (N=116)	Validation (N=124)	Statistics	p-value
<b>Batch</b>				
<b>1</b>	25 (21.55%)	13 (34.21%)		
<b>2</b>	28 (24.14%)	9 (23.68%)		
<b>3</b>	32 (27.59%)	10 (26.32%)		
<b>4</b>	12 (10.34%)	2 (5.26%)	3.37	0.64
<b>5</b>	15 (12.93%)	3 (7.89%)		
<b>6</b>	4 (3.45%)	1 (2.63%)		
<b>RIN</b>	5.42 (1.72)	5.53 (1.58)	-0.36	0.72
<b>Gender (=Male)</b>	86 (74.14%)	28 (73.68%)	-0.98	1
<b>Race</b>				
<b>African American</b>	18 (15.52%)	11 (28.95%)		
<b>White</b>	83 (71.55%)	24 (63.16%)	3.66	0.15
<b>Others</b>	15 (12.93%)	3 (7.89%)		
<b>Age (year)</b>	67.97 (8.65)	68.95 (7.91)	-0.65	0.52
<b>Nodule Size (cm)</b>	1.47 (0.59)	1.62 (0.7)	-1.17	0.25
<b>Pack-year</b>	50.16 (22.54)	55.56 (35.87)	-0.89	0.38
<b>Smoking Status (=Current)</b>	62 (53.45%)	55.56 (44.74%)	1.42	0.45
<b>Cancer Stats (=Benign)</b>	43 (37.07%)	14 (36.84%)	1.01	1

**Table B.9. Clinical annotation for DEAMP I nasal epithelial brushings with both miRNA and gene expression profiles between the discovery and validation cohort.**

Statistical tests for categorical clinical variables (cancer, gender, race and smoking status) were conducted using Chi-square tests. Statistical tests for continuous variables (RIN, age, nodule size and pack-year) were compared using two-sided Student's t-tests. Percentages are reported for categorical variables and mean/standard deviations are reported for the continuous variable.

## BIBLIOGRAPHY

1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer Statistics, 2021. *CA: A Cancer Journal for Clinicians* **71**, 7–33 (2021).
2. Travis, W., Brambilla, E., Burke, A., Marx, A., & Nicholson, A.G. Introduction to the 2015 World Health Organization classification of tumors of the lung, pleura, thymus, and heart. *Journal of Thoracic Oncology* **10**(9), 1240–1242 (2015)
3. Ginsberg, M. S., Grewal, R. K. & Heelan, R. T. Lung cancer. *Radiologic Clinics of North America* **45**, 21–43 (2007).
4. Kenfield, S. A., Wei, E. K., Stampfer, M. J., Rosner, B. A. & Colditz, G. A. Comparison of aspects of smoking among the four histological types of lung cancer. *Tobacco Control* **17**, 198–204 (2008).
5. Health Risk of Radon | US EPA. Available at: <https://www.epa.gov/radon/health-risk-radon>. (Accessed: 10th July 2022)
6. Rogers, A. & Major, G. Letters to the editor: The quantitative risks of mesothelioma and lung cancer in relation to asbestos exposure: The Wittenoom data. *Annals of Occupational Hygiene* **46**, 127–128 (2002).
7. AJ, van L. *et al.* Occupational exposure to carcinogens and risk of lung cancer: results from The Netherlands cohort study. *Occupational and Environmental Medicine* **54**, 817–824 (1997).
8. Pope, C. A. *et al.* Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to FineParticulate Air Pollution. *JAMA: The Journal of the American Medical Association* **287**, 1132–1141 (2002).
9. Lynch, T. J. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *The New England Journal of Medicine* **350**, 2129–2139 (2004).
10. Riely, G. J. *et al.* Frequency and Distinctive Spectrum of KRAS Mutations in Never Smokers with Lung Adenocarcinoma. *Clinical Cancer Research* **14**, 5731–5734 (2008).
11. Brose, M. S. *et al.* BRAF and RAS mutations in human lung cancer and melanoma. *Cancer Research* **62**, 6997–7000 (2002).
12. Soda, M. *et al.* Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. *Nature* **448**(7153), 561–566 (2007).

13. Van Meerbeeck, J. P., Fennell, D. A. & De Ruyscher, D. K. M. Small-cell lung cancer. *Lancet (London, England)* **378**, 1741–1755 (2011).
14. The National Lung Screening Trial Research Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *The New England Journal of Medicine* **365**, 395–409 (2011).
15. Jemal, A. *et al.* Annual Report to the Nation on the Status of Cancer, 1975–2014, Featuring Survival. *JNCI: Journal of the National Cancer Institute* **109**, (2017).
16. Ishizumi, T., McWilliams, A., MacAulay, C., Gazdar, A. & Lam, S. Natural history of bronchial preinvasive lesions. *Cancer and Metastasis Reviews* **29**, 5–14 (2010).
17. Greenberg, A. K., Yee, H. & Rom, W. N. Preneoplastic lesions of the lung. *Respiratory Research* **3**, (2002).
18. Kerr, K.M., Pulmonary preinvasive neoplasia. *Journal of Clinical Pathology* **54**, 257–271 (2001). <https://doi.org/10.1136/jcp.54.4.257>
19. Auerbach, O., Stout, A. P., Hammond, E. C., Garfinkel, L. & Oscar Auerbach, AP Stout, E. C. H. and L. G. Changes in Bronchial Epithelium in Relation to Cigarette Smoking and in Relation to Lung Cancer. *New England Journal of Medicine* **319**, 1374–1378 (1961).
20. Giroux, V. & Rustgi, A. K. Metaplasia: tissue injury adaptation and a precursor to the dysplasia-cancer sequence. *Nature Reviews. Cancer* **17**, 594–604 (2017).
21. Van Boerdonk, R. A. A. *et al.* Close surveillance with long-term follow-up of subjects with preinvasive endobronchial lesions. *American Journal of Respiratory and Critical Care Medicine* **192**, 1483–1489 (2015).
22. Wistuba, I. *et al.* Sequential molecular abnormalities are involved in the multistage development of squamous cell lung carcinoma. *Oncogene* **18**, 643–650 (1999).
23. Brambilla, E. *et al.* p53 mutant immunophenotype and deregulation of p53 transcription pathway (Bcl2, Bax, and Waf1) in precursor bronchial lesions of lung cancer. *Clinical Cancer Research* **4**(7), 1609–1618 (1998).
24. Brambilla, E., Moro, D., Gazzeri, S. & Brambilla, C. Alterations of expression of Rb, p16(INK4A) and cyclin D1 in non-small cell lung carcinoma and their clinical significance. *Journal of Pathology* **188**, 351–360 (1999).
25. Mascaux, C. *et al.* Evolution of microRNA expression during human bronchial squamous carcinogenesis. *The European Respiratory Journal* **33**, 352–359 (2009).

26. Arbour, K. C. & Riely, G. J. Systemic Therapy for Locally Advanced and Metastatic Non–Small Cell Lung Cancer: A Review. *JAMA: The Journal of the American Medical Association* **322**, 764–774 (2019).
27. Mao, J. T. *et al.* Lung cancer chemoprevention with celecoxib in former smokers. *Cancer Prevention Research* **4**, 984–993 (2011).
28. Yuan, J. M. *et al.* Clinical Trial of 2-Phenethyl Isothiocyanate as an Inhibitor of Metabolic Activation of a Tobacco-Specific Lung Carcinogen in Cigarette Smokers. *Cancer Prevention Research* **9**, 396–405 (2016).
29. Clinical Trial of Lung Cancer Chemoprevention With Sulforaphane in Former Smokers - Full Text View - ClinicalTrials.gov. Available at: <https://clinicaltrials.gov/ct2/show/NCT03232138>. (Accessed: 10th July 2022)
30. Electrocautery Ablation for the Prevention of Lung Cancer - Full Text View - ClinicalTrials.gov. Available at: <https://clinicaltrials.gov/ct2/show/NCT03870152>. (Accessed: 10th July 2022)
31. Breuer, R. H. *et al.* The natural course of preneoplastic lesions in bronchial epithelium. *Clinical Cancer Research* **11**, 537–543 (2005).
32. Merrick, D. T. *et al.* Persistence of bronchial dysplasia is associated with development of invasive squamous cell carcinoma. *Cancer Prevention Research* **9**, 96–104 (2016).
33. Merrick, D. T. *et al.* Altered cell-cycle control, inflammation, and adhesion in high-risk persistent bronchial dysplasia. *Cancer Research* **78**, 4971–4983 (2018).
34. Teixeira, V. H. *et al.* Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. *Nature Medicine* **25**, 517–525 (2019).
35. Pennycuik, A. *et al.* Immune surveillance in clinical regression of preinvasive squamous cell lung cancer. *Cancer Discovery* **10**, 1489–1499 (2020).
36. Denisov, E. V. *et al.* Gene Expression Profiling Revealed 2 Types of Bronchial Basal Cell Hyperplasia and Squamous Metaplasia With Different Progression Potentials. *Applied Immunohistochemistry & Molecular Morphology: AIMM* **28**, 477–483 (2020).
37. Chen, S. *et al.* Cancer-associated fibroblasts suppress SOX2-induced dysplasia in a lung squamous cancer coculture. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E11671–E11680 (2018).

38. Beane, J. *et al.* Detecting the presence and progression of premalignant lung lesions via airway gene expression. *Clinical Cancer Research* **23**, 5091–5100 (2017).
39. Beane, J. E. *et al.* Molecular subtyping reveals immune alterations associated with progression of bronchial premalignant lesions. *Nature Communications* **10**, (2019).
40. Ha, M. & Kim, V. N. Regulation of microRNA biogenesis. *Nature Reviews. Molecular Cell Biology* **15**, 509–524 (2014).
41. Lee, Y. *et al.* MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal* **23**, 4051–4060 (2004).
42. Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P. & Blelloch, R. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes & Development* **22**, 2773–2785 (2008).
43. Denli, A. M., Tops, B. B. J., Plasterk, R. H. A., Ketting, R. F. & Hannon, G. J. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**, 231–235 (2004).
44. Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415–419 (2003).
45. Han, J. *et al.* The Drosha-DGCR8 complex in primary microRNA processing. *Genes & Development* **18**, 3016–3027 (2004).
46. Lund, E., Güttinger, S., Calado, A., Dahlberg, J. E. & Kutay, U. Nuclear export of microRNA precursors. *Science* **303**, 95–98 (2004).
47. Park, J. E. *et al.* Dicer recognizes the 5' end of RNA for efficient and accurate processing. *Nature* **475**, 201–205 (2011).
48. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**(6818), 363–366 (2001).
49. Chendrimada, T. P. *et al.* TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* **436**, 740–744 (2005).
50. Bartel, D. P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **136**, 215–233 (2009).
51. Huntzinger, E. & Izaurralde, E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature Reviews. Genetics* **12**, 99–110 (2011).

52. Eichhorn, S. W. *et al.* mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. *Molecular Cell* **56**, 104–115 (2014).
53. Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008).
54. Uhlmann, S. *et al.* Global microRNA level regulation of EGFR-driven cell-cycle protein network in breast cancer. *Molecular Systems Biology* **8**, (2012).
55. Bracken, C. P. *et al.* Genome-wide identification of miR-200 targets reveals a regulatory network controlling cell invasion. *The EMBO Journal* **33**, 2040–2056 (2014).
56. Han, Y. C. *et al.* An allelic series of miR-17 ~ 92-mutant mice uncovers functional specialization and cooperation among members of a microRNA polycistron. *Nature Genetics* **47**, 766–775 (2015).
57. Tsang, J. S., Ebert, M. S. & van Oudenaarden, A. Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Molecular Cell* **38**, 140–153 (2010).
58. Chiang, H. R. *et al.* Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes & Development* **24**, 992–1009 (2010).
59. Vlachos, I. S. *et al.* DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. *Nucleic Acids Research* **40**, W498–W504 (2012).
60. Steinfeld, I., Navon, R., Ach, R. & Yakhini, Z. miRNA target enrichment analysis reveals directly active miRNAs in health and disease. *Nucleic Acids Research* **41**, (2013).
61. Fan, Y. *et al.* miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic Acids Research* **44**, W135–W141 (2016).
62. Backes, C., Khaleeq, Q. T., Meese, E. & Keller, A. miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Research* **44**, W110–W116 (2016).
63. Nam, S. *et al.* MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Research* **37**, W356–W362 (2009).

64. Luscombe, N. M. *et al.* Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308–312 (2004).
65. Bhardwaj, N., Kim, P. M. & Gerstein, M. B. Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. *Science Signaling* **3**, (2010).
66. Califano, A. Rewiring makes the difference. *Molecular Systems Biology* **7**, (2011).
67. Lionetti, M. *et al.* Identification of microRNA expression patterns and definition of a microRNA/mRNA regulatory network in distinct molecular groups of multiple myeloma. *Blood* **114**, e20–e26 (2009).
68. Volinia, S. *et al.* Reprogramming of miRNA networks in cancer and leukemia. *Genome Research* **20**, 589 (2010).
69. Nam, J. W. *et al.* Global analyses of the effect of different cellular contexts on microRNA targeting. *Molecular Cell* **53**, 1031–1043 (2014).
70. Lin, C. C. *et al.* Regulation rewiring analysis reveals mutual regulation between STAT1 and miR-155-5p in tumor immunosurveillance in seven major cancers. *Scientific Reports* **5**, 1–11 (2015).
71. Friedman, R. C., Farh, K. K. H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* **19**, 92–105 (2009).
72. Lin, S. & Gregory, R. I. MicroRNA biogenesis pathways in cancer. *Nature Reviews. Cancer* **15**, 321–333 (2015).
73. O'Donnell, K. A., Wentzel, E. A., Zeller, K. I., Dang, C. V. & Mendell, J. T. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature* **435**, 839–843 (2005).
74. Koralov, S. B. *et al.* Dicer ablation affects antibody diversity and cell survival in the B lymphocyte lineage. *Cell* **132**, 860–874 (2008).
75. Liu, C. *et al.* The microRNA miR-34a inhibits prostate cancer stem cells and metastasis by directly repressing CD44. *Nature Medicine* **17**, 211–216 (2011).
76. Rokavec, M. *et al.* IL-6R/STAT3/miR-34a feedback loop promotes EMT-mediated colorectal cancer invasion and metastasis. *The Journal of Clinical Investigation* **124**, 1853–1867 (2014).
77. Okada, N. *et al.* A positive feedback between p53 and miR-34 miRNAs mediates tumor suppression. *Genes & Development* **28**, 438–450 (2014).

78. Li, L. *et al.* MiR-34a inhibits proliferation and migration of breast cancer through down-regulation of Bcl-2 and SIRT1. *Clinical and Experimental Medicine* **13**, 109–117 (2013).
79. Wang, X. *et al.* Tumor suppressor miR-34a targets PD-L1 and functions as a potential immunotherapeutic target in acute myeloid leukemia. *Cellular Signalling* **27**, 443–452 (2015).
80. Cortez, M. A. *et al.* PDL1 Regulation by p53 via miR-34. *Journal of the National Cancer Institute* **108**, (2015).
81. Wegert, J. *et al.* Mutations in the SIX1/2 pathway and the DROSHA/DGCR8 miRNA microprocessor complex underlie high-risk blastemal type Wilms tumors. *Cancer Cell* **27**, 298–311 (2015).
82. Melo, S. A. *et al.* A Genetic Defect in Exportin-5 Traps Precursor MicroRNAs in the Nucleus of Cancer Cells. *Cancer Cell* **18**, 303–315 (2010).
83. Heravi-Moussavi, A. *et al.* Recurrent somatic DICER1 mutations in nonepithelial ovarian cancers. *The New England Journal of Medicine* **366**, 234–242 (2012).
84. Schembri, F. *et al.* MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 2319–2324 (2009).
85. Perdomo, C. *et al.* MicroRNA 4423 is a primate-specific regulator of airway epithelial cell differentiation and lung carcinogenesis. *Proceedings of the National Academy of Sciences* **110**, 18946–18951 (2013).
86. Justice, R. W., Zilian, O., Woods, D. F., Noll, M. & Bryant, P. J. The *Drosophila* tumor suppressor gene *warts* encodes a homolog of human myotonic dystrophy kinase and is required for the control of cell shape and proliferation. *Genes & Development* **9**, 534–546 (1995).
87. Dey, A., Varelas, X. & Guan, K.-L. Targeting the Hippo pathway in cancer, fibrosis, wound healing and regenerative medicine. *Nature Reviews. Drug Discovery* **19**(7), 480–494 (2020) doi:10.1038/s41573-020-0070-z
88. Dupont, S. *et al.* Role of YAP/TAZ in mechanotransduction. *Nature* **474**, 179–83 (2011).
89. Yu, F. X. *et al.* Regulation of the Hippo-YAP pathway by G-protein-coupled receptor signaling. *Cell* **150**, 780–791 (2012).

90. Zhou, X. *et al.* Estrogen regulates Hippo signaling via GPER in breast cancer. *The Journal of Clinical Investigation* **125**, 2123–2135 (2015).
91. Lehtinen, M. K. *et al.* A conserved MST-FOXO signaling pathway mediates oxidative-stress responses and extends life span. *Cell* **125**, 987–1001 (2006).
92. Geng, J. *et al.* Kinases Mst1 and Mst2 positively regulate phagocytic induction of reactive oxygen species and bactericidal activity. *Nature Immunology* **16**, 1142–1152 (2015).
93. Yang, C. C. *et al.* Differential regulation of the Hippo pathway by adherens junctions and apical-basal cell polarity modules. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 1785–1790 (2015).
94. Boggiano, J. C., Vanderzalm, P. J. & Fehon, R. G. Tao-1 phosphorylates Hippo/MST kinases to regulate the Hippo-Salvador-Warts tumor suppressor pathway. *Developmental Cell* **21**, 888–895 (2011).
95. Poon, C. L. C., Lin, J. I., Zhang, X. & Harvey, K. F. The sterile 20-like kinase Tao-1 controls tissue growth by regulating the Salvador-Warts-Hippo pathway. *Developmental Cell* **21**, 896–906 (2011).
96. Glantschnig, H., Rodan, G. A. & Reszka, A. A. Mapping of MST1 kinase sites of phosphorylation. Activation and autophosphorylation. *The Journal of Biological Chemistry* **277**, 42987–42996 (2002).
97. Hergovich, A., Schmitz, D. & Hemmings, B. A. The human tumour suppressor LATS1 is activated by human MOB1 at the membrane. *Biochemical and Biophysical Research Communications* **345**, 50–58 (2006).
98. Yin, F. *et al.* Spatial organization of Hippo signaling at the plasma membrane mediated by the tumor suppressor Merlin/NF2. *Cell* **154**, 1342 (2013).
99. Meng, Z. *et al.* MAP4K family kinases act in parallel to MST1/2 to activate LATS1/2 in the Hippo pathway. *Nature Communications* **6**, (2015).
100. Zhao, B. *et al.* Inactivation of YAP oncoprotein by the Hippo pathway is involved in cell contact inhibition and tissue growth control. *Genes & Development* **21**, 2747–2761 (2007).
101. Zhao, B. *et al.* TEAD mediates YAP-dependent gene induction and growth control. *Genes & Development* **22**, 1962–1971 (2008).
102. Huang, J., Wu, S., Barrera, J., Matthews, K. & Pan, D. The Hippo signaling pathway coordinately regulates cell proliferation and apoptosis by inactivating

- Yorkie, the Drosophila Homolog of YAP. *Cell* **122**, 421–434 (2005).
103. Camargo, F. D. *et al.* YAP1 increases organ size and expands undifferentiated progenitor cells. *Current Biology: CB* **17**, 2054–2060 (2007).
  104. Dong, J. *et al.* Elucidation of a universal size-control mechanism in Drosophila and mammals. *Cell* **130**, 1120–1133 (2007).
  105. Strano, S. *et al.* Physical interaction with Yes-associated protein enhances p73 transcriptional activity. *The Journal of Biological Chemistry* **276**, 15164–15173 (2001).
  106. Ferrigno, O. *et al.* Yes-associated protein (YAP65) interacts with Smad7 and potentiates its inhibitory activity against TGF-beta/Smad signaling. *Oncogene* **21**, 4879–4884 (2002).
  107. Komuro, A., Nagai, M., Navin, N. E. & Sudol, M. WW domain-containing protein YAP associates with ErbB-4 and acts as a co-transcriptional activator for the carboxyl-terminal fragment of ErbB-4 that translocates to the nucleus. *The Journal of Biological Chemistry* **278**, 33334–33341 (2003).
  108. Zagurovskaya, M. *et al.* EGR-1 forms a complex with YAP-1 and upregulates Bax expression in irradiated prostate carcinoma cells. *Oncogene* **28**, 1121–1131 (2009).
  109. Qiao, Y. *et al.* RUNX3 is a novel negative regulator of oncogenic TEAD-YAP complex in gastric cancer. *Oncogene* **35**, 2664–2674 (2016).
  110. Wang, Y. *et al.* Comprehensive Molecular Characterization of the Hippo Signaling Pathway in Cancer. *Cell Reports* **25**, 1304-1317.e5 (2018).
  111. Kapoor, A. *et al.* Yap1 activation enables bypass of oncogenic Kras addiction in pancreatic cancer. *Cell* **158**, 185–197 (2014).
  112. Zanconato, F. *et al.* Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. *Nature Cell Biology* **2015 17:9** **17**, 1218–1227 (2015).
  113. Bartucci, M. *et al.* TAZ is required for metastatic activity and chemoresistance of breast cancer stem cells. *Oncogene* **34**, 681–690 (2015).
  114. Basu-Roy, U. *et al.* Sox2 antagonizes the Hippo pathway to maintain stemness in cancer cells. *Nature Communications* **6**, (2015).
  115. Cordenonsi, M. *et al.* The Hippo transducer TAZ confers cancer stem cell-related traits on breast cancer cells. *Cell* **147**, 759–772 (2011).

116. Rosenbluh, J. *et al.*  $\beta$ -Catenin-driven cancers require a YAP1 transcriptional complex for survival and tumorigenesis. *Cell* **151**, 1457–1473 (2012).
117. Nallet-Staub, F. *et al.* Pro-invasive activity of the Hippo pathway effectors YAP and TAZ in cutaneous melanoma. *The Journal of Investigative Dermatology* **134**, 123–132 (2014).
118. Hiemer, S. E. *et al.* A YAP/TAZ-Regulated Molecular Signature Is Associated with Oral Squamous Cell Carcinoma. *Molecular Cancer Research* **13**, 957–968 (2015).
119. Cheng, H. *et al.* Functional genomics screen identifies YAP1 as a key determinant to enhance treatment sensitivity in lung cancer cells. *Oncotarget* **7**, 28976–28988 (2016).
120. Lau, A. N. *et al.* Tumor-propagating cells and Yap/Taz activity contribute to lung tumor progression and metastasis. *The EMBO Journal* **33**, 468–481 (2014).
121. Noguchi, S. *et al.* An integrative analysis of the tumorigenic role of TAZ in human non-small cell lung cancer. *Clinical Cancer Research* **20**, 4660–4672 (2014).
122. Lee, B. S. *et al.* Hippo effector YAP directly regulates the expression of PD-L1 transcripts in EGFR-TKI-resistant lung adenocarcinoma. *Biochemical and Biophysical Research Communications* **491**, 493–499 (2017).
123. Szymaniak, A. D., Mahoney, J. E., Cardoso, W. V. & Varelas, X. Crumbs3-Mediated Polarity Directs Airway Epithelial Cell Fate through the Hippo Pathway Effector Yap. *Developmental Cell* **34**, 283–296 (2015).
124. Tilston-Lunel, A. *et al.* Aberrant epithelial polarity cues drive the development of precancerous airway lesions. *Proceedings of the National Academy of Sciences of the United States of America* **118**, (2021).
125. Liu-Chittenden, Y. *et al.* Genetic and pharmacological disruption of the TEAD-YAP complex suppresses the oncogenic activity of YAP. *Genes & Development* **26**, 1300–1305 (2012).
126. Dasari, V. R. *et al.* Verteporfin exhibits YAP-independent anti-proliferative and cytotoxic effects in endometrial cancer cells. *Oncotarget* **8**, 28628–28640 (2017).
127. Zhang, H. *et al.* Tumor-selective proteotoxicity of verteporfin inhibits colon cancer progression independently of YAP1. *Science Signaling* **8**, (2015).
128. Kurppa, K. J. *et al.* Treatment-Induced Tumor Dormancy through YAP-Mediated Transcriptional Reprogramming of the Apoptotic Pathway. *Cancer Cell* **37**, 104–

- 122.e12 (2020).
129. Basu, D. *et al.* Identification, mechanism of action, and antitumor activity of a small molecule inhibitor of hippo, TGF- $\beta$ , and Wnt signaling pathways. *Molecular Cancer Therapeutics* **13**, 1457–1467 (2014).
  130. Noland, C. L. *et al.* Palmitoylation of TEAD Transcription Factors Is Required for Their Stability and Function in Hippo Pathway Signaling. *Structure* **24**, 179–186 (2016).
  131. Chan, P. *et al.* Autopalmitoylation of TEAD proteins regulates transcriptional output of the Hippo pathway. *Nature Chemical Biology* **12**, 282–289 (2016).
  132. Li, Q. *et al.* Lats1/2 Sustain Intestinal Stem Cells and Wnt Activation through TEAD-Dependent and Independent Transcription. *Cell Stem Cell* **26**, 675-692.e8 (2020).
  133. Slaughter, D., Southwick, H., & Smejkal, W. Field cancerization in oral stratified squamous epithelium. Clinical implications of multicentric origin. *Cancer* **6**(5), 963–968 (1953).
  134. Heaphy, C. M. *et al.* Telomere DNA content and allelic imbalance demonstrate field cancerization in histologically normal tissue adjacent to breast tumors. *International Journal of Cancer* **119**(1), 108–116 (2006).
  135. Berman, H., Zhang, J., et al. Genetic and epigenetic changes in mammary epithelial cells identify a subpopulation of cells involved in early carcinogenesis. *Cold Spring Harbor Symposia on Quantitative Biology* **70**, 317–327 (2005).
  136. Lewis, C., Cler, L.R, et al. Promoter hypermethylation in benign breast epithelium in relation to predicted breast cancer risk. *Clinical Cancer Research* **11**(1), 166–172 (2005).
  137. Ogden, G. R. Potential early markers of carcinogenesis in the mucosa of the head and neck using exfoliative cytology. *The Journal of Pathology* **181**, 347 (1997).
  138. Shin, D.M., Ro, J.Y., Hong, W.K., and Hittelman, W.N. Dysregulation of epidermal growth factor receptor expression in premalignant lesions during head and neck tumorigenesis. *Cancer Research* **54**, 3153–3159 (1994).
  139. Parr, R., Dakubo, G., Crandall, K.A. et al. Somatic mitochondrial DNA mutations in prostate cancer and normal appearing adjacent glands in comparison to age-matched prostate samples without malignant. *Journal of Molecular Diagnostics* **8**(3), 312–319 (2006).

140. Jones, A. C. *et al.* Early growth response 1 and fatty acid synthase expression is altered in tumor adjacent prostate tissue and indicates field cancerization. *Prostate* **72**, 1159–1170 (2012).
141. Haaland, C. M. *et al.* Differential gene expression in tumor adjacent histologically normal prostatic tissue indicates field cancerization. *International Journal of Oncology* **35**, 537–546 (2009).
142. Lee, Y. *et al.* Revisit of Field Cancerization in Squamous Cell Carcinoma of Upper Aerodigestive Tract: Better Risk Assessment with Epigenetic Markers. *Cancer Prevention Research* **4**(12), 1982–1992 (2011)
143. Damania, D., Roy, H., Subramanian, H., et al. Nanocytology of rectal colonocytes to assess risk of colon cancer based on field cancerization. *Cancer Research* **72**(11), 2720–2727 (2012).
144. Billatos, E., Vick, J. L., Lenburg, M. E. & Spira, A. E. The airway transcriptome as a biomarker for early lung cancer detection. *Clinical Cancer Research* **24**, 2984–2992 (2018).
145. Hackett, N. R. *et al.* Variability of antioxidant-related gene expression in the airway epithelium of cigarette smokers. *American Journal of Respiratory Cell and Molecular Biology* **29**, 331–343 (2003).
146. Spira, A. *et al.* Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 10143–10148 (2004).
147. Nelson, H. H. *et al.* Implications and prognostic value of K-ras mutation for early-stage lung cancer in women. *Journal of the National Cancer Institute* **91**, 2032–2038 (1999).
148. Franklin, W. A. *et al.* Widely dispersed p53 mutation in respiratory epithelium. A novel mechanism for field carcinogenesis. *The Journal of Clinical Investigation* **100**, 2133–2137 (1997).
149. Tang, X. *et al.* EGFR tyrosine kinase domain mutations are detected in histologically normal respiratory epithelium in lung cancer patients. *Cancer Research* **65**, 7568–7572 (2005).
150. Kadara, H. *et al.* Transcriptomic architecture of the adjacent airway field cancerization in non-small cell lung cancer. *Journal of the National Cancer Institute* **106**, (2014).

151. Gould, M. K. *et al.* Recent Trends in the Identification of Incidental Pulmonary Nodules. *American Journal of Respiratory and Critical Care Medicine* **192**, 1208–1214 (2015).
152. Gesthalter, Y., Koppelman, E., Bolton, R., *et al.* Evaluations of implementation at early-adopting lung cancer screening programs: lessons learned. *Chest* **152**(1), 70–80 (2017)
153. Blomquist, T. *et al.* Pattern of antioxidant and DNA repair gene expression in normal airway epithelium associated with lung cancer diagnosis. *Cancer Research* **69**, 8629–8635 (2009).
154. Crawford, E. L. *et al.* Lung cancer risk test trial: Study design, participant baseline characteristics, bronchoscopy safety, and establishment of a biospecimen repository. *BMC Pulmonary Medicine* **16**, 1–12 (2016).
155. Spira, A. *et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine* **13**, 361–6 (2007).
156. Whitney, D. H. *et al.* Derivation of a bronchial genomic classifier for lung cancer in a prospective study of patients undergoing diagnostic bronchoscopy. *BMC Medical Genomics* **8**, 1–10 (2015).
157. Silvestri, G. A. *et al.* A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. *New England Journal of Medicine* **373**, 243–251 (2015).
158. Ost, D. E. *et al.* Diagnostic Yield and Complications of Bronchoscopy for Peripheral Lung Lesions. Results of the AQUIRE Registry. *American Journal of Respiratory and Critical Care Medicine* **193**, 68–77 (2016).
159. Zhang, X. *et al.* Similarities and differences between smoking-related gene expression in nasal and bronchial epithelium. *Physiological Genomics* **41**, 1–8 (2010).
160. Boudewijn, I. M. *et al.* Nasal gene expression differentiates COPD from controls and overlaps bronchial gene expression. *Respiratory Research* **18**, 213 (2017).
161. Sala, M. A. *et al.* Inflammatory pathways are upregulated in the nasal epithelium in patients with idiopathic pulmonary fibrosis. *Respiratory Research* **19**, 1–10 (2018).
162. Perez-rogers, J. F. *et al.* Shared Gene Expression Alterations in Nasal and Bronchial Epithelium for Lung Cancer Detection. *Journal of the National Cancer Institute* **109**, 1–9 (2017).

163. Mazzone, P. J. *et al.* Early candidate nasal swab classifiers developed using machine learning and whole transcriptome sequencing may improve early lung cancer detection. *Journal of Clinical Oncology* **39**, 8551–8551 (2021).
164. Lamb, C. R. *et al.* A Nasal Genomic Classifier for Assessing Risk of Malignancy in Lung Nodules Demonstrates Similar Performance in Patients That Meet Screening Criteria for High Baseline Risk and Those Who Do Not. *American Thoracic Society International Conference Meetings Abstracts* A5585–A5585 (2022). doi:10.1164/AJRCCM-CONFERENCE.2022.205.1\_MEETINGABSTRACTS.A5585
165. Beane, J. *et al.* A prediction model for lung cancer diagnosis that integrates genomic and clinical features. *Cancer Prevention Research* **1**, 56–64 (2008).
166. Pavel, A. B. *et al.* Alterations in bronchial airway miRNA expression for lung cancer detection. *Cancer Prevention Research* **10**, 651–659 (2017).
167. Su, Y., Fang, H. Bin & Jiang, F. Integrating DNA methylation and microRNA biomarkers in sputum for lung cancer detection. *Clinical Epigenetics* **8**, (2016).
168. Pinsky, P. F. & Berg, C. D. Applying the National Lung Screening Trial eligibility criteria to the US population: What percent of the population and of incident lung cancers would be covered? *Journal of Medical Screening* **19**, 154–156 (2012).
169. Billatos, E. *et al.* Detection of early lung cancer among military personnel (DECAMP) consortium: Study protocols. *BMC Pulmonary Medicine* **19**, 1–9 (2019).
170. National Lung Screening Trial Research Team, T. Lung Cancer Incidence and Mortality with Extended Follow-up in the National Lung Screening Trial. *Journal of Thoracic Oncology* **14**, 1732–1742 (2019).
171. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians* **70**, 7–30 (2020).
172. Oscar Auerbach, AP Stout, E. C. H. and L. G. Changes in Bronchial Epithelium in Relation to Cigarette Smoking and in Relation to Lung Cancer. *New England Journal of Medicine* **319**, 1374–1378 (1961).
173. Moro-Sibilot, D. *et al.* Clinical prognostic indicators of high-grade pre-invasive bronchial lesions. *European Respiratory Journal* **24**, 24–29 (2004).
174. Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).

175. Campbell, J. D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics* **48**, 607–616 (2016).
176. Campbell, J. D. *et al.* The Case for a Pre-Cancer Genome Atlas (PCGA). *Cancer Prevention Research* **9**, 119–124 (2016).
177. Lel, J., Spira, A., Beane, J., Campbell, J. D. & Vick, J. Genomic approaches to accelerate cancer interception. *The Lancet. Oncology* **18**, e494–e502 (2017).
178. Mascaux, C. *et al.* Immune evasion before tumour invasion in early lung squamous carcinogenesis. *Nature* **571**, 570–575 (2019).
179. Jacobsen, A. *et al.* Analysis of microRNA-target interactions across diverse cancer types. *Nature Structural and Molecular Biology* **20**, 1325–1332 (2013).
180. Dhawan, A., Scott, J. G., Harris, A. L. & Buffa, F. M. Pan-cancer characterisation of microRNA across cancer hallmarks reveals microRNA-mediated downregulation of tumour suppressors. *Nature Communications* **9**, 1–13 (2018).
181. Dvinge, H. *et al.* The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature* **497**, 378–382 (2013).
182. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
183. Yang, D. *et al.* Article Integrated Analyses Identify a Master MicroRNA Regulatory Network for the Mesenchymal Subtype in Serous Ovarian Cancer. *Cancer Cell* **23**, 186–199 (2013).
184. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
185. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
186. Campbell, J. D. *et al.* Genomic, Pathway Network, and Immunologic Features Distinguishing Squamous Carcinomas. *Cell Reports* **23**, 194–212.e6 (2018).
187. Yang, Y. *et al.* Progress risk assessment of oral premalignant lesions with saliva miRNA analysis. *BMC Cancer* **13**, 129 (2013).
188. Yerukala Sathipati, S. & Ho, S. Y. Identifying a miRNA signature for predicting the stage of breast cancer. *Scientific Reports* **8**, 1–11 (2018).

189. Shams, R. *et al.* Identification of potential microRNA panels for pancreatic cancer diagnosis using microarray datasets and bioinformatics methods. *Scientific Reports* **10**, 1–15 (2020).
190. Campbell, J. D. *et al.* Assessment of microRNA differential expression and detection in multiplexed small RNA sequencing data. *RNA* **21**, 164–171 (2015).
191. Andrews, S. & others. FastQC: a quality control tool for high throughput sequence data. 2010. <http://www.bioinformatics.babraham.ac.uk/projects/> (2010).
192. Mrxuqdo, Q. H. W., Iurp, V., Wkurxjksxw, K. & Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* **17**, 5–7 (2011).
193. Mackowiak, S. D., Li, N., Chen, W., Friedla, M. R. & Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research* **40**, 37–52 (2012).
194. Kozomara, A. & Griffiths-Jones, S. miRBase : annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research* **42**, 68–73 (2014).
195. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**, 1–17 (2014).
196. Wang, L., Wang, S. & Li, W. RSeQC: Quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
197. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**(7), e47 (2015).
198. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44**, e71–e71 (2016).
199. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
200. Li, J., Liu, S., Zhou, H., Qu, L. & Yang, J. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research* **42**, 92–97 (2014).
201. Chou, C. H. *et al.* MiRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research* **46**, D296–D302 (2018).

202. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, (2013).
203. De Rie, D. *et al.* An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nature Biotechnology* **35**, 872–878 (2017).
204. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–50 (2005).
205. Ludigs, K. *et al.* NLRC5 Exclusively Transactivates MHC Class I and Related Genes through a Distinctive SXY Module. *PLoS Genetics* **11**, e1005088 (2015).
206. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
207. Aliche, H. & Theis, F. J. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. *Cell Systems* **12**, 706-715.e4 (2021).
208. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD, Zhang Y, T. L. sva: Surrogate Variable Analysis. R package version 3.32.1. (2019). Available at: <https://bioconductor.org/packages/release/bioc/html/sva.html>. (Accessed: 5th June 2019)
209. Xu, K. *et al.* Smoking Modulates Different Secretory Subpopulations Expressing SARS-CoV-2 Entry Genes in the Nasal and Bronchial Airways. *Research Square* (2021). doi:10.21203/RS.3.RS-887718/V1
210. Bracken, C. P., Scott, H. S. & Goodall, G. J. A network-biology perspective of microRNA function and dysfunction in cancer. *Nature Reviews. Genetics* **17**, 719–732 (2016).
211. Dragomir, M., Mafra, A. C. P., Dias, S. M. G., Vasilescu, C. & Calin, G. A. Using microRNA networks to understand cancer. *International Journal of Molecular Sciences* **19**, (2018).
212. Corney, D. C., Flesken-Nikitin, A., Godwin, A. K., Wang, W. & Nikitin, A. Y. MicroRNA-34b and MicroRNA-34c are targets of p53 and cooperate in control of cell proliferation and adhesion-independent growth. *Cancer Research* **67**, 8433–8438 (2007).
213. Song, R. *et al.* MiR-34/449 miRNAs are required for motile ciliogenesis by repressing cp110. *Nature* **510**, 115–120 (2014).

214. Guan, T. *et al.* ZEB1, ZEB2, and the miR-200 family form a counterregulatory network to regulate CD8<sup>+</sup> T cell fates. *Journal of Experimental Medicine* **215**, 1153–1168 (2018).
215. He, Y. *et al.* miR-149 in human cancer: A systemic review. *Journal of Cancer* **9**, 375–388 (2018).
216. Ow, S. H., Chua, P. J. & Bay, B. H. miR-149 as a Potential Molecular Target for Cancer. *Current Medicinal Chemistry* **25**, 1046–1054 (2017).
217. Meissner, T. B. *et al.* NLR family member NLRC5 is a transcriptional regulator of MHC class I genes. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 13794–13799 (2010).
218. Kobayashi, K. S. & van den Elsen, P. J. NLRC5: a key regulator of MHC class I-dependent immune responses. *Nature Reviews. Immunology* **12**, 813–820 (2012).
219. Yoshihama, S., Vijayan, S., Sidiq, T. & Kobayashi, K. S. NLRC5/CITA: A Key Player in Cancer Immune Surveillance. *Trends in Cancer* **3**, 28–38 (2017).
220. He, Y., Jiang, X. & Chen, J. The role of miR-150 in normal and malignant hematopoiesis. *Oncogene* **33**, 3887–3893 (2014).
221. Yang, C. *et al.* Integrative analysis of microRNA and mRNA expression profiles in non-small-cell lung cancer. *Cancer Gene Therapy* **23**, 90–97 (2016).
222. Ke, Y., Zhao, W., Xiong, J. & Cao, R. miR-149 Inhibits Non-Small-Cell Lung Cancer Cells EMT by Targeting FOXM1. *Biochemistry Research International* **2013**, (2013).
223. Xiang, F. *et al.* Ursolic Acid Reverses the Chemoresistance of Breast Cancer Cells to Paclitaxel by Targeting MiRNA-149-5p/MyD88. *Frontiers in Oncology* **9**, (2019).
224. Tian, D. *et al.* Anesthetic propofol epigenetically regulates breast cancer trastuzumab resistance through IL-6/miR-149-5p axis. *Scientific Reports* **10**, (2020).
225. Zhang, X. *et al.* Circular RNA circNRIP1 acts as a microRNA-149-5p sponge to promote gastric cancer progression via the AKT1/mTOR pathway. *Molecular Cancer* **18**, (2019).
226. Lv, T., Liu, H., Wu, Y. & Huang, W. Knockdown of lncRNA DLEU1 inhibits the tumorigenesis of oral squamous cell carcinoma via regulation of miR-149-5p/CDK6 axis. *Molecular Medicine Reports* **23**, (2021).

227. Shimada, K. *et al.* Syndecan-1, a new target molecule involved in progression of androgen-independent prostate cancer. *Cancer Science* **100**, 1248–1254 (2009).
228. Pfeffer, S. R. *et al.* Detection of Exosomal miRNAs in the Plasma of Melanoma Patients. *Journal of Clinical Medicine* **4**, 2012–2027 (2015).
229. Srivastava, A. *et al.* Cross-talk between IFN- $\gamma$  and TWEAK through miR-149 amplifies skin inflammation in psoriasis. *Journal of Allergy and Clinical Immunology* **147**, 2225–2235 (2021).
230. Jorge, N. A. N. *et al.* Poor clinical outcome in metastatic melanoma is associated with a microRNA-modulated immunosuppressive tumor microenvironment. *Journal of Translational Medicine* **18**, 56 (2020).
231. Dhatchinamoorthy, K., Colbert, J. D. & Rock, K. L. Cancer Immune Evasion Through Loss of MHC Class I Antigen Presentation. *Frontiers in Immunology* **12**, 469 (2021).
232. Staehli, F. *et al.* NLRC5 deficiency selectively impairs MHC class I- dependent lymphocyte killing by cytotoxic T cells. *Journal of Immunology* **188**, 3820–3828 (2012).
233. Biswas, A., Meissner, T. B., Kawai, T. & Kobayashi, K. S. Cutting edge: impaired MHC class I expression in mice deficient for Nlrc5/class I transactivator. *Journal of Immunology* **189**, 516–520 (2012).
234. Yoshihama, S. *et al.* NLRC5/MHC class I transactivator is a target for immune evasion in cancer. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 5999–6004 (2016).
235. Gettinger, S. *et al.* Impaired HLA class I antigen processing and presentation as a mechanism of acquired resistance to immune checkpoint inhibitors in lung cancer. *Cancer Discovery* **7**, 1420–1435 (2017).
236. Gu, S. S. *et al.* Therapeutically Increasing MHC-I Expression Potentiates Immune Checkpoint Blockade. *Cancer Discovery* **11**, 1524–1541 (2021).
237. Ayukawa, S. *et al.* Epithelial cells remove precancerous cells by cell competition via MHC class I–LILRB3 interaction. *Nature Immunology* **22**(11), 1391–1402 (2021).
238. Rock, J. R. & Hogan, B. L. M. Epithelial progenitor cells in lung development, maintenance, repair, and disease. *Annual Review of Cell and Developmental Biology* **27**, 493–512 (2011).

239. Hynds, R. E. & Janes, S. M. Airway Basal Cell Heterogeneity and Lung Squamous Cell Carcinoma. *Cancer Prevention Research* **10**, 491–493 (2017).
240. Laughney, A. M. *et al.* Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nature Medicine* **26**, 259–269 (2020).
241. Tallerico, R. *et al.* Human NK cells selective targeting of colon cancer-initiating cells: a role for natural cytotoxicity receptors and MHC class I molecules. *Journal of Immunology* **190**, 2381–2390 (2013).
242. Volonté, A. *et al.* Cancer-initiating cells from colorectal cancer patients escape from T cell-mediated immunosurveillance in vitro through membrane-bound IL-4. *Journal of Immunology* **192**, 523–532 (2014).
243. Li, Z. & Rana, T. M. Therapeutic targeting of microRNAs: current status and future challenges. *Nature Reviews. Drug Discovery* **13**, 622–638 (2014).
244. Rupaimoole, R. & Slack, F. J. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nature Reviews. Drug Discovery* **16**(3), 203–222 (2017).
245. Pramanik, D. *et al.* Restitution of tumor suppressor microRNAs using a systemic nanovector inhibits pancreatic cancer growth in mice. *Molecular Cancer Therapeutics* **10**, 1470–1480 (2011).
246. Trang, P. *et al.* Systemic delivery of tumor suppressor microRNA mimics using a neutral lipid emulsion inhibits lung tumors in mice. *Molecular Therapy* **19**, 1116–1122 (2011).
247. Kasinski, A. L. & Slack, F. J. miRNA-34 prevents cancer initiation and progression in a therapeutically resistant K-ras and p53-induced mouse model of lung adenocarcinoma. *Cancer Research* **72**, 5576–5587 (2012).
248. Ma, L. *et al.* Therapeutic silencing of miR-10b inhibits metastasis in a mouse mammary tumor model. *Nature Biotechnology* **28**, 341–347 (2010).
249. Gabriely, G. *et al.* Human glioma growth is controlled by microRNA-10b. *Cancer Research* **71**, 3563–3572 (2011).
250. Hou, X., Zaks, T., Langer, R. & Dong, Y. Lipid nanoparticles for mRNA delivery. *Nature Reviews. Materials* **6**(12), 1078–1094 (2021).
251. DeVeale, B., Swindlehurst-Chan, J. & Blelloch, R. The roles of microRNAs in mouse development. *Nature Reviews. Genetics* **22**(5), 307–323 (2021).

252. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**(7414), 91–100 (2012).
253. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
254. Aibar, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods* **14**, 1083–1086 (2017).
255. Ding, K. F. *et al.* Network rewiring in cancer: Applications to melanoma cell lines and the cancer genome atlas patients. *Frontiers in Genetics* **9**, 228 (2018).
256. Assi, S. A. *et al.* Subtype-specific regulatory network rewiring in acute myeloid leukemia. *Nature Genetics* **51**, 151–162 (2019).
257. McKenzie, A. T., Katsyv, I., Song, W. M., Wang, M. & Zhang, B. DGCA : A comprehensive R package for Differential Gene Correlation Analysis. *BMC Systems Biology* **10**, 1–25 (2016).
258. Zhang, J. *et al.* DiNeR: a Differential graphical model for analysis of co-regulation Network Rewiring. *BMC Bioinformatics* **21**, 281 (2020).
259. Lou, S. *et al.* TopicNet: a framework for measuring transcriptional regulatory network change. *Bioinformatics* **36**, i474–i481 (2020).
260. Tesson, B. M., Breitling, R. & Jansen, R. C. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics* **11**, 497 (2010).
261. Jonas, S. & Izaurralde, E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nature Reviews. Genetics* **16**, 421–433 (2015).
262. Bartel, D. P. Metazoan MicroRNAs. *Cell* **173**, 20–51 (2018).
263. Agarwal, V., Bell, G. W., Nam, J. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).  
doi:10.7554/eLife.05005
264. Gebert, L. F. R. & MacRae, I. J. Regulation of microRNA function in animals. *Nature Reviews. Molecular Cell Biology* **20**(1), 21–37 (2018).
265. Esquela-Kerscher, A. & Slack, F. J. Oncomirs — microRNAs with a role in cancer. *Nature Reviews. Cancer* **6**(4), 259–269 (2006).
266. Nowakowski, T. J. *et al.* Regulation of cell-type-specific transcriptomes by microRNA networks during human brain development. *Nature Neuroscience* **21**,

- 1784–1792 (2018).
267. Hsin, J. P., Lu, Y., Loeb, G. B., Leslie, C. S. & Rudensky, A. Y. The effect of cellular context on miR-155-mediated gene regulation in four major immune cell types. *Nature Immunology* **19**, 1137–1145 (2018).
268. Li, X. *et al.* High-Resolution In Vivo Identification of miRNA Targets by Halo-Enhanced Ago2 Pull-Down. *Molecular Cell* **79**, 167-179.e11 (2020).
269. Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. *Cell* **153**, 654 (2013).
270. Wang, Y., Hicks, S. C. & Hansen, K. D. Co-expression analysis is biased by a mean-correlation relationship. *bioRxiv* 2020.02.13.944777 (2020). doi:10.1101/2020.02.13.944777
271. Tang, J. *et al.* LINE: Large-scale Information Network Embedding. *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web* 1067–1077 (2015). doi:10.1145/2736277.2741093
272. Ru, Y. *et al.* The multiMiR R package and database: integration of microRNA–target interactions along with their disease and drug associations. *Nucleic Acids Research* **42**, e133–e133 (2014).
273. Geraci, M. Linear Quantile Mixed Models: The lqmm Package for Laplace Quantile Regression. *Journal of Statistical Software* **57**, 1–29 (2014).
274. Wenables, W. & Ripley, B. *Modern Applied Statistics with S, Fourth edition.* (Springer, New York, 2002).
275. Edgington, E. S. An additive method for combining probability values from independent experiments. *Journal of Psychology: Interdisciplinary and Applied* **80**, 351–363 (1972).
276. Dewey, M. Meta-Analysis of Significance Values [R package metap version 1.6]. (2021).
277. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**, 4288–4297 (2012).
278. Federico, A. & Monti, S. hypeR: an R package for geneset enrichment workflows. *Bioinformatics* **36**, 1307–1308 (2020).

279. Pellegrino, L. *et al.* miR-23b regulates cytoskeletal remodeling, motility and metastasis by directly targeting multiple transcripts. *Nucleic Acids Research* **41**, 5400–5412 (2013).
280. Liu, Y., Hu, X., Xia, D. & Zhang, S. MicroRNA-181b is downregulated in non-small cell lung cancer and inhibits cell motility by directly targeting HMGB1. *Oncology Letters* **12**, 4181–4186 (2016).
281. Mehta, A. & Baltimore, D. MicroRNAs as regulatory elements in immune system logic. *Nature Reviews. Immunology* **16**, 279–294 (2016).
282. Alivernini, S. *et al.* MicroRNA-155-at the critical interface of innate and adaptive immunity in arthritis. *Frontiers in Immunology* **8**, 1932 (2018).
283. Zhu, F. Q. *et al.* MicroRNA-155 Downregulation Promotes Cell Cycle Arrest and Apoptosis in Diffuse Large B-Cell Lymphoma. *Oncology Research* **24**, 415–427 (2016).
284. Yu, H. *et al.* MicroRNA-155 regulates the proliferation, cell cycle, apoptosis and migration of colon cancer cells and targets CBL. *Experimental and Therapeutic Medicine* **14**, 4053 (2017).
285. Hodge, J. *et al.* Overexpression of microRNA-155 enhances the efficacy of dendritic cell vaccine against breast cancer. *Oncoimmunology* **9**, (2020).
286. Lind, E. F. *et al.* miR-155 Upregulation in Dendritic Cells Is Sufficient To Break Tolerance In Vivo by Negatively Regulating SHIP1. *The Journal of Immunology* **195**, 4632–4640 (2015).
287. Goncalves-Alves, E. *et al.* MicroRNA-155 controls T helper cell activation during viral infection. *Frontiers in Immunology* **10**, 1367 (2019).
288. Chen, L., Gao, D., Shao, Z., Zheng, Q. & Yu, Q. miR-155 indicates the fate of CD4 + T cells. *Immunology Letters* **224**, 40–49 (2020).
289. Bernard, P. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology* **27**, 1160–1167 (2009).
290. Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **33**, 690-705.e9 (2018).
291. Dai, X., Cheng, H., Bai, Z. & Li, J. Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping. *Journal of Cancer* **8**, 3131 (2017).
292. Zhang, H. *et al.* Genome-wide functional screening of miR-23b as a pleiotropic modulator suppressing cancer metastasis. *Nature Communications* **2**, (2011).

293. Helwak, A. & Tollervey, D. Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nature Protocols* **9**(3), 711–728 (2014).
294. Auerbach, O., Stout, A. P., Hammond, E. C. & Garfinkel, L. Changes in bronchial epithelium in relation to cigarette smoking and in relation to lung cancer. *The New England Journal of Medicine* **265**, 253–267 (1961).
295. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).
296. Yang, A. *et al.* p63 is essential for regenerative proliferation in limb, craniofacial and epithelial development. *Nature* **398**(6729), 714–718 (1999).
297. Daniely, Y. *et al.* Critical role of p63 in the development of a normal esophageal and tracheobronchial epithelium. *American Journal of Physiology. Cell Physiology* **287**, (2004).
298. Warner, S. M. B. *et al.* Transcription factor p63 regulates key genes and wound repair in human airway epithelial Basal cells. *American Journal of Respiratory Cell and Molecular Biology* **49**, 978–988 (2013).
299. Lawrence, M. S. *et al.* Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576 (2015).
300. Romano, R. A., Ortt, K., Birkaya, B., Smalley, K. & Sinha, S. An Active Role of the  $\Delta$ N Isoform of p63 in Regulating Basal Keratin Genes K5 and K14 and Directing Epidermal Cell Fate. *PLoS ONE* **4**, e5623 (2009).
301. Keyes, W. M. *et al.*  $\Delta$ Np63 $\alpha$  Is an Oncogene that Targets Chromatin Remodeler Lsh to Drive Skin Stem Cell Proliferation and Tumorigenesis. *Cell Stem Cell* **8**, 164–176 (2011).
302. Tomlinson, V. *et al.* JNK phosphorylates Yes-associated protein (YAP) to regulate apoptosis. *Cell Death and Disease* **1**, e29–e29 (2010).
303. Valencia-Sama, I. *et al.* Hippo component TAZ functions as a co-repressor and negatively regulates  $\Delta$ Np63 transcription through TEA domain (TEAD) transcription factor. *Journal of Biological Chemistry* **290**, 16906–16917 (2015).
304. Huang, H. *et al.* YAP suppresses lung squamous cell carcinoma progression via deregulation of the DNP63-GPX2 axis and ros accumulation. *Cancer Research* **77**, 5769–5781 (2017).

305. Zhao, R. *et al.* Yap Tunes Airway Epithelial Size and Architecture by Regulating the Identity, Maintenance, and Self-Renewal of Stem Cells. *Developmental Cell* **30**, 151–165 (2014).
306. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas Article Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
307. Cordenonsi, M. *et al.* The hippo transducer TAZ confers cancer stem cell-related traits on breast cancer cells. *Cell* **147**, 759–772 (2011).
308. Muramatsu, T. *et al.* YAP is a candidate oncogene for esophageal squamous cell carcinoma. *Carcinogenesis* **32**, 389–398 (2011).
309. Lee, K. W. *et al.* Significant association of oncogene YAP1 with poor prognosis and cetuximab resistance in colorectal cancer patients. *Clinical Cancer Research* **21**, 357–364 (2015).
310. Eun, Y. G. *et al.* Clinical significance of YAP1 activation in head and neck squamous cell carcinoma. *Oncotarget* **8**, 111130–111143 (2017).
311. Lamar, J. M. *et al.* The Hippo pathway target, YAP, promotes metastasis through its TEAD-interaction domain. *Proceedings of the National Academy of Sciences of the United States of America* **109**, (2012).
312. Mahoney, J. E., Mori, M., Szymaniak, A. D., Varelas, X. & Cardoso, W. V. The Hippo Pathway Effector Yap Controls Patterning and Differentiation of Airway Epithelial Progenitors. *Developmental Cell* **30**, 137–150 (2014).
313. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2**, 401–404 (2012).
314. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, (2013).
315. Giorgi FM. aracne.networks: ARACNe-inferred gene networks from TCGA tumor datasets. *R package version 1.20.0* (2021). Available at: <https://bioconductor.org/packages/release/data/experiment/html/aracne.networks.html>. (Accessed: 11th November 2021)
316. Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**, 2233–2235 (2016).

317. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
318. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. In *Bioinformatics: The Impact of Accurate Quantification on Proteomic and Genetic Analysis and Research* 41–74 (2014). doi:10.1201/b16589
319. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* *2012* **9**:4 **9**, 357–359 (2012).
320. Broad Institute. Picard Tools - By Broad Institute. *Github* <http://broadinstitute.github.io/picard> (2009). Available at: <https://broadinstitute.github.io/picard/>. (Accessed: 11th November 2021)
321. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
322. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**, 1–9 (2008).
323. Zhu, L. J. *et al.* ChIPpeakAnno: A Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, 1–10 (2010).
324. Yu, G., Wang, L. G. & He, Q. Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
325. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576–589 (2010).
326. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021). doi:10.1101/060012
327. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
328. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nature Genetics* **51**, 1442–1449 (2019).
329. Mixed-Effects Models in S and S-PLUS. *Mixed-Effects Models in S and S-PLUS* (2000). doi:10.1007/B98882

330. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782–795 (2013).
331. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
332. Deprez, M. *et al.* A single-cell atlas of the human healthy airways. *bioRxiv* 2019.12.21.884759 (2019). doi:10.1101/2019.12.21.884759
333. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
334. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nature Communications* **12**, (2021).
335. Rocco, J. W., Leong, C. O., Kuperwasser, N., DeYoung, M. P. & Ellisen, L. W. p63 mediates survival in squamous cell carcinoma by suppression of p73-dependent apoptosis. *Cancer Cell* **9**, 45–56 (2006).
336. Zhenjiawang, Z. *et al.* BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics* **34**, 2867–2869 (2018).
337. Keenan, A. B. *et al.* ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Research* **47**, W212 (2019).
338. Thomas, Z. V, Wang, Z. & Zang, C. BART Cancer: a web resource for transcriptional regulators in cancer genomes. *NAR Cancer* **3**, (2021).
339. Qu, J. *et al.* Mutant p63 Affects Epidermal Cell Identity through Rewiring the Enhancer Landscape. *Cell Reports* **25**, 3490-3503.e4 (2018).
340. Somerville, T. D. D. *et al.* TP63-Mediated Enhancer Reprogramming Drives the Squamous Subtype of Pancreatic Ductal Adenocarcinoma. *Cell Reports* **25**, 1741 (2018).
341. Li, Y. *et al.* YAP expression and activity are suppressed by S100A7 via p65/NFkB-mediated Repression of DNp63. *Molecular Cancer Research* **15**, 1752–1763 (2017).
342. Wang, R. *et al.* S100A7 promotes lung adenocarcinoma to squamous carcinoma transdifferentiation, and its expression is differentially regulated by the Hippo-YAP pathway in lung cancer cells. *Oncotarget* **8**, 24804 (2017).
343. Shenoy, A. T. *et al.* Antigen presentation by lung epithelial cells directs CD4+ TRM cell function and regulates barrier immunity. *Nature Communications* **12**, 1–16 (2021).

344. Ban, Y. *et al.* Radiation-activated secretory proteins of Scgbl1a1 + club cells increase the efficacy of immune checkpoint blockade in lung cancer. *Nature Cancer* **2**, 919–931 (2021).
345. Steimle, V., Siegrist, C. A., Mottet, A., Lisowska-Grospierre, B. & Mach, B. Regulation of MHC Class II Expression by Interferon- $\gamma$  Mediated by the Transactivator Gene CIITA. *Science* **265**, 106–109 (1994).
346. Chang, C. H., Fontes, J. D., Peterlin, M. & Flavell, R. A. Class II transactivator (CIITA) is sufficient for the inducible expression of major histocompatibility complex class II genes. *The Journal of Experimental Medicine* **180**, 1367–1374 (1994).
347. Stockinger, B. *et al.* A role of Ia-associated invariant chains in antigen processing and presentation. *Cell* **56**, 683–689 (1989).
348. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
349. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, Lahr DL, Hirschman JE, Liu Z, Donahue M, Julian B, Khan M, Wadden D, Smith IC, Lam D, Liberzon A, Toder C, Bagul M, Orzechowski M, Enache OM, Piccio, G. T. A Next Generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2018).
350. Flaherty, K. T. *et al.* Combined BRAF and MEK Inhibition in Melanoma with BRAF V600 Mutations. *The New England Journal of Medicine* **367**, 1694 (2012).
351. Vittoria, M. A. *et al.* Inactivation of the Hippo Tumor Suppressor Pathway Promotes Melanoma. *bioRxiv* 2021.05.04.442615 (2021). doi:10.1101/2021.05.04.442615
352. Saladi, S. V. *et al.* ACTL6A Is Co-Amplified with p63 in Squamous Cell Carcinoma to Drive YAP Activation, Regenerative Proliferation, and Poor Prognosis. *Cancer Cell* **31**, 35–49 (2017).
353. Chang, C. Y. *et al.* Increased ACTL6A occupancy within mSWI/SNF chromatin remodelers drives human squamous cell carcinoma. *Molecular Cell* **81**, 4964–4978.e8 (2021).
354. Johnson, D. B. *et al.* Melanoma-specific MHC-II expression represents a tumour-autonomous phenotype and predicts response to anti-PD-1/PD-L1 therapy. *Nature Communications* **7**, (2016).

355. Gil Del Alcazar, C. R. *et al.* Insights into immune escape during tumor evolution and response to immunotherapy using a rat model of breast cancer. *Cancer Immunology Research* (2022). doi:10.1158/2326-6066.CIR-21-0804
356. Beyaz, S. *et al.* Dietary suppression of MHC class II expression in intestinal epithelial cells enhances intestinal tumorigenesis. *Cell Stem Cell* **28**, 1922-1935.e5 (2021).
357. Sconocchia, G. *et al.* HLA class II antigen expression in colorectal carcinoma tumors as a favorable prognostic marker. *Neoplasia* **16**, 31–42 (2014).
358. Andres, F. *et al.* Expression of the MHC class II pathway in triple-negative breast cancer tumor cells is associated with a good prognosis and infiltrating lymphocytes. *Cancer Immunology Research* **4**, 390–399 (2016).
359. Marjanovic, N. D. *et al.* Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. *Cancer Cell* **38**, 229-246.e13 (2020).
360. Gould, M. K., Ananth, L. & Barnett, P. G. A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. *Chest* **131**, 383–388 (2007).
361. Massion, P. P. & Walker, R. C. Indeterminate pulmonary nodules: risk for having or for developing lung cancer? *Cancer Prevention Research* **7**, 1173–1178 (2014).
362. Farazi, T. A., Spitzer, J. I., Morozov, P. & Tuschl, T. MiRNAs in human cancer. *Journal of Pathology* **223**, 102–115 (2011).
363. Ewels, P., Magnusson, M., Lundin, S. & Källner, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
364. Nathan Wong & Xiaowei Wang. miRDB: an online resource for microRNA target prediction and functional annotations . *Nucleic Acids Research* **43**, D146–D152 (2015).
365. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *Journal of Statistical Software* **36**, 1–13 (2010).
366. CRAN - Package caret. Available at: <https://cran.r-project.org/web/packages/caret/index.html>. (Accessed: 1st July 2022)
367. Marchionni, L., Afsari, B., Geman, D. & Leek, J. T. A simple and reproducible breast cancer prognostic test. *BMC Genomics* **14**, 1–7 (2013).

368. Choi, Y. *et al.* Improving lung cancer risk stratification leveraging whole transcriptome RNA sequencing and machine learning across multiple cohorts. *BMC Medical Genomics* **13**, (2020).
369. Dobzhansky, T. Nothing in Biology Makes Sense except in the Light of Evolution. *The American Biology Teacher* **35**, 125–129 (1973).
370. Junttila, M. R. & De Sauvage, F. J. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* **501**(7467), 346–354 (2013).
371. Binnewies, M. *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nature Medicine* **24**(5), 541–550 (2018).
372. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature Protocols* **15**, 1484–1506 (2020).
373. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature Methods* **17**, 159–162 (2020).
374. Luca, B. A. *et al.* Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell* **184**, 5482–5496.e28 (2021).
375. Jerby-Arnon, L. & Regev, A. Mapping multicellular programs from single-cell profiles. *BioRxiv* (2020).
376. Zhang, A. W. *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nature Methods* **16**, 1007–1015 (2019).
377. Wang, Z. *et al.* Celda: A Bayesian model to perform co-clustering of genes into modules and cells into subpopulations using single-cell RNA-seq data. *bioRxiv* 2020.11.16.373274 (2021). doi:10.1101/2020.11.16.373274
378. Dhainaut, M. *et al.* Spatial CRISPR genomics identifies regulators of the tumor microenvironment. *Cell* **185**, 1223–1239.e20 (2022).
379. Nirmal, A. J. *et al.* The Spatial Landscape of Progression and Immunoediting in Primary Melanoma at Single-Cell Resolution. *Cancer Discovery* **12**, 1518–1541 (2022).
380. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: integrating immunity’s roles in cancer suppression and promotion. *Science* **331**, 1565–1570 (2011).

381. Motzer, R. J. *et al.* Nivolumab plus Ipilimumab versus Sunitinib in Advanced Renal-Cell Carcinoma. *The New England Journal of Medicine* **378**, 1277–1290 (2018).
382. Seewaldt, V. ECM stiffness paves the way for tumor cells. *Nature Medicine* **20**(4), 332–333 (2014).
383. Pothapragada, S. P., Gupta, P., Mukherjee, S. & Das, T. Matrix mechanics regulates epithelial defence against cancer by tuning dynamic localization of filamin. *Nature Communications* **13**(1), 1–12 (2022).
384. Lai, X. *et al.* Epithelial-Mesenchymal Transition and Metabolic Switching in Cancer: Lessons From Somatic Cell Reprogramming. *Frontiers in Cell and Developmental Biology* **8**, 760 (2020).
385. Mazzilli, S. A. *et al.* Vitamin D Repletion Reduces the Progression of Premalignant Squamous Lesions in the NTCU Lung Squamous Cell Carcinoma Mouse Model. *Cancer Prevention Research* **8**, 895 (2015).
386. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, (2017).
387. Naxerova, K. Mutation fingerprints encode cellular histories. *Nature* **597**, 334–336 (2021).
388. Colom, B. *et al.* Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. *Nature Genetics* **52**, 604–614 (2020).
389. Fowler, J. C. & Jones, P. H. Somatic Mutation: What Shapes the Mutational Landscape of Normal Epithelia? *Cancer Discovery* **12**, 1642–1655 (2022).
390. Colom, B. *et al.* Mutant clones in normal epithelium outcompete and eliminate emerging tumours. *Nature* **598**, 510–514 (2021).
391. Griffin, G. K. *et al.* Epigenetic silencing by SETDB1 suppresses tumour intrinsic immunogenicity. *Nature* **595**(7866), 309–314 (2021).

**CURRICULUM VITAE**

