MULTI-OMICS CHARACTERIZATION OF THE MOLECULAR EFFECTS OF SMOKING AND CHRONIC INFLAMMATION ON THE LUNG

by

Emily A. Vucic

B.Sc., University of British Columbia, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Pathology and Laboratory Medicine)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2014

Abstract

Chronic obstructive pulmonary disease (COPD) is a progressive, inflammatory lung disease associated with a 10-fold increased risk of lung cancer (LC), independent of smoking-status. Together these diseases contribute tremendously to morbidity and mortality worldwide. While COPD and lung cancer share common etiologies including genetic susceptibilities and risk factors, the biology driving COPD and LC is largely unknown. No effective treatments exist for either disease, thus a better understanding of the molecular biology underlying these diseases is urgently needed.

The overarching hypothesis of this thesis is that specific risk factors, such as smoking and chronic inflammation lead to selective disruption of genes in exposed tissues and that these selectively disrupted genes contribute directly to COPD and lung cancer pathogenesis. Since selection occurs at the DNA level, and tumour and disease systems may be altered at multiple genetic and epigenetic levels; a major hypothesis of this thesis is that loci which sustain high-level concerted genetic, epigenetic and/or transcriptional disruptions in tissues involved in disease pathology are likely indicative of strong selection and may be identified by applying an integrative multi-omics analysis of these tissues.

Background pertaining to the rationale, objectives and specific aims of this work are described in Chapter 1. Chapters 2-4 detail the main findings of this thesis, which are that: **1)** DNA is altered at the main sites of airflow obstruction in COPD patients (Chapter 2), **2)** smoking status impacts miRNA lung tumour biology and patient prognosis (Chapter 3), **3)** lung tumours from patients with COPD are molecularly distinct at the genetic and epigenetic levels (Chapter 4) and **4)** genes preferentially altered in COPD-related lung tumours are aberrantly methylated in non-malignant airway cells from patients with COPD and lung cancer (Chapter 4). Taken together, this work provides sufficient rationale to explore the clinical application of these findings as potential targets for novel COPD treatments and markers for early lung cancer detection, treatment or targeted chemoprevention. A summary of these key findings, significance, caveats and future directions are discussed in Chapter 5.

Preface

The research in this thesis was conducted with ethics approval from the UBC Research Ethics Board, Certificate Numbers: CIHR H10-02114, NIH H11-03247, EDRN H09-00008, CCSRI H09-00934 and DOD W81XW-10-1-0634.

Portions of Chapters 1 and 5 have been published as: **Vucic EA*†**, Thu KL*, Robison K, Rybaczyk LA, Chari R, Alvarez CE, Lam WL. Translating cancer 'omics to improved outcomes. *Genome Research.* 2012; Feb 22(2):188-95. ***co-first authors, †corresponding author.** This is a perspective article. I co-wrote and am corresponding author of this manuscript.

A version of Chapter 2 has been published as: **Vucic EA†**, Chari R, Thu KL, Wilson IM, Cotton AM, Kennett JY, Zhang M, Lonergan KM, Steiling K, Brown CJ, McWilliams A, Ohtani K, Lenburg ME, Sin DD, Spira A, MacAulay CE, Lam S, Lam WL. Aberrant DNA methylation patterns are widespread in small airways of former smokers with COPD. *American Journal of Respiratory Cell and Molecular Biology*. 2014; May;50(5):912-22 **†corresponding author**. I am first author and corresponding author of this manuscript. I performed data analysis, interpreted results and wrote the manuscript.

A version of Chapter 3 is currently under review for publication: **Vucic EA***, Thu KL*, Pikor LA, Enfield KSS, Yee, J, English JC, Macaulay CE, Lam S, Lam WL. Smoking Status Impacts microRNA Mediated Prognosis and Lung Adenocarcinoma Biology. ***co-first authors, †corresponding author.** I co-designed the study, performed and interpreted data analysis and results, co-wrote the manuscript and am corresponding author.

A manuscript relating to the methods portion of Chapter 4 is being prepared for publication as: Mosslemi, M*, Thu, KL*, **Vucic, EA**, Pikor, LA, Ng RT, MacAulay, CE, Lam, WL. Development of a Multi-dimensional Integrative Tumour gene Ranking Algorithm (MITRA) for the identification of candidate genes in cancer. *co-first authors. I developed the idea for this algorithm, and helped with the design and interpretation of analysis and validation methods of the algorithm.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AC	adenocarcinoma
BCCA	British Columbia Cancer Agency
B-H	Benjamini Hochberg
COPD	chronic obstructive pulmonary disease
CS	current smoker
CSN	current smoker non-malignant tissue
CST	current smoker tumour tissue
DE	differentially expressed
DM	differentially methylated
ES	ever smoker
FEV_1	Forced Expiratory Volume at 1 second
FS	former smoker
FSN	former smoker non-malignant tissue
FST	former smoker tumour tissue
FVC	Forced Vital Capacity
GOLD	Global Initiative for Chronic Obstructive Lung Disease
GWAS	genome-wide association studies
IARC	International Agency for Research on Cancer
ICGC	International Cancer Genome Consortium
InS	Integrated Score
LC	lung cancer
MANOVA	multivariate analysis of variance
miRNA	micro-RNA
mRNA	messenger RNA
ncRNA	non-coding RNA
NS	never smoker
NSCLC	non-small cell lung cancer
NSN	never smoker non-malignant tissue
NST	never smoker tumour tissue
OE	overexpressed
PAH	polycyclic aromatic hydrocarbons
ROS	reactive oxidative species
SAE	small airway epithelia
SCLC	small cell lung cancer
SNP	single nucleotide polymorphism
SqCC	sqamous cell carcinoma
TCGA	The Cancer Genome Atlas
TSG	tumour suppressor gene
UE	underexpressed

Acknowledgements

I would like to sincerely thank the numerous members of the Wan Lam Lab, past and present who have contributed to much of the work presented here and provided invaluable insight, mentorship, support and advice over many hours of discussion throughout the years. I would also like to acknowledge the support and contribution of our many collaborators at St. Paul's hospital (Drs. Don Sin, Raymond Ng, Maen Obeidat, Peter Pare), Vancouver General Hospital (Drs. John English, John Yee), Princess Margaret Cancer Centre in Toronto (Dr. Igor Jurisica), Boston University (Drs. Avi Spira and Marc Lenburg) and The Research Institute at Nationwide Children's Hospital in Columbus (Carlos Alvarez) as well as the support of the Department of Pathology and Laboratory Medicine at UBC (Dr. Haydn Pritchard and Aleya Abdulla).

I owe particular thanks to my supervisor Dr. Wan Lam, my supervisory committee members: Drs. Carolyn Brown, Stephen Lam, Calum MacAulay and David Walker for their mentorship, guidance and support, and to my wonderful family and friends, especially: Bryce Janssens, Danuta Vucic, Ivan Vucic, Leo Janssens and Trudy Janssens.

I would like to also acknowledge generous scholarship support from the Canadian Institutes of Health Research (Frederick Banting and Charles Best Canada Doctoral Award), the University of British Columbia (Four Year Doctoral Fellowship) and the Department of Pathology and Laboratory Medicine.

The research presented in this thesis was funded by the following granting agencies: Canadian Institutes of Health Research, National Institutes of Health, Canadian Cancer Society Research Institute, US Department of Defense, National Cancer Institute Early Detection Research Network and the Canary Foundation.

Dedication

This thesis is dedicated to my scientific mentors. In order of appearance, they are: Ivan Vucic, Mr. Becker, Jamie Doran, Wan Lam, Calum MacAulay, Carolyn Brown, Stephen Lam and David Walker.

1 Chapter: Introduction

1.1 COPD and lung cancer disease burden

Chronic obstructive pulmonary disease (COPD) is a highly prevalent progressive inflammatory lung disease affecting one million Canadians and 300 million people worldwide [1-3]. Globally, COPD kills 3 million people every year and by 2030, this disease will be responsible for over 7 million deaths annually, making it the 3rd leading cause of mortality (currently 4th) and the 5th leading cause of disability in the world (currently 12th) [1, 4, 5]. Nationally, 1 in 5 Canadians 40 years and older have COPD and 10,000 Canadians each year die from it [6]. The accelerated prevalence, morbidity, and mortality of COPD around the world is largely due to: a growing aging population in the developed world, the continued rising prevalence of tobacco smokers worldwide (currently estimated at 1 billion men and 250 million women [7]), and increased exposure to risk factors such as pollution [3].

Lung cancer is the leading cause of cancer mortality worldwide, resulting in 1.37 million deaths each year [8, 9]. By 2030, lung cancer is estimated to be the 5th leading cause of death in the world [10]. In 2012 alone, 1.8 million people were diagnosed with lung cancer, of which 1.6 million people died, representing 20% of all cancer deaths worldwide [10]. While cigarette smoke is the number one risk factor for both COPD and lung cancer; COPD is associated with an up to 10-fold increased risk of lung cancer, independent of smoking-status [11]. The combination of the most common chronic disease and cancer type is alarming and places an enormous burden on patients and the healthcare system.

Relative to the incredible burden of COPD and lung cancer, there exists a gross disparity in funding resources directed towards an improved understanding of these diseases [12]. As such, little is known about i) the molecular biology underlying COPD pathogenesis, ii) how lung cancer develops in the context of different cancer promoting environments (e.g., chronic inflammation and different smoking or non-smoking histories), or iii) how to identify and treat patients at highest cancer risk. As a result there are currently no markers or therapies to predict, treat or prevent COPD or lung cancer progression. There exists an urgent

need for an improved understanding of the molecular mechanisms underlying lung cancer biology specifically in the context of smoking and COPD.

1.2 COPD: definition and disease etiology

COPD is diagnosed based on spirometry measures which assess the amount of air an individual can exhale with force after inhaling as deeply as possible (Forced Vital Capacity (FVC)) and the amount of air an individual can exhale with force in one breath at 1 second (Forced Expiratory Volume, (FEV_1)). A diagnosis of COPD is made as per guidelines set by the Global Initiative for Chronic Obstructive Lung Disease (GOLD), on the basis of a post-bronchodilator irreversible airflow obstruction reading of $FEV_1/FVC < 0.70$, that is not due to bronchiectiasis, cystic fibrosis, or previous tuberculosis [5]. FEV_1% predicted measure classifies patients into stages of COPD severity.

Table 1.1 Classification of severity of airflow obstruction in COPD

In patients with $FEV_1/FVC < 0.70$:		
GOLD 1	Mild	$FEV_1 \geq 80\%$ predicted
GOLD 2	Moderate	$50\% \leq FEV_1 \leq 80\%$ predicted
GOLD 3	Severe	$30\% \leq FEV_1 \leq 50\%$ predicted
GOLD 4	Very Severe	$FEV_1 \leq 30\%$ predicted

COPD is characterized by chronic inflammation, progressive decline in lung function and extensive remodeling of pulmonary tissues including airway walls and lung parenchyma tissues in the form of small airway disease and emphysema [13, 14]. Substantial heterogeneity exists among COPD patients with respect to extent and distribution of emphysema or small airway disease [15, 16]. Airflow obstruction and progressive decline in lung function characteristic of COPD is largely attributed to extensive remodeling of small airways [17-20], which precedes emphysematous destruction of lung parenchyma in COPD patients. Airway wall thickening and obliteration of the number of terminal bronchioles occurs increasingly with increased disease severity [21]. Elucidating the molecular mechanisms underlying airway remodeling in COPD is thus highly relevant to the design of therapeutic and prevention regimes.

1.2.1 COPD susceptibility

Collectively, results from large genome-wide association studies (GWAS) indicate that genetic factors contribute to COPD. For example, COPD risk and spirometry measures have been shown to be heritable [22-25], and genotypes associated with lung function (*FAM13A* and *HHIP* on 4q31) and genes involved in processes integral to COPD pathology including protease/anti-protease balance (*MMP1, MMP12*), inflammation, nicotine response (*CHRNA3*), antioxidant response (*SOD3*, *EPHX1*) and xenobiotic metabolism (*GSTP1*, *CYP1A1*, *EPHX1*, *CYP2A1*) [26-29] have been described in multiple studies [30-33].

Severe alpha-1-antitrypsin deficiency due to homozygous mutations of the *SERPINA1* gene is a documented cause of hereditary and early onset COPD [34, 35]. *SERPINA1* codes for alpha-1 antitrypsin (A1AT) -- a serine peptidase inhibitor critical to protecting lung tissue from enzymatic degradation by proteolytic enzymes released from neutrophils. Based on a family-based linkage analysis in the Boston Early-Onset COPD Study, combined with human gene expression and animal model findings, SNP haplotypes in another serine peptidase inhibitor, *SERPINE2*, were also discovered to be associated with COPD phenotypes [36].

1.2.2 Molecular profiling of tissues involved in COPD

To date, molecular characterization of COPD related tissues has largely focused on correlating array-based mRNA expression profiles of large and small airways and lung parenchymal tissues with various COPD clinical phenotypes, such as COPD severity [37-45]. Collectively, this work provides substantial evidence that impaired oxidative stress response is an important factor in COPD pathology.

Increased oxidative stress and the generation of free radicals (e.g., from cigarette smoke or inflammatory cells), affects nearly all aspects of COPD pathology and pathogenesis. Increased levels of reactive oxidative species (ROS) in airways of COPD patients and smokers is reflected by increased markers of oxidative stress in sputum, breath, lungs, and blood in patients with COPD (reviewed in [46]). Associated damage due to free radicals affects inflammatory, immune and epithelial cells of the airways, resulting in: oxidative inactivation of anti-proteases and surfactants, mucus hypersecretion, membrane lipid peroxidation, mitochondrial respiration, alveolar epithelial injury, remodeling of extracellular matrix and blood vessels, necrosis, apoptosis, and inflammation [46]. Based on its presence and relevance to COPD biology and progression, the development of anti-oxidant based pharmacological strategies for COPD therapeutics is an active area of research [46].

1.3 Lung cancer epidemiology and etiology

Broadly, lung cancer can be subdivided into two major histological groups based on cell type of origin, growth patterns and location in the lung; small cell lung cancer (SCLC) and non small cell lung cancer (NSCLC) account for 15% and 85% of lung cancer cases, respectively [47]. NSCLC can be further subdivided into squamous cell carcinoma (SqCC) which occurs predominantly in the central airways, adenocarcinoma (AC) and large cell carcinoma which occur predominantly in the peripheral airways or lung parenchyma. SqCC and AC account for the vast majority of NSCLC cases. This thesis will focus on AC, which arise in the lung peripheries from type II pneumocytes or clara cells and account for over 40% of all lung cancers overall [7].

Different smoking histories and lung cancer subtypes are associated with distinct and highly clinically relevant histological and molecular pathology [48-54]. Lung cancer incidence, patient demographics and mortality rates closely mirror a population's smoking behaviour including smoking duration, amount of tobacco smoked and duration of smoking cessation (for former smokers) [7]. Thus consideration of smoking history and tumour histology in molecular lung cancer studies is an important consideration to lung cancer research.

Other factors modulating lung cancer risk include: airflow obstruction (specifically low FEV_1% predicted, i.e., severe COPD), a family history of lung cancer, exposure to asbestos, crystalline silica, and chronic exposure to high levels of arsenic, radon gas, heavy metals, second hand smoke and infection by certain viruses (HPV and Epstein Barr Virus) [7].

1.3.1 Lung cancer genetics and targeted therapies

Traditionally, treatment decisions for NSCLC were based solely on histology, but there has been a shift in recent years to incorporate molecular subtype information into treatment paradigms, specifically genetic alterations that are causative or "drive" lung tumourigenesis [55]. Lung cancer driver mutations have been identified for multiple oncogenes and tumour suppressor genes (Table 1.2) [48, 55-58].

These genes map to several common pathways which are frequently deregulated in lung cancer at the mRNA and protein levels, including: EGFR, PI3K-AKT, P53, RB/E2F, and WNT signalling pathways which regulate cellular proliferation, death, cell cycle check-points, angiogenesis, invasion and DNA repair. Therapies targeting several of the alterations are shown (Table 1.2) [58-63].

Table 1.2 Non small cell lung cancer onco- and tumour suppressor genes

Gene	Oncogenes	Alteration	Frequency	Drug
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	m,A	10-30	Tipifarnib, lonafarnib
EGFR	epidermal growth factor receptor	m,A	10-25	Gefitinib, erlotinib, cetuximab
mTOR	mechanistic target of rapamycin (serine/threonine kinase)	A	70-75	Rapamycin, RAD001, CCL-779
MYC	v-myc avian myelocytomatosis viral oncogene homolog	A	5-20	
EML/ALK	echinoderm microtubule associated protein like 4/anaplastic lymphoma receptor tyrosine kinase	fusion	5-13	Crizotinib
HER2	human epidermal growth factor receptor 2	m,A	5-10	Trastuzumab
BRAF	v-raf murine sarcoma viral oncogene homolog B1	m	2-3	Sorafenib
PIK3CA	phosphoinositide-3-kinase, catalytic, α polypeptide	m	1-3	BEZ235, LY294002
AKT1	v-akt murine thymoma viral oncogene homolog 1	m	0.3	
MAP2K1	mitogen-activated protein kinase kinase 1	m	1-2	Trametinib, salumetinib
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	m	0.5	
ROS1	c-ros oncogene 1 , receptor tyrosine kinase	fusion	1	
RET	ret proto-oncogene	fusion	1.3	
	Tumour suppressor genes (TSG)			
CDKN2A	cyclin-dependent kinase inhibitor 2A	m,D,HM	>50	
LKB1	serine/threonine kinase 11	m,D,HM	>30	
RARB	retinoic acid receptor, beta	HM	40	
TP53	tumour protein p53	m,D	>50	
RASSF1A	Ras association (RalGDS/AF-6) domain family member 1	m,D,HM	50-80	
FHIT	fragile histidine triad	m,D,HM	50-70	
RB1	retinoblastoma 1	D	>50	

m: mutated; A: amplified; fusion: gene fusion; D: deleted; HM: hypermethylated

1.4 Rationale for studying the biological links between COPD and lung cancer

Epidemiological and genetic evidence suggests a mechanistic link between COPD and lung cancer [64]. Cigarette smoke is the greatest risk factor for both diseases, accounting for 90% of COPD and 75-80% of lung cancer cases. Chronic inflammation -- which can be defined as an abnormally prolonged protective response to a loss of tissue homeostasis [65, 66], is associated with cancer in almost all affected organs and tissues, including the lung [64]. The association between chronic inflammation and cancer was first described by Virchow over 150 years ago [67, 68]. Today inflammation is a hallmark of cancer [69, 70]. Cellular and systemic responses to cigarette smoke and inflammation are tightly linked.

Cigarette smoke contains over 5000 chemicals, 73 of which have been classified by the International Agency for Research on Cancer (IARC) as carcinogenic [8]. These

chemicals can be broken down into the following classes: polycyclic aromatic hydrocarbons (PAHs), aza-arenes, N-nitrosamines, aromatic amines, heterocyclic aromatic amines, aldehydes, volatile hydrocarbons and nitro compounds (as well as metals and various organic and inorganic compounds) [71, 72]. The metabolism of these compounds by phase I (cytochrome P450 monooxygenase, also called CYPs) and II (e.g., glutathione S-transferases (GSTs)) xenobiotic metabolizing enzymes results in formation of reactive diol epoxides, which can bind directly to DNA, are geno- and cytotoxic, and strong inducers of the cell's ROS response and pro-inflammatory pathways. Inflammation is itself a major source of ROS. Under normal conditions, pro-inflammatory cascades are inhibited by wound-healing/anti-inflammatory signals. However, prolonged exposure to toxic onslaughts from cigarette smoke and ROS leads to abnormal activation of this normal immune response, resulting in COPD pathology and lung tumourigenesis.

The precise mechanisms mediating this abnormal response and that confer increased COPD and cancer risks are largely unknown. Evidence from multiple large scale GWAS linking COPD and lung cancer points to the involvement of genetic variants which modulate activity of xenobiotic enzymes (*GSPT1, CYP1A1*) affecting rate of production of damaging reactive metabolites, ROS response (*EPHX1, SOD3*) and DNA damage repair (XRRC1) [73]-[74, 75].

Given the i) strong association between smoking, lung cancer and COPD, ii) relevance of smoking history to lung tumour biology and patient outcomes, iii) shared genetic risk variants and iv) strong association of inflammation with increased cancer risk in multiple organs; it is plausible that the mechanisms underlying lung tumourigenesis in COPD patients overlaps with the biology of smoking response and COPD pathology. Studying lung tumourigenesis in the context of smoking and chronic inflammation is necessary to elucidate these mechanisms and for informing development of prevention, early detection and treatment strategies of these deadly and highly prevalent diseases.

1.5 Genomic mechanisms disrupted in tumour systems

Cancer genomics refers to the study of tumour genomes using various profiling strategies to obtain sequence, structural, quantitative, qualitative, functional or chemical information about: DNA, non-coding RNA (nc-RNA), mRNA and protein -- molecular levels which may be referred to collectively as 'omics. Background relating to genetic and epigenetic mechanisms interrogated in this thesis is summarized below.

1.5.1 Copy number

Tumour genomes sustain structural aberrations to genome architecture [76, 77]. Structural alterations can result in balanced genomic exchanges, where equal genomic material is exchanged between two genomic regions, or unbalanced alterations, where the number of copies of genomic portions are gained, amplified, lost or homozygously deleted [76]. In tumours, copy number alterations, occur commonly as 1) double minutes-- small amplified acentric DNA fragments, 2) contiguous homogeneously staining regions which are incorporated directly into chromosomes or 3) distributed sporadically throughout the genome. Copy number variation in the normal population and in tumour genomes is due to with high recombination rates between stretches of non allelic homologous flanking repeats [78]. The probability of misalignment between non allelic homologous regions in the genome appears to be strongly influenced by factors such as sequence homology, length, and orientation [78]. In tumours, recombination rates are exaggerated due to impaired DNA damage surveillance and repair mechanisms; thus copy number disruptions in tumour genomes are highly prominent. In tumours, copy number alterations affect genomic stability, mRNA and protein expression and function, particularly that of oncogenes and tumour suppressor genes (TSG) thereby contributing directly to many key aspects of tumour biology [77, 79-82].

1.5.2 DNA methylation

DNA methylation occurs at the 5 carbon position of cytosine in CpG dinucleotide sequences, creating 5'-methylcytosine (5mC). DNA methylation marks are mitotically inherited [83]. These enzymes are instrumental in establishing methylation patterns during normal gastrulation, and are found to be active in cancer cells and throughout establishment

of immortalized cell lines [84, 85]. Short regions enriched for CpG dinucleotides form areas called CpG islands that are often conserved through evolution [86]. CpG islands are generally associated with gene promoter regions where they are usually unmethylated, however promoters without CpG islands are frequently methylated [87]. The majority of 5mC is located in repetitive DNA sequences, such as retrotransposons and satellite DNA, where it is believed to be important in transcriptional silencing. Methylation also serves as a level of transcriptional regulation specific to tissue type [87, 88]. DNA methyl transferase (DNMT) enzymes catalyze the transfer of the methyl group from S-adenosyl L-methionine (SAM) to the cytosine in CpG dinucleotides. The maintenance of methylation patterns in somatic tissue is achieved by DNMT1[89], while *de novo* methylation involves DNMT3a, DNMT3b, and DNMT3L.

CpG dinucleotides are under-represented in the human genome, due to spontaneous deamination of 5mC which results in the conversion of 5mC to thymidine, which is not recognized by repair machinery. Thus methylated CpGs are mutation hotspots [90]. Transcription involves the recruitment of RNA polymerase II (Pol II), and the recruitment of additional cofactors, including histone modifying enzymes. Methylation can directly inhibit transcription by impeding binding of transcription factors to promoters, or indirectly through the recruitment of regulatory proteins, such as histone modifying enzymes or transcriptional repressors and the promotion of chromatin configurations unfavorable to transcription [83]. Disruption to normal patterns of DNA methylation, adversely affects genomic stability and gene expression, and has been described in many developmental diseases and nearly all cancer types [83, 91].

1.5.3 Micro-RNA

MicroRNAs (miRNAs) are small (18 to 25 bps) non-coding RNAs (ncRNA) that negatively regulate translation of mRNA via direct inhibition of translation or induction of mRNA degradation [92]. Genes encoding miRNA-- which are much larger than mature miRNA, are transcribed by RNA polymerase II as large primary transcripts (pri-microRNA) [93]. Pri-miRNAs are processed to double stranded, ~70 nucleotide imperfect hairpin loop precursor microRNAs (pre-microRNAs) by a protein complex containing the RNase III

enzyme Drosha and the double-stranded-RNA-binding protein, Pasha/DGCR8 [94]. Pre-miRNAs are transported to the cytoplasm where they are processed by a second RNase III enzyme, DICER, which forms mature miRNAs. Dicer contains multiple functional domains including two catalytic RNAse III domains (IIIa and IIIb) which independently cut one of the pre-miRNA strands forming single stranded miRNA, one of which (the guide strand) is incorporated into the RNA-induced silencing (RISC) complex. RISC mediates gene silencing by i) translational repression or ii) mediating mRNA cleavage [93, 95]. Translational repression occurs if sequence homology between miRNA and mRNA are low. Conversely, mRNA cleavage occurs when sequence homology between miRNA and mRNA are very high [93].

The ability of miRNAs to exert repressive translational functions through low sequence homology is the more common scenario, and is attributed to the ability of a single miRNA to target up to hundreds of unique mRNAs [93, 95]. Unsurprisingly, miRNAs are involved in almost all biological processes [96]. In nearly all human malignancies, miRNAs have been shown to contribute to tumorigenesis, progression and treatment response, thereby representing promising biologically relevant biomarkers [97-99].

1.6 Rationale for multi-omics approach to analyzing tumour systems

The broad goal of cancer genomics is to survey 'omics data to identify genes and pathways deregulated in cancer that may be useful for the detection and management of disease. Towards this aim, two large international research efforts: the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), have publically released such multi-omics data for thousands of tumours from dozens of cancer types [100-102]. To date, this type of work has improved our understanding of cancer as a disease and revealed clinically relevant diagnostic, prognostic and druggable targets. However, it has also unveiled the immense genomic complexity, and striking inter- and intra- tumour heterogeneity that occurs at most ‘omics levels and even exists between histologically similar tumours [103-107].

As with lung and other cancers, it is now appreciated that the molecular profiles of histologically similar tumours are varied and this is relevant to treatment response [59]. These findings have had immediate implications to the translation of cancer genomics research, particularly to the design and interpretation of clinical drug trials. For example, in the absence of "a genetically simple disease addicted to a single pathway and relatively nontoxic drug", individual biomarkers cannot always predict clinical response to therapeutics [108]. Therefore, for genetically complex tumours, matching a single target in a patient's tumour with a targeted therapy may not be sufficient to predict whether they will do better or worse on targeted *vs.* standard therapy [108]. This is evident in lung cancer, where there exists a stark contrast between treatment response statistics [109] and the number of targetable tumour driver mutations thought to explain biology in a large proportion (> 60%) of lung tumours (Table 1.2) [48]. The necessity to assess where, and by what mechanisms, components of a targeted pathway are disrupted is highly critical to rational therapy design, evaluation and application.

Presently, the field of cancer 'omics research is tasked with distinguishing key genes and pathways driving tumour biology and drug response from a bewildering background of genomic variability. Genomic alterations driving cancer phenotypes, as opposed to those that are reactive or incidental to causative changes, theoretically make ideal candidates for therapeutic intervention. Identifying these causal alterations from backgrounds of immense genomic complexity is challenging but a critical step in the successful translation of 'omics findings, specifically for: design of therapeutics aimed at specific cancer phenotypes, predicting patient response to traditional modalities, and expanding the pool of patients likely to benefit from existing treatments [77, 102, 110].

A promising strategy to identify genes important to clinical phenotypes, is the collective interrogation of multiple 'omics dimensions on appropriate and well annotated tumour specimens, and the interpretation of these molecular data in a biologically meaningful context [111-116]. Availability of multi-omics tumour cohorts with rich clinical annotations is becoming increasingly common in the public domain. However, for many practical and logistic factors, there exists a scarcity of non-tumour (i.e. histologically "normal") tissues

from cancer patients associated with these cohorts. These non-cancerous tissues are a particularly useful reference group for identifying tumour-specific alterations for 'omics levels that are tissue specific (e.g., DNA methylation, miRNA and mRNA expression). The inclusion of patient-matched non-tumour samples is also useful for addressing confounding inter-individual variance associated with all 'omics levels that is not related to disease phenotypes. However, depending on the objective and analytical approach, one caveat of using non-malignant tissues as control or reference relates to the masking of "field effect" or "field cancerization" changes which refer to a "molecular field of injury" in non-malignant tissues that are involved in initiation of frank disease, first described by Slaughter et al. in oral lesions [117] and expanded upon by many others since, including Spira *et al.* in relation to COPD and lung cancer susceptibility [118-120].

1.7 Overarching hypotheses of thesis work

The overarching hypothesis of this thesis is that specific risk factors, such as smoking and chronic inflammation lead to selective DNA level disruption of genes in exposed tissues and that these selectively disrupted genes contribute directly to COPD and lung cancer pathogenesis.

1.8 Thesis objectives and specific hypotheses

Primary objective: develop and apply an integrative and multi-omic approach to the identification of genes and pathways that contribute to the biology of both COPD and lung cancer, and assess the clinical utility of these findings in the context of patient survival and detection in surrogate tissues.

Specific hypotheses:

(1) Selection occurs at the DNA level and tumour and disease systems can be altered at multiple genetic and epigenetic levels; therefore loci sustaining high-level concerted genetic, epigenetic and/or transcriptional disruptions in tissues involved in disease pathology are indicative of strong selection.

(2) Integration of multi-omics information from patient tissues involved in disease is informative to uncovering mechanisms driving disease biology and clinical phenotypes.

(3) Assessment of epigenetic changes (DNA methylation, miRNA) in non-malignant tissues (airway epithelial cells, lung parenchyma) of diseased and non-diseased subjects may indicate biology associated with disease and inform feasibility of using these tissues as surrogate markers for disease phenotypes.

(4) Chronic inflammation in the lung functions as a pro-tumourigenic pressure leading to selective disruption of distinct genes and pathways in lung tumours from COPD patients.

1.9 Specific aims and thesis outline

Aim 1: Evaluate the potential impact of aberrant DNA methylation on the biology of COPD (Chapter 2). DNA methylation is a heritable, tissue-specific, and reversible gene regulatory mark involved in mediating cellular response to environmental stimuli including cigarette smoke and inflammation, and directly involved in development and progression of a myriad of diverse diseases. Small airways are the primary sites of airflow obstruction in COPD; however a genome-wide characterization of DNA methylation events in these highly relevant tissues had not been previously assessed. *We sought to identify genes and pathways affected at level of DNA methylation in small airways by applying an integrated, genome-wide DNA and RNA level characterization to small airway epithelia collected from individuals with and without COPD.*

Aim 2: Investigate the effect of smoking on microRNA expression in lung tumour and adjacent non-tumour tissues (Chapter 3). Cigarette smoke is the greatest risk factor and number one cause of lung cancer, however half of all newly diagnosed lung cancer patients are former smokers, and up to 25% have never smoked. Distinct smoking and non-smoking environments are associated with disparate and clinically relevant molecular, epidemiological and clinical features. microRNAs mediate biological responses to smoking, have extensive

functions in tumourigenesis, and clinical implications as drug targets and diagnostic markers. Patterns of miRNA disruption in lung tumours and non-tumour lung tissues at genome-wide sequence based level in the context of smoking have not been previously performed. *We sought to apply such an approach to the interrogation of lung tumour and patient matched non-malignant lung tissues to investigate the effect of smoking on miRNA expression and prognosis in the context of smoking-status.*

Aim 3: Identify genes and pathways altered in lung adenocarcinoma tumours from patients with and without COPD and assess presence of alterations in non-malignant airway tissues of COPD patients with lung cancer (Chapter 4). Chronic inflammation is associated with increased cancer risk in many organs, including the lung. Chronic inflammation in the lungs may function as a pro-tumourigenic pressure leading to selective disruption of distinct genes and pathways in COPD-related lung tumours. There exists substantial epidemiological and genetic evidence to posit a mechanistic link between COPD and lung cancer. Genes involved in COPD-tumour biology, altered in non-malignant airway tissues may serve as useful epigenetic-based biomarkers informing screening or treatment strategies. *We sought to identify genes important to lung tumour biology in COPD patients using a novel integrative bioinformatics approach based on biological principles governing gene selection in tumour systems. To evaluate the potential clinical application of our findings, we also sought to assess whether COPD-related tumour genes were also disrupted at the level of DNA methylation in small airways from patients with COPD and non small cell lung cancer (NSCLC).*

2 Chapter: Evaluation of DNA methylation and gene expression patterns in COPD small airways

2.1 Introduction

Small airways are critically important sites for the maintenance of normal lung function. This is exemplified in COPD patients, where chronic inflammation leads to extensive small airway remodeling or thickening of the airway walls which occurs at the expense of the lumen, resulting in airflow obstruction. Airflow obstruction in COPD patients is also associated with a reduction in the numbers of small airways and airways per generation of branching. Increased wall thickening and loss of small airways per generation are directly correlated with reduced lung function and increasing COPD severity [21]. Importantly, small airway remodeling is thought to precede emphysematous destruction of lung parenchyma [21]. A better understanding of the molecular mechanisms underlying small airway remodeling is thus relevant to the design of therapeutic and prevention regimes targeting COPD pathogenesis.

To date, efforts to molecularly characterize COPD or identify biomarkers of COPD-predisposition and progression, have largely focused on genome-wide association studies (GWAS) [30, 32, 121, 122], transcriptome profiling of parenchymal lung tissues [37-40, 45] or large and small airways [41-44] and recently studies which integrate genotype (GWAS) and gene expression data for the purpose of identifying expression quantitative trait loci (eQTLs) [123-126]. GWAS have revealed a significant inter-individual genetic variance underlying lung function and response to major COPD risk factors. Transcriptome studies support the hypothesis that impaired protective mechanisms in response to reactive oxidative species (ROS) from cigarette smoke and inflammatory cells result in damage to small airway epithelia (SAE) and promote inflammation in COPD airways that persist years after smoking cessation. eQTL studies are beginning to unravel the functional impact of many disease-related genetic variants in COPD biology, lung function and smoking-response.

Epigenetic mechanisms, including histone modification, DNA methylation and non-coding RNA mediate cellular responses to systemic and environmental stimuli including

those important to COPD, such as inflammation and smoking [127]. Epigenetic changes, which yield somatically heritable changes in gene expression patterns, are important mediators of environmental exposures related to chronic disease. DNA methylation is a heritable, tissue-specific, and reversible gene regulatory mark that is highly modified in response to cigarette smoke and involved in the development and progression of a wide spectrum of diseases (recently reviewed [91]). DNA methylation patterns are established and maintained by DNA methyltransferase enzymes (DNMTs), which methylate the 5' position carbon in the pyrimidine ring of cytosines at cytosine guanine (CpG) dinucleotides. Methylation of CpG dinucleotides often occurs in ~200 bp stretches of CpGs known as CpG islands. Hypermethylation of CpG islands, especially those proximal to gene promoters, are associated with gene silencing whereas hypomethylation of normally methylated promoters is associated with gene activation in some cancers. Globally, methylation functions to silence repetitive elements, a pattern that is often reversed and associated with genomic instability in disease and aging.

While DNA methylation has been explored in the context of COPD in clinically relevant tissues such as sputum and blood [123, 128, 129], like other epigenetic marks, DNA methylation is highly tissue specific therefore exploration of tissues involved in COPD pathology may yield insight into the molecular mechanisms underlying disease pathology. Since DNA methylation is a reversible gene regulatory modification, the exploration of epigenetic drugs to treat inflammatory and malignant disease is an enormous field of study [127, 130].

Whole genome assessment of these marks in small airways of COPD patients has not been previously performed. Given i) the importance of small airways to COPD pathology, ii) the knowledge that epigenetic mechanisms mediate cellular responses to systemic and environmental stimuli such as inflammation and smoking [127] and iii) the highly tissue specific nature of DNA methylation patterns, we hypothesized assessment of these marks in small airway epithelia from individuals with COPD would provide insight into DNA level disruptions associated with small airway remodeling. In this study, we use an integrative multi-omics approach on patient-matched small airway DNA and RNA to evaluate the

potential impact of aberrant DNA methylation on the biology of COPD. Since disruption of epigenetic events may underlie disease-specific gene-expression changes, characterization of DNA methylation is a critical first step towards the development of epigenetic markers and novel epigenetic therapeutic interventions for COPD.

2.2 Methods

2.2.1 Description of cohort and clinical samples

COPD subjects were defined by post-bronchodilator FEV_1 lung function tests as per the 2011 Global Initiative for Chronic Obstructive Lung Disease (GOLD) (Table 3.1) [5]. Small airway epithelia (SAE) were collected as described in Section 2.2.2, from former smokers (FS) with (n = 15) and without (n = 23) COPD (Table 2.1). FS are defined as one who has stopped smoking for ≥ 1 year. Two-tailed Student's t-tests found no significant difference in age, pack years or years since quitting smoking between COPD and non-COPD groups. All COPD subjects were GOLD stage II (n = 9) or III (n = 6) (as per Table 1.1).

Table 2.1 Summary demographics and clinical information for COPD small airway study

	COPD	Normal	p value
n =	15	23	
Age	65±5.76	64±4.8	0.44
Female:Male	5:10	8:15	1
Pack Years	54.77±30.43	46.64±20.53	0.37
Years Quit	10±9.55	14±5.44	0.2
FEV_1act	1.79±0.63	3.06±0.68	5.59E-06
FEV_1%Pred	58±15.59	98±9.84	2.10E-08
FEV_1/ FVC%	58±9.57	75±5.38	4.92E-06

2.2.2 Collection of small airway epithelia

Bronchial small airway epithelial (SAE) specimens were obtained during routine auto-fluorescent bronchoscopy and under local anesthesia and conscious sedation by Dr. Stephen Lam at the British Columbia Cancer Agency. A 1.5-mm Teflon bronchial brush with a sheath is inserted into a peripheral airway with the same luminal diameter as the outer

diameter of the bronchial brush and the brush is gently pushed out from the sheath to collect the bronchial cells as previously described [131, 132]. RNA brushes that were taken from the same patients, on the same day and from the same lobe as DNA brushes, were considered "patient matched". Representative light microscopy images of cells collected using this approach from one patient is shown at 20X and 40X magnifications (Figure 2.1).

Figure 2.1

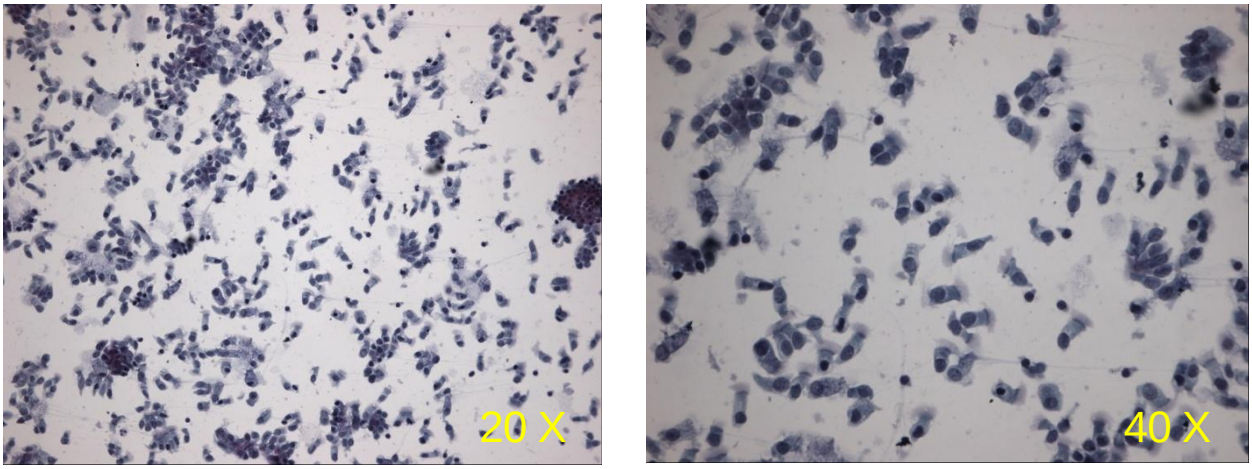

Figure 2.1 Light microscopy images of cells collected with brush from small airways during bronchoscopy

Cytological composition of collected cells described previously reveals these cells are primarily (over 95%) bronchial epithelial cells, with the remainder leukocytes and alveolar macrophages.

2.2.3 Preparation of bronchial epithelial cells for processing

Brushes were removed from -80°C and thawed on ice. Vials were spun at 13000 rpm at 4°C for 10 min. Supernatant was gently removed, followed by a pulse spin and removal of residual supernatant. Brushes and walls of vials were washed twice with 1ml ice cold PBS (made with DEPC water), and spun at 13000 rpm at 4°C for 10 min. Supernatant is removed as described above. Vials were then stored at -80°C.

2.2.4 Extraction of bronchial epithelia DNA from brushings

Brushes stored and frozen in CytoLyt (Holologic Inc., Bedford, MA) were thawed on ice. Vials containing brushes were spun at full speed for 10 min and all supernatant removed. Cells were then lysed in a total of 500 µl of lysis buffer (10 mM Tris pH 8, 100 mM EDTA pH 8, 0.5% SDS, 50 mM NaCl) containing 5 µl of 10 mg/ml Proteinase K (Life Technologies, Carlsbad, CA), at 55°C for 6 hours. Every 2 hours, samples were spiked with 5 µl of 10 mg/ml Proteinase K, and frequently and gently mixed. After cell lysis, brushes were discarded and samples were extracted once with phenol/chloroform extraction to remove Proteinase K. Supernatants from this extraction were then spiked with RNase A for a final concentration of 100 µg/ml, and placed at 37˚C for 1 h. RNase treatment was followed by standard buffered phenol/ chloroform DNA extractions and alcohol precipitation. Briefly, two phenol/chloroform (500 µl phenol/200 µl of chloroform) extractions were performed, followed by one 500 µl chloroform extraction. Each extraction was mixed for 6 min on a rocker, and spun at 13000 rpm for 5 min. DNA was precipitated by 1/10 volume 3 M sodium acetate and equal volume 100% isopronanol. Samples were gently mixed and stored at -20°C for 30 minutes or overnight. Samples were then spun at 13000 rpm at 4°C for 10 min and all supernatant discarded. Pellets were washed with 1 ml ice cold 70% ethanol, spun at 13000 rpm at 4°C for 10 min. All supernatant was removed and pellets were allowed to air dry at room temperature. Pellets were then resuspended in 30 µl sterile water at 37°C overnight. DNA was then quantified using a ND-1000 Spectrophotometer V3.1.0 (NanoDrop Technologies Inc., Wilmington, DE) and stored at -20°C.

2.2.5 DNA methylation profiling

DNA methylation profiles were obtained using the Illumina Infinium Methylation (HM27) chip (Illumina, San Diego, CA) which assesses 27,578 CpG sites of 14,475 genes (GSE55454). DNA samples were bisulfite converted using the Zymo EZ DNA Methylation bisulfite conversion kit (Zymo Research Corporation, Orange, CA) and processed as previously described (7). Methylation data is given as either β-values= Max (methy,0)/[Max (methy,0)+ Max (unmethy, 0) +100], or M-values= log2 ([Max (methy,0) +1] / ([Max (unmethy,0) +1]). Illumina probes were filtered by detection p value ($p > 0.05$) for quality.

Probes that mapped to sex chromosomes, within five base pairs (bp) of known single nucleotide polymorphisms (SNPs), or that contained repeat sequences ≥ 10 bp were also removed, as per the Cancer Genome Atlas Network protocol [133]. Probes were also retained if the number of informative measures (those with detection $p < 0.05$) were available for at least 50% of cases in each group. This left 21,945 probes (12,755 unique genes) for comparative methylation analyses between COPD and non-COPD small airways.

Quantification of percent cytosine methylation for select genes was performed by pyrosequencing on a subset of samples for which adequate material was available from Table 2.1 and on select differentially methylated (DM) genes of interest Table 2.2 for which pyrosequencing probe design was feasible [134]. Each 25 µl PCR contained 1x PCR Buffer (Qiagen Inc.), 0.2 mM dNTPs, 0.025 U Hot Start Taq DNA polymerase (Qiagen Inc.), 0.25 mM forward primer, 0.25 mM reverse primer, and approximately 25 ng bisulfite-converted DNA. Each PCR was performed under the same cycling conditions except for the annealing temperature. Cycling conditions were: 95°C for 15 min, 50 cycles of 94°C for 30 seconds, variable annealing temperature for 30 seconds, 72°C for 60 seconds, followed by 72°C for 10 min. Template preparation and pyrosequencing was performed according to Tost and Gut [135].

Table 2.2 Pyrosequencing validation of Illumina HM27K probes

Gene	**CYP4F11**	**EPHX1**	**IL17RC**	**PTEN**
STATUS	HYPER	HYPER	HYPER	HYPER
HM27K Probe	cg03190825	cg24928687	cg07705835	cg21480743
FC (COPD/Norm)	1.32	1.37	1.29	1.26
Total # CpGs examined	1	2	1	1
% Meth Average HM27 COPD	37	55	28	11
% Meth Average HM27 Normal	7	21	8	4
% Meth Delta (COPD-Norm)	30	34	20	7
% Meth Average Pyro COPD	26	57	9	7
% Meth Average Pyro Normal	9	25	3	5
% Meth PYRO (COPD-Norm)	17	32	6	2
% Meth HM27K- PYRO	13	2	13	5
Forward Primer (5' to 3')	GTTATTTTGAGTTGGTTTTTTTGT	TGGTTATTTTTTTTGGATTTTGTA	Biotin:AGGTTTGTGGGGTTTTAGGA	GGGGTTGTAAATAGATTTGATAGG
Reverse Primer (5' to 3')	Biotin:TGGTTGTTTAGGTTTGGAAGATAT	Biotin:TTAAAAGAAGGGAATTTGGGATAA	AAAGGTTTAGGGTTTAGTTTTTGG	Biotin:TGGTTGAGTTTATAGTAGGTGGG
Sequencing primer (5' to 3')	TGGTTTTTTTGTATTTAGTT	TTTTTGGATTTTGTATAGTA	GAAAGGTTTAGGGTTTAGTT	GATAGGTTTGTTTTTGGG
Annealing temperature (°C)	60	50	50.7	50.7

FC (COPD/Norm): methylation fold change observed in COPD compared to non-COPD small airways; %Meth Average HM27K COPD: average β value COPD airways; %Meth Average HM27K Normal: average β value non-COPD airways; %Meth Delta: average β value COPD airways minus average β value non-COPD airways; % Meth Average Pyro COPD: pyrosequencing percent methylation in COPD airways; % Meth Average Pyro non-COPD: pyrosequencing percent methylation in non-COPD airways; % Meth PYRO (COPD-Norm): pryosenquencing percent methylation from COPD and non-COPD substracted; %Meth HM27K – PYRO; average difference between methylation levels obtained by HM27K and pyrosequencing.

2.2.6 Gene expression profiling

Gene expression profiles for 22 patient matched samples were generated using Affymetrix Human Gene 1.0 ST arrays (GSE56341) (Affymetrix, Santa Clara, CA). Probeset summarization and generation of numerical gene expression levels were performed using the makecdfenv package and RMA normalization using the HuGene-1_0-st-v1.CDF file in *R: A language and environment for statistical computing*. Genes that overlapped between both arrays (after Infinium probe filtering described in Section 2.2.5) resulted in the inclusion of

11,761 unique genes for integrative analyses, gene expression permutation and Spearman correlation tests.

2.2.7 DNA methylation analysis

Sequence dependent color bias correction and SSN normalization algorithms designed for Illumina Infinium HM27 methylation platform were applied [136]. Since commonly used β-values are heteroscedastic, M-values were used for all statistical tests where equal variance is assumed [136, 137]. β-values were used for dimensional reduction by unsupervised principal component analysis (PCA), as recommended [137]. The Illumina Infinium assay was highly reproducible, although less methylated probes were more variable (Figure 2.2 and 2.3). A multivariate ANOVA was used to assess variance in methylation due to disease, age, gender, pack years and years quit. To identify differentially methylated (DM) genes in COPD small airways, we applied a non parametric permutation test, using 10,000 permutations and corrected for multiple testing using the Benjamini and Hochberg (B-H) method (B-H $p < 0.05$ was considered significant). This test is highly powerful for small sample sizes. We further applied standard deviation (SD) ≤ 2, and average fold change (FC) cutoffs of > 1.25 or < 0.75 for probes to be considered differentially hyper or hypomethylated in COPD airways, respectively. A PCA was performed in MatLab. Genes DM between top and bottom pack-year tertiles of our cohort, regardless of disease status, were deemed "smoking-related".

Figure 2.2

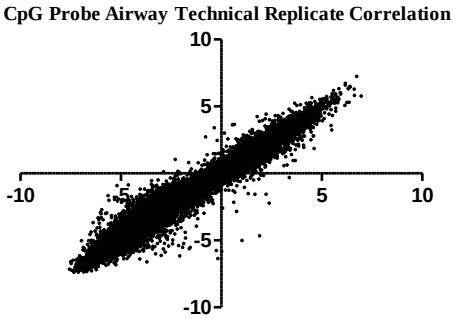

Figure 2.2 Technical reproducibility of Infinium assay using small airway epithelia

DNA methylation profiles from bronchial epithelial cells using the Infinium platform are technically reproducible. Technical replicates from two brushes from the same site in one COPD patient, involving independent bisulfite conversion, array hybridization, color correction and normalization were found to be highly correlative, rho= 0.9740 (95% confidence interval 0.9734 to 9746), $p < 0.0001$, as determined by a non-parametric Spearman test. M-values which = $\log_2$ [(Max (methy,0) +1) / (Max (unmethy,0) +1)] are plotted. Unmethylated probes (negative M values) are more variable than methylated probes (positive M values).

2.2.8 DNA methylation and expression integration

Non-parametric Spearman tests were applied to identify genes likely regulated epigenetically (Spearman's $\rho <= -0.4$ and $p < 0.05$) using patient matched methylation and gene expression profiles. A gene was considered significantly negatively correlated if at least one Illumina and corresponding Affymetrix probe on either array passed the criteria stated. DM genes whose expression levels in COPD airways had: i) a permutation test p value < 0.05 and ii) an average fold change (FC) of > 1.2 or < 0.8 compared to non-COPD profiles, were considered differentially expressed (DE). Here, we focused on genes that sustained concomitant inverse methylation and expression alterations (DM and DE). Recent expression studies report subtle differences (i.e. small effect size) induced by cigarette-smoke in non-malignant tissues thus we employed the same FC criteria to ensure we did not overlook subtle changes [138, 139].

2.2.9 Pathway enrichment analysis

DM and inversely DE genes were selected for IPA (Ingenuity Pathway Analysis®, www.ingenuity.com), which uses a Fisher's exact test to calculate p values corresponding to the probability that enrichment of a canonical pathway is due to chance alone.

2.3 Results

2.3.1 Aberrant DNA methylation patterns affect hundreds of genes in COPD small airways

We hypothesized that patterns of DNA methylation in COPD small airways would be distinct from subjects with normal lung function and similar smoking history. We first evaluated the extent to which DNA methylation was differentially altered in small airways epithelia (SAE) between patients with COPD compared to controls. We detected 1120 unique genes (1260 CpG probes) as DM in COPD SAE, of which 97% were hypermethylated (see Digital Content Supplementary Table 2, available online [140]). Increased variance in lowly methylated probes in combination of our SD cut off threshold may also have contributed to the increased proportion of hypermethylated probes (Figures 2.2 and 2.3). A subset of these genes was validated by pyrosequencing analysis (Figure 2.4). Of the 1120 DM genes, 79 were previously associated with COPD in gene expression studies or GWAS (see Digital Content Supplementary Table 3, available online [140])). These included, for example, hypermethylation of three glutathione S-transferase genes (*GSTP1*, *GSTM1* and *GSTT1*), three cholinergic receptors (*CHRNB1*, *CHRNB2* and *CHRND*), as well as *GPR126*, *HTR4* and *EPHX1*. Hypomethylated COPD associated genes included KSR1, whose over-expression is indicative of increased bacterial colonization frequently associated with COPD phenotypes in humans and in mice [141].

Figure 2.3

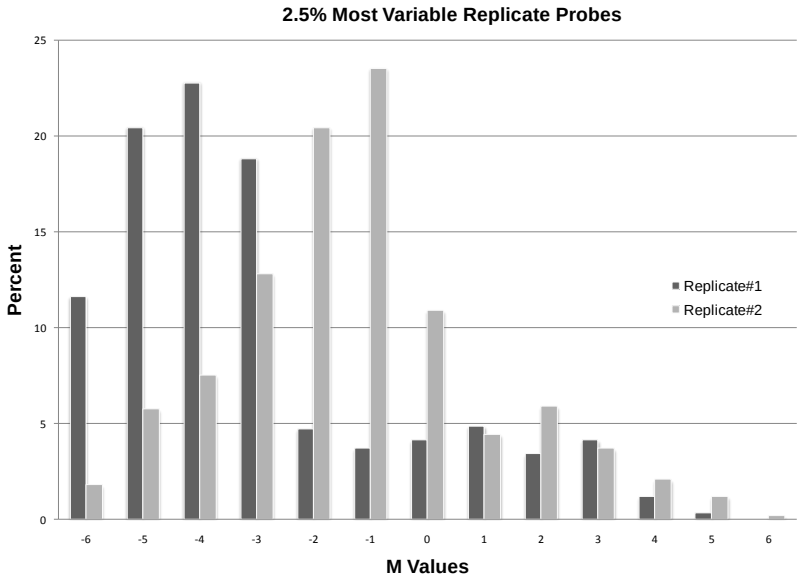

Figure 2.1 Methylation variance between technical replicates and COPD and normal groups

Variance between technical replicates was calculated for each probe. The top 2.5% most variable probes (681 probes), whose variance ranged from 0.93-21.52, are plotted. Most variance occurred for hypomethylated probes (M value ≤ -2), which accounted for >80% of the 2.5% most variable probes. Since approximately 90% of promoters associated with CpG islands are normally unmethylated, and the 27,253 probes we assessed reside primarily in promoters, we expect that many of the differentially methylated (DM) events identified in COPD small airways are aberrantly methylated compared to non-COPD former smoker airways. Likewise, as CpG island-promoters are normally unmethylated, the number of aberrant hypomethylated events is substantially smaller than for hypermethylated events. Increased technical variability of less methylated probes in conjunction with our SD criteria, is also another likely factor in the prevalence of hypermethylated DM probes detected in the present study. For the COPD cases, 70% of DM probes had a standard deviation from the mean (coefficient of variation, COV) of <50%, and for the normal group, 90% of probes had a COV <50%.

Figure 2.4

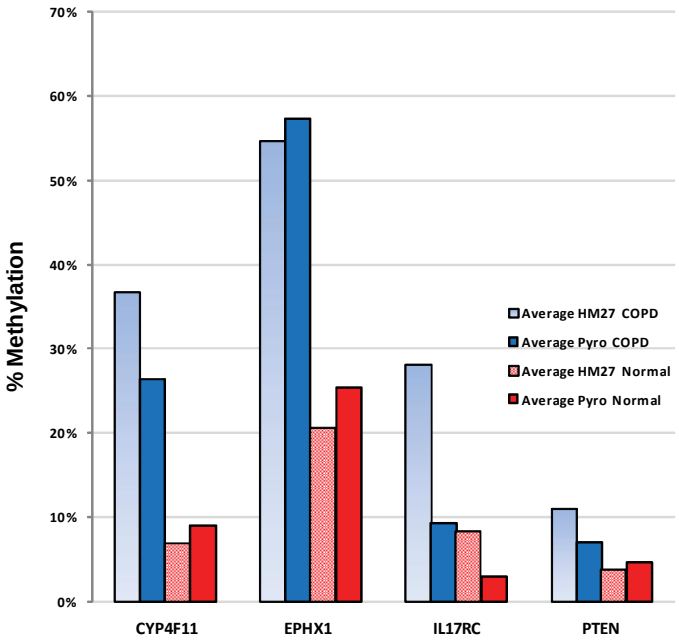

Figure 2.2 Methylation variance between technical replicates and COPD and non-COPD groups

Methylation profiles generated using the Illumina HM27K arraywere validated for a subset of genes by pyrosequencing. M values were converted to Beta values by (2^Mvalue)/ (2^Mvalue + 1). Average % methylation values were calculated for COPD and non-COPD groups (i.e. "Normal") for both HM27K and pyrosequencing results. Pyrosequencing results (dark blue and dark red bars) paralleled those of HM27K (light blue and light red). Average differences between the two assays ranged from 2%-13%.

2.3.2 DNA methylation is correlated with lung function variables

We were next interested in assessing whether methylation may be associated with lung function variables, as opposed to disease status. While we did not detect any significant methylation differences between 9 GOLD II (moderate) and 6 GOLD III (severe) COPD patients, a PCA using only 100 of the most DM genes between 6 severe and 6 control subjects, separated 15 COPD and 23 non-COPD subjects in COPD and non-COPD overall (Figure 2.5). When we considered methylation and lung function as continuous variables, we found methylation levels of 62 genes were significantly (B-H corrected p value < 0.05) correlated with lung function overall, 48% of which overlapped with our 1120 DM COPD genes (from Digital Content Supplementary Table 4 [140]). All significantly correlated lung

function probes were negatively correlated with methylation (i.e. higher methylation was associated with lower lung function).

Figure 2.5

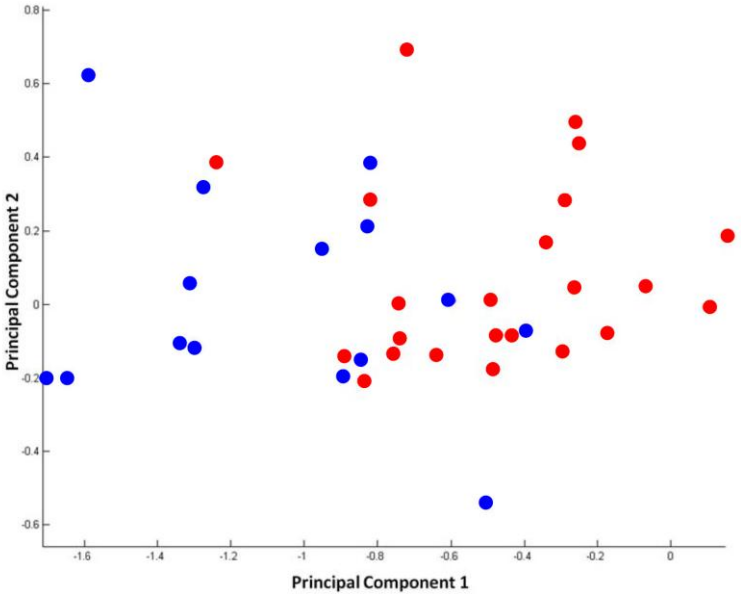

Figure 2.3 Principal component analysis

The 100 most DM probes between GOLD Stage III and a subset of non-COPD cases were applied to clustering and dimension reduction algorithms of 38 COPD (blue dots) and non-COPD (red dots) methylation profiles (β-values). As expected, the top 100 DM genes from severe COPD subjects compared to a subset of normal subjects were capable of clustering COPD and normal methylation profiles using the entire cohort.

2.3.3 COPD related DNA methylation alterations possibly induced by smoking

While two-tailed Student's t-tests found a significant difference in pack-years between our high (n= 11 cases) and low (n= 10 cases) pack-year groups (p= 0.000126); FEV% predicted (p= 0.042852), years quit (p= 0.01543) and age (p= 0.017821) were also significant, therefore we also required "smoking-related" genes to be significantly associated with pack-years ($p< 0.05$) by a MANOVA test. We detected 158 unique genes that passed our criteria for "smoking-related" (i.e. DM between a subset of high and low pack-year patients in our cohort). Of these, 45 overlapped with our 1120 DM COPD, 11 of which were also significantly associated with pack-years and disease status ($p < 0.05$) by a MANOVA in

the high and low pack-year sub-groups. These are: *BAI2*, *C10orf35*, *CD248*, *CDKN2B*, *CHRNB1*, *LIPC*, *PTK9*, *SOX17*, *SUV420H2*, *TREM2* and *ZNF323*, all of which were DM in the same direction in COPD and high pack-year groups; 9 genes were hypermethylated and two (*TREM2* and *ZNF323*) were hypomethylated (see Digital Content Supplementary Table 5, available online [140]).

We also compared our findings to a recent study by Buro-Auriemma et al. who describe the effects of active smoking on the SAE methylome in current (CS) and never (NS) smoker subjects without COPD [142]. Of the top 50 hypomethylated and hypermethylated smoking associated DM genes discovered by Buro-Auriemma et al., none of our smoking-associated genes overlapped, but five of our DM COPD genes did, including: ALDH1A3 and SH3TC2 which were hypermethylated in CS and COPD FS SAE (except one of the two DM ALDH1A3 probes which was hypomethylated in COPD), and CYP1A1, GSTM1 and KCNJ15 which were hypomethylated in smokers, but hypermethylated in COPD SAE.

2.3.4 Pathways affected by DNA methylation in COPD small airways

We next examined what molecular pathways were associated with DM genes from COPD airways. Overall, three pathways were significantly enriched in the 1120 DM gene set (Benjamini-Hochberg (B-H) corrected p value < 0.05), these included: G protein coupled receptor signaling (31 genes), Aryl hydrocarbon receptor signaling (20 genes) and cAMP-mediated signaling (26 genes) (Figure 2.6 and Digital Content Supplementary Table 6, available online [140]). These pathways are known to play a role in small airway biology including COPD small airway remodeling, wound healing and in mediating cellular response to polycyclic aromatic hydrocarbon (a component of cigarette smoke) exposure [143-146].

Figure 2.6

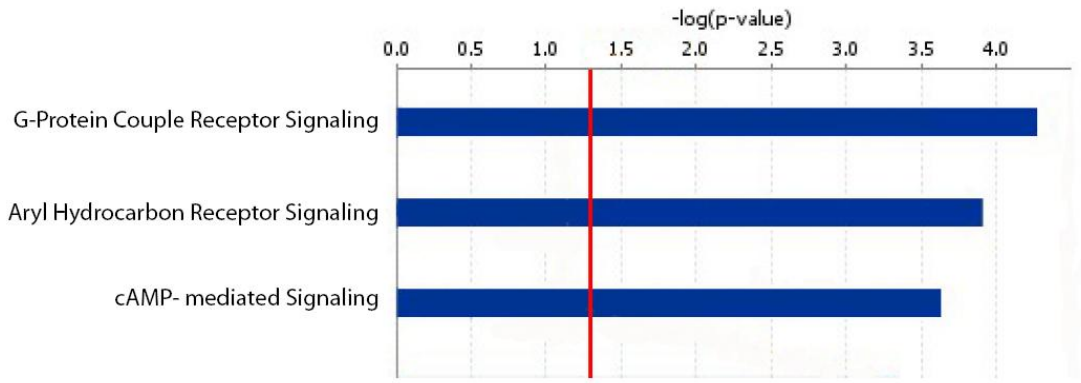

Figure 2.4 Differentially methylated genes in COPD small airways correspond to three significantly enriched pathways

We detected 1120 differentially methylated genes in small airways of COPD patients compared to methylation profiles from individuals without COPD. These genes corresponded to three significantly enriched pathways: G protein coupled receptor signaling (31/272 genes; B-H p = 0.024), Aryl Hydrocarbon Receptor Signaling (20/161 genes; B-H p = 0.0276) and cAMP-mediated signaling (26/224 genes; B-H p= 0.0345). The horizontal axis displays -log of the B-H p value, calculated by Fisher's exact test right-tailed, representing the probability that pathways are enriched in a given gene set by random chance. A B-H p value = 0.05 is indicated by the vertical red line. B-H p: Benjamini-Hochberg corrected p value.

2.3.5 Integration of DNA methylation and gene expression changes to reveal candidate genes and pathways potentially involved in COPD pathogenesis

Epigenetic regulation of gene expression by DNA methylation is dependent upon location and CpG content of regulatory elements. For example, hypermethylation of gene promoter elements with high CpG content is associated with repression of gene expression, whereas hypermethylation within the first exon of gene bodies is associated with activation of gene expression. Since Illumina HM27 CpG probes reside primarily within promoters, we focused our analysis on genes whose methylation and gene expression values were negatively correlated across samples (where matched DNA methylation and gene expression profiles were available) from COPD and non-COPD subjects. We identified 141 such genes; however we note that methylation levels of 335 genes were positively associated with gene expression, 99% of which were hypermethylated and overexpressed. Of the inversely

correlated genes, 130 were hypermethylated and underexpressed and 11 were hypomethylated and overexpressed relative to non-COPD airways (Figure 2.7 and Digital Content Supplementary Table 7, available online [140]). Fifteen of these 141 genes (11%) have been previously associated with COPD, at either the DNA or mRNA level, but none have been previously associated with differential DNA methylation in COPD (Table 2.3). For example, *TFF3*, the trefoil factor which regulates repair of injured human respiratory epithelium, was hypermethylated and underexpressed in COPD SAE [147]. Similarly, the creatine kinase gene CKB, was hypermethylated and underexpressed in COPD airways. Underexpression of CKB has been previously detected in COPD bronchial epithelial cells in association with smoke-induced bronchial epithelial cell senescence [148].

Figure 2.7

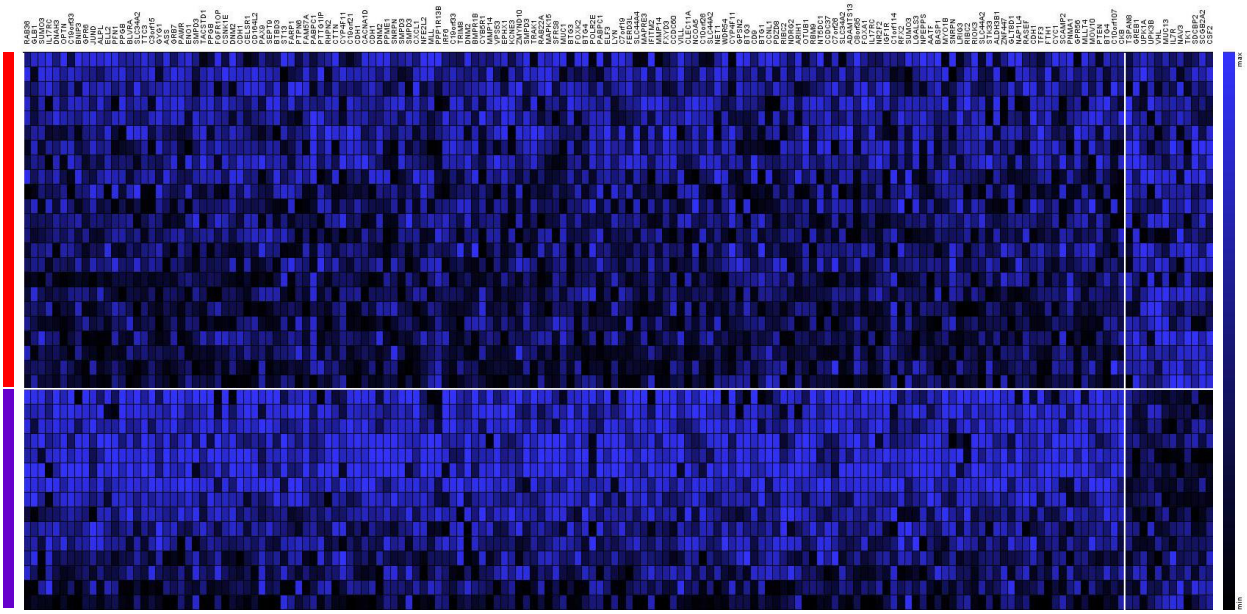

Figure 2.5 Methylation heat map of differentially methylated and inversely differentially expressed genes in COPD airways

141 DM and inversely DE genes in COPD small airways, corresponding to 130 hypermethylated and underexpressed genes and 11 hypomethylated and overexpressed genes are depicted for 38 samples (COPD = 15, purple bar; non-COPD = 23, red bar). M-values are plotted. Positive M-values correspond to more (bright blue) and less (black) methylation. (Genes correspond to Digital Content Supplementary Table 3, available online [140]).

Table 2.3 Differentially methylated and differentially expressed genes in COPD small airways previously associated with COPD

Symbol	Meth	Exp	Meth B-H pval	Meth FC	Exp pval	Exp FC
BNIP3	HYPER	UNDER	7.45E-05	1.76	4.06E-03	0.5
TTC3	HYPER	UNDER	6.73E-07	1.57	3.14E-04	0.35
SMPD3	HYPER	UNDER	1.19E-05	1.42	6.43E-09	0.61
CXCL1	HYPER	UNDER	4.80E-05	1.4	2.87E-06	0.15
EPHX1	HYPER	UNDER	3.10E-05	1.37	1.67E-03	0.21
MUC1	HYPER	UNDER	3.08E-05	1.36	5.62E-11	0.55
NFE2L2	HYPER	UNDER	1.29E-04	1.36	7.61E-07	0.26
MMP14	HYPER	UNDER	3.00E-05	1.35	2.44E-02	0.27
CD9	HYPER	UNDER	2.84E-07	1.31	3.79E-04	0.11
LGALS3	HYPER	UNDER	2.25E-04	1.29	7.64E-04	0.26
TFF3	HYPER	UNDER	9.88E-04	1.26	6.07E-03	0.24
PTEN	HYPER	UNDER	2.38E-05	1.26	2.25E-04	0.34
CKB	HYPER	UNDER	1.04E-07	1.25	4.80E-04	0.55
VHL	HYPO	OVER	4.20E-02	0.7	3.55E-03	2.63
MUC13	HYPO	OVER	3.98E-04	0.71	9.07E-10	1.18

Meth: methylation status of gene in COPD relative to non-COPD airways; Exp: expression status of gene in COPD relative to non-COPD airways; Meth B-H pval: Benjamini-Hochberg corrected permutation test p value of COPD vs non-COPD DNA methylation comparison. Meth FC: methylation fold change of gene in COPD airways over non-COPD airways; Exp pval: permutation test p value of COPD vs non-COPD expression comparison; Exp FC: expression fold change of gene in COPD airways over non-COPD airways.

2.3.6 Validation of gene expression changes in external cohorts

Since this is the first study assessing DNA methylation patterns in small airways of COPD patients, validation of methylation findings in external datasets was not possible. Therefore, to validate our DM and DE gene set, small airway gene expression profiles were downloaded from GSE37147 [41]. After matching for smoking status, this cohort included 39 FS with COPD (n= 32 GOLD II; n= 7 GOLD III) and 63 FS without COPD. We found that 46 out of 141 of our DM and DE genes had similarly altered expression patterns in COPD compared to non-COPD airways in this external dataset (Table 2.4), five of which (EPHX1, IGF1R, LRIG3, MUC13 and SDCBP2) showed statistically significant differential expression between the 39 COPD and 63 non-COPD profiles ($p < 0.05$ by a non-parametric Mann–Whitney U test). Our observation of highly methylated genes exhibiting reduced gene expression levels suggests that aberrant DNA methylation has a concordant effect on gene expression in COPD SAE cells.

Table 2.4 Differentially methylated genes inversely expressed in multiple cohorts

Refseq	Meth	Exp	Meth BH pval	Meth FC	Exp pval	Exp FC
ADAMTS13	HYPER	UNDER	2.45E-04	1.29	2.04E-02	0.81
ALDH3B1	HYPER	UNDER	1.75E-05	1.27	3.52E-03	0.21
ARIH1	HYPER	UNDER	1.36E-05	1.30	4.47E-02	0.59
BLVRA	HYPER	UNDER	1.77E-06	1.60	7.94E-03	0.49
BNIP3	HYPER	UNDER	7.45E-05	1.76	4.06E-03	0.50
*BTG4	HYPER	UNDER	1.09E-06	1.30	4.39E-02	0.72
C10orf26	HYPER	UNDER	2.83E-06	1.33	6.88E-03	0.80
C1orf114	HYPER	UNDER	1.95E-04	1.29	7.36E-03	0.75
C3orf15	HYPER	UNDER	4.57E-08	1.55	3.61E-02	0.20
CCDC60	HYPER	UNDER	6.79E-06	1.33	2.34E-03	0.26
*CDH1	HYPER	UNDER	9.89E-05	1.40	1.03E-03	0.10
CKB	HYPER	UNDER	1.04E-07	1.25	4.80E-04	0.55
CSNK1E	HYPER	UNDER	9.31E-06	1.49	7.69E-03	0.74
*CYP4F11	HYPER	UNDER	3.64E-04	1.37	9.69E-03	0.50
DNAH3	HYPER	UNDER	1.31E-06	1.82	9.60E-04	0.45
†EPHX1	HYPER	UNDER	3.10E-05	1.37	1.67E-03	0.21
FARP1	HYPER	UNDER	3.53E-05	1.45	2.45E-02	0.46
FERD3L	HYPER	UNDER	2.49E-06	1.34	2.70E-02	0.83
FGFR1OP	HYPER	UNDER	2.63E-07	1.50	1.69E-03	0.37
FOXA1	HYPER	UNDER	1.03E-04	1.29	2.19E-03	0.75
FTH1	HYPER	UNDER	2.96E-04	1.26	7.78E-03	0.09
GLT8D1	HYPER	UNDER	1.02E-04	1.27	1.47E-04	0.63
†IGF1R	HYPER	UNDER	2.49E-05	1.29	3.87E-02	0.54
JUND	HYPER	UNDER	6.94E-07	1.72	2.51E-02	0.75
LGALS3	HYPER	UNDER	2.25E-04	1.29	7.64E-04	0.26
†LRIG3	HYPER	UNDER	1.51E-05	1.28	2.90E-02	0.45
MAPRE3	HYPER	UNDER	1.23E-07	1.34	3.94E-02	0.64
MLL	HYPER	UNDER	1.96E-04	1.40	1.17E-02	0.67
MLLT4	HYPER	UNDER	2.88E-05	1.26	1.75E-03	0.16
†MUC13	HYPO	OVER	3.98E-04	0.71	9.07E-10	1.18
NT5DC1	HYPER	UNDER	1.70E-03	1.30	3.17E-02	0.30
*PPGB	HYPER	UNDER	6.35E-07	1.57	1.19E-02	0.75
PPP1R13B	HYPER	UNDER	8.58E-04	1.39	5.33E-03	0.80
PTPN6	HYPER	UNDER	1.65E-04	1.45	4.91E-03	0.72
RFX2	HYPER	UNDER	1.96E-04	1.29	2.44E-03	0.48
†SDCBP2	HYPO	OVER	1.64E-03	0.72	7.27E-05	6.72
SCGB2A2	HYPO	OVER	8.52E-04	0.74	2.57E-04	17.66
SFRS8	HYPER	UNDER	5.23E-07	1.36	1.96E-03	0.52
*SLC44A2	HYPER	UNDER	2.98E-05	1.31	1.01E-02	0.44
*SNRPN	HYPER	UNDER	8.70E-04	1.35	4.37E-02	0.58
ST13	HYPER	UNDER	1.07E-04	1.45	2.59E-02	0.20
STK33	HYPER	UNDER	3.14E-05	1.28	8.92E-03	0.26
*SUMO3	HYPER	UNDER	3.16E-05	1.62	3.46E-02	0.80
VHL	HYPO	OVER	4.20E-02	0.70	3.55E-03	2.63
WDR54	HYPER	UNDER	1.18E-04	1.32	1.09E-02	0.64

Table 2.4 Differentially methylated genes inversely expressed in multiple cohorts

† significantly differentially expressed in FS from GSE37147; *multiple methylation probes map to gene symbol; Meth: methylation status of gene in COPD relative to non-COPD airways; Exp: expression status of gene in COPD relative to non-COPD airways; Meth BH pval: Benjamini-Hochberg corrected permutation test p value of COPD vs non-COPD DNA methylation comparison; Meth FC: methylation fold change of gene in COPD airways over non-COPD airways; Exp pval: permutation test p value of COPD vs non-COPD expression comparison; Exp FC: expression fold change of gene in COPD airways over non-COPD airways.

2.3.7 Methylated genes strongly negatively correlated with gene expression

To identify the aberrantly methylated gene candidates most likely to be controlled by DNA methylation in COPD SAE, we applied a Spearman correlation cut off (ρ < -0.4, p < 0.05) to the 141 DM and DE gene set. The most negatively correlated DM and DE gene was *CYP4F11* (ρ= -0.866742, Spearman p = 0.000001) (Table 2.5). For these genes, the presence of COPD was the only significant factor (p < 0.05) responsible for the observed variance in methylation based on a multivariate ANOVA assessing variance due to COPD, age, gender, pack years and years quit, except for one of the two methylated probes for *CYP4F11* (cg24655310), which was also affected by gender (p= 0.01733). *MUC13* was the gene most significantly associated with hypomethylation in COPD subjects (p= 0.0006442).

Table 2.5 Differentially methylated and expressed genes most likely under epigenetic control in COPD small airways

Symbol	*rho*	p value	Meth	Exp	Meth BH pval	Exp pval	Meth FC	Exp FC	Chr	MapInfo
CYP4F11	-0.87	2.00E-06	HYPER	UNDER	4.75E-04	9.69E-03	1.32	0.50	19	15906788
MUC13	-0.63	1.94E-03	HYPO	OVER	3.98E-04	9.07E-10	0.71	1.18	3	126135480
SNRPN	-0.50	1.88E-02	HYPER	UNDER	1.09E-07	4.37E-02	1.42	0.58	15	22674380
CYP4F11	-0.49	2.20E-02	HYPER	UNDER	2.53E-04	9.69E-03	1.43	0.50	19	15906119
EPHX1	-0.49	2.31E-02	HYPER	UNDER	3.10E-05	1.67E-03	1.37	0.21	1	224079536
BLVRA	-0.48	2.61E-02	HYPER	UNDER	1.77E-06	7.94E-03	1.60	0.49	7	43764312
SDCBP2	-0.45	3.59E-02	HYPO	OVER	1.64E-03	7.27E-05	0.72	6.72	20	1257722
BTG4	-0.45	3.80E-02	HYPER	UNDER	2.04E-06	4.39E-02	1.25	0.72	11	110888125

rho: Spearman correlation coefficient; p value: Spearman test p value; Meth: methylation status of gene in COPD relative to non-COPD airways; Exp: expression status of gene in COPD relative to non-COPD airways; Meth BH pval: Benjamini-Hochberg corrected permutation test p value of COPD vs non-COPD DNA methylation comparison; Exp pval: permutation test p value of COPD vs non-COPD expression comparison; Meth FC: methylation fold change of gene in COPD airways over non-COPD airways; Exp FC: expression fold change of gene in COPD airways over non-COPD airways; Chr: chromosome location of gene; MapInfo: base pair location of CpG assayed.

2.3.8 The Nrf2 signalling pathway is strongly enriched for genes affected by both DNA methylation and mRNA alterations in COPD small airways

We next applied pathway enrichment analysis to the set of 141 DM and DE genes. This 141 gene set was significantly enriched ($p < 0.05$) for three pathways: PTEN signaling, the Nrf2-mediated oxidative stress response pathway and the IL-17F in allergic inflammatory airway diseases (Figure 2.8). Two modulators of the PTEN signaling pathway: *PTEN* and *CSNK2A2* were hypermethylated and underexpressed in COPD airways. Multiple upstream Nrf2 regulators (*PKC*, *MEK1*, *Actin*, *PTEN*, *NRF2*) and downstream effector molecules (*MAF*, *MRP4*, *GST*, *HIP2*, *HSP6*, *EPHX1*, *FTH1*) were found to be differentially affected at the level of DNA methylation and/or gene expression (Figure 2.9). Two of the most strongly negatively correlated genes in our entire study were *EPHX1* and *CYP4F11* (Table 2.5). In the IL-17F pathway, the upstream receptor *IL17RC*, along with downstream *CXCL1* were found to be hypermethylated and underexpressed, while the pro-inflammatory *CSF2*, was hypomethylated and overexpressed.

Figure 2.8

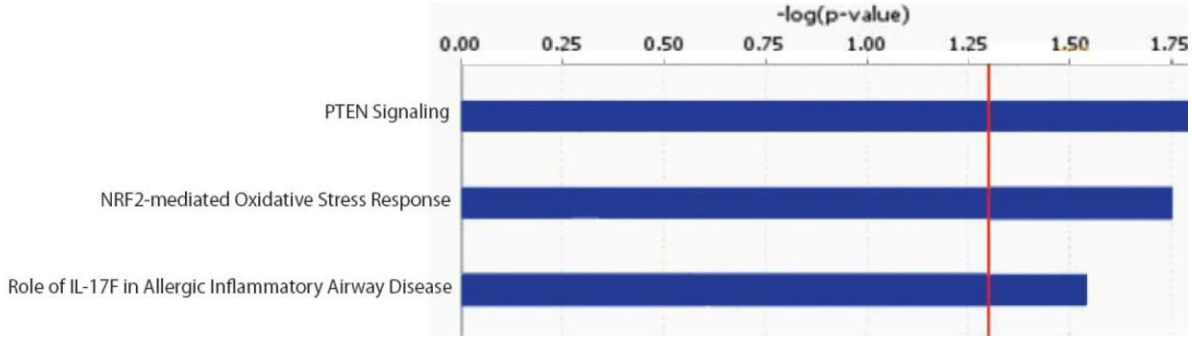

Figure 2.6 Pathways enriched in differentially methylated and differentially expressed COPD airway gene set

Three pathways were significantly ($p < 0.05$) enriched in the 141 DM and DE genes. These included: PTEN signaling ($p= 0.016$), the Nrf2-mediated oxidative stress response pathway ($p= 0.0178$) and the IL-17F in allergic inflammatory airway diseases ($p= 0.0288$). The horizontal axis displays -log of the p-value which was calculated by Fisher's exact test right-tailed, representing the probability that pathways are enriched in a given gene set by random chance. A p value = 0.05 is indicated by the vertical red line.

In our external validation data set, the Nrf2 signaling pathway was the most significantly enriched pathway ($p = 0.00614$) based on the 46 genes that were altered in the same direction and was also the only significantly enriched pathway that overlapped between the DM and the concomitant DM and DE gene sets.

Collectively our integrative analyses indicate that i) expression levels of COPD associated genes are epigenetically deregulated in small airways of patients with COPD and ii) DNA methylation is a likely mechanism through which key pathways of importance to COPD pathology are disrupted.

Figure 2.9

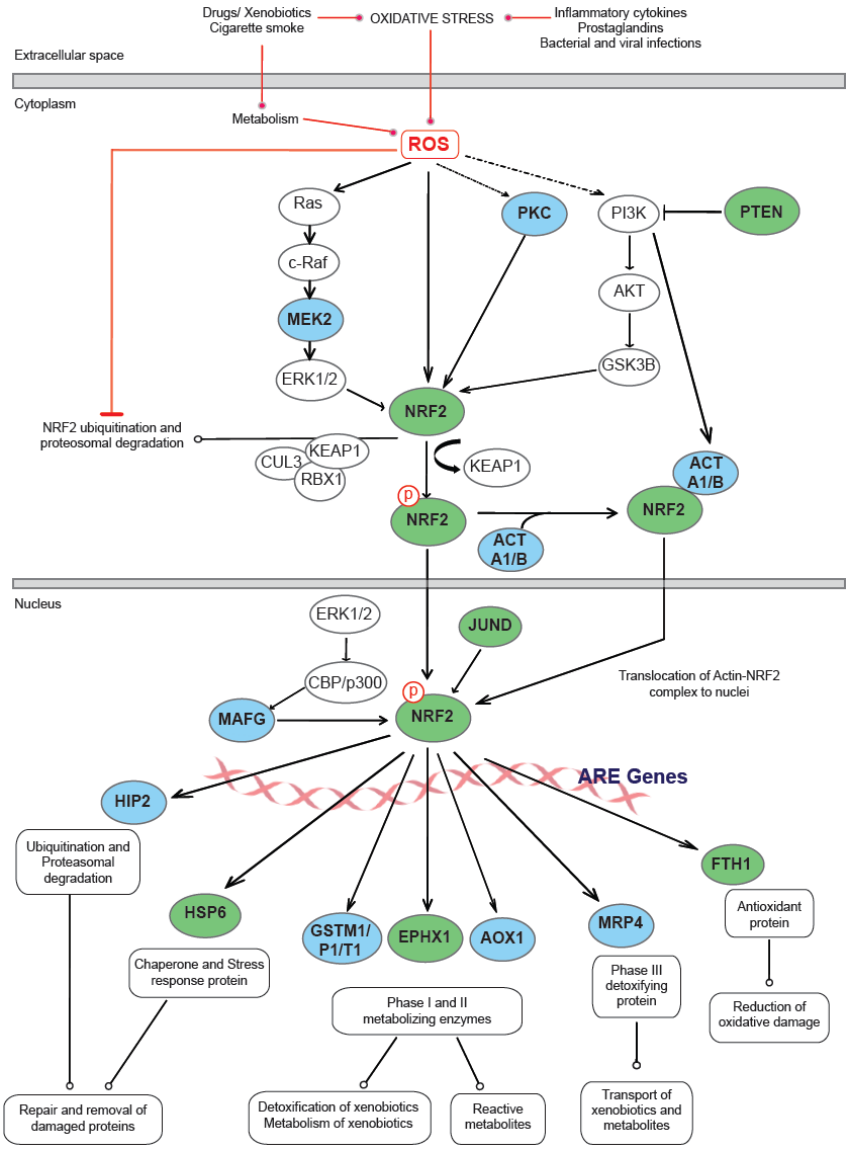

Figure 2.7 The Nrf2-mediated oxidative stress response pathway is altered at multiple levels by DNA methylation in COPD airways

Increased cellular levels of reactive oxidative species (ROS), inhibit KEAP1/CUL3/RBX1 mediated NRF2 ubiquitination and proteosomal degradation, allowing NRF2 nuclear translocation. Antioxidant response element (ARE) genes which are transcriptionally activated by NRF2, mediate processes involved in cellular protection from ROS damage. Genes in the Nrf2 pathway are aberrantly methylated and expressed at multiple points in COPD small airways. An impaired Nrf2 response can result in increased damage from ROS. Hypermethylated genes: light blue; hypermethylated and underexpressed genes: green. ARE: antioxidant response element genes.

2.4 Discussion

DNA methylation is highly modified by inflammation and cigarette smoke in cells of exposed airways and lung tissues, and is directly involved in the development and progression of a wide spectrum of disease. In the context of COPD, DNA methylation has been explored in sputum and blood [123, 128, 129], but not on a genome-wide level in SAE. Since DNA methylation is highly tissue specific, and small airways are the primary sites of airflow obstruction in COPD, assessment of these marks in SAE from patients with COPD is of significant biological and clinical interest. We provide the first genome-wide methylation and integrative 'omics study applied to the analysis of SAE from individuals with COPD. To avoid confounding effects of active cigarette smoking which is known to affect both DNA methylation and gene expression in small airways [142], analyses were restricted to FS.

We found that DNA methylation is widely disrupted in SAE of COPD patients, affecting hundreds of genes which we found predominately hypermethylated relative to SAE of individuals without COPD. Since the majority of gene promoters associated with CpG islands are normally un-methylated [149] and our assay was predominantly restricted to promoter CpGs [150], many of these DM events are likely abnormal. Overall, our DM COPD gene set was enriched for three pathways: G protein coupled receptor signaling, Aryl hydrocarbon receptor signaling and cAMP-mediated signaling. In the context of COPD, deregulation of these pathways has been implicated at the single nucleotide polymorphisms (SNP), mRNA and protein levels [31, 146, 151], but not previously at the level of DNA methylation as described here.

While we did not detect any significant differences between moderate and severe COPD cases, we did detect genes whose methylation status was significantly correlated with lung function overall, all in a negative direction and almost half of which overlapped with our DM COPD genes (from Digital Content Supplementary Table 4 [140]). We found methylation of *GATA4* negatively associated with lung function and DM in COPD; hypermethylation of *GATA4* has been previously associated with lower percent predicted FEV_1 in wood smoke-associated COPD [129]. Methylation levels of genes not detected as DM in COPD but significantly correlated with lung function and of potential interest to

COPD, include *CRABP1* a cellular retinoic acid binding protein [152], and *ITPK1* which has been associated murine tracheal cell model of cystic fibrosis [153].

Given the overwhelming proportion of DM genes that were hypermethylated in COPD SAE, our results contrast those of a study that discovered DNA methylation patterns in blood DNA of large family-based cohorts of COPD patients, were predominantly hypomethylated [123]. However, DNA methylation (and gene expression) patterns are tissue specific, therefore discordance between these studies is expected. COPD is a systemic disease and results from large-scale epigenomic investigations using peripheral blood DNA are indeed of clinical importance given the accessibility of blood and the potential utility of blood based biomarkers.

DNA methylation has been explored in lung and airway cells in the context of other chronic lung and airways diseases, including idiopathic pulmonary fibrosis (IPF) [154] and asthma [155]. Two hypermethylated COPD genes (GRASP and ABCA8) overlapped with the 16 genes discovered by Sanders et al. as DM and DE in IPF lung tissues, although only ABCA8 was in the same direction. Of interest, two of our hypermethylated DM COPD genes (CAV1 and PTEN) have been described elsewhere as IPF suppressor genes [156-158] as well as in COPD [159, 160], highlighting the potential importance of these genes to chronic lung disease, particularly in the context of cigarette smoke [159, 161]. Our COPD SAE results did not have any overlap with six genes described by Stafanowicz et al. as differentially methylated between atopic and asthmatic derived SAE [155] possibly reflecting the distinct biology of these diseases.

In attempting to enrich identification of COPD-specific DNA methylation alterations by only assessing FS, we were not able to assess whether our DM COPD genes may be induced by smoking. Therefore, we compared our results to those of a recent study by Buro-Auriemma *et al.*, who assessed methylation and gene expression differences between SAE from CS and NS without COPD, and found the majority of DM genes hypomethylated in smoker SAE [142]. When we directly compared our results, 5 DM COPD genes overlapped with the most DM in CS SAE, although primarily in opposite directions. Interestingly, while

20% of Buro-Auriemma *et al*'s top hypomethylated genes are involved in aryl hydrocarbon receptor signaling, we found this pathway significantly and almost entirely hypermethylated in COPD SAE (see Digital Content Supplementary Table 6 [140]). Given the importance of these genes and pathways to COPD and smoking response [162], these results could suggest that in individuals without disease, hypomethylation and upregulation of smoking-induced genes and pathways in CS SAE, such as *CYP1A1*, *GSTM1* and genes involved in Aryl hydrocarbon receptor signaling occurs, but abnormally, hypermethylation of these genes may be related to smoking-induced damage associated with COPD.

We attempted to further assess the contribution of smoking to our COPD results, by assessing methylation differences between the individuals with high and low pack years from our cohort, regardless of disease status. Of our "smoking-related" genes, none overlapped with the top smoking-associated methylated genes described by Buro-Auriemma *et al*, but 28% overlapped with our DM COPD genes, including the cholinergic receptor, *CHRNB1*; interestingly SNPs in *CHRNB1* are associated with nicotine dependence and lung cancer [163]. Genes which were not detected as DM in COPD, but detected as smoking-related in our study included *CRYGD*, a member of six gene products required for expression of two important smoking-response genes, *AHR* and *CYP1A1* [164], and *CCL26*, a negative regulator for neutrophils in COPD and whose expression is positively associated with lung function in COPD (but negatively in asthma) [165].

In addition to assessing genome-wide DNA methylation patterns, we further sought to identify genes likely disrupted at the transcriptional level due to aberrant DNA methylation by integrating DNA methylation with gene expression changes using patient-matched DNA and RNA profiles (see Digital Content Supplementary Table 7 [140]). We identified three pathways disrupted at both the DNA methylation and gene expression levels, which we believe are potentially important in COPD pathogenesis, namely: PTEN signaling, the Nrf2-mediated oxidative stress response and the IL-17F inflammatory response pathways.

PTEN is the master inhibitor of the PI3K–AKT–mTOR pathway. Acquired mutations of *PTEN* are evident in airway epithelium of smokers and *PTEN* variants have previously

been associated with COPD [160, 166]. Activation of the PI3K pathway is an important therapeutic target in a wide spectrum of cancers and is increasingly implicated in COPD [167, 168]. In our study, two modulators of this pathway, *PTEN* and *CSNK2A2*, were hypermethylated and underexpressed in COPD airways, suggesting that DNA methylation may be an additional mechanism regulating this pathway in COPD airways.

The IL-17F inflammatory response pathway is also interesting in the context of COPD. Cytokines are important mediators in allergic and non allergic inflammatory airway disease. While overexpression of IL17 is associated with many inflammatory diseases, its role in COPD is ambiguous, likely due to differences in biological tissue or cell type assayed (e.g. serum, lymphocytes, airway epithelia cells) since IL17 and IL17 receptor expression and function varies widely based on cellular context [169, 170]. In COPD SAE, we found the upstream receptor in this pathway, *IL17RC*, along with downstream *CXCL1*, hypermethylated and underexpressed, while the pro-inflammatory *CSF2*, was hypomethylated and overexpressed.

Increased oxidative stress and generation of free radicals, such as that which occurs in response to cigarette smoke exposure, affect nearly all aspects of COPD pathology. The Nrf2 pathway is the major cellular defense system against oxidative stress, mediated through NRF2 nuclear translocation and activation of antioxidant response element (ARE) genes. The Nrf2 pathway is normally up-regulated in airways of healthy smokers, but in smokers with severe COPD, expression of key modulators and downstream ARE genes are underexpressed, resulting in impaired cellular defense mechanisms and increased oxidative damage in airway and lung tissues [42, 44, 171, 172]. At the DNA level, multiple GWAS have identified functional SNPs in promoters of key genes within this pathway including downstream ARE and genes associated with xenobiotic metabolism to be associated with increased COPD and lung cancer risk [173].

Our data strongly suggest the Nrf2 pathway sustains multiple levels of epigenetic disruption. We detected multiple upstream Nrf2 regulators (*PKC*, *MEK1*, *Actin*, *PTEN*, *NRF2*) and downstream effector molecules (*MAF*, *MRP4*, *GST*, *HIP2*, *HSP6*, *EPHX1*, *FTH1*)

differentially affected at the level of DNA methylation alone, or by both DNA methylation and gene expression in COPD small airways (Figure 2.8). Two genes in this pathway, *EPHX1* and *CYP4F11* were amongst the most negatively correlated DM and DE genes overall, strongly suggesting that in COPD airways, reduced expression of *EPHX1* and *CYP4F11* is likely modulated epigenetically by DNA methylation. *EPHX1* functions in the biotransformation of epoxides resulting from degradation of aromatic compounds such as those found in cigarette smoke. Under-expression of *EPHX1* is frequently described in COPD airway and lung tissues, and 'Slow' *EPHX1* SNPs are associated with impaired enzyme activity and increased COPD risk and 'Fast' SNPs potentially confer a protective effect [31]. Cytochrome P450 4F enzymes, such as *CYP4F11*, are involved in cellular protection, xenobiotic metabolism, detoxification, lipid synthesis and metabolic activation of drugs, including those used to treat chronic inflammatory disease [174, 175]. They also have a direct role in inhibiting inflammation through suppression of leukotriene and prostaglandin signals [175]. *CYP4F11* contains both JNK/AP-1 and hormone response element (HRE) binding domains; it is positively regulated by retinoid X receptors (RXR) and JNK (through TNF-α activation), and is negatively regulated by retinoic acid receptors (RARs). However, regulation of CYP4F11 in an environment of chronic inflammation is complex. In human epidermal keratinocytes, while TNF-α leads to immediate activation of *CYP4F11* through JNK, subsequent activation of NFκβ results in direct inhibition of *CYP4F11* [174, 175]. *CYP4F11* regulators are clearly important to COPD biology, although little is known about the function of this enzyme in respiratory tissues. Given the known functions of CYP4F11, it is possible that epigenetic silencing of this enzyme in small airways of COPD patients may lead to impaired cellular protective responses, increased inflammation or altered activation of inhaled steroids.

The study and application of antioxidant inflammation modulators (AIMs) to target the Nrf2 pathway is a growing field of study particularly relevant for COPD therapeutics [46]. Interestingly, overactivation of this pathway through mutation of its key inhibitor, KEAP1 is a frequent event in lung squamous cell carcinoma (SqCC) [176]. Given that SqCC is more frequent in individuals with COPD [177], elucidating the role this pathway plays in

promoting inflammation and tumourigenesis may be critical to the rational application of AIMS therapy to COPD patients.

Our findings suggest that in small airways of COPD patients, aberrant DNA methylation is a genome-wide phenomenon affecting hundreds of genes and several pathways important to smoking response and COPD biology. Since DNA methylation is a reversible gene regulatory modification, further work in this area may contribute to the development of novel treatment strategies or the re-appropriation of existing epigenetic based drugs to the treatment or prevention of COPD.

3 Chapter: Effect of smoking on microRNA expression in lung adenocarcinoma and adjacent non-tumour lung tissues

3.1 Introduction

MicroRNAs (miRNAs) negatively regulate mRNA expression through direct inhibition of translation or induction of mRNA degradation [92]. They are key contributors to smoking response, tumourigenesis, progression and treatment response, and therefore represent promising and biologically relevant biomarkers [97, 98, 178-182]. Cigarette smoke is associated with 75-80% of lung cancer cases. The molecular effects of smoking are widespread, and are associated with specific genetic and epigenetic modifications that alter transcriptional regulation of many lung cancer related genes, including those coding for miRNA [179, 183-185].

We hypothesized that, analogous to distinct smoking-status related patterns of DNA and mRNA alterations, miRNAs display smoking-status specific patterns of disruption in both non-malignant and malignant lung tissues from lung cancer patients. To date, most lung cancer miRNA profiling studies have focused on i) identifying aberrantly expressed miRNA, ii) identifying miRNA with prognostic significance, iii) comparing histological or molecular subtypes and iv) detecting miRNA in blood for use as clinical biomarkers [186-191]. Noticeably absent from the literature is a comprehensive comparison of miRNA deregulation in malignant and non-malignant lung tissues of cancer patients specifically in the context of smoking history. We investigated the effects of smoking on the miRNA transcriptome of lung tumours and parenchymal tissues from current (CS), former (FS) and never (NS) smokers. miRNA-mRNA gene networks were built to determine the potential biological consequences associated with miRNA disruption in these three groups, and we evaluated the potential clinical significance of our findings in relation to patient survival in the context of smoking status.

3.2 Methods

3.2.1 Description of cohort and clinical samples

Fresh-frozen lung adenocarcinoma (LUAC) tumour and patient matched non-malignant lung parenchymal tissue was collected for 94 treatment naïve patients at Vancouver General Hospital under informed, written patient consent and with approval from the University of British Columbia-BC Cancer Agency Research Ethics Board (Table 3.1). Non-malignant samples were collected from areas > 2 cm away from tumour. Tissue microdissection was guided by a lung pathologist to ensure >80% tumour cell or >80% non-malignant cell content. Total RNA was extracted using Trizol reagent.

3.2.2 MiRNA Sequencing

MiRNA-seq transcriptome profiles were obtained using Illumina HiSeq 2000 platform as previously described [102]. Raw miRNA sequence libraries and sample information have been deposited in the NCBI Gene Expression Omnibus (Accession number pending) (http://www.ncbi.nlm.nih.gov/geo/). Reads were aligned to NCBI GRCh37 reference genome and miRBase v18 using the BWA algorithm [192], and multiple alignment locations resolved as previously described [102]. Full description of library construction, sequencing, read pre-processing, alignment and annotation are previously described [102]. MiRNA expression was quantified as reads per kilobase per million (RPKM). In total, 1372 unique miRNAs were detected across 188 libraries. miRNAs with RPKMs < 1 were considered not expressed. miRNAs with RPKMs < 1 across the entire cohort of tumour or non-malignant samples were disregarded, resulting in 927 miRNAs for subsequent statistical analyses.

3.2.3 The Cancer Genome Atlas (TCGA) cohort

miRNA sequencing data were obtained from the TCGA for use as an external cohort for validation purposes as well as for combining with our own dataset to perform miRNA survival association analyses. Expression profiles from the TCGA were processed as described for 'Level 3 data' in the TCGA data compendium (2011 Cancer Genome Atlas

Network). Detailed descriptions of the use of TCGA data are described in sections 3.2.4.2 and 3.2.4.4.

3.2.4 Statistical Analyses

3.2.4.1 Unsupervised hierarchical clustering of miRNA expression profiles

Unsupervised hierarchical clustering using Ward's method was performed on all samples (n=188), tumour samples only (n=94), and non-malignant samples only (n=94) using *Partek Genomics Suite* software. A Fisher's Exact test and Chi-square tests were performed to assess the distribution of tumour and non-malignant profiles, and distribution of the three smoking types within the identified clusters, respectively. A Student's t-test was used to assess differences in pack years and years quit for CS and FS. A multivariate analysis of variance (MANOVA) test was performed to determine which clinical factors were most strongly associated with grouping of non-malignant and tumour miRNA expression profiles into distinct clusters. For all statistical tests, a p-value < 0.05 was considered significant.

3.2.4.2 MiRNAs modulated in response to smoking

To identify miRNAs whose expression is likely modulated in response to smoking, we performed a non-parametric permutation test using 10,000 permutations, between non-malignant CS and NS tissues (CSN and NSN, respectively). Permutation scores were corrected for multiple testing using the Benjamini and Hochberg (B-H) method. miRNAs that had a B-H corrected $p < 0.05$ and an average fold change > 2.0 or < 0.5 were considered differentially expressed (DE) between CSN and NSN tissues. To identify miRNAs recurrently, aberrantly expressed in lung tumours of each smoking group (i.e., CS, FS, and NS), we applied the following criteria: i) pair wise Wilcoxon Sign Rank test B-H multiple testing corrected $p < 0.05$, and ii) tumour/normal fold change > 2 (overexpression) or < 0.5 (underexpression) in at least 25% of the tumours for that particular smoking group. miRNA satisfying these criteria in only one group, were considered smoking status specific and subjected to validation in the TCGA cohort. Tumour tissues from 80 FS, 42 CS, and 16 NS in the TCGA cohort were used to investigate the reproducibility of miRNA we identified as disrupted in a smoking status specific manner. Low numbers of non-malignant lung parenchymal tissues at the time of writing for the various smoking groups (12 FS, 9 CS and 2

NS) precluded us from validating our non-malignant tissue findings and required us to use pooled non-malignant samples of matched smoking history to calculate miRNA fold change for each TCGA tumour. Therefore, miRNAs were considered validated if the frequency of over- or underexpression was found to significantly differ between smoking groups (Fisher's exact test, $p < 0.05$) and there was a minimum disruption frequency difference of 15% between groups.

To identify smoking related miRNAs that may be reversibly expressed upon smoking cessation in non-malignant tissues of lung cancer patients, permutation tests were similarly run between FS non-malignant tissues (FSN), CSN and NSN groups. For each comparison, miRNAs with a B-H corrected $p < 0.05$ and an average fold change > 2.0 or < 0.5 were considered differentially expressed (DE). miRNAs were considered reversibly expressed upon smoking cessation, if they showed i) DE between CSN and NSN and between CSN and FSN, and ii) had a CSN/FSN fold change ≥ 2, but a FSN/NSN fold change < 2. Conversely, miRNAs were considered irreversibly expressed upon smoking cessation, if they were i) DE between CSN and NSN and between FSN and NSN, and ii) had a CSN/FSN fold change < 2, but a FSN/NSN fold change ≥ 2.

3.2.4.3 Generation of predicted miRNA-transcript interaction networks

miRNAs identified as preferentially disrupted in one smoking-status group were input into the *microRNA Data Integration Portal ver 2* (*miRDIP*; http://ophid.utoronto.ca/mirDIP), which integrates 13 microRNA target prediction algorithms and six microRNA prediction databases to predict miRNA-transcript (mRNA) interactions [96]. For this study, we used stringent miRNA target prediction criteria by considering only predictions that were supported by at least six sources. Interactions between miRNAs and their predicted mRNA targets were then visualized as networks using *NAViGaTOR ver 2.14* (http://ophid.utoronto.ca/navigator) [193, 194]. Two interaction networks were generated: 1) a network based on miRNAs specifically deregulated in one smoking group using all significant gene targets identified by miRDIP, and 2) a network based on miRNAs specifically disrupted in CS, FS, or NS and miRNAs commonly disrupted between all groups

using only significant gene targets identified by miRDIP that are known to be associated with lung cancer patient survival [195]. Only the most highly connected miRNA were used to build and visualize the networks. Pathway analysis was performed on biologically validated mRNA targets (miRTarBase v3.5) of miRNA disrupted in a smoking-status specific manner using Ingenuity Pathway Analysis.

3.2.4.4 MiRNA survival associations in lung cancer cohorts

Associations between miRNA expression and patient survival were assessed using a log rank, Mantel-Haenszel test. Patients were divided into tertiles based on miRNA expression and survival for patients in the top and bottom tertiles was compared. Only miRNAs detectably expressed in at least two thirds of patients were assessed to ensure adequate separation between high and low expressing groups for statistical analysis. Mantel-Haenszel p values reflect the probability of randomly selecting subjects whose survival curves are as different as actually observed. p values < 0.05 were considered significant. To enable assessment of smoking status specific survival associations, we combined miRNA expression and outcome data for our own patient cohort (n=91; 22 FS, 42 CS, 27 NS) and the TCGA LUAC cohort (n=127; 80 FS, 33 CS, and 14 NS). In total, the combined cohort contained 218 patients including 102 FS, 75 CS, and 41 NS. Survival analyses were performed on all patients and each specific smoking group. Kaplan-Meier plots were generated using GraphPad Prism 6 software. Expression data for 38 CS LUAC profiles used to calculate the association of EZH2 expression with patient survival was acquired from the Early Detection Research Network (EDRN, http://edrn.nci.nih.gov/science-data), and processed as previously described [54, 196]. EZH2 survival analysis was performed as described above.

Table 3.1 Clinical information for lung adenocarcinoma samples profiled

Characteristic	NS[1]	CS[2]	FS[3]
Number	27	43	24
Sex			
Male	7 (26%)	13 (30%)	9 (38%)
Female	20 (74%)	30 (70%)	15 (62%)
Average Age	71	64	71
Stage			
I	16 (59%)	26 (60%)	16 (67%)
II	6 (22%)	11 (26%)	6 (25%)
III	5 (19%)	4 (9%)	1 (4%)
IV	0	2 (5%)	1 (4%)
Ethnicity			
Caucasian	8 (30%)	11 (26%)	1 (4%)
Asian	16 (59%)	0 (0%)	0 (0%)
Unknown	3 (11%)	32 (74%)	23 (96%)
Average Pack Years	0	46	47
Average Years Quit	n/a	< 1	15

[1] NS: patients who smoked fewer than 100 cigarettes in their lifetime; [2] CS: smokers at the time of diagnosis; [3] FS: patients who stopped smoking at least one year prior to diagnosis date. % in brackets refer to the proportion of NS, CS or FS patients represented by variable indicated.

3.3 Results

3.3.1 MiRNA expression profiles cluster based on malignancy and smoking histories

To determine whether miRNA expression was associated with smoking status in non-malignant and lung tumour tissues, we first performed unsupervised hierarchical clustering on the 927 miRNAs with detectable expression across the 188 lung tumour and non-malignant tissues. Clustering revealed miRNA expression segregated samples based on malignancy and smoking status (Figure 3.1). When all profiles were considered, tumour and non-malignant samples clustered separately, with a significant difference between the two clusters (Figure 3.1A and 3.1D, Fisher's Exact test, $p = 2.2 \times 10^{-16}$). Clustering of non-malignant profiles revealed three clusters that were significantly different in smoking status

composition (Figure 3.1B and 3.1E, Chi-square test, $p = 5.0 \times 10^{-4}$). A similar clustering pattern was observed for tumour profiles (Figure 3.1C and 3.1F, Chi-square test, $p = 0.023$). Multivariate analysis revealed smoking and the number of years following smoking cessation were the clinical variables most strongly associated with cluster grouping in non-malignant tissue (F-value = 8.84, $p = 5.0 \times 10^{-4}$, F-value = 1.89, $p = 0.058$, respectively), whereas in tumours, age was the most significant variable associated with clustering (F-value = 4.83, $p = 0.032$) (Table 3.2). While no significant difference in pack-years was found in non-malignant tissues or CS tumours, we did observe a significant difference in pack-years for FS tumours among the two clusters dominated by CS and FS tumours (Student's t-test, $p = 0.030$). Collectively, these results suggest miRNA expression profiles in both tumour and non-malignant lung tissues are dependent on smoking histories, but that heterogeneity within ever-smoking groups exists.

Figure 3.1

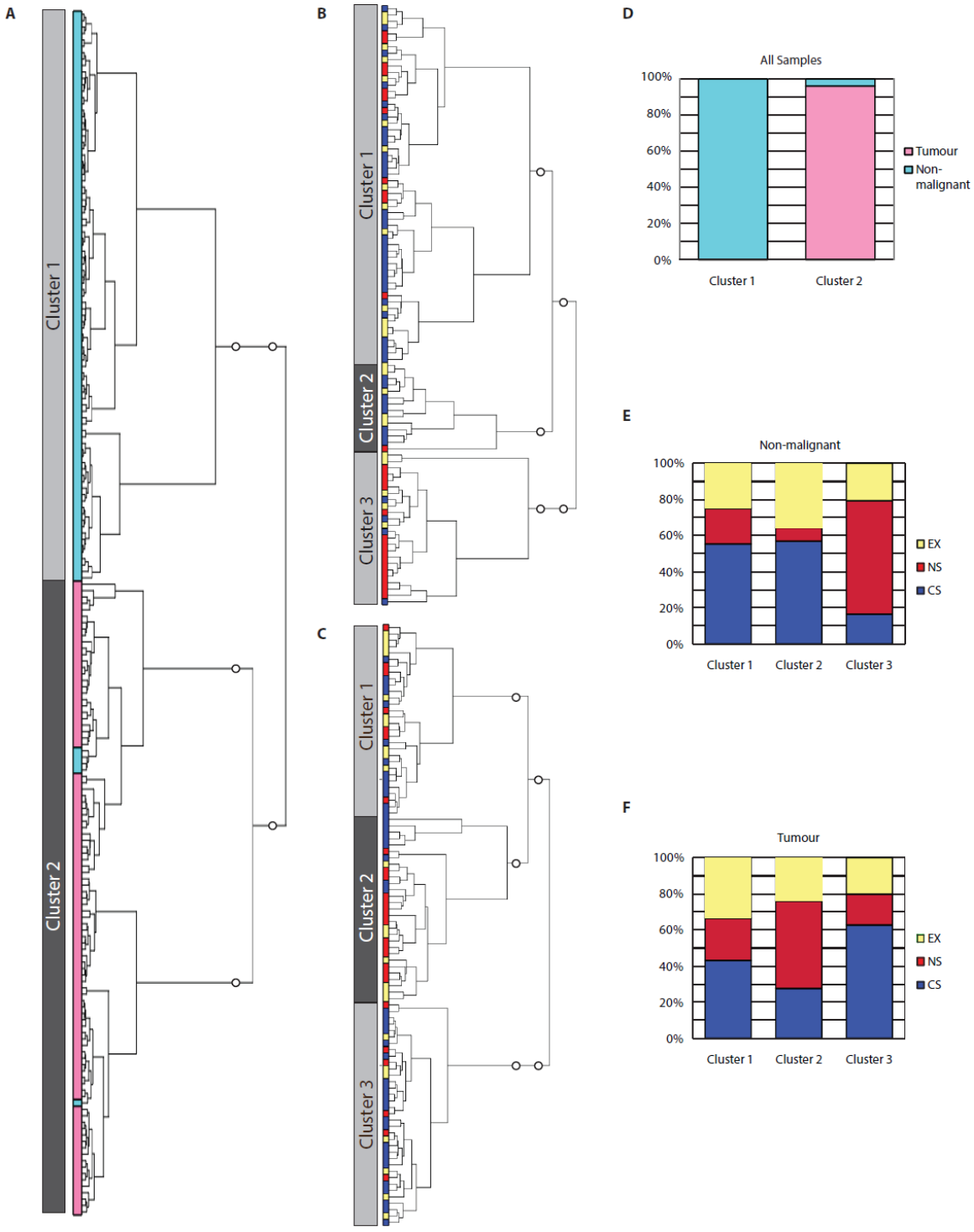

Figure 3.1 Unsupervised hierarchical clustering of lung tumour and non-malignant miRNA expression profiles

Figure 3.1 Unsupervised hierarchical clustering of lung tumour and non-malignant miRNA expression profiles

Clustering of all 188 miRNA expression profiles revealed two distinct clusters, one comprised of non-malignant samples (teal), and the other comprised of mostly tumours (pink) (A). The clusters identified were associated with malignancy, as clusters 1 and 2 were significantly enriched for non-malignant and tumour profiles, respectively (Fisher's Exact test $p = 2.2 \times 10^{-16}$) Clustering of non-malignant tissues only (B) and tumours only (C) revealed three clusters. (D). Assessment of the distribution of CS, FS, and NS within the clusters identified in non-malignant samples revealed enrichment for CS and FS in clusters 1 and 2 compared to cluster 3 (Chi-square test $p = 5.0 \times 10^{-4}$) (E). The same trend was observed in the clusters identified based on tumour profiles (Chi-square test $p = 0.023$) (F).

Table 3.2 MANOVA results for miRNA expression profile clustering

Non-malignant Samples	Df	Sum Squares	Mean Squares	F-value	Pr(>F)
Stage	3	0.70	0.23	0.41	0.75
Gender	1	0.35	0.35	0.62	0.44
Age	1	0.95	0.95	1.67	0.20
Smoking	2	10.05	5.03	8.84	0.00
Race	2	2.96	1.48	2.60	0.08
PackYears	20	11.65	0.58	1.02	0.45
YearsQuit	12	12.87	1.07	1.89	0.06
Residuals	52	29.57	0.57	-	-
Tumour Samples	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Stage	3	2.37	0.79	1.22	0.31
Gender	1	0.69	0.69	1.06	0.31
Age	1	3.13	3.13	4.83	0.03
Smoking	2	0.99	0.49	0.76	0.47
Race	2	1.82	0.91	1.41	0.25
PackYears	20	11.59	0.58	0.89	0.60
YearsQuit	12	10.41	0.87	1.34	0.23
Residuals	52	33.73	0.65	-	-

3.3.2 MiRNAs are differentially expressed between non-malignant lung tissues of CS and NS with lung cancer

Based on the observed clustering patterns, we aimed to identify miRNAs differentially expressed in non-malignant tissues of CS (CSN) and NS (NSN), as these two groups represent the most extreme smoking phenotypes. 37 miRNAs were significantly differentially expressed between CSN and NSN; 25 of which were overexpressed and 12 that were underexpressed in CSN (Table 3.3). Several of these miRNAs have been previously

implicated in lung cancer including miR-106a, miR-107, miR-136, miR-142, miR-19a, miR-212, miR-339, miR-34b, miR-34c, and miR-449a.

Table 3.3 MiRNAs differentially expressed in non-malignant lung tissue of patients with lung adenocarcinoma

miRNA	STATUS IN CSN	FC (CSN/NSN)	Average RPKM CSN	Average RPKM NSN	B-H pval	Other
hsa-mir-378c	OE	5.49	5.49	0	0.00E+00	†
hsa-mir-509	OE	3.49	7.82	2.24	1.76E-08	†
hsa-mir-374c	OE	3.26	3.26	0	0.00E+00	
hsa-mir-514a	OE	3.18	17.4	5.47	2.12E-07	†
hsa-mir-508	OE	3	26.58	8.85	5.07E-08	†
hsa-mir-339	OE	2.97	120.76	40.6	5.85E-11	†
hsa-mir-627	OE	2.96	2.96	0	5.58E-12	†
hsa-mir-539	OE	2.95	2.95	0	4.52E-12	†
hsa-mir-3648	OE	2.74	3.24	1.18	9.91E-05	
hsa-mir-151b	OE	2.73	2.73	0	0.00E+00	
hsa-mir-142	OE	2.54	3723.34	1467.49	0.00E+00	†
hsa-mir-425	OE	2.43	335.41	137.94	6.49E-11	†
hsa-mir-3687	OE	2.37	2.37	0	1.34E-04	
hsa-mir-628	OE	2.36	28.63	12.11	0.00E+00	†
hsa-mir-1277	OE	2.34	2.34	0	1.22E-12	
hsa-mir-106a	OE	2.33	23.05	9.9	0.00E+00	†
hsa-mir-19a	OE	2.27	25.6	11.3	0.00E+00	†
hsa-mir-3130	OE	2.17	5.36	2.46	1.32E-10	
hsa-mir-378a	OE	2.14	313.56	146.41	0.00E+00	†
hsa-mir-369	OE	2.14	9.77	4.57	1.05E-13	
hsa-mir-3613	OE	2.07	30.66	14.81	0.00E+00	
hsa-mir-483	OE	2.06	3	1.45	8.72E-06	†
hsa-mir-184	OE	2.06	34.36	16.66	2.91E-07	†
hsa-mir-136	OE	2.05	29.87	14.56	3.14E-11	†
hsa-mir-154	OE	2.05	5.14	2.51	4.72E-14	†
hsa-mir-934	UE	0.5	2.12	4.28	2.37E-08	
hsa-mir-592	UE	0.46	1.76	3.84	2.53E-03	†
hsa-mir-212	UE	0.42	3.66	8.64	2.60E-02	†
hsa-mir-326	UE	0.42	156.52	374.54	3.51E-05	†
hsa-mir-1180	UE	0.42	13.64	32.68	6.74E-09	
hsa-mir-34b	UE	0.41	39.82	97.96	6.21E-04	†
hsa-mir-1224	UE	0.39	0	2.56	1.54E-02	†
hsa-mir-449a	UE	0.36	1.97	5.44	2.33E-03	†
hsa-mir-34c	UE	0.34	224.8	664.81	3.99E-04	†
hsa-mir-4423	UE	0.33	1.04	3.16	2.46E-02	
hsa-mir-107	UE	0.32	784.26	2482.12	1.61E-10	†
hsa-mir-320b	UE	0.3	33.2	111.25	3.12E-12	†

† previously associated with cancer; B-H p values are Benjamini-Hochberg corrected p values from a permutation test between 43 CS non-malignant tissues (CSN) and 27 NS non-malignant tissues (NSN)

3.3.3 MiRNA expression in non-malignant tissue can be irreversibly altered in FS

Protein coding genes deregulated in response to active smoking display either reversible or irreversible expression in FS [132, 139, 197]. Genes upregulated in response to smoking that remain overexpressed in lung tissues of CS and FS with lung cancer, may indicate those smoking-related events involved in lung tumour development. We investigated

this phenomenon with respect to miRNA expression in FS non-malignant tissue (FSN), and identified two miRNAs exhibiting patterns consistent with reversible expression and 15 with irreversible expression in FSN (Table 3.4). Interestingly, the majority of these miRNAs have been associated with cancer, and four (miR-107, 142, 339, and 34c) specifically in lung cancer.

Table 3.4 MiRNA expression in non-malignant lung tissues can be irreversibly altered in former smokers

STATUS	miRNA	FC CSN/NSN	FC CSN/FSN	FC FSN/NSN	AVG CSN	AVG FSN	AVG NSN	Other
IRREV	hsa-mir-107	0.32	0.66	0.48	784.26	1181.7	2482.1	†
IRREV	hsa-mir-1180	0.42	0.84	0.5	13.64	16.27	32.68	†
IRREV	hsa-mir-1277	2.34	1.05	2.22	2.34	2.22	0	
IRREV	hsa-mir-142	2.54	1.09	2.32	3723.3	3408.5	1467.5	†
IRREV	hsa-mir-151b	2.73	1.3	2.09	2.73	2.09	0	
IRREV	hsa-mir-3130	2.17	0.63	3.47	5.36	8.55	2.46	
IRREV	hsa-mir-339	2.97	1.39	2.14	120.76	86.98	40.6	†
IRREV	hsa-mir-34c	0.34	0.95	0.36	224.8	237.16	664.81	†
IRREV	hsa-mir-374c	3.26	1.61	2.02	3.26	2.02	0	
IRREV	hsa-mir-378c	5.49	1.52	3.61	5.49	3.61	0	†
IRREV	hsa-mir-508	3	1.27	2.37	26.58	20.96	8.85	†
IRREV	hsa-mir-509	3.49	1.21	2.87	7.82	6.44	2.24	†
IRREV	hsa-mir-514a	3.18	1.4	2.27	17.4	12.42	5.47	†
IRREV	hsa-mir-539	2.95	0.94	3.15	2.95	3.15	0	†
IRREV	hsa-mir-628	2.36	0.99	2.39	28.63	28.9	12.11	†
REV	hsa-mir-3648	2.74	3.24	0	3.24	0	1.18	
REV	hsa-mir-3687	2.37	2.37	1	2.37	0	0	

† Expression of miRNA previously associated with cancer; FC= fold change; CSN: current smoker non-malignant tissue; FSN: former smokers non-malignant tissue; NSN: never smoker non-malignant tissue; AVG: average miRNA expression. Please see section 3.2.4.2 for definition of irreversibility (IRREV) and reversibility (REV).

3.3.4 MiRNAs recurrently altered in tumours from CS, FS and NS patients

To identify miRNAs recurrently differentially expressed in tumours from each smoking group, we compared expression profiles for tumour and patient matched non-malignant lung tissues of CS, FS and NS. This analysis revealed 232 overexpressed (OE) and 58 underexpressed (UE) miRNA in current smoker tumours (CST); 257 OE and 47 UE miRNAs in former smoker tumours (FST); and 263 OE and 41 UE miRNAs in never smoker tumours (NST) (Figure 3.2). Overall, the majority of miRNA were OE (304/366, 83%); 65%

(196/304) of OE miRNAs and 58% (36/62) of UE miRNAs were shared between CS, FS and NS tumours, many of which (miR-17, miR-21, miR-106a let-7a, let-7c, miR-101, and miR-143 for example) are well known lung cancer miRNAs (Appendix A and Appendix B). The identification of shared patterns of miRNA deregulation within CS, FS and NS lung tumours suggest that miRNAs likely participate in common mechanisms of tumourigenesis in lung adenocarcinoma.

Figure 3.2

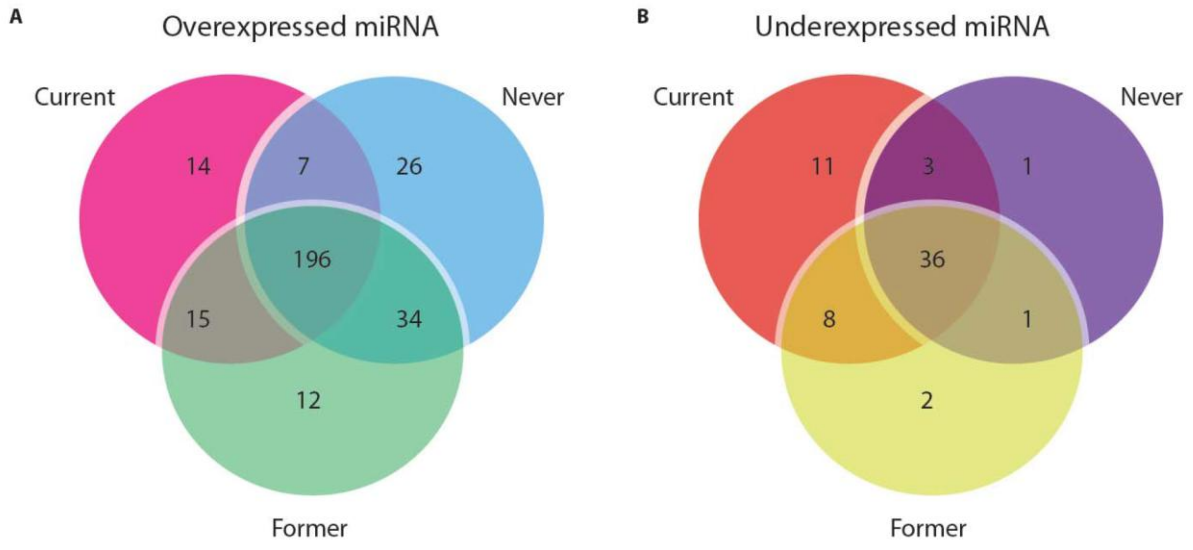

Figure 3.2 Venn diagram illustrating differentially expressed miRNAs in lung tumours relative to matched non-malignant tissues from CS, FS, and NS

miRNAs recurrently (>25%) disrupted and significantly, differentially expressed between paired, tumour and patient matched non-malignant lung tissues were assessed to determine the overlap in disruption between the groups. Overexpressed miRNAs are depicted in (A) and underexpressed miRNAs in (B). The majority of miRNAs differentially expressed between tumour and non-malignant tissues were overexpressed and most of the miRNAs identified were deregulated in all three smoking groups.

66 miRNAs were frequently altered in only one smoking tumour group: 25 in CS (14 OE and 11 UE), 27 in NS (26 OE and 1 UE) and 14 in FS (12 OE and 2 UE) (Table 3.5). We refer to these miRNAs, including those preferentially disrupted in NS, as smoking-status specific. The established cancer related functions of some of the smoking-status specific miRNA, such as overexpression of miR-7, miR-27a, miR-93, miR-372 and underexpression of miR-138, miR-381, miR-582 [198-205], suggest miRNAs are likely involved in promoting tumourigenesis in a smoking status dependent manner.

Table 3.5 MiRNA deregulation in one smoking-status lung tumour group

Overexpressed miRNA			Underexpressed miRNA		
CS	**FS**	**NS**	**CS**	**FS**	**NS**
hsa-mir-129	hsa-mir-105	hsa-mir-1295a	hsa-mir-135a	hsa-mir-381	hsa-mir-582
hsa-mir-18b	hsa-mir-1262	hsa-mir-150	hsa-mir-138	hsa-mir-607	
hsa-mir-215	hsa-mir-151b	hsa-mir-152	hsa-mir-195		
hsa-mir-337	hsa-mir-190b	hsa-mir-185	hsa-mir-3065		
hsa-mir-372	hsa-mir-23c	hsa-mir-1976	hsa-mir-378a		
hsa-mir-3940	hsa-mir-27a	hsa-mir-2114	hsa-mir-378c		
hsa-mir-411	hsa-mir-3187	hsa-mir-216a	hsa-mir-4532		
hsa-mir-545	hsa-mir-320a	hsa-mir-217	hsa-mir-4536		
hsa-mir-5571	hsa-mir-504	hsa-mir-3130	hsa-mir-511		
hsa-mir-576	hsa-mir-514b	hsa-mir-3150b	hsa-mir-532		
hsa-mir-592	hsa-mir-632	hsa-mir-320e	hsa-mir-676		
hsa-mir-654	hsa-mir-944	hsa-mir-329			
hsa-mir-7		hsa-mir-340			
hsa-mir-891a		hsa-mir-3609			
		hsa-mir-3613			
		hsa-mir-4443			
		hsa-mir-4791			
		hsa-mir-500a			
		hsa-mir-532			
		hsa-mir-5701			
		hsa-mir-612			
		hsa-mir-636			
		hsa-mir-652			
		hsa-mir-660			
		hsa-mir-675			
		hsa-mir-93			

Of the 66 miRNA that were altered in a smoking status specific manner, 57 were annotated in the TCGA dataset and 12 of these 57 miRNA were not detectably expressed in the smoking group of interest, resulting in 45 miRNA amenable for validation testing. In addition to a low number of patient matched tumour and non-malignant tissues pairs, inspection of the TCGA data revealed lower overall RPKM counts for most miRNA detected in comparison with our own dataset (Figure 3.3). Therefore we applied a different analysis

strategy for validation (described in section 3.2.4.2), resulting in 4 of the 45 assessable miRNA validating as altered in a smoking specific manner (miR-129 OE in CS, miR-152 OE in NS, miR-3065 UE in CS, and miR-511 UE in CS) (Figure 3.4).

Figure 3.3

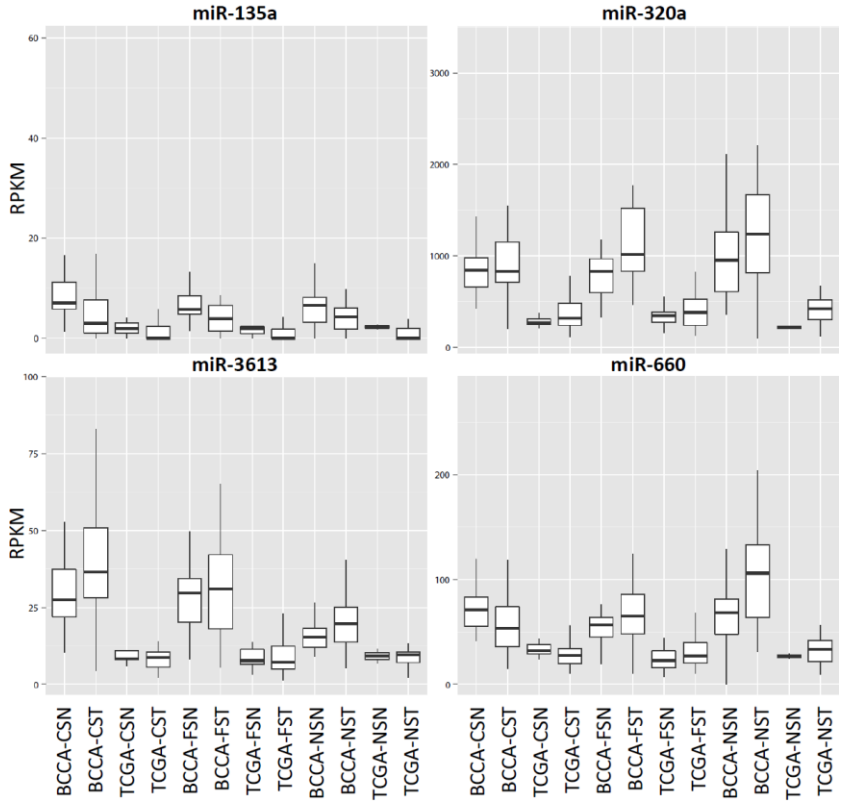

Figure 3.3 Comparison of RPKM values and miRNA detection in the TCGA and BCCA datasets.
Normalized miRNA expression values (RPKM) are plotted for CS, FS, and NS in our BCCA cohort and the TCGA dataset. The four miRNA illustrated demonstrate the difference in miRNA detection between tumour and non-malignant samples between the two datasets; miRNA in the TCGA data have lower expression values. These features of the TCGA data make validation of differentially disrupted miRNA difficult. The low validation rate we observed (4/45 assessable miRNA) likely reflects the compressed nature of the TCGA expression data. The box contains the 25th and 75th percentiles of expression data and the horizontal line indicates the median of the data. Whiskers extending from the box indicate the highest and lowest values within 1.5 times the interquartile range.

Figure 3.4

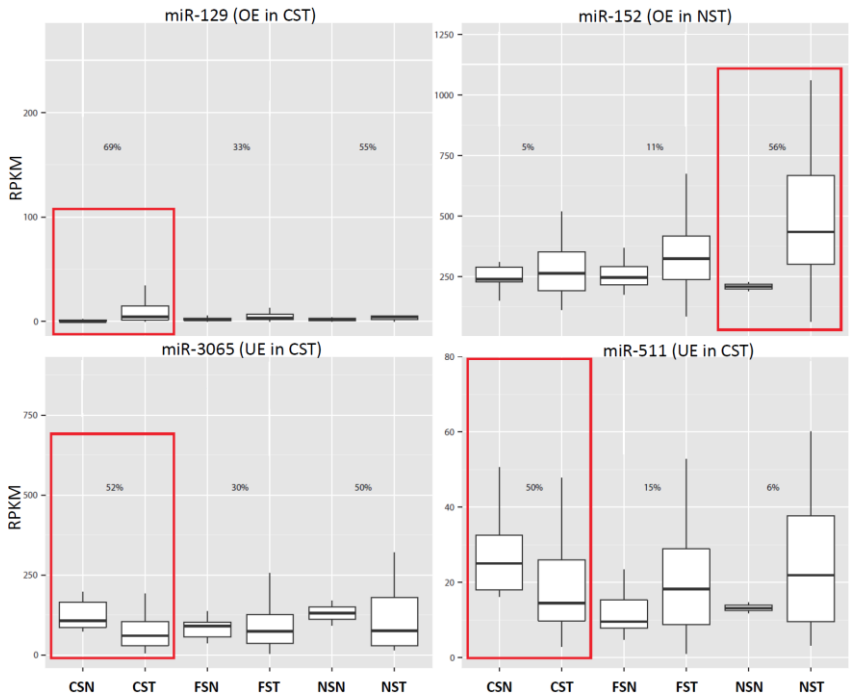

Figure 3.4 Four miRNA validated as specifically disrupted in one smoking group.
Boxplots illustrate expression values of the four miRNA we validated as disrupted in a smoking specific manner in the TCGA cohort. Red boxes indicate the group for which miRNA disruption (over- or underexpression) occurs. Frequencies of miRNA disruption in tumour relative to non-malignant samples are indicated for each smoking group. miRNAs were considered validated if they exhibited a significant difference in alteration frequency between smoking groups (Fisher's exact test, $p < 0.05$) and a minimum 15% frequency difference between smoking groups concordant with our findings (describe in section 3.2.4). CSN, FSN, NSN = CS, FS, NS non-malignant samples, respectively. CST, FST, NST = CS, FS, NS tumour samples, respectively. OE = overexpression, UE = underexpression.

3.3.5 Disrupted miRNA networks in tumours indicate selection of smoking-status specific target genes

To elucidate signaling pathways and biological processes disrupted by smoking-status specific miRNAs, mRNA target genes were identified using miRDIP with stringent filters (i.e., prediction by at least 6 different algorithms). Smoking-status specific miRNAs were predicted to affect a large number of unique mRNA targets in CS (n= 1,162 genes), NS (n= 927 genes) and FS (n= 770 genes), (Figure 3.5) which could indicate that distinct cellular

pathway selection occurs in different smoking and non-smoking environments (Figure 3.5, left side). Conversely, common mRNA target genes (n=1,399) (Figure 3.5, right side), may indicate selection of genes deregulated in lung adenocarcinoma in general. CST-specific miRNAs target mRNAs with numerous Gene Ontology functions, including cellular fate and organization, metabolism, genome maintenance, transcription, and translation. In FST and NST, mRNA targets largely corresponded to similar functions, including transport and sensing (Figure 3.5). As an independent method of assessing the potential biological implications of miRNA disruption, we performed pathway analysis on previously biologically validated targets of miRNAs (as annotated in miRTarBase v3.5) specifically deregulated in one smoking group. We found not only expected commonalities in known cancer pathways across all groups, but also biological pathways that were uniquely disrupted in specific smoking status groups; for example, SAPK/JNK signaling in NS and ERK5 in CS (Figure 3.6).

Figure 3.5

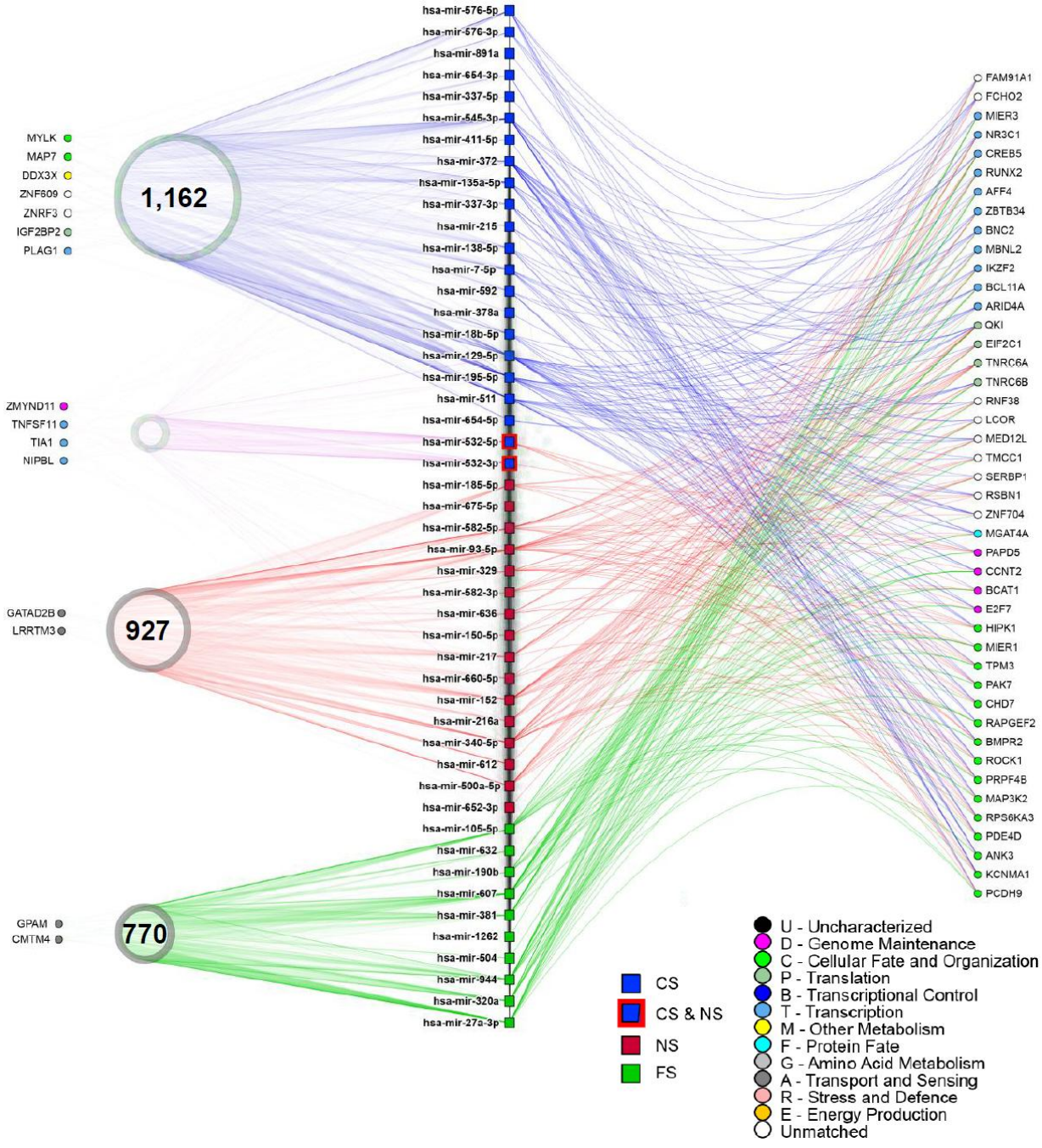

Figure 3.5 Network interactions between miRNAs specifically deregulated in CS, FS and NS lung tumours and their predicted mRNA targets

Figure 3.5 Network interactions between miRNAs specifically deregulated in CS, FS and NS lung tumours and their predicted mRNA targets

miRNAs specifically disrupted in CS, FS, or NS tumours were input into mirDIP to identify their predicted gene targets (i.e., mRNA transcripts predicted by at least six miRNA target prediction algorithms). The network of identified miRNA-mRNA interactions was then generated and visualized using NAViGaTOR. Only the most highly connected miRNA were used to build the network. miRNAs specifically deregulated in CS, FS, or NS are indicated by blue, green, and red colored square nodes, respectively. Predicted mRNA targets are represented as circular nodes. Edges indicate miRNA-mRNA interactions, and are color-coded to match smoking group specificity of miRNA deregulation. Numerous target genes were shared by miRNAs specifically deregulated in CS, FS, and NS, as shown to the right of the miRNAs list in the centre. Conversely, targets unique to specific smoking groups are indicated to the left of the list. Predicted targets uniquely mapping to miRNA uniquely disrupted in one smoking-status group are represented by circles on the left; 1,162 mRNA targets were unique to miRNA altered in CST, 927 to NST miRNA, and 770 to FST miRNA. Conversely, 1,399 mRNA miRNA targets were shared between miRNA altered uniquely in each smoking-status group. miR-532 was underexpressed in CST and overexpressed in NST. Gene Ontology terms associated with predicted target genes are indicated by target gene shading.

Figure 3.6

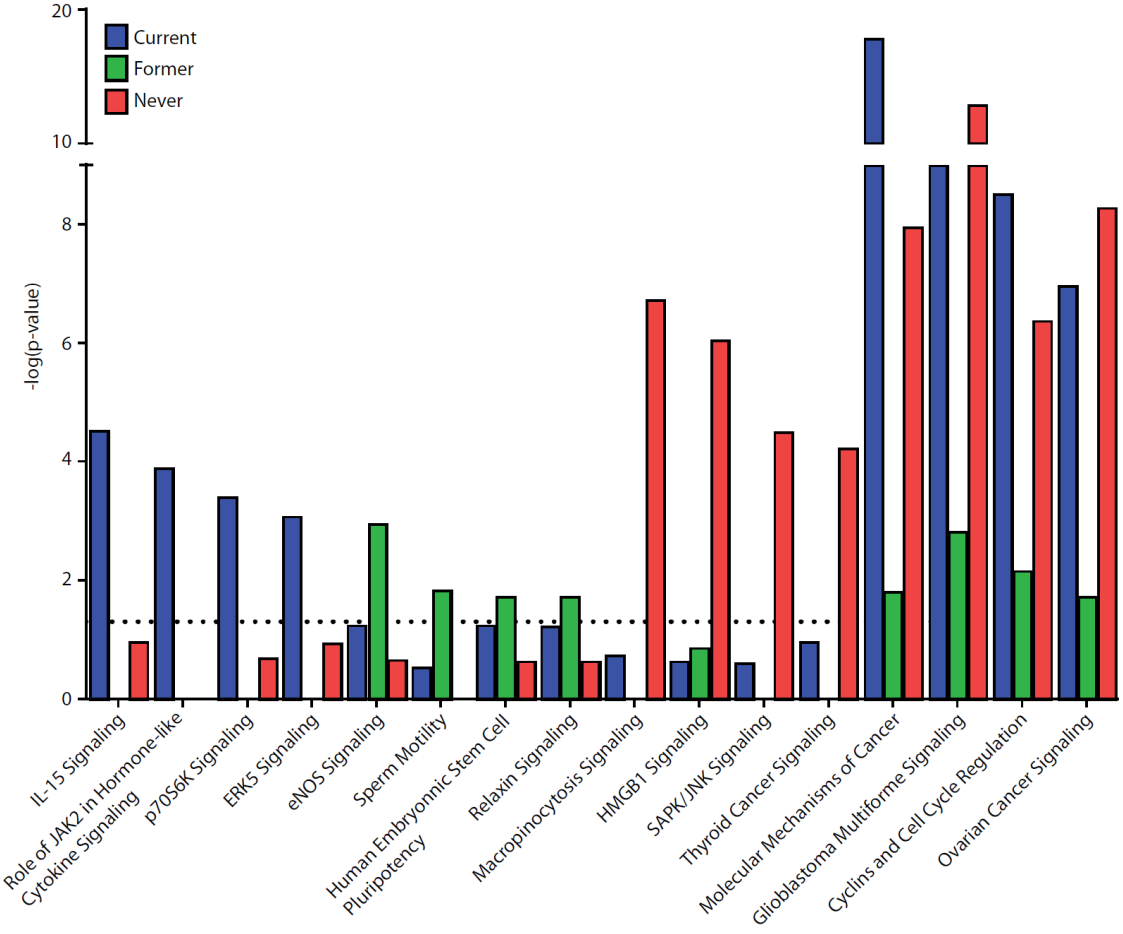

Figure 3.6 Canonical pathways differentially and commonly enriched for biologically validated target genes of miRNA specifically deregulated in one smoking group

Pathway analysis was performed on biologically validated gene targets (based on miRTarBase v3.5) of miRNAs specifically deregulated in one smoking-status group. Analyses were conducted separately for targets of miRNA specific to CS, FS, and NS. Pathways specific to CS, FS, and NS, and significantly across all smoking groups. Dotted horizontal line indicates threshold for significant pathway enrichment (Fisher's Exact test $p < 0.05$).

3.3.6 Prognostically relevant lung cancer genes are targeted by miRNAs disrupted in a smoking-status specific manner

To assess the potential prognostic implications of miRNA deregulation in lung cancer, we used a curated list of 1,066 lung cancer prognostic genes compiled by Zhu *et al.* [195], to build a miRNA-transcript interaction network comprised of both smoking-status specific miRNAs (Figure 3.7, colored square nodes) and miRNAs frequently altered across all LUAC groups (Figure 3.7, white square nodes). Of the 1,066 prognostic genes, 358 (34%) were predicted targets of the most highly connected miRNAs (n=75) used to derive the network. Prognostic genes predicted to be targeted by four or more of the miRNA are listed in Table 3.6. Interestingly, the majority of miRNAs were highly connected to the same lung cancer prognostic genes, and vice versa. For instance, miR-372, a miRNA identified as specifically overexpressed in CST, was connected to 31 different lung cancer prognostic genes. Conversely, *nuclear transcription factor I, beta* (*NFIB*), is a predicted target of 16 unique miRNAs, highlighting the potential importance of this gene to LUAC. Moreover, of the miRNAs disrupted in a smoking-status specific manner, CS (blue) and FS (green) miRNAs demonstrated a higher number of connections to lung cancer prognostic genes than NS-specific miRNAs (red). Taken together, these data re-emphasize the biological and prognostic relevance of miRNA disruption in lung cancer, and highlight potential clinical and biological differences based on smoking status.

Figure 3.7

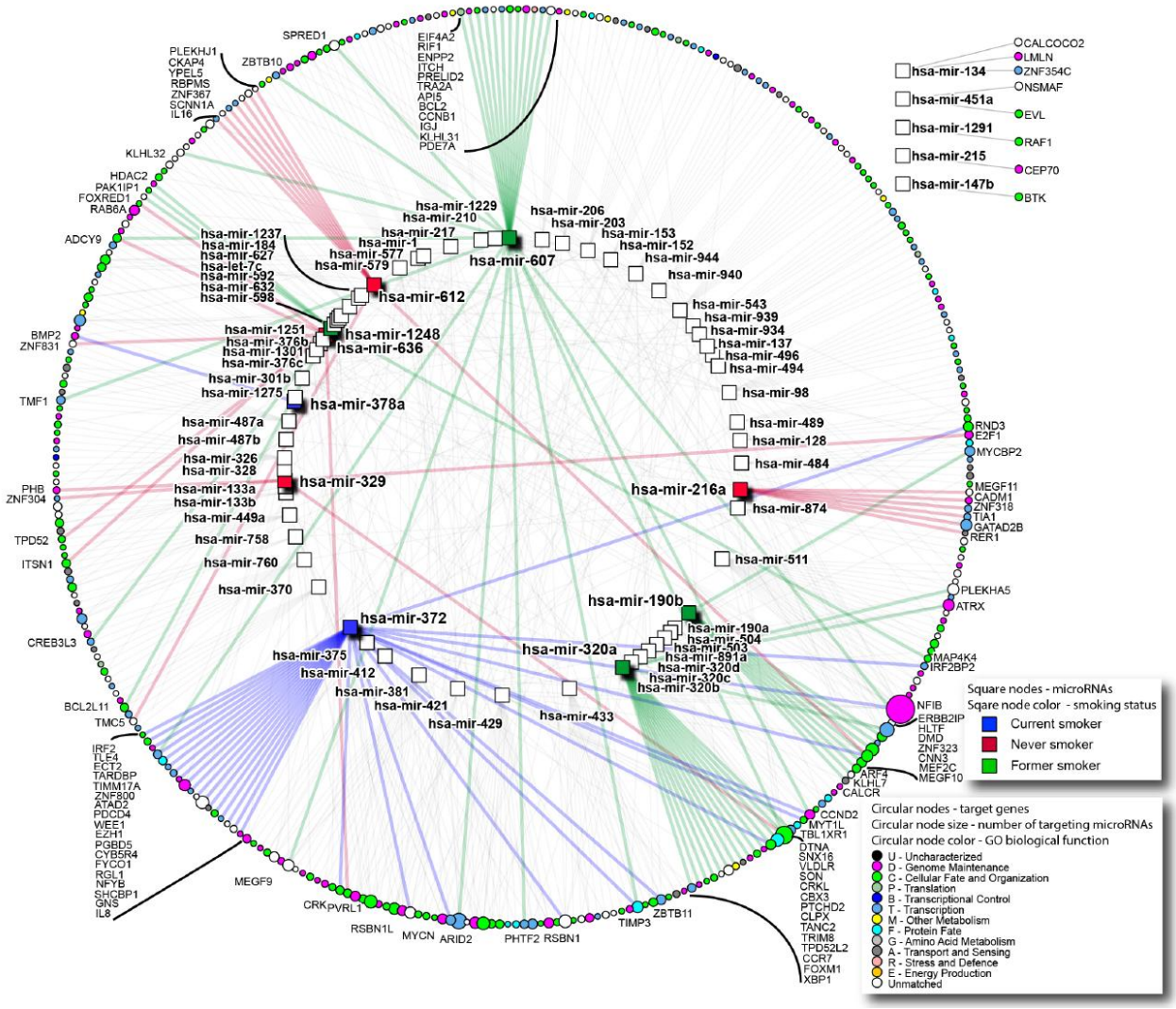

Figure 3.7 Predicted interaction between prognostic lung cancer genes and miRNAs deregulated in tumours from CS, FS and NS

Figure 3.7 Predicted interaction between prognostic lung cancer genes and miRNAs deregulated in tumours from CS, FS and NS

MiRNAs specifically disrupted in CS, FS, or NS tumours as well as miRNA frequently disrupted across the groups were inputted into mirDIP to identify their predicted gene targets. The network of identified miRNA-mRNA interactions was then generated and visualized using NAViGaTOR, but was restricted to predicted target genes that are known to have prognostic significance in lung cancer. miRNAs specifically deregulated in a single smoking group are indicated by colored square nodes. miRNAs disrupted in multiple groups are indicated by white square nodes. Connections for miRNAs commonly disrupted among the smoking groups are indicated by grey edges, while blue, green and red edges indicate miRNA-mRNA interactions specific to CS, FS, and NS, respectively. Predicted targets are depicted as circular nodes, with shading corresponding to Gene Ontology terms associated with gene function. The degree of connectivity for gene targets is depicted by the target node size, where larger circular nodes indicate genes targeted by a greater number of different miRNAs. In total, the network is comprised of 75 miRNAs and 385 prognostic target genes. Most miRNAs are well connected to prognostic genes, with more connections for CS and FS specific miRNAs and fewer connections for NS specific miRNAs. miR-372, miR-607, and miR-543 were among the miRNAs most highly connected to lung cancer prognostic gene targets.

Table 3.6 Prognostic lung cancer genes targeted by multiple miRNA

Prognostic lung cancer gene	Number of predicted targeting miRNA	Target Gene Name	GO Biological Function
NFIB	16	Nuclear factor 1 B-type	D
DTNA	9	Dystrobrevin alpha	C
ARID2	8	AT-rich interactive domain-containing protein 2	T
HLTF	7	Helicase-like transcription factor	T
MAP4K3	6	Mitogen-activated protein kinase kinase kinase kinase 3	C
ABCA1	6	ATP-binding cassette sub-family A member 1	C
CNN3	6	Calponin-3	C
SNX16	6	Sorting nexin-16	F
CYB5R4	6	Cytochrome b5 reductase 4	N
RSBN1	6	Round spermatid basic protein 1	N
JARID2	5	Protein Jumonji	C
MEF2C	5	Myocyte-specific enhancer factor 2C	C
AGFG1	5	Arf-GAP domain and FG repeats-containing protein 1	D
WEE1	5	Wee1-like protein kinase	D
ATRX	5	Transcriptional regulator ATRX	D
TIMP3	5	Metalloproteinase inhibitor 3	F
PLEKHA5	5	Pleckstrin homology domain-containing family A member 5	N
RSBN1L	5	Round spermatid basic protein 1-like protein	N
MEGF9	5	Multiple epidermal growth factor-like domains protein 9	N
GATAD2B	5	Transcriptional repressor p66-beta	T
MAF	5	Transcription factor Maf	T
MEGF10	4	Multiple epidermal growth factor-like domains protein 10	C
RND3	4	Rho-related GTP-binding protein RhoE	C
DMD	4	Dystrophin	C
NPTN	4	Neuroplastin	C
VLDLR	4	Very low-density lipoprotein receptor	C
FGFR2	4	Fibroblast growth factor receptor 2	D
CCNE2	4	G1/S-specific cyclin-E2	D
CUL4B	4	Cullin-4B	D
CCND2	4	G1/S-specific cyclin-D2	D
HMGA2	4	High mobility group protein HMGI-C	D
CDC73	4	Parafibromin	D
TANC2	4	Protein TANC2	N
UBL3	4	Ubiquitin-like protein 3	N
SPRED1	4	Sprouty-related, EVH1 domain-containing protein 1	N
RUNX1T1	4	Protein CBFA2T1	T
TBL1XR1	4	F-box-like/WD repeat-containing protein TBL1XR1	T
ZBTB11	4	Zinc finger and BTB domain-containing protein 11	T
HLF	4	Hepatic leukemia factor	T
MYCBP2	4	Probable E3 ubiquitin-protein ligase MYCBP2	T
MYCN	4	N-myc proto-oncogene protein	T

Gene Ontology (GO) functions: D - Genome Maintenance; C - Cellular Fate and Organization; T - Transcription; F - Protein Fate; N - Not Matched

3.3.7 MiRNAs disrupted specifically in CS, FS, or NS tumours are associated with outcome

Due to the scarcity of publically available cohorts with smoking status large enough to enable statistical analysis and comparison of survival, analysis of miRNA expression in relation to CS, FS or NS lung cancer patient survival has to our knowledge not been previously assessed. By combining miRNA expression and patient survival data from our lung adenocarcinoma (LUAC) cohort (n=91 tumours), with the TCGA's LUAC cohort (n=127 tumours), for a total of 218 LUAC corresponding to 102 FST, 75 CST and 41 NST, we addressed these challenges and performed the first smoking status-specific miRNA survival analysis.

Of the miRNAs connected to lung cancer prognostic genes (Figure 3.7), 15 were significantly associated with survival (Mantel Haenszel, logrank $p < 0.05$), including miR-1 and miR-153 (Figure 3.8A and 3.8B and Table 3.7). Considering all miRNAs altered in a smoking-status specific or shared manner, when all smoking groups were combined we identified 76 miRNAs as significantly associated with LUAC patient survival (Mantel-Haenszel, logrank $p < 0.05$), 22 of which were significant after correcting for multiple testing (B-H $p < 0.05$) (Appendix C). These included miRNAs previously associated with LUAC patient survival (miR-1247, let-7g, miR-146a, miR-126) [206-209] and recurrence (miR-200b) [210]. miR-187 which has been previously associated with brain metastasis in lung cancer patients [206] was the most significant miRNA associated with survival and disrupted in that same smoking group overall (Figure 3.9). Within individual smoking groups, 71 miRNAs were associated with patient survival in FS, 12 in NS and 11 in CS (Table 3.8 and Appendix C). Low expression of the CST-specific tumour suppressor miR-138 was also associated with poor survival in CS ($p = 0.009$) (Figure 3.8C). High expression of EZH2, a recently biologically validated target of miR-138 in LUAC cells, was concordantly associated with poor patient survival ($p = 0.021$, Figure 3.8D). Collectively, these survival data provide further rationale for the stratification of lung cancer patients based on detailed smoking histories, and the evaluation of miRNAs in LUAC biology in the context of smoking.

Figure 3.8

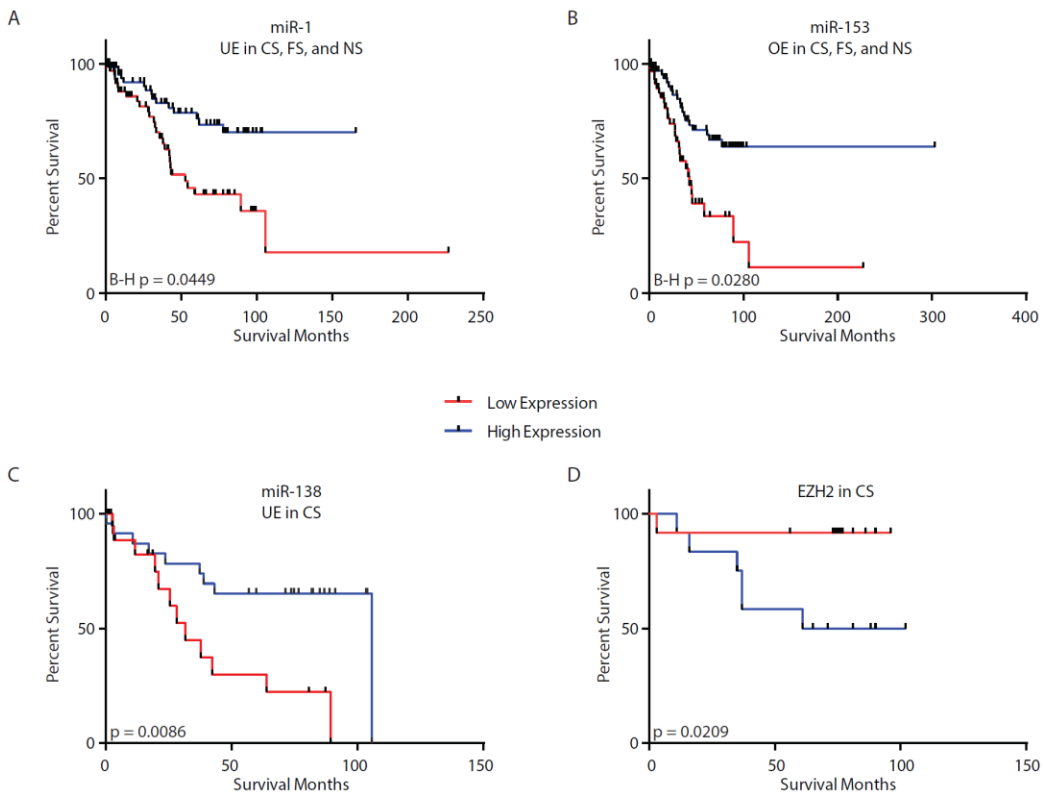

Figure 3.8 MiRNAs frequently deregulated in CS, FS and NS related lung adenocarcinomas are associated with patient survival

Associations between miRNA expression and LUAC patient survival were assessed for miRNAs identified as deregulated in LUAC using a logrank, Mantel-Haenszel test. Survival analyses were performed independently for tumours from all smoking status groups, CS, FS, and NS patients. Numerous miRNAs commonly disrupted across CS, FS, and NS were significantly associated with LUAC patient survival, including miR-1 (A) and miR-153 (B) (B-H $p < 0.05$) which were both connected to multiple lung cancer prognostic genes from Figure 3.7. miR-138, which was preferentially underexpressed in CS tumours, was also significantly associated with CS LUAC patient outcome, with low expression associated with poor survival (C) ($p < 0.05$). High expression of EZH2, a biologically validated target of miR-138, showed a significant association with poorer survival in CS LUAC patients (D) ($p < 0.05$). OE, overexpressed. UE, underexpressed. NS, never smokers. CS, current smokers. FS, former smokers. B-H, Benjamini-Hochberg multiple test corrected.

Table 3.7 MiRNA associated with lung AC patient survival that target lung cancer prognostic genes

miRNA	Status in Tumours			Mantel-Haenszel p value				Other
	CST	FST	NST	All Lung AC	CS	FS	NS	
hsa-mir-1	UE	UE	UE	0.0021	0.8863	0.0179	0.0565	†
hsa-mir-1301	OE	OE	OE	0.0214	0.2644	0.0122	0.3624	-
hsa-mir-133a	UE	UE	UE	0.0096	0.6305	0.0001	0.3569	-
hsa-mir-133b	UE	UE	UE	0.0414	na	na	0.0304	†
hsa-mir-153	OE	OE	OE	0.0006	0.9752	0.0006	0.3777	-
hsa-mir-184	UE	UE		0.0036	na	0.2470	0.0103	†
hsa-mir-320d	OE	OE	OE	0.0371	0.5347	na	0.6182	†
hsa-mir-326	OE	OE		0.2404	0.0131	0.5943	0.6563	†
hsa-mir-328	OE	OE	OE	0.0139	0.0520	0.0043	0.9766	†
hsa-mir-375	OE	OE	OE	0.0177	0.7243	0.0387	0.2538	†
hsa-mir-376c	OE	OE	OE	0.0965	0.9687	0.0076	0.4326	-
hsa-mir-429	OE	OE	OE	0.0107	0.7818	0.0002	0.6960	†
hsa-mir-484	OE	OE	OE	0.0394	0.3633	0.0110	0.7589	†
hsa-mir-598	UE	UE	UE	0.2408	0.6718	0.0447	0.0326	†
hsa-mir-940	OE	OE	OE	0.0289	0.4283	0.0656	0.6866	-

p values with "na" refer to miRNA that were not sufficienty variably expressed for inclusion in survival analysis† previously associated with cancer prognosis; OE or UE refer to frequently over- or underexpressed miRNA in lung tumours from current (CST), former (FST) or never (NST) smoker tumours. Survival analysis was performed in either all tumour as a group (All Lung AC), current smokers only (CS), former smokers only (FS) or never smokers only (NS).

Figure 3.9

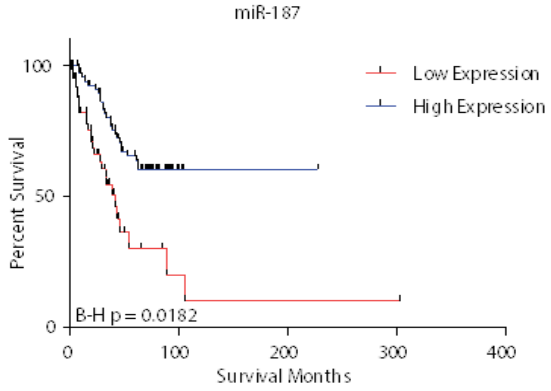

Figure 3.9 MiR-187 is the most significant miRNA associated with patient survival.
Associations between miRNA expression and LUAC patient survival were assessed using a logrank, Mantel-Haenszel test. Considering all patient samples combined, regardless of smoking histories, miR-187 was the disrupted miRNA most significantly associated with patient survival after multiple testing correction.

Table 3.8 Top 10 miRNA significantly associated with lung adenocarcinoma patient survival in CS, FS, and NS

Tumour Group Surival Assoc.	miRNA	p-value	Status in Tumours
CS	hsa-mir-1287	0.0022	OE in ALL
	hsa-mir-138	0.0086	UE in CST
	hsa-mir-326	0.0131	OE in FST and NST
	hsa-mir-331	0.0146	OE in ALL
	hsa-mir-30d	0.0282	UE in ALL
	hsa-mir-204	0.0291	UE in ALL
	hsa-mir-664	0.0331	OE in FST and NST
	hsa-mir-148a	0.0429	OE in ALL
	hsa-mir-195	0.0436	UE in CST
	hsa-mir-1270	0.0462	OE in ALL
FS	hsa-mir-133a	0.0001	UE in ALL
	hsa-mir-429	0.0002	OE in ALL
	hsa-mir-642a	0.0005	OE in ALL
	hsa-mir-153	0.0006	OE in ALL
	hsa-mir-187	0.001	OE in ALL
	hsa-mir-21	0.0013	OE in ALL
	hsa-mir-26b	0.0018	OE in ALL
	hsa-mir-135b	0.0021	OE in ALL
	hsa-mir-3607	0.0022	OE in ALL
	hsa-mir-99b	0.0027	OE in CST and FST
NS	hsa-mir-338	0.0006	UE in ALL
	hsa-let-7g	0.0036	OE in ALL
	hsa-mir-184	0.0103	UE in CST and FST
	hsa-mir-150	0.0143	OE in NST
	hsa-mir-139	0.02	UE in ALL
	hsa-mir-133b	0.0304	UE in ALL
	hsa-mir-664	0.0307	OE in FST and NST
	hsa-mir-598	0.0326	UE in ALL
	hsa-mir-10a	0.0342	UE in CST and FST
	hsa-mir-92b	0.0351	OE in ALL

3.4 Discussion

Cigarette smoke is associated with specific modifications to the genomic and epigenomic landscapes of airways and lung tissues [8, 142], affecting the transcriptional regulation of both genes and microRNAs (miRNAs) [179, 183-185]. Recent evidence suggests that histologically similar lung tumours harbour distinct molecular profiles based on smoking status, and that these alterations underlie observed clinical disparities between lung tumours in smokers and NS [50]. Since large LUAC cohorts with well annotated smoking histories and multi-omics data in both tumour and non-malignant tissues are only very recently becoming available, few studies have directly investigated the effect of smoking on molecular features of lung tumours on a global 'omics level, particularly for miRNAs. The LUAC cohort we have compiled is the largest lung tumour data set with well-defined smoking status annotation and matched non-malignant tissue for every patient, to date (n=94 patients, 188 miRNA sequencing profiles).

Hierarchical clustering revealed that while smoking status and malignancy were associated with miRNA expression patterns, heterogeneity amongst CS and FS is present (Figure 3.1). It is likely that in addition to other non-miRNA molecular alterations, inter-individual genetic variants involved in the biological response to smoking may underlie this observed heterogeneity [211-213]. Analogous to observations for protein coding genes [132, 139], we identified miRNAs differentially expressed between CSN and NSN and either irreversibly (miR-107, -378c, -142 and -34c) or reversibly (miR-3648 and miR-3687) expressed in FSN (Table 3.4). It is plausible that altered expression of miRNAs in CSN and FSN tissues may be an early event related to smoke-associated tumourigenesis, although without interrogation of lung tissues from CS, FS and NS individuals *without* lung cancer, it is difficult to distinguish smoking induced alterations from those that are related to lung cancer itself.

The majority of frequently disrupted miRNAs in LUAC relative to non-malignant lung tissues were commonly disrupted across all tumour groups, indicating that despite different smoking histories, common biological mechanisms largely underlie LUAC tumourigenesis. However, our findings of smoking-status specific miRNAs with unique

mRNA targets, suggests that miRNAs may underlie LUAC tumourigenesis in distinct smoking environments (Figure 3.5 and 3.7). For example, gene targets unique to miRNAs disrupted in CST, FST or NST could be indicative of the importance of these miRNAs to LUAC biology in distinct smoking environments. Conversely, genes heavily targeted by different miRNAs distinctly altered in CST, FST or NST, may indicate the importance of these genes to LUAC biology, irrespective of smoking status (Figure 3.5 and 3.7).

A large number of lung cancer prognostic genes were identified as predicted targets for both commonly disrupted and smoking-status specific miRNAs [195]. CST specific miR-372 targeted the most lung cancer prognostic genes and is a known oncogenic miRNA associated with poor outcome and aggressive disease in multiple cancers, including lung cancer, where it is a strong candidate for use as an early detection sputum-based biomarker [198, 199, 214-216] (Figure 3.7). The numerous targets of miR-372 were recently described in a LUAC comparative proteomic analysis, further alluding to the extensive pro-tumourigenic role of this miRNA in lung cancer [217]. However, despite being overexpressed in CST (26%), the low variability in its expression levels across CST prevented us from statistically assessing the association between miR-372 expression levels and lung cancer patient survival in our study.

The degree of overlap between miRNAs and prognostic mRNA targets was particularly high for miRNAs specifically disrupted in CST or FST, a finding we suspect is due to the fact that the lung cancer prognostic signatures we analyzed were based primarily on typical lung cancer patient cohorts which contain small numbers of NST. These findings underscore the potential clinical significance of miRNAs frequently altered in LUAC and illustrate the complexity that disruption of even few miRNAs can introduce in the study of disease biology.

In contrast to previous studies where the association of miRNA expression and survival has been conducted in relation to NSCLC histological subtypes, mutational status, or tumour stage [186, 188, 190, 191, 206], we conducted an analysis of miRNA expression associated with lung cancer patient survival in relation to smoking status. We identified

numerous miRNAs significantly associated with patient prognosis in different smoking-status groups or in LUAC in general (Table 3.7 and 3.8). Of interest, three miRNAs, miR-195, miR-138 and miR-150, which demonstrated recurrent, aberrant expression in a specific smoking-status group, were significantly associated with survival in that same smoking group.

Low expression levels of miR-195 are associated with poor patient prognosis in glioblastoma and colon cancer [218, 219], consistent with the prognostic association we identified for this miRNA in LUAC for CST. In the context of cigarette smoke and non-malignant lung disease, miR-138 may have a role in hypoxic pulmonary vascular remodelling and pulmonary arterial hypertension through its role in the negative regulation of pulmonary artery smooth muscle cell apoptosis [220]. A recent study by Zhang *et al*. not only validated an anti-tumourigenic role for miR-138, but also demonstrated that this action occurred through targeted inhibition of EZH2 by miR-138 [221]. This study provides independent validation of our target prediction methods and provides further biological evidence of the importance of miR-138 to lung cancer (Figure 3.8).

miR-150 is a candidate oncogenic miRNA, although its role in lung cancer is ambiguous [222-224]. Sun *et al.* report that expression of miR-150 is significantly downregulated in tumour tissues and embryonic lung tissues compared to normal lung tissues, although preferentially in tumours from smokers [222]. Two additional studies however identified upregulation of miR-150 in lung tumours, demonstrating a link between lung cancer cell proliferation and miR-150 through targeted inhibition of TP53 [223, 224]. We observed frequent (40%) overexpression of miR-150 specifically in NST, and found its overexpression was associated with better prognosis in NS LUAC patients. Thus, the mechanisms contributing to prognostic significance of miR-150 in LUAC may be related to the biology underlying lung tumourigenesis in NS.

In conclusion, our study suggests that patterns of miRNA deregulation promote smoking-specific LUAC biology, but also highlights shared biology underlying LUAC tumourigenesis across all smoking and non-smoking groups. Collectively, our results

reaffirm the extent to which miRNAs can contribute to the molecular complexity of cancer genomes and suggest that miRNA disruption may contribute to the distinct clinical features and outcomes observed in CS, FS, and NS lung cancer patients.

4 Chapter: Identification of genes and pathways disrupted in lung adenocarcinoma tumours from patients with and without COPD using a multi-omics approach

4.1 Introduction

Worldwide, the combined burden of COPD and lung cancer is staggering [1, 225, 226]. These diseases share many of the same genetic and environmental risk factors including smoking. However COPD alone is associated with an up to 10 fold increased risk of lung cancer that is correlated with lung function decline, independent of smoking history, suggesting a mechanistic link between COPD progression and lung tumourigenesis [64]. Despite considerable efforts towards understanding lung cancer biology at the molecular level and the increasing availability of large multi-omics data in the public domain; lung tumour 'omics datasets with associated lung function measurements are extremely rare, thus little is known about the molecular biology of lung cancer in the context of COPD. There are currently no therapies that prevent or inhibit COPD progression or lung cancer, and no clinically viable molecular markers for early detection. An improved understanding of the molecular biology underlying COPD-related lung cancer is thus urgently needed so that effective prevention, treatment and early detection regimes may be developed.

Chronic inflammation is causally associated with cancer development in a variety of tissue types including the lung in the context of COPD. We hypothesized that lung tumours arising in an environment of COPD-- a chronic inflammatory lung disease, would harbour distinct and clinically relevant molecular alterations compared to lung tumours from non-COPD patients. Moreover, since tumour systems are altered at multiple 'omic levels, we posited that interrogation of integrated multi-omics datasets would be conducive to elucidating genes selectively altered in tumour systems. We applied such an approach to the identification of genes and pathways disrupted in lung adenocarcinoma tumours from patients with and without COPD. We present the first multi-omic interrogation of COPD-related lung tumours using the largest COPD-associated lung tumour cohort to date.

To evaluate the potential clinical application of our findings, we assessed whether genes heavily disrupted in COPD-related lung tumours were also disrupted at the level of DNA methylation in small airways from patients with COPD and non small cell lung cancer (NSCLC). We propose that genes important to COPD-related lung tumour biology altered in non-malignant airway tissues of COPD patients with NSCLC warrant further exploration as epigenetic-based early detection biomarkers or as potential targets for chemoprevention therapies in COPD patients.

4.2 Methods

4.2.1 Description of cohort and clinical samples

Fresh-frozen lung adenocarcinoma tumours and patient matched non-malignant lung parenchymal tissue were collected for 73 treatment naïve patients at Vancouver General Hospital under informed, written patient consent and with approval from the University of British Columbia-BC Cancer Agency Research Ethics Board (Table 4.1). Patient matched non-malignant control lung parenchymal tissues samples were collected > 2 cm away from the tumour site. Microdissection was guided by hematoxylin and eosin stained sections graded by a lung pathologist for > 80% tumour cell or > 80% non-malignant cell content. DNA was extracted using standard phenol-chloroform procedures. RNA was extracted from tumour and matched non-malignant normal tissue using RNeasy Mini Kits (Qiagen Inc.) or Trizol reagent (Invitrogen, CA). Quality and quantity of genomic material was assessed using a NanoDrop 1000 spectrophotometer and by gel electrophoresis and/or by Agilent 2100 Bioanalyzer.

Table 4.1 Discovery lung tumour cohort patient demographics

	COPD	Non-COPD	p value
n =	**29**	**44**	
Sex			3.00E-01
Male	11	11	
Female	18	33	
Average Age	69	69	9.78E-01
Lung Function			
FEV_1 Act	1.9	2.11	8.77E-02
FEV_1% Pred	75	90	6.38E-04
FEV_1/FVC%	61	76	2.89E-15
Smoking			2.00E-03
Current	18	15	
Former	7	5	
Never	4	24	
Average Pack Years	55	28	1.09E-03
Stage			3.04E-01
I	17	25	
II	9	10	
III	1	6	
IV	2	2	
Ethnicity			2.00E-03
Caucasian	26	24	
Asian	3	19	
Native American	0	1	

p values refer to two tailed students t-test or Fisher's exact test for continuous and categorical variables respectively

4.2.2 Genome-wide multi-omics profiling

4.2.2.1 DNA copy number platform

The Affymetrix SNP 6.0 array was used to obtain DNA copy number status of over 900,000 non-polymorphic probes. DNA from tumours and non-malignant parenchymal tissues were hybridized to this array. Raw CEL probe intensity files were processed and normalized using Partek Genomics Suite Software and probe sequence, fragment length, GC content and background adjustments were applied to correct for biases in signal intensities as described by Thu *et al.* [54]. Each matched non-malignant profile was used as a copy number baseline for a respective tumour and final copy number values were generated in Partek software using the Copy Number, Paired Analysis Workflow.

4.2.2.2 DNA methylation platform

DNA methylation profiles for tumour and non-malignant tissues were obtained using the Illumina Infinium Methylation (HM27) beadchip as detailed in Section 2.2.5. Processing

and filtering for quality, colour correction and quantile normalization were also performed as described in Section 2.2.5.

4.2.2.3 mRNA expression platform

The Illumina HT-12 Whole Genome 6, v3 BeadChip array was used to obtain expression level data for over 48,000 probes corresponding to 25,000 genes. Total RNA from tumour and matched non-malignant lung tissues were hybridized to separate arrays according to the manufacturer's instructions. Bead-level data were pre-processed using the R package mbcb to perform background correction and probe summarization and pre-processed data quantile normalized and $\log_2$ transformed [227].

4.2.3 Analytical approach: Identification of genes likely selectively activated or inactivated in each tumour

Currently available informatics methods for integrating 'omics data are based on correlative or regression based models whereby 'omics levels are analyzed independently or simultaneously, and then alterations are statistically compared between tumour and normal groups [103, 228-233]. We have previously shown that independent analyses of ‘omic dimensions can overlook biologically relevant genes and pathways disrupted through different mechanisms in individual tumours [77, 110]. Therefore we developed a multi-'omic integrative gene prioritization algorithm to generate “Integrated Scores” for each gene based on the magnitude of concomitant DNA and mRNA alteration. Our algorithm simultaneously i) assesses multi-omic mechanisms of gene disruption, ii) weighs the impact of disruption on gene expression, and iii) provides an integrated score for each gene based on the extent of disruption within individual tumours.

4.2.3.1 Calculation of gene scores

Scores indicate how prominently disrupted a gene is in a tumour compared to its control tissue. For each data dimension scores are binned into categories based on commonly used thresholds for defining gene alterations in tumour systems. Each bin is assigned a direction (+ or -) to signify the likely putative effect on mRNA.

Copy number: For gene dosage, we utilized commonly applied $\log_2$ copy number ratio thresholds to call genes gained (ratio > 0.3) or lost (ratio < -0.3) (Table 4.2) [133, 234]. We specified two categories for copy number scores: low level changes (e.g. single copy gains or losses, $\log_2$ ±0.3-0.8) and high level changes (e.g. DNA amplifications or deletions $\log_2$ > ±0.8) and assigned a two point difference between categories.

DNA Methylation: Delta β-values (dβV) refer to the difference in percent methylation between a tumour and matched non-malignant sample (Table 4.2). dβV ≥ 0.2 is commonly considered aberrant in tumour systems [133]. Genes with dβV ≥ 0.2 are considered hypermethylated or dβV ≤ -0.2, hypomethylated, and binned into two categories (Bin 1: ±0.2-0.6 and Bin 2: ±0.6-1), with a two point score difference for the dβV bins. The range of scores for both methylation and copy number are the same (0 to 4), as we do not assume the effects of copy number alterations to be more or less important than those of methylation alterations.

mRNA Expression: We applied a minimum fold change (FC) (tumour/normal) of two for defining aberrant expression. A FC > 2 or < 0.5 is over or underexpression, respectively (Table 4.2). Four scoring bins spanning the range of fold changes observed (Bin1: 2-4, Bin2: 4-10, Bin3: 10-50, Bin4: > 50) were specified for both over and underexpression. To limit the inclusion of "expression only" genes in the top ranking genes, we assigned expression scores such that expression changes falling within the highest magnitude bins would produce a total score comparable to scores for genes with simultaneous DNA and associated mRNA changes. However, high magnitude mRNA changes may be indicative of underlying genetic or epigenetic alterations we are not assessing; therefore their inclusion is warranted.

Table 4.2 Scoring system indicates magnitude of DNA or mRNA level change of a gene in a tumour

COPY NUMBER	SCORE BIN	e.g. SCORES
No change (e.g. copy neutral)	$-0.3 < CN \leq 0.3$	0
Low level changes (e.g. single copy gains or losses)	$0.3 < CN \leq 0.8$ or $-0.3 \leq CN < -0.8$	+2 or -2
High level changes (e.g. amplification or homozygous deletion)	$CN > 0.8$ or $CN < -0.8$	+4 or -4
DNA METHYLATION		
No difference	$-0.2 < d\beta V \leq 0.2$	0
Methylation difference magnitude 1	$-0.2 < d\beta V \leq -0.4$ or $0.2 < d\beta V \leq 0.4$	+2 or -2
Methylation difference magnitude 2	$-0.4 < d\beta V \leq -0.6$ or $0.4 < d\beta V \leq 0.6$	+4 or -4
GENE EXPRESSION		
No difference	$0.5 < FC \leq 2$	0
Fold change magnitude 1	$2 < FC \leq 4$ or $0.25 < FC \leq 0.5$	+2 or -2
Fold change magnitude 2	$4 < FC \leq 10$ or $0.01 < FC \leq 0.25$	+4 or -4
Fold change magnitude 3	$10 < FC \leq 50$ or $0.02 < FC \leq 0.1$	+8 or -8
Fold change magnitude 4	$FC > 50$ or $FC \leq 0.02$	+16 or -16

4.2.3.2 Calculation of gene weights

A gene is weighted based on whether concurrent DNA and mRNA expression changes occur. We calculated the weight as the average effect on mRNA of a DNA alteration as observed in >100 multi-omic tumour/non-malignant paired tissues. Specifically, we estimated "the effect" of DNA alterations (i.e. copy number and DNA methylation) on gene expression in 104 tumours and matched normal samples from four different cancer types, for which copy number, DNA methylation, and mRNA data were available for all tumour and normal pairs [54, 112, 196, 235, 236]. Data was acquired from the TCGA (https://tcga-data.nci.nih.gov/tcga/) and the Early Detection Research Network (EDRN, http://edrn.nci.nih.gov/science-data). TCGA data were normalized and processed as described in the TCGA Data Primer (https://wiki.nci.nih.gov/display/TCGA/TCGA+Data+Primer). EDRN data were processed as previously described [54, 196]. The average gene expression fold change in 104 samples was 2.77 times higher for genes with DNA disruption compared to genes without DNA alterations (see Appendix D).

Table 4.3 Weighting system indicates effect of DNA change on mRNA of a gene in a tumour

	WEIGHT
< 2 fold expression Δ	0
> 2 fold expression Δ but no DNA Δ	1
> 2 fold expression Δ w/concordant DNA Δ	2.77

4.2.3.3 Calculation of integrated scores

For every gene (*j*) and tumour (*i*), scores (*S*) and weights (*W*) for DNA (*CN*, *Meth*) and expression (*Exp*) level alterations are combined to form an Integrated Score (*I*):

[Formula 1] $$I_{ij} = S(Exp)_{ij} + W_{ij}\,[S(CN)_{ij} + S(Meth)_{ij}]$$

This alogirthm was run in MATLAB (*version R2012a*, *MathWorks Incorporated*). Genes that sustain a) both high level DNA and mRNA changes or b) high mRNA change only (> 50 fold) with no DNA level change in tumour relative to normal will have the highest Integrated Scores (*I*). Genes that sustain a) DNA but no mRNA changes, or b) < 2 fold mRNA change without DNA level alteration in tumour relative to normal are given a InS of 0. $I > 0$ refer to upregulated and overexpressed genes, and $I < 0$ to downregulated and underexpressed genes.

4.2.3.4 Normalization of integrated scores

Integrated Scores (I_{ij}) were normalized for every gene (*j*) by scaling to +1 to -1, by dividing positive scores by the maximum Integrated Score (Max I_i) and negative scores by the minimum Integrated Score (Min I_i) in each tumour (*i*):

[Formula 2] $$\text{If } I_{ij} > 0\text{: } I_{ij} \div \text{Max } I_i\text{; If } I_i < 0\text{; } I_{ij} \div \text{Min } I_i$$

Normalized Integrated Scores were applied to all subsequent analyses and from this point forward will be referred to as InS.

4.2.4 Dimensional reduction and analysis of gene sets

4.2.4.1 Clustering

Unsupervised dimensional reduction was performed by non-negative matrix factorization (NMF) using the *'NMF'* algorithm in *R* [237]. NMF finds a small number of metagenes, defined as a positive linear combination of genes in the input matrix. The input matrix consisted of absolute values of normalized InS for COPD and non-COPD related lung tumours. Genes with InS of '0' across > 80% of all samples were excluded. NMF groups tumours into clusters based on patterns of metagenes. NMF differs from other component

reduction algorithms, such as PCA, in that it allows components to overlap (as genes would across different signaling pathways) or other clustering algorithms such as hierarchical clustering which imposes a stringent tree structure on data [238]. The Kullback-Leibler divergence based 'brunet' algorithm was applied using the non-negative double singular value decomposition ('NNDSVD') method for initialization, and an optimal factorization rank (of k=2) as estimated by the cophenetic correlation coefficient using 50 random permutations (and r ranges 2:6), as recommended [237].

4.2.4.2 Pathway and gene set enrichment analyses

Gene set enrichment analysis (GSEA) was used to assess whether a given *a priori* set of genes (stored in Molecular Signatures Database (MSigDB) version 4.0) are randomly distributed or primarily found at the top or bottom of a ranked set of genes between two states, in this case COPD *vs* non-COPD related lung tumours, using InS as input (instead of typical gene expression matrices) [239]. Enrichment scores (ES) reflect the degree to which a gene set is overrepresented at the extremes (top or bottom) of the entire ranked list. A nominal p value (NOM p-val) is calculated by permuting (500 times) the phenotype labels (COPD or non-COPD) and re-computing the ES of the gene set for the permuted data based on a null distribution for the ES. Multiple hypothesis testing is accounted for by normalizing the ES gene set size to get a normalized enrichment score (NES) and then applying a Benjamini and Hochberg false discovery rate (FDR) correction to get an FDR q-value [239]. To determine which transcription factor (TF) gene sets were differentially enriched in the COPD tumour group, we assessed the “c3 regulatory motif gene set”, which is a database of conserved cis-regulatory motifs that are known or likely regulatory elements in promoters and 3’-UTRs described by Xie et al. [240]. The number of genes assessed in the gene set after filtering out those not present in the input matrix is indicated by “Size”.

Pathway enrichment analyses was performed using IPA (Ingenuity Pathway Analysis®, www.ingenuity.com), as described in Chapter 2 (section 2.2.9). IPA does not take into account gene ranks within or across groups, so the top 1st percentile of up (InS > 0) and down (InS < 0) regulated genes altered at a frequency of > 15% uniquely in COPD or non-COPD tumour group were used as input for this analysis. Three such analyses were run

composed of genes frequently altered in 1) COPD tumours alone, 2) non-COPD tumours alone and 3) both COPD and non-COPD tumour groups. A cross-comparison analysis of these results was then performed in IPA.

An enrichment analysis for targets of upstream regulators (i.e. transcription factors) was also performed in IPA-- a similar yet independent analysis to the transcription factor gene set enrichment performed in GSEA. p value scores are calculated similarly to canonical pathway enrichment (i.e., Fisher's exact test $-\log_{10}$(p-value)).

4.2.5 Validation strategy

External multi-omics data, including copy number (Affymetrix SNP 6.0 described in section 4.2.2.1), DNA methylation (Illumina HumanMethylation450 BeadChip (HM450K)) and gene expression data (Illumina HiSeq 2000 RNA sequencing Version 2) as well as associated post bronchodilator FEV_1/FVC data were available for 70 lung adenocarcinoma tumours (Table 4.4). All data were obtained from the TCGA (https://tcga-data.nci.nih.gov/tcga/). Molecular data were processed as described for 'Level 3 data' in the TCGA data compendium [133]. Copy number ratios and RPKM expression data for genes of interest were extracted. While the Illumina platforms used to derive whole genome DNA methylation profiles for TCGA tumours (HM450K) differed from the one we used in our discovery set (HM27K) in many respects[241], a large number of the HM27K probes (> 90%) are present on the HM450K array (Type II probes) [242]. Therefore, we extracted from the TCGA methylation dataset, the same probes we were interested in validating from our discovery set and averaged β values within the COPD and non-COPD tumour groups.

There are a scarcity of non-malignant lung tissues in the TCGA cohort (n=1 COPD subject, n= 10 for non-COPD subjects). However given our interest in validating genes specifically altered in our COPD tumour group, we reasoned that one validation approach would be to determine if copy number status, DNA methylation levels and expression fold change in TCGA COPD tumours (n= 16) were significantly different on average from those in non-COPD tumours (n= 54) (Table 4.4). Therefore, permutation tests and multiple comparisons testing correction were performed separately for each 'omics dimension as

described in section 2.2.7. Genes with B-H corrected permutation test p values < 0.05 in i) at least one DNA 'omics comparison (copy number and/or DNA methylation) and ii) mRNA level in concordant directions (as in section 4.2.3.1), were considered differentially altered between COPD and non-COPD tumours.

Table 4.4 TCGA validation lung tumour cohort patient demographics

	COPD	Non-COPD
n=	**16**	**54**
Sex		
Male		
Female	9	31
Average Age	66	66
Lung Function		
FEV_1% Pred	66	86
FEV_1/FVC%	52	88
Smoking		
Current	8	12
Former	8	42
Never	0	0
Average Pack Years	39	42
Stage		
I	11	40
II	1	12
III	3	1
IV	1	0
Ethnicity		
Caucasian	15	47
Asian	1	1
African American	0	6

4.2.6 DNA methylation profiling and pre-processing of small airway epithelia from COPD patients with and without lung cancer

The 1st percentile of up and down regulated genes in COPD-related tumours based on normalized InS, that were also altered at a frequency of >15% only in the COPD tumour group were assessed at the level of DNA methylation in small airway epithelia (SAE) from COPD patients with non small cell lung cancer (n=10), COPD patients without cancer (n=15) and subjects without COPD or lung cancer (n=23) (Table 4.5). SAE cells were collected and processed as described in Chapter 2 (section 2.2.2-2.2.5). Similarly, a non parametric permutation test, using 10,000 permutations described in Section 2.2.7 was also performed, correcting for multiple testing using the Benjamini and Hochberg (B-H) method (B-H p

value < 0.05 was considered significant) between the group with cancer (COPD + lung cancer) compared to each group without cancer (COPD alone and non-COPD, no lung cancer). Probes mapping to COPD tumour genes also had to exceed a stringent fold change threshold of > 4, calculated as the ratio of average M values (described in section 2.2.5) in SAE from COPD patients with lung cancer compared to COPD and non-COPD groups without lung cancer.

Table 4.5 Demographics for DNA methylation small airway cohort

	COPD + LC	**COPD alone**	**non-COPD, no LC**
n =	10	15	23
Age	61±7.80	65±5.76	64±4.8
Female:Male	4:06	5:10	8:15
Pack Years	48.43±14.03	54.77±30.43	46.64±20.53
Years Quit	3.7±6.40	10±9.55	14±5.44
FEV_1act	1.88±0.47	1.79±0.63	3.06±0.68
FEV_1%Pred	60±13.00	58±15.59	98±9.84
FEV_1/ FVC%	58±5.94	58±9.57	75±5.38

4.3 Results

4.3.1 COPD and non-COPD lung tumours cluster differentially

Dimensional reduction by NMF using absolute value InS from all tumours was capable of distinguishing COPD and non-COPD tumour types (Figure 4.1A). Assessment of the distribution of COPD and non-COPD tumours within the largest three clusters revealed enrichment COPD tumours in clusters 2 and 3 compared to cluster 1 (Chi-square test p = 3.39E-06) (4.1B).

Figure 4.1

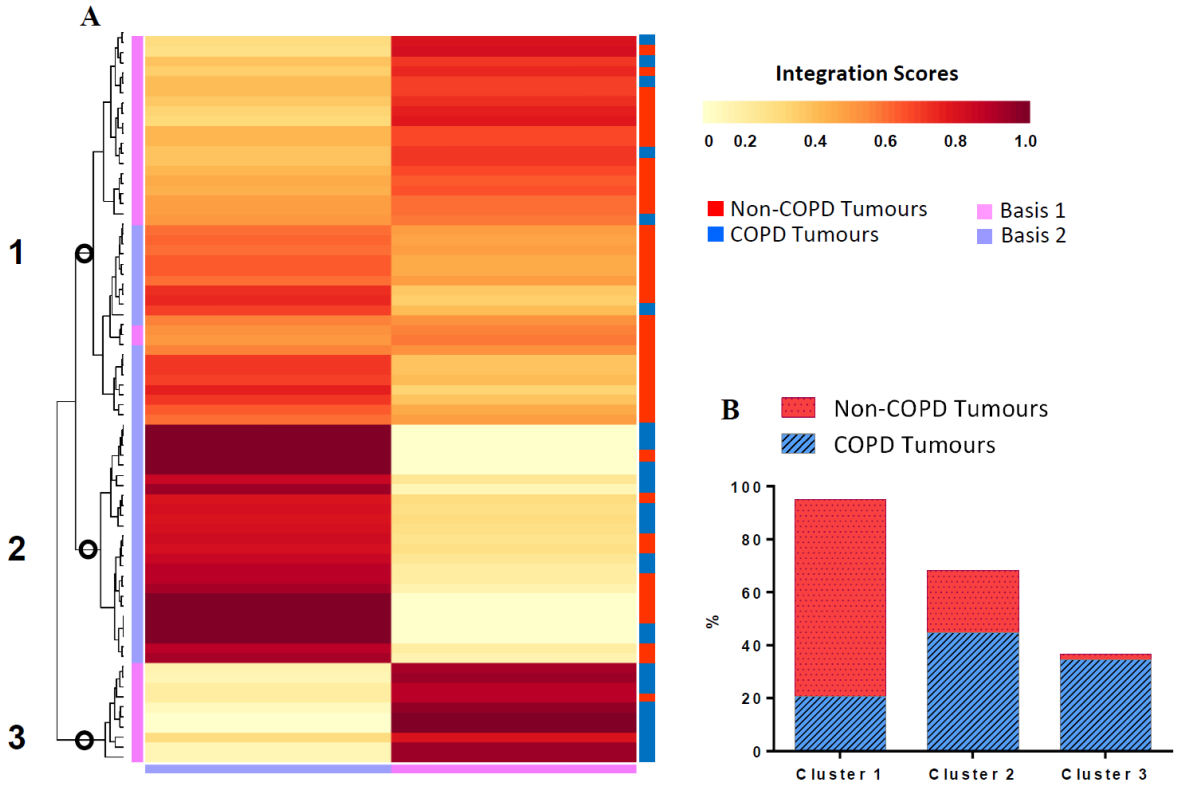

Figure 4.1 NMF clustering of COPD and non-COPD tumour integration scores

A) Absolute values of normalized Integration Scores were applied to non-negative matrix factorization dimensional reduction. Meta-genes with high or low integration scores are indicated by maroon or light yellow coloured boxes in the matrix. Red and blue boxes to right of matrix indicate COPD (blue) or non-COPD (red) related lung tumours. Pink and purple bars refer to meta-gene sets referred to as basis. **B)** Percentage of tumour types in each of three clusters, indicated by bold black circles on cluster plot to the left of matrix in A.

4.3.2 COPD and non-COPD tumours are differentially enriched for distinct transcription factor gene sets

We next assessed whether gene sets corresponding to upstream transcription factors were enriched between COPD and non-COPD tumour groups. Five transcription factor gene sets were significantly (p value < 0.05) and all positively enriched in COPD tumours compared to non-COPD tumours (Table 4.6). Gene sets corresponded to targets of the following transcription factors: *PITX2* (*paired-like homeodomain 2*), *NR2F2* (*nuclear receptor subfamily 2, group F, member 2*), *SP1* (*Sp1 transcription factor*), *PPARG* (*peroxisome proliferator-activated receptor gamma*) and *HNF4A* (*hepatocyte nuclear factor 4, alpha*).

Table 4.6 Transcription factor gene sets enriched in COPD-related lung tumours

Headers defined in Section 4.2.4.2

TF	MSigDB TF Targets Name	SIZE	ES	NES	NOM p-val	FDR q-val
PITX2	V$COUP_DR1_Q6	211	3.58E-01	1.61E+00	0.00E+00	3.98E-01
NR2F2	V$DR1_Q3	213	3.29E-01	1.53E+00	7.59E-03	5.94E-01
PPARG	V$PPARG_01	41	4.31E-01	1.52E+00	2.14E-02	5.37E-01
SP1	V$SP1_Q6	231	3.23E-01	1.45E+00	2.28E-02	6.91E-01
HNF4A	V$HNF4_DR1_Q3	221	3.19E-01	1.41E+00	3.20E-02	5.32E-01

4.3.3 Genes recurrently altered in tumours from COPD and non-COPD patients

We were next interested in determining which biological pathways may be enriched in genes uniquely, frequently and highly disrupted in COPD or non-COPD related tumours. For each tumour and each tumour group we first determined the top 1st percentile of up and down regulated genes per tumour (based on normalized InS), altered at a frequency of > 15% in either COPD or non-COPD groups. Genes were aligned by symbol and those fulfilling the above criteria, altered in the same direction in each tumour group were considered "shared". Genes fulfilling these criteria and only appearing in one tumour group were considered unique to that group (Figure 4.2).

Figure 4.2

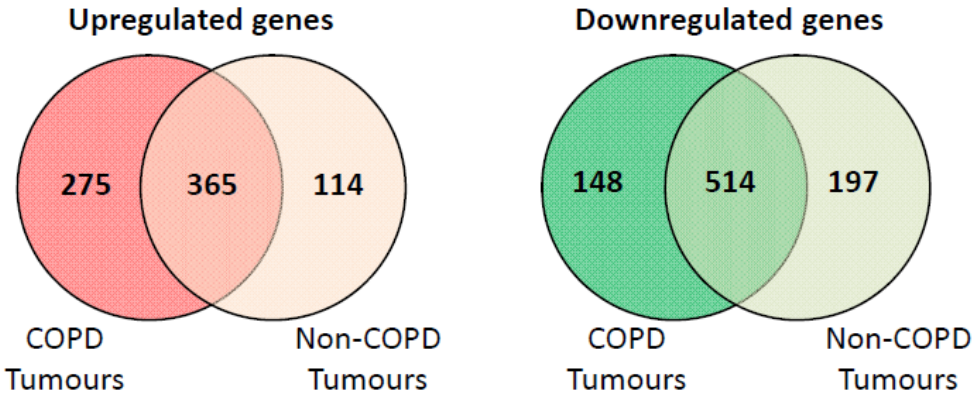

Figure 4.2 Frequently up- and downregulated genes in COPD and non-COPD related tumours

4.3.4 COPD and non-COPD tumours are differentially enriched for distinct canonical pathways involved in inflammation, DNA damage and metabolism

To further evaluate the biological significance of these findings we assessed which canonical pathways were enriched in the most highly and frequently uniquely and commonly disrupted genes in COPD and non-COPD tumours. In COPD tumours, we observed striking enrichment of genes involved in inflammation including (atherosclerosis signaling, retinoic acid response and activation, liver fibrosis, glucocorticoid regulation, IL-17A signaling), DNA damage (DNA damage-induced 14-3-3σ signaling) and metabolic pathways involving pyrimidine (thymine, cytosine and uracil) salvage and critical to epigenetic maintenance of

DNA and histone methylation marks (S-adenosyl-L-methionine biosynthesis) in the most highly altered gene sets in COPD tumours (Figure 4.3).

Figure 4.3

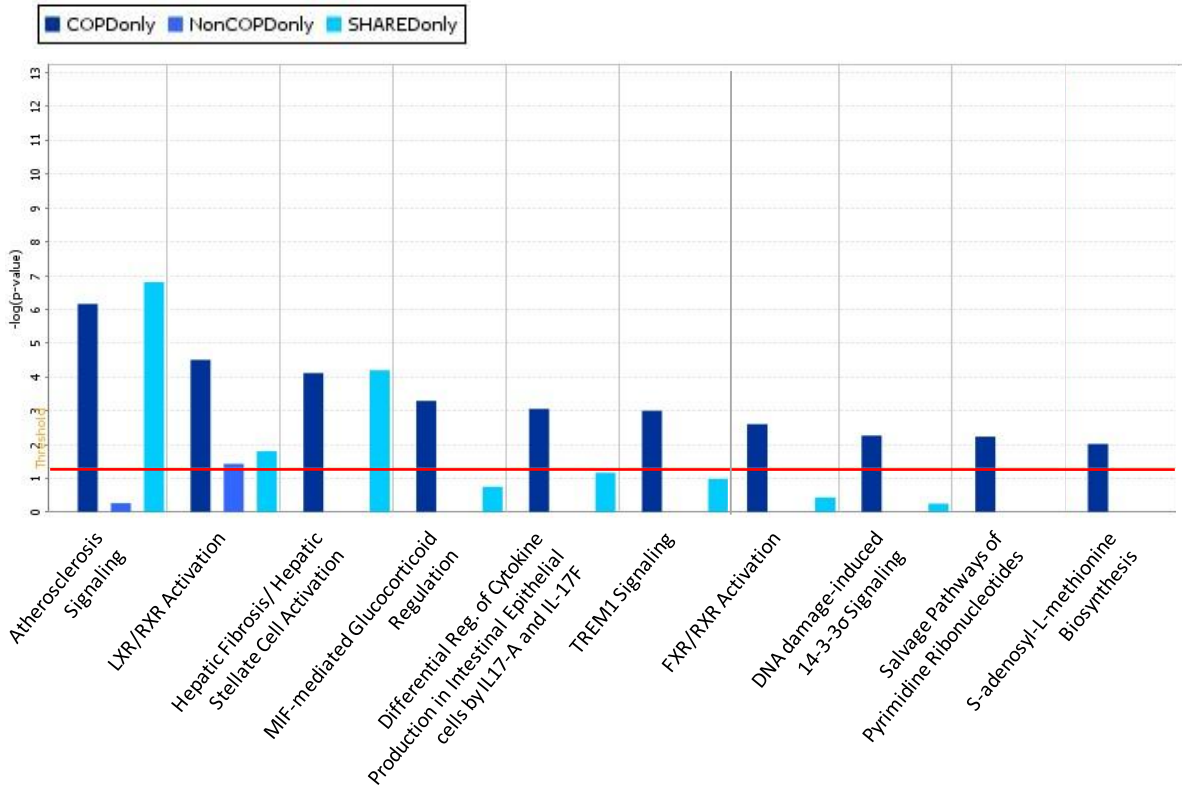

Figure 4.3 Pathways enriched in top disrupted gene sets from COPD and non-COPD related tumours

Canonical pathways enriched in the most highly and frequently altered genes (from figure 4.2) uniquely altered in COPD (navy blue bars) and non-COPD (light blue bars) tumours, as well as those enriched in genes commonly altered (i.e. shared) between tumour groups (aqua bars), were calculated in IPA. A p value = 0.05 is indicated by the horizontal red line. Bars above this line are considered significantly enriched in the corresponding tumour group.

Table 4.7 Genes highly altered in lung tumours enriched in atherosclerosis signaling

p values for pathway enrichment are indicated next to each group header

Total number of genes tested for each group is indicated in Figure 4.2

Symbol	Entrez Gene Name	Dir. in Tumour
COPD ONLY (p = 6.89E-07)		
CCL2	chemokine (C-C motif) ligand 2	DOWN
CD40	CD40 molecule, TNF receptor superfamily member 5	DOWN
IL8	interleukin 8	DOWN
IL1A	interleukin 1, alpha	DOWN
IL1B	interleukin 1, beta	DOWN
PLA2G3	phospholipase A2, group III	DOWN
PLA2G2A	phospholipase A2, group IIA (platelets, synovial fluid)	DOWN
RBP4	retinol binding protein 4, plasma	DOWN
SELE	selectin E	DOWN
SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	DOWN
TNFSF12	tumour necrosis factor (ligand) superfamily, member 12	DOWN
APOE	apolipoprotein E	UP
PLA2G12B	phospholipase A2, group XIIB	UP
NON-COPD ONLY (p = 5.35E-01)		
APOD	apolipoprotein D	DOWN
MMP1	matrix metallopeptidase 1 (interstitial collagenase)	UP
SHARED (p = 1.56E-07)		
ALOX5	arachidonate 5-lipoxygenase	DOWN
ALOX15	arachidonate 15-lipoxygenase	DOWN
CD36	CD36 molecule (thrombospondin receptor)	DOWN
CXCL12	chemokine (C-X-C motif) ligand 12	DOWN
CXCR4	chemokine (C-X-C motif) receptor 4	DOWN
IL6	interleukin 6 (interferon, beta 2)	DOWN
IL33	interleukin 33	DOWN
LPL	lipoprotein lipase	DOWN
LYZ	lysozyme	DOWN
MSR1	macrophage scavenger receptor 1	DOWN
PLA2G1B	phospholipase A2, group IB (pancreas)	DOWN
S100A8	S100 calcium binding protein A8	DOWN
SELP	selectin P (granule membrane protein 140kDa, antigen CD62)	DOWN
COL10A1	collagen, type X, alpha 1	UP
COL1A1	collagen, type I, alpha 1	UP
IL37	interleukin 37	UP
MMP9	matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)	UP
MMP13	matrix metallopeptidase 13 (collagenase 3)	UP
PLA2G2D	phospholipase A2, group IID	UP
TNFSF14	tumour necrosis factor (ligand) superfamily, member 14	UP

One pathway enriched only in the top disrupted genes unique to COPD tumours was the IL-17A and F signaling pathway, which we also detected as aberrantly methylated in non-malignant airways in the context of COPD in the absence of cancer (Figure 2.8 and2 and Figure 4.4).

Figure 4.4

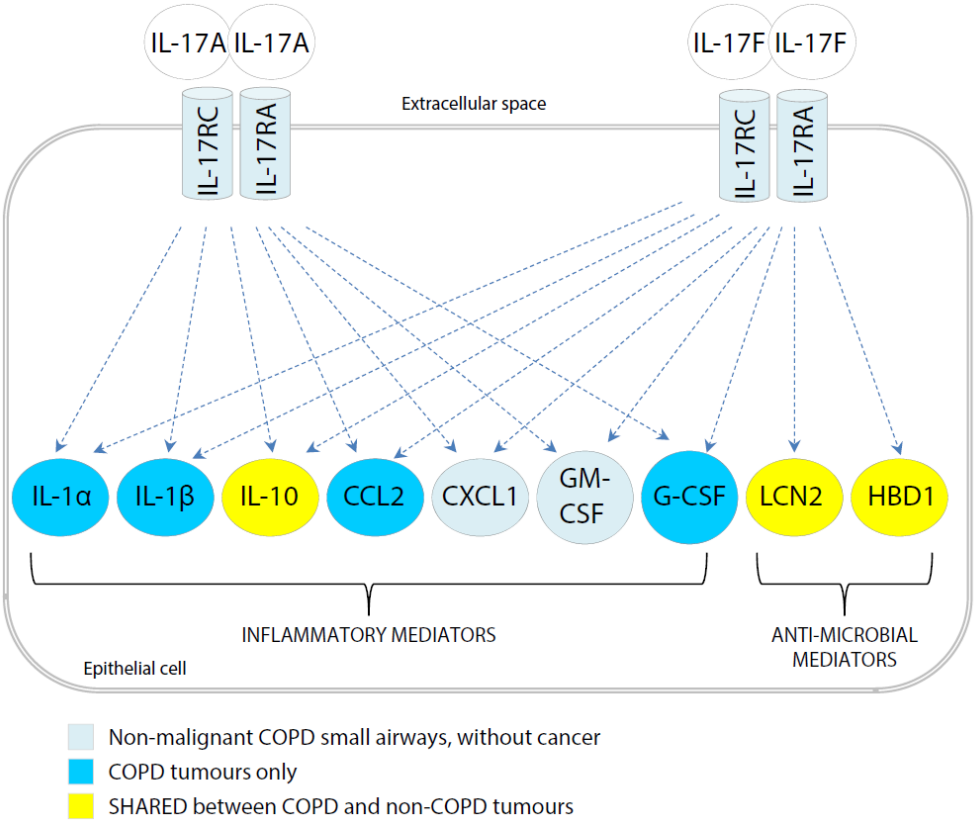

Figure 4.4 IL-17A and IL-17F signaling pathway is disrupted uniquely in airways of COPD patients without cancer and differentially in COPD and non-COPD lung tumours

The IL-17A and F signaling pathway was significantly enriched in the most highly disrupted genes in COPD tumours (bright blue molecules). Genes highly altered in both tumour groups are yellow. This pathway was also significantly enriched in genes altered at the level of DNA methylation and mRNA expression in non-malignant small airway epithelia from COPD patients without lung cancer (light blue molecules). All coloured molecules are downregulated in respective groups, except *GM-CSF*, *IL-10* and *LCN2*. Gene symbols: *IL*: *interleukin*; *IL-17RC*: *interleukin 17 receptor C*; *IL-17RA*: *interleukin 17 receptor A*; *CCL2*: *chemokine (C-C motif) ligand 2*; *GM-CSF*: *granulate-macrophage stimulating factor*; *G-CSF*: *colony stimulating factor 3*; *LCN2*: *lipocalin 2*; *HBD1*: *defensin, beta 1*.

4.3.5 COPD-related lung cancer genes associated with and without smoking status

Since our COPD tumour cohort was heavily biased towards smokers (Table 4.1), and smoking is associated with distinct and clinically relevant molecular features in lung adenocarcinoma (Chapter 1, 3 and [50, 54]) we were interested in determining which of the genes we deemed uniquely and highly altered in our COPD tumour group were also associated with i) smoking regardless of COPD status and ii) COPD excluding never smokers. For each comparison, we performed a similar analysis described in section 4.2.4.3, except the top 1st percentile of up (InS > 0) and down (InS < 0) regulated genes altered at a frequency of > 15% was calculated for tumours from i) ever (CS and FS) (n= 45) and never (NS) (n= 28) smokers, irrespective of COPD status and ii) for ever smoker COPD (n= 25) and ever smoker non-COPD (n= 20) subjects (i.e. excluding NS).

4.3.5.1 Overlap of COPD and smoking associated genes

A large number of genes fulfilled our criteria for "smoking-status specific" (Figure 4.5), of which a large number overlapped with those we deemed "COPD and lung cancer related" (Table 4.8). Therefore, it is possible that cigarette smoking alone accounts for approximately 40% of our uniquely upregulated COPD genes, and over 50% of our uniquely downregulated COPD genes. Of interest, it is also possible that 55% and 50% of our unique up and downregulated non-COPD tumour genes are related to never smoker tumour biology.

Figure 4.5

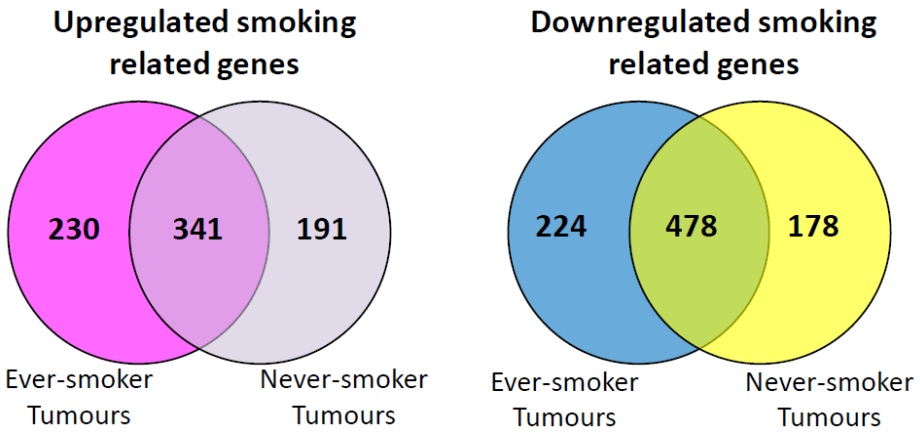

Figure 4.5 Frequently up- and downregulated genes in ever and never smoker related tumours regardless of COPD status

4.3.5.2 Analysis of COPD and non-COPD tumours from ever-smokers

We next re-analyzed our dataset after removing NS. While this greatly reduced our sample size, we detected a number of genes as uniquely and commonly disrupted in COPD and non-COPD tumours (see row 2 in Table 4.8). Of these 69% overlapped with those detected in the initial analysis that included NS. While only 24% of the genes identified as uniquely disrupted in COPD tumours in the ever smoker (ES) cohort overlapped with those detected as uniquely altered in COPD tumours in the analysis that included NS; the most significantly enriched pathways in gene sets derived from our ES COPD vs ES non-COPD analysis were strikingly similar to results derived from gene sets from our COPD vs non-COPD analysis (which included NS). In each separate pathway analyses, two out of three of the most highly enriched pathways were atherosclerosis signaling and hepatic fibrosis. These pathways also displayed the same patterns of enrichment depicted in Figure 4.3: i.e., atherosclerosis signaling was significant in ES COPD unique (p= 5.5E-06) and ES non-COPD unique (p= 3.94E-06) gene sets, whereas hepatic fibrosis was only significant in the ES COPD unique gene set (p= 1.77E-05) but not the ES non-COPD unique gene set (p= 0.9E-02). Moreover, almost all (95%) of the genes that overlapped between our smoking specific and ES COPD vs ES non-COPD analysis (Table 4.8, row 5) overlapped with those we deemed associated with smoking in our COPD vs non-COPD analysis which included NS (Table 4.8, row 6).

Therefore, while our discovery cohort was biased towards ES (i.e., current and former smokers) in the COPD tumour group and never smokers in the non-COPD tumour group, our results indicate that the biology related to cigarette smoke response is likely important to both COPD and lung cancer biology.

Table 4.8 Genes frequently upregulated in COPD and non-COPD related tumours and overlap with smoking specific results

	Genes freq UP regulated in tumours			Genes freq DOWN regulated in tumours		
Tumour group comparison	COPD unique	Shared	Non-COPD unique	COPD unique	Shared	Non-COPD unique
1. All COPD vs All non-COPD	275	365	114	148	514	197
2. ES COPD vs ES non-COPD	60	77	117	76	135	123
3. Overlap ES and All results	20	75	15	12	133	16
4. Overlap smoking specific and All	112 (CS)	60	63 (NS)	76 (CS)	437	98 (NS)
5. Overlap smoking specific and ES results	18 (CS)	60	1 (NS)	16 (CS)	128	1 (NS)
6. Overlap smoking specific, All and ES results	11 (CS)	60	1 (NS)	11 (CS)	128	1 (NS)

All: analysis includes current, former and never smokers; ES: analysis includes current and former smokers; All COPD vs All non-COPD: analysis described in methods section 4.2 using entire discovery cohort in Table 4.1 (includes never smokers); ES COPD vs ES non-COPD: analysis described in section 4.3.5 in lung tumours from ever-smokers with and without COPD; Smoking specific: analysis described in section 4.3.5 in lung tumours from ever-smokers compared to never smokers (regardless of COPD status); Overlap: genes within top percentile InS, altered > 15% in associated group occurring in the same direction in all analyses indicated.

4.3.6 Validation of findings in external datasets

Since we were most interested in validating the 423 genes uniquely and highly disrupted in our COPD tumour group (275 upregulated and 148 downregulated genes indicated in Figure 4.2) we reasoned that one approach would be to determine if copy number status, methylation levels and expression fold change in TCGA COPD tumours (n= 16) were significantly different on average from those in non-COPD tumours (n= 54) (Table 4.4) using the strategy described in section 4.2.5. After processing of TCGA Level 3 data described in the TCGA data compendium, we were able to assess copy number status, DNA methylation and expression levels of 356 out of the 423 genes of interest. However, of these only 34 genes (~10%) validated as differentially altered at the expression level and by at least one DNA level. At the level of gene expression alone, 81 genes (23%) were differentially expressed between TCGA COPD and non-COPD tumours in the same direction as in our discovery cohort. While this emphasized the i) preliminary nature of our findings and ii) the

importance of validating our results in future studies, it also emphasized certain strengths of our cohort related to the size and availability of clinical information.

4.3.7 A subset of genes disrupted in COPD related tumours are hypermethylated in airways of patients with COPD and lung cancer

To assess the potential clinical relevance of genes possibly important to COPD-related lung tumourigenesis, we assessed the methylation level of genes uniquely disrupted in COPD tumours, in small airway epithelia (SAE) from COPD patients with and without lung cancer, and non-COPD patients without lung cancer (Table 4.5). Of the 339 genes, five genes passed our criteria for "differentially methylated" in SAE from COPD patients with lung cancer, in the same direction as in COPD tumours compared to subjects with and without COPD, but with no lung cancer (Table 4.9). While we only assessed a small number of SAE tissues from COPD patients with lung cancer, and did not include profiles from lung cancer patients with lung cancer but without COPD, these results have intriguing clinical implications and highly relevant biological functions.

Table 4.9 COPD-related lung tumour genes altered by DNA methylation in airways of lung cancer patients with COPD

Symbol	Alter. in COPD Tumor	Meth Status LC+COPD SAE	SAE LC+COPD vs COPD-alone B-H p val	SAE FoldΔ LC+COPD/ COPD-alone	SAE LC+COPD vs No Disease B-H p val	SAE FoldΔ LC+COPD/ No Disease
CCNDBP1	DOWN	HYPER	8.25E-05	4.58	0.00E+00	8.81
DUT	DOWN	HYPER	1.42E-02	4.47	4.78E-06	39.22
HPGD	DOWN	HYPER	1.48E-02	4.23	3.67E-13	5.51
MAT2B	DOWN	HYPER	1.53E-03	4.97	0.00E+00	12.15
PPARGC1A	DOWN	HYPER	1.04E-02	5.01	8.19E-14	14.43

LC: lung cancer; SAE: small airway epithelia; B-H p val: Benjamini-Hochberg corrected p value; Alter. In COPD Tumour: Direction of InS in COPD tumour group; Meth Status LC+COPD SAE: direction of methylation difference in small airways from subjects with LC and COPD compared to COPD and non-COPD subjects without LC; No Disease: subjects without COPD or LC. Gene Symbols: CCNDBP1: *Cyclin D-type binding-protein 1*; DUT: *Deoxyuridine triphosphatase*; HPGD: *Hydroxyprostaglandin dehydrogenase 15-(NAD)*; MAT2B: *Methionine adenosyltransferase 2B*; PPARGC1A: *Peroxisome proliferator-activated receptor gamma, coactivator 1α*

4.1 Discussion

Due to the scarcity of 'omics tumour data from cohorts with annotated lung function measurements, little is known about the molecular biology of COPD-related lung tumours, or if these tumours differ from non-COPD related lung tumours of the same subtype. Since chronic inflammation is causally associated with cancer development in a variety of tissue types including the lung in the context of COPD, we hypothesized that lung tumours arising in an environment of COPD-- a chronic inflammatory lung disease, would harbour distinct and clinically relevant molecular alterations compared to lung tumours from non-COPD patients. Since tumour systems are altered at multiple 'omic levels, interrogation and integration of multi-omics data is conducive to the elucidation of mechanisms driving cancer biology. We applied such an approach to the analysis of lung adenocarcinoma tumours from patients with and without COPD. To evaluate the potential clinical application of our findings, we assessed whether genes altered in COPD-related lung tumours were also disrupted at the level of DNA methylation in small airway epithelia from patients with COPD and non small cell lung cancer (NSCLC). We provide the first 'omics interrogation of lung tumours in the context of COPD to date, as well the first study assessing the methylation status of COPD-related lung tumour genes in airways of COPD patients with lung cancer.

Our integrative multi-omic analytical approach was based on the notion that since DNA is a heritable molecule propagated through cellular divisions: 1) cancer cells select DNA alterations that confer a clonal expansion advantage, thus genes with DNA level changes are more likely to be biologically relevant than genes only altered at the mRNA level; 2) cellular maintenance of high level events is metabolically expensive, therefore larger magnitude alterations may be indicative of selection; and 3) biologically relevant DNA level alterations will likely be accompanied by consequential mRNA expression changes. We applied these hypotheses to the generation of an algorithm which integrates DNA methylation, copy number and mRNA data for paired tumour and non-malignant tissues and yields an "Integration Score" (InS) for every gene on a per tumour basis based on: the i) magnitude of DNA level alteration relative to a matched non-malignant sample, and ii) presence of both DNA and mRNA level alterations within a tumour. We applied InS to the identification of genes and pathways important to COPD-related lung tumours.

Overall, InS were capable of broadly distinguishing COPD and non-COPD tumours by clustering. Since clustering was based on InS, and high InS are those that sustain high-level DNA and RNA disruptions; distinct clusters based on these scores may be indicative of unique biological processes underlying COPD-related lung tumourigenesis. Indeed, enrichment analysis of transcription factor gene sets in InS matrices revealed enrichment of well known mediators involved in lung cancer, inflammation and other malignancies in COPD compared to non-COPD tumours. These included genes previously found to be associated with COPD (*PITX2*), multiple oncogenic processes in lung and other malignancies (*NR2F2*, *SP1*, *PPARG*) and inflammation-associated fibrosis and cancer of the liver (*HNF4A*), which were all positively enriched in COPD compared to non-COPD tumours [243-246]. Only one of these (*PITX2*) was significant after correcting for multiple testing. *PITX2* is a putative methylation biomarker in lung cancer, previously associated with senescence gene networks in COPD [247].

At the gene level, we found that overall, more genes were commonly altered between COPD and non-COPD related tumours than were unique to any group suggesting that common biological mechanisms likely underlie lung cancer biology between these groups in general. However, a large number of genes, particularly upregulated genes, were uniquely disrupted in COPD tumours which may also be indicative of selection of biological processes unique to COPD-related lung cancer. We attempted to validate these findings in external lung datasets, however lung tumour cohorts with both ‘omics and patient lung function data are extremely rare. The TCGA consortium is the largest such dataset to date; of the approximately 500 lung adenocarcinoma tumours with available ‘omics data (copy number, DNA methylation and gene expression) accrued as part of this consortium, only 70 subjects have post-bronchodilator FEV_1/FVC airflow measurements, of which only 16 have COPD based on GOLD guidelines outlined in Table 1.1. Even rarer are such datasets that include non-malignant lung tissues from lung cancer patients; TCGA includes only 1 non-malignant sample from a single COPD patient with lung cancer. Non-malignant ‘omics data is particularly important for assessing tissue-specific aberrant events related to DNA methylation and gene expression. Therefore we were only able to assess whether our genes

of interest were different between COPD and non-COPD TCGA tumour groups at the DNA and mRNA levels, but were unable to determine whether these genes were differentially methylated or expressed in tumours compared to non-malignant tissues in each group. Using this approach only 10% of the genes we found validated. The lack of tumour 'omics cohorts with associated lung function measures highlights the uniqueness of our own cohort, although emphasizes the need to validate our results in future studies.

There were multiple, potential confounding factors in our dataset. For example, our COPD tumour cohort was heavily biased towards smokers, and our non-COPD cohort towards never smokers (Table 4.1). We observed substantial overlap among gene sets derived from analyses between 1) all COPD and non-COPD related lung tumours (i.e. including NS subjects), 2) lung tumours from ever smokers compared to NS, regardless of COPD status and 3) COPD and non-COPD related tumours from ever smokers (excluding NS). While we note that our primary analysis is likely confounded by the biological effects of smoking, we also noticed overlap in gene sets and enriched pathways derived from our COPD and non-COPD analyses which did and did not include NS (Table 4.8), indicating that genes altered in response to cigarette smoke could be involved in COPD related lung cancer. Indeed, strong epidemiological and genetic association studies link cigarette smoke to both diseases [248-251], therefore we did not remove smoking-related genes from our subsequent analyses.

Pathway enrichment analysis of the top percentile of genes altered uniquely and commonly in COPD and non-COPD tumour groups revealed striking enrichment of functions involved in inflammatory response (FXR/RXR activation, MIF-mediated glucocorticoid regulation, IL17-A and IL-17F signaling and TREM1 signaling), DNA damage (DNA damage-induced 14-3-3σ signaling) and metabolism (Pyrimidine ribonucleotide salvage pathways and S-adenosyl-L-methionine biosynthesis) in gene sets uniquely altered in COPD tumours. COPD tumour genes and genes shared between tumour groups were commonly associated with the most significantly enriched pathways overall including atherosclerosis signaling, LXR/RXR activation and hepatic fibrosis which are involved in chronic inflammatory disease and inflammatory response. This indicated to us

that while disruption to these pathways is common in both tumour groups, these processes may be further deregulated in COPD tumours, possibly at different levels.

The most significantly enriched pathway overall in the most deregulated tumour genes, was atherosclerosis signaling. Atherosclerosis -- a condition in which artery walls thicken as the result of a build-up of fatty materials such as cholesterol, is a chronic inflammatory process stemming from interactions between cells (macrophages, endothelial, smooth muscle and T cells), plasma lipoproteins and arterial wall extracellular matrix (ECM) [252]. Cardiovascular disease (heart attacks, strokes and peripheral vascular disease) commonly caused by atherosclerosis, is the leading cause of death amongst COPD patients [253]. Genes uniquely and commonly altered in this pathway in each group are listed in Table 4.7. Multiple genes in this pathway have been well described in COPD notably, *SERPINA1* defects which cause α1-Antitrypsin (A1AT) deficiency) -- a hereditary form of COPD, and among the most frequently downregulated genes unique to COPD related tumours. Overexpression and SNPs in matrix metallopeptidases are frequently described in COPD [38, 254, 255]. We detected MMP9 and MMP13 commonly disrupted among both tumour groups, whereas MMP1 was only disrupted in non-COPD tumours. COPD and heart disease share common risk factors, however airflow limitation is an independent risk factor for cardiovascular diseases [253]. Understanding the mechanisms linking atherosclerosis and COPD with the goal of reducing cardiovascular risk in COPD patients through targeted therapies is a major field of study [253]. An improved understanding of the molecular biology underlying COPD pathogenesis may contribute to this aim.

The IL-17A and IL-17F signaling pathway was significantly enriched in genes highly disrupted uniquely in COPD tumours (Figure 4.4). This pathway was also aberrantly disrupted at the level of DNA methylation and gene expression in non-malignant SAE from COPD subjects without lung cancer (Chapter 2). In addition to the inflammatory and anti-microbial functions indicated in Figure 4.4, polymorphisms associated with multiple effector molecules in this pathway are promising predictors of cancer risk in multiple cancers, particularly in the context of inflammation associated colorectal and gastric cancer [256,

257]. Further exploration of this pathway in COPD model systems may yield mechanistic insight into the biology underlying COPD progression and lung risk in COPD patients.

Two independent analyses of our dataset indicated enrichment of genes in COPD tumours involved in liver cirrhosis and inflammation-related cancer of the liver; transcription factor gene set enrichment utilized complete gene matrices of InS from COPD and non-COPD tumours (Table 4.6), whereas pathway enrichment was performed only on the top percentile of altered genes in each tumour group (Figure 4.3). HNF4A controls diverse metabolic functions and is highly expressed in the liver, kidney, pancreas and intestine [258]. A role for HNF4A has been described in alveolar differentiation where it is regulated by DNA methylation [259], in mucinous adenocarcinoma lung cancer [260, 261] and overexpression of *HNF4A* has recently implicated in emphysema [262]. While the mechanistic links between COPD and lung cancer have only recently begun to be explored, the tumour-promoting effects of chronic inflammation in other organs are better understood [263]. In the liver, viral infection and chronic, excessive exposure to alcohol, sugar, fat and cigarette smoke induces expression of pro-inflammatory cytokines, leading to accumulation of ECM proteins (mainly collagen), tissue remodeling, fibrosis and eventually liver cirrhosis which is associated with an up to a 55 fold increased risk of hepatocellular carcinoma (HCC). HNF4A is a key regulatory transcription factor involved in this process. Of note, *α1-Antitrypsin* (*A1AT*) deficiency -- the cause of hereditary COPD, is also associated with liver disease, cirrhosis and HCC and accounts for a high proportion of liver transplants in children [264]. Our data provide further rationale for exploring this pathway in the context of COPD-related lung cancer.

To assess the potential clinical relevance of our findings we assessed our results for presence in small airway epithelia (SAE) tissues from COPD patients with and without lung cancer. Since our cohort was small, and we did not have SAE material from lung cancer patients without COPD, we applied stringent criteria to our query. The biological functions of the five genes we detected as aberrantly methylated in SAE from lung cancer patients appear highly relevant to COPD and lung cancer biology. For example, *CCNDBP1* is a tumour suppressor gene (TSG) in breast, prostate and colon cancer that interacts with the

class III histone deacetylase, *sirtuin 6* (*SIRT6*) [265]. Reduced expression of sirtuins is implicated in aging, senescence and cancer [266]. *SIRT6* is a TSG, involved in protecting cells from senescence, telomere dysfunction and DNA damage [267], is underexpressed in COPD lungs, and mediates cigarette smoke induced senescence in bronchial epithelial cells [268]. SNPs in *HPGD* are associated with colon cancer risk. Genes involved in metabolic processes included *DUT*, an essential nucleotide metabolism enzyme, and *MAT2B* -- the enzyme that catalyzes synthesis of S-adenosylmethionine (SAM) which is the cell's primary methyl donor for histone and DNA methyltransferases. Disruption of *MAT2B* could have profound implications to the maintenance of normal epigenetic patterns at both the DNA and histone level, in tumour and non-malignant tissues. MAT2B also interacts with SIRT1, another sirtuin previously implicated in aging, oxidative stress, senescence and COPD [269, 270]. *PPARGC1A* is a central inducer of mitochondrial biogenesis in cells that increases expression of ROS-detoxifying enzymes [271], and intriguingly, decreased expression of PPARGC1A has been previously correlated with increased COPD severity [272].

Our results support the notion that lung tumourigenesis in COPD patients exhibit distinct molecular aberrations compared to lung cancer in patients without COPD, and intriguingly, may be related to tumour-inducing mechanisms of chronic inflammation in other organs. While our discovery cohort was limited in sample size and validation of these findings in appropriate COPD and lung tumour models [273] are necessary, lung tumour 'omics cohorts with annotated lung function measures are extremely scarce in the public domain and our study represents the largest 'omic survey of COPD-related lung tumours performed to date. Our findings provide rationale for exploring the molecular biology of COPD-related lung cancer in future studies. Moreover, genes altered in non-cancerous airway cells that are associated with lung cancer in COPD patients warrant further study as lung cancer risk biomarkers or as targets for novel chemoprevention therapies in this high risk population. Validation of this aspect is of particular clinical relevance, and may be performed in prospectively collected cohorts or in banked specimens with longitudinal sampling and monitoring.

5 Chapter: Conclusions

5.1 Summary of thesis findings

The goal of this thesis work was to uncover the genetic and epigenetic mechanisms underlying COPD biology, related to smoking response in lung tumours and non-malignant tissues and the mechanistic links between COPD and lung cancer. Collectively, this work revealed that **1)** DNA methylation is a mechanism involved in processes important to COPD biology in small airways (Chapter 2) and **2)** different smoking histories are associated with common and divergent miRNA expression patterns in lung tumour and non-malignant tissues, a subset of which are associated with patient outcome (Chapter 3). We applied an integrative, multi-omics strategy to the discovery of **3)** genes and pathways involved in COPD-related lung cancer (Chapter 4), and showed that **4)** a number of these genes could be detected in non-malignant airways of patients with lung cancer and COPD (Chapter 4).

Taken together these findings provide evidence supporting our overarching hypothesis: that specific risk factors, such as smoking and chronic inflammation lead to selective DNA level disruption of genes in exposed tissues and that these selectively disrupted genes likely contribute to COPD and lung cancer biology. Section 5.2 states the potential clinical significance and summarizes these findings in the context of the research goals stated in Section 1.9 of this thesis.

5.2 Conclusions regarding thesis hypotheses

5.2.1 DNA methylation is globally disrupted and associated with expression changes in COPD small airways

DNA methylation is a tissue specific, reversible gene regulatory mark, associated with development and progression of a wide spectrum of diseases, including malignant and non-malignant respiratory disease. Since small airways are the primary sites of airflow obstruction in COPD, we believe that an integrated DNA and RNA level characterization of this tissue is highly relevant to understanding how pathways underlying airflow obstruction are disrupted at the molecular level, and is a critical first step in the development of novel

therapeutic interventions for COPD. To our knowledge this is the first genome-wide assessment of DNA methylation disruptions in COPD small airways.

In **Chapter 2**, we discovered disruption to normal DNA methylation patterns was widespread in small airway epithelia and that a subset of methylation alterations was associated with corresponding changes to expression of key genes and pathways important to COPD pathology, such as the NRF2 pathway-- the cell's primary oxidative response protective mechanism. Since DNA methylation is tissue specific, reversible and may underlie disease-specific gene expression changes, the characterization of these events in small airways may be a critical first step towards development of novel treatment strategies or re-appropriation of existing epigenetic based drugs to treat or prevent COPD.

5.2.2 Smoking status impacts miRNA mediated prognosis and lung tumour biology

Distinct smoking and non-smoking environments are associated with disparate and clinically relevant molecular, epidemiological and clinical features, such that lung cancers in current and never smokers are widely accepted to be different diseases. Due to the success of smoking cessation programs in developed countries, smoking incidence continues to drop, and consequently the proportion of lung cancer patients who are former and never smokers is growing. Thus, it is increasingly imperative to understand the molecular biology of lung cancer in relation to distinct smoking histories.

Given the importance of microRNAs to tumourigenesis and mediating biological response to tobacco smoke, in **Chapter 3** we sought to investigate the contribution of microRNA disruption to lung tumour biology and patient outcome in the context of smoking status. We 1) performed miRNA transcriptome sequencing on 188 lung tissues comprised of 94 tumours and 94 paired non-malignant samples from current, former and never smoker lung adenocarcinomas, 2) derived smoking-status driven miRNA-mediated gene networks, and 3) performed the first smoking-status specific miRNA prognostic association study. We discovered that miRNA expression patterns in lung tumours are dependent upon smoking status, consequently disrupting distinct cellular pathways and associated with patient prognosis in current, former and never smokers. These findings provide novel insight into

how smoking status affects miRNA expression and impacts lung tumour biology and patient prognosis.

5.2.3 Lung tumours from patients with COPD are molecularly distinct at the genetic and epigenetic levels

Chronic inflammation is causally associated with cancer development in a variety of tissue types including the lung. We hypothesized that lung tumours arising in an environment of COPD-- a chronic inflammatory lung disease, would harbour distinct and clinically relevant molecular alterations compared to lung tumours from non-COPD patients. Since tumour systems are altered at multiple 'omic levels, interrogation and integration of multi-omics data is conducive to the elucidation of mechanisms driving cancer biology. In **Chapter 4**, we developed and applied such an approach to the analysis of lung adenocarcinoma tumours from patients with and without COPD. Multi-omics profiling (copy number, DNA methylation and gene-expression) was performed on 76 tumour and non-malignant lung tissues using genome-wide array platforms. To identify genes/pathways which sustain high DNA and RNA level changes in tumours, we applied our integrative multi-omics algorithm (section 4.2.3), which generated Integration Scores (InS) for each gene based on magnitude of concomitant DNA and mRNA alterations. InS were applied to all downstream analyses.

Overall, our analysis revealed that genes frequently and differentially disrupted at both DNA and RNA levels in COPD tumours were differentially enriched for transcription factor targets involved in inflammation-related cancer, oxidative stress and smoking response; the most significant being *PITX2*-- a putative methylation biomarker in LC, previously associated with senescence gene networks in COPD. Pathway analysis of the top percentile of disrupted genes in each tumour group revealed an enrichment of pathways involved in chronic inflammatory disease (atherosclerosis and hepatic fibrosis), DNA damage and key metabolic processes involved establishment and maintenance of epigenetic patterns. One of the pathways (IL-17A and –F signaling) we detected as significantly enriched for aberrantly methylated genes in airways from COPD patients without lung cancer in Chapter 2 was enriched in genes exclusively and highly altered in COPD-related lung tumours. Our findings also support the notion that lung tumourigenesis in COPD patients

share common and divergent lung tumourigenic processes. This is the first multi-omic characterization of lung tumours in the context of lung function. Taken together our findings provide mechanistic insight into lung tumour biology associated with inflammation and COPD biology. The further assessment of these findings in larger cohorts and in relation to tumours from inflammation-associated tumours in other organs (e.g. the colon or liver) may yield important insight in the biology underlying the link between chronic inflammation and cancer. This could have broad implications to the development of prevention and early detection regimes for multiple malignancies and chronic conditions.

5.2.4 Genes preferentially altered in COPD-related lung tumours are aberrantly methylated in non-malignant airway cells from patients with COPD and lung cancer

To evaluate the potential utility of our findings as diagnostic and lung-cancer-risk-assessment biomarkers in non-surgical patients, we assessed whether genes preferentially disrupted in COPD lung tumours could be detected as aberrantly methylated in small airway cells from COPD patients with lung cancer. We detected hypermethylation of five COPD-related lung cancer genes in these tissues. These included putative cancer epigenetic diagnostic biomarkers (*CCNDBP1*, *HPGD*), genes associated with oxidative response and COPD progression (*PPARGC1A*) and the enzyme (*MAT2B*) that catalyzes synthesis of S-adenosylmethionine (SAM) -- the cell's primary methyl donor for histone and DNA methyltransferases.

Taken together, these findings provide rationale to assess whether alterations specific to tumourigenic processes can be used as targets for identifying patients before the development of lung cancer. Such discoveries will lead to the discovery of novel diagnostic, prognostic, and therapeutic markers that will ultimately improve patient outcomes. The National Lung Screening Trial (NLST) sponsored by the National Institutes of Health (NIH) was a large scale trial (n= 53,454 current or heavy former smokers, across 33 centres across the United States) to assess the effects of two screening methods: low-dose helical computed tomography (CT) scans and conventional care (i.e. chest X-rays). This study found that amongst high risk individuals, there was a 20.6% lower risk of dying from lung cancer compared to conventional care, if individuals received CT scans [274, 275]. The cost of

screening all current and former smokers is prohibitive therefore there is a need for strategies to stratify individuals based on risk. This is especially true for never smokers without a family history of lung cancer, who are excluded from screening under standard lung cancer risk criteria. An improved understanding of the biology underlying lung cancer risk, and the detection of these markers in surrogate tissues in a minimally-invasively manner could have important implications to improving lung cancer survival.

5.3 Strengths and limitations of thesis work

5.3.1 Chapter 2

We provided the first genome-wide methylation and integrative 'omics study applied to the analysis of SAE from individuals with COPD. However, this study was severely limited by sample size, therefore further investigation of aberrant methylation in small airways across a larger cohort of subjects for which COPD phenotypes are defined by both CT and symptoms in addition to lung function, is warranted. Moreover, it is possible that populations of cells obtained from small airways of COPD patients may contain more inflammatory cells than those from individuals without COPD, and should therefore be considered in the interpretation of gene expression and methylation comparisons. Our results provide rationale for further assessment of the involvement of DNA methylation to COPD biology that ideally considers functional elements beyond gene promoters, such as the HumanMethylation450 BeadChip (Illumina, San Diego, CA) which assesses methylation status of over 450,000 CpG sites spanning 99% of RefSeq genes and 96 of all CpG islands [242].

We also note that since DNA sequence variants can affect normal methylation patterns affecting gene expression [276], the integration of information from COPD GWAS results with methylation and expression data may help elucidate those genes and mechanisms contributing to COPD biology.

5.3.2 Chapter 3

The lung adenocarcinoma cohort we have compiled is the largest lung tumour data set with well-defined smoking status annotation and matched non-malignant tissue for every patient, to date. We also applied these findings to the first smoking- status specific survival analyses. We acknowledge that the lack of validation of our findings in external datasets is a limitation of this study. Due to i) small sample sizes, ii) lack of smoking status annotation, iii) lack of patient matched non-malignant tissue profiles for defining miRNAs as over- or under-expressed in individual tumours, and iv) use of miRNA expression arrays for profiling, which drastically reduces the number of measureable miRNAs, existing external miRNA adenocarcinoma expression datasets are not directly appropriate for validation. TCGA, which represents the largest public repository for lung adenocarcinoma miRNA expression data generated by sequencing, contains a small number of never smoker tumours (n= 16) and fewer than 50 cases with patient matched tumour and non-malignant profiles. We also noticed that the available TCGA data had lower detection sensitivity (Figure 3.3). We were only able to validate four smoking status specific miRNA in the TCGA cohort. Moreover, our NS lung tumour cohort was biased towards patients of Asian ethnicity. Thus further assessment of our findings in additional lung tumour and non-malignant 'omic cohorts from ethnically balanced cohorts is necessary. Since survival analyses are not dependent on non-malignant profiles or absolute miRNA expression values because patient outcome was assessed on rank based tertiles of tumour miRNA expression, the large numbers of well annotated tumour miRNA sequencing expression profiles from the TCGA were integral in increasing our sample size to enable us to perform smoking status specific survival analyses.

5.3.3 Chapter 4

We were able to conduct the first multi-omic characterization of lung tumours in the context of COPD. However there were multiple caveats to this study. There was a significant bias towards smokers in our COPD cohort and towards never smokers in our non-COPD cohort. While we noted substantial overlap among gene sets derived from our i) primary cohort and ii) "smoking- associated" analyses, we also note a substantial degree of overlap between both of these gene sets, with results from our COPD tumour analysis which did not include never smokers. Thus while results from our primary analyses are likely confounded by the biological effects of smoking, overlap among gene sets and enriched pathways

generated from these three analyses indicate to us that genes altered in response to cigarette smoke are likely important to COPD related lung cancer. This is supported by multiple epidemiological and genetic association studies that link cigarette smoke to both diseases [248-251]. Understanding why some current and former smokers are at increased risk of COPD and lung cancer, while others remain free of disease does however have important implications to the design of markers which stratify patients based on risk (section 5.2.4). The type of smoke (tobacco vs. biomass fuels) has also recently been associated with differences in clinical presentation of COPD [277], therefore there are further therapeutic implications to clarifying this interaction.

Lung tumour 'omics cohorts (any 'omics dimension) with annotated lung function measures are extremely scarce in the public domain. Given the growing global burden of COPD and lung cancer, and the relatively recently recognized importance of clinical factors such as smoking status and lung function to lung cancer biology and patient outcome; this type of resource is sure to expand in the public domain. Larger tumour and non-malignant cohorts with annotated lung function measures will accord analytical strategies beyond two group comparisons. As mentioned in Chapter 1 (section 1.6), the inclusion of non-malignant tissues as references for identifying tumour-specific alterations has the potential of masking potentially important molecular alterations occurring in the "field of cancerization" [117-120]. These "field effect" changes, also referred to as the "molecular field of injury" may be involved in initiation of frank disease and are thus potential highly relevant. Compared to the prominence of genomic alterations detected in tumour tissues, the magnitude of genomic disruptions occurring in non-malignant tissues is relatively small. Effect size-- which can be defined as a signal to noise ratio, where noise includes both biological and technical variation [137], is small in non-malignant tissues, thus large sample sizes improve discovery of alterations specific to the molecular field of injury. In the case of a large, annotated tumour and non-malignant tissue 'omics cohort, an alternate analytical strategy based on the molecular field of injury hypothesis would be to identify 'omics changes associated with lung function decline (as a continuous variable) in non-malignant tissues of patients with COPD and lung cancer compared to COPD alone, and then assess whether associated genes

are further deregulated in COPD-related lung tumours and/or in non- or minimally invasive tissues such as airway, nasal or buccal cells.

We acknowledge the importance of other 'omic levels known to be involved in COPD and lung cancer biology, that occur at the genetic (whole-genome sequence), epigenetic (non-coding RNA, modification to histones) and proteomic levels [278-281], which we did not assess in our current study. We also acknowledge that validation of our findings was lacking both functionally (in lung tumour and COPD models) and clinically (in prospectively collected cohorts).

5.4 Future research directions and considerations

5.4.1 Incorporating epidemiological evidence into cancer 'omics research

Chronic inflammation in different organs share common causes (e.g., chronic insult of cyto- and genotoxic agents), pathology (e.g., tissue remodeling) and consequences (e.g., high cancer risk). It is therefore plausible that inflammatory-related cancers also share common underlying molecular biology. Given the increasing availability of a very large number of clinically annotated, high resolution (i.e. sequence level) multi-omic tumour profiles; a meta- multi-omic analysis of inflammatory-related tumours across different tissue types is one such possible approach to address this hypothesis. An improved understanding of the molecular events mediating the progression of chronic inflammatory states to cancer is required for development of targeted therapies which can halt the damaging effects of chronic inflammation while preserving the protective effects of the immune system, in the lung and other organs.

5.4.2 Anticipating shifts in patient demographics

With the success of smoking prevention and cessation campaigns, in the coming decades lung cancer and COPD in North America will increasingly become a disease of former and never smokers (Figure 5.1). Presently, in North America, half of all newly diagnosed lung cancer patients are FS and 25% are NS [50, 282]. Statistics for British

Columbia are presented in Figure 5.1. As clinically relevant molecular differences exist for both COPD and lung cancer in relation to duration, type and amount of smoking exposure, understanding the biology of these diseases in the context of different smoking environments is thus increasingly important.

Figure 5.1

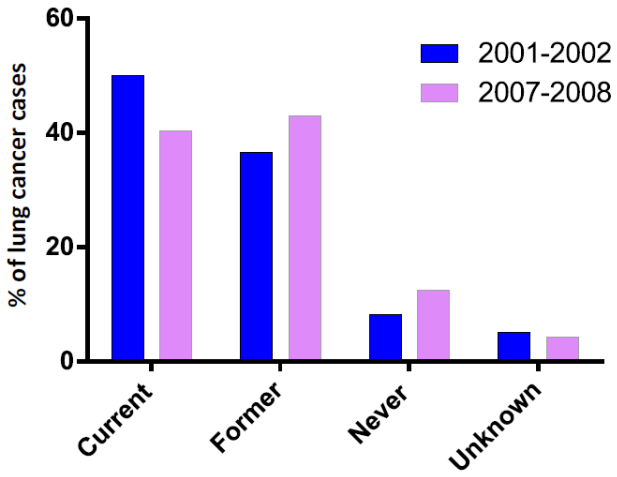

Figure 5.1 Smoking status of lung cancer patients in British Columbia
Data from British Columbia Cancer Agency

Another shift relates to advances and widespread implementation of cancer screening and early detection programs. Currently, lung cancer patient survival after 5 years of diagnosis is dismal at 16%, although this is largely due to advanced stage disease at diagnosis; when stratified by stage, 50% of lung cancer patients who are diagnosed early (stage IA) survive after five years, compared to only 2% of those diagnosed late (stage IV). Encouragingly, early stage tumours are becoming more readily detectable and survival rates for many cancers are improving [283]. These advances will also bring new opportunities for genomic analyses on pre-invasive tumours, and with it, emphasis on the genomics of early disease and new strategies for detection and early intervention. A shift towards 'omics cohorts with early stage patients will likely result in an improved understanding of cancer initiation and development. Translation of these studies will vastly accelerate improved gene

signatures or pathway disruption patterns which could for example, predict which lesions progress to lung cancer or whether a CT detected lung nodule would become malignant. Strategies to predict which nodules detected by low-dose CT will progress based on nodule characteristics and patient information, as described by McWilliams et al., are an increasingly important aspect to lung cancer patient care [284].

5.4.3 Challenges to translation of 'omics findings

Many of the challenges in translational 'omics can be summed up by the "cancer biomarker problem", which refers to the great disparity in 'omics information produced, to number of successfully translated diagnostic, prognostic and especially predictive biomarkers derived from this massive body of work [285, 286]. Translational success of a biomarker may be defined as a diagnostic test "that will change clinical practice and reduce costs, either by improving people's health or by eliminating ineffective, expensive treatments" [286, 287]. Practical challenges hindering translation of 'omics-generated biomarkers include: availability of well defined clinically characterized cohorts and lack of standardization regarding how specimens are collected, handled and stored. Ultimately, these roadblocks relate to how, or if, biomarkers are validated for cancer specificity in well controlled cohorts [286, 288]. As described by Liotta and Petricoin, biomarker translation has also been impeded by a lack of mechanistic links to the tumour itself, which the authors suggest may be overcome by i) initial discovery of the biomarker in animal tumour models, ii) showing a functional role for the biomarker in tumourigenesis or in response to treatment and iii) validating the marker in humans, as exemplified by a recent study by Taguchi *et al.*[289].

The application of high throughput 'omics techniques, particularly sequencing level technologies to tumour tissues or model systems to identify molecular features driving associated phenotypes, holds great potential to accelerate translation of 'omics research. The wealth of this type of data in the public domain is rapidly growing. Today, multi-omics data for thousands of tumours with detailed clinical information across dozens of cancer types is increasingly available to researchers. When these data are considered in a specific biological context -- for example, in relation to distinct clinical phenotypes including treatment response, outcome, acquired phenotypes such as drug resistance, or towards understanding

cancer development under unique selective pressures such as cigarette smoke or inflammation -- they hold potential to greatly improve discovery and accurate assessment of biomarkers, perhaps especially so for predictive markers informing rational application of targeted therapeutics [290-293].

For clinical utility to be established, a biomarker must be reproducible, specific and sensitive [294]. However, most published biomarkers fail at the point of reproducibility. This can be attributed to i) small sample size of discovery cohorts, ii) assessment of relatively few markers or 'omic levels or iii) unaccounted for confounding bias in discovery and test cohorts, which can be easily introduced during sample collection or processing [294]. One potential hazard particularly pertinent to high-dimensional data is over-fitting which occurs when model fitting exploits characteristics of data related to noise or experimental artifacts rather than biology [295]. The risk of over-fitting increases when the model has a large number of measurements relative to number of samples, as is the case with whole genome sequence and array based approaches to marker discovery.

Analysis of multi-omics data, integrated on a per tumour basis, in the context of distinct clinical phenotypes and treatment exposure for that tumour specimen -- is an analytical concept in line with the fundamental goals of personalized medicine. The clinical utility of such information could be significant; causative DNA level events could serve as biologically relevant biomarkers related to COPD pathogenesis or lung cancer risk, or as targets for therapeutic interventions. Since some of these mechanisms are reversible (e.g., DNA methylation), further work in this area may contribute to the development of novel treatment strategies or the re-appropriation of existing epigenetic based drugs to the treatment or prevention of COPD.

Encouragingly, the elapsed time between target discovery and clinical utilization of targeted therapies has decreased significantly in the past five years [296, 297]. The translation of ALK inhibitors, which are used to treat the ~7% of NSCLCs patients whose tumours harbour *EML4-ALK* rearrangements, was achieved in a remarkable three years [296, 298, 299]. Speed of translation will likely increase further as classic drug development and

trial regimes are reshaped to include prospective characterization of patients, and collection and interrogation of biological materials throughout clinical trials to assess patient response [296, 297, 300, 301].

The central tenet of 'omics investigation is that it allows for open (not only hypothesis driven) discovery. The integration of ‘omics data with epidemiological data from well defined cohorts, improves our ability to associate genetic alterations with environmental exposures and specific clinical phenotypes. This has the potential to improve our current understanding of cancer biology and ultimately patient management. Now that technological developments have enabled such multi-dimensional studies, much of the focus will shift to study design, interpretation and clinical applicability. Crucial to furthering the goals of this field and to the continued support and funding of this multidisciplinary work, is the communication of findings to the public by the scientific community. While translational success of cancer research is judged by improved survival for cancer patients, its effective implementation will require educating the medical establishment and the public at large about the power of ‘omics to transform medicine and improve patient outcomes.

Bibliography

1. Lozano, R., et al., *Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010.* Lancet, 2012. **380**(9859): p. 2095-128.
2. Murray, C.J. and A.D. Lopez, *Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study.* Lancet, 1997. **349**(9064): p. 1498-504.
3. Murray, C.J. and A.D. Lopez, *Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study.* Lancet, 1997. **349**(9063): p. 1436-42.
4. Mathers, C.D. and D. Loncar, *Projections of global mortality and burden of disease from 2002 to 2030.* PLoS Med, 2006. **3**(11): p. e442.
5. (GOLD), G.I.f.C.O.L.D., *Global strategy for the diagnosis, management and prevention of chronic pulmonary disease.* http://www.goldcopd.org/Guidelines/guidelines-resources.html., 2011.
6. Tan, W.C., et al., *Can age and sex explain the variation in COPD rates across large urban cities? A population study in Canada.* Int J Tuberc Lung Dis, 2011. **15**(12): p. 1691-8.
7. International Agency for Research on Cancer (IARC), W.H.O., *World Health Organization Classification of Tumours: Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart*, E.B. William D. Travis, H.Konrad Muller-hermelink, Curtis C. Harris, Editor. 2004, IARC*Press*: Lyon, France. p. 10-124.
8. Hecht, S.S., *Lung carcinogenesis by tobacco smoke.* Int J Cancer, 2012. **131**(12): p. 2724-32.
9. Jemal, A., et al., *Global cancer statistics.* CA Cancer J Clin, 2011. **61**(2): p. 69-90.
10. Brambilla E, T.W., Brennan P, Harris CC, Padilla JRP, *Lung cancer*. World cancer report 2014. 2014: International Agency for Research on Cancer.
11. El-Zein, R.A., et al., *Genetic predisposition to chronic obstructive pulmonary disease and/or lung cancer: important considerations when evaluating risk.* Cancer Prev Res (Phila), 2012. **5**(4): p. 522-7.
12. Hanna, M., *Matching Taxpayer Funding to Taxpayer Health Needs.* Am J Med, 2014.
13. Turato, G., R. Zuin, and M. Saetta, *Pathogenesis and pathology of COPD.* Respiration, 2001. **68**(2): p. 117-28.
14. Wright, J.L., et al., *The structure and function of the pulmonary vasculature in mild chronic obstructive pulmonary disease. The effect of oxygen and exercise.* Am Rev Respir Dis, 1983. **128**(4): p. 702-7.
15. Nagai, A., W.W. West, and W.M. Thurlbeck, *The National Institutes of Health Intermittent Positive-Pressure Breathing trial: pathology studies. II. Correlation between morphologic findings, clinical findings, and evidence of expiratory air-flow obstruction.* Am Rev Respir Dis, 1985. **132**(5): p. 946-53.
16. Snider, G.L., *Chronic obstructive pulmonary disease--a continuing challenge.* Am Rev Respir Dis, 1986. **133**(5): p. 942-4.
17. Hogg, J.C., P.T. Macklem, and W.M. Thurlbeck, *Site and nature of airway obstruction in chronic obstructive lung disease.* N Engl J Med, 1968. **278**(25): p. 1355-60.

18. Van Brabandt, H., et al., *Partitioning of pulmonary impedance in excised human and canine lungs.* J Appl Physiol, 1983. **55**(6): p. 1733-42.
19. Leopold, J.G. and J. Gough, *The centrilobular form of hypertrophic emphysema and its relation to chronic bronchitis.* Thorax, 1957. **12**(3): p. 219-35.
20. Burgel, P.R., et al., *Update on the roles of distal airways in COPD.* Eur Respir Rev, 2011. **20**(119): p. 7-22.
21. McDonough, J.E., et al., *Small-airway obstruction and emphysema in chronic obstructive pulmonary disease.* N Engl J Med, 2011. **365**(17): p. 1567-75.
22. Hauser, E., et al., *The incidence of Duchenne muscular dystrophy in eastern Austria. The controversy regarding CK screening.* Wien Klin Wochenschr, 1993. **105**(15): p. 433-6.
23. Silverman, E.K., et al., *Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis.* Am J Respir Crit Care Med, 1998. **157**(6 Pt 1): p. 1770-8.
24. Givelber, R.J., et al., *Segregation analysis of pulmonary function among families in the Framingham Study.* Am J Respir Crit Care Med, 1998. **157**(5 Pt 1): p. 1445-51.
25. Wilk, J.B., et al., *Evidence for major genes influencing pulmonary function in the NHLBI family heart study.* Genet Epidemiol, 2000. **19**(1): p. 81-94.
26. He, J.Q., et al., *Antioxidant gene polymorphisms and susceptibility to a rapid decline in lung function in smokers.* Am J Respir Crit Care Med, 2002. **166**(3): p. 323-8.
27. Joos, L., et al., *The role of matrix metalloproteinase polymorphisms in the rate of decline in lung function.* Hum Mol Genet, 2002. **11**(5): p. 569-76.
28. Cho, M.H., et al., *Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis.* Lancet Respir Med, 2014. **2**(3): p. 214-25.
29. Kirkham, P.A. and P.J. Barnes, *Oxidative stress in COPD.* Chest, 2013. **144**(1): p. 266-73.
30. Artigas, M.S., et al., *Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function.* Nat Genet, 2011. **43**(11): p. 1082-1090.
31. Smolonska, J., et al., *Meta-analyses on suspected chronic obstructive pulmonary disease genes: a summary of 20 years' research.* Am J Respir Crit Care Med, 2009. **180**(7): p. 618-31.
32. Repapi, E., et al., *Genome-wide association study identifies five loci associated with lung function.* Nat Genet, 2010. **42**(1): p. 36-44.
33. Hancock, D.B. and S.J. London, *Determinants of lung function, COPD, and asthma.* N Engl J Med, 2011. **364**(1): p. 86-7.
34. Brebner, J.A. and R.A. Stockley, *Recent advances in alpha-1-antitrypsin deficiency-related lung disease.* Expert Rev Respir Med, 2013. **7**(3): p. 213-29; quiz 230.
35. Laurell, C.B. and S. Eriksson, *The electrophoretic alpha1-globulin pattern of serum in alpha1-antitrypsin deficiency. 1963.* COPD, 2013. **10 Suppl 1**: p. 3-8.
36. Demeo, D.L., et al., *The SERPINE2 gene is associated with chronic obstructive pulmonary disease.* Am J Hum Genet, 2006. **78**(2): p. 253-64.
37. Campbell, J.D., et al., *A gene expression signature of emphysema-related lung destruction and its reversal by the tripeptide GHK.* Genome Med, 2012. **4**(8): p. 67.
38. Ning, W., et al., *Comprehensive gene expression profiles reveal pathways related to the pathogenesis of chronic obstructive pulmonary disease.* Proc Natl Acad Sci U S A, 2004. **101**(41): p. 14895-900.

39. Golpon, H.A., et al., *Emphysema lung tissue gene expression profiling.* Am J Respir Cell Mol Biol, 2004. **31**(6): p. 595-600.
40. Spira, A., et al., *Gene expression profiling of human lung tissue from smokers with severe emphysema.* Am J Respir Cell Mol Biol, 2004. **31**(6): p. 601-10.
41. Steiling, K., et al., *A Dynamic Bronchial Airway Gene Expression Signature of COPD and Lung Function Impairment.* Am J Respir Crit Care Med, 2013.
42. Tilley, A.E., et al., *Biologic phenotyping of the human small airway epithelial response to cigarette smoking.* PLoS One, 2011. **6**(7): p. e22798.
43. Ammous, Z., et al., *Variability in small airway epithelial gene expression among normal smokers.* Chest, 2008. **133**(6): p. 1344-53.
44. Pierrou, S., et al., *Expression of genes involved in oxidative stress responses in airway epithelial cells of smokers with chronic obstructive pulmonary disease.* Am J Respir Crit Care Med, 2007. **175**(6): p. 577-86.
45. Wang, I.M., et al., *Gene expression profiling in patients with chronic obstructive pulmonary disease and lung cancer.* Am J Respir Crit Care Med, 2008. **177**(4): p. 402-11.
46. Rahman, I., *Pharmacological antioxidant strategies as therapeutic interventions for COPD.* Biochim Biophys Acta, 2011.
47. Pikor, L.A., et al., *Genetic alterations defining NSCLC subtypes and their therapeutic implications.* Lung Cancer, 2013. **82**(2): p. 179-89.
48. Pao, W. and K.E. Hutchinson, *Chipping away at the lung cancer genome.* Nat Med, 2012. **18**(3): p. 349-51.
49. Lockwood, W.W., et al., *Divergent genomic and epigenomic landscapes of lung cancer subtypes underscore the selection of different oncogenic pathways during tumor development.* PLoS One, 2012. **7**(5): p. e37775.
50. Sun, S., J.H. Schiller, and A.F. Gazdar, *Lung cancer in never smokers--a different disease.* Nat Rev Cancer, 2007. **7**(10): p. 778-90.
51. Liu, J., et al., *Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events.* Genome Res, 2012. **22**(12): p. 2315-27.
52. Wu, X., et al., *Genome-wide association study of genetic predictors of overall survival for non-small cell lung cancer in never smokers.* Cancer Res, 2013. **73**(13): p. 4028-38.
53. Dasgupta, S., et al., *Mitochondrial DNA mutations in respiratory complex-I in never-smoker lung cancer patients contribute to lung cancer progression and associated with EGFR gene mutation.* J Cell Physiol, 2012. **227**(6): p. 2451-60.
54. Thu, K.L., et al., *Lung adenocarcinoma of never smokers and smokers harbor differential regions of genetic alteration and exhibit different levels of genomic instability.* PLoS One, 2012. **7**(3): p. e33003.
55. Pao, W., A.J. Iafrate, and Z. Su, *Genetically informed lung cancer medicine.* J Pathol, 2011. **223**(2): p. 230-40.
56. Nana-Sinkam, S.P. and C.A. Powell, *Molecular biology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines.* Chest, 2013. **143**(5 Suppl): p. e30S-9S.
57. Toyooka, S., et al., *Molecular oncology of lung cancer.* Gen Thorac Cardiovasc Surg, 2011. **59**(8): p. 527-37.

58. Brambilla, E. and A. Gazdar, *Pathogenesis of lung cancer signalling pathways: roadmap for therapies.* Eur Respir J, 2009. **33**(6): p. 1485-97.
59. Palmer, J.D., et al., *Molecular markers to predict clinical outcome and radiation induced toxicity in lung cancer.* J Thorac Dis, 2014. **6**(4): p. 387-398.
60. Ray, M.R., D. Jablons, and B. He, *Lung cancer therapeutics that target signaling pathways: an update.* Expert Rev Respir Med, 2010. **4**(5): p. 631-45.
61. Wistuba, II, A.F. Gazdar, and J.D. Minna, *Molecular genetics of small cell lung carcinoma.* Semin Oncol, 2001. **28**(2 Suppl 4): p. 3-13.
62. Dearden, S., et al., *Mutation incidence and coincidence in non small-cell lung cancer: meta-analyses by ethnicity and histology (mutMap).* Ann Oncol, 2013. **24**(9): p. 2371-6.
63. Pan, Y., et al., *ALK, ROS1 and RET fusions in 1139 lung adenocarcinomas: A comprehensive study of common and fusion pattern-specific clinicopathologic, histologic and cytologic features.* Lung Cancer, 2014. **84**(2): p. 121-6.
64. Vermaelen, K. and G. Brusselle, *Exposing a deadly alliance: novel insights into the biological links between COPD and lung cancer.* Pulm Pharmacol Ther, 2013. **26**(5): p. 544-54.
65. Medzhitov, R., *Origin and physiological roles of inflammation.* Nature, 2008. **454**(7203): p. 428-35.
66. Elinav, E., et al., *Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms.* Nat Rev Cancer, 2013. **13**(11): p. 759-71.
67. Coussens, L.M., L. Zitvogel, and A.K. Palucka, *Neutralizing tumor-promoting chronic inflammation: a magic bullet?* Science, 2013. **339**(6117): p. 286-91.
68. Balkwill, F. and A. Mantovani, *Inflammation and cancer: back to Virchow?* Lancet, 2001. **357**(9255): p. 539-45.
69. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
70. Colotta, F., et al., *Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability.* Carcinogenesis, 2009. **30**(7): p. 1073-81.
71. Pfeifer, G.P., et al., *Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers.* Oncogene, 2002. **21**(48): p. 7435-51.
72. Hecht, S.S., *Cigarette smoking and lung cancer: chemical mechanisms and approaches to prevention.* Lancet Oncol, 2002. **3**(8): p. 461-9.
73. Hubert, H.B., et al., *Genetic and environmental influences on pulmonary function in adult twins.* Am Rev Respir Dis, 1982. **125**(4): p. 409-15.
74. Yang, I.A., J.W. Holloway, and K.M. Fong, *Genetic susceptibility to lung cancer and co-morbidities.* J Thorac Dis, 2013. **5**(Suppl 5): p. S454-S462.
75. Tang, W., et al., *Genetic variation in antioxidant enzymes, cigarette smoking, and longitudinal change in lung function.* Free Radic Biol Med, 2013. **63**: p. 304-12.
76. Albertson, D.G., et al., *Chromosome aberrations in solid tumors.* Nat Genet, 2003. **34**(4): p. 369-76.
77. Chari, R., et al., *Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer.* Cancer Metastasis Rev, 2010. **29**(1): p. 73-93.
78. Vucic, E.A., et al., *Copy number variations in the human genome and strategies for analysis.* Methods Mol Biol, 2010. **628**: p. 103-17.

79. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature, 2012. **486**(7403): p. 346-52.
80. Craddock, K.J., W.L. Lam, and M.S. Tsao, *Applications of array-CGH for lung cancer.* Methods Mol Biol, 2013. **973**: p. 297-324.
81. Campbell, P.J., et al., *The patterns and dynamics of genomic instability in metastatic pancreatic cancer.* Nature, 2010. **467**(7319): p. 1109-13.
82. Cox, C., et al., *A survey of homozygous deletions in human cancer genomes.* Proc Natl Acad Sci U S A, 2005. **102**(12): p. 4542-7.
83. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nat Rev Genet, 2012. **13**(7): p. 484-92.
84. Haaf, T., *Methylation dynamics in the early mammalian embryo: implications of genome reprogramming defects for development.* Curr Top Microbiol Immunol, 2006. **310**: p. 13-22.
85. Jones, P.A., et al., *De novo methylation of the MyoD1 CpG island during the establishment of immortal cell lines.* Proc Natl Acad Sci U S A, 1990. **87**(16): p. 6117-21.
86. Feinberg, A.P., R. Ohlsson, and S. Henikoff, *The epigenetic progenitor origin of human cancer.* Nat Rev Genet, 2006. **7**(1): p. 21-33.
87. Weber, M., et al., *Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.* Nat Genet, 2007. **39**(4): p. 457-66.
88. Esteller, M. and J.G. Herman, *Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours.* J Pathol, 2002. **196**(1): p. 1-7.
89. Jaenisch, R. and A. Bird, *Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.* Nat Genet, 2003. **33 Suppl**: p. 245-54.
90. Jones, P.A. and S.B. Baylin, *The fundamental role of epigenetic events in cancer.* Nat Rev Genet, 2002. **3**(6): p. 415-28.
91. Teperino, R., A. Lempradl, and J.A. Pospisilik, *Bridging epigenomics and complex disease: the basics.* Cell Mol Life Sci, 2013. **70**(9): p. 1609-21.
92. Yates, L.A., C.J. Norbury, and R.J. Gilbert, *The long and short of microRNA.* Cell, 2013. **153**(3): p. 516-9.
93. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.
94. Denli, A.M., et al., *Processing of primary microRNAs by the Microprocessor complex.* Nature, 2004. **432**(7014): p. 231-5.
95. Filipowicz, W., et al., *Post-transcriptional gene silencing by siRNAs and miRNAs.* Curr Opin Struct Biol, 2005. **15**(3): p. 331-41.
96. Shirdel, E.A., et al., *NAViGaTing the micronome--using multiple microRNA prediction databases to identify signalling pathway-associated microRNAs.* PLoS One, 2011. **6**(2): p. e17429.
97. Baer, C., R. Claus, and C. Plass, *Genome-wide epigenetic regulation of miRNAs in cancer.* Cancer Res, 2013. **73**(2): p. 473-7.
98. Nana-Sinkam, S.P. and C.M. Croce, *Clinical applications for microRNAs in cancer.* Clin Pharmacol Ther, 2013. **93**(1): p. 98-104.
99. Croce, C.M., *Oncogenes and cancer.* N Engl J Med, 2008. **358**(5): p. 502-11.
100. Hudson, T.J., et al., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993-8.

101. Verhaak, R.G., et al., *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.* Cancer Cell, 2010. **17**(1): p. 98-110.
102. Consortium, T.C.G.A., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
103. Ding, L., et al., *Somatic mutations affect key pathways in lung adenocarcinoma.* Nature, 2008. **455**(7216): p. 1069-75.
104. Stephens, P.J., et al., *Complex landscapes of somatic rearrangement in human breast cancer genomes.* Nature, 2009. **462**(7276): p. 1005-10.
105. Bozic, I., et al., *Accumulation of driver and passenger mutations during tumor progression.* Proc Natl Acad Sci U S A, 2010. **107**(43): p. 18545-50.
106. Pleasance, E.D., et al., *A small-cell lung cancer genome with complex signatures of tobacco exposure.* Nature, 2010. **463**(7278): p. 184-90.
107. Swanton, C., R.A. Burrell, and P.A. Futreal, *Breast cancer genome heterogeneity: a challenge to personalised medicine?* Breast Cancer Res, 2011. **13**(1): p. 104.
108. Fojo, T. and D.R. Parkinson, *Biologically targeted cancer therapy and marginal benefits: are we making too much of too little or are we achieving too little by giving too much?* Clin Cancer Res, 2010. **16**(24): p. 5972-80.
109. Kim, E.S. and K.J. Pandya, *Advances in personalized therapy for lung cancer.* Expert Opin Med Diagn, 2013. **7**(5): p. 475-85.
110. Chari, R., et al., *An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer.* BMC Syst Biol, 2010. **4**: p. 67.
111. Hood, L., et al., *Systems biology and new technologies enable predictive and preventative medicine.* Science, 2004. **306**(5696): p. 640-3.
112. Cancer Genome Atlas, N., *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.
113. Modrek, B., et al., *Oncogenic activating mutations are associated with local copy gain.* Mol Cancer Res, 2009. **7**(8): p. 1244-52.
114. Segditsas, S., et al., *APC and the three-hit hypothesis.* Oncogene, 2009. **28**(1): p. 146-55.
115. Wang, K., et al., *Systems biology and the discovery of diagnostic biomarkers.* Dis Markers, 2010. **28**(4): p. 199-207.
116. Vidal, M., M.E. Cusick, and A.L. Barabasi, *Interactome networks and human disease.* Cell, 2011. **144**(6): p. 986-98.
117. Slaughter, D.P., H.W. Southwick, and W. Smejkal, *Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin.* Cancer, 1953. **6**(5): p. 963-8.
118. Gomperts, B.N., et al., *Evolving concepts in lung carcinogenesis.* Semin Respir Crit Care Med, 2011. **32**(1): p. 32-43.
119. Kadara, H., et al., *Transcriptomic architecture of the adjacent airway field cancerization in non-small cell lung cancer.* J Natl Cancer Inst, 2014. **106**(3): p. dju004.
120. Steiling, K., M.E. Lenburg, and A. Spira, *Airway gene expression in chronic obstructive pulmonary disease.* Proc Am Thorac Soc, 2009. **6**(8): p. 697-700.
121. Cho, M.H., et al., *A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13.* Hum Mol Genet, 2012. **21**(4): p. 947-57.

122. Kong, X., et al., *Genome-wide association study identifies BICD1 as a susceptibility gene for emphysema.* Am J Respir Crit Care Med, 2011. **183**(1): p. 43-9.
123. Qiu, W., et al., *Variable DNA methylation is associated with chronic obstructive pulmonary disease and lung function.* Am J Respir Crit Care Med, 2012. **185**(4): p. 373-81.
124. Bowler, R.P., et al., *Integrative Omics Approach Identifies Interleukin-16 as a Biomarker of Emphysema.* OMICS, 2013.
125. Obeidat, M., et al., *A comprehensive evaluation of potential lung function associated genes in the SpiroMeta general population sample.* PLoS One, 2011. **6**(5): p. e19382.
126. Lamontagne, M., et al., *Refining susceptibility loci of chronic obstructive pulmonary disease with lung eqtls.* PLoS One, 2013. **8**(7): p. e70220.
127. Kabesch, M. and I.M. Adcock, *Epigenetics in asthma and COPD.* Biochimie, 2012.
128. Wan, E.S., et al., *Systemic Steroid Exposure is Associated with Differential Methylation in Chronic Obstructive Pulmonary Disease.* Am J Respir Crit Care Med, 2012.
129. Sood, A., et al., *Wood smoke exposure and gene promoter methylation are associated with increased risk for COPD in smokers.* Am J Respir Crit Care Med, 2010. **182**(9): p. 1098-104.
130. Nebbioso, A., et al., *Trials with 'epigenetic' drugs: an update.* Mol Oncol, 2012. **6**(6): p. 657-82.
131. Lonergan, K.M., et al., *Identification of novel lung genes in bronchial epithelium by serial analysis of gene expression.* Am J Respir Cell Mol Biol, 2006. **35**(6): p. 651-61.
132. Chari, R., et al., *Effect of active smoking on the human bronchial epithelium transcriptome.* BMC Genomics, 2007. **8**: p. 297.
133. *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-15.
134. Cotton, A.M., et al., *Inactive X chromosome-specific reduction in placental DNA methylation.* Hum Mol Genet, 2009. **18**(19): p. 3544-52.
135. Tost, J. and I.G. Gut, *DNA methylation analysis by pyrosequencing.* Nat Protoc, 2007. **2**(9): p. 2265-75.
136. Du, P., W.A. Kibbe, and S.M. Lin, *lumi: a pipeline for processing Illumina microarray.* Bioinformatics, 2008. **24**(13): p. 1547-8.
137. Zhuang, J., M. Widschwendter, and A.E. Teschendorff, *A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform.* BMC Bioinformatics, 2012. **13**: p. 59.
138. Hackett, N.R., et al., *RNA-Seq Quantification of the Human Small Airway Epithelium Transcriptome.* BMC Genomics, 2012. **13**(1): p. 82.
139. Bosse, Y., et al., *Molecular Signature of Smoking in Human Lung Tissues.* Cancer Res, 2012.
140. Vucic, E.A., et al., *DNA Methylation is Globally Disrupted and Associated with Expression Changes in COPD Small Airways.* Am J Respir Cell Mol Biol, 2013.
141. Zhang, Y., et al., *Kinase suppressor of Ras-1 protects against pulmonary Pseudomonas aeruginosa infections.* Nat Med, 2011. **17**(3): p. 341-6.

142. Buro-Auriemma, L.J., et al., *Cigarette smoking induces small airway epithelial epigenetic changes with corresponding modulation of gene expression.* Hum Mol Genet, 2013.
143. Kistemaker, L.E., et al., *Muscarinic M(3) receptors contribute to allergen-induced airway remodeling in mice.* Am J Respir Cell Mol Biol, 2014. **50**(4): p. 690-8.
144. Karakiulakis, G. and M. Roth, *Muscarinic receptors and their antagonists in COPD: anti-inflammatory and antiremodeling effects.* Mediators Inflamm, 2012. **2012**: p. 409580.
145. Tsay, J.J., et al., *Aryl hydrocarbon receptor and lung cancer.* Anticancer Res, 2013. **33**(4): p. 1247-56.
146. Allen-Gipson, D.S., et al., *Smoke Extract Impairs Adenosine Wound Healing: Implications of Smoke-Generated Reactive Oxygen Species.* Am J Respir Cell Mol Biol, 2013.
147. Chwieralski, C.E., et al., *Epidermal growth factor and trefoil factor family 2 synergistically trigger chemotaxis on BEAS-2B cells via different signaling cascades.* Am J Respir Cell Mol Biol, 2004. **31**(5): p. 528-37.
148. Hara, H., et al., *Involvement of creatine kinase B in cigarette smoke-induced bronchial epithelial cell senescence.* Am J Respir Cell Mol Biol, 2012. **46**(3): p. 306-12.
149. Vinson, C. and R. Chatterjee, *CG methylation.* Epigenomics, 2012. **4**(6): p. 655-63.
150. Bibikova, M., et al., *Genome-wide DNA methylation profiling using Infinium(R) assay.* Epigenomics, 2009. **1**(1): p. 177-200.
151. Fujimoto, J., et al., *G-protein coupled receptor family C, group 5, member A (GPRC5A) expression is decreased in the adjacent field and normal bronchial epithelia of patients with chronic obstructive pulmonary disease and non-small-cell lung cancer.* J Thorac Oncol, 2012. **7**(12): p. 1747-54.
152. Plantier, L., et al., *Dysregulation of elastin expression by fibroblasts in pulmonary emphysema: role of cellular retinoic acid binding protein 2.* Thorax, 2008. **63**(11): p. 1012-7.
153. Yang, L., et al., *Apical localization of ITPK1 enhances its ability to be a modifier gene product in a murine tracheal cell model of cystic fibrosis.* J Cell Sci, 2006. **119**(Pt 7): p. 1320-8.
154. Sanders, Y.Y., et al., *Altered DNA methylation profile in idiopathic pulmonary fibrosis.* Am J Respir Crit Care Med, 2012. **186**(6): p. 525-35.
155. Stefanowicz, D., et al., *DNA methylation profiles of airway epithelial cells and PBMCs from healthy, atopic and asthmatic children.* PLoS One, 2012. **7**(9): p. e44213.
156. Wang, X.M., et al., *Caveolin-1: a critical regulator of lung fibrosis in idiopathic pulmonary fibrosis.* J Exp Med, 2006. **203**(13): p. 2895-906.
157. Miyoshi, K., et al., *Epithelial Pten controls acute lung injury and fibrosis by regulating alveolar epithelial cell integrity.* Am J Respir Crit Care Med, 2013. **187**(3): p. 262-75.
158. Xia, H., et al., *Pathologic caveolin-1 regulation of PTEN in idiopathic pulmonary fibrosis.* Am J Pathol, 2010. **176**(6): p. 2626-37.
159. Ryter, S.W., et al., *Deadly triplex: smoke, autophagy and apoptosis.* Autophagy, 2011. **7**(4): p. 436-7.

160. Hosgood, H.D., 3rd, et al., *PTEN identified as important risk factor of chronic obstructive pulmonary disease.* Respir Med, 2009. **103**(12): p. 1866-70.
161. Shaykhiev, R., et al., *Cigarette smoking reprograms apical junctional complex molecular architecture in the human airway epithelium in vivo.* Cell Mol Life Sci, 2011. **68**(5): p. 877-92.
162. Lakhdar, R., et al., *Combined analysis of EPHX1, GSTP1, GSTM1 and GSTT1 gene polymorphisms in relation to chronic obstructive pulmonary disease risk and lung function impairment.* Dis Markers, 2011. **30**(5): p. 253-63.
163. Saccone, N.L., et al., *Multiple distinct risk loci for nicotine dependence identified by dense coverage of the complete family of nicotinic receptor subunit (CHRN) genes.* Am J Med Genet B Neuropsychiatr Genet, 2009. **150B**(4): p. 453-66.
164. Solaimani, P., R. Damoiseaux, and O. Hankinson, *Genome Wide RNAi High Throughput Screen Identifies Proteins Necessary for the AHR-Dependent Induction of CYP1A1 by 2,3,7,8-Tetrachlorodibenzo-rho-dioxin.* Toxicol Sci, 2013.
165. Paplinska, M., et al., *Expression of eotaxins in the material from nasal brushing in asthma, allergic rhinitis and COPD patients.* Cytokine, 2012. **60**(2): p. 393-9.
166. Anderson, G.P. and S. Bozinovski, *Acquired somatic mutations in the molecular pathogenesis of COPD.* Trends Pharmacol Sci, 2003. **24**(2): p. 71-6.
167. Wong, K.K., J.A. Engelman, and L.C. Cantley, *Targeting the PI3K signaling pathway in cancer.* Curr Opin Genet Dev, 2010. **20**(1): p. 87-90.
168. Bozinovski, S., et al., *Akt in the pathogenesis of COPD.* Int J Chron Obstruct Pulmon Dis, 2006. **1**(1): p. 31-8.
169. Zhang, X., et al., *Increased interleukin (IL)-8 and decreased IL-17 production in chronic obstructive pulmonary disease (COPD) provoked by cigarette smoke.* Cytokine, 2011. **56**(3): p. 717-25.
170. Kramer, J.M. and S.L. Gaffen, *Interleukin-17: a new paradigm in inflammation, autoimmunity, and therapy.* J Periodontol, 2007. **78**(6): p. 1083-93.
171. Hubner, R.H., et al., *Coordinate control of expression of Nrf2-modulated genes in the human small airway epithelium is highly responsive to cigarette smoking.* Mol Med, 2009. **15**(7-8): p. 203-19.
172. Tomaki, M., et al., *Decreased expression of antioxidant enzymes and increased expression of chemokines in COPD lung.* Pulm Pharmacol Ther, 2007. **20**(5): p. 596-605.
173. Cho, H.Y. and S.R. Kleeberger, *Nrf2 protects against airway disorders.* Toxicol Appl Pharmacol, 2010. **244**(1): p. 43-56.
174. Bell, J.C. and H.W. Strobel, *Regulation of cytochrome P450 4F11 by nuclear transcription factor-kappaB.* Drug Metab Dispos, 2012. **40**(1): p. 205-11.
175. Wang, Y., et al., *Gene regulation of CYP4F11 in human keratinocyte HaCaT cells.* Drug Metab Dispos, 2010. **38**(1): p. 100-7.
176. Hammerman, P.S., et al., *Comprehensive genomic characterization of squamous cell lung cancers.* Nature, 2012. **489**(7417): p. 519-25.
177. Papi, A., et al., *COPD increases the risk of squamous histological subtype in smokers who develop non-small cell lung carcinoma.* Thorax, 2004. **59**(8): p. 679-81.
178. Farazi, T.A., et al., *miRNAs in human cancer.* J Pathol, 2011. **223**(2): p. 102-15.

179. Schembri, F., et al., *MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium.* Proc Natl Acad Sci U S A, 2009. **106**(7): p. 2319-24.
180. Izzotti, A., et al., *Dose-responsiveness and persistence of microRNA expression alterations induced by cigarette smoke in mouse lung.* Mutat Res, 2011. **717**(1-2): p. 9-16.
181. Perdomo, C., A. Spira, and F. Schembri, *MiRNAs as regulators of the response to inhaled environmental toxins and airway carcinogenesis.* Mutat Res, 2011. **717**(1-2): p. 32-7.
182. Melkamu, T., et al., *Alteration of microRNA expression in vinyl carbamate-induced mouse lung tumors and modulation by the chemopreventive agent indole-3-carbinol.* Carcinogenesis, 2010. **31**(2): p. 252-8.
183. Chin, L.J., et al., *A SNP in a let-7 microRNA complementary site in the KRAS 3' untranslated region increases non-small cell lung cancer risk.* Cancer Res, 2008. **68**(20): p. 8535-40.
184. De Flora, S., et al., *Smoke-induced microRNA and related proteome alterations. Modulation by chemopreventive agents.* Int J Cancer, 2012. **131**(12): p. 2763-73.
185. Xi, S., et al., *Cigarette smoke mediates epigenetic repression of miR-487b during pulmonary carcinogenesis.* J Clin Invest, 2013. **123**(3): p. 1241-61.
186. Landi, M.T., et al., *MicroRNA expression differentiates histology and predicts survival of lung cancer.* Clin Cancer Res, 2010. **16**(2): p. 430-41.
187. Hu, Z., et al., *Serum microRNA signatures identified in a genome-wide serum microRNA expression profiling predict survival of non-small-cell lung cancer.* J Clin Oncol, 2010. **28**(10): p. 1721-6.
188. Dacic, S., et al., *miRNA expression profiling of lung adenocarcinomas: correlation with mutational status.* Mod Pathol, 2010. **23**(12): p. 1577-82.
189. Heegaard, N.H., et al., *Circulating micro-RNA expression profiles in early stage nonsmall cell lung cancer.* Int J Cancer, 2012. **130**(6): p. 1378-86.
190. Jang, J.S., et al., *Increased miR-708 expression in NSCLC and its association with poor survival in lung adenocarcinoma from never smokers.* Clin Cancer Res, 2012. **18**(13): p. 3658-67.
191. Zhang, Y.K., et al., *miRNAs expression profiling to distinguish lung squamous-cell carcinoma from adenocarcinoma subtypes.* J Cancer Res Clin Oncol, 2012. **138**(10): p. 1641-50.
192. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
193. Brown, K.R., et al., *NAViGaTOR: Network Analysis, Visualization and Graphing Toronto.* Bioinformatics, 2009. **25**(24): p. 3327-9.
194. McGuffin, M.J. and I. Jurisica, *Interaction techniques for selecting and manipulating subgraphs in network visualizations.* IEEE Trans Vis Comput Graph, 2009. **15**(6): p. 937-44.
195. Zhu, C.Q., et al., *Understanding prognostic gene expression signatures in lung cancer.* Clin Lung Cancer, 2009. **10**(5): p. 331-40.
196. Selamat, S.A., et al., *Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression.* Genome Res, 2012. **22**(7): p. 1197-211.

197. Beane, J., et al., *Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression.* Genome Biol, 2007. **8**(9): p. R201.
198. Zhou, C., et al., *microRNA-372 maintains oncogene characteristics by targeting TNFAIP1 and affects NFkappaB signaling in human gastric carcinoma cells.* Int J Oncol, 2013. **42**(2): p. 635-42.
199. Roa, W.H., et al., *Sputum microRNA profiling: a novel approach for the early detection of non-small cell lung cancer.* Clin Invest Med, 2012. **35**(5): p. E271.
200. Yu, Z., et al., *Identification of miR-7 as an oncogene in renal cell carcinoma.* J Mol Histol, 2013.
201. Tang, W., et al., *miR-27a regulates endothelial differentiation of breast cancer stem like cells.* Oncogene, 2013.
202. Rothschild, S.I., et al., *MicroRNA-381 represses ID1 and is deregulated in lung adenocarcinoma.* J Thorac Oncol, 2012. **7**(7): p. 1069-77.
203. Fang, L., et al., *MiR-93 enhances angiogenesis and metastasis by targeting LATS2.* Cell Cycle, 2012. **11**(23): p. 4352-65.
204. Hesse, J.E., et al., *Genome-wide small RNA sequencing and gene expression analysis reveals a microRNA profile of cancer susceptibility in ATM-deficient human mammary epithelial cells.* PLoS One, 2013. **8**(5): p. e64779.
205. Uchino, K., et al., *Therapeutic effects of microRNA-582-5p and -3p on the inhibition of bladder cancer progression.* Mol Ther, 2013. **21**(3): p. 610-9.
206. Lu, Y., et al., *MicroRNA profiling and prediction of recurrence/relapse-free survival in stage I lung cancer.* Carcinogenesis, 2012. **33**(5): p. 1046-54.
207. Jusufovic, E., et al., *let-7b and miR-126 are down-regulated in tumor tissue and correlate with microvessel density and survival outcomes in non--small--cell lung cancer.* PLoS One, 2012. **7**(9): p. e45577.
208. Yang, J., et al., *MicroRNA-126 inhibits tumor cell growth and its expression level correlates with poor survival in non-small cell lung cancer patients.* PLoS One, 2012. **7**(8): p. e42978.
209. Capodanno, A., et al., *Let-7g and miR-21 expression in non-small cell lung cancer: Correlation with clinicopathological and molecular features.* Int J Oncol, 2013. **43**(3): p. 765-74.
210. Patnaik, S.K., et al., *Evaluation of microRNA expression profiles that may predict recurrence of localized stage I non-small cell lung cancer after surgical resection.* Cancer Res, 2010. **70**(1): p. 36-45.
211. Alexandrov, K., M. Rojas, and S. Satarug, *The critical DNA damage by benzo(a)pyrene in lung tissues of smokers and approaches to preventing its formation.* Toxicol Lett, 2010. **198**(1): p. 63-8.
212. Anttila, S., H. Raunio, and J. Hakkola, *Cytochrome P450-mediated pulmonary metabolism of carcinogens: regulation and cross-talk in lung carcinogenesis.* Am J Respir Cell Mol Biol, 2011. **44**(5): p. 583-90.
213. Bartsch, H., et al., *Expression of pulmonary cytochrome P4501A1 and carcinogen DNA adduct formation in high risk subjects for tobacco-related lung cancer.* Toxicol Lett, 1992. **64-65 Spec No**: p. 477-83.
214. Skrzypek, K., et al., *Interplay Between Heme Oxygenase-1 and miR-378 Affects Non-Small Cell Lung Carcinoma Growth, Vascularization, and Metastasis.* Antioxid Redox Signal, 2013. **19**(7): p. 644-60.

215. Yamashita, S., et al., *MicroRNA-372 is associated with poor prognosis in colorectal cancer.* Oncology, 2012. **82**(4): p. 205-12.
216. Gu, H., et al., *Upregulation of microRNA-372 associates with tumor progression and prognosis in hepatocellular carcinoma.* Mol Cell Biochem, 2013. **375**(1-2): p. 23-30.
217. Lai, J.H., et al., *Comparative proteomic profiling of human lung adenocarcinoma cells (CL 1-0) expressing miR-372.* Electrophoresis, 2012. **33**(4): p. 675-88.
218. Lakomy, R., et al., *MiR-195, miR-196b, miR-181c, miR-21 expression levels and O-6-methylguanine-DNA methyltransferase methylation status are associated with clinical outcome in glioblastoma patients.* Cancer Sci, 2011. **102**(12): p. 2186-90.
219. Wang, X., et al., *Downregulation of miR-195 correlates with lymph node metastasis and poor prognosis in colorectal cancer.* Med Oncol, 2012. **29**(2): p. 919-27.
220. Li, S., et al., *MicroRNA-138 plays a role in hypoxic pulmonary vascular remodelling by targeting Mst1.* Biochem J, 2013. **452**(2): p. 281-91.
221. Zhang, H., et al., *MiR-138 inhibits tumor growth through repression of EZH2 in non-small cell lung cancer.* Cell Physiol Biochem, 2013. **31**(1): p. 56-65.
222. Sun, Y., et al., *Expression of miR-150 and miR-3940-5p is reduced in non-small cell lung carcinoma and correlates with clinicopathological features.* Oncol Rep, 2013. **29**(2): p. 704-12.
223. Zhang, N., X. Wei, and L. Xu, *miR-150 promotes the proliferation of lung cancer cells by targeting P53.* FEBS Lett, 2013. **587**(15): p. 2346-51.
224. Wang, D.T., et al., *miR-150, p53 protein and relevant miRNAs consist of a regulatory network in NSCLC tumorigenesis.* Oncol Rep, 2013. **30**(1): p. 492-8.
225. Schluger, N.W. and R. Koppaka, *Lung disease in a global context. A call for public health action.* Ann Am Thorac Soc, 2014. **11**(3): p. 407-16.
226. Ferkol, T. and D. Schraufnagel, *The global burden of respiratory disease.* Ann Am Thorac Soc, 2014. **11**(3): p. 404-6.
227. Allen, J.D., M. Chen, and Y. Xie, *Model-Based Background Correction (MBCB): R Methods and GUI for Illumina Bead-array Data.* J Cancer Sci Ther, 2009. **1**(1): p. 25-27.
228. Dees, N.D., et al., *MuSiC: identifying mutational significance in cancer genomes.* Genome Res, 2012. **22**(8): p. 1589-98.
229. Youn, A. and R. Simon, *Identifying cancer driver genes in tumor genome sequencing studies.* Bioinformatics, 2011. **27**(2): p. 175-81.
230. Lahti, L., et al., *Cancer gene prioritization by integrative analysis of mRNA expression and DNA copy number data: a comparative review.* Brief Bioinform, 2012. **14**(1): p. 27-35.
231. Louhimo, R., et al., *Comparative analysis of algorithms for integration of copy number and expression data.* Nat Methods, 2012. **9**(4): p. 351-5.
232. Zhang, S., et al., *Discovery of multi-dimensional modules by integrative analysis of cancer genomic data.* Nucleic Acids Res, 2012. **40**(19): p. 9379-91.
233. Gevaert, O. and S. Plevritis, *Identifying master regulators of cancer and their downstream targets by integrating genomic and epigenomic features.* Pac Symp Biocomput, 2013: p. 123-34.
234. Weir, B.A., et al., *Characterizing the cancer genome in lung adenocarcinoma.* Nature, 2007. **450**(7171): p. 893-8.

235. Cancer Genome Atlas Research, N., *Comprehensive genomic characterization of squamous cell lung cancers.* Nature, 2012. **489**(7417): p. 519-25.
236. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
237. Gaujoux, R. and C. Seoighe, *A flexible R package for nonnegative matrix factorization.* BMC Bioinformatics, 2010. **11**: p. 367.
238. Brunet, J.P., et al., *Metagenes and molecular pattern discovery using matrix factorization.* Proc Natl Acad Sci U S A, 2004. **101**(12): p. 4164-9.
239. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
240. Xie, X., et al., *Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.* Nature, 2005. **434**(7031): p. 338-45.
241. Aryee, M.J., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.* Bioinformatics, 2014.
242. Sandoval, J., et al., *Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome.* Epigenetics, 2011. **6**(6): p. 692-702.
243. Qin, J., et al., *COUP-TFII inhibits TGF-beta-induced growth barrier to promote prostate tumorigenesis.* Nature, 2013. **493**(7431): p. 236-40.
244. Deacon, K., et al., *Elevated SP-1 transcription factor expression and activity drives basal and hypoxia-induced vascular endothelial growth factor (VEGF) expression in non-small cell lung cancer.* J Biol Chem, 2012. **287**(47): p. 39967-81.
245. Meng, X., et al., *Transcriptional regulatory networks in human lung adenocarcinoma.* Mol Med Rep, 2012. **6**(5): p. 961-6.
246. Han, E.J., et al., *Combined treatment with peroxisome proliferator-activated receptor (PPAR) gamma ligands and gamma radiation induces apoptosis by PPARgamma-independent up-regulation of reactive oxygen species-induced deoxyribonucleic acid damage signals in non-small cell lung cancer cells.* Int J Radiat Oncol Biol Phys, 2013. **85**(5): p. e239-48.
247. Acquaah-Mensah, G.K., et al., *Suppressed expression of T-box transcription factors is involved in senescence in chronic obstructive pulmonary disease.* PLoS Comput Biol, 2012. **8**(7): p. e1002597.
248. Zhai, R., et al., *Smoking and smoking cessation in relation to the development of co-existing non-small cell lung cancer with chronic obstructive pulmonary disease.* Int J Cancer, 2014. **134**(4): p. 961-70.
249. Reuter, S., et al., *Oxidative stress, inflammation, and cancer: how are they linked?* Free Radic Biol Med, 2010. **49**(11): p. 1603-16.
250. Houghton, A.M., M. Mouded, and S.D. Shapiro, *Common origins of lung cancer and COPD.* Nat Med, 2008. **14**(10): p. 1023-4.
251. Young, R.P. and R.J. Hopkins, *How the genetics of lung cancer may overlap with COPD.* Respirology, 2011. **16**(7): p. 1047-55.
252. National Heart, L., and Blood Institute. *Coronary Heart Disease*. [cited 2014; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/cad/.
253. Ghoorah, K., A. De Soyza, and V. Kunadian, *Increased cardiovascular risk in patients with chronic obstructive pulmonary disease and the potential mechanisms linking the two conditions: a review.* Cardiol Rev, 2013. **21**(4): p. 196-202.

254. Mocchegiani, E., R. Giacconi, and L. Costarelli, *Metalloproteases/anti-metalloproteases imbalance in chronic obstructive pulmonary disease: genetic factors and treatment implications.* Curr Opin Pulm Med, 2011. **17 Suppl 1**: p. S11-9.
255. Gosselink, J.V., et al., *Differential expression of tissue repair genes in the pathogenesis of chronic obstructive pulmonary disease.* Am J Respir Crit Care Med, 2010. **181**(12): p. 1329-35.
256. Omrane, I., et al., *Significant association between IL23R and IL17F polymorphisms and clinical features of colorectal cancer.* Immunol Lett, 2014. **158**(1-2): p. 189-94.
257. Wu, X., et al., *Association between polymorphisms in interleukin-17A and interleukin-17F genes and risks of gastric cancer.* Int J Cancer, 2010. **127**(1): p. 86-92.
258. Wang, Z., E.P. Bishop, and P.A. Burke, *Expression profile analysis of the inflammatory response regulated by hepatocyte nuclear factor 4alpha.* BMC Genomics, 2011. **12**: p. 128.
259. Marconett, C.N., et al., *Integrated transcriptomic and epigenomic analysis of primary human lung epithelial cell differentiation.* PLoS Genet, 2013. **9**(6): p. e1003513.
260. Sugano, M., et al., *HNF4alpha as a marker for invasive mucinous adenocarcinoma of the lung.* Am J Surg Pathol, 2013. **37**(2): p. 211-8.
261. Snyder, E.L., et al., *Nkx2-1 represses a latent gastric differentiation program in lung adenocarcinoma.* Mol Cell, 2013. **50**(2): p. 185-99.
262. Savarimuthu Francis, S.M., et al., *MicroRNA-34c is associated with emphysema severity and modulates SERPINE1 expression.* BMC Genomics, 2014. **15**: p. 88.
263. Mu, W. and L. Hui, *Establishing a cancer cell in the inflammatory tissue: an epigenetic circuit.* Acta Biochim Biophys Sin (Shanghai), 2012. **44**(4): p. 279-80.
264. Needham, M. and R.A. Stockley, *Alpha 1-antitrypsin deficiency. 3: Clinical manifestations and natural history.* Thorax, 2004. **59**(5): p. 441-5.
265. Ma, W., et al., *GCIP/CCNDBP1, a helix-loop-helix protein, suppresses tumorigenesis.* J Cell Biochem, 2007. **100**(6): p. 1376-86.
266. Trapp, J. and M. Jung, *The role of NAD+ dependent histone deacetylases (sirtuins) in ageing.* Curr Drug Targets, 2006. **7**(11): p. 1553-60.
267. Cardus, A., et al., *SIRT6 protects human endothelial cells from DNA damage, telomere dysfunction, and senescence.* Cardiovasc Res, 2013. **97**(3): p. 571-9.
268. Takasaka, N., et al., *Autophagy induction by SIRT6 through attenuation of insulin-like growth factor signaling is involved in the regulation of human bronchial epithelial cell senescence.* J Immunol, 2014. **192**(3): p. 958-68.
269. Yao, H., et al., *SIRT1 redresses the imbalance of tissue inhibitor of matrix metalloproteinase-1 and matrix metalloproteinase-9 in the development of mouse emphysema and human COPD.* Am J Physiol Lung Cell Mol Physiol, 2013. **305**(9): p. L615-24.
270. Sundar, I.K., H. Yao, and I. Rahman, *Oxidative stress and chromatin remodeling in chronic obstructive pulmonary disease and smoking-related diseases.* Antioxid Redox Signal, 2013. **18**(15): p. 1956-71.
271. Austin, S. and J. St-Pierre, *PGC1alpha and mitochondrial metabolism--emerging concepts and relevance in ageing and neurodegenerative disorders.* J Cell Sci, 2012. **125**(Pt 21): p. 4963-71.

272. Li, J., et al., *Positive correlation between PPARgamma/PGC-1alpha and gamma-GCS in lungs of rats and patients with chronic obstructive pulmonary disease.* Acta Biochim Biophys Sin (Shanghai), 2010. **42**(9): p. 603-14.
273. Vlahos, R. and S. Bozinovski, *Recent advances in pre-clinical mouse models of COPD.* Clin Sci (Lond), 2014. **126**(4): p. 253-65.
274. National Lung Screening Trial Research, T., et al., *Reduced lung-cancer mortality with low-dose computed tomographic screening.* N Engl J Med, 2011. **365**(5): p. 395-409.
275. National Lung Screening Trial Research, T., et al., *Results of initial low-dose computed tomographic screening for lung cancer.* N Engl J Med, 2013. **368**(21): p. 1980-91.
276. Aran, D. and A. Hellman, *DNA methylation of transcriptional enhancers and cancer predisposition.* Cell, 2013. **154**(1): p. 11-3.
277. Camp, P.G., et al., *COPD phenotypes in biomass smoke- versus tobacco smoke-exposed Mexican women.* Eur Respir J, 2014. **43**(3): p. 725-34.
278. Molina-Pinelo, S., et al., *MicroRNA clusters: dysregulation in lung adenocarcinoma and COPD.* Eur Respir J, 2014.
279. Leung, J.M. and D.D. Sin, *Biomarkers in airway diseases.* Can Respir J, 2013. **20**(3): p. 180-2.
280. Barnes, P.J., *Histone deacetylase-2 and airway disease.* Ther Adv Respir Dis, 2009. **3**(5): p. 235-43.
281. Mercado, N., et al., *Decreased histone deacetylase 2 impairs Nrf2 activation by oxidative stress.* Biochem Biophys Res Commun, 2011. **406**(2): p. 292-8.
282. Couraud, S., et al., *Lung cancer in never smokers--a review.* Eur J Cancer, 2012. **48**(9): p. 1299-311.
283. Aberle, D.R., et al., *Reduced lung-cancer mortality with low-dose computed tomographic screening.* N Engl J Med, 2011. **365**(5): p. 395-409.
284. McWilliams, A., et al., *Probability of cancer in pulmonary nodules detected on first screening CT.* N Engl J Med, 2013. **369**(10): p. 910-9.
285. Sawyers, C.L., *The cancer biomarker problem.* Nature, 2008. **452**(7187): p. 548-52.
286. Poste, G., *Bring on the biomarkers.* Nature, 2011. **469**(7329): p. 156-7.
287. Meckley, L.M. and P.J. Neumann, *Personalized medicine: factors influencing reimbursement.* Health Policy, 2010. **94**(2): p. 91-100.
288. Liotta, L.A. and E. Petricoin, *Cancer Biomarkers: Closer to Delivering on their Promise.* Cancer Cell, 2011. **20**(3): p. 279-80.
289. Taguchi, A., et al., *Lung cancer signatures in plasma based on proteome profiling of mouse tumor models.* Cancer Cell, 2011. **20**(3): p. 289-99.
290. Jones, S.J., et al., *Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors.* Genome Biol, 2010. **11**(8): p. R82.
291. Lunshof, J.E., et al., *Personal genomes in progress: from the human genome project to the personal genome project.* Dialogues Clin Neurosci, 2010. **12**(1): p. 47-60.
292. Pasche, B. and D. Absher, *Whole-genome sequencing: a step closer to personalized medicine.* JAMA, 2011. **305**(15): p. 1596-7.
293. Welch, J.S., et al., *Use of whole-genome sequencing to diagnose a cryptic fusion oncogene.* JAMA, 2011. **305**(15): p. 1577-84.

294. Dancey, J., *Biomarker Discovery and Development through Genomics*, in *Cancer Genomics: From Bench to Personalized Medicine*, B.a.A. Dellaire, Editor. 2014, Elsevier.
295. Kristensen, V.N., et al., *Principles and methods of integrative genomic analyses in cancer.* Nat Rev Cancer, 2014. **14**(5): p. 299-313.
296. Gerber, D.E. and J.D. Minna, *ALK inhibition for non-small cell lung cancer: from discovery to therapy in record time.* Cancer Cell, 2010. **18**(6): p. 548-51.
297. Chin, L., J.N. Andersen, and P.A. Futreal, *Cancer genomics: from discovery science to personalized medicine.* Nat Med, 2011. **17**(3): p. 297-303.
298. Soda, M., et al., *Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer.* Nature, 2007. **448**(7153): p. 561-6.
299. Kwak, E.L., et al., *Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer.* N Engl J Med, 2010. **363**(18): p. 1693-703.
300. Haber, D.A., N.S. Gray, and J. Baselga, *The evolving war on cancer.* Cell, 2011. **145**(1): p. 19-24.
301. Lynch, T.J., et al., *Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib.* N Engl J Med, 2004. **350**(21): p. 2129-39.

Appendices

Appendix A miRNA frequently altered in all tumour smoking groups

† Previously described in lung cancer

Overexpressed in all tumours	hsa-mir-187†; hsa-mir-376a; hsa-mir-376c; hsa-mir-136†; hsa-mir-196b†; hsa-mir-224†; hsa-mir-154†; hsa-mir-370; hsa-mir-432; hsa-mir-487b; hsa-mir-494; hsa-mir-31†; hsa-mir-149†; hsa-mir-627†; hsa-mir-95; hsa-mir-409; hsa-mir-431; hsa-mir-377†; hsa-mir-412†; hsa-mir-1180†; hsa-mir-1277; hsa-mir-323b; hsa-mir-365b†; hsa-mir-376b; hsa-mir-493; hsa-mir-127†; hsa-mir-539†; hsa-mir-382; hsa-mir-134†; hsa-mir-487a; hsa-mir-193b†; hsa-mir-214†; hsa-mir-1248; hsa-mir-339†; hsa-mir-3607; hsa-mir-3676; hsa-mir-642a; hsa-mir-503; hsa-mir-1296†; hsa-mir-551a; hsa-mir-616; hsa-mir-320d; hsa-mir-4454; hsa-mir-485; hsa-mir-92b†; hsa-mir-205†; hsa-mir-219†; hsa-mir-3127; hsa-mir-3200; hsa-mir-425†; hsa-mir-489; hsa-mir-19b†; hsa-mir-3651; hsa-mir-887; hsa-mir-1269a; hsa-mir-196a†; hsa-mir-222†; hsa-mir-491; hsa-mir-500b; hsa-mir-577; hsa-mir-625; hsa-mir-148a†; hsa-mir-192†; hsa-mir-2110†; hsa-mir-212†; hsa-mir-24†; hsa-mir-26b†; hsa-mir-3615; hsa-mir-483†; hsa-mir-1271†; hsa-mir-3677; hsa-mir-3909; hsa-mir-5010; hsa-mir-29a†; hsa-mir-365a†; hsa-mir-4661; hsa-mir-4724; hsa-mir-98; hsa-mir-450b; hsa-mir-939; hsa-mir-629; hsa-mir-130a†; hsa-mir-3157; hsa-mir-4677; hsa-mir-1251; hsa-mir-125b†; hsa-mir-210†; hsa-mir-135b†; hsa-mir-188; hsa-mir-96; hsa-mir-1226†; hsa-mir-130b†; hsa-mir-141†; hsa-mir-182†; hsa-mir-200a†; hsa-mir-331†; hsa-mir-33b; hsa-mir-345†; hsa-mir-877; hsa-mir-1301†; hsa-mir-1306; hsa-mir-183†; hsa-mir-29b†; hsa-mir-424; hsa-mir-429; hsa-mir-708; hsa-mir-874; hsa-mir-147b†; hsa-mir-191†; hsa-mir-200b†; hsa-mir-301a†; hsa-mir-301b†; hsa-mir-324; hsa-mir-33a†; hsa-mir-3605; hsa-mir-423; hsa-mir-4326; hsa-mir-4728; hsa-mir-671; hsa-mir-940; hsa-mir-1307; hsa-mir-296†; hsa-mir-34a†; hsa-mir-450a; hsa-mir-744; hsa-mir-766; hsa-mir-937; hsa-mir-200c†; hsa-mir-3648; hsa-mir-3653; hsa-mir-421; hsa-mir-505; hsa-mir-1228†; hsa-mir-1249†; hsa-mir-1266; hsa-mir-5001; hsa-mir-644b; hsa-mir-9; hsa-mir-1291†; hsa-mir-18a†; hsa-mir-197†; hsa-mir-19a†; hsa-mir-21†; hsa-mir-2277; hsa-mir-342†; hsa-mir-375; hsa-mir-454; hsa-mir-455; hsa-mir-5698; hsa-mir-589; hsa-mir-590; hsa-mir-628†; hsa-mir-320b†; hsa-mir-323a; hsa-mir-484; hsa-mir-574; hsa-mir-760; hsa-mir-769; hsa-let-7i†; hsa-mir-1229†; hsa-mir-3170; hsa-mir-3194; hsa-mir-4449; hsa-mir-4668; hsa-mir-615; hsa-mir-146a†; hsa-mir-153†; hsa-mir-199a†; hsa-mir-3189; hsa-mir-328†; hsa-mir-3928; hsa-mir-548v; hsa-mir-106a†; hsa-mir-1287; hsa-mir-3687; hsa-mir-4787; hsa-mir-1270; hsa-mir-1275†; hsa-mir-148b†; hsa-mir-17†; hsa-mir-199b†; hsa-mir-5699; hsa-let-7g†; hsa-mir-1343; hsa-mir-137†; hsa-mir-3917; hsa-mir-16†; hsa-mir-320c; hsa-mir-3617; hsa-mir-550a; hsa-mir-106b†; hsa-mir-1254†; hsa-mir-128†; hsa-mir-2116†; hsa-mir-4638; hsa-mir-4652
Underexpressed in all tumours	hsa-mir-5683; hsa-mir-143†; hsa-mir-144†; hsa-mir-30a†; hsa-mir-451a†; hsa-mir-374a†; hsa-mir-486; hsa-mir-584; hsa-mir-139†; hsa-mir-101†; hsa-mir-190a†; hsa-mir-100†; hsa-mir-1258†; hsa-mir-1†; hsa-mir-218†; hsa-mir-223†; hsa-mir-1247; hsa-mir-10b†; hsa-mir-204†; hsa-let-7c†; hsa-mir-338†; hsa-mir-5586; hsa-mir-206†; hsa-mir-4732; hsa-mir-133a†; hsa-let-7a†; hsa-mir-30d†; hsa-mir-4772; hsa-mir-99a†; hsa-mir-202†; hsa-mir-126†; hsa-mir-133b†; hsa-mir-598; hsa-mir-374b; hsa-mir-4521; hsa-mir-490

Appendix B miRNA frequently altered in one or more smoking tumour groups

	Tumour Subgroup	miRNA
Overexpressed	CS	hsa-mir-654; hsa-mir-891a; hsa-mir-18b†; hsa-mir-215; hsa-mir-592†; hsa-mir-129; hsa-mir-576; hsa-mir-337; hsa-mir-411; hsa-mir-545; hsa-mir-7†; hsa-mir-372; hsa-mir-3940; hsa-mir-5571
	CS, FS	hsa-mir-449a†; hsa-mir-758; hsa-mir-20a†; hsa-mir-335†; hsa-mir-99b; hsa-mir-556; hsa-mir-27b†; hsa-mir-3664; hsa-mir-579; hsa-mir-433; hsa-mir-449b; hsa-mir-570; hsa-mir-659; hsa-mir-3136; hsa-mir-4664
	CS, NS	hsa-mir-369; hsa-mir-496; hsa-mir-142†; hsa-mir-3913; hsa-mir-4746; hsa-mir-543; hsa-mir-1269b
	FS	hsa-mir-504†; hsa-mir-1262; hsa-mir-944; hsa-mir-27a†; hsa-mir-23c; hsa-mir-151b; hsa-mir-190b†; hsa-mir-320a†; hsa-mir-3187; hsa-mir-514b; hsa-mir-105†; hsa-mir-632
	FS, NS	hsa-mir-3614; hsa-mir-4461; hsa-mir-548b; hsa-mir-4709; hsa-mir-326†; hsa-mir-452; hsa-mir-502; hsa-mir-125a†; hsa-mir-155†; hsa-mir-3934; hsa-mir-501; hsa-mir-1224†; hsa-mir-146b†; hsa-mir-362; hsa-mir-181b†; hsa-mir-186†; hsa-mir-2355; hsa-mir-542; hsa-mir-221†; hsa-mir-4685; hsa-mir-1227†; hsa-mir-4444; hsa-mir-3064; hsa-mir-4758; hsa-mir-664; hsa-mir-1237; hsa-mir-3610; hsa-mir-3652; hsa-mir-1292†; hsa-mir-4446; hsa-mir-4690; hsa-mir-4742; hsa-mir-330†; hsa-mir-3620
	NS	hsa-mir-150†; hsa-mir-185†; hsa-mir-675; hsa-mir-217†; hsa-mir-660; hsa-mir-3130; hsa-mir-500a; hsa-mir-3150b; hsa-mir-3613; hsa-mir-1976; hsa-mir-652†; hsa-mir-152†; hsa-mir-5701; hsa-mir-340; hsa-mir-216a†; hsa-mir-3609; hsa-mir-1295a†; hsa-mir-329†; hsa-mir-4791; hsa-mir-636; hsa-mir-93†; hsa-mir-2114; hsa-mir-4443; hsa-mir-320e; hsa-mir-612
Underexpressed	CS	hsa-mir-511; hsa-mir-135a†; hsa-mir-4532; hsa-mir-3065; hsa-mir-138†; hsa-mir-4536; hsa-mir-378a†; hsa-mir-676; hsa-mir-195†; hsa-mir-378c†
	CS, FS	hsa-let-7b†; hsa-mir-10a†; hsa-mir-184†; hsa-let-7f†; hsa-mir-4662a; hsa-mir-140†; hsa-mir-5000; hsa-mir-1468†; hsa-mir-532; hsa-mir-34c†; hsa-mir-34b; hsa-mir-3926
	FS	hsa-mir-381; hsa-mir-607
	FS, NS	hsa-mir-203†
	NS	hsa-mir-582

Appendix C miRNA identified as having significant associations between miRNA expression and lung AC patient survival

*B-H corrected p-value
‡ considering uncorrected p-value for the All Lung AC group

	Status in Tumours			Mantel-Haenszel p value					
miRNA	**CST**	**FST**	**NST**	**All Lung AC**	**All Lung AC***	**CS**	**FS**	**NS**	**Group‡**
hsa-let-7a	UE	UE	UE	0.0664	0.1901	0.1493	0.019	0.197	FS
hsa-let-7f	UE	UE		0.0244	0.1121	0.3043	0.0727	0.7148	ALL
hsa-let-7g	OE	OE	OE	0.0006	0.0335	0.1894	0.0438	0.0036	ALL+FS+NS
hsa-let-7i	OE	OE	OE	0.0417	0.1416	0.0887	0.4869	0.2138	ALL
hsa-mir-1	UE	UE	UE	0.0021	0.0449	0.8863	0.0179	0.0565	ALL+FS
hsa-mir-106a	OE	OE	OE	0.1324	0.3059	0.3771	0.3858	0.0383	NS
hsa-mir-10a	UE	UE		0.7948	0.8639	0.4552	0.7071	0.0342	NS
hsa-mir-1226	OE	OE	OE	0.1119	0.2675	0.9905	0.0493	0.7862	FS
hsa-mir-1247	UE	UE	UE	0.0027	0.0498	0.2228	0.0066	0.6707	ALL+FS
hsa-mir-1249	OE	OE	OE	0.0203	0.1055	0.1338	0.4165	0.16	ALL
hsa-mir-125a		OE	OE	0.0136	0.0889	0.2276	0.0122	0.4137	ALL+FS
hsa-mir-126	UE	UE	UE	0.0167	0.0975	0.6402	0.0902	0.2882	ALL
hsa-mir-1270	OE	OE	OE	0.8638	0.8964	0.0462	na	0.854	CS
hsa-mir-1287	OE	OE	OE	0.0048	0.0473	0.0022	0.169	0.4399	ALL+CS
hsa-mir-1301	OE	OE	OE	0.0214	0.1051	0.2644	0.0122	0.3624	ALL+FS
hsa-mir-1306	OE	OE	OE	0.1964	0.3942	0.5442	0.048	0.9494	FS
hsa-mir-130a	OE	OE	OE	0.0677	0.1881	0.1222	0.028	0.4554	FS
hsa-mir-133a	UE	UE	UE	0.0096	0.0756	0.6305	0.0001	0.3569	ALL+FS
hsa-mir-133b	UE	UE	UE	0.0414	0.1423	na	na	0.0304	ALL+NS
hsa-mir-135b	OE	OE	OE	0.079	0.213	0.7256	0.0021	0.1	FS
hsa-mir-136	OE	OE	OE	0.1847	0.3877	0.7652	0.0235	0.1303	FS
hsa-mir-138	UE			0.0145	0.0905	0.0086	0.2916	0.6801	ALL+CS
hsa-mir-139	UE	UE	UE	0.0089	0.072	0.7048	0.0567	0.02	ALL+NS
hsa-mir-142	OE		OE	0.0247	0.1115	0.8207	0.0165	0.3855	ALL+FS
hsa-mir-143	UE	UE	UE	0.0177	0.1014	0.0997	0.2432	0.906	ALL
hsa-mir-1468	UE	UE		0.0031	0.0453	0.0567	0.0301	0.2268	ALL+FS
hsa-mir-146a	OE	OE	OE	0.0048	0.0487	0.1486	0.1216	0.2677	ALL
hsa-mir-148a	OE	OE	OE	0.0268	0.119	0.0429	0.1372	0.2013	ALL+CS
hsa-mir-149	OE	OE	OE	0.0342	0.1305	0.2459	0.0348	0.4066	ALL+FS
hsa-mir-150			OE	0.0073	0.067	0.8039	0.0311	0.0143	ALL+FS+NS
hsa-mir-153	OE	OE	OE	0.0006	0.028	0.9752	0.0006	0.3777	ALL+FS
hsa-mir-16	OE	OE	OE	0.0378	0.1333	0.1836	0.0109	0.6305	ALL+FS
hsa-mir-184	UE	UE		0.0036	0.0445	na	0.247	0.0103	ALL+NS
hsa-mir-186		OE	OE	0.1288	0.3001	0.9525	0.02	0.0534	FS
hsa-mir-187	OE	OE	OE	0.0002	0.0182	0.1882	0.001	0.7241	ALL+FS
hsa-mir-18a	OE	OE	OE	0.0276	0.1204	0.3858	0.5285	0.6604	ALL
hsa-mir-191	OE	OE	OE	0.0338	0.1308	0.532	0.0403	0.4667	ALL+FS
hsa-mir-195	UE			0.0071	0.0671	0.0436	0.0151	0.8953	ALL+CS+FS
hsa-mir-200a	OE	OE	OE	0.0205	0.1045	0.4595	0.0047	0.262	ALL+FS
hsa-mir-200b	OE	OE	OE	0.0023	0.0461	0.5559	0.0063	0.1569	ALL+FS
hsa-mir-204	UE	UE	UE	0.0035	0.0462	0.0291	0.0093	0.7435	ALL+CS+FS
hsa-mir-21	OE	OE	OE	0.0082	0.0687	0.323	0.0013	0.4412	ALL+FS
hsa-mir-2110	OE	OE	OE	0.003	0.0457	0.0636	0.011	0.1423	ALL+FS
hsa-mir-212	OE	OE	OE	0.1104	0.2686	0.2352	0.0056	0.3233	FS
hsa-mir-24	OE	OE	OE	0.0224	0.108	0.3106	0.0117	0.5308	ALL+FS
hsa-mir-26b	OE	OE	OE	0.0001	0.0194	0.0634	0.0018	0.5666	ALL+FS
hsa-mir-27a		OE		0.0321	0.128	0.1226	0.2255	0.2672	ALL
hsa-mir-27b	OE	OE		0.013	0.0874	0.8206	0.011	0.1543	ALL+FS
hsa-mir-296	OE	OE	OE	0.0039	0.0462	0.0854	0.0208	0.1344	ALL+FS

*B-H corrected p-value
‡ considering uncorrected p-value for the All Lung AC group

	Status in Tumours			Mantel-Haenszel p value					
miRNA	**CST**	**FST**	**NST**	**All Lung AC**	**All Lung AC***	**CS**	**FS**	**NS**	**Group‡**
hsa-mir-29a	OE	OE	OE	0.0229	0.1067	0.2196	0.0387	0.1769	ALL+FS
hsa-mir-29b	OE	OE	OE	0.0127	0.0874	0.1006	0.0054	0.1964	ALL+FS
hsa-mir-301a	OE	OE	OE	0.0647	0.1891	0.4031	0.0271	0.2524	FS
hsa-mir-3065	UE			0.0021	0.0484	0.09	0.0128	0.2516	ALL+FS
hsa-mir-30d	UE	UE	UE	0.4038	0.5876	0.0282	0.2891	0.2239	CS
hsa-mir-320d	OE	OE	OE	0.0371	0.1342	0.5347	na	0.6182	ALL
hsa-mir-324	OE	OE	OE	0.2906	0.4995	0.0996	0.0069	0.2922	FS
hsa-mir-326		OE	OE	0.2404	0.4497	0.0131	0.5943	0.6563	CS
hsa-mir-328	OE	OE	OE	0.0139	0.0889	0.052	0.0043	0.9766	ALL+FS
hsa-mir-331	OE	OE	OE	0.0007	0.0254	0.0146	0.0054	0.6023	ALL+CS+FS
hsa-mir-338	UE	UE	UE	0.0332	0.1304	0.7063	0.1734	0.0006	ALL+NS
hsa-mir-339	OE	OE	OE	0.0003	0.0189	0.0783	0.006	0.7159	ALL+FS
hsa-mir-33a	OE	OE	OE	0.0353	0.1313	0.3898	0.0667	0.885	ALL
hsa-mir-33b	OE	OE	OE	0.0291	0.1195	0.812	0.0054	0.9763	ALL+FS
hsa-mir-342	OE	OE	OE	0.0281	0.1206	0.2682	0.0127	0.2081	ALL+FS
hsa-mir-345	OE	OE	OE	0.0844	0.2149	0.6497	0.0082	0.2331	FS
hsa-mir-34a	OE	OE	OE	0.0029	0.0503	0.1892	0.0065	0.168	ALL+FS
hsa-mir-34b	UE		UE	0.0177	0.0975	0.0849	0.0078	0.0663	ALL+FS
hsa-mir-3607	OE	OE	OE	0.0013	0.0319	0.2897	0.0022	0.9309	ALL+FS
hsa-mir-3613			OE	0.0545	0.1685	0.9888	0.0052	0.6258	FS
hsa-mir-362		OE	OE	0.0152	0.0931	0.4064	0.0875	0.7775	ALL
hsa-mir-3653	OE	OE	OE	0.0044	0.0485	0.2089	0.0067	0.3956	ALL+FS
hsa-mir-3676	OE	OE	OE	0.0047	0.0498	0.0855	0.0602	0.4167	ALL
hsa-mir-369	OE		OE	0.7974	0.8633	0.4857	0.0155	0.2601	FS
hsa-mir-374b	UE	UE	UE	0.008	0.0688	0.0483	0.0885	0.41	ALL+CS
hsa-mir-375	OE	OE	OE	0.0177	0.0993	0.7243	0.0387	0.2538	ALL+FS
hsa-mir-376c	OE	OE	OE	0.0965	0.239	0.9687	0.0076	0.4326	FS
hsa-mir-423	OE	OE	OE	0.0181	0.0956	0.1645	0.0476	0.6897	ALL+FS
hsa-mir-425	OE	OE	OE	0.0161	0.0962	0.1765	0.0415	0.3394	ALL+FS
hsa-mir-429	OE	OE	OE	0.0107	0.0796	0.7818	0.0002	0.696	ALL+FS
hsa-mir-432	OE	OE	OE	0.4154	0.592	0.3711	0.0048	0.5628	FS
hsa-mir-454	OE	OE	OE	0.0427	0.1413	0.3653	0.2962	0.7588	ALL
hsa-mir-484	OE	OE	OE	0.0394	0.1371	0.3633	0.011	0.7589	ALL+FS
hsa-mir-486	UE	UE	UE	0.03	0.1213	0.9594	0.0565	0.0457	ALL+NS
hsa-mir-491	OE	OE	OE	0.0079	0.0701	0.1354	0.0137	0.2151	ALL+FS
hsa-mir-493	OE	OE	OE	0.441	0.6004	0.9771	0.0344	0.1376	FS
hsa-mir-502		OE	OE	0.01	0.0764	0.0883	0.1188	0.8083	ALL
hsa-mir-505	OE	OE	OE	0.0661	0.1914	0.3847	0.0257	0.1765	FS
hsa-mir-539	OE	OE	OE	0.4244	0.5895	0.0526	0.0056	0.3633	FS
hsa-mir-548b		OE	OE	0.0363	0.1333	0.0512	0.0498	0.5446	ALL+FS
hsa-mir-574	OE	OE	OE	0.0352	0.1326	0.0521	0.1069	0.8262	ALL
hsa-mir-590	OE	OE	OE	0.0212	0.1058	0.3105	0.2123	0.9051	ALL
hsa-mir-598	UE	UE	UE	0.2408	0.4475	0.6718	0.0447	0.0326	FS+NS
hsa-mir-628	OE	OE	OE	0.018	0.0972	0.295	0.0374	0.179	ALL+FS
hsa-mir-642a	OE	OE	OE	0.0113	0.0816	0.8041	0.0005	0.3554	ALL+FS
hsa-mir-664		OE	OE	0.0012	0.0342	0.0331	0.0128	0.0307	ALL
hsa-mir-7	OE			0.1757	0.3774	0.9604	0.0379	0.7983	FS
hsa-mir-769	OE	OE	OE	0.012	0.0844	0.1469	0.0852	0.3701	ALL
hsa-mir-887	OE	OE	OE	0.0433	0.1402	0.6015	0.0085	0.1952	ALL+FS
hsa-mir-92b	OE	OE	OE	0.6739	0.7721	0.6029	0.0378	0.0351	FS+NS
hsa-mir-940	OE	OE	OE	0.0289	0.1205	0.4283	0.0656	0.6866	ALL
hsa-mir-95	OE	OE	OE	0.0034	0.0463	0.4213	0.0306	0.2897	ALL+FS
hsa-mir-96	OE	OE	OE	0.0428	0.1401	0.2864	0.037	0.0631	ALL+FS
hsa-mir-99b	OE	OE		0.0639	0.191	0.0639	0.0027	0.8974	FS

Appendix D Assessment of DNA change on expression fold change and difference in expression fold change between scoring bins

CN: copy number; HypoMeth: Hypomethylation; HyperMeth: Hypermethylation; Expr OE: overexpressed; Expr UE: underexpressed

WEIGHTS	Colon	Lung AC	Breast	Lung SQ	Average
CN Gain on Expr	1.91	2.17	2.27	2.25	2.15
CN Loss on Expr	2.06	2.66	3.39	3.16	2.82
HypoMeth on Expr	2.25	2.25	2.76	2.66	2.48
HyperMeth on Expr	2.77	2.69	4.30	4.82	3.65
SCORES	**Colon**	**Lung AC**	**Breast**	**Lung SQ**	**Average**
CN Gain Bin 1	1.81	2.11	2.24	2.11	2.07
CN Gain Bin 2	2.26	2.15	2.39	2.97	2.44
CN Loss Bin 1	1.97	2.65	3.34	3.12	2.77
CN Loss Bin 2	2.06	2.69	4.18	4.11	3.26
Meth Hypo Bin 1	2.23	2.25	2.74	2.64	2.47
Meth Hypo Bin 2	2.46	2.22	3.02	4.48	3.04
Meth Hyper Bin 1	2.69	2.68	4.23	4.79	3.60
Meth Hyper Bin 2	2.88	3.09	9.20	6.59	5.44
Expr OE Bin 1	2.16	2.29	2.33	2.30	2.27
Expr OE Bin 2	4.41	4.88	4.79	4.98	4.77
Expr OE Bin 3	13.68	13.23	13.72	18.00	14.66
Expr OE Bin 4	66.32	80.91	69.54	107.46	81.06
Expr UE Bin 1	2.23	2.48	2.60	2.53	2.46
Expr UE Bin 2	4.37	4.84	5.32	5.37	4.97
Expr UE Bin 3	12.35	13.00	14.19	16.48	14.01
Expr UE Bin 4	78.83	63.58	71.83	91.55	76.45
Expression Fold Change Between Bins	**Colon**	**Lung AC**	**Breast**	**Lung SQ**	**Average**
CN Gain Bin 1 vs 2	1.25	1.02	1.07	1.40	1.18
CN Loss Bin 1 vs 2	1.05	1.02	1.25	1.32	1.16
Hypomethylation Bin 1 vs 2	1.10	0.98	1.10	1.70	1.22
Hypermethylation Bin 1 vs 2	1.07	1.15	2.17	1.38	1.44

Appendix E List of Publications

This Appendix lists all of the publications I contributed to during my degree that were either published, accepted or are currently under review. First or co-first authorships are underlined.

1. Garnis C, Lockwood WW, **Vucic E**, Ge Y, Girard L, Minna JD, Gazdar AF, Lam S, MacAulay C, Lam WL. High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH. *International Journal of Cancer*. 2006; 118: 1556-1564.

2. Kuo A, Wilson IM, **Vucic E**, Lee E, Davies JJ, MacAulay C, Brown C, Lam WL. Comparative cancer epigenomics. In *Comparative Genomics: Fundamental and Applied Perspectives.* 2008; James R. Brown (ed.) CRC Press, Taylor & Francis Publisher. 261-279.

3. **Vucic EA**[†], Brown CJ, Lam WL. Epigenomics of Cancer Progression. *Pharmacogenomic*s. 2008; 9: 215-234. [†]**corresponding author**

4. Chari R, Coe BP, Wedseltoft C, Benetti M, Wilson IM, **Vucic E**, MacAulay C, Ng RT, Lam WL. SIGMA2: A system for the integrative genomic multi-dimensional analysis of cancer genomes, epigenomes and transcriptomes. *BMC Bioinformatics*. 2008; 9: 422, 1-12.

5. Thu KL*, **Vucic EA***, Kennett JY, Heryet C, Brown CJ, Lam WL, Wilson IM. Methylated DNA Immunoprecipitation *Journal of Visualized Experiments*. 2009; 23. http://www.jove.com/index/details.stp?id=935, doi: 10.3791/935. ***co-first author**

6. **Vucic EA**[†], Wilson IM, Campbell JM, Lam WL. Methylation analysis by DNA immunoprecipitation (MeDIP). *Methods in Molecular Biology*. 2009; 556: 141-153. [†]**corresponding author**

7. Chari R, Coe BP, **Vucic EA**, Lockwood WW, Lam WL. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Systems Biology*. 2010; 4:67, 1-14.

8. **Vucic EA**[†], Thu K, Williams AC, Lam WL, Coe BP. Copy Number Variations in the Human Genome and Strategies for Analysis. *Methods in Molecular Biology*. 2010; 628: 103-117. [†]**corresponding author**

9. Martinez VD, **Vucic EA**, Adonis M, Gil L, Lam WL. Arsenic Biotransformation as a Cancer Promoting Factor by Inducing DNA Damage and Disruption of Repair Mechanisms. *Molecular Biology International*. 2011; 718974.

10. Gibb EA, **Vucic EA**, Enfield KSS, Stewart GL, Lonergan KM, Kennett JY, Becker Santos DD, MacAulay CE, Lam S, Brown CJ, Lam WL. Human cancer long non-coding RNA transcriptomes. *PLoS One*. 2011; 6(10):e25915. Epub 2011 Oct 3.

11. Martinez VD, **Vucic EA**, Becker-Santos DD, Gil L, Lam WL. Arsenic Exposure and the Induction of Human Cancers. *Journal of Toxicology*. 2011; Article ID 431287.

12. Martinez VD, Becker-Santos DD, **Vucic EA**, Lam S, Lam WL. Induction of Human Squamous Cell-Type Carcinomas by Arsenic. *Journal of Skin Cancer*. 2011; Article ID 454157.

13. **Vucic EA*†**, Thu KL*, Robison K, Rybaczyk LA, Chari R, Alvarez CE, Lam WL. Translating cancer 'omics to improved outcomes. *Genome Research.* 2012; Feb 22(2):188-95.
***co-first authors, †corresponding author**

14. Thu, KL, **Vucic EA**, Chari R, Zhang W, Lockwood WW, English JC, MacAulay CE, Gazdar AF, Lam S, Lam WL. Lung adenocarcinomas of never smokers and smokers are genomically distinct. *PLoS One.* 2012; 7(3):e33003. Epub 2012 Mar 7.

15. Martinez VD, **Vucic EA**, Lam S, Lam WL. Arsenic and lung cancer in never smokers: Lessons from Chile. *American Journal of Respiratory and Critical Care Medicine.* 2012; May 15 (vol. 185) no. 10, 1131-1132.

16. Dong X, Dong L, Low C, **Vucic E**, English J, Yee J, Lam WL, Ling V, Lam S, Gout PW, Wang YZ. Elevated expression of BIRC6 protein in non-small cell lung cancers is associated with cancer recurrence and chemoresistance. *Journal of Thoracic Oncology.* 2013; Feb; 8(2):161-70.

17. Coe BP*, Thu KL*, Aviel-Ronen S, **Vucic EA**, Gazdar AF, Lam S, Tsao MS, Lam WL. Genomic deregulation of the E2F/Rb pathway leads to activation of the oncogene EZH2 in small cell lung cancer. *PLoS One.* 2013; 8(8):e71670. *co-first authors

18. Wilson I*, **Vucic EA***, Enfield KSS, Zhang YA, Chari R, Thu KL, Lockwood WW, Radulovich N, Starczynowski D, Banath JP, Zhang M, Pusic A, Fuller M, Lonergan KM, Yee J, English JC, Buys TPH, Selamat SA, Laird-Offringa IA, Liu P, Anderson M, You M, Tsao MS, Brown CJ, Bennewith KL, MacAulay CE, Karson A, Gazdar AF, Lam S, Lam WL. EYA4 is inactivated biallelically at a high frequency in sporadic lung cancer and is associated with familial lung cancer risk. *Oncogene.* 2013; Oct 7. doi: 10.1038/onc.2013.396. ***co-first authors**

19. Martinez VD, Thu KL, **Vucic EA**, Hubaux R, Adonis MI, Gil L, MacAulay CE, Lam S, Lam WL. Whole genome sequencing analysis identifies a distinctive mutational spectrum in an arsenic-related lung tumor. *Journal of Thoracic Oncology.* 2013; Nov;8(11):1451-5.

20. Martinez VD*, **Vucic EA***, Pikor LA, Thu KL, Hubaux R, Lam WL. Frequent concerted genetic mechanisms disrupt multiple components of the NRF2 inhibitor KEAP1/CUL3/RBX1 E3-ubiquitin ligase complex in thyroid cancer. *Molecular Cancer.* 2013; Oct 20;12(1):124. ***co-first authors**

21. Pikor LA, Lockwood WW, Thu KL, **Vucic EA**, Chari R, Gazdar AF, Lam S, Lam WL YEATS4 is a novel oncogene amplified in non-small cell lung cancer that regulates the p53 pathway. *Cancer Research.* 2013; Oct 29. [Epub ahead of print]

22. Martinez VD, **Vucic EA**, Lam S, Lam WL. Emerging Arsenic Threat in Canada. *Science.* 2013; Nov 1; 342(6158):559.

23. Pikor L, Thu K, **Vucic E**, Lam WL. The detection and implication of genome instability in cancer. *Cancer Metastasis Reviews.* 2013; Dec; 32(3-4):341-52.

24. Martinez VD*****, **Vucic EA***, Thu KL, Pikor LA, Lam S, Lam WL. Disruption of KEAP1/CUL3/RBX1 E3-ubiquitin ligase complex components by multiple genetic mechanisms is associated with poor prognosis in head and neck cancer. *Head & Neck.* 2014; Mar 5. doi: 10.1002/hed.23663. [Epub ahead of print] ***co-first authors**

25. Huebbers CU, Olthof NC, KolligsJ, Haesevoets A, Henfling M, Ramaekers FCS, Preuss SF, Drebber U, Lam WL, **Vucic EA**, Kremer B, Speel EJM, Klussmann JP. Human papillomavirus type 16 integration in oropharyngeal squamous cell carcinomas often occurs in genes involved in tumorigenesis. *PLoS One.* 2014; Feb 24;9(2):e88718. doi: 10.1371/journal.pone.0088718.

26. Martinez VD, **Vucic EA**, Thu KL, Pikor LA, Hubaux R, Lam WL. Unique Pattern of Component Gene Disruption in the NRF2 Inhibitor KEAP1/CUL3/RBX1 E3-Ubiquitin Ligase Complex in Serous Ovarian Cancer. *Biomed Research International.* 2014; Article ID 159459

27. **Vucic EA**[†], Chari R, Thu KL, Wilson IM, Cotton AM, Kennett JY, Zhang M, Lonergan KM, Steiling K, Brown CJ, McWilliams A, Ohtani K, Lenburg ME, Sin DD, Spira A, MacAulay CE, Lam S, Lam WL. Aberrant DNA methylation patterns are widespread in small airways of former smokers with COPD. *American Journal of Respiratory Cell and Molecular Biology*. 2014; May;50(5):912-22
[†]**corresponding author**

28. Rowbotham D*, Enfield KSS*, Martinez VD*, Thu KL, **Vucic EA**, Stewart GL, Bennewith KL, Lam WL. Multiple components of the VHL tumor suppressor complex are frequently affected by DNA copy number losses in pheochromocytoma. *International Journal of Endocrinology.* 2014; In press. * co-first authors.

29. Hubaux R, Thu KL**, Vucic EA**, Pikor LA, Kung SHY, Martinez VD, Mosslemi M, Becker-Santos DD, Gazdar AF, Lam S, Lam WL. Microtubule affinity-regulating kinase 2 contributes to cisplatin sensitivity through modulation of the DNA damage response in non-small cell lung cancer. 2014; Under review at *Oncotarget*

30. **Vucic EA***, Thu KL*, Pikor LA, Enfield KSS, Yee, J, English JC, Macaulay CE, Lam S, Lam WL. Smoking Status Impacts microRNA Mediated Prognosis and Lung Adenocarcinoma Biology. Under review at *BMC Cancer* ***co-first authors**