MOLECULAR MECHANISMS UNDERLYING TUMORIGENESIS OF NON-SMALL CELL LUNG CANCER SUBTYPES

by

Larissa A. Pikor

BScH., BPHE, Queen's University at Kingston, 2009

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

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Interdisciplinary Oncology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

July 2014

© Larissa Pikor, 2014

Abstract

Lung cancer is the leading cause of cancer death in Canada and worldwide. A late stage of diagnosis in conjunction with a lack of effective treatment options are largely responsible for the poor survival rates of lung cancer. Histological subtypes of lung cancer are known to respond differently to standard therapies, suggesting they are distinct diseases. A better understanding of the molecular alterations and underlying biology of lung cancer subtypes is therefore necessary for the development of novel detection and treatment strategies in order to improve patient prognosis. We hypothesize that lung adenocarcinoma (AC) and squamous cell carcinoma (SqCC) arise through disparate patterns of molecular alterations and that these differences underlie unique biological mechanisms that contribute to subtype development, phenotypes and response to therapy.

In this thesis, I apply multidimensional integrative 'omics approaches to characterize the genomic and epigenomic landscapes of AC and SqCC and elucidate differential patterns of alterations and subtype specific gene disruptions causal to NSCLC and the development of specific subtypes. The integration of DNA copy number, methylation, gene and miRNA expression data on AC and SqCC tumors and patient matched non-malignant tissue identified several subtype specific alterations and revealed unique oncogenic pathways associated with AC and SqCC that can be successfully targeted by existing therapies. By combining genomic analyses with manipulation of candidate genes *in vitro* and *in vivo*, we validated the contribution of candidate genes to tumorigenesis and determined the mechanisms through which they contribute to disease pathogenesis. In addition to revealing differentially disrupted genes and pathways we also identified numerous alterations common to both subtypes.

Collectively, this work has further characterized the landscape of molecular alterations that define AC and SqCC, and the mechanisms through which these alterations contribute to subtype tumorigenesis. This work has identified novel candidate genes involved in subtype tumorigenesis as well as miRNAs with potential as diagnostic biomarkers for lung cancer. Taken together, these findings underscore the importance of tailoring treatment strategies to the histological subtype based on the underlying biology of that subtype.

Preface

The research in this thesis was conducted with ethics approval from the UBC Research Ethics Board, Certificate Numbers: EDRN H09-00008, CCSRI H09-00934, W81XW-10-1-0634

Portions of Chapter 1 have been published as:

1. [Pikor LA], Ramnarine VR, Lam S, Lam WL (2013). Genetic alterations defining NSCLC subtypes and their therapeutic implications. Lung Cancer 82(2):179-89.
2. Enfield KSS, [Pikor LA], Martinez VD, Lam WL (2012). Mechanistic roles of non-coding RNAs in lung cancer biology and their clinical implications. Genetics Research International 2012:737416

Chapters 2,3 and 4 were co-authored as manuscripts for publication. The following author lists apply for these chapters:

A version of Chapter 2 has been published and presented by Will Lockwood in his thesis. Lockwood WW, Wilson IM, Coe BP, Chari R, [Pikor LA], Thu KL, Yee J, English J, Murray N, Tsao MS, Minna J, Gazdar AF, MacAulay CE, Lam S, Lam WL (2012). Divergent genomic and epigenomic landscapes of lung cancer subtypes underscore the selection of different oncogenic pathways during tumor development. *PLoS One* 2012;7(5). This chapter presents the necessary description of datasets (genomic, epigenomic and expression) used in subsequent chapters to investigate the molecular biology of lung cancer subtypes. As this chapter is quite detailed, it was included as a data chapter rather than integrated into the introduction. I conducted and analyzed dose response experiments, performed SNP6 data analysis and wrote sections of the manuscript pertaining to the experiments I performed.

A version of Chapter 3 has been published. Thu KL, [Pikor LA], Chari R, Wilson IM, MacAulay CE, Gazdar AF, Lam S, Lam WL, Lockwood WW (2011) Disruption of KEAP1/CUL3 E3 ubiquitin ligase complex components is a key mechanism for NFkB pathway activation in lung cancer. *Journal of Thoracic Oncology* 6(9):1521-9. I was co-first author of this manuscript and

participated in the design of the study, performed experiments, data analysis and interpretation of the results, and wrote and edited the manuscript. R Chari and IM Wilson contributed to project design and data analysis. Patient samples were collected by S. Lam and cell lines provided by AF Gazdar. WL Lam and WW Lockwood oversaw the project.

A version of Chapter 4 has been published. [Pikor LA], Lockwood WW, Thu KL, Vucic EA, Chari R, Gazdar AF, Lam S, Lam WL (2013). YEATS4 is a novel oncogene amplified in non-small cell lung cancer that regulates the p53 pathway. *Cancer Research 73(24):7301-12.* I designed the study, conducted the experiments, performed data analysis and interpretation and wrote the manuscript. Patient samples were collected by S. Lam, cell lines were provided by AF Gazdar and mouse work was performed by WW Lockwood. KL Thu, EA Vucic and R Chari aided in data analysis and interpretation of results.

A version of Chapter 5 is being prepared for publication. [Pikor LA], Kennett JY, Thu KL, Vucic EA, Lam S, Lam WL (2014). Characterizing miRNA expression in lung adenocarcinoma and squamous cell carcinoma identifies miR-944 and miR-1287 as subtype specific miRNAs. I conceived the experimental design for this project, performed data analysis and interpretation, *in vitro* experiments and wrote the manuscript. JY Kennett provided flow cytometry technical assistance and KL Thu and EA Vucic aided in data analysis, project design and edited the manuscript.

Table of Contents

Abstract ii

Preface iii

Table of Contents v

List of Tables xi

List of Figures xii

List of Abbreviations xiii

Acknowledgements xv

Dedication xvi

Chapter 1: Introduction 1

1.1 Lung cancer 1

1.2 Lung cancer subtypes/ histopathology of NSCLC 1

1.3 Molecular pathology of lung cancer 3

1.3.1 Gene expression 3

1.3.2 Copy number alterations 4

1.3.3 DNA methylation 4

1.3.4 DNA mutations 5

1.3.5 Non-coding RNAs 5

1.4 Integrative genomics and a systems biology approach to gene discovery 6

1.5 Thesis theme and rationale 7

1.6 Objectives and hypotheses 8

1.7 Specific aims and thesis outline 8

Chapter 2: Divergent genomic and epigenomic landscapes underscore the selection of different oncogenic pathways during subtype development 11

2.1 Introduction 11

2.2 Methods 12

2.2.1 DNA samples 12

2.2.2 Tiling path array comparative genomic hybridization 13

2.2.3 DNA methylation analysis 13

2.2.4 Comparison of subtype alteration frequencies 13

2.2.5 Gene expression microarray analysis 14
2.2.6 Statistical analysis of gene expression data 15
2.2.7 Survival analysis 16
2.2.8 Network identification 16
2.2.9 Human lung tissue microarray case selection 16
2.2.10 Immunohistochemical analysis 17
2.2.11 Trichostatin A dose-response analysis 17
2.3 Results 18
2.3.1 Assessment of global genomic instability in AC and SqCC 18
2.3.2 Disparate genetic and epigenetic landscapes characterize lung SqCC and AC 21
2.3.3 Gene disruptions are selected in a subtype specific manner 23
2.3.4 Different oncogenic pathways are associated with AC & SqCC 24
2.3.5 Epigenetically regulated genes complement genetically regulated genes 24
2.3.6 Concerted genetic and epigenetic disruption of subtype- specific genes 29
2.3.7 Subtype specific genes are responsible for AC and SqCC phenotypes 31
2.3.8 Subtype-specific genetic differences are translated to the protein level 33
2.3.9 Subtype-specific genes are associated with distinct clinical characteristics 35
2.3.10 Defining putative treatment strategies tailored to lung cancer subtypes using *in silico* screening of candidate therapeutic compounds 44
2.4 Discussion 47
2.5 Conclusions 51
Chapter 3: Genetic disruption of KEAP1/CUL3 E3 ubiquitin ligase complex components is a key mechanisms of NF-κB pathway activation in lung cancer 52
3.1 Introduction 52
3.2 Methods 55
3.2.1 Non-small cell lung cancer (NSCLC) samples 55
3.2.2 Determination of gene dosage and expression levels 55
3.2.3 Copy number analysis of external cohorts 56
3.2.4 Multi-dimensional analysis of complex component disruption in the TCGA 56

3.2.5 Western blot assessment of total and phospho- IKBKB and total and phospho- NF-κB protein levels 57
3.2.6 Immunohistochemistry (IHC) staining for IKBKB protein levels 57
3.2.7 siRNA-mediated complex component knockdowns 58
3.2.8 NF-κB target gene analysis 58
3.2.9 IKBKB inhibition in NSCLC cell lines 59
3.3 Results 60
3.3.1 Disruption of E3-ubiquitin ligase complex components in NSCLC 60
3.3.2 Complex components are differentially altered in AC and SqCC subtypes 63
3.3.3 Other genetic and epigenetic mechanisms of complex component disruption 64
3.3.4 Functional consequences of genetic complex disruption 65
3.3.5 Pharmacological inhibition of IKBKB 68
3.4 Discussion 70
3.5 Conclusions 73
Chapter 4: YEATS4 is a novel oncogene amplified in non-small cell lung cancer that regulates the p53 pathway 74
4.1 Introduction 74
4.2 Methods 75
4.2.1 NSCLC tumor samples and cell lines 75
4.2.2 Array comparative genomic hybridization and GISTIC analysis 76
4.2.3 Gene expression profiling and data integration 76
4.2.4 Analysis of external cohorts 77
4.2.5 Plasmids and generation of stably expressing YEATS4 lines 77
4.2.6 Quantitative reverse transcriptase PCR 78
4.2.7 siRNA mediated knockdown of CDKN1A 79
4.2.8 Western blot 79
4.2.9 Cell viability 79
4.2.10 Colony formation 80
4.2.11 Cellular senescence 80
4.2.12 Dose response assay 80

4.2.13 Mouse models 81
4.2.14 Pathway analysis and gene set enrichment 81
4.2.15 Survival analysis 81
4.3 Results 82
4.3.1 Recurrently amplified regions in NSCLC 82
4.3.2 Identification of YEATS4, the target of 12q15 amplification 84
4.3.3 YEATS4 is frequently amplified and overexpressed in NSCLC 84
4.3.4 YEATS4 displays oncogenic properties *in vitro* and *in vivo* 88
4.3.5 YEATS4 suppresses p53 and p21 92
4.3.6 YEATS4 alters the sensitivity of cell lines to cisplatin and nutlin 92
4.3.7 Sensitivity to cisplatin is not mediated solely through the p53-p21 pathway 96
4.3.8 Knockdown phenotypes are independent of p21 signaling in mutant p53 cells 96
4.3.9 Identification of additional cellular networks regulated by YEATS4 98
4.4 Discussion 101
Chapter 5: Characterizing miRNA expression in lung adenocarcinoma and squamous cell carcinoma identifies miR-944 and miR-1287 as subtype specific miRNAs 104
5.1 Introduction 104
5.2 Methods 106
5.2.1 NSCLC patients and samples 106
5.2.2 MicroRNA sequencing and data analysis 106
5.2.3 Identification of subtype specific miRNAs and NSCLC miRNAs 106
5.2.4 Statistical analysis of miRNA expression data 107
5.2.5 Validation in external cohorts 108
5.2.6 Integration of DNA copy number and methylation with miRNA expression 108
5.2.7 Target prediction, interaction networks and pathway analysis 108
5.2.8 Cell culture and manipulation of miR-1287 109
5.2.9 3'-UTR reporter assays 109
5.2.10 qPCR of miRNA expression 109
5.2.11 Western blotting 110
5.2.12 MTT assay 110

5.2.13 Apoptosis, DNA damage and cell proliferation 110
5.3 Results 111
5.3.1 MiRNAs are differentially expressed between NSCLC tumors and non-malignant tissue and accurately segregate profiles based on malignancy. 111
5.3.2 Lung AC and SqCC display different patterns of miRNA expression 114
5.3.3 Validation of differentially expressed miRNAs 117
5.3.4 Copy number alterations contribute to aberrant miRNA expression 119
5.3.5 Predicted targets of subtype miRNAs are associated with distinct pathways 119
5.3.6 Lung cancer prognostic genes are targeted by NSCLC miRNAs 120
5.3.7 MiR-944 overexpression is highly specific to SqCC 124
5.3.8 MiR-1287 inhibits expression of RAD9A by binding its 3'-UTR 126
5.3.9 Overexpression of miR-1287 enables proliferation of damaged cells in LCCL 126
5.4 Discussion 128
Chapter 6: Conclusion 133
6.1 Summary 133
6.1.1 Identify molecular alterations common or specific to lung cancer subtypes. 133
6.1.2 Delineate the biological significance and clinical relevance of alterations. 134
6.2 Conclusions 135
6.3 Significance and clinical implications 136
6.4 Future directions 137
6.4.1 Mechanisms of YEATS4 mediated tumorigenesis 138
6.4.2 Validation of miRNAs as biomarkers for the detection of lung cancer 138
6.4.3 Contribution of miR-944 to SqCC tumorigenesis 139
6.4.4 Global assessment of histone modifications in SqCC 139
6.4.5 Integration of molecular and immunologic alterations to understand the contribution of the tumor microenvironment to tumorigenesis. 140
References 141
Appendix 153
Appendix A Sample sets and demographic data of cohorts used in this thesis 153
A.1 Samples sets used in Chapter 2 153

A.2 Demographic data of BCCRC 261 AC and 92 SqCC with copy number profiles . 153
A.3 Datasets used to determine the prevalence and subtype specificity of YEATS4 amplification and overexpression 154
Appendix B List of publications 155

List of Tables

Table 2.1 Top subtype specific networks associated with genes deregulated by copy number ... 26
Table 2.2. Genes demonstrating concerted genetic and epigenetic disruption in SqCC 30
Table 2.3 Genes demonstrating a significant association with survival 36
Table 3.1 Copy Number status of complex components in multiple independent cohorts 62
Table 3.2 Frequency of KEAP1 complex copy number alterations in AC and SqCC 64
Table 4.1 Frequency of YEATS4 gain and amplification across multiple datasets 87
Table 4.2 Cisplatin and Nutlin IC_{50}s 95
Table 4.3 Differentially altered genes following knockdown of YEATS4 used for Ingenuity Pathway Analysis 99
Table 5.1 Patient demographics for all cases with miRNA sequencing data 113
Table 5.2 Alteration frequencies of the most frequently deregulated NSCLC miRNAs..... 113
Table 5.3 AUC results of all BCCRC subtype specific miRNA 116
Table 5.4 Subtype specific and NSCLC miRNAs that validated in the TCGA cohort 118

List of Figures

Figure 2.1. Global genetic and epigenetic instability of AC and SqCC 20
Figure 2.2. Genomic landscapes of lung AC and SqCC..... 22
Figure 2.3. Gene networks involved in the development of SqCC and AC. 25
Figure 2.4. Epigenetically altered SqCC genes are significantly enriched for SCLC signaling. . 28
Figure 2.5. Subtype specific genes explain AC and SqCC phenotypes..... 32
Figure 2.6. Subtype-specific genomic differences are reflected at the protein level. 34
Figure 2.7 CD9 expression is associated with differences in survival in AC and SqCC tumors. 44
Figure 2.8 SqCC cell lines are more sensitive to Trichostatin A than AC cell lines. 46
Figure 3.1 The role of the KEAP1/CUL3 ubiquitin ligase complex 53
Figure 3.2 Frequent disruption of the KEAP1 E3-ligase complex and IKBKB in NSCLC. 61
Figure 3.3. Association between gene expression and copy number in clinical lung tumors. 62
Figure 3.4. IKBKB protein expression in NSCLC. 66
Figure 3.5. Activation of NF-κB targets in tumors with complex components disruption 68
Figure 3.6. Pharmacological inhibition of IKBKB in NSCLC cell lines..... 69
Figure 4.1. Recurrent amplifications in NSCLC. 83
Figure 4.2 *YEATS4* is the target of 12q15 amplification. 86
Figure 4.3. Overexpression of YEATS4 induces a malignant phenotype. 89
Figure 4.4. YEATS4 knockdown impairs growth and induces senescence. 91
Figure 4.5. YEATS4 alters p21 and p53 protein levels. 94
Figure 4.6. Senescence induced by YEATS4 is dependent on p21 in wildtype p53 cells..... 97
Figure 4.7. Gene networks associated with YEATS4 knockdown. 100
Figure 5.1. miRNA expression profiles accurately segregate tumor and non-malignant tissue. 112
Figure 5.2. miRNAs are disrupted in a subtype specific manner that can discriminate between subtypes..... 115
Figure 5.3 miRNA-mRNA interaction network of overexpressed SqCC specific miRNA 121
Figure 5.4. Bi-directional miRNAs target genes involved in cellular fate and organization..... 122
Figure 5.5 Predicted interaction between NSCLC miRNAs and lung cancer prognostic genes 123
Figure 5.6 miR-944 overexpression is specific to SqCC and targets ERK signaling. 125
Figure 5. 7. miR-1287 induces a malignant phenotype through the suppression of RAD9A 127

List of Abbreviations

3' UTR	3 prime untranslated region
AC	adenocarcinoma
aCGH	array comparative genomic hybridization
BAC	bacterial artificial chromosome
BrdU	bromodeoxyuridine
CMAP	Connectivity Map
CNA	copy number alteration
CS	current smoker
BCCA	British Columbia Cancer Agency
BH	Benjamini Hochberg
COSMIC	Catalogue of Somatic Mutations in Cancer
DNA	deoxyribonucleic acid
EGFR	Epidermal Growth Factor Receptor
Freq	frequency
FFPE	formalin fixed paraffin embedded
FS	former smoker
GEO	Gene Expression Omnibus
GISTIC	Genomic Identification of Significant Targets in Cancer
HBEC	human bronchial epithelial cells
IHC	immunohistochemistry
KEAP1	Kelch-like ECH-associated protein 1
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene
LCCL	lung cancer cell lines
MANOVA	multivariate analysis of variance
miRNA/miR	microRNA
ncRNA	non-coding RNA
NF-κB	Nuclear Factor Kappa B
NS	never smoker
NSCLC	non-small cell lung cancer

OE	overexpressed
PCA	proportion of genome altered
PCR	polymerase chain reaction
PI	propidium iodide
RPKM	reads per kilobase of exon model per million mapped reads
RT-qPCR	reverse transcription quantitative PCR
RNA	ribonucleic acid
SCLC	small cell lung cancer
SMRT	sub megabase resolution tiling
SqCC	squamous cell carcinoma
TCGA	The Cancer Genome Atlas
TKI	tyrosine kinase inhibitor
UE	underexpressed
YEATS4	YEATS domain containing 4

Acknowledgements

I would like to acknowledge the contributions of the many members of the Wan Lam Lab, both past and present who contributed to this work, and provided useful insight and discussion, especially the co-authors of each of the manuscript chapters presented herein. In addition, I would also like to acknowledge our collaborators and generous scholarship support from the Interdisciplinary Oncology Program Training Incentive Award, University of British Columbia Graduate Entrance Scholarship, Canadian Institutes of Health Research Frederick Banting and Charles Best Canada Graduate Scholarship Master's Award, University of British Columbia Four Year Doctoral Fellowship, and the Vanier Canada Graduate Scholarship.

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

I owe particular thanks to my supervisor Dr. Wan Lam and my supervisory committee members Dr. Stephen Lam and Dr. Calum MacAulay for their mentorship, guidance and continued support.

Dedication

This thesis is dedicated to my family and friends; in particular my Nan who was always my biggest supporter and liked to call me Dr. P. long before it was official, and my childhood friend Kristen who serves as a reminder of why I decided to pursue a career in cancer research, and to live each day to the fullest, appreciate the small things in life and never forget to smile.

Chapter 1: Introduction

1.1 Lung cancer

Lung cancer is the leading cause of cancer death in the world, with an estimated 25,500 new cases and 20,200 deaths in Canada in 2013, accounting for 27% of all cancer related deaths- more than breast, prostate and colon cancer combined [1, 2]. Surgery is the best curative option and standard of care for patients with localized disease. However, the majority of patients (>80%) present with locally advanced or metastatic disease for which platinum doublet chemotherapy and radiation are the standard treatment. Despite recent advances in diagnosis and treatment, such as the development of targeted therapies, the 5-year survival rate of lung cancer has failed to improve significantly over the last 30 years, and remains a meager 18% [1, 2]. This modest improvement is largely attributed to two factors; a lack of early detection strategies, and the modest survival benefits and inability of current treatment strategies to cure patients with disseminated disease.

Tobacco smoke exposure is the main etiological factor associated with lung cancer, with smokers having a 14 fold increased risk of developing lung cancer compared to never smokers [3]. However only 10-20% of smokers develop lung cancer, and with the success of smoking cessation programs, there is an increasing percentage of lung cancers (25%) that arise in individuals who never smoked [4, 5]. Other factors known to influence lung cancer risk include carcinogens such as arsenic, radon and asbestos, a family history of lung cancer, second hand smoke and viral infection (Human papillomavirus and Epstein Barr Virus) [4, 6].

1.2 Lung cancer subtypes/ histopathology of NSCLC

Histologically, lung cancer is classified into two broad categories; small-cell lung cancer (SCLC), occurring in approximately 15% of patients and the more prevalent NSCLC, which accounts for roughly 85% of cases [7]. NSCLC itself is a heterogeneous disease comprised of phenotypically diverse and regionally distinct neoplasias that differ in their cell of origin, growth pattern and location within the lung, and can be further subdivided into 3 major histological subtypes: adenocarcinoma (AC), squamous cell carcinoma (SqCC) and large cell carcinoma, of

which AC and SqCC are the predominant subtypes, and the focus of this thesis [7]. Within the last few decades, there has been a dramatic shift in the global trends of lung cancer histology, with a steady decline in SCLC and SqCC, making AC the most prevalent subtype of lung cancer. These changes are largely believed to be due to widespread changes in cigarette composition (lower tar and nicotine content) which has led to a change in smoking behavior, with smokers smoking more frequently and inhaling deeper in an attempt to achieve the same effect, causing tobacco carcinogens to be deposited further into the lung periphery.

Adenocarcinoma is the most common histological subtype of lung cancer, accounting for 40-50% of all lung cancer cases and typically arises in the glandular epithelium of the lung periphery from type II pneumocytes or Clara cells [4, 8]. Substantial heterogeneity exists within AC, with >90% of adenocarcinomas consisting of two or more histological subtypes and being classified as mixed subtype [9]. With advances in the field of lung adenocarcinoma, including the development of targeted therapies and the identification of prognostic subsets, there was a need for improvement in histological stratification. Therefore, in 2011 the International Association for the Study of Lung Cancer (IASLC), American Thoracic Society (ATS) and European Respiratory Society (ERS) developed a new classification for AC in which invasive ACs are classified according to the predominant subtype: acinar, lepidic, micropapillary, papillary or solid and have significant differences in prognosis [8].Lepidic tumors are the most indolent subtype. Acinar and papillary predominant tumors have an intermediate clinical behaviour while solid and micropapillary tumors are associated with poor prognosis suggesting early stage patients with these histologies may benefit from adjuvant therapy [8].

Squamous cell carcinoma comprises 20% of all lung cancer cases, is strongly associated with a history of smoking and develops primarily in the central airways and segmental bronchi [7, 9]. It is believed that SqCC develops through a stepwise transformation of the normal epithelium. This process begins with hyperplasia of basal cells, followed by squamous metaplasia, various degrees of dysplasia (mild, moderate and severe) and then carcinoma in situ (CIS) before eventually developing into a malignant carcinoma [10]. Each stage can be characterized by increased morphological and cytological changes. Well differentiated squamous

tumors are characterized by cell keratinization, intracellular bridges and keratin pearl formation, and while less heterogeneous than AC, can also be further sub-classified into basaloid, papillary, clear cell and small cell types [9].

Until recently, NSCLC was treated as a single disease with a "one size fits all" therapeutic approach determined exclusively by disease stage due to the similar therapeutic effects of conventional chemotherapeutic agents. However, evidence from clinical trials has demonstrated that tumor histology influences response rates, toxicity and progression free survival of chemotherapy and targeted drugs such as bevacizumab, pemetrexed and epidermal growth factor receptor tyrosine kinase inhibitors (EGFR-TKI's) [11, 12], such that histology is now recognized as an important factor in treatment selection [13]. The disparate clinical behaviours of AC and SqCC suggest distinct molecular mechanisms underlie these phenotypic differences and highlight the importance of understanding the biology and molecular origins of lung cancer subtypes.

1.3 Molecular pathology of lung cancer

Lung tumors harbor hundreds to thousands of molecular alterations that result in the activation of oncogenes and inactivation of tumor suppressor genes, leading to the deregulation of fundamental cellular processes and promoting malignancy [14-18]. These include but are not limited to gene and miRNA expression changes, copy number alterations, aberrant DNA methylation, and sequence mutations. In the past two decades, there has been extensive effort to identify and characterize the landscape of driver alterations that contribute to lung tumorignesis in an attempt to better understand the underlying biology of lung tumors and identify novel therapeutic targets [15, 18, 19].

1.3.1 Gene expression

Gene expression signatures have shown the ability to define and distinguish histological subtypes, [20-24] morphological subtypes within AC [20, 21] and SqCC [25] as well as distinguish tumor from non-malignant tissue [26-29], yet their clinical utility is limited due to the lack of overlap between subtype signatures. Interestingly, functional overlap between subtype

specific signatures has been observed, suggesting disruption of specific pathways is selected for rather than specific genes. Deregulation of antioxidant proteins, detoxification genes and overexpression of cytokeratins and cytokeratin-regulatory genes (*GSTT1, CEL*, and *PRDX6*) often characterize SqCC tumors [20-24], whereas disruption of surfactant-related and small airway-associated genes (*SFTPA2, SFTPB, MUC1, and NAPSA*) are typically altered in AC [20-24, 30, 31]. These functions are largely associated with the histological properties of the cells of origin from which these subtypes develop, further highlighting the contribution of histology to tumorigenesis. Due to the large number of passenger alterations within tumors, gene expression studies alone are limited in their ability to discriminate genes and pathways implicated in tumorigenesis.

1.3.2 Copy number alterations

DNA copy number alterations (CNAs) are a prominent mechanism of gene disruption in NSCLC [18, 32-39]. Although very few CNAs occur exclusively in a single subtype, many regions are altered at significantly different frequencies between subtypes and therefore deemed regions of subtype specific CNA [33, 34, 36]. For example, a recent analysis of over 2000 tumors identified 13 subtype-specific regions with at least a 25% difference in the frequency of alteration between subtypes [39]. Amidst all copy number studies, the most prominent and consistent difference between subtypes is amplification of 3q in SqCC [32, 33, 35, 38, 40-43]. Given the large size of some amplicons/regions of deletion, integration of gene expression data is useful in identifying of candidate oncogenes or tumor suppressors.

1.3.3 DNA methylation

Epigenetic marks such as DNA methylation are important regulators of somatically heritable changes in gene expression. DNA methylation is a tissue-specific and inherently reversible gene regulatory alteration targeted for chemoprevention, treatment and as potential diagnostic and prognostic biomarkers in malignant and non-malignant tissues [44]. DNA methylation profiling of NSCLC has identified hundreds of aberrantly methylated genes [45-49]. However, to date most genome-wide epigenetic studies lack-corresponding gene expression level data, which in the context of determining functional consequences of DNA methylation

alterations to lung cancer biology, is limiting. In SqCC, integration of global DNA methylation and expression profiles indicate methylation of *HOXA2* and *HOXA10* may have prognostic relevance[14]. In AC, aberrantly methylated genes are enriched for cell differentiation, cell cycle regulation, epithelial to mesenchymal transition and RAS and WNT signaling pathways [50].

1.3.4 DNA mutations

Advances in whole genome sequencing technologies have enabled high throughput identification of mutations, copy number aberrations, and structural alterations such as gene fusions and chromosomal rearrangements in a genome-wide, unbiased manner. Sequencing studies have revealed lung cancer to have one of the highest mutation rates of all cancers, as well as immense mutational heterogeneity both within and between patients [17, 51-53]. For example, a single AC tumor was found to have over 50,000 variants, 391 of which affected coding sequences [54]. Frequently mutated genes in lung cancer include *TP53, BRAF, ERBB2, KRAS, EGFR, PIK3CA, PTEN, CDKN2A, NF1, FGFR4, KEAP1* and *RB1* [15, 55]. In addition to mutations, a number of gene fusions have been identified, including *EML4-ALK, KIF5B-RET* and multiple *ROS1* fusions. The discovery of improved responses and outcomes with EGFR TKIs in lung cancer patients harboring *EGFR* mutations, and the profound clinical benefit of targeted therapies in other cancers launched the search for additional actionable alterations in lung cancer and marked the beginning of a new era in NSCLC in which NSCLC are further defined by their driver alterations.

1.3.5 Non-coding RNAs

MicroRNAs (miRNAs) are small 18-25nt long non-coding RNAs (ncRNAs) that negatively regulate gene expression post-transcriptionally through transcript degradation or translational repression [56]. A single miRNA is capable of regulating hundreds of protein coding genes, and a gene can similarly be targeted by numerous miRNAs. miRNAs regulate numerous biological process, including but not limited to proliferation, apoptosis, metabolism, epithelial to mesenchymal transition, differentiation and cellular development, and are frequently deregulated in tumorigenesis [56, 57]. The pathogenesis of lung cancer has been associated with the deregulation of several miRNAs, including loss of the well known tumor suppressive miRNA

let-7, which targets *KRAS* and is associated with poor prognosis in lung cancer. Similar to gene expression, miRNA signatures can accurately separate histological subtypes and are thought to be as good or even superior to global mRNA expression profiles in their ability to accurately classify NSCLC subtypes [57]. Array based miRNA profiling studies have shown *miR-205* to be a highly specific marker for SqCC, while AC specific miRNAs have been shown to associate with mutation patterns. For example, *miR-155* is upregulated exclusively in AC with wildtype *EGFR* and *KRAS*, while *miR-21* and *miR-25* are upregulated in *EGFR* mutant AC and *miR-495* is up-regulated in *KRAS* positive AC [58-60]. Based on these and other findings, miRNAs may be just as important to tumour biology and therapeutics as protein coding transcripts.

1.4 Integrative genomics and a systems biology approach to gene discovery

As a gene can be disrupted by a variety of mechanisms, to accurately determine gene disruption status, multiple mechanisms of disruption need to be interrogated. While the application of single dimensional analyses (expression, copy number, or mutation studies alone) are informative for identifying disrupted genes, they often overlook genes disrupted at low frequencies and are not capable of distinguishing causal from passenger events [61]. The integration of multiple dimensions of 'omics data provides a more comprehensive understanding of the genetic mechanisms affecting a tumor as it not only enables the identification of genes with concurrent DNA and expression alterations which are more likely to be driver alterations, but also genes disrupted by multiple mechanisms (indicative of selection) but at low frequencies by any single mechanism [61]. This is particularly relevant for the identification of frequently altered pathways, as signaling pathways are often disrupted at multiple and distinct points in different tumors as opposed to the frequent alteration of a single gene within a given pathway. Large scale, high throughput genomic profiling studies such as that of the National Cancer Institute's The Cancer Genome Atlas (TCGA) project reinforces the concept of assessing multiple 'omics dimensions to characterize tumors and further our understanding of the alterations that underlie their development.

Given the survival benefit associated with targeted therapies, the ability to accurately identify driver/actionable alterations within a tumor is critical. However, due to the substantial heterogeneity, high mutational load of lung cancer and tissue specific nature of molecular alterations such as DNA methylation, this can only be truly accomplished through the individual analysis of patient genomes with reference to that patient's matched non-malignant tissue. An individualized approach reduces the probability of overlooking driver genes that are disrupted in only a small subset of tumors, and would otherwise be missed using conventional approaches in which tumor and normal samples are grouped. Moreover, consideration of multiple dimensions enables the identification of genes with concordant DNA and expression alterations. A multi-dimensional integrative approach on a tumor by tumor basis is therefore ideal for identifying the molecular mechanisms and signaling pathways that contribute to tumorigenesis.

1.5 Thesis theme and rationale

The theme of this thesis is to characterize the genetic and epigenetic alterations and signaling pathways that define and distinguish AC and SqCC in order to improve our understanding of the genetic basis of the major subtypes of NSCLC. As histological subtypes display distinct clinical features and differential responses to treatment, knowledge of the molecular mechanisms underlying these differences will lead to the development of novel subtype specific therapeutic and detection strategies, and potentially alter clinical management and patient prognosis. Previous attempts to discern the genetic differences between AC and SqCC have been limited in their ability to identify unique causal subtype specific disruptions due to small sample sizes, a lack of patient matched non-malignant tissue and single dimension analyses in which tumors are grouped for analysis. As a result, our understanding of the molecular mechanisms that contribute to tumorigenesis and how these differ between subtypes is limited. A genome wide, integrative multi-'omics approach in combination with functional analysis will provide 1) a comprehensive analysis of the recurrent alterations characteristic of AC and SqCC and 2) insight into how these alterations deregulate specific signaling pathways and biological processes and contribute to subtype pathogenesis.

1.6 Objectives and hypotheses

The objective of this work is to characterize the genetic alterations, genes and pathways that contribute to tumorigenesis and the differential development of histological subtypes using a multi-dimensional integrative analysis in combination with functional validation.

This is based on the following hypotheses:

1. AC and SqCC arise through distinct molecular mechanisms, which can be identified through integrative analysis of AC and SqCC genomes.

2. These different patterns of genetic and epigenetic alterations underlie unique biological mechanisms that contribute to subtype development, phenotypes and response to therapy.

1.7 Specific aims and thesis outline

To address the questions of 1) what are the genetic, molecular and biological similarities and differences between AC and SqCC, 2) how these differences contribute to tumorigenesis and subtype biology and 3) whether any of the identified subtype specific alterations have potential therapeutic, prognostic or diagnostic implications, we devised the following specific aims.

Aim 1: Identify DNA and miRNA alterations that are shared or specific to lung cancer subtypes.

Chapters 2 and 5 describe genome wide comparisons of DNA alterations (copy number and methylation) with concordant expression changes and miRNA expression differences between subtypes, respectively. At the time Chapter 2 was submitted for publication, a large scale multi-dimensional integrative analysis and comparison of subtypes on high resolution platforms had yet to be performed. Given the recent observations of subtype specific indications for *EGFR* TKIs, pemetrexed and bevacuzimab, a comprehensive analysis of the genetic and epigentic alterations and subsequent genes and pathways differentially disrupted between AC and SqCC was warranted. Our integrative analysis identified numerous subtype specific alterations in concordance with hypothesis 1. To date, miRNA profiling studies of lung cancer

subtypes have been limited by the use of microarray platforms which depend on an *a priori* knowledge of miRNAs and as a result are limited to a few hundred miRNAs, and a lack of profiles for patient matched non-malignant tissue. At the time of publication of this thesis, an unbiased, genome wide comparison of miRNA expression between subtypes had yet to be performed. In Chapter 5, we performed miRNA sequencing on 22 SqCC and 66 AC tumor and non-malignant tissue pairs and identified miRNAs differentially altered and expressed between subtypes, as well as a list of miRNAs deregulated in greater than 90% of all cases irrespective of histology. In all chapters of this thesis, genomic findings were validated in multiple independent datasets in order to ensure findings were reproducible and not merely an artifact of our dataset or analysis.

Aim 2: Elucidate genes and pathways involved in subtype tumorigenesis and delineate the biological significance of their disruption.

Upon the identification of subtype specific alterations, I next sought to understand how these alterations contribute to tumorigenesis. This was achieved through *in vitro* and *in vivo* manipulation of candidate genes followed by phenotype assessment and interrogation of the pathways and biological processes affected by gene disruption. Chapter 3 is an in depth analysis of one of the subtype specific gene findings from Chapter 2. In this study we assessed the frequency and functional consequences of disruption of our candidate gene (*KEAP1*) as well as the other components of the E3 ubiquitin ligase complex of which it is a part of. The results from this analysis revealed that disruption of any one complex component was sufficient to impair complex function and that differential patterns of complex components disruption characterize AC and SqCC.

In chapter 4, I focused on the most significant region of amplification in our AC cohort and identified a novel candidate oncogene (*YEATS4*) that is frequently amplified and overexpressed in both AC and SqCC. Manipulation of this gene in lung cancer cell lines and xenograft models impaired tumor growth, while overexpression in bronchial epithelial cells was sufficient to induce malignant transformation. Downstream analysis revealed a role for this gene

in the regulation of the p21-p53 pathway. Although initially identified in AC, our candidate gene was found to be amplified and overexpressed at similar frequencies in both subtypes in multiple independent cohorts, highlighting one of the many shared alterations between subtypes.

Our genome wide analysis of miRNA expression in Chapter 5 identified several subtype specific miRNAs. miRNAs can target hundreds of mRNAs and stringent target prediction of the only AC specific miRNA to validate in the TCGA identified two intriguing predicted targets. The final aim of this chapter was to determine whether these predicted targets are true targets of our AC specific miRNA and assess the effect of miRNA overexpression on tumorigenesis. The findings in this chapter demonstrate the ability of our analysis approach to identify biologically relevant miRNAs involved in tumor biology and subtype development, and provide evidence in support of hypothesis 2.

Aim 3: Determine the clinical relevance of gene/pathway disruption

The final aim of this thesis was to assess the potential clinical relevance of alterations identified in Aim 1 and 2. In Chapter 2 we confirmed *in silico* findings that suggested SqCC tumors would be sensitive to HDAC inhibitors. Similarly, dose response assays performed in Chapters 3 and 4 demonstrate that gene disruption significantly sensitized cell lines to a IKBKB inhibitor and cisplatin, respectively. The identification of miRNAs ubiquitously altered in NSCLC tumors and the stability of miRNAs in blood highlights the potential of these miRNAs as blood based biomarkers for the detection of lung cancer. Taken together, these findings demonstrate the clinical relevance and potential application of characterizing shared and subtype specific alterations.

Chapter 2: Divergent genomic and epigenomic landscapes underscore the selection of different oncogenic pathways during subtype development

2.1 Introduction

Evidence from recent clinical trials has demonstrated that histological subtypes of NSCLC respond differently to both targeted drugs and newly developed chemotherapies, possibly related to differences in cell derivation and pathogenetic origins[12, 62, 63]. One of the most striking examples is the folate antimetabolite pemetrexed, which exhibits superior efficacy and is restricted for use in patients with non-SqCC, presumably due to the higher expression of thymidylate synthase in SqCC tumors [12, 64, 65]. Likewise, numerous studies have associated a higher response rate upon treatment of AC with the *EGFR* tyrosine kinase inhibitors Gefitinib and Erlotinib, reflecting the higher prevalence of *EGFR* mutations in this subtype [13, 62]. These discrepancies in tumor biology and clinical response emphasize the need to determine the underlying genetic, epigenetic and metabolic similarities and differences between the NSCLC subtypes in order to define more appropriate avenues for therapeutic intervention.

Initial gene expression profiling studies were able to segregate AC and SqCC tumors into their respective histologic groupings based on multi-gene models; however, critical events in tumorigenesis may be masked by reactive changes when examining expression profiles alone [20, 66]. Conversely, DNA copy number or DNA methylation changes corresponding with gene expression changes are often regarded as evidence of causality. These alterations are critical events driving progression and other cancer phenotypes [67-69]. Since SqCC and AC develop from distinct cell lineages in different regions of the lung, the range of genetic alterations required for tumor initiation may occur in a lineage-restricted manner. For example, the amplification of the lineage survival oncogenes *SOX2* and *TITF1/NKX2-1* have recently been identified as key events specific to the development of lung SqCC and AC, respectively [18, 70]. However, these genes alone are insufficient to explain the phenotypic diversity of the subtypes, suggesting that the vast majority of genes responsible for their differential development remain unknown. Although subtype specific differences have been observed across all 'omics levels, low genome coverage and/or small sample sizes have been limiting [33-35, 43, 49]. Moreover, gene

discovery on its own provides little information regarding tumor biology, and studies interrogating how differentially disrupted genes interact to perturb signaling pathways within subtypes are rare.

In this study, we performed a large-scale analysis of primary NSCLC tumors (261 total;169 AC and 92 SqCC), integrating DNA copy number, methylation and gene expression profiles to identify critical subtype-specific molecular features. The characterization of the genomic and epigenomic landscapes of AC and SqCC revealed an astounding number of differences at the DNA level with subsequent gene expression changes that are selected for during subtype-specific lung tumor development. Importantly, we identified key oncogenic pathways disrupted by these alterations that likely serve as the basis for differential behaviors in tumor biology and clinical outcomes. Lastly, through prognostic analysis and *in silico* screening of candidate therapeutic compounds using subtype-specific pathway components, we show how these findings may influence our approach to the clinical management of NSCLC.

2.2 Methods

2.2.1 DNA samples

Formalin-fixed, paraffin embedded and fresh-frozen tissues were collected from St. Paul's Hospital, Vancouver General Hospital and Princess Margaret Hospital following approval by the Research Ethics Boards. Hematoxylin and eosin stained sections for each sample were graded by a lung pathologist for use in selecting regions for microdissection. DNA was isolated using standard procedure with proteinase K digestion followed by phenol-chloroform extraction as previously described [71]. All samples were collected under informed, written patient consent and anonymized as approved by the University of British Columbia-British Columbia Cancer Agency Research Ethics Board (REB number H04-60060). Patient information is located in Appendix A2.

2.2.2 Tiling path array comparative genomic hybridization

Array hybridization was performed as previously described [72-74]. Briefly, equal amounts (200-400 ng) of sample and single male reference genomic DNA were differentially labelled with cyanine-5 dCTP and cyanine-3 dCTP, respectively and cohybridized to the SMRT array v.2, which includes over 32,000 BACs spotted in triplicate and has a resolution of approximately 80kb (BCCRC Array Laboratory, Vancouver, BC). Hybridized arrays were imaged using a charge-coupled device (CCD) camera system and analyzed using *SoftWoRx Tracker Spot Analysis* software (Applied Precision). Systematic biases were removed from all array data files using a stepwise normalization procedure as previously described [75, 76]. *SeeGH* software was used to combine replicates and visualize all data as $\log_2$ ratio plots [77]. Stringently, all replicate spots with a standard deviation above 0.075 or signal to noise ratios below three were removed from further analysis. The probes were then positioned based on the human March 2006 (hg18) genome assembly, with removal of the X and Y chromosomes. Genomic imbalances (gains and losses) within each sample were identified using *aCGH-Smooth* [78] with lambda and breakpoint per chromosome settings of 6.75 and 100, respectively.

2.2.3 DNA methylation analysis

For 30 AC samples, 30 patient-matched non-malignant lung samples, 13 SqCC samples and 18 non-patient matched bronchial epithelia samples, all of which were fresh frozen (Sample Set # 2, Appendix A1), DNA methylation profiling was performed using the Illumina Human Methylation27 chip. Five hundred nanograms of DNA from each sample was analyzed by this technology. Normalized β-values were obtained and only those with a detection *p*-value of ≤ 0.05 were used. When comparing tumor samples (AC/SqCC) and normal non-malignant samples (AC non-malignant parenchyma and bronchial epithelia), probes were deemed aberrantly methylated if the absolute difference between tumor and the average of the appropriate normal samples was ≥ 0.15.

2.2.4 Comparison of subtype alteration frequencies

Regions of differential copy number alteration between AC and SqCC genomes were identified as follows. Each array element was scored as 1 (gain/amplification), 0 (neutral/retention), or -1

(loss/deleted) for each individual sample. Values for elements filtered based on quality control criteria were inferred by using neighbouring probes within 10 Mb. Probes were then aggregated into genomic regions if the similarity in copy number status between adjacent probes was at least 90% across all samples from the same subtype. The occurrence of copy number gain/amplification, loss/deletion, and retention at each locus was then compared between AC and SqCC data sets using the Fisher's exact test. Testing was performed using the *R* statistical computing environment on a 3 x 2 contingency table as previously described, generating a p-value for each probe [72]. Benjamini-Hochberg multiple hypothesis testing correction based on the number of distinct regions was applied and resulting p-values ≤0.01 were considered significant. Adjacent regions within 1 Mb which matched both the direction of copy number difference and statistical significance were then merged. Finally, regions had to be altered in >20% of samples in a group and the difference between groups >10% to be considered subtype specific.

A similar approach was used for determining subtype-specific DNA methylation alterations. Frequencies of hypermethylation and hypomethylation for each probe were compared using a Fisher's exact test, followed by a Benjamini-Hochberg multiple testing correction. A corrected *p*-value cut-off of $p<0.05$ was used to deem a probe differentially methylated between the two groups. Due to the smaller number of probes that passed multiple testing correction, no frequency criteria was applied.

2.2.5 Gene expression microarray analysis

Fresh-frozen lung tumors were obtained from Vancouver General Hospital as described above. Microdissection of tumor cells was performed and total RNA was isolated using RNeasy Mini Kits (Qiagen Inc). Samples were labeled and hybridized to a custom Affymetrix microarray according to the manufacture's protocols (Affymetrix Inc, Sample Set # 3,Appendix A1). In addition, RNA was obtained from exfoliated bronchial cells of lung cancer free individuals obtained during fluorescence bronchoscopy (Sample Set # 5, Appendix A1) [79]. All individuals were either current or former smokers. Expression profiles were generated for all cases using the Affymetrix U133 Plus 2 platform (Affymetrix Inc). All data was normalized using the Robust

Multiarray Average (RMA) algorithm in R [80]. In addition, publically available datasets downloaded from the *Gene Expression Omnibus* were used: Affymetrix U133 Plus 2 expression data was downloaded for accession numbers GSE3141 (Sample Set # 4, Appendix A1) [81] and GSE8894 (Sample Set #6, Appendix A1) [82].

2.2.6 Statistical analysis of gene expression data

Gene expression probes were mapped to March 2006 (hg18) genomic coordinates and those within the regions of copy number difference between the subtypes were determined. Comparisons between expression levels for AC and SqCC tumors were performed using the Mann-Whitney U test. As the direction of gene expression difference was predicted to match the direction of copy number difference, one tailed *p*-values were calculated. A Benjamini-Hochberg multiple hypothesis testing correction was applied based on the total number of gene expression probes analyzed for each region. Probes with a corrected *p*-value ≤ 0.001 were considered significant. If multiple probes mapped to the same gene, the one with the lowest p-value was used. Resulting genes were then mapped to the corresponding probes on the Affymetrix U133 Plus 2 array in order to compare their expression in a second set of NSCLC tumors (GSE3141, Sample Set # 4) against normal bronchial epithelial cells. If multiple probes were present for a gene, the one with the strongest *p*-value was used. All comparisons were performed using a one-tailed t-test with unequal variances in *Excel* and genes with a $p<0.001$ were considered significant. The fold-change for tumors versus normal tissues was then determined in order to determine genes expressed in the direction predicted by copy number.

Principal component analysis was performed using expression data for the three independent tumor data sets (described above, Sample Sets #3, 4, 6) in MATLAB. All genes of interest with probes on the corresponding arrays were used. Briefly, the first and second principal components were generated from the original dataset. In the subsequent validation in secondary datasets, these principal components are then used to weight the expression data for a gene based on the original distribution. The Receiver Operating Characteristic (ROC) area under the curve (AUC) analysis was performed to determine the ability of principle component 1 to separate the AC and

SqCC samples into their appropriate histological groups. Calculations were performed using *GraphPad Prism* software.

Connectivity Map (http://www.broad.mit.edu/cmap/) analysis was performed using the up and downregulated genes specific to each subtype as previously described [83].

2.2.7 Survival analysis

Survival analysis was performed using the statistical toolbox in *Matlab*. Expression data for each gene were sorted and survival times compared between the top 1/3 and bottom 1/3 in expression using a publicly available gene expression microarray dataset with survival data (Sample Set #6, Appendix A1). Two tailed *p*-values were generated using a Mantel-Cox log test and those < 0.05 were considered significant. Kaplan-Meier plots were then generated for each gene of interest.

2.2.8 Network identification

Functional identification of gene networks and canonical signalling pathways was performed using Ingenuity Pathway Analysis program (Ingenuity Systems). AC and SqCC specific gene lists were imported as individual experiments using the Core Analysis tool. The analysis was performed using Ingenuity Knowledge Database with the Affymetrix U133 Plus 2 platform as the reference set and was limited to direct and indirect relationships.

2.2.9 Human lung tissue microarray case selection

To determine the expression of ERCC1, KEAP1 and SOX2 in primary NSCLC, we selected 330 NSCLCs (AC, n=220; SqCC, n=110) from surgically resected lung cancer specimens from the Lung Cancer Specialized Program of Research Excellence Tissue Bank at The University of Texas M.D. Anderson Cancer Center. We used archived, formalin-fixed, paraffin-embedded (FFPE) tumor tissue samples placed in tissue micro-array (TMA). The tumor tissue samples were collected between 1997 and 2003, and were histologically analyzed and classified using the 2004 WHO classification system [7]. The characteristics of these TMAs have been previously described in detail [84, 85].

2.2.10 Immunohistochemical analysis

The immunohistochemical analysis was done using commercially available antibodies against KEAP1 (dilution1:25; Proteintech), ERCC1 (dilution 1:25; Labvision) and SOX2 (dilution 1:50; R&D system). Immunohistochemical staining was done using an automated stainer (Dako, Inc.) with 5-µm-thick TMA sections from FFPE tissues. Tissue sections were deparaffinized and hydrated. Antigen retrieval was done in pH 6.0 citrate buffer in a decloaking chamber (121°C × 30 seconds, 90°C × 10 seconds) and washed on Tris buffer. Peroxide blocking was done at ambient temperature with 3% H2O2 in methanol. The slides were incubated with primary antibody (KEAP1 and ERCC1 for 60 minutes; SOX2 for 90 minutes) at ambient temperature and washed with Tris buffer, followed by incubation with biotin-labeled secondary antibody for 30 minutes (EnVision Dual Link System-HRP-Dako for KEAP1 and ERCC1; LSAB system-Dako for SOX2). The immunostaining was developed with 0.5% 3,3′- diaminobenzidine, freshly prepared with imidazole-HCl buffer (pH 7.5) containing hydrogen peroxide and an antimicrobial agent (Dako) for 5 minutes, and then the slides were counterstained with hematoxylin, dehydrated, and mounted.

Nuclear ERCC1, cytoplasmic KEAP1, and nuclear SOX2 expressions were quantified using a four-value intensity score (0, 1+, 2+, or 3+) and the percentage (0-100%) of the extent of reactivity. An immunohistochemical expression score was obtained by multiplying the intensity and reactivity extension values (range, 0-300), and these expression scores were used to determine expression levels.

2.2.11 Trichostatin A dose-response analysis

The effect of HDAC inhibitor Trichostatin A, (Cayman Chemicals) on six NSCLC cell lines; three AC (H3255, H1395 and A549) and three SqCC lines (HCC95, HCC15 and H520) was assessed by cell viability assays. Cells were plated in triplicate in 96 well plates at optimal densities for growth (A549 at 2000 cells/well, HCC95, HCC15 and H520 at 3000 cells/well, and H3255 and H1395 at 6000 cells/well). Cells were subjected to a series of 2-fold dilutions of Trichostatin A prepared in cell growth media with DMSO. The experimental concentrations ranged from 100 uM to 109 pM and the final DMSO concentration for treated and untreated

(control) cells was 1%. Blank wells contained equal volumes of growth media with 1% DMSO. Cells were incubated for 72 hours at 37°C and then treated with 10 µl of Alamar Blue cell viability reagent (Invitrogen) according to manufacturer's instructions. The reaction product was quantified by measuring absorbance at 570 nm with reference to 600 nm using an EMax plate reader (Molecular Devices). The response of treated cells was measured as a proportion of the viability of untreated cells, with the mean background subtracted treatment absorbance divided by the mean background subtracted untreated absorbance for each concentration. Dose response curves and IC50 values were generated in Graph Pad v5 using the proportionate response of all 20 drug concentrations. Experiments were repeated in quadruplicate and differences in IC50 values were assessed by a student's t-test with a p-value <0.05 considered significant.

2.3 Results

2.3.1 Assessment of global genomic instability in AC and SqCC

Based on the differing exposure of cells in the central (SqCC) and peripheral (AC) airways to tobacco carcinogens, which has been linked to the induction of DNA mutations and broad chromosomal instability, we first sought to determine whether global genetic or epigenetic instability was more prevalent in one subtype. Whole genome copy number profiles and DNA methylation profiles were generated and compared for 261 NSCLC FFPE tumors; 169 AC and 92 SqCC (Sample Set #1, Appendix A1) and 30AC, 13SqCC, 30 non-malignant lung parenchyma samples (AC reference) and 18 histologically normal exfoliated bronchial epithelial cell samples (SqCC reference), from patients with NSCLC, respectively (Sample Set #2, Appendix A1).

Unlike the genome, which is identical for most normal cells in the body, the epigenome differs between tissue types [86, 87]. Cancer genomes exhibit global hypomethylation to varying degrees depending on the tissue of origin [88]. DNA methylation profiles are also influenced by mutational profiles within different cancer types, as DNA hyper- and hypomethylation alterations are known to be related to tissue and genetic background [89] as well as smoking

behavior [90]. Given the differing mutational spectra of the two NSCLC subtypes and their differing cells of origin, we investigated the overall DNA methylation level of 30 AC and 13 SqCC samples relative to appropriately matched normal cells (exfoliated broncial epithelial cells). Analysis of 27,578 CpG dinucleotides probes within >13000 CpG islands revealed DNA methylation in the bronchial epithelia and SqCC tumors to be slightly lower than in the normal lung or AC tumors (Figure 2.1A). This trend is mirrored and exaggerated in the CpG dinucleotides outside of CpG islands, suggesting that the cells of the central airway are globally hypomethylated relative to the cells of the peripheral airways, whether cancerous or not (Figure 2.1B, Mann Whitney U test, $p<0.0001$).

The relative genomic instability in AC and SqCC, calculated as the average number of altered probes per sample; gained, lost, and neutral probes, and differentially methylated probes (Tumor βvalue- respective normal β value) was assessed for each tumor using a Mann-Whitney U test. No significant differences in differential methylation at CpG and non CpG island probes or copy number between the two subtypes were found (Figure 2.1C-F), consistent with previous findings [91]. This analysis demonstrates that neither subtype has a proclivity for gain, loss, hypo- or hypermethylation of DNA and therefore any observed differences in alteration frequency at a given locus can be attributed to subtype specific selection of genes within altered regions and not to different degrees of random genomic instability associated with tumor development.

Figure 2.1

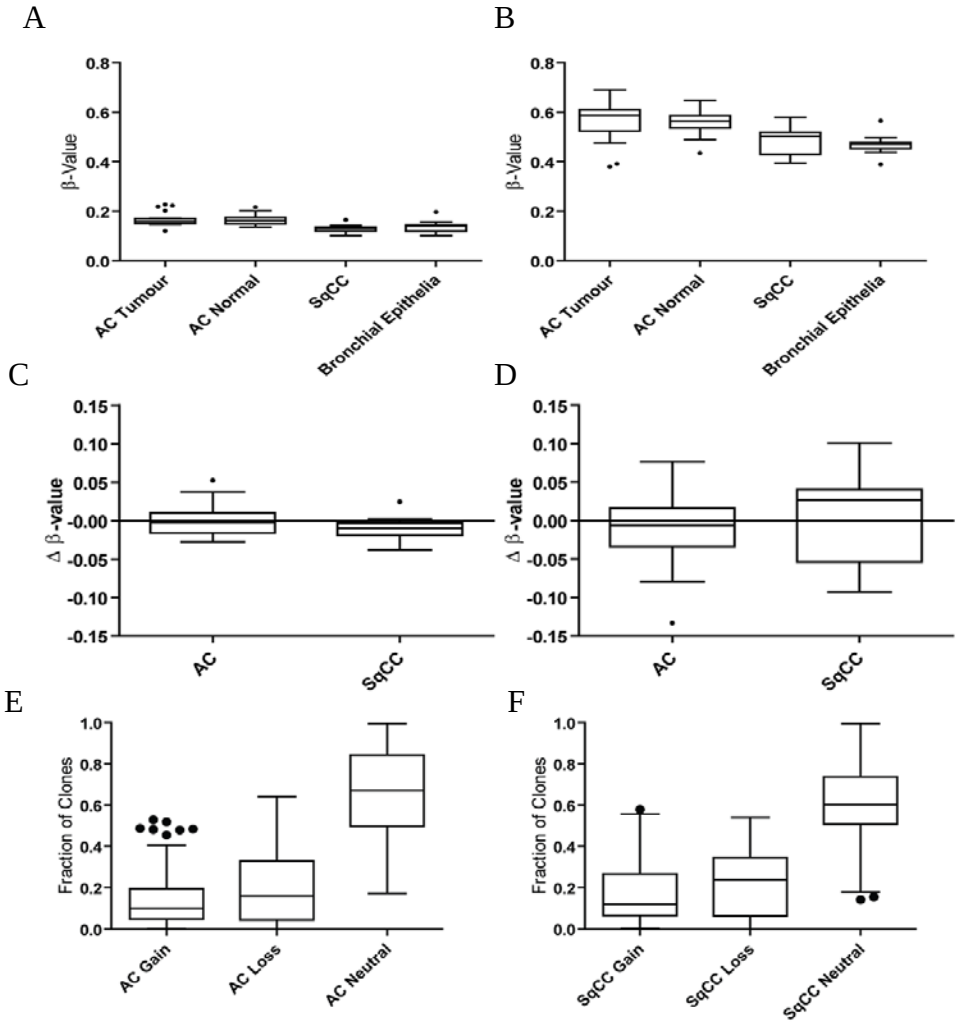


Figure 2.1. Global genetic and epigenetic instability of AC and SqCC

Comparison of DNA methylation averages between AC, histologically normal lung parenchyma, SqCC, and bronchial epithelia for CpG island probes (A) and non-CpG island probes (B). β-value is defined by the methylated signal/total signal for each probe. Average differential methylation levels at CpG islands (C) and non CpG islands (D) for 30AC and 13 SqCC. Percentage of BAC clones of each state calculated for each sample. Box plots illustrate the percentage of clones with gain/amplification, loss/deletion or neutral, in AC (E) and SqCC (F). No significant differences in genetic or epigenetic instability between subtypes was observed, Mann-Whitney U test, $p>0.05$. Box plots depict the 25th, median and 75th percentile, whiskers represent the 5th and 95th percentile and dots show those samples outside theses cutoffs.

2.3.2 Disparate genetic and epigenetic landscapes characterize lung SqCC and AC

Although NSCLC subtypes exhibit similar levels of genomic instability, if specific genetic pathways are involved in their differential development, differences in the genomic alterations selected during tumorigenesis should be present. To determine if genetic alterations unique to each NSCLC subtype exist, we looked for recurrent regions of aberration in each group. For copy number, samples were grouped by subtype and probes were aggregated into regions based on similar copy number status. The frequency of alteration across autosomes was determined and compared between subtypes using the Fisher's exact test with multiple testing correction and a cut-off of <0.01 considered significant. To be considered subtype specific, we required regions to be altered in >20% of samples from a subtype group and have a difference between groups of >10%. Figure 2.2 displays the resulting genomic landscapes of AC and SqCC and highlights the corresponding regions of difference between subtypes. 294 regions of copy number disparity between SqCC and AC were identified, 205 of which were SqCC specific, and 89 that were AC specific (For specific regions, see supplemental digital content, available online http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0037775) . Although some regions overlapped, the character of the alteration (i.e. gain vs. loss) was specific to a subtype. Since the alteration status between the subtypes differed strongly, we classified these as subtype-specific copy number alterations. In total, these alterations covered approximately 550 Mbp of the genome, ranging in size from large segments on chromosome arms (64.8 Mbp on 4q) to discrete peaks only kilobases in size (0.05 Mbp in multiple places).

For DNA methylation, tumor methylation levels were compared to the average of available normal reference tissue profiles. The frequency of probe hypermethylation and hypomethylation (tumor–normal ≥ ±0.15) in AC and SqCC samples was compared using the Fisher's exact test. Following correction for multiple comparisons, 2708 probes corresponding to 2384 genes were found to be differentially methylated ($p<0.05$) (See supplemental digital content for probes). The SqCC group contained markedly more recurrently hyper- and hypomethylated loci than the AC group, similar to the disparity in the numbers of subtype-specific copy number-regulated genes. In fact, only 8% (217) of the 2708 significant probes were more frequently altered in AC, the rest being more commonly hyper- or hypomethylated in SqCC.

Figure 2.2

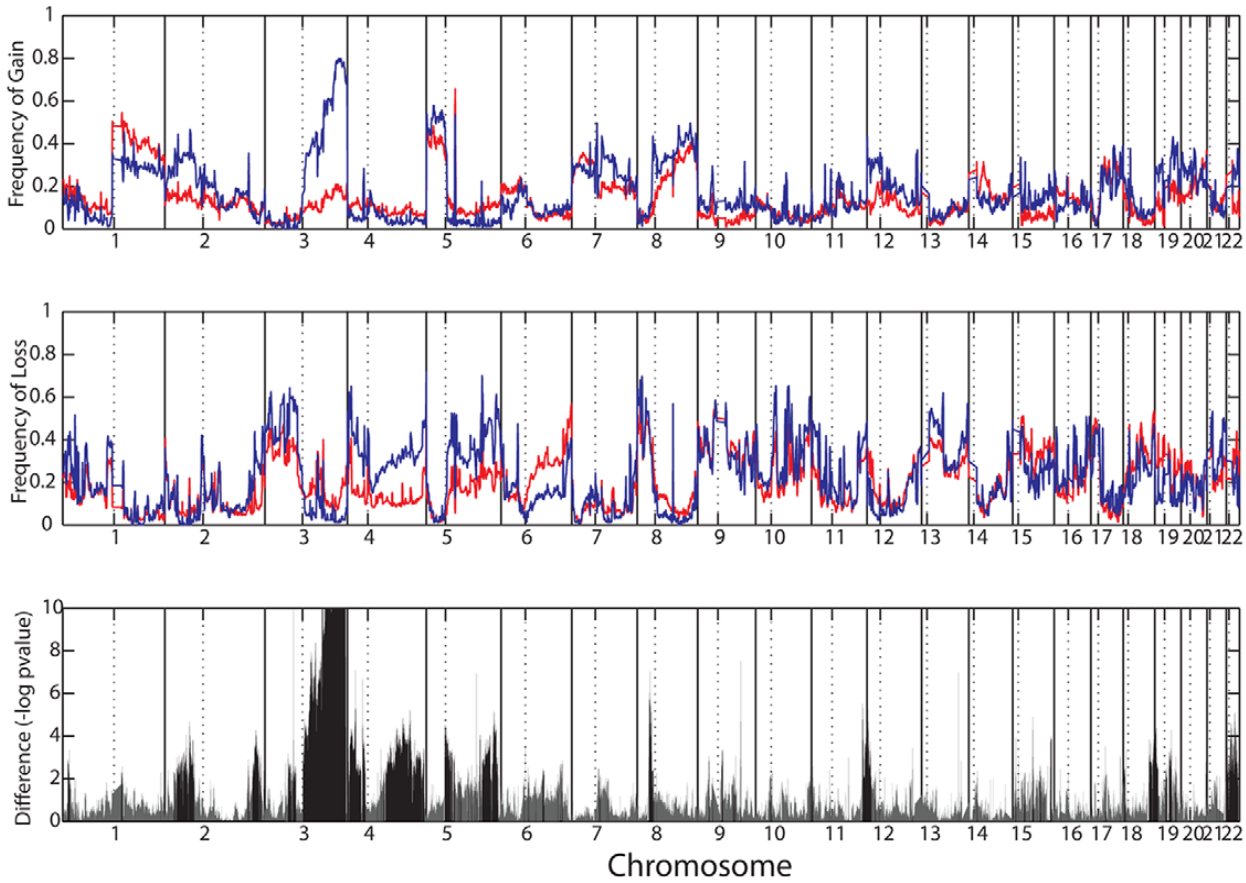


Figure 2.2. Genomic landscapes of lung AC and SqCC.

Alteration frequencies for 169 AC (red) and 92 SqCC (blue) across the human genome. Solid vertical black lines represent chromosome boundaries whereas dotted black lines represent chromosome arm boundaries. The frequency of copy number gain and loss are denoted in the top and middle panel, respectively. The significance of copy number disparity (inverse p-value corrected for multiple comparisons) between subtypes is depicted in the bottom panel. Black lines represent statistically different regions ($p<0.01$) whereas grey lines are not significant.

2.3.3 Gene disruptions are selected in a subtype specific manner

The identification of DNA copy number and methylation disparities between NSCLC subtypes suggests that genes within these regions may be preferentially selected during tumorigenesis and thus responsible for the differential development and pathological characteristics of the subtypes. To identify the potential target genes of these alterations, we integrated DNA copy number and methylation with gene expression levels, with the hypothesis that genes targeted by subtype-specific alterations would 1) be differentially expressed between AC and SqCC, 2) have an expression pattern matching the direction of the alteration and 3) be aberrantly expressed in tumors relative to normal tissue (matching the direction of the alteration), suggesting a role in tumor development. Gene expression profiles were generated for a subset of tumors analyzed by array CGH and DNA methylation (20 SqCC and 29 AC tumors, Sample Set #3, Appendix A1). An independent cohort of 53 SqCC and 58 AC lung tumors and 67 exfoliated bronchial cell samples from cancer-free individuals (Sample Sets #4 and #5, Appendix A1) were used to identify expression differences between tumors and normal cells.

Genes located within each subtype-specific copy number alteration were identified and the expression levels between SqCC and AC and tumor and normal samples compared to determine those that were differentially expressed (Mann Whitney U test, $p<0.001$, after multiple testing correction). 4669 and 2050 unique genes mapped to the SqCC and AC subtype- specific copy number alterations, respectively. After applying all three criteria for defining candidate subtype-specific copy number alteration regulated targets, 447 SqCC-specific and 71 AC-specific genes, corresponding to a total of 492 unique genes were identified as subtype specific. Expression analysis of the 2384 genes with differential methylation revealed 32 AC candidate genes and 297 SqCC genes that were differentially expressed between subtypes and tumor and normal tissue (See supplemental digital content for lists). In both the copy number and methylation analysis, some genes were disrupted in both subtypes, however their patterns of disruption were specific to a subtype, suggestive of opposing roles (oncogenic vs. tumor suppressive) depending on cellular context and therefore considered to be subtype-specific targets. Genes meeting all our criteria may represent critical alterations driving the development of each subtype.

2.3.4 Different oncogenic pathways are associated with AC & SqCC

Cellular pathways and processes specifically disrupted in individual subtypes may reveal key oncogenic mechanisms driving the differential development of AC and SqCC. After identifying the genes differentially altered between subtypes, we next wanted to investigate their biological functions. To discover subtype-related networks of biologically related genes we performed Ingenuity Pathway Analysis (IPA) on the 71 AC and 447 SqCC specific copy number alteration regulated targets. SqCCs exhibited disruptions in gene networks that function in regulating DNA replication, recombination and repair, with additional roles in lymphoid tissue structure and development (Figure 2.3, Table 2.1). Genes involved in the top SqCC network were associated with the binding and modification of histone protein H4, as well as the regulation of the NFKB complex (Figure 2.3b). In contrast, the primary networks in AC displayed functions associated with cell-to-cell signaling, development, and drug metabolism (Table 2.1).

2.3.5 Epigenetically regulated genes complement genetically regulated genes

To determine whether the subtype specific epigenetically-regulated genes carried out functions similar to those subtype-specific genes discovered by the DNA copy number analysis, pathway disruption analysis was performed on the methylation regulated genes. The most significant epigenetically regulated gene network in AC was involved in cell cycle, cell death, and cellular development. This is partly in contrast to the top AC network of copy number regulated genes, which have functions associated with tissue development, but also possess cell signaling and hematological system functions in common. The overall degree of similarity between AC- specific genes that are genetically or epigenetically regulated is quite small, likely due to the low number of AC-specific genes identified (potential reasons for this are discussed below). In contrast, the SqCC gene networks in both analyses are very similar. For example, DNA replication, recombination and repair are highly featured functions of genes identified by both DNA copy number and DNA methylation analyses of SqCC. Genes involved in immunological disease and lymphoid tissue structure and development were also prominent.

Figure 2.3

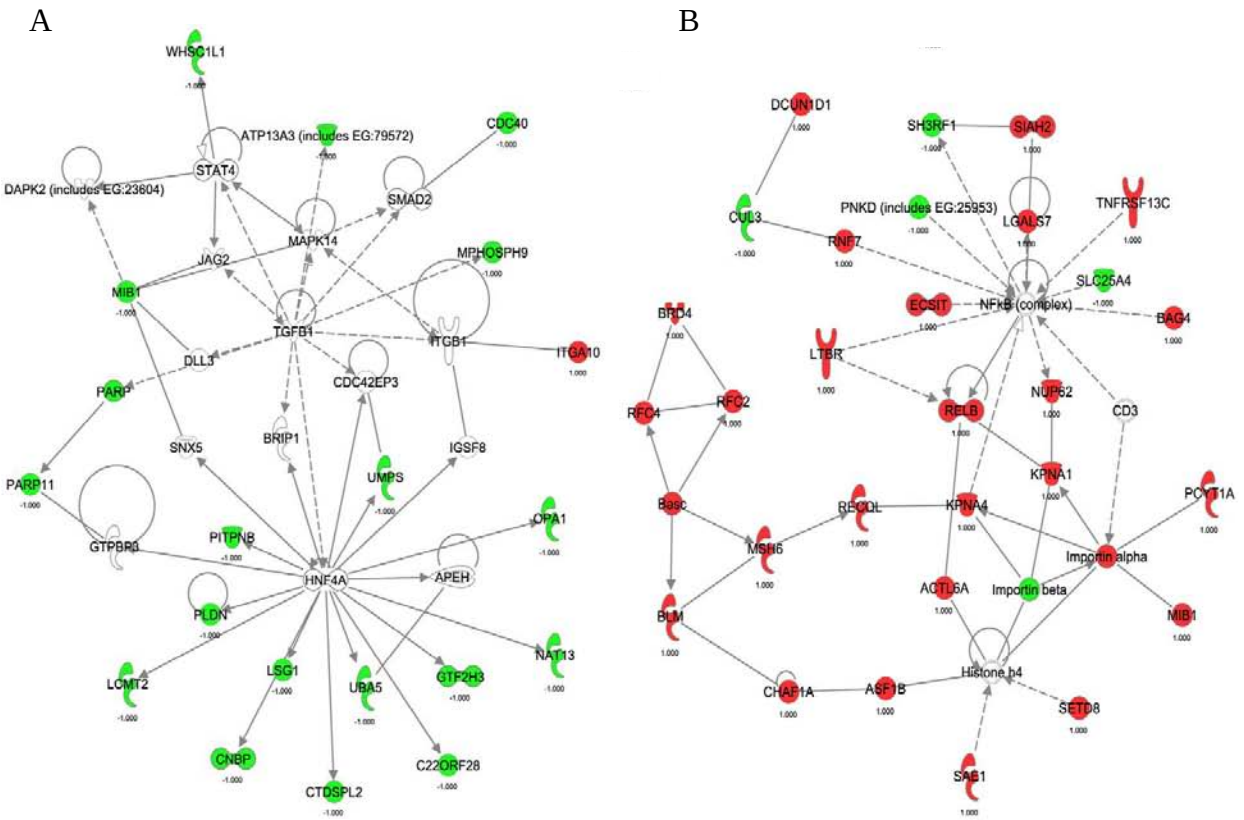


Figure 2.3. Gene networks involved in the development of SqCC and AC.

Ingenuity Pathway Analysis identified biologically related networks from the subtype specific genes deregulated by subtype-specific copy number alterations. The top resultant gene networks for each subtype are displayed; (A) AC network related to HNF4 signaling. (B) SqCC network of potential interactions between multiple histone regulating genes. Solid lines denote direct interactions, dotted lines represent indirect gene interactions. Network components highlighted in red are upregulated whereas those in green are downregulated. Those not highlighted are used to display relationships. Molecules are represented as follows; corkscrews represent enzymes, y-shaped molecules are transmembrane receptors, barbells are regulators of transcription, thimble-shaped molecules are transporters, kinases are triangular, and circular molecules encompass all other gene products.

Table 2.1 Top subtype specific networks associated with genes deregulated by copy number

Network	Molecules in Network	Score	Focus Molecules	Top Functions
AC 1	APEH, ATP13A3, BRIP1, C22ORF28, CDC40, CDC42EP3, CNBP, CTDSPL2, DAPK2, DLL3, GTF2H3, GTPBP3, HNF4A, IGSF8, ITGA10, ITGB1, JAG2, LCMT2, LSG1, MAPK14, MIB1, MPHOSPH9, NAT13, OPA1, PARP, PARP11, PITPNB, PLDN, SMAD2, SNX5, STAT4, TGFB1, UBA5, UMPS, WHSC1L1	40	19	Tissue Development, **Cell-To-Cell Signaling and Interaction**, Hematological System Development and Function
AC 2	ATP2C1, beta-estradiol, BMP4, BRWD1, Ca2+, CD9, CD53, CD160, CENPM, cyclic AMP, EIF3J, EIF4A2, FYTTD1, GJA1, GJA5, hCG, IL15, MAPT, MIR21, MRPL3, MYO1B, NR3C1, NUDT17, PTPRU, S1PR5, SGPP1, SHBG, TCEB1, TIPIN, TM2D2, TTBK1, TTBK2, UBR2, UBXN7, VEGFA	30	16	**Drug Metabolism**, Endocrine System Development and Function, Lipid Metabolism
AC 3	ADAP1, ASH2L, BAG4, Beta ark, C16ORF53, CSDA, DENND4A, DPY30, DUSP3, DUSP9, ELANE, FSH, Histone h3, HIVEP1, LSP1, LY96, MAPK1, MLL3 ,NFkB (complex), NOTCH2NL, PCYT1A, PDC, PDZK1, Pka, PPAP2A, PPP1R2, PPP1R9A, PRLH, PTPN7, RGS5, RIOK3, RLN2, S1PR3, SETD1A, ST8SIA1	17	10	Cardiovascular System Development and Function, Organismal Development, Tissue Morphology
SqCC 1	ACTL6A, ASF1B, BAG4, Basc, BLM, BRD4, CD3, CHAF1A, CUL3, DCUN1D1, ECSIT, Histone h4, Importin alpha, Importin beta, KPNA1, KPNA4, LGALS7, LTBR, MIB1, MSH6, NFkB (complex), NUP62, PCYT1A, PNKD, RECQL, RELB, RFC2, RFC4, RNF7, SAE1, SETD8, SH3RF1, SIAH2, SLC25A4, TNFRSF13C	44	29	**DNA Replication, Recombination, and Repair**, Cell Cycle, **Lymphoid Tissue Structure and Development**
SqCC 2	Alcohol group acceptor phosphotransferase, Alpha tubulin, ATF5, CAMK2N2, CaMKII, CLINT1, COPS7A, CSDA, CSTA, DNAJB11, FBXO45, FOXL2, HDL, HNRNPM, Hsp70, Hsp90, IL11, IL6ST, MAPK1, MAPK9, MAPK12, OTX1, Pak, PAK2, POMC, PPM1F, PPP5C, Sapk, SORBS2, STAR, STAT, TNPO1, TNPO2, TOM1L1, Ubiquitin	34	25	Nervous System Development and Function, Organismal Development, Endocrine System Development and Function
SqCC 3	ACTR2, ARHGAP26, Arp2/3, CLDND1, CYFIP2, DOT1L, EFNA5, EHBP1, EHD3, FCHO2, FXR1, H1FX, Histone H1, Histone h3,histone-lysine N-methyltransferase, KLHL24, LARP1B, LRRC58, MIR124, Nfat, NMD3, P38 MAPK, Pka, PLEKHA2, PPM1L, PTMS, Rac, RANBP1, Ras homolog, SETD7, SH3BP1, SMC4, UHRF1, WHSC1L1, XPO1	29	25	Cell Morphology, Cell-To-Cell Signaling and Interaction, Cellular Development

Of particular interest was the enrichment of aberrantly methylated genes in the small cell lung cancer signaling pathway (genes known to be deregulated in small cell lung cancer) (Figure 2.4a). This was the most significantly enriched canonical pathway affected by DNA methylation alterations and is of interest because both of these lung cancers (SCLC and SqCC) arise in the central airways with similar exposure to cigarette smoke carcinogens. *E2F1* is among the hypomethylated and overexpressed genes represented in this pathway, and is known to be overexpressed in SCLC and to drive expression of *EZH2* which is also overexpressed in SCLC [92, 93]. To explore this pathway further, we investigated whether *EZH2* was more highly expressed in SqCC than AC tumors. As expected, we found that *EZH2* was expressed at a significantly higher level in SqCC tumors than AC tumors (Mann Whitney U test $p<0.0001$), demonstrating the biological consequence of *E2F1* disruption (Figure 2.4B). These findings could reflect the key role of EZH2 and the polycomb group, a protein complex involved in DNA methylation [92] in the pathogenesis of SqCC.

Figure 2.4

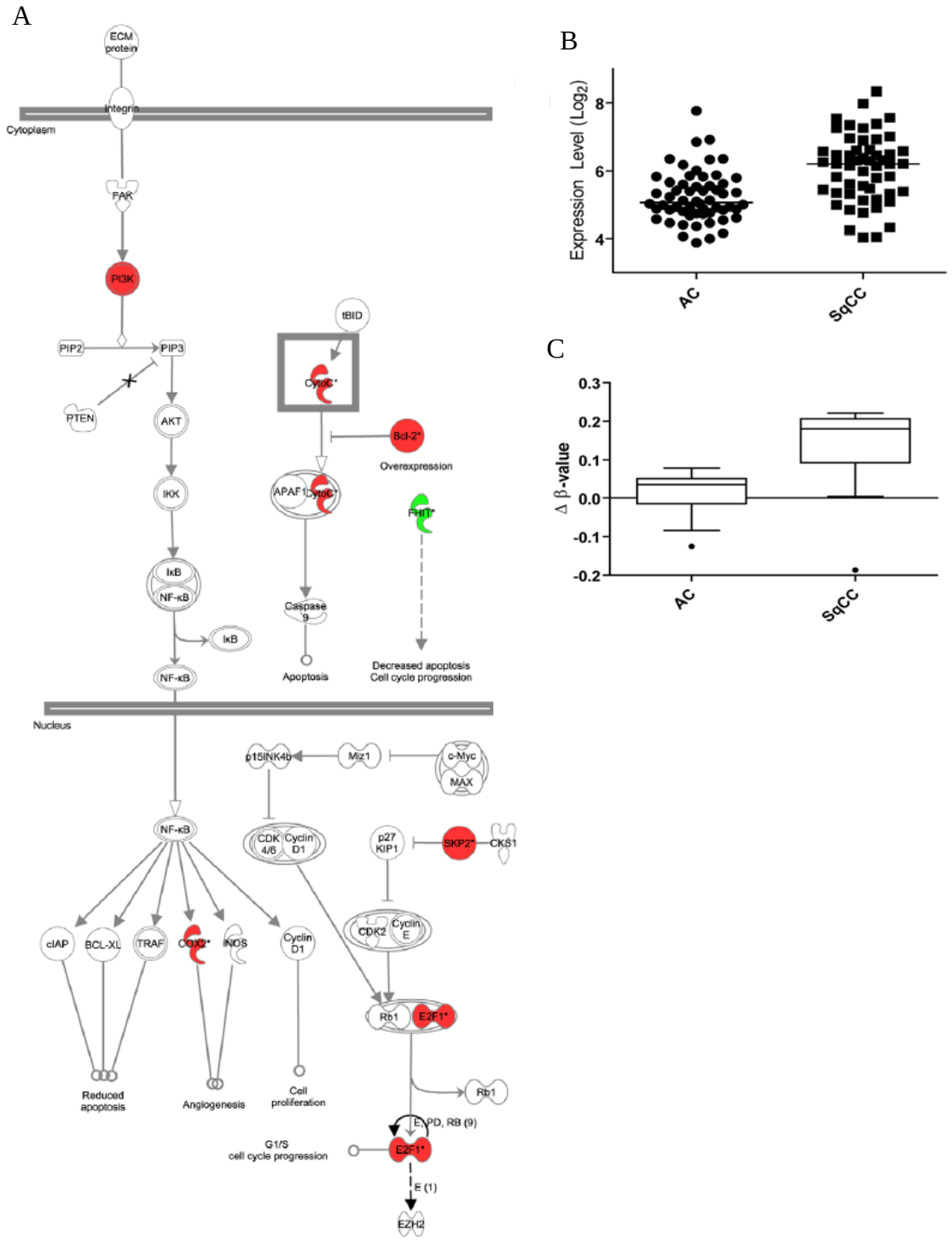


Figure 2.4. Epigenetically altered SqCC genes are significantly enriched for SCLC signaling.

Figure 2.4. Epigenetically altered SqCC genes are enriched for SCLC signaling.
(A) SCLC signaling components are altered at levels of the pathway by DNA methylation in SqCC. Hypomethylated and overexpressed genes are shown in red, hypermethylated and underexpressed genes are shown in green. (B) *EZH2* expression in 58 AC tumors and 53 SqCC tumors. As predicted, *EZH2* expression is higher in SqCC tumors compared to AC tumors, Mann-Whitney U test ($p<0.0001$). (C) *FHIT* is significantly more methylated in SqCC (n=13) and AC (n=30) tumors (Mann-Whitney U test, $p<0.05$).

2.3.6 Concerted genetic and epigenetic disruption of subtype- specific genes

In order to determine if both DNA copy number and DNA methylation aberrations simultaneously disrupted any genes, we combined the subtype-specific gene lists derived using the two analytical approaches described above. None of the 71 genes altered by copy number in AC or the 32 genes associated with DNA methylation overlapped. This was not surprising given the observed lack of similarity at the level of function/network analysis, and could be due in part to the extensive heterogeneity within AC. Combining the 447 copy-number associated genes with the 297 DNA methylation genes from SqCC yielded an overlap of 38 genes (Table 2.2). These genes exhibit frequent concurrent genetic, epigenetic, and subsequent gene-expression alterations that discriminate them from AC tumors. Notably, loss of the well-known 3p tumor suppressor gene (TSG) *FHIT* which is associated with smoking and has been investigated as a potential biomarker for centrally-occurring lung cancers (Figure 2.4c), and amplification of the recently identified lineage specific oncogene *BRF2* were among these genes, highlighting their importance in the development of SqCC. In addition to genes previously associated with lung cancer, TSGs and oncogenes known to be deregulated in other cancer types such as *PRDM2* and *SIAH2* were also altered at the genetic, epigenetic and gene expression levels. Interestingly, seven genes (*ATP2C1, PCYT1A, ZWILCH, CENTB2, BAG4, PARP11* and *CSDA*) were disrupted by gene-dosage in one subtype and DNA methylation in the other. Such multidimensional disruption is indicative of strong selective pressures, suggesting these genes may play a pivotal role in tumor biology.

Table 2.2. Genes demonstrating concerted genetic and epigenetic disruption in SqCC

Gene	SqCC Copy Number Status	SqCC Methylation Status
CFI	Loss and Underexpressed	Hypermethylated and Underexpressed
CYFIP2	Loss and Underexpressed	Hypermethylated and Underexpressed
FHIT	Loss and Underexpressed	Hypermethylated and Underexpressed
GLRB	Loss and Underexpressed	Hypermethylated and Underexpressed
LRP2BP	Loss and Underexpressed	Hypermethylated and Underexpressed
LTC4S	Loss and Underexpressed	Hypermethylated and Underexpressed
OTUD4	Loss and Underexpressed	Hypermethylated and Underexpressed
PRDM2	Loss and Underexpressed	Hypermethylated and Underexpressed
SGPP2	Loss and Underexpressed	Hypermethylated and Underexpressed
ARMC8	Gained and Overexpressed	Hypomethylated and Overexpressed
BAG4	Gained and Overexpressed	Hypomethylated and Overexpressed
BRF2	Gained and Overexpressed	Hypomethylated and Overexpressed
C3orf26	Gained and Overexpressed	Hypomethylated and Overexpressed
CENTB2	Gained and Overexpressed	Hypomethylated and Overexpressed
CSDA	Gained and Overexpressed	Hypomethylated and Overexpressed
EXOSC5	Gained and Overexpressed	Hypomethylated and Overexpressed
FBXO27	Gained and Overexpressed	Hypomethylated and Overexpressed
KLK13	Gained and Overexpressed	Hypomethylated and Overexpressed
MGC2408	Gained and Overexpressed	Hypomethylated and Overexpressed
MRPL51	Gained and Overexpressed	Hypomethylated and Overexpressed
NCK1	Gained and Overexpressed	Hypomethylated and Overexpressed
PCYT1A	Gained and Overexpressed	Hypomethylated and Overexpressed
PDXP	Gained and Overexpressed	Hypomethylated and Overexpressed
PPAN	Gained and Overexpressed	Hypomethylated and Overexpressed
RAD51AP1	Gained and Overexpressed	Hypomethylated and Overexpressed
RECQL	Gained and Overexpressed	Hypomethylated and Overexpressed
RELB	Gained and Overexpressed	Hypomethylated and Overexpressed
RNF185	Gained and Overexpressed	Hypomethylated and Overexpressed
RPL35A	Gained and Overexpressed	Hypomethylated and Overexpressed
RPS27A	Gained and Overexpressed	Hypomethylated and Overexpressed
RSRC1	Gained and Overexpressed	Hypomethylated and Overexpressed
SAE1	Gained and Overexpressed	Hypomethylated and Overexpressed
SFRS10	Gained and Overexpressed	Hypomethylated and Overexpressed
SHOX2	Gained and Overexpressed	Hypomethylated and Overexpressed
SIAH2	Gained and Overexpressed	Hypomethylated and Overexpressed
SMEK2	Gained and Overexpressed	Hypomethylated and Overexpressed
SNRPB	Gained and Overexpressed	Hypomethylated and Overexpressed
SNRPD2	Gained and Overexpressed	Hypomethylated and Overexpressed

2.3.7 Subtype specific genes are responsible for AC and SqCC phenotypes

Next, we sought to confirm that the genes differentially disrupted at the genomic and epigenomic level are responsible for the different biological characteristics of AC and SqCC. Since they are regulated by subtype specific alterations, we hypothesized that the expression levels of these genes should be able to accurately segregate NSCLC tumors into distinct AC and SqCC groups. As predicted, when using the expression values for the 49 NSCLC tumors from our data set, principle component analysis with the 778 unique genetically and/or epigenetically deregulated genes clearly delineated distinct subtype specific clusters (Figure 2.5a). A receiver operating characteristic (ROC) area under the curve (AUC) value of 0.9690 ($P<0.0001$) confirmed that principle component 1 was a strong discriminator of the subtypes (Figure 2.5a). Given that these genes were uncovered based on differences between the subtypes, this finding was not surprising. Therefore, to validate the role of these genes in subtype development, we applied the same analysis to two independent publically available sample sets. The first consisted of 111 (58 AC and 53 SqCC, Sample Set #3) and the second of 138 clinical lung tumors (62 AC and 76 SqCC, Sample Set #6) [82]. Strikingly, this analysis was also able to separate the AC and SqCC samples with a substantial accuracy (ROC AUC values of 0.9076 and 0.9442, $P<0.0001$, respectively) (Figure 2.5b and c). Validation of our subtype specific genes in these large, independent panels of NSCLC tumors from separate institutions provides further evidence that the genes regulated by subtype specific genomic and epigenomic disruptions are responsible for driving the differential development of AC and SqCC. Furthermore, our results highlight the impact of this novel integrative genome, epigenome and transcriptome analysis in identifying robust target genes that can be used as biomarkers of disease.

Figure 2.5

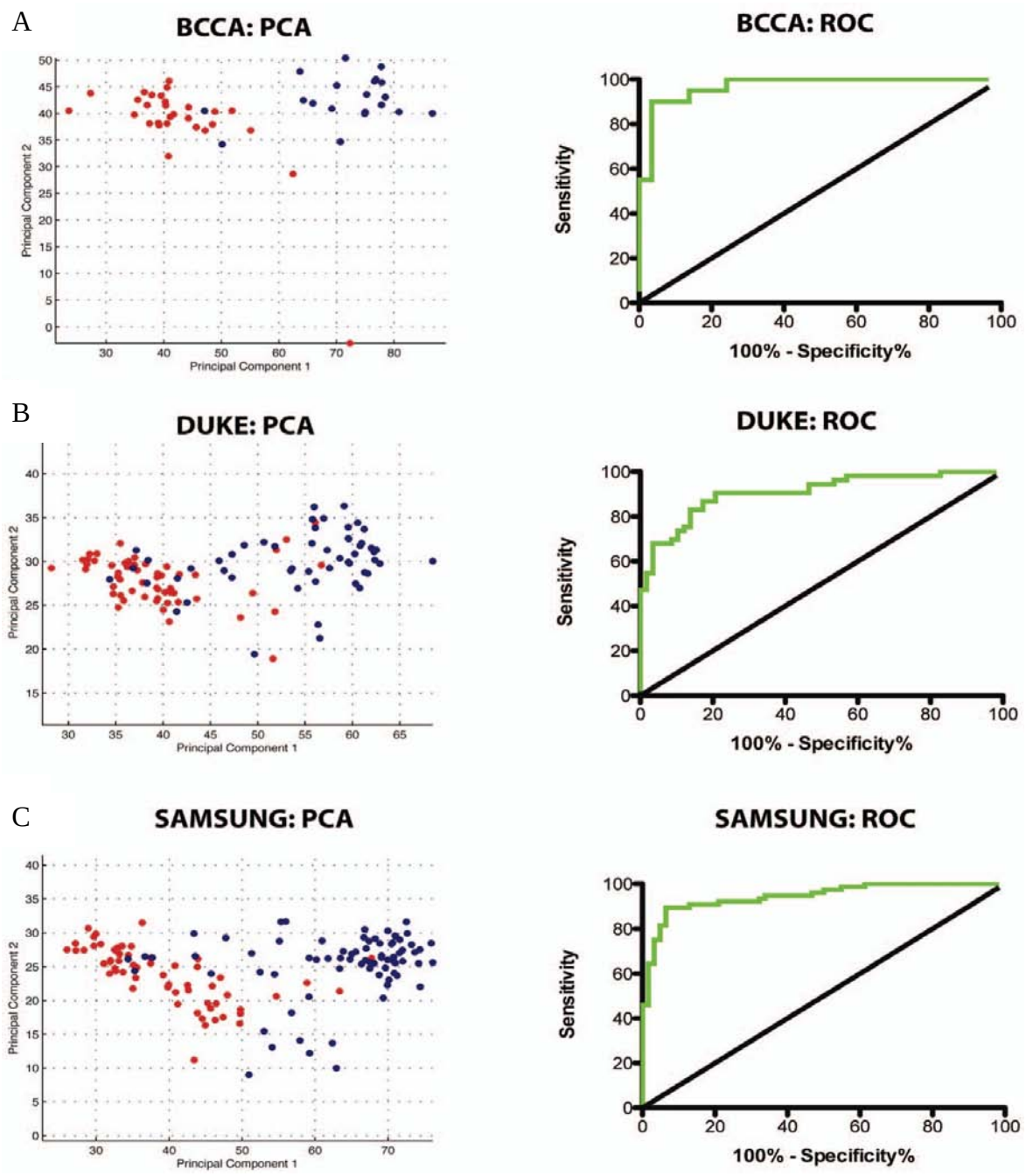


Figure 2.5. Subtype specific genes explain AC and SqCC phenotypes.

Figure 2.5. Subtype specific genes explain AC and SqCC phenotypes.
Principal components analysis of all genes demonstrating expression differences between subtypes as a result of genetic and/or epigenetic alterations in: (A) 49 NSCLC tumors (29 AC, 20 SqCC) used for gene discovery in this study, publically available data from (B) 111 NSCLC tumors (58 AC and 53 SqCC, Sample Set #3-Duke,) and (C) 138 NSCLC tumors (62 AC and 76 SqCC, Sample Set #6- Samsung,). Red circles indicate AC samples, blue circles indicate SqCC samples. Strong separation of tumors along principal component 1 is observed in all sets, demonstrating the contribution of these genes to the differential phenotypes. The respective ROC curves for each dataset using the respective principle component 1 values for each sample are shown on the right. AUC values of 0.9690, 0.9076 and 0.9442 for A, B and C, respectively, suggest the gene expression signature is accurately discriminates subtypes.

2.3.8 Subtype-specific genetic differences are translated to the protein level

To confirm that the genome and transcriptome differences between subtypes affects the relative protein levels of the identified genes, we performed immunohistochemical (IHC) analysis on an independent panel of >200 lung tumors. Protein levels for three subtype specific genes with available antibodies validated for IHC were analyzed: ERCC1 (inactivated in AC), KEAP1 (inactivated in AC) and SOX2 (activated in SqCC) (Figure 2.6). Average protein levels for all three genes were significantly different between subtypes in the direction predicted by our integrative analysis (Figure 2.6a, d, and g, $p<0.001$, unpaired t test with Welch's correction). The average nuclear ERCC1 expression and cytoplasmic KEAP1 expression were significantly lower in AC tumors (ERCC1:43.45±5.389, N = 175 and KEAP1:126.5 ±4.179, N = 184) compared to SqCC tumors (ERCC1: 79.99±9.095, N = 106 and KEAP1:160.9 ±5.401, N = 110) consistent with these genes being inactivated in AC. Conversely, nuclear levels of SOX2, a SqCC lineage specific oncogene were significantly higher in SqCC (206.5±8.839, N = 106) than AC (70.39± 6.342, N = 170). These data demonstrate that the genomic, epigenomic and expression differences between subtypes are translated to the protein level, supporting the hypothesis that these changes have a functional consequence on the phenotypes of AC and SqCC.

Figure 2.6

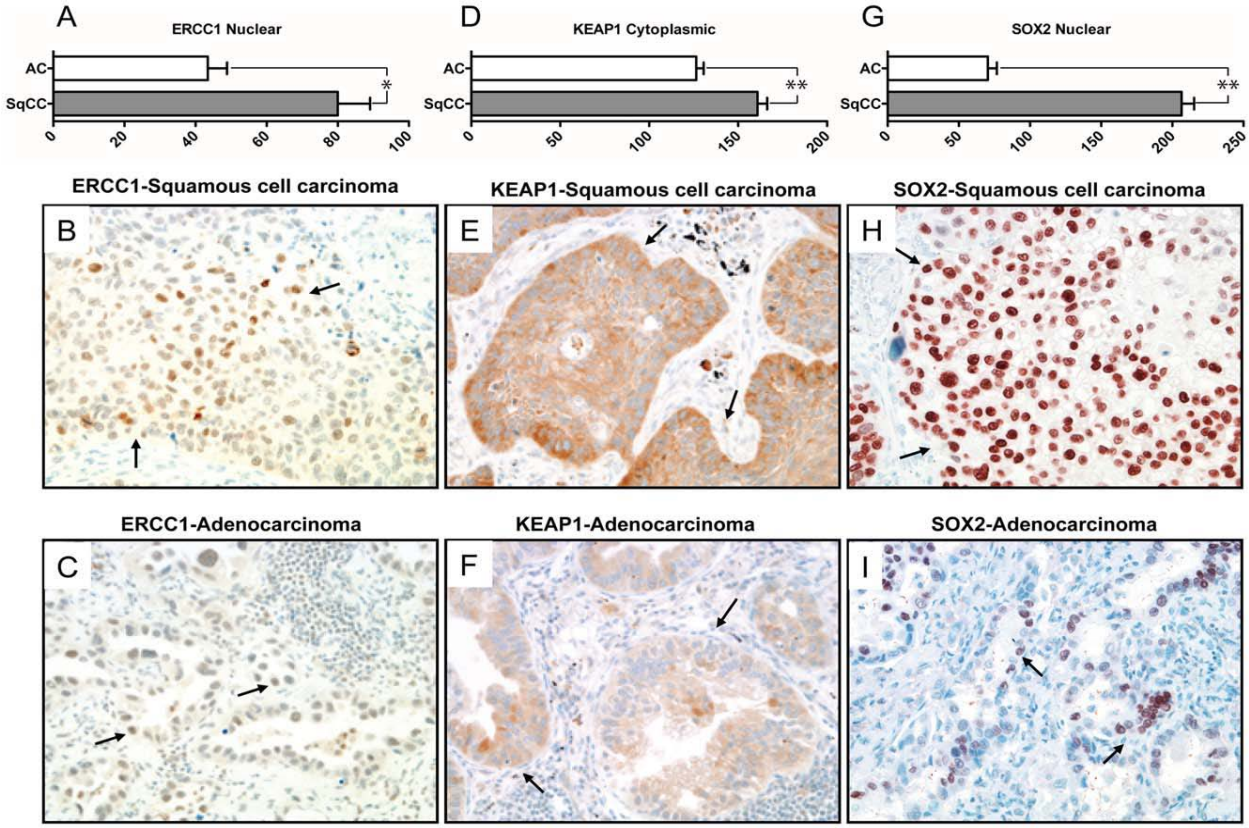


Figure 2.6. Subtype-specific genomic differences are reflected at the protein level.
Immunohistochemical analysis of protein levels for ERCC1 (A–C), KEAP1 (D–F) and SOX2 (G–I) in squamous and adenocarcinoma lung tumors. Average immunohistochemical protein expression levels for each subtype are plotted ± SEM of each group. Representative microphotographs showing tumoral cells (arrows) with higher levels of expression of nuclear ERCC1 (B and C), cytoplasmic KEAP1 (E and F) and nuclear SOX2 (H and I) in squamous cell carcinomas (B, E and H) compared to lung adenocarcinomas (C, F, and I). Images are of samples reflecting the average protein expression for each group (ERCC1: SqCC = ~80, AC =~43; KEAP1: SqCC = ~161, AC = ~126; SOX2: SqCC = ~207, AC = ~70). Magnification 200x. * and ** = $p<0.001$ and $p<0.0001$, two-tailed unpaired t test with Welch's correction, respectively.

2.3.9 Subtype-specific genes are associated with distinct clinical characteristics

In an attempt to elucidate the influence of our subtype specific genes on the clinical characteristics of AC and SqCC we determined the survival associations using a Mantel-Cox log rank test for each of the 778 subtype specific genes in AC, SqCC and NSCLC as a whole in sample set #6 which has overall survival information available for 138 cases (62 AC and 76 SqCC). Since these genes are responsible for defining the distinct biology of these diseases, we reasoned that their expression should only correlate with clinical features in one subtype and not the other subtype or NSCLC (AC + SqCC) as a whole. 131 AC and 46 SqCC specific genes had significant ($p < 0.05$) associations with overall survival (Table 2.3). Remarkably, the associations were completely specific to an individual subtype as no genes were correlated with survival in the same manner across both subtypes. Six genes (*DSG2*, *PLAC2*, *ATP9A*, *TPM4*, *CD9* and *PSMD11*) were significantly associated with survival in both subtypes; however, they displayed opposing survival patterns, with low expression associated with poor survival in one subtype and high expression with poor survival in the other (Figure 2.7). Thus, although associated with survival in both subtypes, these genes exhibit distinct subtype-specific associations. Interestingly, low levels of *CD9* expression have been previously implicated in the poor prognosis of lung cancer patients [94]. However, we now show that this association is subtype specific, with low levels of *CD9* correlated with poor prognosis in SqCC and high levels with poor prognosis in AC (Figure 2.7). Importantly, only eight genes that were associated with survival in one of the subtypes were also significant when analyzing NSCLC as a whole, providing further evidence to the importance of treating the subtypes as separate disease entities and underscoring the potential clinical relevance of subtype specific alterations.

Table 2.3 Genes demonstrating a significant association with survival

		AC (n=62)			NSCLC (n=138)			SqCC (n=76)		
Probe Set ID	**Gene Symbol**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**
235053_at	---	0.0144	16.9405	49.44151	0.17068	24.50015	49.4357	0.53886	39.27745	Undefined
238277_at	---	0.03738	33.81425	Undefined	0.71096	49.43922	45.26797	0.3616	51.27054	34.00295
235288_at	---	0.00853	24.33483	Undefined	0.51286	51.27393	Undefined	0.25026	51.27537	34.00406
235772_at	---	0.00036	21.33842	Undefined	0.2415	34.00001	42.80249	0.30358	45.27746	31.14234
240969_at	---	0.14949	24.50005	Undefined	0.57797	51.27273	39.00354	0.01258	Undefined	23.63333
1559360_at	---	0.37749	49.4449	27.2	0.0207	Undefined	29.13404	0.03775	Undefined	28.8752
240527_at	---	0.38033	49.44523	Undefined	0.3844	55.36737	39.00058	0.04488	Undefined	23.63478
238580_at	---	0.31036	33.8017	42.806	0.10849	Undefined	28.10062	0.02382	Undefined	28.10747
225318_at	---	0.01158	21.33602	Undefined	0.00396	24.50067	Undefined	0.13721	28.86847	Undefined
236010_at	---	0.02674	33.80404	Undefined	0.85402	39.26732	42.80081	0.32808	39.26707	31.14145
238281_at	---	0.00073	19.83424	Undefined	0.30314	28.86707	42.80193	0.12523	Undefined	31.13674
1560625_s_at	---	0.35725	39.60002	24.5	0.10421	49.43598	28.36731	0.00727	Undefined	28.1056
202394_s_at	ABCF3	0.04375	Undefined	24.13333	0.73065	49.43428	51.27379	0.24315	34.00969	Undefined
212895_s_at	ABR	0.73575	42.83241	28.36667	0.95439	42.8039	39.26733	0.03852	17.0087	Undefined
1554390_s_at	ACTR2	0.0024	Undefined	22.57662	0.41497	42.80818	34.00004	0.42825	31.13862	51.27873
209321_s_at	ADCY3	0.37955	39.00968	24.50383	0.07318	39.0035	Undefined	0.01947	14.46667	Undefined
205268_s_at	ADD2	0.0208	24.33711	Undefined	0.05083	34.00914	51.26791	0.18153	34.00443	Undefined
223145_s_at	AKIRIN2	0.0042	Undefined	19.83365	0.39962	45.26702	39.27015	0.32392	39.60958	Undefined
224982_at	AKT1S1	0.03429	Undefined	21.33333	0.26164	Undefined	33.80263	0.42188	Undefined	39.27398
236626_at	ALG1	0.28678	33.80035	55.3923	0.38036	Undefined	34.00052	0.04702	Undefined	28.1
1559640_at	ANKFN1	0.03854	24.1401	Undefined	0.20954	28.36686	51.2675	0.6613	Undefined	39.6
208074_s_at	AP2S1	0.02846	Undefined	24.33346	0.0408	Undefined	31.1347	0.55993	Undefined	39.27003
206632_s_at	APOBEC3B	0.47521	49.4339	33.80355	0.10034	39.60002	Undefined	0.01601	28.87687	Undefined

		AC (n=62)			NSCLC (n=138)			SqCC (n=76)		
Probe Set ID	**Gene Symbol**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**
223665_at	ARPM1	0.0041	55.37556	16.93581	0.03967	Undefined	28.86709	0.81962	33.8	31.90115
202872_at	ATP6V1C1	0.04422	39.60013	21.33399	0.60693	45.27069	33.80082	0.79845	45.27323	51.29902
216129_at	ATP9A	0.0126	27.20467	Undefined	0.85486	51.26829	42.80025	0.01832	Undefined	28.1
220488_s_at	BCAS3	0.0401	27.21179	55.39337	0.37157	28.36738	39.60042	0.75453	39.60025	33.80295
229177_at	C16orf89	0.02698	29.14723	55.37189	0.53278	31.90096	45.2675	0.7107	39.27301	34.00405
219114_at	C3orf18	0.02553	20.30096	Undefined	0.14933	28.86671	42.80078	0.80837	45.273	Undefined
220218_at	C9orf68	0.03902	29.13582	Undefined	0.80718	49.44241	39.60069	0.21869	Undefined	31.13333
1553693_s_at	CBR4	0.10425	33.57563	Undefined	0.94272	39.26718	29.13397	0.0402	Undefined	23.63478
202047_s_at	CBX6	0.03347	24.33355	Undefined	0.00724	24.3334	Undefined	0.20137	28.86884	Undefined
1553886_at	CCDC108	0.2563	27.2001	49.43969	0.37501	51.26785	42.80266	0.00775	Undefined	14.4692
226723_at	CCDC23	0.84863	49.46235	39.00205	0.0259	Undefined	28.3667	0.00063	Undefined	21.00575
201005_at	CD9	0.02904	Undefined	22.56753	0.53697	39.00059	51.2691	0.03468	28.87073	Undefined
235117_at	CHAC2	0.02397	49.43635	19.83681	0.85201	49.44391	51.26889	0.52374	39.27526	Undefined
218566_s_at	CHORDC1	0.0101	Undefined	19.1687	0.50217	42.8012	29.13335	0.14139	31.90955	51.28971
208925_at	CLDND1	0.0018	Undefined	22.57001	0.4322	Undefined	45.2716	0.67328	39.61269	Undefined
230609_at	CLINT1	0.03465	24.3392	Undefined	0.57718	39.27082	49.43703	0.05689	39.27138	17.00452
216295_s_at	CLTA	0.00431	Undefined	24.13618	0.75029	45.26681	39.60227	0.15425	23.63333	Undefined
206158_s_at	CNBP	0.02958	Undefined	22.56982	0.86081	49.43538	51.27565	0.36265	31.13497	Undefined
221676_s_at	CORO1C	0.00315	Undefined	21.33395	0.57564	39.60005	51.27283	0.30983	28.1	51.28618
217889_s_at	CYBRD1	0.01382	Undefined	29.13997	0.90114	45.26687	51.27134	0.25109	28.1	51.26834
223385_at	CYP2S1	0.68764	28.36735	29.13414	0.16294	28.86693	Undefined	0.02874	17	Undefined
226745_at	CYP4V2	0.3883	Undefined	39.61238	0.04372	Undefined	39.00442	0.6401	28.10529	39.27164
1555301_a_at	DIP2A	0.88767	Undefined	55.38224	0.08636	Undefined	31.13481	0.02417	Undefined	31.14654
202514_at	DLG1	0.00817	Undefined	24.33561	0.65233	Undefined	51.27229	0.75011	34.00781	Undefined

		AC (n=62)			NSCLC (n=138)			SqCC (n=76)		
Probe Set ID	**Gene Symbol**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**
213707_s_at	DLX5	0.02338	20.30062	42.80028	0.44252	33.57163	33.80075	0.23382	34.00095	Undefined
209187_at	DR1	0.11236	Undefined	39.00756	0.4124	Undefined	39.26692	0.04499	Undefined	23.63478
1553105_s_at	DSG2	0.00731	Undefined	21.33333	0.57155	Undefined	51.27133	0.01533	31.13771	Undefined
204455_at	DST	0.02439	Undefined	24.5	0.47177	55.36794	45.27146	0.94676	39.60211	51.30379
227103_s_at	ECE2	0.03216	Undefined	24.50024	0.61781	45.2674	39.26885	0.27278	23.63333	51.28227
219787_s_at	ECT2	0.04794	Undefined	27.20217	0.96564	39.60003	45.27416	0.9103	39.27167	31.9
227540_at	EEFSEC	0.12983	33.80965	55.37667	0.08076	33.80235	55.36909	0.03373	23.63478	Undefined
208264_s_at	EIF3J	0.01706	Undefined	19.83377	0.45044	Undefined	33.80004	0.78744	39.26773	33.80045
210213_s_at	EIF6	0.01797	Undefined	21.33505	0.24002	42.80152	28.86696	0.64135	34.00444	28.86746
1555274_a_at	EPT1	0.00473	Undefined	24.13482	0.79192	55.36992	51.26965	0.37729	39.27466	51.27142
224576_at	ERGIC1	0.13149	Undefined	28.36758	0.0231	Undefined	29.1339	0.59119	39.27742	33.80355
218481_at	EXOSC5	0.01974	49.43491	21.33395	0.31024	49.43857	33.80179	0.96215	45.27775	33.8022
215133_s_at	FAM153A/B	0.10274	Undefined	22.56667	0.03727	Undefined	23.63367	0.07532	Undefined	31.13787
211623_s_at	FBL	0.0136	Undefined	22.56795	0.26212	55.37016	33.8006	0.69323	Undefined	51.2858
225737_s_at	FBXO22	0.02196	Undefined	19.16994	0.60894	39.60087	33.80066	0.83565	34.00913	39.27489
228220_at	FCHO2	0.00771	42.80089	20.3	0.53555	33.80002	34.00015	0.11278	17.00242	51.27968
218880_at	FOSL2	0.03719	42.80044	21.3341	0.11227	Undefined	31.1347	0.86562	34.00245	39.27724
219170_at	FSD1	0.48502	39.00755	24.5	0.19504	33.80163	Undefined	0.02367	28.10747	Undefined
205384_at	FXYD1	0.02207	24.50712	Undefined	0.11755	28.36939	51.26727	0.62167	33.81213	45.27641
217398_x_at	GAPDH	0.00599	Undefined	21.33395	0.29288	Undefined	33.80167	0.94368	Undefined	Undefined
225161_at	GFM1	0.00928	Undefined	19.1687	0.21403	Undefined	39.26803	0.52363	34.01822	31.13856
218473_s_at	GLT25D1	0.02424	Undefined	20.3	0.73697	45.27217	49.44348	0.24779	33.80179	51.26929
235678_at	GM2A	0.23687	42.81498	19.83333	0.61717	42.80336	Undefined	0.0219	22.5	Undefined
214431_at	GMPS	0.0474	Undefined	27.2023	0.43364	49.43901	33.80168	0.29773	Undefined	31.90064

		AC (n=62)			NSCLC (n=138)			SqCC (n=76)		
Probe Set ID	**Gene Symbol**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**
210761_s_at	GRB7	0.01647	Undefined	28.37302	0.33651	45.27118	33.80045	0.23033	39.60983	Undefined
213911_s_at	H2AFZ	0.01466	Undefined	24.50124	0.94486	49.43829	39.60273	0.48193	39.27751	33.80045
206194_at	HOXC4	0.05237	24.33779	Undefined	0.02445	24.33492	Undefined	0.18757	22.50201	39.60027
217805_at	ILF3	0.00013	Undefined	24.33919	0.06694	Undefined	33.80253	0.80721	34.00397	28.1
201389_at	ITGA5	0.02179	Undefined	20.3	0.09358	42.80019	28.86712	0.79504	33.80209	39.27456
239695_at	JAK1	0.0454	29.13355	Undefined	0.48185	34.00013	42.80262	0.68832	31.9	45.27641
202417_at	KEAP1	0.05737	Undefined	27.2	0.91753	45.26791	49.4404	0.04781	39.6006	Undefined
213208_at	KIAA0240	0.43129	24.50348	39.62181	0.02893	22.50118	55.36901	0.03511	28.10483	Undefined
226328_at	KLF16	0.01845	Undefined	22.56667	0.79878	39.00306	45.27031	0.06192	33.81203	Undefined
221986_s_at	KLHL24	0.03703	21.33523	Undefined	0.02229	24.13362	Undefined	0.11447	28.87458	51.27145
225267_at	KPNA4	0.03218	Undefined	24.3359	0.64512	Undefined	39.60009	0.49027	45.27188	33.80006
34031_i_at	KRIT1	0.08207	42.80441	33.57252	0.92975	39.6014	39.60017	0.04527	28.87493	Undefined
209008_x_at	KRT8	0.0298	Undefined	22.56667	0.18157	55.37109	28.1	0.81381	39.27259	28.1
216952_s_at	LMNB2	0.01254	Undefined	12.46667	0.91158	55.37024	Undefined	0.29715	34.00048	Undefined
240936_at	LOC100287290	0.0142	22.56894	Undefined	0.92029	45.27692	39.00242	0.36183	51.27015	31.90132
229187_at	LOC283788	0.28723	28.36682	Undefined	0.20841	Undefined	39.60002	0.027	Undefined	34.00403
202736_s_at	LSM4	0.00474	Undefined	24.34862	0.71931	39.60217	33.80062	0.45598	34.0006	Undefined
224656_s_at	LUZP6/MTPN	0.00987	Undefined	29.13705	0.60054	Undefined	34.00004	0.09512	39.61736	Undefined
242838_at	MAP6D1	0.3869	42.81631	Undefined	0.05706	33.56682	Undefined	0.03499	34.00369	Undefined
200712_s_at	MAPRE1	0.00286	Undefined	12.46868	0.40758	49.43642	34.00003	0.29926	17	39.26873
204825_at	MELK	0.02125	55.37232	22.5686	0.53408	49.43576	33.80078	0.58503	51.30354	Undefined
224725_at	MIB1	0.00118	Undefined	19.17769	0.3179	55.36692	33.80026	0.62303	34.0042	Undefined
231975_s_at	MIER3	0.00445	Undefined	28.37263	0.0452	Undefined	33.80143	0.40189	39.60096	34.00044
224784_at	MLLT6	0.58994	29.13854	39.00102	0.06093	31.90083	49.43717	0.00102	17	Undefined

		AC (n=62)			NSCLC (n=138)			SqCC (n=76)		
Probe Set ID	**Gene Symbol**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**
223086_x_at	MRPL51	0.00434	Undefined	21.33395	0.45839	42.80053	33.80126	0.5265	39.60044	51.27282
204331_s_at	MRPS12	0.00082	Undefined	24.13416	0.45616	49.43749	39.27097	0.09454	34.02697	Undefined
213380_x_at	MST1P9	0.81968	22.57145	33.80789	0.04798	Undefined	33.80322	0.00081	Undefined	22.50201
205455_at	MST1R	0.69461	24.50304	33.8026	0.19801	Undefined	39.00408	0.01155	Undefined	34.0004
226856_at	MUSTN1	0.03969	21.33717	49.43585	0.62207	42.80249	51.26764	0.4065	Undefined	51.27849
228846_at	MXD1	0.00339	Undefined	19.83333	0.01847	Undefined	22.50033	0.6546	Undefined	39.63075
212462_at	MYST4	0.10006	28.37113	Undefined	0.01361	28.86671	Undefined	0.64988	39.60051	Undefined
217745_s_at	NAA50	0.03094	Undefined	28.36976	0.89708	49.43981	45.27204	0.8809	39.60787	33.8019
204725_s_at	NCK1	0.0232	Undefined	29.14091	0.45144	45.26871	Undefined	0.08955	31.13754	Undefined
208969_at	NDUFA9	0.01447	49.43586	21.33395	0.52469	42.8016	33.80242	0.82327	39.26987	39.60899
219396_s_at	NEIL1	0.04625	27.22196	55.37556	0.20662	33.56729	49.43815	0.40991	31.1358	Undefined
218036_x_at	NMD3	0.00253	Undefined	24.33426	0.24537	Undefined	33.80111	0.73953	45.2738	39.27156
205129_at	NPM3	0.00325	Undefined	24.5	0.16436	Undefined	31.90049	0.80558	51.27316	33.80546
207740_s_at	NUP62	0.03487	Undefined	24.33683	0.18146	55.37659	31.90033	0.95715	34.00119	39.27398
215952_s_at	OAZ1	0.03153	Undefined	28.37069	0.83043	45.26775	51.26712	0.14832	31.90753	Undefined
1567245_at	OR5J2	0.55999	39.00086	Undefined	0.29774	39.00273	24.33406	0.00638	Undefined	14.13568
220669_at	OTUD4	0.28954	39.61079	27.2	0.00609	Undefined	31.13369	0.03617	Undefined	31.90056
207634_at	PDCD1	0.05585	28.36714	Undefined	0.03491	31.13376	Undefined	0.48531	39.62476	Undefined
202212_at	PES1	0.00027	Undefined	19.16994	0.1771	Undefined	39.26986	0.76431	34.00568	33.80504
201600_at	PHB2	0.00212	Undefined	12.46749	0.26775	49.43491	31.90047	0.58147	39.6005	33.8036
226846_at	PHYHD1	0.85006	49.47308	33.80646	0.12844	Undefined	33.80158	0.02709	Undefined	31.90501
204297_at	PIK3C3	0.01335	Undefined	24.14054	0.14753	45.26707	29.13394	0.59912	Undefined	33.8
202522_at	PITPNB	0.00512	42.80212	11.6	0.44106	42.80204	31.90069	0.61958	34.00213	39.27623
219584_at	PLA1A	0.1342	24.13645	Undefined	0.01409	28.1017	Undefined	0.30541	39.28778	51.273

		AC (n=62)			NSCLC (n=138)			SqCC (n=76)		
Probe Set ID	**Gene Symbol**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**
244374_at	PLAC2	0.0089	24.33367	Undefined	0.9082	39.00066	34.00004	0.02157	Undefined	31.14163
201411_s_at	PLEKHB2	0.02874	Undefined	27.20319	0.15134	Undefined	29.13356	0.37245	33.80401	51.27426
217841_s_at	PPME1	0.02953	Undefined	28.36886	0.13805	Undefined	39.26963	0.69668	Undefined	33.80063
201979_s_at	PPP5C	0.01205	Undefined	27.20299	0.12645	Undefined	33.80025	0.83953	39.6096	33.80045
203056_s_at	PRDM2	0.03028	29.13442	Undefined	0.26908	31.13423	45.27053	0.26807	34.00161	28.10422
200707_at	PRKCSH	0.0038	Undefined	27.20319	0.94375	39.60022	45.26917	0.38645	33.80392	45.27737
206445_s_at	PRMT1	0.01499	Undefined	27.20435	0.14026	Undefined	33.80131	0.29468	Undefined	33.80116
201267_s_at	PSMC3	0.00053	Undefined	29.14001	0.09155	Undefined	34	0.97923	39.60516	39.2773
208777_s_at	PSMD11	0.03359	Undefined	28.36795	0.91385	49.43483	55.36745	0.03752	34.01822	Undefined
204748_at	PTGS2	0.12298	Undefined	29.13333	0.03916	Undefined	31.90193	0.15653	Undefined	28.87493
222981_s_at	RAB10	0.03622	49.44053	19.83377	0.98263	49.43854	51.27597	0.58024	34.00554	Undefined
202483_s_at	RANBP1	0.00028	Undefined	19.16994	0.29456	55.37286	39.27082	0.69617	39.60904	33.8021
1552482_at	RAPH1	0.85588	28.37231	39.00947	0.24756	Undefined	39.60089	0.0052	Undefined	34.00609
205091_x_at	RECQL	0.01306	Undefined	29.14178	0.65705	45.26784	39.60024	0.56676	45.27521	39.27466
205205_at	RELB	0.00024	Undefined	12.46826	0.0613	42.80261	28.10416	0.42162	31.90878	51.28158
220334_at	RGS17	0.9146	49.44602	29.13333	0.19007	49.43744	28.36678	0.00478	Undefined	21.00575
202129_s_at	RIOK3	0.01461	Undefined	29.13465	0.93382	42.80218	34.00016	0.14088	23.63333	51.29828
203022_at	RNASEH2A	0.01996	Undefined	27.20329	0.91781	39.60158	39.60411	0.86278	39.27691	31.1415
204208_at	RNGTT	0.03237	Undefined	29.13469	0.12068	Undefined	33.80109	0.63752	Undefined	51.28576
234243_at	RPF1	0.01228	19.83474	Undefined	0.91676	39.26867	42.80095	0.10884	39.26785	22.5
200022_at	RPL18	0.04181	Undefined	27.20692	0.02266	Undefined	28.10041	0.63984	Undefined	Undefined
230695_s_at	RSPH9	0.21092	33.81375	33.56745	0.28838	33.80153	33.57037	0.04879	31.13333	Undefined
230378_at	SCGB3A1	0.59364	33.80097	55.38584	0.09802	31.13412	55.36957	0.00959	28.88063	Undefined
238017_at	SDR16C5	0.24231	42.81232	24.33362	0.03907	Undefined	34.00241	0.39325	Undefined	34.00689

		AC (n=62)			NSCLC (n=138)			SqCC (n=76)		
Probe Set ID	**Gene Symbol**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**
57703_at	SENP5	0.00984	Undefined	22.56849	0.88425	49.43804	39.6003	0.4657	34.00108	31.13936
205637_s_at	SH3GL3	0.03148	22.56814	Undefined	0.06137	29.13336	Undefined	0.99496	51.2827	45.27023
214437_s_at	SHMT2	0.00877	Undefined	24.5004	0.45624	49.43966	33.80165	0.71676	39.27181	31.14163
227791_at	SLC9A9	0.41764	33.80973	55.38054	0.05746	28.36677	Undefined	0.04263	28.10127	Undefined
233759_s_at	SMEK2	0.04774	Undefined	29.13447	0.57193	39.0003	51.27423	0.11077	39.27633	Undefined
202690_s_at	SNRPD1	0.02425	Undefined	27.20323	0.42348	55.36969	33.80201	0.75276	51.27334	39.60742
200826_at	SNRPD2	0.00192	Undefined	24.13475	0.0512	Undefined	31.90002	0.20704	Undefined	28.86667
208608_s_at	SNTB1	0.00764	Undefined	24.50541	0.11671	Undefined	33.56895	0.31805	39.27284	Undefined
234005_x_at	STK36	0.4891	Undefined	29.13392	0.04193	Undefined	33.80256	0.2361	Undefined	31.90206
200870_at	STRAP	0.01224	Undefined	22.56753	0.88866	55.37286	51.27355	0.30975	39.60044	Undefined
217834_s_at	SYNCRIP	0.00745	Undefined	27.20213	0.09284	49.43463	31.13375	0.83177	45.29099	51.2942
217839_at	TFG	0.00023	Undefined	11.6	0.10494	Undefined	33.8009	0.78479	Undefined	51.28579
207332_s_at	TFRC	0.00253	Undefined	19.1687	0.93278	42.80181	51.26735	0.50376	39.6108	33.8
203235_at	THOP1	0.54289	39.60594	27.20092	0.56371	39.60616	49.43707	0.03206	34.01432	Undefined
1552522_at	TIGD4	0.08266	24.14291	49.43914	0.46013	Undefined	34.00004	0.0356	Undefined	34.00185
224413_s_at	TM2D2	0.04324	42.80067	28.36803	0.56527	45.2677	33.80005	0.60526	31.9124	39.60534
236430_at	TMED6	0.03763	24.50075	55.37892	0.39659	34.00011	55.37017	0.62181	33.80344	39.62128
225766_s_at	TNPO1	0.00087	Undefined	24.1406	0.05921	Undefined	31.13516	0.82285	Undefined	33.80204
217960_s_at	TOMM22	0.00386	Undefined	24.34083	0.42893	Undefined	39.60017	0.71248	39.60655	39.27185
201512_s_at	TOMM70A	0.00499	Undefined	19.1687	0.20854	Undefined	39.26963	0.88668	34.0042	39.27171
213011_s_at	TPI1	0.0212	Undefined	22.56767	0.26103	Undefined	31.13388	0.79194	Undefined	Undefined
212481_s_at	TPM4	0.02596	Undefined	22.57022	0.63006	55.37739	39.27207	0.02658	17	Undefined
236020_s_at	TRUB1	0.0006	24.33376	Undefined	0.01604	29.13403	Undefined	0.76752	31.13333	39.60318
1557073_s_at	TTBK2	0.65558	55.46281	39.00417	0.72421	39.60268	51.27788	0.03372	31.13436	Undefined

		AC (n=62)			NSCLC (n=138)			SqCC (n=76)		
Probe Set ID	**Gene Symbol**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**	**P Value**	**Median Survival (Low)**	**Median Survival (High)**
211337_s_at	TUBGCP4	0.0071	Undefined	12.46667	0.68162	55.37233	51.27741	0.20628	34.00365	Undefined
214007_s_at	TWF1	0.02118	Undefined	11.60108	0.7189	42.80104	39.60032	0.09158	31.13565	Undefined
222601_at	UBA6	0.04354	39.60059	22.56767	0.11044	Undefined	33.80161	0.46036	51.29911	45.27548
227413_at	UBLCP1	0.01242	42.80171	24.13996	0.11705	Undefined	34.00001	0.80738	45.26912	Undefined
225655_at	UHRF1	0.02933	Undefined	27.20192	0.44667	Undefined	39.60386	0.7462	39.27203	33.80063
216775_at	USP53	0.37925	27.20326	33.80577	0.05508	51.26675	29.13395	0.00277	Undefined	17
229369_at	VSIG2	0.01852	20.30096	Undefined	0.30675	31.90264	45.26791	0.60017	39.27796	34.00913
219060_at	WDYHV1	0.02266	Undefined	24.33353	0.68566	49.44061	39.267	0.19027	39.60983	Undefined
208775_at	XPO1	0.0335	Undefined	29.13592	0.84698	Undefined	51.27005	0.79137	34.01252	33.80342
207757_at	ZFP2	0.00983	21.33607	Undefined	0.65835	31.90012	42.80369	0.17654	39.26719	34.01328
241793_at	ZMYND17	0.44512	33.56949	28.36771	0.55572	49.43697	27.20028	0.02879	Undefined	22.50407
232117_at	ZNF471	0.16299	Undefined	22.56779	0.01829	Undefined	33.57015	0.02274	Undefined	22.5
238454_at	ZNF540	0.02493	28.37374	Undefined	0.15849	31.13421	42.80043	0.81279	33.81307	34.00037
218349_s_at	ZWILCH	0.01393	Undefined	27.20322	0.70174	42.80079	39.26778	0.44344	34.01252	51.27434

Figure 2.7

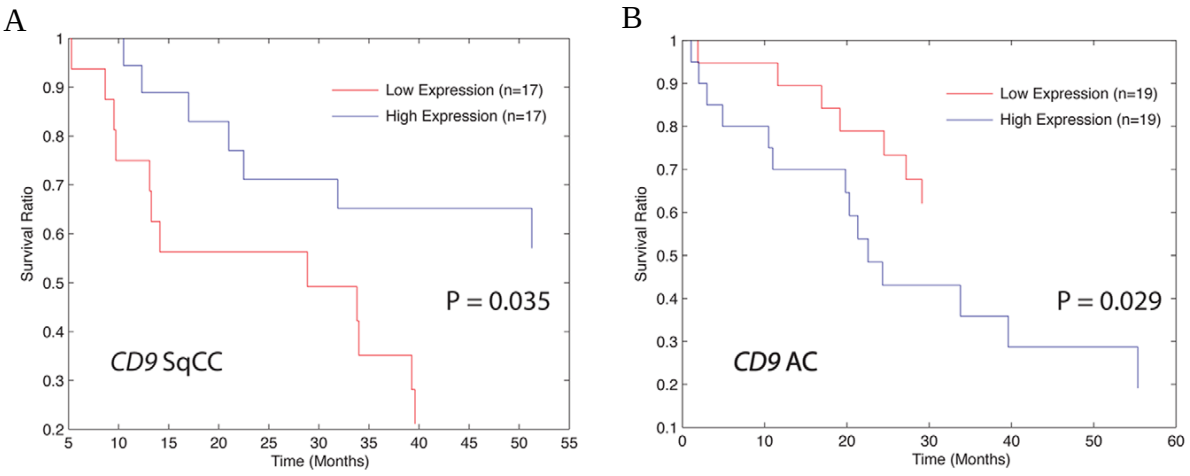


Figure 2.7 CD9 expression is associated with differences in survival in AC and SqCC tumors.
The prognostic value of *CD9* expression levels was evaluated in 53 SqCC tumors and 58 AC tumors. Poor prognosis is significantly associated (Mantel-Cox log test, $p<0.05$) with low *CD9* levels in SqCC (A) and high *CD9* in AC (B). Survival of the 1/3 lowest *CD9* expressers is shown in red, and the top 1/3 is shown in blue.

2.3.10 Defining putative treatment strategies tailored to lung cancer subtypes using *in silico* screening of candidate therapeutic compounds

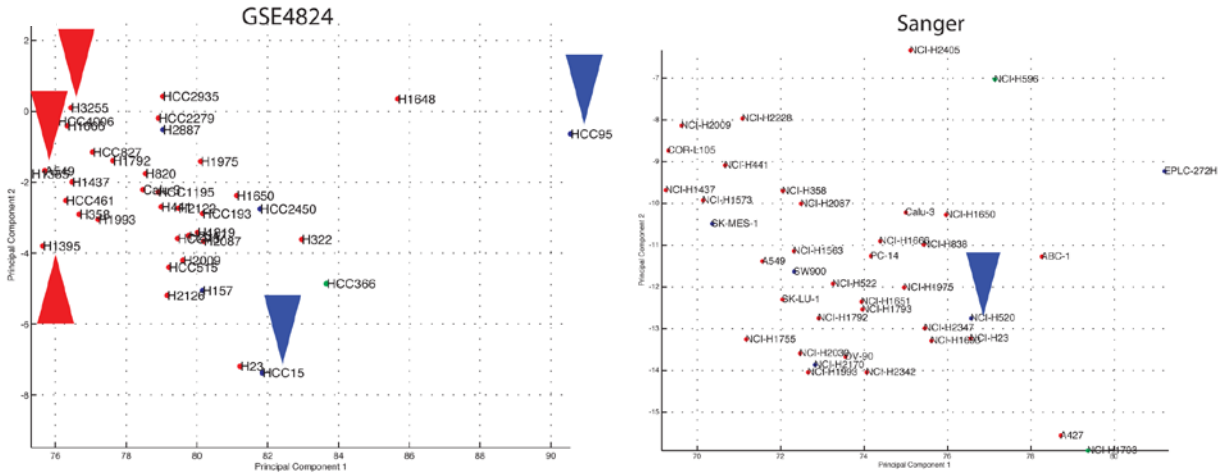
Lastly, after defining and validating our AC and SqCC specific cancer genes, we applied these findings to define potential treatment strategies tailored to each lung cancer subtype. Using our subtype specific genes, we queried the Connectivity Map (CMAP) database to identify compounds that could "reverse" the expression direction of each subtype signature. CMAP consists of thousands of gene expression profiles from different cancer cell lines treated with a vast collection of small molecules [83]. By comparing the subtype specific signatures of up and down regulated genes with preexisting response signatures, CMAP identifies small molecules whose effects on gene expression changes are positively or negatively correlated. Negative correlation scores imply that the matched molecules have a mode of action that can reverse the expression direction of query genes, and therefore serve as potential therapeutic compounds. Using this *in silico* screening approach, we identified numerous instances (cell line/treatment

combination) that were significantly correlated with both our AC and SqCC specific gene signatures. SqCC had an expression signature that was negatively correlated with multiple HDAC and PI3K/ mTOR inhibitors including trichostatin A, vorinostat (also known as SAHA), LY-294002 and MS-275 - all HDAC inhibitors as well as quinostatin, sirolimus (also known as Rapamycin) and wortmannin - all PI3K/mTOR inhibitors. These findings were interesting given that the alteration of histone modifying enzymes was the major network disrupted in SqCC and we observed concerted disruption of PRC2 components, which are responsible for *de novo* methylation. In addition, *PIK3CA* activation (mutation and/or amplification) is known to occur more frequently in SqCC than AC [95] and many downstream components of this pathway were also altered specifically in SqCC. Conversely, CMAP analysis for AC was not very informative, as none of the negatively correlated molecules shared the same functions.

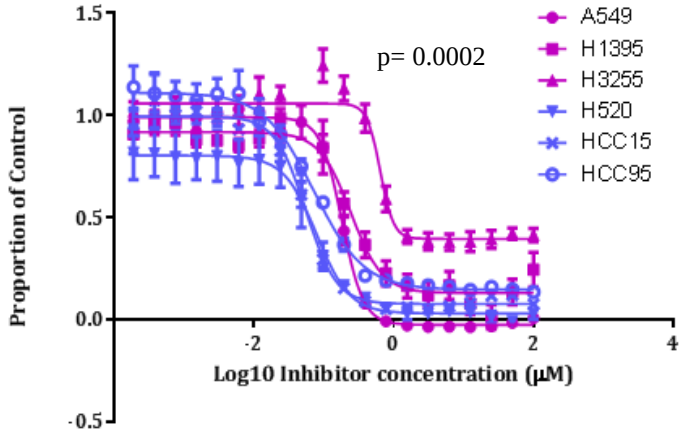
To confirm the results of the CMAP analysis, we treated a panel of six NSCLC cell lines (three AC and three SqCC) with the HDAC inhibitor Trichostatin A, the most significant negatively correlated HDAC inhibitor from the SqCC analysis. Importantly, we selected the available cell lines that best represented their respective clinical tumor subtypes by performing principle component analysis with the subtype-specific genes using publically available gene expression profiles for a large panel of NSCLC cell lines (Figure 2.8A). As predicted by the *in silico* analysis, SqCC cell lines were on average, five times more sensitive to Trichostatin A than AC cell lines (SqCC avg. IC_{50} =69nM, AC avg. IC_{50} = 346 nM, Mann-Whitney U Test, $p < 0.05$) (Figure 2.8B), validating the clinical relevance of subtype specific alterations.

Figure 2.8

A



B



Cell Line	Subtype	IC_{50}	SEM
H3255	Adeno	0.62	0.06232
H1395	Adeno	0.24	0.02246
A549	Adeno	0.18	0.02496
H520	SqCC	0.08	0.01186
HCC95	SqCC	0.073	0.01415
HCC15	SqCC	0.055	0.00858

Figure 2.8 SqCC cell lines are more sensitive to Trichostatin A than AC cell lines.

(A) Selection of cell lines for in vitro assays by PCA analysis of two lung cancer cell line datasets. Red indicates AC cell lines, blue SqCC cell lines and green NSCLC cell lines. (B) Dose-response analysis of the HDAC inhibitor Trichostatin A on the relative viability of AC (A549, H3255 and H1395) and SqCC (H520, HCC15 and HCC95) cell lines. Each curve was generated from the average data points from four separate experiments. Vertical error bars represent SEM. Student's t test p =0.0002; all AC replicate IC50 values vs. all SqCC replicate IC50 values. (C) Table with the average IC50 and SEM for each cell line tested derived from four separate experiments. Replicate experiments were highly repeatable.

2.4 Discussion

The emergence of tumor cells from normal precursors is thought to involve a complex interplay between genetics and cell lineage [96]. Due to the different cell types involved as well as the attributes of an individual cell's local environment, it is logical to assume different mechanisms are required for tumorigenesis of each lung cancer subtype. Previous studies suggest that distinct patterns of DNA alteration exist for AC and SqCC; however, the specific genes responsible for the different tumor phenotypes are largely unknown [33-35, 38]. At the time of publication, this study provide the first comprehensive investigation of the key genetic and epigenetic alterations distinguishing AC and SqCC lung tumors at the gene level. Through the integration of whole-genome DNA copy number, DNA methylation, and gene expression data, we identified 778 genes altered in a subtype-specific manner. These genes are associated with distinct gene networks, providing insight into the signaling pathways that contribute to subtype tumorigenesis. Furthermore, subtype-type specific changes were found to be correlated with clinical outcomes and revealed novel putative treatment strategies for SqCC.

While no difference in the percentage of AC or SqCC genomes altered by copy number was observed, SqCC tumors were found to be more hypomethylated, suggesting that the epigenetic machinery is highly deregulated in SqCC (Figure 2.1). There is precedent for this finding, as altered global methylation is thought to be a consequence of exposure to the carcinogens found in tobacco smoke [89, 90, 97]. Global hypomethylation, such as that caused by cigarette smoke, is also known to be associated with chromosomal instability. We identified a greater number of subtype specific alterations linked to both DNA copy number and DNA methylation in SqCC than AC. The reason for this is unclear, but may be indicative of similar selective pressures in the SqCC tumors that facilitate the development of recurrent alterations, whereas increased cellular and/or genetic heterogeneity in AC due to the different histological subtypes results in a greater diversity of alterations. Heterogeneity of clinical characteristics may also contribute to this discrepancy, as lung cancer in non-smokers are more likely to appear as AC tumors, and cigarette smoke can contribute to specific genetic or epigenetic alterations [4, 79]. Although 22.5% of our AC tumors were from never smokers, no significant differences in copy number were identified between AC tumors from ever and never smokers (data not shown).

The identification of subtype-specific copy number and methylation alterations with concordant expression changes demonstrate that different genetic pathways are involved in the pathogenesis of AC and SqCC. Previously identified lineage specific oncogenes including *SOX2* and *BRF2* were identified, validating our approach [70, 98]. Although some of the regions and genes altered by copy number have previously been shown to be important in NSCLC development, our findings suggest their newfound importance to a specific lung cancer subtype. For example, previously identified oncogenes *NOTCH3* and *FOXM1* were gained and overexpressed specifically in SqCC while the tumor suppressor *KEAP1* was deleted and underexpressed specifically in AC [99-101]. This is the first report suggesting these previously established lung cancer-associated genes are actually involved in subtype-specific tumorigenesis.

A gene network-based analysis of our subtype specific alterations revealed additional insights into the differential oncogenic mechanisms driving the pathogenesis of AC and SqCC. The top SqCC gene network associated with subtype specific copy number alterations was associated with DNA replication, recombination and repair, while SqCC specific genes altered by methylation were associated with the small cell signaling pathway. Of particular interest was the finding that the transcription factor *E2F1*,one of the deregulated components of this pathway, exhibited SqCC-specific hypomethylation and overexpression. *E2F1* is upregulated in SCLC tumors [93], and suppresses apoptosis and induces expression of *EZH2*, an oncogenic polycomb histone-methyltransferase [92]. Our observation that *EZH2* expression is significantly higher in SqCC than AC, along with the preferential disruption of polycomb group proteins in SqCC further supports the relevance of this pathway to SqCC, especially given the identification of SqCC specific deregulation of numerous histone modifying enzymes by copy number alterations. Histones are fundamental building blocks of eukaryotic chromatin and are involved in a myriad of cellular processes, including replication, repair, recombination and chromosome segregation [102-104]. Recently, global alterations of histone modification patterns have been reported in human cancers, with alterations occurring more frequently in SqCC than AC, consistent with our findings [105, 106]. Our data suggest that direct deregulation of histone modification enzymes including *ASF1B, PRMT1, SAE1, SET8, CHAF1A and UHRF1* may drive this phenomenon and play a key role during the development of lung SqCC. As histone modifications also play an

essential role in DNA replication, there may be a synergistic effect between the histone modifying genes and replication/recombination associated genes that contribute to tumor development.

The top gene network detected as perturbed in AC tumors contained genes mainly involved in regulating tissue development and cell-to-cell signaling and known to be targeted by the transcription factor *HNF4a*. *HNF4a* regulates a large set of genes in a cell-specific manner and is necessary for cell differentiation and maintenance of a differentiated epithelial phenotype [107]. In other carcinomas, deregulation of *HNF4a* leads to increased cellular proliferation, progression and dedifferentiation [108-110]. This suggests that *HNF4a* may act as a tumor suppressor in epithelial carcinogenesis [107]. Interestingly, although *HNF4a* itself was not affected, we found that numerous downstream targets of this gene are downregulated specifically in AC. This may have the same net affect as inactivation of HNF4a itself and lead to increased cellular proliferation during AC tumorigenesis.

Interestingly, we identified numerous genomic regions that showed opposite patterns of alteration in each lung cancer subtype. For example, a discrete alteration spanning 2.4 Mbp on chromosome bands 8p12-11.23 was commonly gained in SqCC and lost in AC, while *PARP11* was upregulated in SqCC by DNA hypomethylation and downregulated in AC by copy number loss, implying that these regions/genes may play opposite roles during the development of the individual NSCLC subtypes, acting as TSGs in AC and as oncogenes in SqCC. Importantly, these differentially altered genes may be indicative of disparate clinical outcomes depending on which subtype they are disrupted in. *CD9* was one of six genes that displayed opposite patterns of alteration (gained/overexpressed in SqCC and copy loss/underexpression AC) along with differential survival; high expression of this gene was correlated with favorable survival in SqCC, whereas low expression was associated with good survival in AC (Figure 2.7). Together, these results indicate that the genes involved in defining clinical characteristics are largely exclusive to individual NSCLC subtypes and influenced by the acquisition of distinct genetic alterations during tumor development, underscoring the importance of separating AC and SqCC when assessing genes involved in predicting patient prognosis and other clinical outcomes. This

information will become particularly important as targeted therapeutic strategies based around these genes develop. For example, since activated MEK1 and MEK2 phosphorylate and activate ERK (MAPK1) [111], the differential deregulation of *MAPK1* in AC (inactivated) and SqCC (activated) tumors may be an important consideration in determining the efficacy of MEK inhibitors in lung cancer subtypes.

The specific alterations selected during the development of each subtype may influence treatment outcomes and therefore play a role in clinical management. To demonstrate how our genomic findings can be used to define treatment strategies tailored to lung cancer subtypes and attempt to elucidate novel subtype specific treatment strategies, we performed CMAP analysis on our AC and SqCC specific gene signatures to identify compounds that can potentially reverse the expression of these genes. While the results for AC were uninformative, CMAP analysis of SqCC genes identified numerous HDAC and PI3K/mTOR inhibitors as compounds that could potentially induce a gene expression signature negatively correlated with that associated with SqCC. The HDAC inhibitor result was remarkable as the alteration of histone modifying enzymes was the most prominent network disrupted in this subtype, providing a biological basis for this finding. Furthermore, cancer cells with elevated activity of *E2F1* have been shown to be highly susceptible to HDAC inhibitor induced cell death and HDAC inhibitors such as SAHA have been shown to suppress the activity of *EZH2* [112, 113]. As *E2F1* and *EZH2* are both upregulated in SqCC (Figure 2.4b & c), this data suggests that treatment with HDAC inhibitors, in conjunction with standard chemotherapy, could be a promising avenue for disease treatment. In addition, since *PIK3CA* activation (mutation and/or amplification) is known to occur more frequently in SqCC the finding of multiple PI3K/mTOR inhibitors as potential therapeutics for SqCC is logical [95, 114]. Together, this data demonstrates the potential to use information about the underlying molecular biology of cancer subtypes to make informed decisions about clinical management strategies, and suggests that HDAC and PI3K/mTOR inhibitors, in combination with current treatment regimes, may provide a novel treatment tailored to lung SqCC.

2.5 Conclusions

Fundamental differences in tumor biology may be a primary factor determining the differential outcomes and response to therapies of lung cancer subtypes. A better understanding of the molecular mechanisms underlying subtype development is therefore essential to improving the poor prognosis of lung cancer. Our high-resolution integrative analysis of NSCLC genomes and epigenomes delineated novel tumor subtype-specific genetic and epigenetic alterations responsible for driving the differential pathogenesis and phenotypes of AC and SqCC. The specific genes and networks identified in this study provide essential starting points for elucidating mechanisms of tumor differentiation and developing tailored therapeutics for lung cancer treatment. More generally, our results confirm at the molecular level that these lung cancer subtypes are distinct disease entities and highlight how biological differences between AC and SqCC can influence patient outcome and response to therapy. When designing new treatment strategies and testing new drugs in clinical trials, these subtype differences as well as the biological pathways should be taken into account.

Chapter 3: Genetic disruption of KEAP1/CUL3 E3 ubiquitin ligase complex components is a key mechanisms of NF-κB pathway activation in lung cancer

3.1 Introduction

Our multi-dimensional integrative analysis of AC and SqCC genomes identified for the first time, subtype specific patterns of alteration of several previously established lung cancer associated genes, highlighting their newfound importance in subtype tumorigenesis. Among these genes was *Kelch-like ECH-associated protein 1* (*KEAP1*), a substrate adaptor protein that binds substrates to an E3-ubiquitin ligase complex comprised of *Cullin 3 (CUL3)* and *Ring box 1 (RBX1)* and was found to be preferentially lost in AC(Figure 3.1). The ubiquitin-proteasome pathway plays an essential role in maintaining normal cellular functions by controlling the abundance of several proteins and preventing undesired downstream effects. The most well characterized substrate of the KEAP1/CUL3 ubiquitin E3 ligase complex is *NRF2* [115-117]. In response to oxidative stress, *NRF2* stimulates transcription of cytoprotective genes that scavenge harmful reactive molecules, preventing cellular damage [118]. Interestingly, lung specific *Keap1* knockout in mice was shown to protect against cell damage caused by cigarette smoke by enabling *Nrf2* accumulation and increased expression of its target genes [119]. *NRF2* has also been implicated in cancer cell resistance to chemotherapeutics by its activation of drug-metabolizing and drug-efflux proteins [118].

NF-κB is a transcription factor that acts as a critical regulator of genes implicated in cell proliferation and survival, angiogenesis, epithelial to mesenchymal transition as well as inflammatory and immune responses [120-122]. The NF-κB pathway is activated in over 60% of lung cancers, however, the genetic mechanisms underlying its activation remain largely unknown [120, 123-126]. In the cytoplasm, NF-κB is bound by inhibitory proteins (I-kappaB), but upon stimulation, the kinase IKBKB phosphorylates IκB, releasing its inhibition, and enabling NF-κB translocation to the nucleus where it exerts its effects (Figure 3.1) [15]. Recently, Lee *et al.* implicated the KEAP1 E3-ligase complex in the regulation of NF-κB signaling, by demonstrating that KEAP1 binds IKBKB drawing it to the E3-ligase complex for ubiquitination and degradation [101]. As IKBKB is known to promote tumorigenicity through

phosphorylation-mediated inhibition of tumor suppressors and upregulation of NF-κB signaling, these findings further implicate the dysregulation of KEAP1 and the E3-ligase complex in NSCLC tumorigenesis and suggest disruption of this complex may underlie the high frequency of NF-κB activation in lung cancer.

Figure 3.1

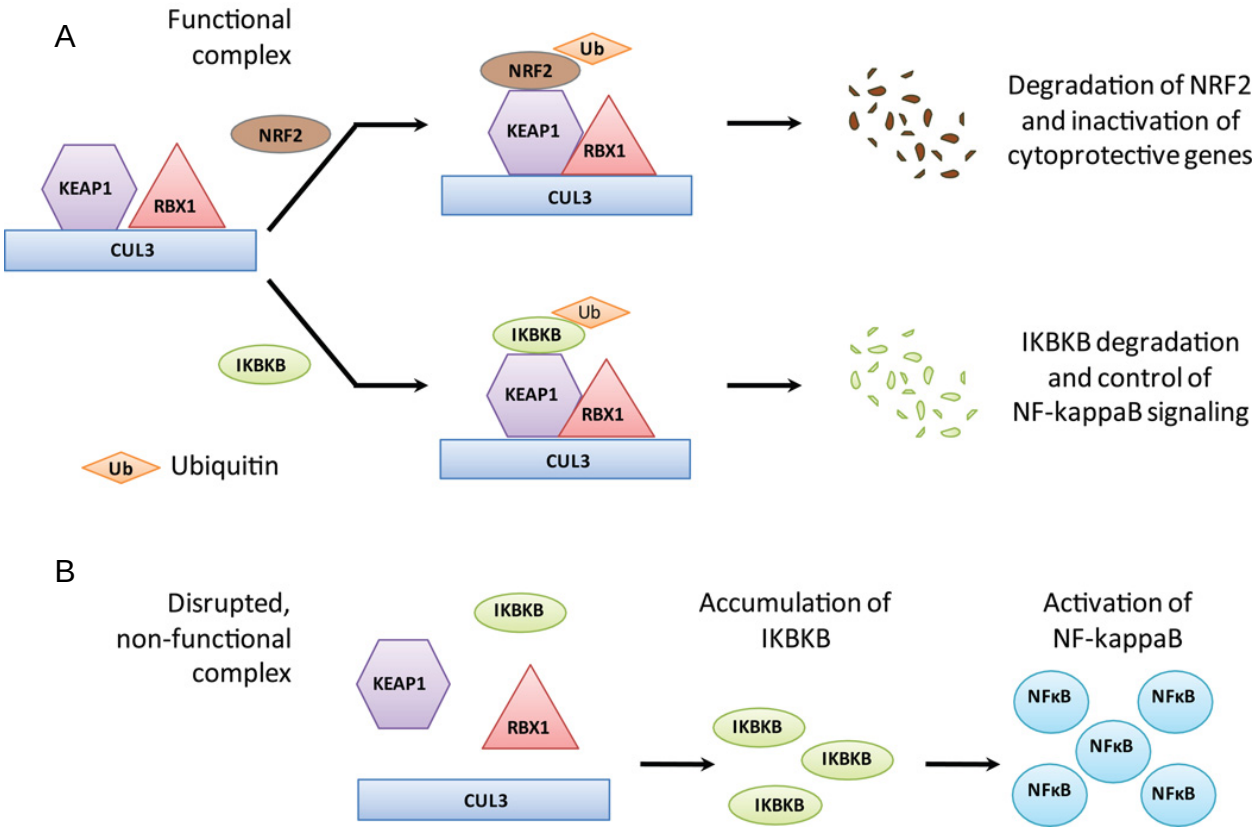


Figure 3.1 The role of the KEAP1/CUL3 ubiquitin ligase complex

(A) When complex components are intact, KEAP1 facilitates binding of NRF2 or IKBKB which promotes their ubiquitination. This complex prevents accumulation of IKBKB and subsequent NF-κB activation. (B) Disruption of any complex component compromises function leading to stabilization and accumulation of IKBKB, and aberrant activation of NF-κB.

Inactivating somatic mutations, loss of heterozygosity and hypermethylation of *KEAP1* have been reported at varying frequencies (3-41%) in lung tumors and cell lines, and low KEAP1 expression is associated with poor patient outcome [84, 127-132]. However, the moderate frequency of *KEAP1* gene disruption alone is not sufficient to explain the high (>60%) frequency of NF-κB activation observed in lung cancer [124]. The ability of the E3-ligase complex to ubiquitinate IKBKB was found to be most efficient when all three complex components were expressed and intact, suggesting disruption of even a single component compromises function. Although somatic DNA alterations have been observed in the genes encoding some of these complex components, the frequency of genetic and/or epigenetic disruption of complex components and whether complex component gene disruptions are a key mechanism of NF-κB activation in lung cancer is unknown [101]. Given the importance of the NF-κB pathway in lung cancer and the lack of inquiry into the role of the other complex components in lung cancer, we sought to determine the subtype specific patterns and biological effect of component disruption.

We hypothesize that somatic disruptions of *CUL3* (2q36.2) and *RBX1* (22q13.2), in addition to *KEAP1* (19p13.2) occur frequently in lung tumors, representing a prominent genetic mechanism that may be responsible for IKBKB accumulation and stimulation of NF-κB, and that genetic disruption of any one E3-ubiquitin ligase complex component is sufficient to result in tumorigenic NF-κB activation due to loss of complex function and subsequent accumulation of IKBKB [15, 101, 133]. In this study, we 1) investigated whether these complex components and *IKBKB* (8p11.21) exhibit gene dosage and expression alterations and the frequencies at which they occur in multiple independent tumor cohorts, 2) investigated whether the complex components display subtype specific disruption, and 3) assessed the functional consequence of complex disruption on NF-κB activity, as these genetic events may be significant contributors to the NF-κB activation commonly observed in lung cancer.

3.2 Methods

3.2.1 Non-small cell lung cancer (NSCLC) samples

261 lung tumors (169 adenocarcinomas (AC) and 92 squamous cell carcinomas (SqCC)) were accrued from Vancouver General Hospital (Vancouver) and Princess Margaret Hospital (Toronto) following ethics approval with patient consent (Chapter 2 and Sample Set #1, Appendix A1). Tissue sections were microdissected with the guidance of lung pathologists. Matched non-malignant lung tissue was also obtained for a subset of the primary tumors collected. DNA for all 261 samples was extracted using standard phenol-chloroform procedures. RNA was extracted from tumor and matched non-malignant tissues using RNeasy Mini Kits (Qiagen Inc.). NSCLC cell lines (H1650, HCC827, H3255, H358, H23, HCC95, H2347, and H2122) were obtained from ATCC or the laboratory of AFG and cultivated as previously described [134]. These cell lines were fingerprinted to confirm their identity [61]. Human bronchial epithelial cells (HBEC-KT) were provided by Dr. John Minna (University of Texas Southwestern Medical Center at Dallas) and maintained as previously described [98]. Primary non-malignant human bronchial epithelial (NHBE) lung cells were obtained from Lonza.

3.2.2 Determination of gene dosage and expression levels

Copy number status (gain, loss, or neutral) for the *KEAP1*, *CUL3*, *RBX1*, and *IKBKB* loci was determined for each tumor sample and 63 NSCLC cell lines by array comparative genomic hybridization (array CGH) using the whole genome tiling path array (SMRT v.2, BCCRC Array Laboratory, Vancouver, BC) as previously described in Chapter 2 and [76, 135, 136]. Gene expression levels of *KEAP1*, *CUL3*, *RBX1*, and *IKBKB* were determined using custom Agilent gene expression microarrays in 35 AC and 13 SqCC lung tumors and corresponding matched non-malignant tissues. Genes were classified as over or underexpressed if the fold change in mRNA expression levels in tumors relative to matched non-malignant tissues was greater or less than 2 fold. Gene expression for *KEAP1*, *CUL3*, *RBX1*, and *IKBKB* was also assessed in an additional, distinct cohort of 49 NSCLC (29 AC and 20 SqCC) tumors with matched CGH profiles using custom Affymetrix arrays to determine the contribution of copy number on gene expression (Sample Set #3, Appendix A1). The association between copy number and gene

expression was assessed by segregating these 49 NSCLC tumors into those with and without copy number alterations for each gene as previously described [76]. Expression levels were compared in both groups using a *U* test with a p-value less than 0.05 considered significant. The probe with the highest median intensity across all tumor samples was assessed for each gene.

3.2.3 Copy number analysis of external cohorts

Publically available NSCLC data was downloaded to further explore the frequency and subtype specificity of genomic disruption at the *KEAP1*, *CUL3*, *RBX1*, and *IKBKB* loci. Affymetrix SNP 250K data for 383 matched tumor non-malignant NSCLC pairs were accessed from the dbGaP Genotypes and Phenotypes database (Study Accession: phs000144.v1.p1). Affymetrix SNP 6.0 array profiles were obtained for 232 NSCLC tumors and matched non-malignant tissue (155 SqCC, 77 AC, GSE25016) and 54 NSCLC cell lines, 27 of which overlapped with cell lines profiled on the SMRT array, from the Wellcome Trust Sanger Institute CGP Data Archive (sanger.ac.uk/genetics/CGP/Archive/). Copy number profiles were generated using Partek Genomics Suite software. For tumor profiles, the patient matched non-malignant sample was used as a baseline for defining copy number alterations, whereas SNP profiles derived from 72 cytogenetically normal HapMap individuals were used as a reference for the cell line data [137]. In total, 570 tumors and 90 NSCLC cell lines (combination of lines profiled on the SNP and SMRT array) were analyzed for copy number alterations. The Broad Institute's Tumorscape database (www.broadinstitute.org/tumorscape) was also accessed to investigate copy number status at these four gene loci [138].

3.2.4 Multi-dimensional analysis of complex component disruption in the TCGA

Level 3 Affymetrix SNP6 copy number, HM27K DNA methylation and whole exome sequencing data was downloaded for all available AC and SqCC tumors. Analysis of DNA methylation data was limited to those cases with profiles of the matched non-malignant tissue (54 SqCC and 48 AC). Genes were identified as aberrantly methylated if the delta beta value (normal Beta value subtracted from the tumor Beta value) was greater than ± 0.15.

3.2.5 Western blot assessment of total and phospho- IKBKB and NF-κB protein levels

IKBKB protein levels were assessed in 8 NSCLC cell lines with various combinations of *KEAP1*, *CUL3*, or *RBX1* genomic loss or *IKBKB* genomic gain using antibodies from Cell Signaling (IKBKB #2678, p-IKBKB #2697, NF-κB: #4764, p-NF-κB: #3033, and GAPDH #2118). Western blots were performed following standard procedures as previously described [98]. Cells were washed with cold PBS and lysed in RIPA buffer with complete protease inhibitor cocktail (Roche, Basel Switzerland). Protein lysates were quantified using the BCA assay (Fisher Scientific). Lysates were diluted and boiled for electrophoresis then transferred to a polyvinylidene membrane. Membranes were blocked in 5% skim milk or 5% BSA in Tris buffered saline containing Tween 20 (TBS-T) (according to the manufacturer's instructions) and then incubated with primary antibody (1:1000) at 4°C overnight. Following three washes in TBS-T, membranes were incubated with HRP conjugated secondary antibody (Cell Signaling, cat. #7074, 1:20000) for 1 hour at room temperature. Antibody binding was visualized by enhanced chemiluminescence (Thermo Scientific) after three washes in TBS-T.

3.2.6 Immunohistochemistry (IHC) staining for IKBKB protein levels

5 µm thick sections were cut from 13 formalin fixed, paraffin embedded tumor specimens with various states of genomic disruption to the E3-ubiquitin ligase complex genes or *IKBKB*,based on copy number profiles. IHC to determine protein expression of IKBKB was performed as previously described [98]. Briefly, slides were deparaffinized in xylene and rehydrated with graded ethanol washes. Antigen retrieval was performed using a decloaking chamber with sodium citrate buffer pH 6.0, after which endogenous peroxidase activity was blocked using 3% H_2O_2 for 30 minutes at room temperature. Sections were blocked with goat serum for 3 hours at room temperature and then incubated overnight at 4°C with 32 µg/ml of anti-IKBKB mouse monoclonal primary antibody (EMD4 Biosciences, cat. OP134, San Diego, CA, USA). Prior to incubation with an anti-mouse HRP-streptavidin conjugated secondary antibody (DAKO, cat. K4000), four five minute washes in TBS-T were performed to remove unbound primary antibody. Detection of antibody binding was assessed using diaminobenzidine (Sigma Aldrich, cat. D4293). Slides were counterstained with hematoxylin for visualization. Intensity of staining was scored using a 0-3+ system based on the consensus of 3 observers (KT, LP, JCE). The mean

staining intensity for each tumor section was judged as follows: 0 - no staining, 1 - weak intensity, 2 - moderate intensity, 3 - strong staining.

3.2.7 siRNA-mediated complex component knockdowns

On-Target plus SMART pool siRNAs targeting *KEAP1, RBX1, CUL3*, and a non-targeting control pool of siRNAs were purchased from Thermo Scientific. One day prior to transfection, HBEC-KT cells were plated in regular growth media (antibiotic free) at a density of 200,000 cells/well in six well plates. Transfections were performed using an siRNA concentration of 100 nm according to the Thermo Scientific DharmaFECT siRNA transfection protocol. DharamaFECT 1 transfection reagent (Dharmacon) was used at a concentration of 0.2 µl/100 µl media. After 12 hours, transfection media was replaced with regular growth media. RNA was harvested from cells at 48 hours post transfection using the Trizol method (Invitrogen) and protein lysates were prepared using RIPA buffer 72 hours post transfection. Knockdown efficiencies were measured by qPCR with the following TaqMan assays from Applied Biosystems and using 18S rRNA as an endogenous control: Hs99999901_s1 (18S), Hs00202227_m1 (*KEAP*1), Hs00180183_m1 (*CUL3*), and Hs00360274_m1 (*RBX1*). Western blots were performed as above to measure total and phospho-IKBKB and NF-κB protein levels.

3.2.8 NF-κB target gene analysis

Expression of nine genes transcriptionally controlled by NF-κB (*CCND1, CXCR4, MMP2, TRAF1, MMP9, BCL2L11, CXCL13, PTGS2* and *CXCL12*) - as annotated in the *Ingenuity Pathway Analysis* database (Ingenuity® Systems) were analyzed in 48 NSCLC tumors with deregulated *KEAP1, CUL3, RBX1* or *IKBKB* expression levels (2-fold or greater). Wilcoxon signed-rank tests were used to compare NF-κB target gene expression in tumors versus their matched non-malignant lung tissues. NF-κB target genes were considered upregulated if expression levels were significantly elevated in tumors relative to matched non-malignant tissues (Wilcoxon $p < 0.05$).

3.2.9 IKBKB inhibition in NSCLC cell lines

Cell viability assays were performed to measure the effect of IKBKB inhibition by a cell permeable, competitive ATP inhibitor, IKK-2 inhibitor IV (Calbiochem cat. #401481) on five NSCLC cell lines (H3255, H2122, H23, HCC95 and H1650). Cells were plated in triplicate in 96 well plates at optimal densities for growth (H2122 and H23 at 2000 cells/well, H1650 and HCC95 at 3000 cells/well, and H3255 at 5000 cells/well) and subjected to a series of 2-fold dilutions of IKBKB inhibitor prepared in cell growth media and DMSO. The experimental inhibitor concentrations ranged from 500 µM to 244 nM and the final DMSO concentration for treated and untreated (control) cells was 1%. Blank wells contained equal volumes of growth media with 1% DMSO. Cells were incubated for 72 hours at 37°C and then treated with 10µl of Alamar Blue cell viability reagent (Invitrogen) according to manufacturer's instructions. The reaction product was quantified by measuring absorbance at 570 nm with reference to 600 nm using an EMax plate reader (Molecular Devices). The average absorbance readings for blank wells were subtracted from all treatment and control wells and technical replicates were averaged. The response of treated cells was measured as a proportion of the viability of untreated cells, with the mean background subtracted treatment absorbance divided by the mean background subtracted untreated absorbance for each inhibitor concentration. Dose response curves and IC_{50} values were generated in Graph Pad v5 using the proportionate response of all 12 drug concentrations. Experiments were repeated in triplicate and differences in IC_{50} values were determined using a student's t-test with a p-value < 0.05 considered significant.

3.3 Results

3.3.1 Disruption of E3-ubiquitin ligase complex components in NSCLC

261 tumors were screened for DNA copy number alterations at the *KEAP1*, *CUL3*, and *RBX1* loci. A significant proportion of tumor samples (103 of 261, 39%) showed genomic loss in at least one of the complex associated genes (Figure 3.2A). *KEAP1* was the most frequently disrupted complex component, undergoing genomic loss in 23% of lung tumors analyzed (Figure 3.2A). Strikingly, 69% (71 of 103) of the tumors harboring copy number alterations had alterations affecting only one of the genes assessed. Gene expression analysis revealed aberrant expression in at least one of the complex component genes in 40% of tumors analyzed (19 of 48) and of those with aberrant expression, 84% (16 of 19) had only one of the genes affected (Figure 3.2A). At the expression level, *RBX1* was the most frequently altered complex component, underexpressed in 21% of tumors analyzed (Figure 3.2A). We also detected a high frequency of *IKBKB* DNA copy number gain and mRNA overexpression (23% and 35% of tumors, respectively) (Figure 3.2A). When *IKBKB* status was taken into account, the frequency of NSCLC disruption at any of the *KEAP1*, *CUL3*, *RBX1* and/or *IKBKB* loci rose to 54% (141 of 261 tumors, Figure 3.2A) at the gene copy number level and 63% (30 of 48 tumors, Figure 3.2A) at the expression level. Even with the inclusion of *IKBKB*, the majority of tumors exhibited complex or *IKBKB* genetic disruption at only one gene locus (Figure 3.2B and 3.2C). Frequent genetic disruption of E3-ubiquitin ligase complex components or *IKBKB* (73%) was also evident in NSCLC cell lines (Table 3.1).

To determine whether complex component and IKBKB gene dosage alterations are regulating gene expression, we integrated DNA copy number and gene expression data for the complex genes, and *IKBKB* in an additional set of 49 NSCLC tumors. *KEAP1*, *RBX1* and *CUL3*, expression was significantly lower in tumors with genomic loss compared to those without loss (*U* test, p=0.00076, p=0.00116, and p=0.00339, respectively, Figure 3.3A-C), whereas *IKBKB* gene expression was elevated in tumors with gain compared to those without (*U* test, p=0.0143, Figure 3.3D). These findings demonstrate that dosage alterations affect mRNA expression levels, and therefore, likely contribute to E3-ligase complex disruption.

Figure 3.2

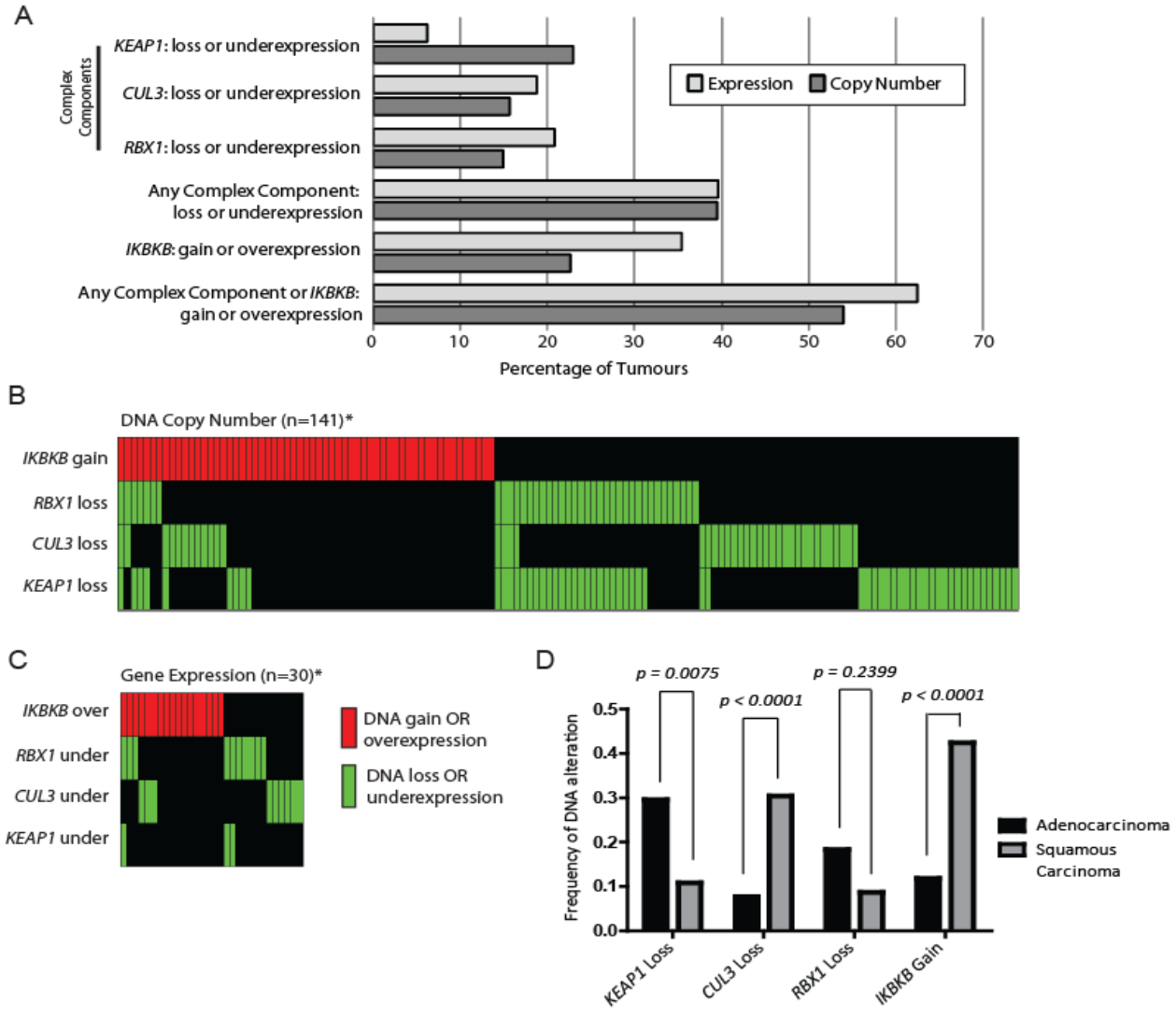


Figure 3.2 Frequent disruption of the KEAP1 E3-ligase complex and IKBKB in NSCLC.

(A) Summary of DNA copy number and gene expression alterations in NSCLC tumors. (B) Copy number analysis of 261 lung tumors revealed frequent loss of *KEAP1*, *RBX1*, and *CUL3*, as well as frequent gain of *IKBKB* (141/261, 54%). Vertical columns indicate individual tumor samples and only samples with ≥1 alterations are shown. (C) mRNA expression profiles for 48 lung tumors revealed frequent underexpression of complex components and overexpression of *IKBKB* (30/48, 62.5%). Expression was considered altered if tumor/matched non-malignant tissue was changed >2 fold. (D) *KEAP1*, *CUL3*, and *IKBKB* exhibit significant differences in copy number alteration patterns between AC and SqCC (Fisher's exact test, $p < 0.05$).

Table 3.1 Copy Number status of complex components in multiple independent cohorts

Gene	Cell Lines (n=90)	dbGAP (n=383)	GSE25016 (n=232)
KEAP1	39% loss	23% loss	18.5% loss
CUL3	17% loss	3% loss	18.5% loss
RBX1	39% loss	11% loss	10% loss
IKBKB	28% gain	11% gain	23% gain
Any	73%	34%	52%

Figure 3.3

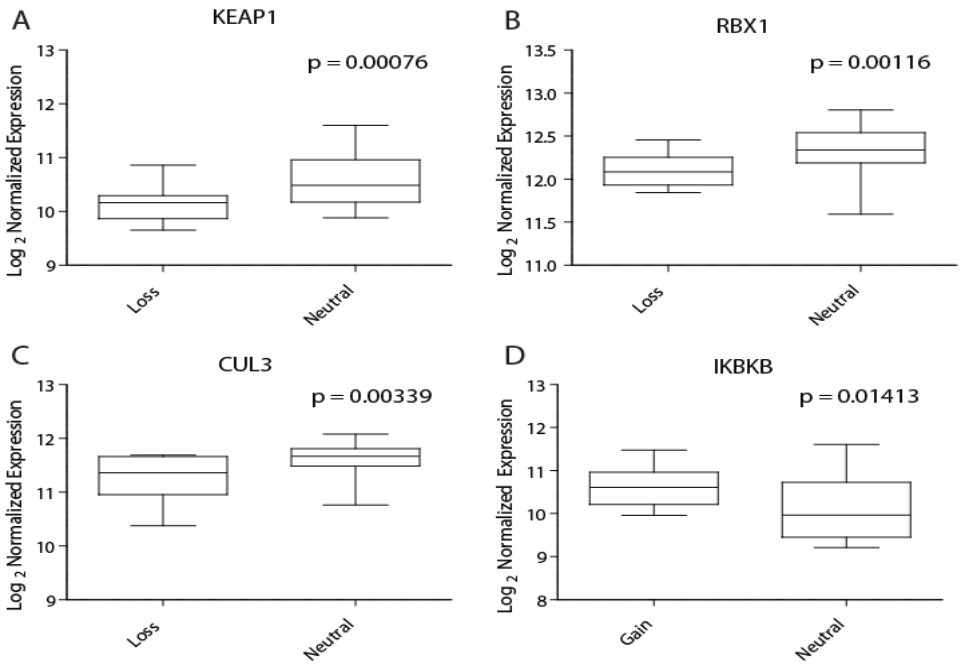


Figure 3.3. Association between gene expression and copy number in clinical lung tumors.
Box and whisker plots demonstrating *KEAP1* (A), *RBX1* (B), *CUL3* (C) and *IKBKB* (D) expression levels in tumors with and without genetic disruption. Tumors were grouped based on copy number status for each gene (gain, loss, neutral) and expression levels were compared between groups using a *U* test with a p-value < 0.05 considered significant. Whiskers show the min and max, while boxes illustrate the 25th, median and 75th percentile.

In addition to the tumor data generated using array comparative genomic hybridization, we analyzed copy number profiles derived from publically available SNP array data from the dbGaP Genotypes and Phenotypes database (n=383) and GSE25016 (n=232) both of which had matched non-malignant tissue for all cases, and a panel of 90 NSCLC cell lines. Consistent with our findings, frequent genomic disruption to the E3-ubiquitin ligase complex components and *IKBKB* was observed in all external cohorts (Table 3.1). 73% of cell lines harboured DNA alterations encompassing at least one complex component or *IKBKB*, whereas 34% and 52% of tumors had disruption. We further interrogated the copy number status of *KEAP1, CUL3, RBX1*, and *IKBKB* genes in the Broad Institute's Tumorscape database [138]. This revealed that *CUL3* was significantly deleted in 12% of all 3131 tumors in the database and in 13% percent of all NSCLC specimens (n=733), of which 5.4% had focal *CUL3* deletions. Similarly, *IKBKB* was significantly amplified in 22% of the all tumors and 28% of NSCLC specimens, of which 12% contained focal *IKBKB* DNA amplifications. *KEAP1* and *RBX1* were not significantly deleted in the Tumorscape database. Together, these results support our observation that KEAP1 complex components and IKBKB undergo frequent copy number alterations in NSCLC.

3.3.2 Complex components are differentially altered in AC and SqCC subtypes

As distinct patterns of DNA alterations exist for AC and SqCC, we sought to determine whether complex component disruption displays subtype specific patterns of alteration. While both subtypes of NSCLC showed high frequency of complex component and *IKBKB* gene disruption, the pattern of gene disruption differed between these subtypes (Figure 3.2D and Table 3.2). *KEAP1* loss appears to be the main mechanism of complex disruption in lung AC, accounting for 64% of cases with complex disruption, and is more prevalent in AC than in the SqCC subtype (Fisher's exact test, $p = 0.0075$, Table 3.2). In contrast, *CUL3* loss and *IKBKB* gain occurred more often in the SqCC than AC subtype of lung tumors (Fisher's exact test, $p = 4.926 \times 10^{-6}$ and $p = 8.446 \times 10^{-7}$, respectively, Table 3.2). Gain of *IKBKB* was the most frequent alteration in SqCC, occurring in 57% of cases, followed closely by *CUL3* loss in 44% of cases with complex component disruption.

To confirm these subtype specific findings, we interrogated complex component and *IKBKB* copy number in The Cancer Genome Atlas (TCGA) and GSE25016 cohorts, both of which have CN data for AC and SqCC. The same trends in component disruption were observed across all three datasets; with *KEAP1* and *RBX1* loss being more prevalent in AC and *CUL3* loss and *IKBKB* gain more frequent in SqCC (Table 3.2). Statistically significant differences between the frequency of component alteration was observed for all components in the GSE25016 dataset and for *KEAP1* and *CUL3* in the TCGA (Table 3.2). Analysis of external cohorts corroborated the subtype specific patterns of alteration we observed and demonstrate for the first time that the mechanisms of KEAP1/CUL3 complex disruption are subtype specific.

Table 3.2 Frequency of KEAP1 complex copy number alterations in AC and SqCC

	Frequency in AC			Frequency in SqCC			Fishers Exact Test		
Gene	**BCCRC**	**TCGA**	**GSE**	**BCCRC**	**TCGA**	**GSE**	**BCCRC**	**TCGA**	**GSE**
KEAP1	50/169 (29.6%)	48/277 (17.3%)	27/77 (35.1%)	10/92 (10.9%)	20/201 (10%)	16/155 (10.3%)	0.0075	0.02427	$1.1x10^{-5}$
CUL3	13/169 (7.7%)	4/277 (1.4%)	6/77 (7.8%)	28/92 (30.4%)	27/201 (13.4%)	37/155 (24%)	$4.93x10^{-6}$	$1.36x10^{-7}$	0.00233
RBX1	31/169 (18.3%)	29/277 (10.5%)	14/77 (18.2%)	8/92 (8.7%)	11/201 (5.5%)	9/155 (5.8%)	0.24	0.06495	0.0047
IKBKB	20/169 (11.8%)	44/277 (16.2%)	11/77 (14.3%)	39/92 (42.4%)	46/201 (22.9%)	42/155 (27.1%)	$8.45x10^{-7}$	0.05248	0.0312
ANY	78/169 (45.6%)	99/277 (35.7%)	41/77 (53.2%)	64/92 (69.6%)	78/201 (38.8%)	79/155 (51%)			

3.3.3 Other genetic and epigenetic mechanisms of complex component disruption

Given our hypothesis, we focused on measuring gene dosage alterations that could account for disruption of the E3-ubiquitin ligase complex and its downstream consequences. However, as mutations and hypermethylation of *KEAP1* have been described in lung cancer and these events are known to downregulate *KEAP1* expression [84, 128, 129, 132], we next sought to determine whether *IKBKB* and complex components were frequently altered by other genetic or epigenetic mechanisms. To assess the contribution of DNA methylation and gene mutation to complex disruption, we analyzed 408 tumors (230 AC and 178 SqCC) for which mutation data

was available and 102 tumors with matched non-malignant tissue (54 SqCC and 48 AC) and HM27K methylation profiling from the TCGA. *KEAP1* was the most frequently mutated component, mutated in 17% (40/230) of AC and 12% (22/178) of SqCC. All other complex components were infrequently mutated; 3.4% for *CUL3*, 0.5% for *RBX1* and 2.2% for *IKBKB*. Interestingly, mutations in *CUL3* were significantly more frequent in SqCC (6% vs. 1.5%, Fisher's Test p=0.011), further supporting the notion that CUL3 is preferentially disrupted in SqCC. None of the complex components or *IKBKB* were hypo- or hypermethylated in any of the samples assessed. The rarity of aberrant methylation across all complex components and the lack of mutations in *CUL3*, *RBX1* and *IKBKB* further supports the notion that copy number is the primary genetic mechanism through which these genes are deregulated. The high frequency of *KEAP1* mutations, and mutual exclusivity with copy loss underscores the selective inactivation of *KEAP1*and highlights its importance in lung tumor biology.

3.3.4 Functional consequences of genetic complex disruption

Since IKBKB protein levels would be directly affected by E3-complex function, we assessed IKBKB protein levels in NSCLC cell lines with and without complex disruption to measure the consequence of complex disruption. Immunoblotting for IKBKB in non-malignant human bronchial epithelial (NHBE) cells and a panel of 8 NSCLC cell lines revealed elevated expression levels in lines with genomic loss of *KEAP1, CUL3, RBX1,* or gain of *IKBKB* (Figure 3.4A). No correlation between the number of disrupted components and the levels of IKBKB were observed. To directly assess the effect of complex component integrity on IKBKB and NF-κB activity, we performed siRNA mediated knockdowns of *KEAP1, CUL3* and *RBX1* in non-malignant, bronchial epithelial cells. We achieved at least 80% knockdown for all three genes and observed an increase in phospho-IKBKB and phospho- NF-κB levels in the knockdowns relative to the non-targeting control, providing evidence that complex disruption directly regulates NF-κB signaling (Figure 3.4B).

Figure 3.4.

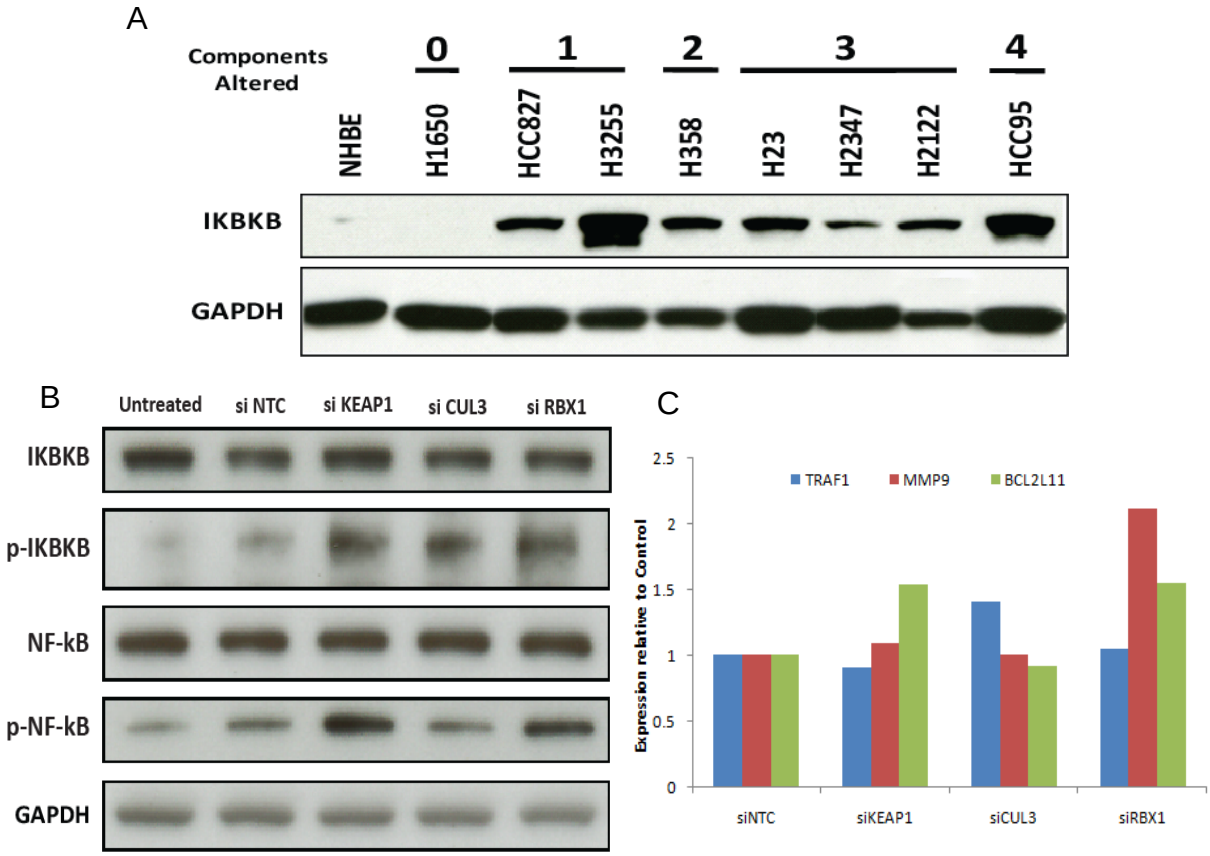


Figure 3.4. IKBKB protein expression in NSCLC.

(A) Western blot depicting IKBKB protein expression in eight cell lines with varying degrees of genetic disruption to KEAP1 E3-ligase complex components and/or *IKBKB* (as determined from copy number profiles of 90 NSCLC cell lines). NHBE is a non-malignant human bronchial epithelial lung line that provides a baseline for IKBKB expression. The number of disrupted genes for each line is indicated above the cell line. (B) Western blot depicting the effects of transient siRNA knockdown of *KEAP1, RBX1,* and *CUL3*, on total and phospho- IKBKB and NF-κB protein levels. (C) qRT-PCR analysis of three NFkB transcriptional targets upon siRNA knockdown of individual complex components.

In an attempt to determine whether genetic disruption of complex components affects the relative protein levels of IKBKB, we performed immunohistochemistry (IHC) on a panel of 13 lung tumors with various states of complex component disruption. Staining for IKBKB revealed protein expression in both complex disrupted and undisrupted tumors; however, the vast inter- and intratumor heterogeneity in staining and potential non-specific antibody binding limited our ability to identify significant correlations between IKBKB protein levels and E3-complex genetic disruption (data not shown, available as Supplemental Digital Content online at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3164321/).

As we were unable to discriminate differences in IKBKB protein levels between complex disrupted and non disrupted tumors by IHC, we sought to determine whether loss and underexpression of E3-ligase complex components activates NF-κB signaling. Expression of nine well characterized transcriptional targets of NF-κB were analyzed in the 48 NSCLC tumor pairs for which complex component disruption could be determined. Of the nine genes, four - *BCL2L11*, *CXCL13*, *MMP9*, and *TRAF1* were significantly up-regulated in tumors with underexpression of a complex component or overexpression of *IKBKB* as compared to their matched non-malignant lung tissue (Figure 3.5). Assessment of the same NF-κB target genes following siRNA knockdown confirmed these genes to be overexpressed relative to non-targeting controls (Figure 3.4C), supporting our hypothesis that complex disruption leads to activation of NF-κB signaling.

Figure 3.5

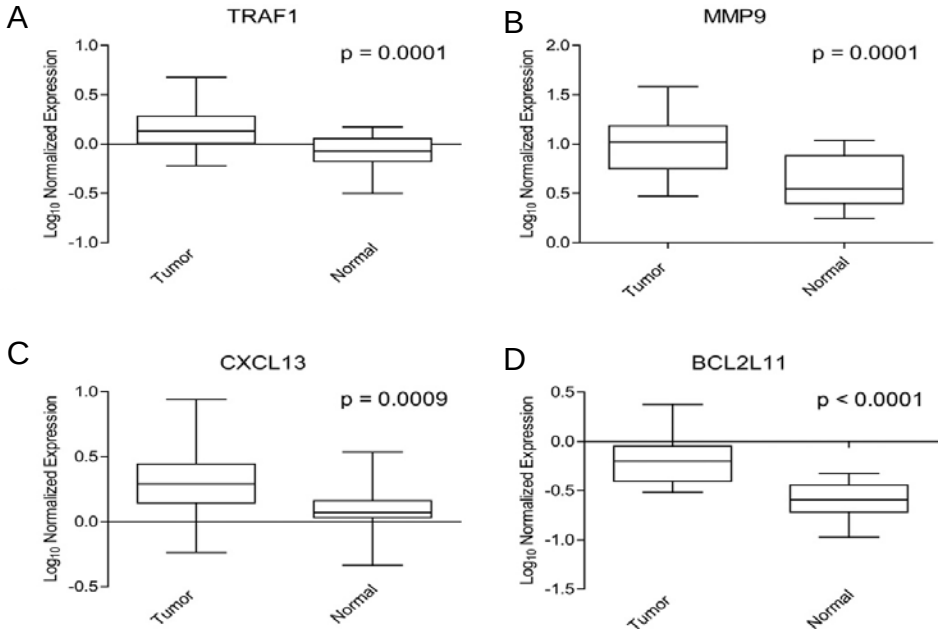


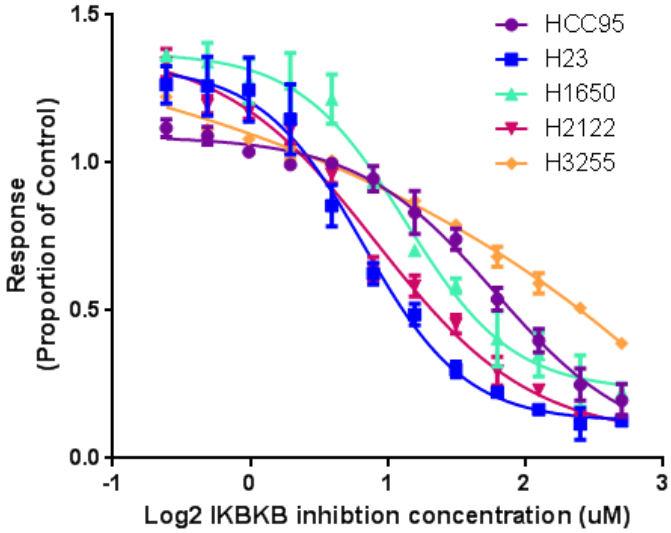
Figure 3.5. Activation of NF-κB targets in tumors with complex components disruption
(A-D) Box plots demonstrating four examples of NF-κB target genes (*BCL2L11*, *CXCL13*, *MMP9*, *TRAF1*) that are significantly upregulated in tumors with complex component disruption, relative to matched non-malignant lung tissues (Wilcoxon sign rank test, $p \leq 0.001$). Copy number status of the NF-κB target gene loci was determined in these tumors to ensure that overexpression was not due to dosage changes of the target genes. Whiskers show the min and max values, while boxes illustrate the 25th, median and 75th percentile.

3.3.5 Pharmacological inhibition of IKBKB

Overexpression of IKBKB can contribute to a malignant phenotype through activation of the NF-κB signaling pathway and consequential effects on cell growth. We hypothesized cells dependent on this pathway would be more sensitive to IKBKB inhibition than those without complex alterations and normal IKBKB protein levels. IKBKB inhibition experiments revealed H1650, an AC line without genomic loss of any complex component, had reduced sensitivity (as measured by cell viability) to IKBKB inhibition (IC_{50} 13.41) than H2122 and H23 (IC_{50} 8.19 and 6.59, respectively) which have multiple components altered (student's t-test, $p < 0.05$) (Figure

3.6). H3255 and HCC95 showed insensitivity to inhibition compared to H1650 (student's t-test, $p < 0.05$), however these two cell lines have the highest IKBKB protein expression levels of the cell lines we tested, which likely contributed to their relative resistance to this competitive inhibitor. Replicate experiments were highly reproducible, with similar trends in sensitivity observed across all replicates.

Figure 3.6



Cell Line	Copy Number Alteration status	IKBKB protein levels	IC50
H1650	No alterations	-	13.4
H2122	3 components	+	8.19
H23	3 components	+	6.59
H3255	1 component	+++	1.27x10^18
HCC95	4 components	++	60.02

Figure 3.6. Pharmacological inhibition of IKBKB in NSCLC cell lines

Dose response assays were performed to measure the effect of IKBKB inhibition on viability of 5 NSCLC cell lines. H2122, H23, H3255, and HCC95 harboured genetic alterations to either *IKBKB* and/or one or more complex components while H1650 was not altered at the DNA level. Error bars represent the standard error of the mean of replicate experiments. IKBKB protein expression as measured by western blot and the IC50s of each cell line are shown in the table.

3.4 Discussion

The NF-κB pathway is aberrantly activated in the majority of lung cancers and is essential in *KRAS* driven mouse models of lung tumorigenesis [123-126]. NF-κB signaling contributes to tumorigenesis via its promotion of cell proliferation and survival [120-122]. In order for NF-κB to become active, inhibition by IκB must be released. This is achieved through phosphorylation of IκB by the kinase, IKBKB [139]; hence, IKBKB has a critical role in NF-κB activation [101, 139]. In fact, constitutive IKBKB activity has been postulated to drive the aberrant NF-κB activation observed in cancer [121]. Despite what is known about the cascade of protein signaling events that result in NF-κB activation, the genetic mechanisms responsible for the aberrant activation of NF-κB signaling in lung cancer are not well understood. In this study, we hypothesized that genetic disruption and loss of function of the KEAP1 E3-ubiquitin ligase complex, which regulates IKBKB protein levels, is a major mechanism of IKBKB accumulation and consequential NF-κB activation in lung cancer.

These results provide evidence that somatic E3-ligase complex disruption is a prominent genetic mechanism of NF-κB activation in lung cancer that compromises the ability of cells to degrade the NF-κB activator, IKBKB. We discovered a remarkably high frequency of both genetic disruption and gene expression changes for the genes encoding E3-ligase protein components (*KEAP1*, *RBX1*, and *CUL3*) as well as the gene encoding the complex's oncogenic substrate, *IKBKB*. We found that genetic disruption of the complex genes alters mRNA expression and results in elevated IKBKB protein levels, demonstrating the consequence of complex disruption. Moreover, we demonstrated evidence of NF-κB activation in complex compromised lung tumors and showed the importance of IKBKB protein expression in driving the lung cancer phenotype.

Although genetic and epigenetic disruption of *KEAP1* has been reported in lung cancer before, to our knowledge, this is the first study to comprehensively characterize somatic gene dosage alterations to the *CUL3*, *RBX1*, and *IKBKB* loci in a large cohort of clinical lung tumors. The strikingly high frequency of copy number and gene expression alterations observed in our study highlights the importance of these E3-ubiquitin ligase complex components and also

IKBKB in lung cancer. The recurrent nature of DNA copy number alterations at the complex component loci and their effects on gene expression are strong evidence that these genes are targeted for dosage alterations as opposed to passengers of alterations targeting other genes. In addition, the high proportion of disrupted lung tumors observed to have genetic alterations affecting a single component only, at both the copy number (67%, Figure 3.2B) and gene expression levels (73%, Figure 3.2C), suggests that disruption of only a single complex component is sufficient to compromise complex function and promote NF-κB signaling through abnormal IKBKB accumulation. Interestingly, we observed differential complex component disruption patterns in AC and SqCC subtypes (Figure 3.2D). Although E3-ubiquitin ligase complex disruption occurs in both subtypes, the differences in the component genes preferentially altered suggests that complex disruption is achieved by different means.

Examination of IKBKB protein levels in NSCLC cell lines revealed high expression in lines harboring genetic disruption to at least one complex component or IKBKB, whereas the non-malignant lung line (NHBE) and a line without genetic disruption (H1650) showed very low or undetectable levels. The E3-ligase complex was considered to be genetically intact in H1650 as neither underexpression of *KEAP1, RBX1*, and *CUL3,* or overexpression of *IKBKB* relative to NHBE cells was observed. This suggests there are no genetic or epigenetic alterations affecting the complex components or *IKBKB* in this cell line and the observed IKBKB levels were consistent with H1650 having a functioning E3-ligase complex, supporting our hypothesis. Therefore, in addition to affecting gene expression, copy number losses of the loci coding for complex components and gains of *IKBKB* appear to influence IKBKB protein expression. A trend towards higher IKBKB expression in lines with more complex components/*IKBKB* alterations was not evident, suggesting genetic disruption of a single component is sufficient to result in loss of complex function and IKBKB accumulation (Figure 3.4A). This finding is consistent with the observation that the majority of tumors exhibiting complex disruption have only one complex component altered, further supporting the idea that single component disruption is sufficient to produce an oncogenic effect.

We have conclusively demonstrated elevated IKBKB protein expression in NSCLC cell lines with complex disruption, however, measuring this effect directly in tumor tissue sections was not a straightforward task due to the extent of heterogeneity in tumor staining intensity across and within individual tumors which likely reflects the heterogeneous nature of lung tumor specimens. Due to this innate tumor heterogeneity, unlike cell lines, we were unable to conclude whether or not there was a significant correlation between E3-ligase complex disruption and IKBKB protein levels *in vivo*. Examples of tumor staining are available in Supplemental Digital Content 6, available online at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3164321/.

To investigate the direct consequence of complex disruption on IKBKB accumulation and NF-κB activity, we performed siRNA knockdowns on the individual complex coding genes (*KEAP1*, *RBX1*, and *CUL3)*. Consistent with our hypothesis, we observed elevated levels of activated IKBKB and NF-κB as well as elevated expression of NF-κB target genes upon complex component disruption in HBEC cells (Figure 3.4B and C). Since phospho-NF-κB is an indicator of active NF-κB signaling, these results clearly illustrate the functional consequence of E3-ligase complex disruption. Our work provides evidence to support the hypothesis that genetic loss of the complex component encoding genes causes downregulation in their expression, and that loss of expression of these genes results in increased levels of activated IKBKB and NF-κB.

A number of reports have detailed the critical role of IKBKB protein in driving NF-κB activation, and the importance of IKBKB to cancer cell viability is emphasized by the development of IKBKB inhibitors as a strategy for tempering NF-κB signaling [121, 133, 139]. We found that IKBKB inhibition reduced NSCLC cell viability and that cells without complex or *IKBKB* disruption, which we hypothesized to be less dependent on IKBKB expression for growth, were indeed more insensitive to IKBKB inhibition, as were cells with high endogenous levels of IKBKB protein (Figure 3.6). In addition to cell experiments to verify the importance of E3-ligase complex disruption in lung cancer, we analyzed the expression levels of several NF-κB target genes in tumors with complex disruption to measure its effect on NF-κB activity. Despite the possibility that other mechanisms could also contribute to the transcription of the NF-κB target genes, we observed a significant increase in the expression of NF-κB target genes in

complex compromised tumors (Figure 3.5). Together, these findings support our hypotheses and demonstrate the biological significance of complex disruption and subsequent IKBKB overexpression in lung cancer biology.

3.5 Conclusions

Our analyses have revealed remarkably frequent genetic disruption and aberrant expression of not only *KEAP1*, but all members of the KEAP1 E3-ubiquitin ligase complex and *IKBKB* in lung cancer. For the first time we show that AC and SqCC acquire copy number alterations to different components of the E3-complex or *IKBKB* which suggests the genetic mechanisms of complex disruption that promote NF-κB activation may be subtype specific. Moreover, we have shown that IKBKB protein expression is elevated in NSCLCs with genetic loss of *KEAP1*, *CUL3* or *RBX1* or gain of *IKBKB*, and that knockdown of complex components leads to an accumulation of active IKBKB and NF-κB, demonstrating the functional consequence and significance of complex disruption. We have also provided evidence of NF-κB activity, a downstream effect of IKBKB accumulation, in complex disrupted tumors and cell lines. Collectively, our findings suggest that prominent genetic disruption to the E3-ubiquitin ligase complex and its oncogenic substrate, *IKBKB*, play a major role in driving the aberrant NF-κB activation that is characteristic of lung tumorigenesis but that this activation occurs via different mechanisms depending on the subtype.

Chapter 4: YEATS4 is a novel oncogene amplified in non-small cell lung cancer that regulates the p53 pathway

4.1 Introduction

Within the last decade, characterization of lung cancer genomes has revealed a number of genes critical to tumorigenesis, resulting in significant changes to lung cancer treatment and a subsequent increase in progression free and overall survival for a subset of these patients. These successes have prompted a search for additional driver alterations, and have identified a number of recurrently mutated genes including *TP53,EGFR, CDKN2A, PTEN, NRAS, BRAF, PIK3CA, DDR2, KEAP1*and *NRF2* as well as gene fusions encompassing *RET* and *ROS* tyrosine kinases [15, 55, 140-142]. Despite these discoveries, approximately 50% of lung cancers harbor no known targetable alterations, highlighting the need for a better understanding of the biology underlying lung tumorigenesis [55, 142].

In addition to somatic mutations, copy number alterations such as recurrent amplifications and deletions occur in almost all lung cancers [18, 76]. DNA amplification directly contributes to oncogene activation and the promotion of tumorigenesis, particularly for tumors driven by oncogene addiction. Oncogenes amplified at the DNA level therefore make ideal therapeutic targets as unlike loss of function tumor suppressor genes (TSG), they have the potential to be targeted directly. In NSCLC, recurrent amplifications of several regions activate known oncogenes. These include; 1q21.2 (*ARNT*), 3q26.3-q27 (*PIK3CA & SOX2*), 5p15.33 (*TERT*), 7p11.2 (*EGFR*), 7q31.1(*MET*), 8p12 (*FGFR1*) 8q24.21 (*MYC*), 12q14.1 (*CDK4*), 14q13.3 (*NKX2-1*) [18, 70, 95, 122, 143-145]. In chapter 2, numerous regions of frequent amplification were identified, some of which were highly specific to a subtype like 3q and others such as 12q and 20q that were frequently altered in both AC and SqCC. In an attempt to identify novel oncogenes involved in NSCLC tumorigenesis, we integrated DNA copy number and gene expression data in order to identify candidate driver genes within highly recurrent amplicons.

Our approach was based on the rationale that oncogenes selectively amplified and biologically relevant to NSCLC tumor biology would: i) span regions of frequent high level amplification, ii) undergo frequent overexpression and iii) exert pro-tumorigenic functions *in vitro* and *in vivo*. Our analysis identified a recurrent amplicon at 12q15, within which we identified the candidate oncogene *YEATS4/GAS41* (*YEATS domain containing 4, glioma-amplified sequence 41*). *In vivo* and *in vitro* functional assays were performed to characterize the biologic effects and investigate the oncogenic mechanism of YEATS4 in lung tumorigenesis. Based on the frequency of *YEATS4* amplification and overexpression in NSCLC tumors and cell lines, its role in viability, anchorage independent growth, senescence and tumor formation, we propose that *YEATS4* is novel candidate oncogene in lung cancer.

4.2 Methods

4.2.1 NSCLC tumor samples and cell lines

261 formalin-fixed paraffin embedded and fresh-frozen lung tumors (169 AC and 92 SqCC) were obtained under informed, written consent with approval from the University of British Columbia-BC Cancer Research and University of Toronto Ethics Board from patients undergoing surgical resection at the Vancouver General Hospital and the Princess Margaret Hospital in Toronto [146]. Tissue sections were micro-dissected with the guidance of lung pathologists and matched non-malignant lung tissue obtained for a subset of the primary tumors. DNA was extracted using standard phenol-chloroform procedures. RNA was extracted from tumor and matched non-malignant normal tissue using RNeasy Mini Kits (Qiagen) or Trizol reagent (Invitrogen). Quality and quantity of genomic material was assessed using a NanoDrop 1000 spectrophotometer and by gel electrophoresis and/or by Agilent 2100 Bioanalyzer. Demographic information for this cohort is summarized in Appendix A2. NSCLC cell lines H1993, H1355, H226, A549 were obtained from American Type Culture Collection and HCC4011 from Dr. Adi Gazdar (University of Texas Southwestern Medical Center at Dallas) and fingerprinted to confirm their identity [147]. All lines were cultured in RPMI-1640 medium supplemented with 10% fetal bovine serum and 0.1% Penicillin-Streptomycin (Invitrogen).

Immortalized normal human bronchial epithelial cells (HBEC) with (HBEC-KT53) and without p53 knockdown (HBEC-KT), courtesy of Dr. John Minna, were cultured in K-SFM media supplemented with 50ng/µl BPE and 5 ng/µl EGF (Invitrogen). Demographic data for the panel of cell lines used in this study can be found at http://edrn.jpl.nasa.gov/ecas/data/dataset/urn:edrn:UTSW_MutationData

4.2.2 Array comparative genomic hybridization and GISTIC analysis

Copy number profiles were generated for 261 NSCLC tumors using whole-genome tiling path array comparative genomic hybridization (aCGH), and were processed as previously described in Chapter 2. Probes were mapped to the March 2006 (Hg18) genomic coordinates, X and Y chromosomes removed and aCGH-Smooth was used to segment and smooth log2 ratio values [78]. The corresponding segments and ratio values were analyzed using the GISTIC algorithm [148] and gene pattern software (http://www.broadinstitute.org/cancer/software/genepattern/) to identify regions of significant amplification across samples. Amplification threshold of 0.8, join segment size of 2, qv threshold 0.05 and removal of the X chromosome were the settings applied for analysis.

4.2.3 Gene expression profiling and data integration

Gene expression profiles were generated using custom Agilent microarrays for a subset (35 AC and 13SqCC) of the 261 tumors which had sufficient quantity and quality material for both tumor and matched non-malignant tissue. Data was normalized using the Robust Multiarray Average algorithm in R [80]. Genes were classified as over- or underexpressed if the mRNA fold change in tumors relative to matched non-malignant was greater or less than 2-fold. Mann-Whitney U tests with Benjamini Hochberg correction $p<0.05$ were used to compare expression of 12q15 genes between tumor and non-malignant tissue in 83 AC pairs (EDRN cohort) and determine whether increased gene dosage resulted in increased gene expression. A Spearman's correlation conducted using MATLAB software was used to determine the strength of the correlation between copy number and expression, with a coefficient >0.55 considered significant.

4.2.4 Analysis of external cohorts

Amplification of *YEATS4* was assessed in six publically available datasets. Affymetrix SNP 6 data from GSE25016, Early Detection Research Network (EDRN), the Broad Institute and TCGA were downloaded and segmented using Partek Genomics Suite Copy Number, Paired Analysis Workflow, using the matched non-malignant profiles as a copy number baseline for each tumor and the following parameters; signal to noise >0.3, minimum of 50 markers per segment, p-value threshold of 10^{-7} for the statistical difference between intensities of adjacent segments and for significance of deviation of intensities in tumor tissue from intensities in non-malignant lung. An additional dataset with array CGH profiles from Memorial Sloan Kettering was segmented using the break point algorithm FAÇADE [149]. Data from the Sanger Institute's Cancer Cell Line Project was used to investigate the prevalence of *YEATS4* amplification in human cancer cell lines. Copy gain was defined as 2.3 to 5 copies and amplification as greater than 5 copies. *YEATS4* expression was also assessed in the EDRN and TCGA data sets. The details of all datasets used in this chapter are listed in Appendix A3.

4.2.5 Plasmids and generation of stably expressing YEATS4 lines

Lentiviral short hairpin RNA constructs targeting YEATS4 were purchased from Open Biosystems. Lentiviral production and infection were performed as previously described [98]. Five independent pLKO.1 lentiviral shRNA constructs targeting YEATS4; shY4-1 (TRCN0000013143) shY4-2 (TRCN0000013144) shY4-3 (TRCN0000013145), shY4-4 (TRCN0000013146) and shY4- 5 (TRCN0000013147) were tested for their ability to reduce YEATS4 mRNA and protein expression. NSCLC cell lines were transfected with constructs containing a single shRNA targeting YEATS4 (shY4) or an empty vector control (PLKO). YEATS4 knockdown was confirmed by quantitative real-time PCR (RT-qPCR). shY4-1 resulted in the greatest degree of knockdown and was used for all subsequent experiments. Stably transfected lines were maintained in growth media supplemented with puromycin, 0.8 μg/ml (A549), 1 μg/ml (H1355, HCC4011, H226) or 1.5 μg/ml (H1993).

Ectopic expression of *YEATS4* was achieved using Invitrogen's Ultimate ORF (clone IOH2880) following manufacturer's instructions. The *YEATS4* gene insert was shuttled from a pENTR™221 entry vector to the lentiviral destination vector pLenti6.3/V5-DEST by LR recombination. Lentiviral vector viral stock was produced by transfecting H293FT cells using Invitrogen's ViraPower ™HiPerform™ Lentiviral Expression System. HBEC-KT and HBEC-KT53 lines seeded in 6 well plates were transformed with virus expressing *YEATS4* or an empty vector (EV) and stable transformants were selected following two weeks of treatment with 2µg/ml Blasticidin. Transfected lines were maintained in growth media containing 2µg/ml blasticidin. *YEATS4* expression was confirmed by qRT-PCR and western blotting.

4.2.6 Quantitative reverse transcriptase PCR

RT-qPCR was performed on SDS7900HT (Applied Biosystems) using the ΔΔCt method with 18S rRNA expression levels used as a reference for normalization. Validation of 18S as an appropriate reference gene for lung cancer is described in [150]. Reverse transcription was performed using the High Capacity cDNA Reverse Transcription kit (Applied Biosystems) according to manufacturer's instructions. TaqMan gene expression assays and master mix used were: YEATS4 (Hs00232423_m1), 18S (Hs99999901_s1) and Gene Expression Master Mix (4369016). RT-qPCR was performed on a reaction volume of 15µl containing 0.2 ng of cDNA and using default thermal cycling conditions (2 mins at 50°C, 10 min at 95°C and 40 cycles of 15sec at 95°C and 1 min at 60°C). Data was analyzed using 7500 Fast System Software v1.4 with auto calibration and outliers removed. Samples were analyzed with reference to their matched control, with an RQ of greater than 2 used to define overexpression. Expression array findings were validated in a cohort of 59AC pairs for which RNA was available for both the tumor and matched non-malignant tissue. Of the 59 AC pairs analyzed, 35 of the sample pairs were profiled by expression array. A Pearson correlation was used to determine the concordance between array data and RT-qPCR results.

4.2.7 siRNA mediated knockdown of CDKN1A

On-Target plus SMART pool siRNA targeting *CDKN1A* (L-003471-00) and a non-targeting control (NTC) pool (D-001810-10) were purchased from Thermo Scientific. siRNA knockdown of CDKN1A was performed on H1993, H1355 and H226 PLKO and shY4 cell lines as described in Chapter 3. One day prior to infection, cells were seeded in six well plates at a density of 125,000 cells/well for senescence and 200,000 cells/well for lysates and collected 48 and 72 hours post transfection, respectively. Knockdown efficiency was assessed by western blot.

4.2.8 Western blot

Protein lysates were collected and western blots performed as previously described [151]. Membranes were incubated with primary antibodies against YEATS4 (W-21, Santa Cruz), MDM2 (HDM2-232, Santa Cruz), p53 (ab7757 AbCam), p21 (#2946 Cell Signaling), p14 (#2407 Cell Signaling), phospho p53 (#9284 Cell Signaling), cleaved PARP (#9541 Cell Signaling), pRb807/811 (Cell Signaling# #9308), p27/kip1 (#2552 Cell Signaling) and GAPDH or B-Actin as loading controls (#2118 and #4970 respectively, Cell Signaling). Following primary antibody incubation, membranes were incubated with anti-mouse or anti-rabbit HRP conjugated secondary antibodies (Cell Signaling #7074, 7076) and visualized by enhanced chemiluminescence.

4.2.9 Cell viability

MTT assays were used to assess cell viability following shRNA knockdown and overexpression of *YEATS4*. Cells were seeded in triplicate in 96 well plates at optimal densities for growth (2000 cells/well for H1993, HBEC-KT, HBEC KT-53, 2500 cells/well for H1355, H226, A549 and 4000 cells/well for HCC4011). Media only wells were plated to serve as absorbance blanks to normalize wells. Viability was measured over five consecutive days by the addition of 10ul MTT reagent (every 24 hours) and incubated for an additional four hours, followed by the addition of 100ul of 20% SDS to solubilize the MTT dye. Plates were quantified by spectrophotometry (EMax plate reader, Molecular Devices) at 570nm with reference to 650nm. Experiments were performed in triplicate.

4.2.10 Colony formation

Anchorage independent growth was assessed in all transfected cell lines by the soft agar method. Single cell suspensions were prepared in growth medium with 10% FBS and 0.3% low melting point agarose, and seeded in triplicate at 1000 cells/well in 12 well plates and cultured for 14 days at 37°C. Colonies were stained with MTT and counted, with colony formation is reported as the mean ± SEM normalized to the average of the control (PLKO) of triplicate experiments.

4.2.11 Cellular senescence

Beta- galactosidase (βgal) activity at pH 6 was used as a measure of cellular senescence. Cells were plated in triplicate in 6 well plates at concentrations such that 24 hours later they were roughly 50% confluent and processed according to the manufacturer's instructions (Senescent Cells Histochemical staining kit, Sigma). Cells were visualized by a phase contrast microscope and images taken of three areas within each well. The percent of senescent cells was calculated and differences between PLKO and shY4 were assessed by a student's t-test with a p-value <0.05 considered significant.

4.2.12 Dose response assay

To measure the effect of *YEATS4* manipulation on cisplatin and nutlin sensitivity, cell viability assays were performed as described in Chapter 3. Cells were plated in triplicate in 96 well plates at optimal densities for growth (H226, A549, HBEC KT and KT53 at 2500 cells/well, H1993 and H1355 at 4000 cells/well, and HCC4011 at 6000 cells/well) and subjected to a series of 20, 2-fold dilutions of cisplatin or nutlin inhibitor prepared in cell growth media with final concentration of 1% DMSO. The experimental inhibitor concentrations ranged from 333 µM to 635 pM for cisplatin and 100µm to 191pm for nutlin. Experiments were repeated in triplicate and differences in IC_{50} values were determined using a student's t-test with a p-value < 0.05 considered significant.

4.2.13 Mouse models

Tumor forming ability of H1993 and H1355 shY4 and PLKO cells was assessed in Crl:nu-foxn1nu mice. Subcutaneous flank injections of $3x10^6$ cells in 50µl of PBS were injected into 6-week old mice obtained from The Jackson Laboratory. Tumor size and volume was measured by palpation 11 days following injection and every 6-8 days thereafter. Experiments were terminated once tumors reached a volume of 400 mm^3 or became ulcerated. Tumor burden between mice injected with YEATS4 KD (n=10) cells and those expressing and empty vector (PLKO) (n=10) was assessed using a student's T-test.

4.2.14 Pathway analysis and gene set enrichment

Following knockdown, gene expression profiles for H1993, H1355 and H226 shY4 and PLKO cells were generated using Illumina HT12 expression microarrays and normalized using BRB array tools. A list of over- and underexpressed genes (2-fold difference) for each knockdown line relative to empty vector controls were generated. Genes altered in all lines in the same direction were considered for target gene analysis using Ingenuity Pathway Analysis software. A pre-ranked gene set enrichment analysis was performed on fold change data for all genes for each cell line using the C3 transcription factor gene set to determine which transcription factor target genes were affected by YEATS4 knockdown.

4.2.15 Survival analysis

Survival analysis was performed using a Mantel-Cox log test in *Matlab* with *p*-values <0.05 considered significant. Director's Challenge data was sorted by *YEATS4* expression and survival times between the top and bottom tertiles of expression compared and Kaplan-Meier plots generated.

4.3 Results

4.3.1 Recurrently amplified regions in NSCLC

Copy number profiles for 169 AC and 92 SqCC were generated using aCGH as described in Chapter 2. Significant regions of high level amplification ($\log_2$ ratio >0.8) were identified using the Genomic Identification of Significant Targets in Cancer (GISTIC) algorithm which calculates significance scores by considering both the amplitude and frequency of copy number alterations [148]. GISTIC analysis of all 261 samples (NSCLC) identified 3 significant regions of focal amplification; 7p11.2 (q=0.00075), 8p12 (q= 0.036) and 12q15 (q=0.036). Subtype specific analysis revealed 2 regions of amplification across the 169 AC tumors; 12q15 (q= 4.5×10^{-5}) and 20q13.33 (q=0.017) and 6 regions across the 92 SqCC tumors; 1p34.2 (q= 0.044), 3q27.1 (q=1.4×10^{-10}), 7p11.2 (q=0.029), 8p11.23 (q=0.0042), 8p12 (q=0.0042) and 14q13.3 (q=0.03) (Fig. 4.1A-C). Amplification of these regions have been previously described in NSCLC indicating our tumors display patterns of alteration characteristic of lung cancer [18, 55, 152, 153].

While none of the regions identified were common between all three analyses, all of the regions identified in NSCLC were also significant in a subtype specific manner (Figure 4.1A-C). Further examination of these amplicons revealed that known oncogenes EGFR and BRF2, both of which are known to be preferentially amplified in SqCC, [114] [98] were driving selection of the 7p11.2 and 8q12 amplicons, respectively. Intriguingly, the primary target of 12q15 amplification, which is believed to be MDM2- a ubiquitin ligase that targets TP53 for proteasomal degradation, and when overexpressed results in aberrant p53 inactivation, was excluded from both the focal and wide peak boundaries. The exclusion of MDM2 from this focal region suggested that a gene other than MDM2 may be driving selection of this amplicon. This combined with the fact that all other regions harbored known oncogenes 7p11.2 (*EGFR*), 8p11.23 (*FGFR1*), 8p12 (*BRF2*) 14q13.3 (*NKX2-1*), 20q13.3 (*EEF1A*) or are known to be subtype specific regions of amplification (1p34.2 and 3q in SqCC) [18, 55] prompted us to further explore the 12q15 amplicon.

Figure 4.1

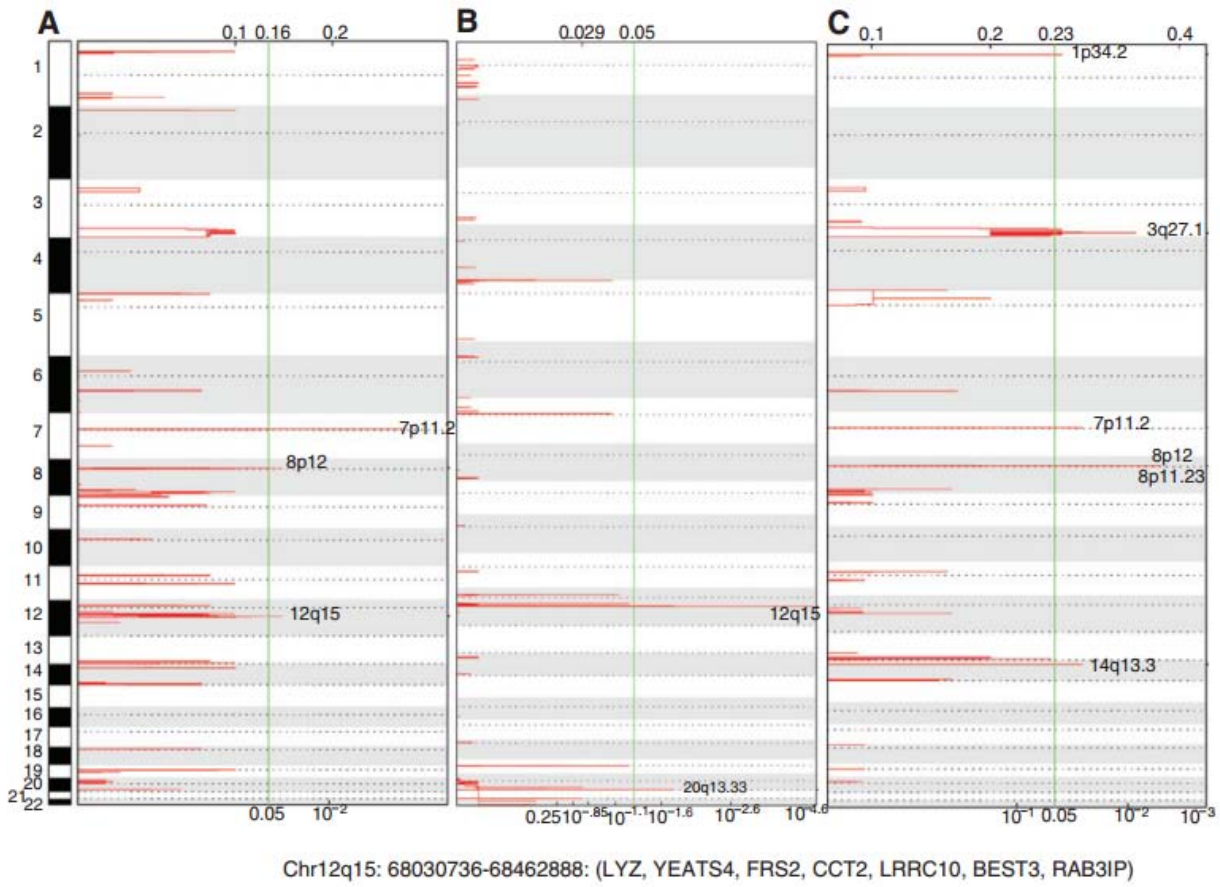


Figure 4.1. Recurrent amplifications in NSCLC.

GISTIC plots for (A) 261 NSCLC, (B) 169 AC and (C) 92 SqCC. Chromosomes are depicted as rows and chromosome numbers are indicated. Red peaks indicate frequently amplified regions and the green vertical line indicates the false discovery rate threshold (q=0.05). Peaks extending beyond this line indicates a significant region. X-axis indicates the GISTIC score scale. Genomic coordinates and the genes located within the 12q15 amplicon are shown below.

4.3.2 Identification of YEATS4, the target of 12q15 amplification

The peak amplified region of 12q15 spanned a 432 kb interval (68,030,736-68,462,888) and contained 7 genes; *LYZ, YEATS4, FRS2, CCT2, LRRC10, BEST3, RAB3IP*, none of which have been previously implicated in lung tumorigenesis. Based on the notion that selectively amplified oncogenes would demonstrate elevated expression, we integrated copy number and gene expression data for adenocarcinoma tumors and matched non-malignant tissue. Due to the limited size of our dataset with both copy number and expression data, identification of the 12q15 driver gene was performed in the largest dataset available (EDRN, n=83). Of the 7 genes within the amplicon, only *YEATS4* was both gained/amplified and concomitantly overexpressed in lung tumors relative to matched non-malignant tissues (Figure. 4.2A-C). While *YEATS4* has not been previously described in lung cancer, it is a well-established oncogene in cancers of neural origin [154, 155] and frequently amplified in liposarcomas [156].

4.3.3 YEATS4 is frequently amplified and overexpressed in NSCLC

YEATS4 was amplified in 18% (47/261) and overexpressed in 31% (15/48) of cases from our cohort. While 12q15 was not significant in the GISTIC analysis of our 92 SqCC cases, to conclusively determine whether amplification of YEATS4 was specific to AC, we compared copy number and expression data for both subtypes. Although no statistical difference in *YEATS4* copy number or expression was observed between subtypes (Figure. 4.2D-F), AC tumors had a higher number of copies and greater fold change in expression compared to SqCC tumors. This suggests that while copy gain is a frequent event in both subtypes, it is likely a broader amplification event that occurs at a lower amplitude in SqCC relative to AC, which is why 12q15 failed to be identified by GISTIC in SqCC tumors. Analysis of external datasets with both AC and SqCC data supported our findings, with gain /amplification and overexpression occurring at similar frequencies in both data sets (Table 4.1), indicating that amplification and overexpression of *YEATS4* is not subtype specific.

To gain further insight into the prevalence of *YEATS4* amplification, we investigated *YEATS4* copy number and expression in publically available NSCLC tumor datasets. *YEATS4* was gained (2.3-5 copies) or amplified (> 5 copies) at various frequencies across the five datasets, ranging from 5-22% and 0.4-5% respectively (Table 4.1). A broader analysis of 508 human cancer cell lines revealed *YEATS4* copy gain/amp in 43/128 (33.6%) of lung cancer cell lines and in 122/508 (24%) of all cancer cell lines (Table 4.1). Expression analysis of the EDRN and TCGA data sets, revealed YEATS4 was overexpressed at comparable frequencies to our dataset; 18% (15/83) and 33% (14/42), respectively (Table 4.1). Taken together, these results show *YEATS4* is frequently gained and overexpressed in NSCLC, irrespective of subtype.

To validate array findings and verify *YEATS4* is upregulated at the transcript level, we assessed YEATS4 expression by quantitative reverse transcriptase PCR (RT-qPCR) in a panel of 59 lung ACs relative to matched non-malignant tissue and in 18 NSCLC cell lines (2 SqCC and 16 AC) with reference to an immortalized normal human bronchial epithelial (HBEC) line. 15/59 (25.4%) tumors and 8/18 (44.4%) cell lines showed a two-fold or greater increase in *YEATS4* expression relative to their matched control. Moreover, analysis of the 35 AC samples with expression data revealed a strong correlation between array findings and PCR results ($r=0.75$, $P<0.001$, Pearson Correlation, data not shown), validating array findings and confirming frequent overexpression of *YEATS4*. Western blotting of cell lines with and without *YEATS4* amplification revealed increased *YEATS4* expression in lines with amplification, demonstrating that amplification drives overexpression at both the mRNA and protein level (Figure. 4.2G). Multivariate analysis of *YEATS4* revealed no significant associations with age, sex, stage, smoking status or race. Survival analysis of the Director's challenge expression datasets [157] revealed a trend towards poorer survival in patients with *YEATS4* amplification, however failed to reach statistical significance ($p>0.05$, data not shown).

Figure 4.2

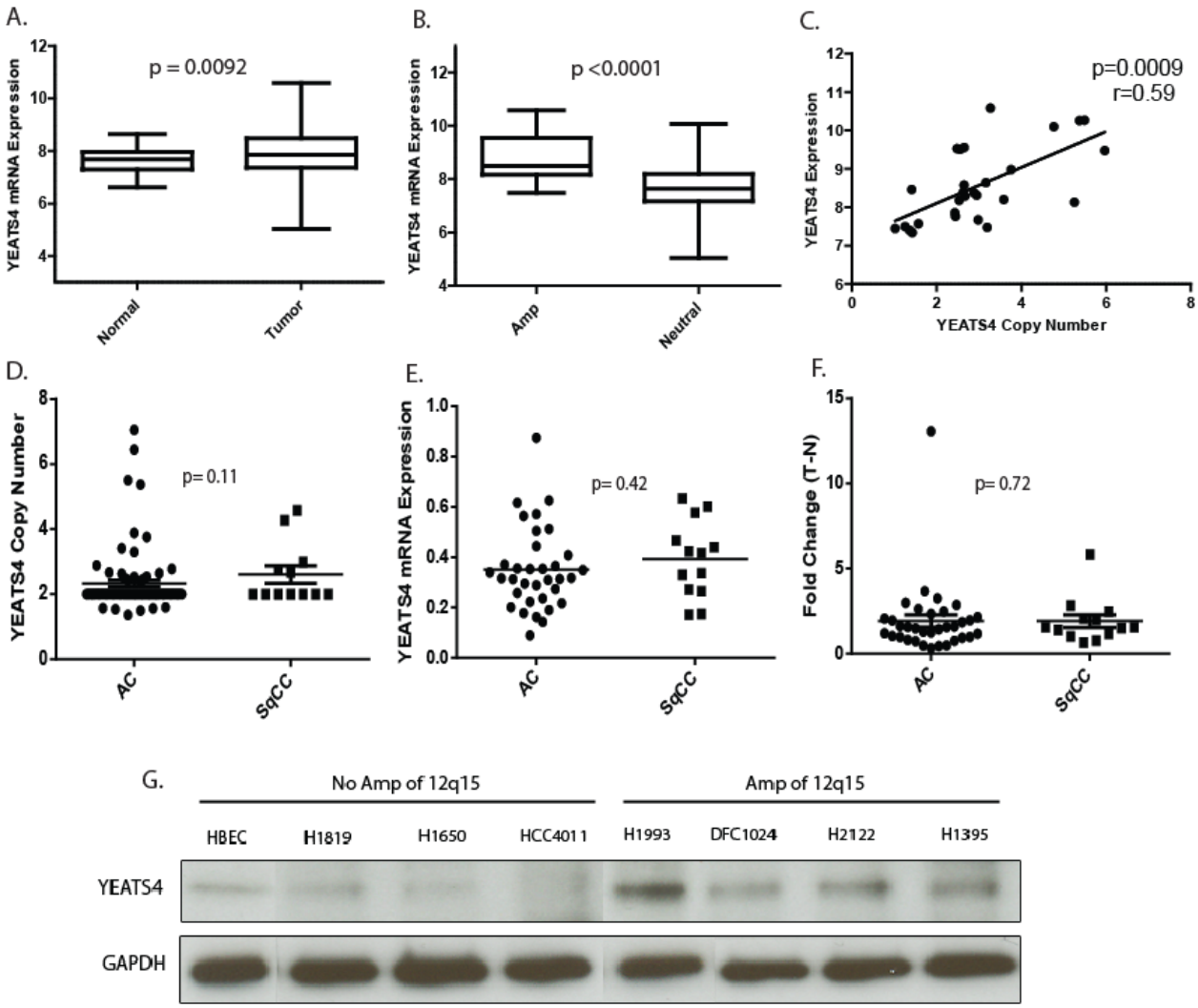


Figure 4.2 *YEATS4* is the target of 12q15 amplification.

(A) Comparison of mRNA expression in 83 AC tumors and matched non-malignant tissue from the EDRN (p=0.0092). (B) *YEATS4* expression between tumors with gain/amplification and tumors with neutral copy number (p<0.0001). (C) Spearman's correlation of copy number and expression for tumors with copy number alterations of *YEATS4* (r=0.59, p=0.009). Comparison of YEATS4 copy number (D), expression (E) and fold change (F) between AC and SqCC. Expression values for all plots are in log2 units. (G) Immunoblot of YEATS4 in NSCLC lines with and without amplification of 12q15. GAPDH was used as a loading control.

Table 4.1 Frequency of YEATS4 gain and amplification across multiple datasets

Data Set	Histology	# Samples with CN	Gain (n) (2.3-5 copies)	Freq Gain (%)	Amplification (n) (>5 copies)	Freq Amp (%)	Samples with Expression (n)	Samples with OE	Freq OE (%)
BCCRC	AC & SqCC	169,92	28,19	16.5,20.1	5,2	3.0, 2.2	35,13	11, 4	31,30
EDRN	AC	83	18	21.7	4	4.8	83	15	18
GSE25016	AC & SqCC	77,155	10,22	13, 14.2	0,1	0, 0.6	N/A	N/A	N/A
dbGAP	AC	354	18	5.1	4	1.1	N/A	N/A	N/A
MSKCC	AC	199	21	10.6	N/A	N/A	N/A	N/A	N/A
Broad	NSCLC	473	28	5.9	8	1.7	N/A	N/A	N/A
TCGA	AC & SqCC	277,201	19,22	6.9,10.9	15,6	5.4,3.0	17,25	8, 6	32, 35
Sanger LCCL	NSCLC/SCLC	128	43	33.6	1	0.78	N/A	N/A	N/A
Sanger cell lines	All cancers	508	122	24	8	1.5	N/A	N/A	N/A

4.3.4 YEATS4 displays oncogenic properties *in vitro* and *in vivo*

YEATS4 encodes a protein found in a number of multi-subunit protein complexes involved in chromatin modification and transcriptional regulation and has also been shown to be involved in the regulation of TP53. To assess its oncogenic potential, *YEATS4* was stably transfected into two immortalized HBEC lines; HBEC-KT and HBEC-KT53 (KT-YEATS and KT53-YEATS). Empty vector transfected cells were used as controls (KT-EV and KT53-EV). YEATS4 gene and protein expression was confirmed by qPCR and western blot (Fig. 4.3A-B). Relative to controls, ectopic expression of YEATS4 had no effect on viability and failed to induce anchorage independent growth in HBECs (data not shown), indicating that in normal, immortalized cells, YEATS4 overexpression alone is incapable of inducing colony formation. However, a dramatic inhibition of senescence in overexpressing cells relative to controls was observed in both lines (Student's t-test, $p<0.05$) (Fig. 3C-D), suggesting elevated *YEATS4* expression is capable of inducing a phenotype associated with malignant transformation.

Figure 4.3

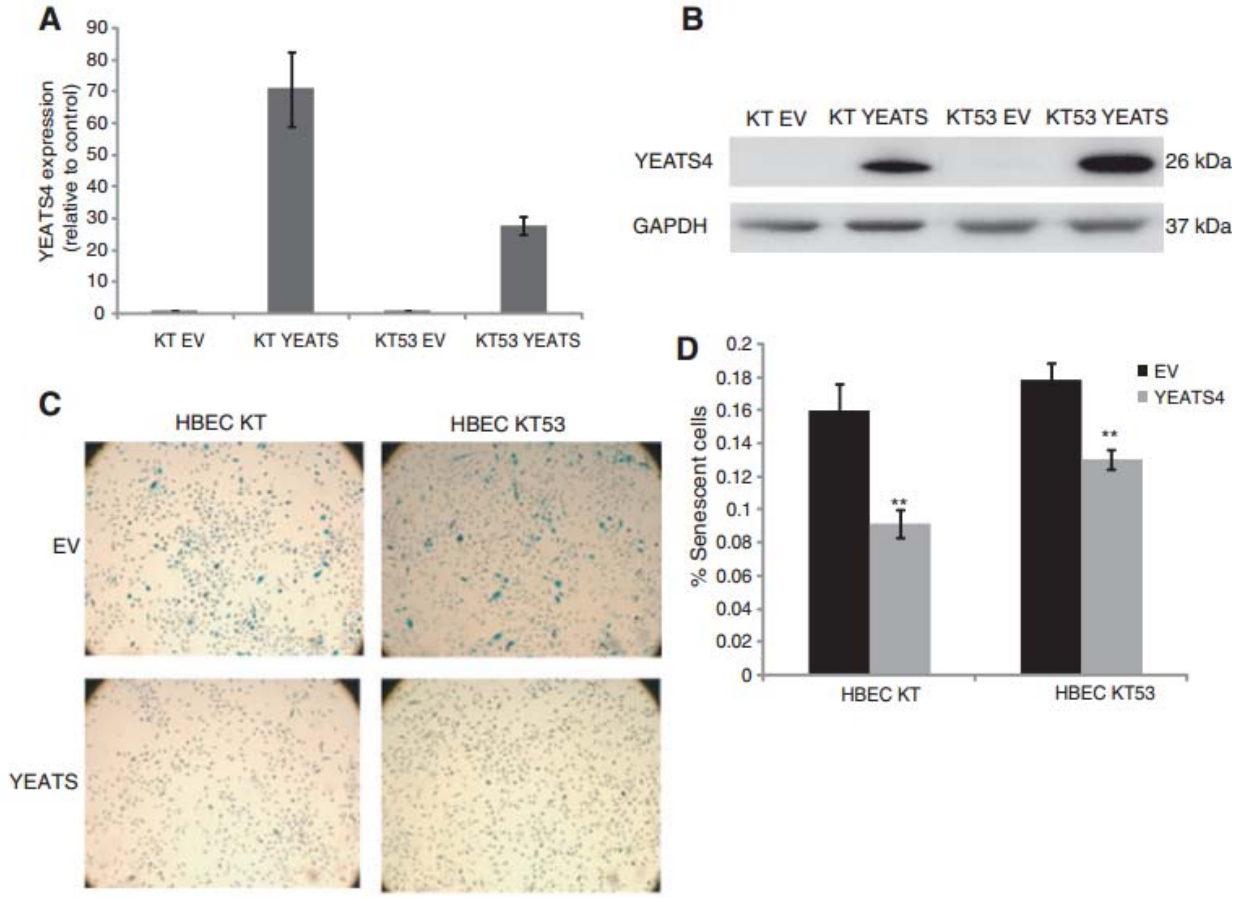


Figure 4.3. Overexpression of YEATS4 induces a malignant phenotype.

Ectopic expression of YEATS4 increases (A) mRNA expression (mean± SEM of triplicate replicates) and (B) protein levels relative to EV controls. GAPDH was used as a loading control. (C) β-Gal staining for cellular senescence in EV and YEATS4 expressing HBECs. Cells stained blue indicate senescence. Original magnification, 10x. (D) Quantification of cellular senescence in YEATS4 and control cells. The mean of the proportion of senescent cells (senescent cells/total cells) for YEATS4 and EV lines is shown for triplicate experiments ± SEM. ** $p<0.01$, Student's t-test.

Complimentary knockdown experiments using lentiviral shRNAs were performed in lung cancer cell lines with (H1993, H1355, H226) and without (A549, HCC4011) *YEATS4* amplification and various p53 backgrounds. Empty vector transfected cells were used as controls (PLKO) and knockdown was confirmed by qPCR and western blotting (Figure 4.4A-B). Knockdown significantly decreased cell viability in H1993 and H1355 ($p=0.0127$ and $p=0.0172$, respectively), both of which harbour *YEATS4* amplification and mutant *p53* (Figure 4.4C), but had no effect on A549, HCC4011 or H226 lines ($p = 0.428$, $p = 0.45$ and $p = 0.49$, respectively) which do not harbor *YEATS4* amplification (A549 &H4011), or have *YEATS4* amplification with wild type (wt) *p53* (H226) (Figure 4.4C). Similarly, knockdown resulted in a significant decrease in anchorage-independent colony formation in H1993 ($p = 7.26 \times 10^{-6}$) and H1355 ($p = 6.06 \times 10^{-10}$) cells, but not in A549 ($p = 0.97$), H4011 ($p = 0.21$) or H226 ($p = 0.74$) cells, indicating wt *p53* may abrogate the effect of *YEATS4* knockdown on viability and colony formation in lines with amplification (Figure 4.4D). A significant increase in senescence was observed in all three lines with amplification; H1993 ($p = 5.71 \times 10^{-6}$), H1355 ($p = 0.0012$) and H226 ($p = 1.21 \times 10^{-13}$) as well as moderate increase in A549 ($p = 2.99 \times 10^{-7}$). No difference in senescence was observed in HCC4011 ($p = 0.06$) (Figure 4.4E). The finding that A549 cells showed a modest increase in senescence is not surprising given the role of YEATS4 in the p53 pathway (discussed below) and the wt p53 background of this line, which enables pathway activation and cellular senescence.

To explore the oncogenic potential of YEATS4 *in vivo*, tumor formation in SCID mice was examined by subcutaneous flank injections of H1993 and H1355 control and shY4 cells. Tumor formation was significantly reduced in shY4 cells of both cell lines at all time points (Figure 4.4F,G). Our results demonstrate that knockdown of YEATS4 in cell lines with amplification effectively inhibits tumor growth, with a significant inhibition in viability, tumor and anchorage independent growth and increased cellular senescence, strongly supporting YEATS4 as an oncogene in NSCLC.

Figure 4.4

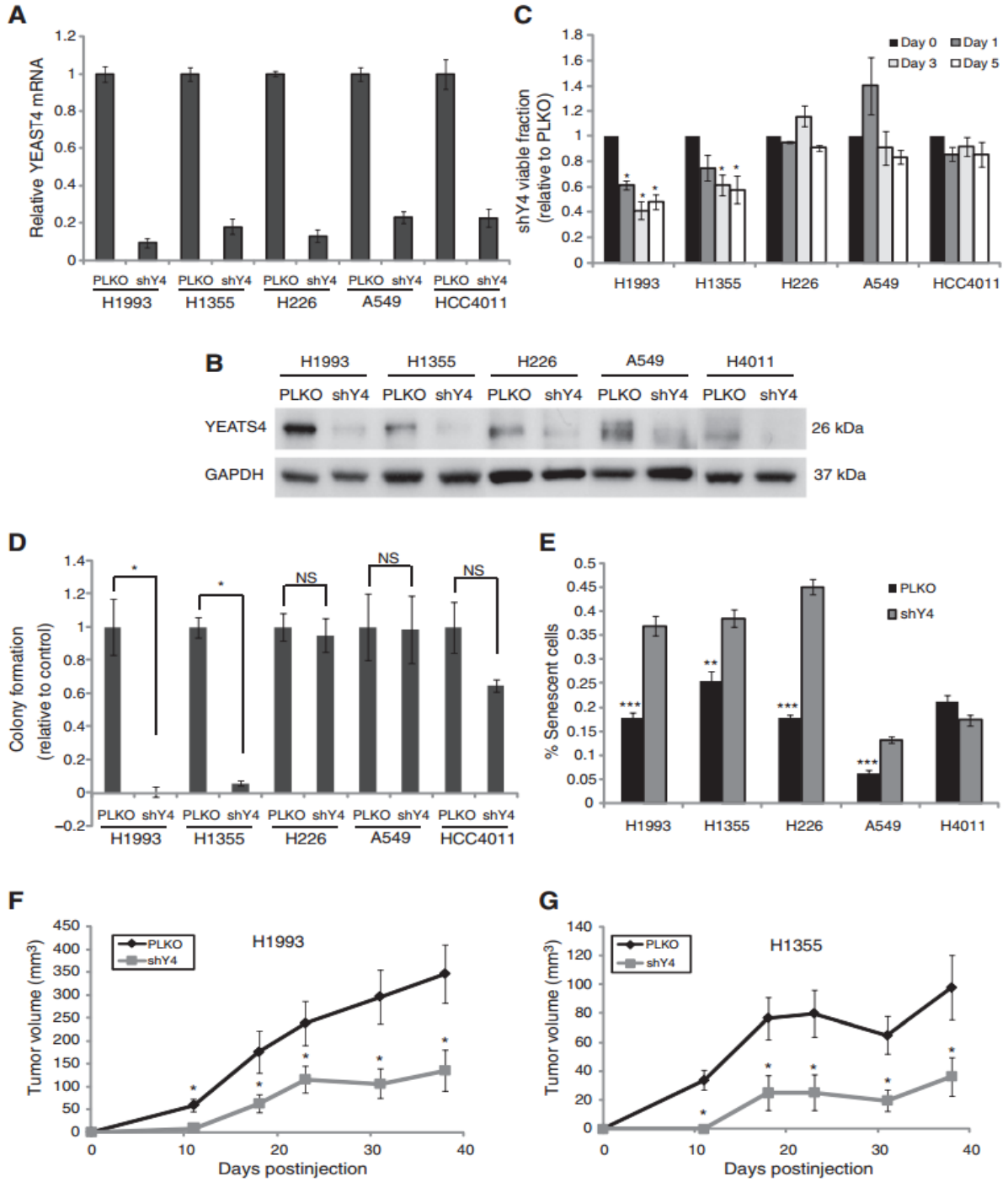


Figure 4.4. YEATS4 knockdown impairs growth and induces senescence.

Figure 4.4. YEATS4 knockdown impairs growth and induces senescence.
shRNA targeting YEATS4 significantly reduces (A) mRNA expression and (B) protein levels in all cell lines relative to controls (PLKO). GAPDH was used as a loading control. (C) Viability of cell lines with knockdown (shY4) relative to controls as measured by MTT. (D) Colony formation ability of shY4 cell lines relative to controls. (E) Quantification of cellular senescence based of β-Gal staining. Values reported as mean ± SEM of triplicate experiments.* $p<0.05$, ** $p<0.01$, *** $p<0.001$, Student's *t*-test of shY4 cells relative to PLKO. (F,G) Effect of YEATS4 knockdown on tumor growth in mice injected with H1993 or H1355 PLKO and shY4 cells. Error bars indicate SEM of each group of 10 mice, * $p<0.05$

4.3.5 YEATS4 suppresses p53 and p21

Inactivation of the p53 pathway is one of the most frequent alterations in lung cancer, with somatic mutations occurring in approximately 50% of all cases [114, 158]. *p53* is a key tumor suppressor that regulates cell cycle, DNA repair, apoptosis, and senescence and inhibits aberrant proliferation and the propagation of damaged cells. A study by Park *et al*, showed that under normal, unstressed conditions, *YEATS4* binds to and inhibits the promoters of *p14* and *p21*, subsequently repressing the p53 tumor suppressor pathway [159]. To assess whether this interaction occurs in NSCLC, we assessed these proteins in cell lines with *YEATS4* manipulation. Upon *YEATS4* knockdown, p21 and p53 protein levels were increased, with the greatest increases in expression of p21 and p53 observed in cell lines harbouring *YEATS4* amplification or wt *p53* respectively (Figure 4.5A). No change in p14 levels was observed upon knockdown. Overexpressing lines showed a modest reduction of p21 and p14 as well as a reduction of p53 levels in HBEC-KT (Figure 4.5B). MDM2 levels remained unchanged following knockdown or overexpression of *YEATS4*, indicating that the observed changes in p21, p14 and p53 were a direct result of *YEATS4* manipulation.

4.3.6 YEATS4 alters the sensitivity of cell lines to cisplatin and nutlin

To determine whether the downstream effects of *YEATS4* manipulation alters cellular sensitivity to chemotherapy, cell lines were treated with serial dilutions of cisplatin, a commonly

prescribed first line chemotherapy for lung cancer patients that crosslinks DNA triggering apoptosis, or nutlin, a cis-imidazoline analog that inhibits the interaction of p53 and MDM2, stabilizing p53. Based on the observed effects on p53 and p21 protein levels following manipulation of *YEATS4* expression and the notion that cells with *YEATS4* amplification may be dependent on *YEATS4* for growth and survival, we hypothesized that HBEC-KT/KT53-Y cells would be more resistant to treatment, while shY4 cells harbouring *YEATS4* amplification would be more sensitive.

As expected, HBEC-KT-YEATS and HBEC-KT53-YEATS lines were significantly more resistant to both cisplatin and nutlin than their control counterparts (Figure 4.5C; Table 4.2). Differences in sensitivity were less consistent in the lung cancer cell lines, likely due to the fact these cell lines harbour numerous genomic alterations which could influence drug sensitivities. While H1993 shY4 cells were significantly more sensitive to cisplatin (IC_{50} PLKO:11.45 vs. shY4:8.65) (Figure 4.5D) supporting our hypothesis, knockdown in both H1355 and H226, showed the opposite trend resulting in greater resistance relative to controls (Table 4.2). As anticipated, A549 and HCC4011 shY4 cells showed no difference in sensitivity. As specimens with mutant p53 are resistant to nutlin, only A549 and H226 were treated. Similar to the cisplatin results, A549 shY4 cells showed no significant difference in sensitivity to nutlin (PLKO: 7.58 vs. shY4:6.91, p=0.84), while H226 shY4 cells were unexpectedly significantly more resistant (PLKO:3.27, shY4:4.33, p=0.033) (Table 4.2). Analysis of lung cancer cell line IC50 data from the Sanger drug sensitivity project failed to reveal a significant association between *YEATS4* amplification and response to cisplatin or nutlin. However, based on the fact that transformed bronchial epithelial cells which harbour minimal genetic alterations were significantly more resistant to cisplatin and nutlin following overexpression of YEATS4, and H1993 shY4 cells (which harbor the greatest amplification of *YEATS4*) were more sensitive to cisplatin compared to controls, we feel this data supports the notion that YEATS4 alters the *in vitro* sensitivity of lung cells to cisplatin and nutlin.

Figure 4.5

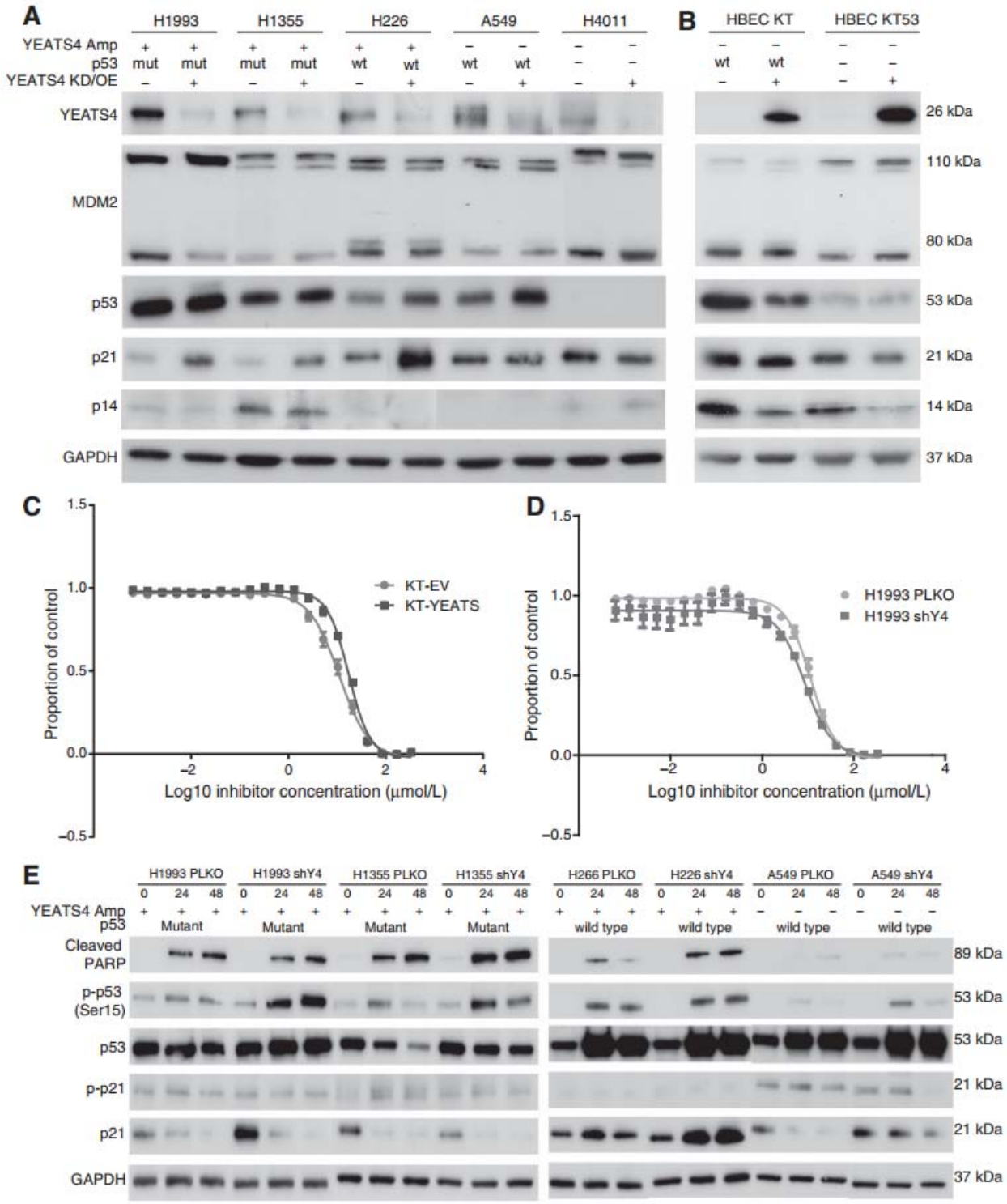


Figure 4.5. YEATS4 alters p21 and p53 protein levels.

Figure 4.5. YEATS4 alters p21 and p53 protein levels.

(A) Knockdown of YEATS4 increases expression of p21 in cell lines with YEATS4 amplification and increases p53 in all lines that express p53. (B) Overexpression of YEATS4 reduces p14 and p21 levels in both HBEC lines, and p53 only in the HBEC KT line. Dose-response curves of HBEC KT (C) and H1993 (D) cells treated with 2-fold dilutions of cisplatin for 72 hours. Viability is shown as a proportion of treated cells against untreated controls (mean ± SEM of triplicate experiments). (E) Immunoblot of PLKO and shY4 cell lines treated with 40μM of Cisplatin for 0, 24 or 48 hours. Cisplatin treatment induces apoptosis as measured by the increase in cleaved PARP, p53 and phosphorylated p53 (Ser15). GAPDH was used as a loading control for all blots.

Table 4.2 Cisplatin and Nutlin IC_{50}s

		Cisplatin			Nutlin			
Cell Line	**YEATS4**	**IC50**	**SEM**	***t* Test**	**IC50**	**SEM**	***t* Test**	**Trend**
H1993	PLKO	11.45	1.11					
H1993	shY4	8.649	0.488	0.004				Sensitive
H1355	PLKO	9.111	0.491					
H1355	shY4	15.91	0.905	1.6E-06				Resistant
H226	PLKO	5.788	0.276		3.266	0.130		
H226	shY4	9.965	0.716	0.0003	4.325	0.376	0.033	Resistant in both
A549	PLKO	10.88	0.452		7.58	0.316		
A549	shY4	11.4	0.705	0.204	6.908	0.123	0.084	Not significant
H4011	PLKO	8.952	0.326					
H4011	shY4	10.32	0.566	0.055				Not significant
HBEC KT	EV	11.09	1.472		16.55	1.405		
HBEC KT	YEATS	17.32	0.696	0.007	26.01	2.696	0.023	Resistant in both
HBEC KT53	EV	15.41	1.612		19.63	1.367		
HBEC KT53	YEATS	20.38	0.718	0.029	24.85	1.385	0.034	Resistant in both

4.3.7 Sensitivity to cisplatin is not mediated solely through the p53-p21 pathway

To gain further insight into the potential mechanisms of altered sensitivity to cisplatin, cell lines were treated with 40μM cisplatin and protein lysates collected at 0, 24 and 48 hours post treatment. As expected, cisplatin treatment resulted in an increase in p53, p53 Ser15 phosphorylation (a marker of stabilization), p21 and induced apoptosis as measured by cleaved PARP in HBECs. However, no differences between HBEC-EV and HBEC-YEATS cells were observed for any of the proteins examined (data not shown). In shY4 cells with amplification, treatment with cisplatin led to a greater induction of p53 and phospho-p53 (Ser15), and in H226 also led to a significant increase in p21 levels relative to control cells (Figure 4.5E). As no significant differences in protein levels were observed, despite a significant increase in resistance following OE, our results suggests that while the p53-p21 signaling pathway may be involved, resistance is likely mediated through the interaction of *YEATS4* with other signaling pathways.

4.3.8 Knockdown phenotypes are independent of p21 signaling in mutant p53 cells

To explore the effect of increased p21 expression on the observed phenotypes, siRNA knockdown of *CDKN1A* was performed on shY4 and PLKO cells for cell lines with *YEATS4* amplification. Knockdown of *CDKN1A* showed no effect on viability or colony formation in any of the lines (data not shown), but significantly altered senescence levels in the presence of wt p53 (Figure 4.6A). *CDKN1A* siRNA reduced senescence in both H226 shY4 and PLKO cells relative to non-targeting control siRNA treated cells, such that the percent of senescent H226 shY4-p21 cells was similar to H226 PLKO-NTC (Figure 4.6A). These experiments suggest that in a wildtype p53 background, the increase in senescence following *YEATS4* knockdown occurs in a p53 dependent manner and is the direct result of increased p21expression. As *CDKN1A* knockdown failed to rescue viability, colony formation and senescence in cell lines with mutant p53, these findings further support the notion that the phenotypes observed following *YEATS4* knockdown are not solely due to changes in p53-p21 signaling. Based on these findings, and the prominent role of *Rb* in senescence, we next investigated whether the increased senescence following *YEATS4* knockdown could be due to altered *Rb* signaling. Modest reductions in Rb Ser807/811 phosphorylation following *YEATS4* knockdown were observed, however in mutant p53 cell lines this does not appear to be due to increased levels of p27 (Figure 4.6B).

Figure 4.6

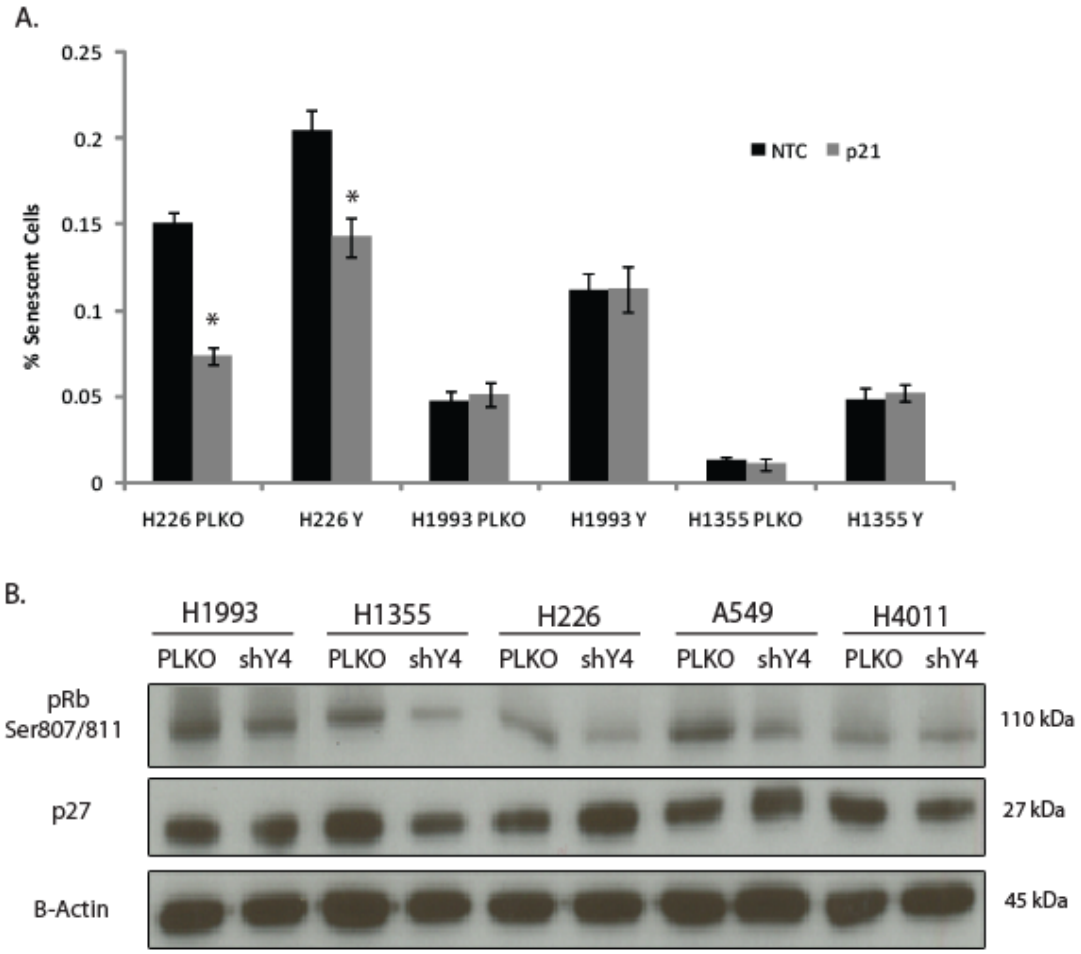


Figure 4.6. Senescence induced by YEATS4 is dependent on p21 in wildtype p53 cells.

(A) Quantification of percent senescent cells following siRNA mediated knockdown of p21 in PLKO and shY4 cells. * $p<0.05$, Student's t-test. Values reported as mean ± SEM of triplicate experiments. (B) Knockdown of YEAST4 leads to reduced Rb Ser807/811 phosphorylation and p27 in p53 mutant cells. B-actin was used as a loading control.

4.3.9 Identification of additional cellular networks regulated by YEATS4

In an attempt to gain a better understanding of other pathways *YEATS4* is involved in, we performed expression profiling on shY4 and PLKO cells for the three cell lines with *YEATS4* amplification. To identify significantly enriched pathways/networks and gene sets affected by *YEATS4* knockdown, Ingenuity Pathway Analysis (IPA) and Gene Set Enrichment Analysis (GSEA) were performed. A total of 32 genes (27 overexpressed and 5 underexpressed, Table 4.3) were differentially expressed between knockdown and control cells across all three cell lines. Due to the small number of input genes, none of the significantly enriched canonical pathways passed multiple testing correction. However, network analysis, which assesses regulatory relationships existing between genes and proteins, identified two networks associated with pro-tumorigenic functions; (1) cancer and (2) cell death, survival, cell cycle and cell morphology. These networks were centered around known targets or binding partners of *YEATS4* including *p53*, *CDKN1A* and *MYC* (Figure 4.7), further supporting our *in vitro* findings. Pre-ranked GSEA revealed significant enrichment of a number of transcription factor gene sets including *MYCN*, which has been shown to be a binding partner of *YEATS4* and all 6 *serum response factor (SRF)* gene sets. *SRF* is a ubiquitously expressed transcription factor implicated in cell proliferation, differentiation, metastasis and clinically associated with castration-resistant prostate cancer [160, 161]. Interestingly, *PDLIM7* which contains a serum response element and is transcribed upon induction of *SRF*, was shown to inhibit p53 and p21 through the inhibition of MDM2 self ubiquitination. While neither *MYCN*, *SRF* or *PDLIM7* were differentially disrupted at the mRNA level following knockdown, our downstream analysis suggests the target genes of these two transcription factors could be involved in *YEATS4* mediated tumorigenesis and warrant investigation in future studies to elucidate additional mechanisms through which YEATS4 promotes tumorigenesis.

Table 4.3 Differentially altered genes following knockdown of YEATS4 used for Ingenuity Pathway Analysis

Gene	Direction of change
YEATS4	Underexpressed
MUC1	Underexpressed
SCNN1A	Underexpressed
ST3GAL5	Underexpressed
KLHL5	Underexpressed
CRABP2	Underexpressed
LOC283267	Underexpressed
TINP1	Underexpressed
FNDC3A	Underexpressed
LOC100128196	Underexpressed
NSA2	Underexpressed
ACOT7	Underexpressed
S100A4	Underexpressed
NKD2	Underexpressed
MRPL35	Underexpressed
SULT1A3	Underexpressed
ISCA1	Underexpressed
LOC100132658	Underexpressed
DENND2D	Underexpressed
MGC42367	Underexpressed
LIN7C	Underexpressed
LOC400061	Underexpressed
SLC1A6	Underexpressed
DDX12	Underexpressed
RRAD	Underexpressed
PPIL5	Underexpressed
TNFSF12	Underexpressed
CAMTA1	Overexpressed
RRAGD	Overexpressed
TMEM158	Overexpressed
TAGLN	Overexpressed
GLIPR1	Overexpressed

Figure 4.7

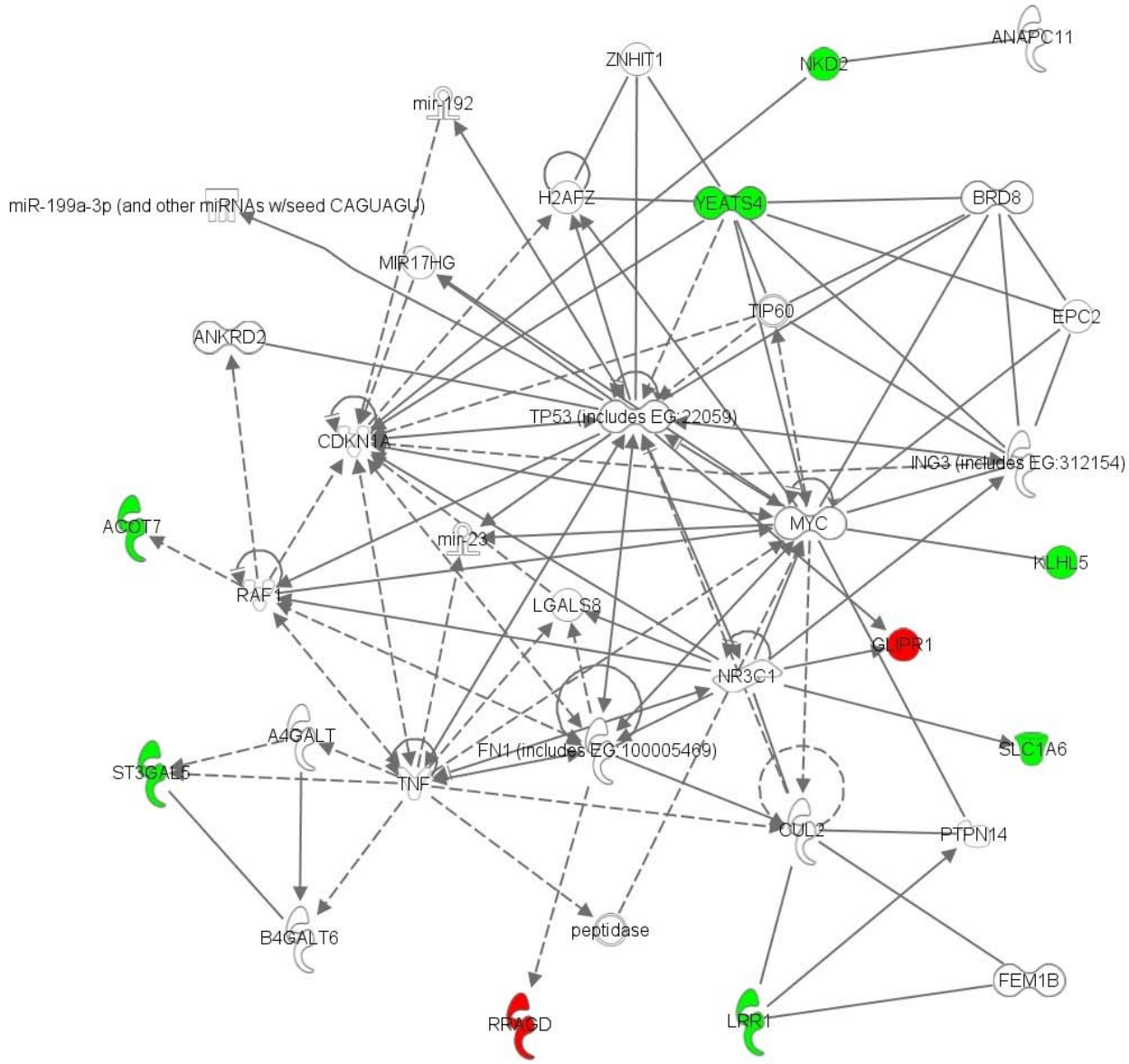


Figure 4.7. Gene networks associated with YEATS4 knockdown.

Ingenuity Pathway Analysis was used to identify biologically related gene networks from the 34 genes with differential expression across all three cell lines (H1993, H1355, H226) following knockdown. One of the top networks- cell death, survival, cell cycle and cell morphology related to p53, CDKN1A, TNF and MYC signaling is displayed. Solid lines denote direct interactions while dotted lines represent indirect relationships. Components highlighted in red are upregulated while those in green are downregulated.

4.4 Discussion

While single dimensional genomic analyses have been instrumental in cancer gene discovery, this type of analysis often overlooks genes disrupted at low frequencies, and is unlikely to distinguish causal from passenger events. The integration of multiple parallel genomic dimensions enables the identification of genes with concurrent DNA and expression alterations, which are likely selected for due to their roles in driving cancer phenotypes [61]. Towards this end, we integrated copy number and gene expression data in an attempt to identify novel oncogenes important in lung tumorigenesis. While our analysis revealed gains/ amplifications in a number of regions previously reported in NSCLC, the amplicon at 12q15 was the only one without a candidate driver gene located within the amplicon boundaries and was therefore the only regions we pursued further. Integration of expression and copy number data for the 7 genes located within 12q15 identified *YEATS4* as the candidate target gene of this amplicon.

First identified and isolated in the glioblastoma multiforme cell line TX3868, *YEATS4* is a highly conserved nuclear protein essential for cell viability that is frequently amplified in gliomas, astrocytomas and liposarcomas [154, 156, 162]. A member of a protein family characterized by the presence of an N-terminal YEATS domain, *YEATS4* shares high homology with transcription factor family members *AF-9* and *ENL* [163]. Like other family members, *YEATS4* is involved in chromatin modification and transcriptional regulation through its incorporation into multi-subunit complexes; specifically the human TIP60/TRRAP and SRCAP complexes [164, 165], which mediate the incorporation of an H2A variant histone protein into nucleosomes, altering chromatin structure and controlling transcriptional regulation.

In addition to its role in transcriptional regulation, yeast two hybrid screens have revealed a number of *YEATS4* binding partners. These include *MYC*, *MYCN*, *TACC1*, *TACC2*, *NuMa*, *AF10*, *PFDN1* and *KIAA1009* [166-170]. Analysis of expression data before and after *YEATS4* knockdown showed no effect on expression of any binding partners, suggesting that *YEATS4* does not control the expression of its binding partners at the mRNA level. To date, the majority of work surrounding *YEATS4* has focused primarily on the identification of *YEATS4* binding

partners with only a few studies having explored the phenotypic effects of *YEATS4* amplification, none of which have been performed in lung [159, 170].

Our study is the first to show gain/amplification and overexpression of *YEATS4* in NSCLC and the first to implicate amplification of *YEATS4* in lung cancer tumorigenesis. We observed frequent gain/amplification of *YEATS4* in multiple independent tumor cohorts in addition to our own, as well as a strong correlation between gain and overexpression in both tumors and cell lines (Figure 4.2). Analysis of the catalogue of somatic mutations in cancer (COSMIC) revealed YEATS4 is rarely mutated in lung (0.23%) or any cancer type (0.17%) suggesting that DNA amplification is the predominant mechanism of activation. In addition to the genomic evidence supporting selection of *YEATS4* in NSCLC, we demonstrate the oncogenic potential of *YEATS4* both *in vitro* and *in vivo* (Figures 4.3&4.4). Ectopic expression resulted in a significant reduction in senescence, suggesting overexpression of *YEATS4* is sufficient to induce phenotypic changes characteristic of malignant transformation, (Figure 4.3) while knockdown in cell lines with amplification and mutant p53 showed reduced viability and colony formation along with an increase in senescence, consistent with oncogenic function. While wt p53 abrogates the effects on viability and colony formation on *YEATS4* knockdown lines with amplification, a significant increase in senescence was still observed. In addition to these phenotypic effects, we also demonstrated that *YEATS4* inhibits p21 thereby repressing p53 activity, consistent with the findings of Park and Roeder who demonstrated this interaction in unstressed conditions [159]. siRNA-mediated knockdown of *CDKN1A* failed to rescue viability, colony formation and senescence in mutant p53 backgrounds suggesting the phenotypic effects of *YEATS4* amplification occur through a mechanism other than p21.

MDM2, an E3 ubiquitin ligase, is the major negative regulator of p53, mediating its ubiquitination and subsequent degradation [171, 172]. Overexpression results in inactivation of p53 and is a common mechanism of *p53* inactivation in cancer. *MDM2* is frequently amplified and overexpressed in human cancers including lung cancer, and is largely considered to be the driver gene of the 12q15 amplicon [18, 173]. We were therefore intrigued to discover that despite being frequently gained in our dataset, MDM2 did not fall within the boundaries of the

12q15 amplicon identified in our cohort. This led us to postulate that an alternative oncogene was being selected in this region. When looking at high resolution copy number profiles, while the majority of cases showed identical copy number for both *YEATS4* and *MDM2*, a small number of cases (3/83) had more copies of *YEATS4* than *MDM2*, suggesting *YEATS4* is selected as the target of amplification in these samples and that amplification of *YEATS4* is not merely a passenger event of *MDM2* amplification. Of note, 4/83 cases had higher level gain/amplification of *MDM2* relative to *YEATS4*. For cancers with amplification of 12q15 spanning both *YEATS4* and *MDM2*, these genes may work synergistically to suppress p53, however further experimentation is required to investigate this hypothesis. Along with the many tumor promoting effects of *YEATS4*, of immediate clinical interest is our discovery of a YEATS4 dependent mechanism of reduced cisplatin and nutlin sensitivity, which appears to occur at least in part through inhibition of p21 and subsequent suppression of the p53 pathway.

In summary, we have shown that 12q15 is frequently amplified in both AC and SqCC and suggest that YEATS4, given its multiple oncogenic functions, is a novel oncogene involved in lung tumorigenesis. Our findings imply that *MDM2* is not the sole driver gene targeted by amplification of 12q15 and that suppression of the p53 pathway can be achieved through amplification of *YEATS4* via inhibition p21. For the first time, we demonstrate that YEATS4 expression alters cisplatin and nutlin sensitivity. Additional investigation into the signaling pathways altered as a result of YEATS4 amplification will provide further insight into the mechanism underlying *YEATS4*-mediated tumorigenesis and its potential clinical relevance.

Chapter 5: Characterizing miRNA expression in lung adenocarcinoma and squamous cell carcinoma identifies miR-944 and miR-1287 as subtype specific miRNAs

5.1 Introduction

The human genome is comprised of less than 2% protein coding genes, however more than 90% of the genome is transcribed, suggesting that the majority of the transcriptome is comprised of non-coding RNAs (ncRNAs) - transcripts that lack an open reading frame and as such do not encode a protein [174-176]. This by no means implies that ncRNAs lack function, but rather highlights the importance of looking beyond protein coding genes in order to improve our knowledge of normal and disease biology. While some are known to play important roles in the regulation of gene expression, splicing, epigenetic control, chromatin structure and nuclear transport, the function of most ncRNAs remains unknown [177, 178]. miRNAs are the most thoroughly investigated class of ncRNAs and are often located at chromosomal breakpoint regions, fragile sites, and minimal regions of loss of heterozygosity or amplification, making miRNA loci highly susceptible to genomic alterations and subsequently, de-regulated expression [179, 180]. To date, over 2000 human miRNAs have been identified [181] and a single miRNA is capable of affecting multiple protein coding genes. It is believed that over one third of the genome is regulated by at least one miRNA [182]. With roles in cellular functions from proliferation and apoptosis to cellular development and epithelial to mesenchymal transition, it is not surprising that miRNA deregulation has been linked to human diseases including cancer.

MicroRNAs (miRNAs) are small 18-25 nucleotide non-coding RNAs that negatively regulate gene expression post-transcriptionally through transcript degradation or translational repression [56]. miRNAs are known to be frequently deregulated in cancer as well as pre-invasive lesions and circulate in bodily fluids (blood, sputum, urine etc.) with substantial stability, making them ideal candidates for non-invasive, early detection biomarkers. While protein coding genes remain the primary focus of current genomic and proteomic studies, the deregulation of several miRNAs (*let-7a, miR-34, -125, -126 -221, -222*, to name a few) have

been implicated in lung tumorigenesis [183-187], and in the past decade miRNA expression profiles have been associated with lung cancer prognosis, disease progression, survival, and outcome prediction, as well as subtype discrimination [58, 183, 188, 189]. However, these studies are limited by the use of microarray profiling which restricts the number of miRNAs assessed, small sample sizes, or the absence of patient matched non-malignant tissue. Moreover, expression profiling studies aimed at distinguishing differences between AC and SqCC have used minimal statistical criteria to identify differentially expressed miRNA, rarely including tumor/non-malignant fold change criteria and typically just correlating miRNA expression with clinical features. As such, an understanding of the recurrently altered miRNAs defining each subtype and how these miRNAs contribute to subtype specific tumorigenesis through the genes and pathways they affect is lacking. We hypothesize that similar to DNA level alterations, the differential deregulation of miRNAs and their consequential effects on genes and pathways may contribute to the disparate clinical phenotypes observed in AC and SqCC.

In this study we performed a genome wide, unbiased analysis of miRNA expression patterns in NSCLC subtypes. miRNA sequencing was performed on a cohort of 88 NSCLC tumors with matched non-malignant tissue (66 AC and 22 SqCC) and detailed clinical information. We identified miRNA panels capable of distinguishing tumor from non-malignant tissue (n=85) and AC from SqCC (n=47), and validated our findings in independent, external datasets. Network and pathway analysis was performed on the most robust mRNA targets of our validated miRNAs based on *in silico* target prediction algorithms. Consistent with our observations of protein coding genes (Chapter 2), pathway analysis of target genes disrupted by SqCC specific miRNAs revealed disruption of multiple histone modifying enzymes. *In vitro* analysis of *miR-1287*, an AC specific miRNA, confirmed RAD9A as a biological target and implicated *miR-1287* in AC tumorigenesis. Overexpression of *miR-1287* in HBECs and lung cancer cell lines resulted in reduced apoptosis and DNA damage repair, validating the ability of our approach to identify biologically relevant miRNAs. To our knowledge, this is the first study to 1) perform a genome wide unbiased comparison of miRNA expression profiles in AC and SqCC, and 2) identify the pathways and functions disrupted by these subtype specific miRNAs.

5.2 Methods

5.2.1 NSCLC patients and samples

88 fresh-frozen lung tumors (66AC and 22SqCC) with matched non-malignant tissue were obtained from treatment naïve patients undergoing surgical resection with curative intent (BCCRC cohort). Samples were collected from the Tumor Tissue Repository of British Columbia Cancer Agency under informed, written patient consent. Tissue sections were microdissected with the guidance of a lung pathologist to obtain >70% tumor cell content. Total RNA was extracted using Trizol reagent (Invitrogen) and then size fractionated to isolate small RNAs including miRNA.

5.2.2 MicroRNA sequencing and data analysis

miRNA sequencing libraries for the 88 fresh-frozen tumor pairs were constructed, bar-coded for multiplex sequencing and sequenced on the Illumina HiSeq 2000 platform using a plate-based protocol developed by the British Columbia Genome Sciences Center (BCGSC) [190]. Raw sequence reads were separated into individual samples based on their assigned indexes, adapter sequences removed and reads trimmed based on quality control metrics. High quality reads were aligned to the NCBI GRCh37 reference genome and miRBase v18 using the BWA algorithm. Reads were normalized as reads per kilobase of exon model per million mapped reads (RPKM).

5.2.3 Identification of subtype specific miRNAs and NSCLC miRNAs

Expression levels for identical miRNAs from different genomic locations were summed, leaving 1372 unique miRNAs for examination. miRNAs with read counts <1 were considered undetected/not expressed and miRNAs with undetectable expression in all samples within a group were excluded from further analysis, resulting in 916 miRNAs for all tumors and normals, 859 miRNA in AC and 753 miRNAs in SqCC for further analysis. The fold change in miRNA expression between tumor and normal pairs was calculated for each miRNA in each sample, and a 2-fold change in expression was used as a cutoff to define over- and underexpression. miRNAs without 2-fold over- or underexpression in any of the samples within a subtype were removed, leaving 513 miRNAs in AC and 419 miRNAs in SqCC.

To be considered subtype specific, miRNAs had to meet three criteria: i) differential expression in tumors versus non-malignant tissue, as defined by a Wilcoxon Sign Rank with Benjamini-Hochberg (BH) multiple testing correction $p<0.05$, ii) over- or underexpressed in at least 25% of samples within the subtype and a difference in frequency of aberrant expression >15% between subtypes, and iii) differential tumor expression and frequency of alteration between AC and SqCC as defined by permutation and Fisher's Exact tests with BH correction $p<0.05$. NSCLC miRNA were defined as i) differentially altered in tumors versus non-malignant tissue based on a Wilcoxon Sign Rank with Benjamini-Hochberg (BH) multiple testing correction $p<0.05$, ii) over- or underexpressed in at least 25% of samples and <15% difference in frequency between subtypes OR over-/underexpression in >50% of both subtypes, and iii) no difference in expression or frequency of alteration between AC vs SqCC, based on permutation and Fisher's Exact tests with BH correction $p>0.05$.

5.2.4 Statistical analysis of miRNA expression data

Unsupervised hierarchical clustering using Ward's method was performed on tumor samples only (n=88), and non-malignant samples only (n=88) using *Partek Genomics Suite* software. A Fisher's Exact test was performed to assess the distribution of subtypes in tumor and non-malignant profiles, with a $p<0.05$ considered significant.

Principal component analysis was performed using miRNA sequencing data for AC and SqCC tumors in *Matlab*. All 47 subtype specific miRNAs were used to generate the principal components. Receiver Operating Characteristic (ROC) area under the curve (AUC) analysis was performed in *GraphPad Prism* Software on subtype specific miRNAs to determine the ability of these miRNAs to accurately discriminate AC from SqCC ($p<0.05$).

To identify miRNAs associated with clinical features (gender, smoking status, stage and histology), a multivariate analysis of variance (MANOVA) on the 88 tumors with miRNA sequencing data was performed in *R*. A $p<0.05$ was considered significant. MANOVA results were used to confirm that the variance in subtype specific miRNA expression was in fact due to histology.

5.2.5 Validation in external cohorts

Level 3 miRNA sequencing data was obtained from the TCGA for use as an external cohort for validation. Expression profiles were processed as described for 'Level 3 data' in the TCGA data compendium (2011 Cancer Genome Atlas Network). Data was downloaded for all patients with tumor and matched non-malignant tissue, which at the time included 35 AC and 35 SqCC. Subtype specific and NSCLC miRNAs were assessed using the same criteria applied in our initial discovery cohort, and only those passing all criteria were considered validated.

5.2.6 Integration of DNA copy number and methylation with miRNA expression

Affymetrix SNP 6.0 copy number and Illumina HM27K DNA methylation profiles were available for 77 (62 AC, 15SqCC) and 58 (45 AC, 13 SqCC) cases with miRNA sequencing, respectively. miRNA copy number and methylation status were integrated with expression data in order to determine the frequency of concerted DNA and expression alterations. Array data was processed as described in Chapter 2 and standard thresholds were used to define copy gain (>2.3 copies), loss (<1.7 copies), DNA hypomethylation ($\Delta\beta$ <-0.15) and hypermethylation ($\Delta\beta$>0.15). miRNAs were considered to be controlled by genetic or epigenetic mechanisms if concerted DNA disruption (copy number or DNA methylation changes) and miRNA expression changes were observed in a concordant direction (e.g. copy loss and underexpression) in >20% of cases.

5.2.7 Target prediction, interaction networks and pathway analysis

miRNAs identified as NSCLC or subtype specific were input into the *microRNA Data Integration Portal v.2* (http://ophid.utoronto.ca/mirDIP), which integrates 13 microRNA target prediction algorithms and six microRNA prediction databases to predict miRNA-transcript (mRNA) interactions [191, 192]. 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 v.2.14* (http://ophid.utoronto.ca/navigator) [193, 194]. Interaction networks were generated for subtype specific miRNAs as well as NSCLC miRNAs disrupted in >90% of cases. Pathway analysis was performed on all predicted mRNA targets (miRTarBase v3.5) of miRNAs disrupted in a subtype specific manner using Ingenuity Pathway Analysis.

5.2.8 Cell culture and manipulation of miR-1287

Immortalized human bronchial epithelial cells courtesy of Dr. John Minna were maintained in KSFM supplemented with 5ng/µl EGF and 50ng/µl BPE. AC cell lines A549 and H1993 were obtained from the American Type Culture Collection and cultured and maintained in RPMI supplemented with 10% FBS and 0.1% penicillin-streptomycin (Invitrogen). HBECs and lung cancer cell lines were transfected with 20nM of miR-1287 mimic or a non-targeting control (Thermo Scientific) according to manufacturer's instructions. miR-1287 expression was verified by qRT-PCR 48 hours post-transfection and target gene expression by western blot 72 hours post-transfection.

5.2.9 3'-UTR reporter assays

A commercially available RAD9A 3'-UTR dual luciferase reporter plasmid (pEZX-RAD9A-UTR) was purchased from GeneCopoeia. HBECs and lung AC cell lines were transfected with miR-1287 mimic or the non-targeting control for 24 hours and then the cells were transfected with the reporter vector. Luciferase assays were performed 48 hours later using the Luc-Pair miR Luciferase Assay kit (GeneCopoeia) and luminescence was measured on a luminometer. Firefly luciferase activity was normalized to Renilla luminescence to account for differences in transfection efficiency.

5.2.10 qPCR of miRNA expression

Quantitative real-time PCR analysis of miR-1287 was performed using the miRNA-specific TaqMan MicroRNA Assay Kit (Applied Biosystems) and an Applied Biosystems 7500 Fast Real Time PCR system. 10 nanograms of sample RNA was converted to cDNA using the TaqMan MicroRNA Reverse Transcription kit and 10µg of cDNA was combined with TaqMan Universal PCR Master Mix without AmpErase Uracil N-glycosylase and miRNA specific primers as per the manufacturer's instructions (Applied Biosystems). U6 small nuclear RNA was used as an endogenous control to normalize cDNA input.

5.2.11 Western blotting

Protein lysates were collected and western blots performed as previously described [151]. Membranes were incubated with primary antibodies against RAD9A (ab70810, AbCam at 1:1000) and TEAD3 (#13224, Cell Signaling at 1:1000) and GAPDH as a loading control (#2118, Cell Signaling at 1:50,000). Following primary antibody incubation, membranes were incubated with anti-rabbit HRP conjugated secondary antibodies (Cell Signaling #7076) and visualized by enhanced chemiluminescence.

5.2.12 MTT assay

MTT assays were performed as previously described [195] to assess cell viability following miR-1287 overexpression. Cells were seeded in triplicate in 96 well plates at optimal densities for growth (1000 cells/well for A549 and 1500 cells/well for H1993 and HBECs). 48 hours post transfection, viability was measured over five consecutive days and quantified by spectrophotometry. Experiments were performed in triplicate.

5.2.13 Apoptosis, DNA damage and cell proliferation

Apoptosis, DNA damage and cell proliferation were assessed by measurement of propidium iodide (PI) and Annexin staining, phospho-γH2AX and 5-bromo-2'-deoxyuridine (BrdU) incorporation, respectively (FITC Annexin V Apoptosis Detection Kit and Apoptosis, DNA Damage and Cell proliferation kit, BD Pharmigen). Cells were seeded in 6-well plates at a density of $1.5x10^5$ for A549 and $2x10^5$ for HBEC and H1993 and 24 hours after seeding they were transfected with miR-1287 mimic or non-targeting control. 48 hours after transfection, cells were treated with 10uM cisplatin for 4 hours, stained with BrdU and then allowed to recover for 24 hours. Following recovery, cells were trypsinized, washed, and processed as per the manufacturer's instructions. Flow cytometry was performed using a FACSCanto II flow cytometer and analyzed with FACSDiva software (BD Biosciences). Results are reported as the average percent of positive staining cells ± SEM of triplicate experiments.

5.3 Results

5.3.1 MiRNAs are differentially expressed between NSCLC tumors and non-malignant tissue and accurately segregate profiles based on malignancy.

Expression levels for 1372 unique miRNAs generated by miRNA sequencing of 176 fresh-frozen NSCLC tumors and matched non-malignant tissues (66AC and 22 SqCC pairs, Table 5.1) were examined. To assess the similarity of miRNA expression profiles across tumors and matched non-malignant tissue, unsupervised hierarchical clustering was performed on 916 miRNAs with detectable expression. Clustering revealed two distinct clusters; one comprised of all tumors and the other of all non-malignant tissues and two tumor samples, demonstrating that miRNA expression strongly differentiates tumors from normals (Figure 5.1A).

Based on the accuracy of clustering, we next sought to determine which miRNAs were differentially expressed between tumors and matched non-malignant tissue. 215 miRNAs met our criteria of being differentially altered in tumors vs. matched non-malignant tissue (Sign Rank with Benjamini-Hochberg correction $p<0.05$), and aberrantly expressed (2 fold or greater between tumor and normal) in at least 25% of samples. However, only 85 miRNAs had no significant difference in expression or frequency of alteration between subtypes (Fishers Exact test, BH $p>0.05$) and were deemed "true" NSCLC miRNAs. 18 miRNAs (12 OE, 6 UE, Table 5.2) were altered in greater than 90% of all tumors, and principal component analysis of these miRNAs confirmed their ability to accurately discriminate tumors from non-malignant tissue (Figure 5.1B), highlighting the potential of these 18 miRNAs as biomarkers for NSCLC detection. Interestingly, PCA revealed tumors cluster more closely together than non-malignant tissues, suggesting there is greater inter-individual variance in miRNA expression in non-malignant tissues than tumors (Figure 5.1B).

Figure 5.1

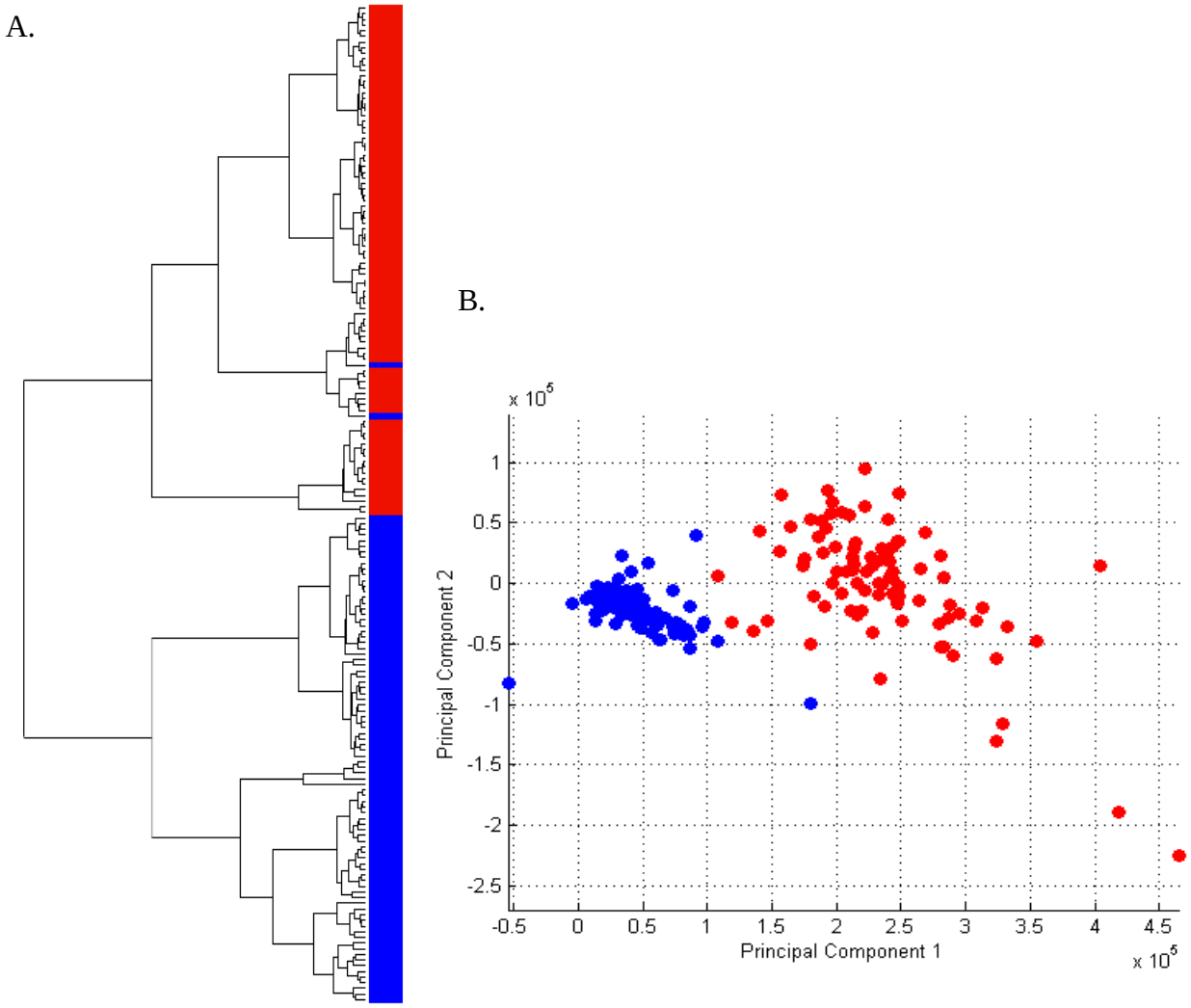


Figure 5.1. miRNA expression profiles accurately segregate tumor and non-malignant tissue

Clustering of 176 miRNA expression profiles (66 AC and 22 SqCC with matched non malignant tissue) revealed two distinct clusters associated with malignancy, one comprised of tumor samples (blue) and the other of all non-malignant tissues (red) and two tumor samples (A). Principal component analysis using expression of the 18 miRNAs altered in >90% of all cases accurately separated tumor and non-malignant tissue (B). Blue dots represent non-malignant tissue while red dots represent tumor samples.

Table 5.1 Patient demographics for all cases with miRNA sequencing data

Feature	AC (n=66)	SqCC (n=22)	Total (n=88)
Stage			
I	41(62.1%)	7 (31.8%)	48 (54.5%)
II	17 (25.8%)	10 (45.5%)	27 (30.7%)
III	5 (7.6%)	4 (13.6%)	8 (9.1%)
IV	2 (3.0%)	1 (4.5%)	3 (3.4%)
Sex			
Female	44 (66.7%)	7 (31.8%)	51 (58%)
Male	22(33.3%)	15 (68.2%)	37 (42%)
Average Age	67 (45-90)	71 (58-86)	68 (45-90)
Smoking Pack Years	47	48.6	47.4
Years Quit	5.5	10.2	6.64

Table 5.2 Alteration frequencies of the most frequently deregulated NSCLC miRNAs

		AC		SqCC		All Samples	
miRNA	**Alteration**	**Freq OE**	**Freq UE**	**Freq OE**	**Freq UE**	**Freq OE**	**Freq UE**
hsa-mir-210	OE	100%	0%	95%	0%	99%	0%
hsa-mir-96	OE	98%	0%	100%	0%	99%	0%
hsa-mir-130b	OE	95%	0%	100%	0%	97%	0%
hsa-mir-183	OE	94%	0%	100%	0%	95%	0%
hsa-mir-345	OE	94%	2%	95%	0%	94%	1%
hsa-mir-877	OE	94%	2%	95%	0%	94%	1%
hsa-mir-331	OE	94%	0%	91%	5%	93%	1%
hsa-mir-182	OE	94%	0%	86%	0%	92%	0%
hsa-mir-708	OE	89%	0%	100%	0%	92%	0%
hsa-mir-141	OE	92%	0%	86%	0%	91%	0%
hsa-mir-193b	OE	89%	3%	95%	0%	91%	2%
hsa-mir-301b	OE	89%	0%	95%	0%	91%	0%
hsa-mir-144	UE	0%	98%	0%	100%	0%	99%
hsa-mir-30a	UE	0%	97%	0%	100%	0%	98%
hsa-mir-451a	UE	0%	97%	0%	91%	0%	95%
hsa-mir-143	UE	0%	97%	0%	82%	0%	93%
hsa-mir-486	UE	0%	94%	0%	86%	0%	92%
hsa-mir-101	UE	0%	89%	0%	95%	0%	91%

5.3.2 Lung AC and SqCC display different patterns of miRNA expression

AC and SqCC develop through distinct patterns of genomic alterations affecting unique protein coding genes [146]. Therefore, we next sought to investigate whether miRNAs display subtype specific patterns of expression that can discriminate subtypes. Unsupervised hierarchical clustering of tumor and non malignant tissue independently revealed similar results; one of the clusters contained primarily AC samples, while the other cluster contained a mix of both AC and SqCC (Figure 5.2A-B). The distribution of AC and SqCC was significantly different between clusters in both tumors (Fisher's Exact test, $p=0.0063$) and non-malignant tissue ($p=0.0097$) suggesting that histological subtypes influence miRNA expression in tumors and non-malignant lung tissue. Analysis of AC and SqCC non-malignant tissue revealed 17 differentially expressed miRNAs (8 OE and 9 UE, permutation test BH $p<0.05$ and fold change >2), several of which have been implicated in cancer and two (*miR-107* and *miR-429*) in NSCLC [196, 197].

To identify miRNAs most likely involved in subtype specific tumorigenesis, we applied a series of statistical tests and fold change criteria to the 916 miRNAs with detectable expression. 47 miRNAs met all of our criteria and were deemed subtype specific; 12 OE and 1 UE in AC and 25 OE and 4 UE in SqCC (Table 5.3). The majority of subtype specific miRNAs were overexpressed (37 vs. 5) and SqCC specific (29 vs. 13, Figure 5.2C&D). The increased number of SqCC specific miRNAs relative to AC specific miRNAs mirrors the findings in Chapter 2, in which a greater number of SqCC specific alterations were identified. In addition, we identified 5 miRNAs (*miR-203, -326, -375,-378a* and *-4662a*) with bi-directional disruption (OE in one subtype and UE in the other or vice versa), suggesting that while these miRNAs are disrupted in both subtypes, they display subtype specific patterns of alteration. Area under the curve (AUC) analysis of individual miRNAs revealed 43/47 (91%) subtype specific miRNAs were able to distinguish between subtypes ($p<0.05$, Table 5.3), with 13 miRNAs having an AUC >0.8. Two miRNAs, miR-944 (OE in SqCC), and miR-375 (OE in AC and UE in SqCC), had AUCs over 0.9 (0.9669 and 0.9001 respectively) suggesting that these miRNAs may be the most robust in discriminating subtypes (Figure 5.2 E&F). Interestingly, 4 of the 5 miRNAs with the highest AUCs (*miR-944, -30b, -152* and *-4652*) were all overexpressed in SqCC indicating that miRNAs overexpressed in SqCC may be the most useful for differentiating subtypes.

Figure 5.2

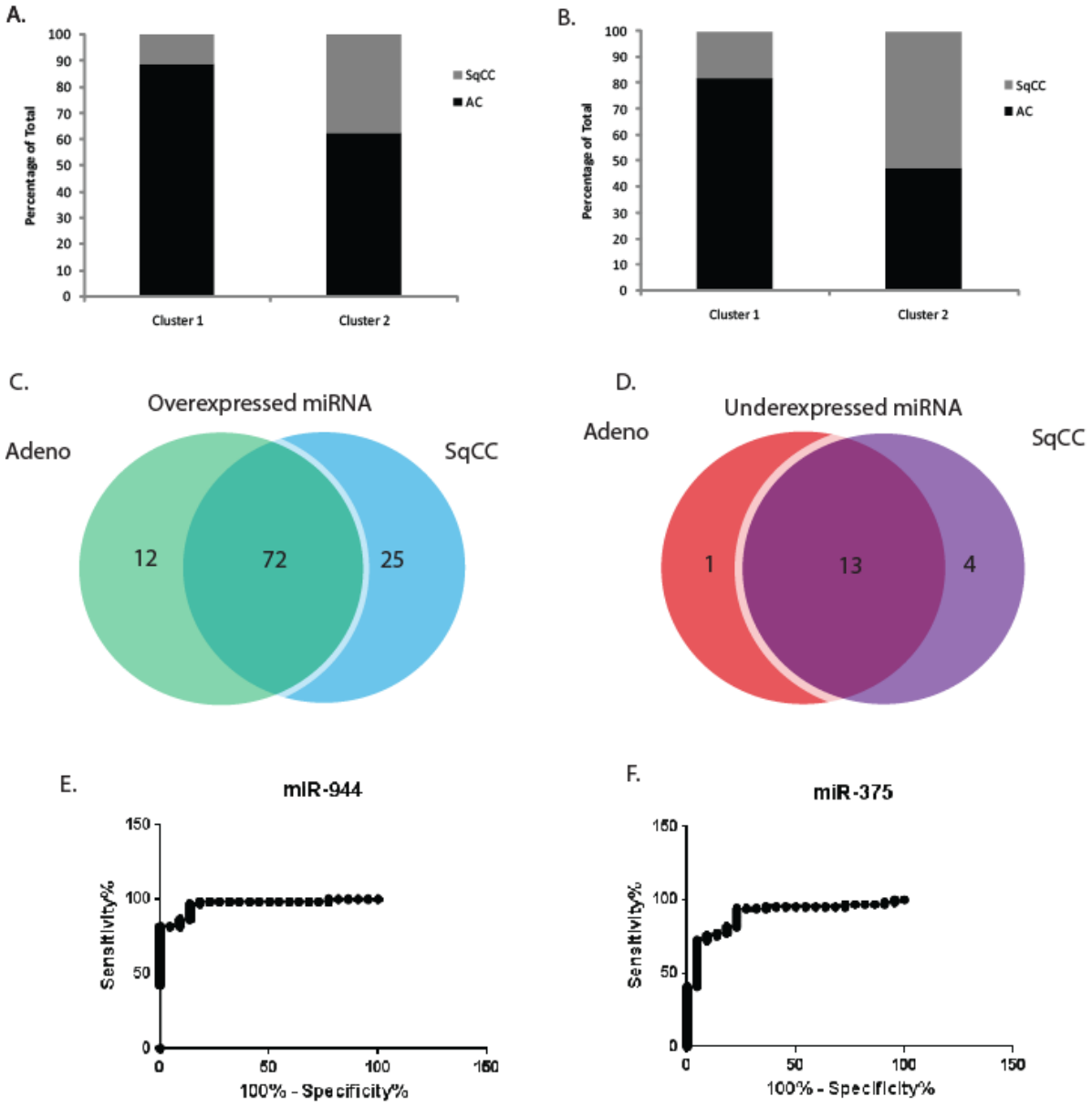


Figure 5.2. miRNAs are disrupted in a subtype specific manner that can discriminate between subtypes
Hierarchical clustering of tumors (A) and non-malignant tissue (B) identifies two clusters, one comprised of primarily AC and the other of a mix of AC and SqCC. Venn diagram illustrating the differentially expressed miRNA in AC and SqCC relative to matched non-malignant tissue. Overexpressed miRNA are depicted in (C) and underexpressed miRNA in (D). AUC curves of miR-944 (E) and miR-375 (F), the two subtype specific miRNA most accurate at discriminating AC and SqCC tumors.

Table 5.3 AUC results of all BCCRC subtype specific miRNA

miRNA	Alteration	AUC	pValue
hsa-mir-1251	AC OE	0.6873	0.008792
hsa-mir-1287	AC OE	0.7989	<0.0001
hsa-mir-26b	AC OE	0.7018	0.00477
hsa-mir-3189	AC OE	0.8695	<0.0001
hsa-mir-320b	AC OE	0.7163	0.002492
hsa-mir-320d	AC OE	0.8457	<0.0001
hsa-mir-4724	AC OE	0.7924	<0.0001
hsa-mir-4728	AC OE	0.8526	<0.0001
hsa-mir-489	AC OE	0.7225	0.001865
hsa-mir-491	AC OE	0.7211	0.001991
hsa-mir-551a	AC OE	0.7255	0.001865
hsa-mir-5698	AC OE	0.8354	<0.0001
hsa-mir-204	AC UE	0.6784	0.0126
hsa-mir-326	AC OE/SqCC UE	0.6419	0.04719
hsa-mir-375	AC OE/SqCC UE	0.9001	<0.0001
hsa-mir-203	AC UE/SqCC OE	0.7287	0.001386
hsa-mir-378a	AC UE/SqCC OE	0.7658	0.000202
hsa-mir-4662a	AC UE/SqCC OE	0.8119	<0.0001
hsa-mir-105	SqCC OE	0.8233	<0.0001
hsa-mir-1227	SqCC OE	0.836	<0.0001
hsa-mir-1237	SqCC OE	0.7803	<0.0001
hsa-mir-1295a	SqCC OE	0.6798	0.01193
hsa-mir-1343	SqCC OE	0.7107	0.003206
hsa-mir-152	SqCC OE	0.8691	<0.0001
hsa-mir-15b	SqCC OE	0.7679	0.00018
hsa-mir-1910	SqCC OE	0.7321	0.001172
hsa-mir-3619	SqCC OE	0.6102	0.1232
hsa-mir-3682	SqCC OE	0.7417	0.000724
hsa-mir-378b	SqCC OE	0.8281	<0.0001
hsa-mir-4488	SqCC OE	0.6674	0.01924
hsa-mir-4497	SqCC OE	0.6921	0.0072
hsa-mir-4652	SqCC OE	0.8561	<0.0001
hsa-mir-4713	SqCC OE	0.7314	0.001212
hsa-mir-4778	SqCC OE	0.6646	0.02131
hsa-mir-4787	SqCC OE	0.7094	0.003411
hsa-mir-651	SqCC OE	0.6784	0.0126
hsa-mir-663a	SqCC OE	0.6977	0.005702
hsa-mir-767	SqCC OE	0.8058	<0.0001
hsa-mir-873	SqCC OE	0.7066	0.003858
hsa-mir-876	SqCC OE	0.7163	0.002492
hsa-mir-93	SqCC OE	0.6236	0.08374
hsa-mir-942	SqCC OE	0.5634	0.3754
hsa-mir-944	SqCC OE	0.9669	<0.0001
hsa-mir-135a	SqCC UE	0.7693	0.000166
hsa-mir-181d	SqCC UE	0.5909	0.2034
hsa-mir-30b	SqCC UE	0.75	0.000473
hsa-mir-338	SqCC UE	0.6908	0.007626

5.3.3 Validation of differentially expressed miRNAs

Although a number of miRNA profiling studies have been performed in lung cancer, miRNA sequencing data is still quite limited. In order to validate the findings from our study and ensure we proceed with only the most robust targets, we downloaded miRNA sequencing data for 219 lung AC and 367 SqCC from the Cancer Genome Atlas (TCGA). We limited the validation cohort to only those tumors with patient matched non-malignant tissue and current or former smokers in order to replicate our original analysis; resulting in 33AC and 35 SqCC pairs for validation. Of our 85 NSCLC miRNAs, 76 were measurable in TCGA cohorts and 51% of these were validated as common to NSCLC. Of the 18 miRNAs that were altered in >90% of all NSCLC cases, 16 (89%) validated in the TCGA, with the other 2 (*miR-301b*, and *-101*) identified as subtype specific miRNAs in SqCC. Impressively, of the 16 miRNAs that validated as NSCLC miRNAs, all were altered in >70% of cases, and 3 (*miR-210, -130, -183*) were altered in >90%, suggesting this panel of miRNAs are robustly and recurrently altered in NSCLC (Table 5.4). Despite identifying a greater number of subtype specific miRNAs in the TCGA set (63 vs. 47), likely because of the increased SqCC cohort size, subtype specific miRNA validation in the TCGA dataset was surprisingly low, with only 27% of the 33 assessed miRNAs validating. All miRNAs, both subtype specific and NSCLC (n=48) that validated in the TCGA dataset are shown in Table 5.4.

Table 5.4 Subtype specific and NSCLC miRNAs that validated in the TCGA cohort

		Frequency of Alteration in BCCRC				Frequency of Alteration in TCGA			
miRNA	**Alteration**	**AC OE**	**AC UE**	**SqCC OE**	**SqCC UE**	**AC OE**	**AC UE**	**SqCC OE**	**SqCC UE**
miR-1287	AC OE	50.0%	1.5%	9.1%	9.1%	48.5%	0%	11.4%	22.9%
miR-326	AC OE/SqCC UE	30.3%	12.1%	9.1%	50.0%	60.6%	6.1%	5.7%	65.7%
miR-375	AC OE/ SqCC UE	65.2%	6.1%	13.6%	68.2%	69.7%	9.1%	8.6%	71.4%
miR-1227	SqCC OE	22.7%	1.5%	59.1%	0%	9.1%	0%	42.9%	0%
miR-1910	SqCC OE	15.2%	0%	54.5%	0%	0%	0%	42.9%	0%
miR-651	SqCC OE	30.3%	15.2%	68.2%	0%	33.3%	0%	57.1%	11.4%
miR-944	SqCC OE	31.8%	12.1%	95.5%	0%	39.4%	12.1%	85.7%	8.6%
miR-1295a	SqCC OE	12.1%	0%	40.9%	0%	9.1%	0%	45.7%	0%
miR-338	SqCC UE	1.5%	66.7%	0%	100.0%	0%	54.5%	2.9%	85.7%
let-7c	NSCLC UE	1.5%	72.7%	0%	50.0%	0%	63.6%	2.9%	60.0%
miR-1247	NSCLC UE	3.0%	77.3%	0%	59.1%	9.1%	63.6%	11.4%	51.4%
miR-126	NSCLC UE	4.5%	50.0%	9.1%	40.9%	0%	51.5%	5.7%	68.6%
miR-144	NSCLC UE	0.0%	98.5%	0%	100.0%	0%	87.9%	2.9%	80.0%
miR-223	NSCLC UE	1.5%	74.2%	0%	50.0%	3.0%	36.4%	8.6%	42.9%
miR-451a	NSCLC UE	0%	97.0%	0%	90.9%	3.0%	81.8%	2.9%	85.7%
miR-486	NSCLC UE	0%	93.9%	0%	86.4%	0%	81.8%	2.9%	85.7%
miR-1180	NSCLC OE	40.9%	10.6%	27.3%	0%	51.5%	0%	51.4%	0%
miR-1301	NSCLC OE	87.9%	0%	86.4%	0%	78.8%	0%	71.4%	0%
miR-130b	NSCLC OE	95.5%	0%	100.0%	0%	93.9%	0%	97.1%	0%
miR-141	NSCLC OE	92.4%	0%	86.4%	0%	84.8%	0%	71.4%	2.9%
miR-148a	NSCLC OE	54.5%	1.5%	72.7%	0%	45.5%	0%	31.4%	8.6%
miR-17	NSCLC OE	43.9%	0%	54.5%	0%	48.5%	0%	40.0%	0%
miR-18a	NSCLC OE	68.2%	0%	77.3%	0%	45.5%	6.1%	57.1%	2.9%
miR-193b	NSCLC OE	89.4%	3.0%	95.5%	0%	84.8%	0%	88.6%	2.9%
miR-199a	NSCLC OE	50.0%	3.0%	59.1%	0%	39.4%	0%	31.4%	8.6%
miR-199b	NSCLC OE	40.9%	3.0%	50.0%	9.1%	42.4%	0%	31.4%	5.7%
miR-200c	NSCLC OE	63.6%	0%	68.2%	0%	45.5%	0%	54.3%	0%
miR-20a	NSCLC OE	34.8%	3.0%	45.5%	0%	45.5%	6.1%	42.9%	0%
miR-214	NSCLC OE	83.3%	3.0%	90.9%	4.5%	45.5%	6.1%	31.4%	11.4%
miR-219	NSCLC OE	57.6%	6.1%	68.2%	0%	51.5%	3.0%	57.1%	0%
miR-301a	NSCLC OE	84.8%	0%	81.8%	0%	78.8%	0%	71.4%	0%
miR-3127	NSCLC OE	59.1%	3.0%	59.1%	0%	81.8%	0%	71.4%	0%
miR-324	NSCLC OE	80.3%	0%	81.8%	0%	72.7%	0%	57.1%	2.9%
miR-335	NSCLC OE	31.8%	4.5%	40.9%	0%	33.3%	9.1%	40.0%	5.7%
miR-3607	NSCLC OE	75.8%	1.5%	59.1%	9.1%	93.9%	0%	57.1%	8.6%
miR-3677	NSCLC OE	45.5%	7.6%	54.5%	4.5%	69.7%	0%	71.4%	2.9%
miR-4326	NSCLC OE	77.3%	0%	90.9%	0%	48.5%	0%	62.9%	0%
miR-484	NSCLC OE	56.1%	1.5%	59.1%	0%	30.3%	3.0%	34.3%	5.7%
miR-505	NSCLC OE	60.6%	1.5%	72.7%	0%	81.8%	0%	54.3%	2.9%
miR-625	NSCLC OE	60.6%	4.5%	77.3%	0%	75.8%	0%	54.3%	2.9%
miR-654	NSCLC OE	42.4%	19.7%	50.0%	9.1%	48.5%	6.1%	40.0%	5.7%
miR-744	NSCLC OE	68.2%	0%	77.3%	0%	57.6%	0%	65.7%	0%
miR-758	NSCLC OE	40.9%	15.2%	54.5%	4.5%	42.4%	3.0%	31.4%	8.6%
miR-760	NSCLC OE	65.2%	0%	0%	77.3%	57.6%	0%	85.7%	0%
miR-766	NSCLC OE	77.3%	1.5%	81.8%	0.0%	81.8%	0%	57.1%	5.7%
miR-874	NSCLC OE	69.7%	3.0%	90.9%	4.5%	51.5%	0%	45.7%	8.6%
miR-877	NSCLC OE	93.9%	1.5%	95.5%	0%	66.7%	0%	74.3%	0%
miR-940	NSCLC OE	80.3%	3.0%	100.0%	0%	66.7%	0%	74.3%	0%

5.3.4 Copy number alterations contribute to aberrant miRNA expression

To determine whether differential expression of any of the validated miRNAs could be the result of recurrent DNA level alterations, we examined copy number status and DNA methylation levels of the 48 differentially expressed miRNAs. Due to the lack of miRNAs represented on the HM27K Illumina Methylation platform (only 110 miRNAs have probes), only 4 of our validated miRNAs (*miR-17, -18a, 20a* and *-219*) were measurable, and none were frequently hypo- or hypermethylated in any of the cases. 62/66 AC and 15/22 SqCC samples with miRNA sequencing data also had copy number profiles. A number of the validated miRNAs mapped to similar genomic locations, including 1q24.3, 12p13.31, 13q31.3, 14q32.31, 16p13 and 17q11.2, of which 1q, 14q32.31 and 17q11.2 are known to be recurrently amplified in NSCLC [37-39]. 9 miRNAs (*miR-130b, -141, -148a, -301a, -484, -651, -940, -944* and *-4326*) had copy number alterations and concordant miRNA expression changes affecting >20% of cases, however, only *miR-141* and *miR-301a* had a statistically significant association between copy number and expression (Mann Whitney U test, $p<0.05$). Together, our results suggest that while copy number alterations occur, they are not the driving force of over- or underexpression for the majority of the NSCLC and subtype specific miRNAs we identified.

5.3.5 Predicted targets of subtype miRNAs are associated with distinct pathways

In an attempt to elucidate the signaling pathways and biological processes disrupted by our validated subtype specific miRNAs, we identified their predicted mRNA targets using miRDIP with stringent filtering criteria (prediction by 6 algorithms). The 5 OE SqCC miRNAs were predicted to target 849 unique genes, including known tumor suppressors and oncogenes *HIF1a*, *EYA4*, *BCL2* and *IGF1R*. Visualization of miRNA-mRNA interactions in *NAViGaTOR* showed a number of mRNAs to be targets of multiple miRNAs, with 14 predicted to be targets of 4 SqCC miRNAs (Figure 5.3). Pathway analysis of all SqCC predicted targets revealed enrichment of target genes in the epithelial to mesenchymal transition, molecular mechanisms of cancer, TGFB, PTEN and VEGF signaling pathways as well as post-translational modification, cell morphology and cellular movement and DNA replication, recombination and repair functions. The networks of these functions were centered around CUL3, JNK and histone proteins, respectively.

Interestingly, miRNA-mRNA interactions of the two validated, bi-directionally altered miRNAs *miR-326* and *miR-375*, revealed a large number (37/112) and significant enrichment of genes involved in cellular fate and organization (Fisher's Exact Test, $p=2.2x10^{-12}$, Figure 5.4). Ingenuity Pathway Analysis of the 112 target genes further supported this finding, with tissue development, cell morphology, assembly and organization and cellular function and maintenance ranking as the top functions of these genes. These findings suggest that *miR-326* and *miR-375* have differing roles within AC and SqCC and may play a role in maintaining cell function and lineage, providing insight into the reason for their differential disruption.

5.3.6 Lung cancer prognostic genes are targeted by NSCLC miRNAs

To assess the prognostic implications of the NSCLC miRNAs altered in >90% of cases, we aligned the list of predicted mRNA targets for each miRNA from miRDIP to a curated list of 1,066 lung cancer prognostic genes [198]. Of the 1,066 prognostic genes, 42 (4%) were predicted targets of the 16 validated miRNAs (Figure 5.5). The majority of miRNAs were highly connected to multiple prognostic genes. The RNA bindings protein *QKI* which regulates pre-mRNA splicing and mRNA stability was the most targeted mRNA, predicted to be a target of 5 unique miRNAs, highlighting the potential importance of this gene to NSCLC (Figure 5.6). Conversely, *miR-141* and *miR-182*, both of which are frequently overexpressed were connected to the most prognostic genes (n=9). This data emphasizes the biological and prognostic relevance of these NSCLC miRNAs and further supports the potential of these miRNAs as biologically based biomarkers for lung cancer detection.

Figure 5.3

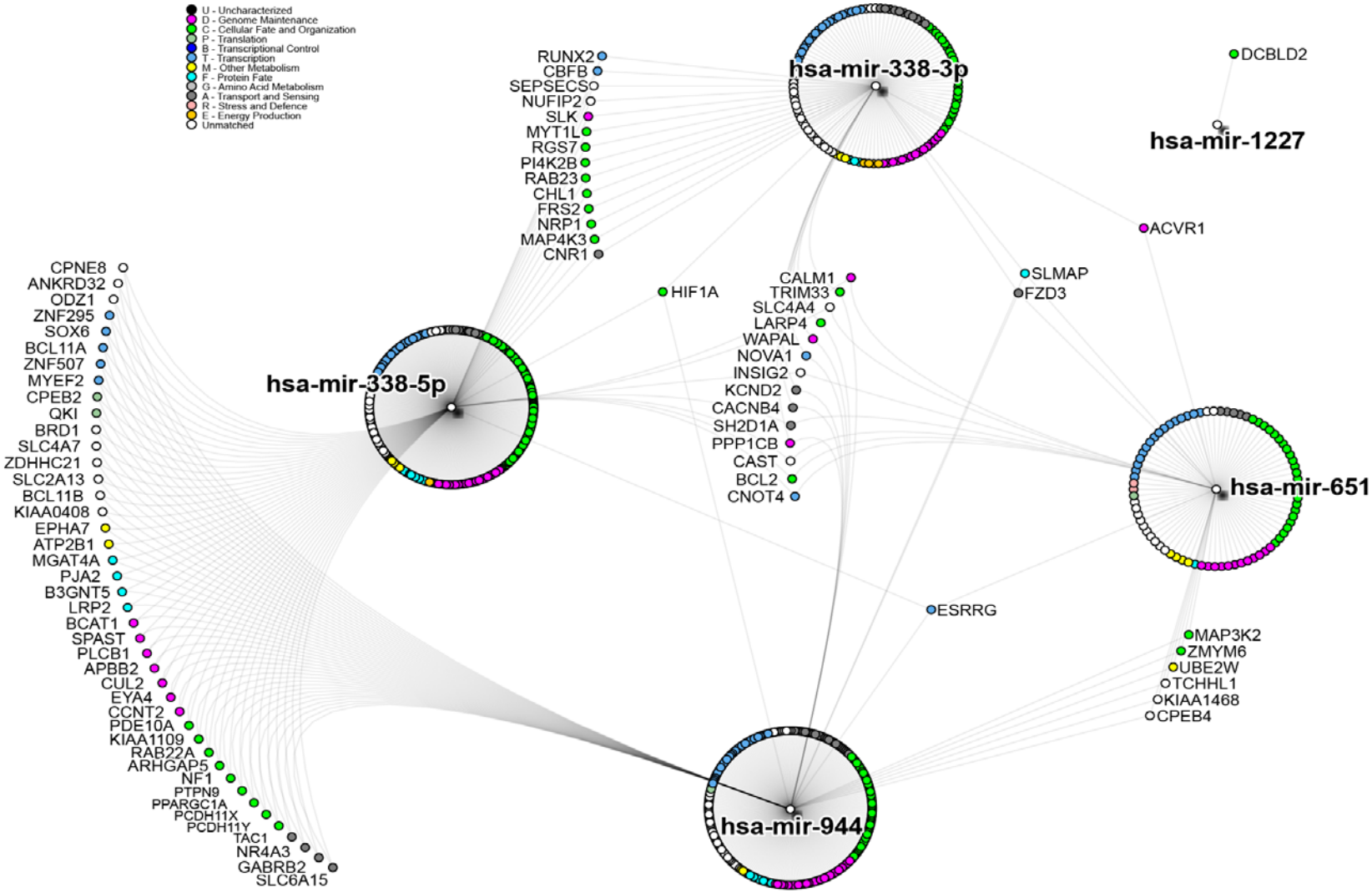


Figure 5.3 miRNA-mRNA interaction network of overexpressed SqCC specific miRNA

Figure 5.3 miRNA-mRNA interaction networks of overexpressed SqCC specific miRNAs

Predicted mRNA targets of overexpressed SqCC specific miRNAs were identified using miRDIP and miRNA-mRNA interactions visualized in NAViGaTOR. Grey lines indicate miRNA-mRNA interactions and predicted targets are depicted as circular nodes with colouring corresponding to Gene Ontology terms associated with gene function. SqCC specific miRNA were predicted to target 849 unique genes, but only those genes targeted by multiple miRNAs are displayed.

Figure 5.4

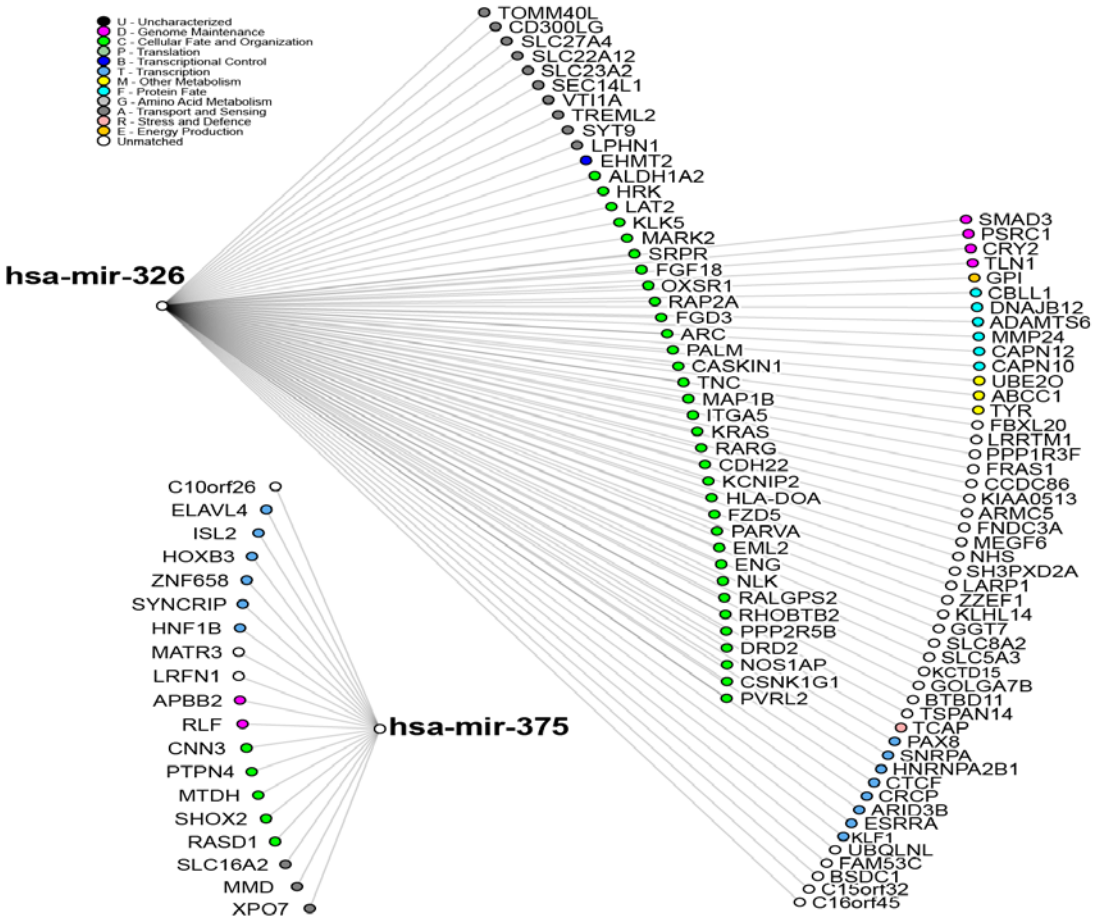


Figure 5.4. Bi-directional miRNAs target genes involved in cellular fate and organization

miRNA-mRNA interactions of the bi-directionally altered miRNA miR-326 and miR-375. Target genes are coloured based on their Gene Ontology functions and were significantly enriched for genes involved in cellular fate and organization (Fisher's test $p<0.0001$).

Figure 5.5

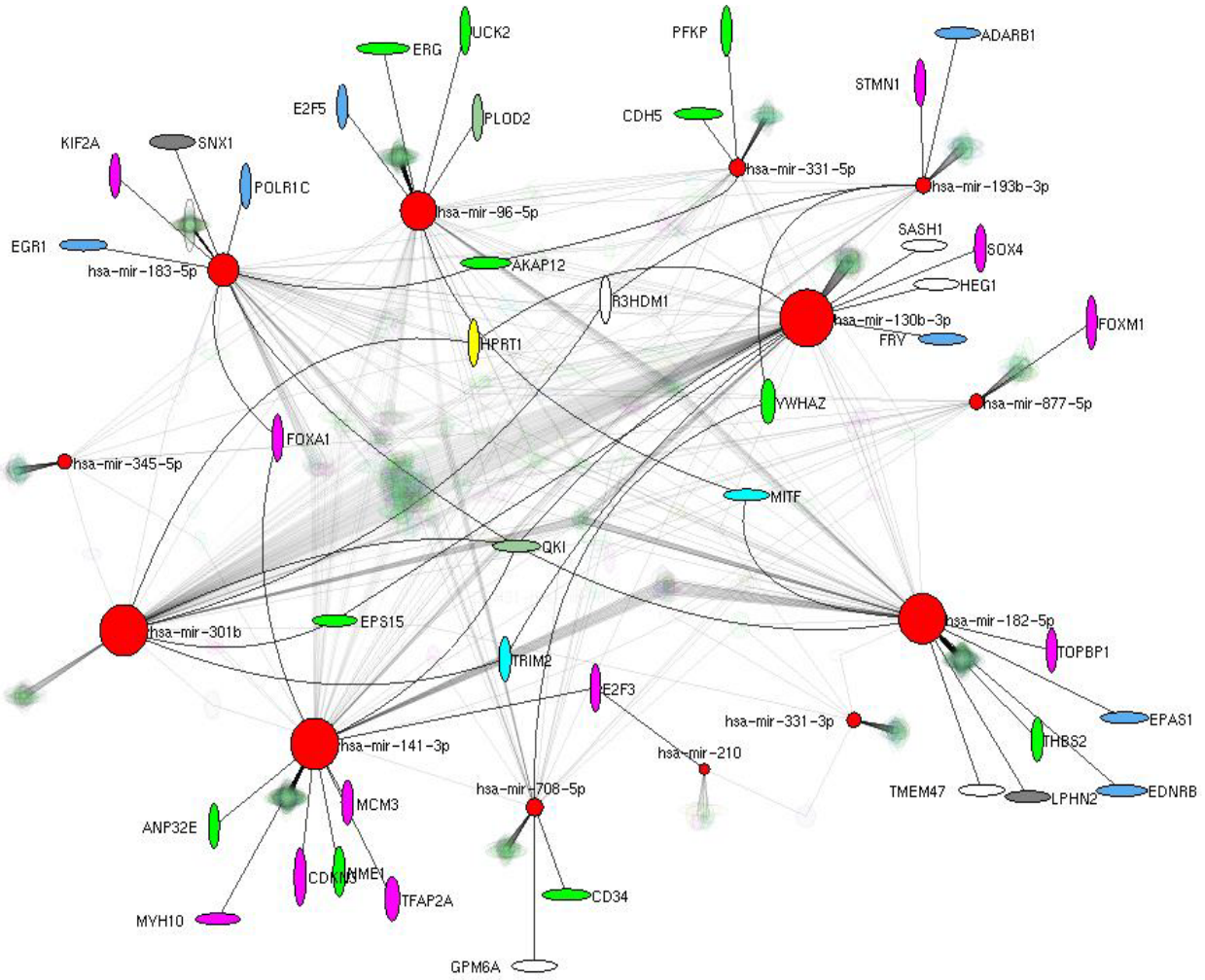


Figure 5.5 Predicted interaction between NSCLC miRNAs and lung cancer prognostic genes

miRNAs frequently disrupted in NSCLC were input into miRDIP to identify predicted targets. The network of miRNA-mRNA interactions was visualized in NAViGaTOR, but restricted to those targets with known prognostic significance in lung cancer (n=1066). In total, the network is comprised of 13 miRNAs and 42 prognostic target genes with most miRNAs well connected to prognostic genes. *miR-141* and *miR-182* are the most highly connected to prognostic gene targets. Connections between miRNAs and mRNAs are illustrated by grey lines. Predicted targets are depicted as elliptical nodes with colouring corresponding to Gene Ontology terms associated with gene function. Vertical mRNA nodes represent genes frequently upregulated in the signature while wide nodes represent downregulated genes. Fuzzy green clusters depict all other predicted mRNA targets not associated with prognostic genes.

5.3.7 MiR-944 overexpression is highly specific to SqCC

Of the miRNAs regulated by copy number, *miR-944* was the most frequently altered; gained and overexpressed in over 50% of SqCC cases and overexpressed without any known DNA alteration in another 40% of cases. *miR-944* is lowly expressed in both AC and SqCC non-malignant tissue (1.35 ± 0.23 RPKM in AC and 1.42 ± 0.34 RPKM in SqCC, Student's t test p=0.87), with no significant difference in expression between AC tumors and non-malignant tissue (7.7 ± 3.99 RPKM, Student's t- test, p=0.11, Figure 5.6A). However, in SqCC tumors, *miR-944* expression is dramatically increased (258.8 ± 75.13 RPKM,) relative to both AC tumors (Student's t-test p=0.0014) and SqCC non-malignant tissue (Student's t-test p<0.001, Figure 5.6A), with an average fold change of 172.8. A similar pattern of expression was seen in samples from the TCGA (Figure 5.6B). Given this dramatic difference in expression, it is not surprising *miR-944* was the best miRNA to accurately distinguish between subtypes (BCCRC: AUC =0.967, p<0.0001, Figure 5.2E, TCGA: AUC=0.858, p<0.001, data not shown).

While *miR-944* may be a useful marker for subtype discrimination given its subtype specificity and frequent overexpression in SqCC, we were curious as to the target genes and pathways predicted to be disrupted by this miRNA, and the contribution of *miR-944* to SqCC tumorigenesis. *miR-944* was predicted by miRDIP to target 305 unique mRNAs, including the tumor suppressors *APC*, *EYA4* and *NF1*. Pathway analysis of *miR-944* predicted targets revealed multiple networks associated with cellular development and cell death and survival, that were centered around ERK signaling, although ERK itself was not a predicted target (Figure 5.6C). miR-944 targets were also significantly enriched for the immune related functions of NF-κB and TGFβ signaling. The high frequency of *miR-944* overexpression and the predicted disruption of key signaling pathways suggests that *miR-944* may play a pivotal role SqCC tumorigenesis

Figure 5.6

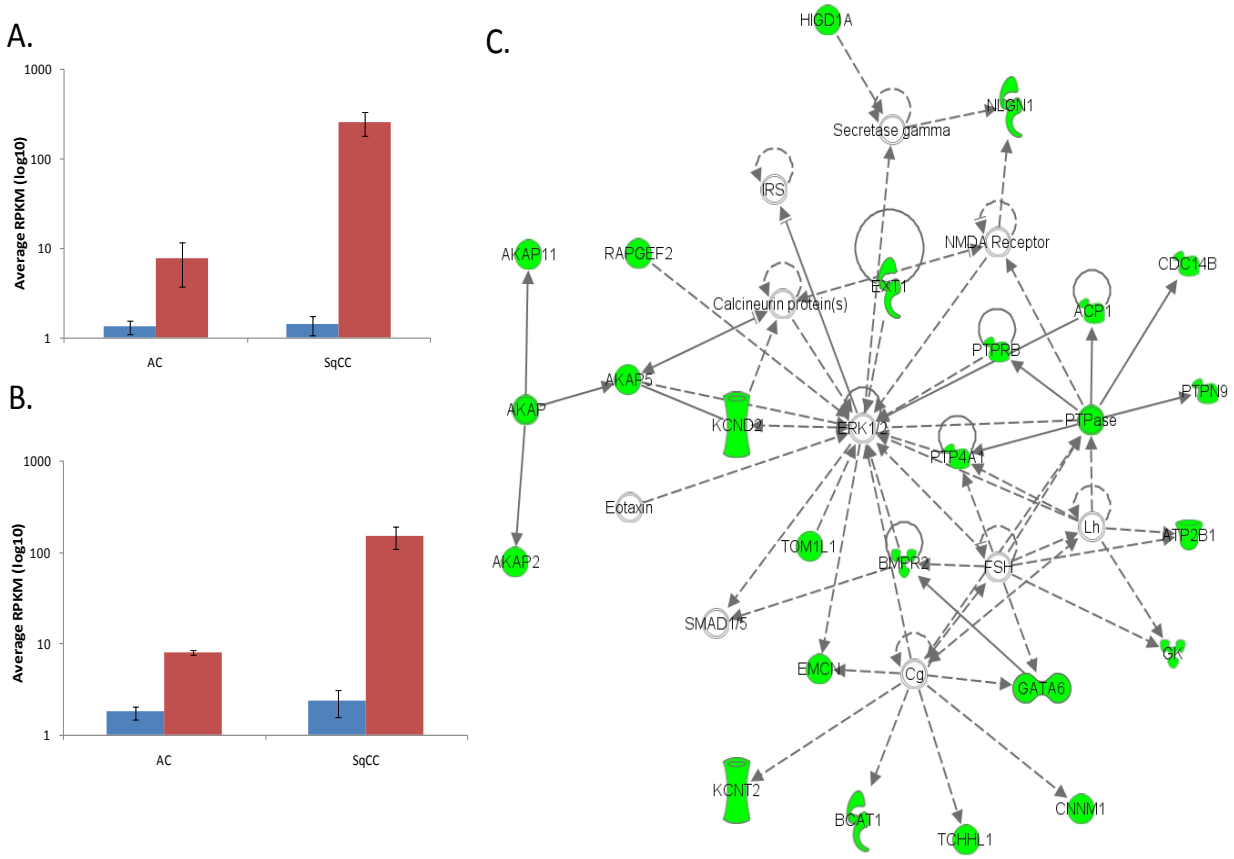


Figure 5.6 miR-944 overexpression is specific to SqCC and targets ERK signaling.
miR-944 expression in AC and SqCC tumors and non-malignant tissue in the BCCRC (A) and TCGA (B) cohorts. Blue bars represent non-malignant tissue and red bars tumors. Error bars depict the standard error of the mean. (C) ERK signaling components are downregulated (shown in green) by the predicted targets of miR-944. Solid lines represent direct interactions while dotted lines represent indirect gene interactions. Corkscrews represent enzymes, three pronged shapes are kinases, thimble shapes are transporters, cylinders are ion channels and circular molecules encompass all other gene products.

5.3.8 MiR-1287 inhibits expression of RAD9A by binding its 3'-UTR

miR-1287 which was overexpressed in 50% of AC, was the only AC specific miRNA to validate in the TCGA. Target prediction of *miR-1287* identified 5 predicted targets, 3 of which (*RAD9A*, *TEAD3* and *SLC8A1*) were targeted exclusively in AC. Both *RAD9A*, a component of the 9-1-1 cell cycle checkpoint complex which plays a central role in DNA damage repair and cell cycle arrest, and *TEAD3*, a transcription factor involved in the Hippo signaling pathway have been implicated in tumorigenesis. Transfection of *miR-1287* mimic oligonucleotides significantly reduced RAD9A but not TEAD3 protein levels compared to non-targeting control oligonucleotides (Figure 5.7A). To determine whether RAD9A is a direct target of *miR-1287*, a RAD9A 3'UTR dual luciferase reporter construct was co-transfected into cell lines with *miR-1287* mimic or control oligonucleotides and luciferase activity measured 48 hours post-transfection. Overexpression of miR-1287 significantly reduced luciferase activity relative to the control olignonucleotides, confirming *miR-1287* directly inhibits RAD9A (Figure 5.7B). Taken together, these results strongly suggest that *miR-1287* post-transcriptionally inhibits RAD9A by directly binding the 3'-UTR of RAD9A mRNA and preventing translation.

5.3.9 Overexpression of miR-1287 enables proliferation of damaged cells in LCCL

To investigate whether overexpression of *miR-1287* is biologically relevant to the development of a malignant phenotype, we overexpressed *miR-1287* in HBEC, A549 and H1993 cell lines (Figure 5.7C). In HBECs, treatment of transfected cells with 10uM of cisplatin for four hours reduced proliferation and apoptosis, suggesting these cells may be undergoing senescence (Figure 5.7D and E). Conversely, overexpression in LCCLs had the opposite effect, inducing apoptosis and increasing proliferation relative to controls (Figure 5.7D and E). A greater proportion of miR-1287 OE cells were positive for both phospho γH2AX and BrdU in LCCLs, suggesting that in these lines, miR-1287 inhibits DNA damage repair through suppression of RAD9A, enabling proliferation of damaged cells. Induction of phenotypic changes in non-malignant HBECs and AC cell lines suggests that overexpression of *miR-1287* is associated with malignant changes.

Figure 5.7

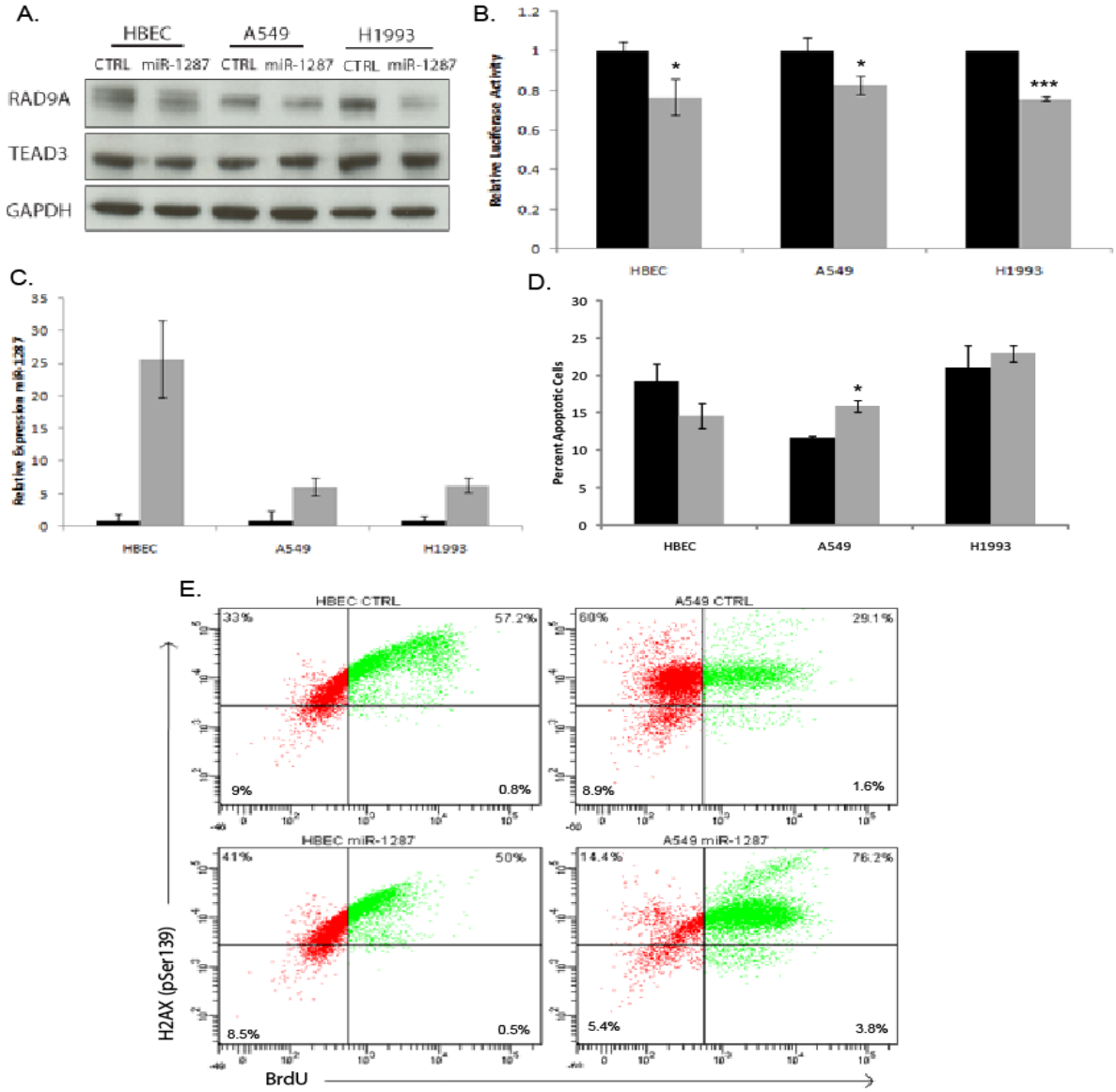


Figure 5. 7. miR-1287 induces a malignant phenotype through the suppression of RAD9A

(A) Western blot analysis of TEAD3 and RAD9A following treatment with 20nM control or miR-1287 oligonucleotides. (B) Luciferase assays of pEZX-RAD9A-3'UTR reporter in indicated cells transfected with control olignonucleotides (black) or miR-1287 oligonucleotides (grey). (C) Levels of miR-1287 expression following transfection measured by qRT-PCR. Representative images of the effect of miR-1287 overexpression on apoptosis (D) and DNA damage repair (pH2AX) and proliferation (BrdU incorporation) following treatment with cisplatin. (E) Error bars represent standard error of the mean of triplicate experiments, *p<0.05, ***p<0.0001.

5.4 Discussion

Within the last decade, ncRNAs especially miRNAs, have emerged as key players in tumorigenesis due to their involvement in numerous biological processes. While several studies have shown that miRNA signatures can distinguish AC and SqCC, the majority of these studies focus only on miRNAs that are differentially expressed between tumor tissues, with little use of non-malignant tissue as a baseline [199-201]. A lack of investigation into the biological roles of these differentially altered miRNAs has resulted in a limited understanding of how miRNAs contribute to subtype tumorigenesis. In this study, we performed miRNA sequencing on a cohort of AC and SqCC tumors with patient matched non-malignant tissue in an attempt to generate the first unbiased and comprehensive analysis of miRNA deregulation in NSCLC subtypes and to identify subtype specific miRNAs that may contribute to the distinct clinical phenotypes of AC and SqCC. As smoking is known to lead to different patterns of genomic alterations and has been shown to influence miRNA expression patterns (Vucic et al., Unpublished), we limited our analysis to current and former smokers in order to eliminate the confounding effects of smoking.

Validation of profiling results is essential to ensure robust findings and eliminate potential artifacts. The TCGA is the largest public repository of lung AC and SqCC miRNA expression data, and the only one generated by sequencing. Despite hundreds of tumor profiles, fewer than 80 patient matched non-malignant tissues have profiles, significantly limiting the number of available cases for validation. Nevertheless, we used this cohort to validate the 85 NSCLC and 47 subtype specific miRNAs we identified. The overall validation rate of our subtype specific and NSCLC miRNAs was surprisingly low (48/132, 36%). Analysis of the TCGA data revealed lower detection sensitivity (Vucic *et al*., unpublished), which could partially explain the poor validation rates, especially in AC where heterogeneity is known to be a major factor. Discrepancies in collection of non-malignant tissue and microdissection between our cohort and the TCGA's may also contribute to this disparity. Despite a low validation rate, we feel confident that the miRNAs that did validate represent the most robust candidates for further examination. As additional sequencing data emerges, validation in other cohorts will provide further insight into the most recurrently altered and differentially expressed miRNAs between subtypes.

The identification of miRNAs that were aberrantly expressed in greater than 90% of cases suggests that these miRNAs are important in lung tumor biology, regardless of subtype. The high validation rate of these miRNAs in the TCGA (16/18), and the fact that several of these miRNAs have been previously implicated in tumorigenesis, confirms the importance of these miRNAs to lung tumor biology. Notably, *miR-126, -130b, -144, -451a* and *-486* were identified in meta-analyses of miRNA profiling studies comparing tumor tissue to non-malignant tissue, further validating the prevalence of their disruption in NSCLC [202, 203]. These findings in conjunction with the observation that a number of lung cancer prognostic genes were predicted targets of these aberrantly expressed miRNAs (Figure 5.6), provides additional support to the relevance of these miRNAs in tumor biology and suggests they may be biologically relevant biomarkers. As miRNAs are known to circulate throughout the body with substantial stability, analysis of these miRNAs in blood samples and large sample cohorts to determine their potential as prognostic and/or diagnostic blood based biomarkers of lung cancer is warranted.

Hierarchical clustering of tumor and non-malignant tissue miRNA sequencing profiles revealed miRNA expression is associated with histology, in agreement with previous studies [199, 201, 204, 205]. Not surprisingly, it also suggests that heterogeneity amongst subtypes is present. The identification of subtype specific miRNAs demonstrates that like protein coding genes, disruption of distinct miRNAs underlie tumorigenesis of AC and SqCC. Despite a larger AC cohort, the majority of subtype specific miRNAs identified were SqCC specific, mimicking our findings in Chapter 2. The greater prevalence of overexpressed subtype specific miRNAs, suggests that most deregulated miRNAs function to suppress tumor suppressor genes. Of the 9 validated subtype specific miRNAs, 4 (miR-326 OE in AC and UE in SqCC, -338 UE in SqCC, -375 OE in AC and UE in SqCC and -944 OE in SqCC) have been previously identified as subtype specific, and all in the same subtype and same direction as we observed [204-207].

miR-375 is the only subtype specific miRNA to be implicated in tumorigenesis, let alone lung cancer. Loss of *miR-375* expression has been observed in multiple cancer types and is associated with suppression of cancer hallmarks through the targeting of oncogenes including *IFG1R*, *PDK1* and *JAK2* [207]. While underexpression of *miR-375* is most common, in breast

and prostate cancer, *miR-375* is up-regulated relative to normal tissues and alters proliferation [208]. Interestingly, studies of esophageal cancer observed a similar pattern of subtype specific expression, with AC tumors having a 6-fold higher expression than SqCC tumors [209], however stable introduction of *miR-375* into A549 lung AC cells failed to induce a significant inhibition in growth [210]. Taken together, these findings suggest that *miR-375* functions as both a tumor suppressive and oncogenic miRNA and that its function is likely context dependent, supporting our observation of its bi-directional alteration in NSCLC subtypes.

To date, miRNA profiling studies have identified numerous miRNAs differentially expressed between AC and SqCC, and correlated these with clinical features such as survival and prognosis. However, none of these studies have assessed the functional consequences of this disruption. To provide insights into the differential oncogenic mechanisms disrupted by subtype specific miRNAs, we performed network and pathway analysis on miRDIP predicted targets of subtype specific miRNAs. The top SqCC gene networks were associated with post-translational modification and DNA replication, recombination and repair, corroborating the findings in Chapter 2 and further supporting the relevance of disruption to histone modifying enzymes in SqCC. In Chapter 3, analysis of the KEAP1-CUL3 E3 ubiquitin ligase complex revealed this complex to be differently altered between subtypes, with loss of *CUL3* occurring preferentially in SqCC [151]. Interestingly, the most significant SqCC network disrupted by SqCC specific miRNAs was centered around multiple CUL proteins, specifically CUL2, CUL3 and CUL5. All three of these CUL proteins are components of different E3 ubiquitin ligase complexes that ubiquitinate key tumor suppressors and oncogenes including *JAK2*, *TP53*, *HIF1a* and *NRF2* [115, 117, 211-213]. As we demonstrated, loss of the Cullin component disrupts the function of the complex, leading to disruption of substrate regulation and promotion of tumorigenesis [151]. Mutation and copy loss of *CUL3* in SqCC has been shown by us and others, however, our findings suggest that multiple cullin proteins may be underexpressed in SqCC due to aberrant miRNA expression. Validation of these proteins as direct targets of the predicted miRNA will be essential in further establishing the role of cullin proteins in SqCC tumorignesis.

miR-944 was by far the most frequently deregulated SqCC specific miRNA we identified, overexpressed in 96% and 86% of the BCCRC and TCGA cohorts, respectively. Examination into the genomic coordinates of *miR-944* revealed it to be located at 3q28. Amplification of 3q in SqCC is one of the most well characterized and consistent copy number differences between AC and SqCC [41], and offers an explanation into the frequent copy gain, overexpression and subtype specificity of *miR-944*. Interestingly, the miR-944 transcript is located within the TP63 transcript, a widely used immunohistochemical marker to define lung SqCC. This overlap with *TP63* may explain why *miR-944* was so accurate in discriminating subtypes and supports the notion that *miR-944* could be a useful diagnostic biomarker for SqCC. The identification of disrupted ERK signaling by miR-944 predicted targets highlights a potential mechanism through which *miR-944* may contribute to SqCC tumorigenesis. Given the high frequency of *miR-944* disruption in SqCC, investigation into its biological relevance may provide novel insight into mechanisms of SqCC pathogenesis.

With only one AC specific miRNA validating, pathway analysis would be uninformative, especially given that we identified only 5 predicted targets. As two of the predicted targets have been implicated in tumorigenesis but little is known about *miR-1287*, we explored the role of *miR-1287* in the pathogenesis of lung AC. We have shown a tumor promoting role of *miR-1287*, with overexpression in LCCLs promoting the survival and proliferation of damaged, genetically unstable cells. Conversely, overexpression in HBECs resulted in a reduction of apoptosis and proliferation, suggesting that non-malignant cells may induce a senescent state as a tumor suppressive mechanism, however this needs to be further explored. Moreover, we validated RAD9A as a target of *miR-1287*. RAD9A is a cell cycle checkpoint protein, and together with RAD1 and HUS1 forms the 9-1-1 complex, which plays a central role in activation of DNA damage induced checkpoints [214, 215]. In response to DNA damage, RAD17 loads the 9-1-1 complex around DNA lesions, facilitating phosphorylation and activation of CHK1 kinase- the key event that determines cell survival following genotoxic stress [214, 215]. RAD9A is the key component of the 9-1-1 complex due to the presence of a nuclear localization sequence [215]. Single nucleotide polymorphisms within *RAD9A* are associated with an increased susceptibility for lung AC, further supporting a role for this gene in lung AC [216]. The identification of a

malignant phenotype associated with *miR-1287* overexpression confirms that our analysis approach accurately identifies miRNAs involved in subtype tumor biology. Importantly, our findings highlight the contribution of miRNA dysregulation to lung cancer biology and suggests that analysis of the other subtype specific miRNAs, the functions of which are largely unknown, will provide further insight into the molecular mechanisms underlying AC and SqCC.

While our analysis identified numerous subtype specific and NSCLC miRNAs, we recognize that the stringent criteria we used may have excluded true subtype specific miRNAs. For example, *miR-205* and *miR-21* which have been reported to be overexpressed in SqCC and NSCLC respectively in numerous studies [60, 188, 202], failed to meet our criteria. Despite having significantly different expression between subtypes, *miR-205* was not differentially altered and therefore was not considered subtype specific. Similarly *miR-21*, despite being frequently overexpressed in both subtypes (68% in AC and 77% in SqCC), was found to have significantly different expression between subtypes. The goal of this work was to identify miRNAs that were either significantly different between subtypes, or extremely similar. Therefore, although our list of miRNAs may not be comprehensive, we have shown that our criteria identifies biologically relevant miRNAs that may have use as biomarkers for the detection of lung cancer.

In conclusion, our study provides the first comprehensive sequencing analysis of miRNA expression in lung AC and SqCC. Collectively, our study confirms that similar to protein coding genes, miRNA deregulation occurs in both a subtype specific and non specific manner and that these disruption patterns can reveal novel insight into the biology and molecular mechanisms underlying tumorigenesis and contributing to their disparate clinical phenotypes. We have identified ubiquitously deregulated miRNAs that could be promising biomarkers for the detection of lung cancer as well as subtype specific miRNAs that accurately discriminate subtypes. For the first time, we illustrate the oncogenic pathways and functions associated with subtype specific miRNA deregulation, and implicate *miR-1287* and its target RAD9A in AC tumorigenesis. Our findings underscore the importance of miRNAs to lung cancer biology and highlight the need to look beyond protein coding gene alterations.

Chapter 6: Conclusion

6.1 Summary

Lung cancer is the leading cause of cancer related mortality, accounting for more deaths than breast, prostate and colon cancer combined [1, 2]. The five year survival rate of lung cancer has failed to improve significantly over the past decades, from 14% in 1994 to 18% in 2008, emphasizing the need for novel diagnostic and therapeutic strategies [2]. Genomic analyses of tumor genomes, especially recent characterization using high throughput sequencing technologies have identified numerous driver alterations contributing to tumorigenesis. However, driver alterations for up to 50% of cases remain to be elucidated, highlighting the need to improve our understanding of the molecular mechanisms underlying lung tumorigenesis.

Histological subtypes of lung cancer display disparate clinical phenotypes and responses to therapy, suggesting they are distinct diseases that arise through divergent molecular mechanisms. The goal of this work was to characterize the landscape of molecular alterations in AC and SqCC through the integration of multiple 'omics dimensions, and elucidate the mechanisms through which frequently deregulated genes and signaling pathways common to NSCLC or altered in a subtype specific manner contribute to tumorigenesis. Our analysis revealed several molecular alterations specific to a given subtype. Importantly, the findings from this work provide insights into novel potential therapeutic strategies for SqCC and identified potential diagnostic biomarkers for lung cancer treatment and prognosis.

6.1.1 Identify molecular alterations common or specific to lung cancer subtypes.

NSCLC is a heterogeneous disease comprised of regionally distinct and phenotypically diverse tumors. Fundamental differences in tumor biology are thought to underlie the phenotypic differences of these tumors, and although previous studies have identified distinct patterns of alterations, the specific genes and signaling pathways responsible for the disparate clinical phenotypes are only just beginning to be identified. Therefore in Chapter 2, we integrated genome wide copy number and methylation with gene expression profiles for a cohort of 169 AC and 92 SqCC tumors to characterize the landscape of molecular alterations with concordant

expression changes specific to each subtype as these are most likely to be causal in tumorigenesis. Our analysis revealed 294 regions of copy number disparity of which 205 were specific to SqCC and 89 were AC specific as well as 2384 differentially methylated genes. After excluding genes without differential expression between tumors and non-malignant tissue, we identified 778 genetically or epigenetically deregulated subtype specific genes. From these genes we identified key oncogenic pathways disrupted in each subtype that likely serve as the basis for their differential biology and clinical outcomes. Downregulation of HNF4α target genes was the most common pathway specific to AC while SqCC tumors demonstrated disruption of multiple histone modifying enzymes and *E2F1*. In addition, several genes previously implicated in NSCLC such as *ERCC1* and *KEAP1* were found to display subtype specific patterns of alteration.

Within the past decade, miRNAs have emerged as critical regulators of gene expression, implicated in normal biological processes as well as tumorigenesis [56]. To date, miRNA profiling studies of histological subtypes have been exclusively array based, limiting the number of miRNAs for interrogation. In Chapter 5 we performed miRNA sequencing on 66 AC and 22 SqCC tumors with matched non-malignant tissue. We identified miRNAs recurrently altered in all cases regardless of subtype as well as 48 subtype specific miRNAs; 29 of which were SqCC specific, 13 that were AC specific and 5 that were disrupted in both subtypes, but in opposing directions. Similar to protein coding genes, subtype specific miRNAs were associated with distinct pathways and functions in each subtype and their expression accurately discriminates subtypes. Together these profiling studies revealed novel subtype specific patterns of alterations and provide insight into the specific genes that contribute to subtype tumorigenesis, further demonstrating that AC and SqCC develop through distinct pathways.

6.1.2 Delineate the biological significance and clinical relevance of alterations.

Despite the existence of numerous profiling studies examining the pattern of alterations defining AC and SqCC, few studies have investigated the function and molecular mechanisms through which these subtype specific alterations contribute to subtype development and tumorigenesis. In chapters 3, 4 and 5 we assessed the clinical relevance of subtype specific

alterations by manipulating candidate genes identified by integrative analysis or miRNA sequencing *in vitro* and *in vivo* to delineate the biological significance of their disruption and elucidate the mechanism by which they contribute to disease pathogenesis.

Specifically, we show that the complex components of the KEAP1-CUL3 E3 ubiquitin ligase as well as its substrate *IKBKB* are differentially disrupted in AC and SqCC, with *KEAP1* loss characterizing AC tumors and *CUL3* loss and *IKBKB* gain being preferential in SqCC. Importantly, disruption of a single complex component was sufficient to impair function, leading to the accumulation and aberrant activation of NF-κB signaling. In chapter 4 we identified 12q13-15 and more specifically *YEATS4* as amplified and overexpressed in >20% of NSCLCs. Overexpression of *YEATS4* abrogated senescence while attenuation of *YEATS4* in lung cancer cell lines reduced proliferation, colony formation and tumor growth and induced cellular senescence. Furthermore we demonstrate that *YEATS4* acts as a negative regulator of the p21-p53 pathway by inhibiting p21 and that *YEATS4* expression affected the cellular response to cisplatin with increased levels associated with resistance and lower levels with sensitivity. Finally in chapter 5, we demonstrate the ability of our analysis approach to identify biologically relevant miRNAs by showing the effect of *miR-1287* overexpression in normal immortalized bronchial epithelial cells and lung cancer cell lines. In addition to the identification of subtype specific miRNAs, we identified a panel of 16 miRNAs recurrently aberrantly expressed in tumors (>90%) regardless subtype and suggest that these miRNAs may be useful biomarkers for the detection of lung cancer.

6.2 Conclusions

We hypothesized that 1) AC and SqCC arise through distinct molecular mechanisms, which can be identified through integrative analysis of AC and SqCC genomes and 2) that these different patterns of genetic and epigenetic alterations underlie unique biological mechanisms that contribute to subtype development, phenotypes and response to therapy. Our characterization of AC and SqCC genomes using an integrative 'omics approach revealed a number of subtype specific alterations with concordant expression changes as well as differentially expressed miRNAs. These differences were found to affect distinct oncogenic

pathways and accurately separate AC and SqCC samples, providing further evidence that these subtype specific genes are responsible for driving the differential development of AC and SqCC. These findings, support our hypotheses and are consistent with previous reports, but importantly they provide new insight into specific genes underlying subtype tumorigenesis. Functional analysis of frequently deregulated subtype specific and non-specific genes revealed the phenotypes associated with their disruption, and the molecular mechanisms through which they promote tumor development and progression. Through the characterization of these mechanisms, we were able to identify drugs whose sensitivity was altered by gene disruption, highlighting the clinical relevance of these genes. In addition to subtype specific genes, genes and miRNAs frequently disrupted across both subtypes were identified. This suggest that certain biological functions or cellular processes are essential to all lung tumors, regardless of subtype. Collectively, the findings from this thesis confirm that while certain molecular alterations are shared amongst histological subtypes, distinct alterations underlie their differential development, and clinical phenotypes.

6.3 Significance and clinical implications

The success of EGFR TKIs in *EGFR* mutant lung AC, and the profound benefit of targeted therapies such as trastuzumab and imatinib in breast cancer and chronic myeloid leukemia, respectively demonstrated the potential survival benefit of targeted therapies and launched the search for additional actionable alterations in lung cancer. As a result, the landscape of genomic alterations defining AC and SqCC, has been extensively studied [15, 17, 18, 33-36, 39, 51, 55, 153]. While a subset of the genes involved in lung cancer have been well characterized, including *NKX2-1*, *PTEN*, *CDKN1A*, *PIK3CA*, *SOX2*, *KRAS* and *EGFR*, the majority of genes responsible for tumorigenesis remain unknown. This is exemplified by the findings that roughly 50% of NSCLC tumors harbour no known driver gene/targetable alteration [142]. Moreover, the majority of studies describing the patterns of alterations in NSCLC subtypes focus primarily on gene identification, with little emphasis on elucidating the pathways and molecular mechanisms underlying subtype development. Therefore one of the most important findings from this thesis is the identification of novel subtype specific genes and uncovering the molecular mechanisms through which these genes are involved in tumorigenesis.

Diverse patterns of genomic alterations suggests that AC and SqCC are distinct diseases. These findings have had significant impact on the clinical management on NSCLC, with the one size fits all approach no longer considered practical. Targeted therapies significantly improve survival of patients harbouring the specific alteration compared to standard chemotherapy; however patients will inevitably recur, such that targeted therapies are administered without curative intent. The development of novel, effective treatment approaches is desperately needed, especially for SqCC in which there are currently no targeted therapies approved for clinical use. This thesis identified several patterns of gene disruption associated with treatment sensitivity, the most clinically relevant of which was the finding that SqCC tumors/cell lines are more sensitive to HDAC inhibition than AC, offering a novel treatment strategy for this subtype with few effective treatment options. PI3K inhibitors were also identified as potentially effective in SqCC and the TCGAs characterization of SqCC tumors found the PI3K pathway was altered in 47% of cases [55], further supporting the role of this pathway in SqCC tumorigenesis. While the dependence of SqCC tumors on these pathways and specific alterations has yet to be functionally defined, our analysis has identified new therapeutic avenues worth exploring. Overall, this thesis illustrates the molecular differences that underlie AC and SqCC and highlights the importance of not only stratifying patients based on histology but also underlying tumor biology.

6.4 Future directions

The findings of this work have generated several additional research avenues worth pursuing: i) understanding the mechanism of *YEATS4* mediated tumorigenesis ii) discerning the contribution of miR-944 to SqCC tumorigenesis , iii) validation of the most frequently deregulated miRNAs as biomarkers for the early detection of lung cancer, iv) analysis of histone modifications in SqCC and v) integration of genomics data with features of the tumor microenvironment. Future work investigating these issues will be essential to further define the mechanisms underlying lung tumorigenesis and assess the potential of candidates as robust biomarkers for diagnosis or novel therapeutic targets.

6.4.1 Mechanisms of YEATS4 mediated tumorigenesis

Our analysis of *YEATS4* revealed it contributes to tumorigenesis by altering numerous cellular processes. While we show that *YEATS4* negatively regulates the p53-p21 pathway, in p53 mutant cell lines, knockdown of *CDKN1A* had no effect on viability, senescence or anchorage independent growth, and cisplatin sensitivity could not be associated with differences in p21 or p53, suggesting that these phenotypes are mediated through other pathways. Analysis of differentially expressed genes following knockdown identified the targets of *MYC* and *SRF* transcription factors to be significantly enriched. These findings provide a logical starting point to further investigate the mechanisms of *YEATS4* mediated tumorigenesis. Elucidation of these mechanisms will be important in determining the clinical relevance of *YEATS4* amplification in NSCLC.

6.4.2 Validation of miRNAs as biomarkers for the detection of lung cancer

NSCLC patient survival improves dramatically to over 70% when tumors are detected at the earliest localized stage, emphasizing that the greatest survival benefits for lung cancer will be achieved through early detection [1]. Due to the time consuming and invasive nature of current early detection techniques such as bronchoscopy and low-dose computed tomography, as well as a lack of criteria (other than smoking) for identifying individuals at risk of lung cancer, early detection strategies are not widely available to the general public. Non-invasive diagnostic biomarkers therefore hold immense promise in improving lung cancer survival. miRNAs circulate in the blood with substantial stability, making them attractive targets for biomarkers. The identification of 16 miRNAs that were frequently disrupted in our NSCLC cases, validated in the TCGA and previously reported by others, suggests they may have potential as biomarkers. There are a number of criteria for robust, highly sensitive and specific biomarkers; however for this panel of miRNAs, the first step should involve analysis in blood samples to determine whether they are detectable, and the concordance between tumor and blood expression. Successful candidates should then be assessed in blood samples from patients with and without cancer as well as those from patients with benign lung disease to assess the ability of these miRNAs to discriminate between benign and malignant disease. Finally, validation in multiple

large independent cohorts will be essential for determining the clinical feasibility of remaining candidates.

6.4.3 Contribution of miR-944 to SqCC tumorigenesis

miR-944 expression was frequently deregulated in SqCC (2 fold overexpression in >95%) and highly specific for SqCC tumors, accurately separating AC from SqCC with significant accuracy in both our datasets and the TCGA (AUC values of 0.9669 and 0.858, $p<0.0001$, respectively). Predicted targets of miR-944 include known tumor suppressors *EYA4* and *NF1*. Together, these findings provide strong evidence supporting a role for *miR-944* in SqCC tumorigenesis. In order to determine the biological relevance of *miR-944* overexpression, its function and mechanism of action should be characterized in lung cancer cell lines and immortalized human bronchial epithelial cells by ectopic expression. There are currently no targeted therapies approved for clinical use in SqCC. The ubiquitous disruption of miR-944 could therefore have significant clinical implications if it is found that *miR-944* does in fact contribute to SqCC development.

6.4.4 Global assessment of histone modifications in SqCC

Although this thesis integrated multiple 'omics dimensions (copy number, DNA methylation, gene and miRNA expression), several other mechanisms of gene disruption exist. These include sequence mutations (point mutations, insertions and deletions) which can alter gene function, gene fusions which create new oncogenic proteins, histone modifications which affect chromatin structure and accessibility and ncRNAs (long non-coding, antisense RNAs) which modulate gene expression. Post-translational modification of histone tails by acetylation and methylation play crucial roles in regulating chromatin structure and gene expression [104]. The polycomb repressive complex 2 (PRC2) is a histone methyltransferase comprised of numerous polycomb proteins that methylates histone H3 on Lysine 27 (H3K27me, me2 or me3), a mark that transcriptionally silences chromatin [217]. Components of this complex including *EZH2*- the catalytic subunit, were found to be frequently deregulated in SqCC likely leading to aberrant patterns of gene expression. High throughput approaches for the analysis of histone modifications are widely available, and a large public repository of histone modification data is

available through ENCODE [218]. Given the prominent disruption of histone modifying enzymes in SqCC, interrogation of the polycomb group proteins and histone modifications throughout the genome would likely reveal novel insights into SqCC pathogenesis.

6.4.5 Integration of molecular and immunologic alterations to understand the contribution of the tumor microenvironment to tumorigenesis.

The immune system plays a key role in the early elimination of tumors, and its importance in tumorigenesis is becoming increasingly appreciated, with immune evasion now considered a hallmark of cancer [219]. The tumor microenvironment is a complex milieu in which cancer cells, stromal cells and tumor infiltrating leukocytes interact. Targeting immune cells within the microenvironment is a growing field, and immunotherapies for the treatment of lung cancers have shown clinical activity in NSCLC suggesting this immune modulation may be an effective treatment in lung cancer [220-222]. However, preliminary data show that only a subset of patients respond to PD-1 blockade. While *EGFR* mutations were recently found to activate the PD-1 pathway contributing to the immune escape of these tumors [223], it is largely unknown whether specific genomic subsets of lung tumors activate this or other immune pathways and may confer sensitivity to immunotherapies. The identification of immune related signaling pathways disrupted by subtype specific alterations suggests that the immune cell compartments and activation/suppression of these cells within the tumor microenvironement may differ between subtypes. Integration of genomics data with tumor microenvironment profiles will provide insight into how genomic alterations influence the immune compartment and may reveal novel therapeutic strategies for NSCLC.

References

1. Siegel, R., et al., *Cancer statistics, 2014.* CA Cancer J Clin, 2014. **64**(1): p. 9-29.
2. Statistics, C.C.S.s.A.C.o.C., *Canadian Cancer Statistics 2013.* 2013.
3. Sato, M., et al., *A translational view of the molecular pathogenesis of lung cancer.* J Thorac Oncol, 2007. **2**(4): p. 327-43.
4. 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.
5. Hecht, S.S., *Tobacco carcinogens, their biomarkers and tobacco-induced cancer.* Nat Rev Cancer, 2003. **3**(10): p. 733-44.
6. Hajdu, S.I., *Much overlooked causes of lung cancer.* Ann Clin Lab Sci, 2011. **41**(1): p. 97-101.
7. Travis, W.D., et al., *Pathology and genetics of tumours of the lung, pleura, thymus and heart*. World Health Organization classification of tumours2004:
8. Travis, W.D., et al., *International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society International Multidisciplinary Classification of Lung Adenocarcinoma.* Journal of Thoracic Oncology, 2011. **6**(2): p. 244-285.
9. Travis, W.D., *Pathology of lung cancer.* Clin Chest Med, 2002. **23**(1): p. 65-81, viii.
10. Wistuba, II and A.F. Gazdar, *Lung cancer preneoplasia.* Annu Rev Pathol, 2006. **1**: p. 331-48.
11. Johnson, D.H., et al., *Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer.* J Clin Oncol, 2004. **22**(11): p. 2184-91.
12. Scagliotti, G., et al., *The differential efficacy of pemetrexed according to NSCLC histology: a review of two Phase III studies.* Oncologist, 2009. **14**(3): p. 253-63.
13. Langer, C.J., et al., *The evolving role of histology in the management of advanced non-small-cell lung cancer.* J Clin Oncol, 2010. **28**(36): p. 5311-20.
14. Heller, G., et al., *Genome-wide CpG island methylation analyses in non-small cell lung cancer patients.* Carcinogenesis, 2013. **34**(3): p. 513-21.
15. Ding, L., et al., *Somatic mutations affect key pathways in lung adenocarcinoma.* Nature, 2008. **455**(7216): p. 1069-75.
16. Liu, P., et al., *Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing.* Carcinogenesis, 2012. **33**(7): p. 1270-6.
17. Seo, J.S., et al., *The transcriptional landscape and mutational profile of lung adenocarcinoma.* Genome Res, 2012. **22**(11): p. 2109-19.
18. Weir, B.A., et al., *Characterizing the cancer genome in lung adenocarcinoma.* Nature, 2007. **450**(7171): p. 893-8.
19. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
20. Bhattacharjee, A., et al., *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13790-5.
21. Garber, M.E., et al., *Diversity of gene expression in adenocarcinoma of the lung.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13784-9.

22. McDoniels-Silvers, A.L., et al., *Differential gene expression in human lung adenocarcinomas and squamous cell carcinomas.* Clin Cancer Res, 2002. **8**(4): p. 1127-38.
23. McDoniels-Silvers, A.L., et al., *Differential expression of critical cellular genes in human lung adenocarcinomas and squamous cell carcinomas in comparison to normal lung tissues.* Neoplasia, 2002. **4**(2): p. 141-50.
24. Nacht, M., et al., *Molecular characteristics of non-small cell lung cancer.* Proc Natl Acad Sci U S A, 2001. **98**(26): p. 15203-8.
25. Wilkerson, M.D., et al., *Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types.* Clin Cancer Res, 2010. **16**(19): p. 4864-75.
26. Hou, J., et al., *Gene expression-based classification of non-small cell lung carcinomas and survival prediction.* PLoS One, 2010. **5**(4): p. e10312.
27. Yamagata, N., et al., *A training-testing approach to the molecular classification of resected non-small cell lung cancer.* Clin Cancer Res, 2003. **9**(13): p. 4695-704.
28. Virtanen, C., et al., *Integrated classification of lung tumors and cell lines by expression profiling.* Proc Natl Acad Sci U S A, 2002. **99**(19): p. 12357-62.
29. Rohrbeck, A., et al., *Gene expression profiling for molecular distinction and characterization of laser captured primary lung cancers.* J Transl Med, 2008. **6**: p. 69.
30. Beer, D.G., et al., *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nat Med, 2002. **8**(8): p. 816-24.
31. Giordano, T.J., et al., *Organ-specific molecular classification of primary lung, colon, and ovarian adenocarcinomas using gene expression profiles.* Am J Pathol, 2001. **159**(4): p. 1231-8.
32. Balsara, B.R. and J.R. Testa, *Chromosomal imbalances in human lung cancer.* Oncogene, 2002. **21**(45): p. 6877-83.
33. Pei, J., et al., *Genomic imbalances in human lung adenocarcinomas and squamous cell carcinomas.* Genes Chromosomes Cancer, 2001. **31**(3): p. 282-7.
34. Sy, S.M., et al., *Distinct patterns of genetic alterations in adenocarcinoma and squamous cell carcinoma of the lung.* Eur J Cancer, 2004. **40**(7): p. 1082-94.
35. Tonon, G., et al., *High-resolution genomic profiles of human lung cancer.* Proc Natl Acad Sci U S A, 2005. **102**(27): p. 9625-30.
36. Yakut, T., et al., *Assessment of molecular events in squamous and non-squamous cell lung carcinoma.* Lung Cancer, 2006. **54**(3): p. 293-301.
37. Massion, P.P., et al., *Genomic copy number analysis of non-small cell lung cancer using array comparative genomic hybridization: implications of the phosphatidylinositol 3-kinase pathway.* Cancer Res, 2002. **62**(13): p. 3636-40.
38. Petersen, I., et al., *Patterns of chromosomal imbalances in adenocarcinoma and squamous cell carcinoma of the lung.* Cancer Res, 1997. **57**(12): p. 2331-5.
39. Staaf, J., et al., *Landscape of somatic allelic imbalances and copy number alterations in human lung carcinoma.* Int J Cancer, 2013. **132**(9): p. 2020-31.
40. Bass, A.J., et al., *SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas.* Nature Genetics, 2009. **41**(11): p. 1238-U105.
41. Bjorkqvist, A.M., et al., *DNA gains in 3q occur frequently in squamous cell carcinoma of the lung, but not in adenocarcinoma.* Genes Chromosomes Cancer, 1998. **22**(1): p. 79-82.

42. Cancer Genome Atlas Research, N., et al., *Comprehensive genomic characterization of squamous cell lung cancers.* Nature, 2012. **489**(7417): p. 519-25.
43. Luk, C., et al., *Molecular cytogenetic analysis of non-small cell lung carcinoma by spectral karyotyping and comparative genomic hybridization.* Cancer Genet Cytogenet, 2001. **125**(2): p. 87-99.
44. Baylin, S.B., *The cancer epigenome: its origins, contributions to tumorigenesis, and translational implications.* Proc Am Thorac Soc, 2012. **9**(2): p. 64-5.
45. Tessema, M. and S.A. Belinsky, *Mining the epigenome for methylated genes in lung cancer.* Proc Am Thorac Soc, 2008. **5**(8): p. 806-10.
46. Heller, G., C.C. Zielinski, and S. Zochbauer-Muller, *Lung cancer: from single-gene methylation to methylome profiling.* Cancer Metastasis Rev, 2010. **29**(1): p. 95-107.
47. Shiraishi, M., et al., *HOX gene clusters are hotspots of de novo methylation in CpG islands of human lung adenocarcinomas.* Oncogene, 2002. **21**(22): p. 3659-62.
48. Zochbauer-Muller, S., et al., *Aberrant promoter methylation of multiple genes in non-small cell lung cancers.* Cancer Res, 2001. **61**(1): p. 249-55.
49. Toyooka, S., et al., *Mutational and epigenetic evidence for independent pathways for lung adenocarcinomas arising in smokers and never smokers.* Cancer Res, 2006. **66**(3): p. 1371-5.
50. 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.
51. Imielinski, M., et al., *Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing.* Cell, 2012. **150**(6): p. 1107-20.
52. 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.
53. Kandoth, C., et al., *Mutational landscape and significance across 12 major cancer types.* Nature, 2013. **502**(7471): p. 333-9.
54. Lee, W., et al., *The mutation spectrum revealed by paired genome sequences from a lung cancer patient.* Nature, 2010. **465**(7297): p. 473-7.
55. Hammerman, P.S., et al., *Comprehensive genomic characterization of squamous cell lung cancers.* Nature, 2012. **489**(7417): p. 519-25.
56. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.
57. Calin, G.A. and C.M. Croce, *MicroRNA signatures in human cancers.* Nat Rev Cancer, 2006. **6**(11): p. 857-66.
58. Lebanony, D., et al., *Diagnostic assay based on hsa-miR-205 expression distinguishes squamous from nonsquamous non-small-cell lung carcinoma.* J Clin Oncol, 2009. **27**(12): p. 2030-7.
59. Dacic, S., et al., *miRNA expression profiling of lung adenocarcinomas: correlation with mutational status.* Mod Pathol, 2010. **23**(12): p. 1577-82.
60. Seike, M., et al., *MiR-21 is an EGFR-regulated anti-apoptotic factor in lung cancer in never-smokers.* Proc Natl Acad Sci U S A, 2009. **106**(29): p. 12085-90.
61. Chari, R., et al., *Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer.* Cancer Metastasis Rev, 2010. **29**(1): p. 73-93.
62. Mok, T.S., et al., *Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma.* N Engl J Med, 2009. **361**(10): p. 947-57.

63. Broet, P., et al., *Prediction of clinical outcome in multiple lung cancer cohorts by integrative genomics: implications for chemotherapy selection.* Cancer Res, 2009. **69**(3): p. 1055-62.
64. Igawa, S., et al., *Pemetrexed for previously treated patients with non-small cell lung cancer and differences in efficacy according to thymidylate synthase expression.* Chemotherapy, 2012. **58**(4): p. 313-20.
65. Ceppi, P., et al., *Squamous cell carcinoma of the lung compared with other histotypes shows higher messenger RNA and protein levels for thymidylate synthase.* Cancer, 2006. **107**(7): p. 1589-96.
66. Thomas, R.K., B. Weir, and M. Meyerson, *Genomic approaches to lung cancer.* Clin Cancer Res, 2006. **12**(14 Pt 2): p. 4384s-4391s.
67. Feinberg, A.P. and B. Tycko, *The history of cancer epigenetics.* Nat Rev Cancer, 2004. **4**(2): p. 143-53.
68. Hyman, E., et al., *Impact of DNA amplification on gene expression patterns in breast cancer.* Cancer Res, 2002. **62**(21): p. 6240-5.
69. Pollack, J.R., et al., *Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.* Proc Natl Acad Sci U S A, 2002. **99**(20): p. 12963-8.
70. Bass, A.J., et al., *SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas.* Nat Genet, 2009. **41**(11): p. 1238-42.
71. Garnis, C., et al., *Chromosome 5p aberrations are early events in lung cancer: implication of glial cell line-derived neurotrophic factor in disease progression.* Oncogene, 2005. **24**(30): p. 4806-12.
72. Coe, B.P., et al., *Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer.* Br J Cancer, 2006. **94**(12): p. 1927-35.
73. Baldwin, C., et al., *Multiple microalterations detected at high frequency in oral cancer.* Cancer Res, 2005. **65**(17): p. 7561-7.
74. Lockwood, W.W., et al., *Whole genome tiling path array CGH analysis of segmental copy number alterations in cervical cancer cell lines.* Int J Cancer, 2007. **120**(2): p. 436-43.
75. Khojasteh, M., et al., *A stepwise framework for the normalization of array CGH data.* BMC Bioinformatics, 2005. **6**: p. 274.
76. Lockwood, W.W., et al., *DNA amplification is a ubiquitous mechanism of oncogene activation in lung and other cancers.* Oncogene, 2008. **27**(33): p. 4615-24.
77. Chi, B., et al., *SeeGH--a software tool for visualization of whole genome array comparative genomic hybridization data.* BMC Bioinformatics, 2004. **5**: p. 13.
78. Jong, K., et al., *Breakpoint identification and smoothing of array comparative genomic hybridization data.* Bioinformatics, 2004. **20**(18): p. 3636-7.
79. Chari, R., et al., *Effect of active smoking on the human bronchial epithelium transcriptome.* BMC Genomics, 2007. **8**: p. 297.
80. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
81. Bild, A.H., et al., *Oncogenic pathway signatures in human cancers as a guide to targeted therapies.* Nature, 2006. **439**(7074): p. 353-7.

82. Lee, E.S., et al., *Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression.* Clin Cancer Res, 2008. **14**(22): p. 7397-404.
83. Lamb, J., et al., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.* Science, 2006. **313**(5795): p. 1929-35.
84. Solis, L.M., et al., *Nrf2 and Keap1 abnormalities in non-small cell lung carcinoma and association with clinicopathologic features.* Clin Cancer Res, 2010. **16**(14): p. 3743-53.
85. Yuan, P., et al., *Sex determining region Y-Box 2 (SOX2) is a potential cell-lineage gene highly expressed in the pathogenesis of squamous cell carcinomas of the lung.* PLoS One, 2010. **5**(2): p. e9112.
86. Doi, A., et al., *Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts.* Nat Genet, 2009. **41**(12): p. 1350-3.
87. Choi, S.H., et al., *Changes in DNA methylation of tandem DNA repeats are different from interspersed repeats in cancer.* Int J Cancer, 2009. **125**(3): p. 723-9.
88. Hoffmann, M.J. and W.A. Schulz, *Causes and consequences of DNA hypomethylation in human cancer.* Biochem Cell Biol, 2005. **83**(3): p. 296-321.
89. Yang, H.H., et al., *Influence of genetic background and tissue types on global DNA methylation patterns.* PLoS One, 2010. **5**(2): p. e9355.
90. Vaissiere, T., et al., *Quantitative analysis of DNA methylation profiles in lung cancer identifies aberrant DNA methylation of specific genes and its association with gender and cancer risk factors.* Cancer Res, 2009. **69**(1): p. 243-52.
91. Petersen, I., et al., *Core classification of lung cancer: correlating nuclear size and mitoses with ploidy and clinicopathological parameters.* Lung Cancer, 2009. **65**(3): p. 312-8.
92. Wu, Z.L., et al., *Polycomb protein EZH2 regulates E2F1-dependent apoptosis through epigenetically modulating Bim expression.* Cell Death Differ, 2010. **17**(5): p. 801-10.
93. Eymin, B., et al., *Distinct pattern of E2F1 expression in human lung tumours: E2F1 is upregulated in small cell lung carcinoma.* Oncogene, 2001. **20**(14): p. 1678-87.
94. Higashiyama, M., et al., *Reduced motility related protein-1 (MRP-1/CD9) gene expression as a factor of poor prognosis in non-small cell lung cancer.* Cancer Res, 1995. **55**(24): p. 6040-4.
95. Yamamoto, H., et al., *PIK3CA mutations and copy number gains in human lung cancers.* Cancer Res, 2008. **68**(17): p. 6913-21.
96. Garraway, L.A. and W.R. Sellers, *Lineage dependency and lineage-survival oncogenes in human cancer.* Nat Rev Cancer, 2006. **6**(8): p. 593-602.
97. Hillemacher, T., et al., *Global DNA methylation is influenced by smoking behaviour.* Eur Neuropsychopharmacol, 2008. **18**(4): p. 295-8.
98. Lockwood, W.W., et al., *Integrative genomic analyses identify BRF2 as a novel lineage-specific oncogene in lung squamous cell carcinoma.* PLoS Med, 2010. **7**(7): p. e1000315.
99. Dang, T.P., et al., *Constitutive activation of Notch3 inhibits terminal epithelial differentiation in lungs of transgenic mice.* Oncogene, 2003. **22**(13): p. 1988-97.
100. Wang, I.C., et al., *Transgenic expression of the forkhead box M1 transcription factor induces formation of lung tumors.* Oncogene, 2008. **27**(30): p. 4137-49.

101. Lee, D.F., et al., *KEAP1 E3 ligase-mediated downregulation of NF-kappaB signaling by targeting IKKbeta.* Mol Cell, 2009. **36**(1): p. 131-40.
102. Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications.* Nature, 2000. **403**(6765): p. 41-5.
103. Esteller, M., *Epigenetics in cancer.* N Engl J Med, 2008. **358**(11): p. 1148-59.
104. Esteller, M., *Cancer epigenomics: DNA methylomes and histone-modification maps.* Nat Rev Genet, 2007. **8**(4): p. 286-98.
105. Barlesi, F., et al., *Global histone modifications predict prognosis of resected non small-cell lung cancer.* J Clin Oncol, 2007. **25**(28): p. 4358-64.
106. Van Den Broeck, A., et al., *Loss of histone H4K20 trimethylation occurs in preneoplasia and influences prognosis of non-small cell lung cancer.* Clin Cancer Res, 2008. **14**(22): p. 7237-45.
107. Lazarevich, N.L. and D.I. Fleishman, *Tissue-specific transcription factors in progression of epithelial tumors.* Biochemistry (Mosc), 2008. **73**(5): p. 573-91.
108. Grigo, K., et al., *HNF4 alpha orchestrates a set of 14 genes to down-regulate cell proliferation in kidney cells.* Biol Chem, 2008. **389**(2): p. 179-87.
109. Watt, A.J., W.D. Garrison, and S.A. Duncan, *HNF4: a central regulator of hepatocyte differentiation and function.* Hepatology, 2003. **37**(6): p. 1249-53.
110. Lucas, B., et al., *HNF4alpha reduces proliferation of kidney cells and affects genes deregulated in renal cell carcinoma.* Oncogene, 2005. **24**(42): p. 6418-31.
111. Gollob, J.A., et al., *Role of Raf kinase in cancer: therapeutic potential of targeting the Raf/MEK/ERK signal transduction pathway.* Semin Oncol, 2006. **33**(4): p. 392-406.
112. Zhao, Y., et al., *Inhibitors of histone deacetylases target the Rb-E2F1 pathway for apoptosis induction through activation of proapoptotic protein Bim.* Proc Natl Acad Sci U S A, 2005. **102**(44): p. 16090-5.
113. Orzan, F., et al., *Enhancer of Zeste 2 (EZH2) is up-regulated in malignant gliomas and in glioma stem-like cells.* Neuropathol Appl Neurobiol, 2011. **37**(4): p. 381-94.
114. Herbst, R.S., J.V. Heymach, and S.M. Lippman, *Lung cancer.* N Engl J Med, 2008. **359**(13): p. 1367-80.
115. Cullinan, S.B., et al., *The Keap1-BTB protein is an adaptor that bridges Nrf2 to a Cul3-based E3 ligase: oxidative stress sensing by a Cul3-Keap1 ligase.* Mol Cell Biol, 2004. **24**(19): p. 8477-86.
116. Furukawa, M. and Y. Xiong, *BTB protein Keap1 targets antioxidant transcription factor Nrf2 for ubiquitination by the Cullin 3-Roc1 ligase.* Mol Cell Biol, 2005. **25**(1): p. 162-71.
117. Kobayashi, A., et al., *Oxidative stress sensor Keap1 functions as an adaptor for Cul3-based E3 ligase to regulate proteasomal degradation of Nrf2.* Mol Cell Biol, 2004. **24**(16): p. 7130-9.
118. Hayes, J.D. and M. McMahon, *NRF2 and KEAP1 mutations: permanent activation of an adaptive response in cancer.* Trends Biochem Sci, 2009. **34**(4): p. 176-88.
119. Blake, D.J., et al., *Deletion of Keap1 in the lung attenuates acute cigarette smoke-induced oxidative stress and inflammation.* Am J Respir Cell Mol Biol, 2010. **42**(5): p. 524-36.
120. Karin, M., et al., *NF-kappaB in cancer: from innocent bystander to major culprit.* Nat Rev Cancer, 2002. **2**(4): p. 301-10.

121. Nakanishi, C. and M. Toi, *Nuclear factor-kappaB inhibitors as sensitizers to anticancer drugs.* Nat Rev Cancer, 2005. **5**(4): p. 297-309.
122. Weiss, J., et al., *Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer.* Sci Transl Med, 2010. **2**(62): p. 62ra93.
123. Okamoto, T., T. Sanda, and K. Asamitsu, *NF-kappa B signaling and carcinogenesis.* Curr Pharm Des, 2007. **13**(5): p. 447-62.
124. Tang, X., et al., *Nuclear factor-kappaB (NF-kappaB) is frequently expressed in lung cancer and preneoplastic lesions.* Cancer, 2006. **107**(11): p. 2637-46.
125. Meylan, E., et al., *Requirement for NF-kappaB signalling in a mouse model of lung adenocarcinoma.* Nature, 2009. **462**(7269): p. 104-7.
126. Basseres, D.S., et al., *Requirement of the NF-kappaB subunit p65/RelA for K-Ras-induced lung tumorigenesis.* Cancer Res, 2010. **70**(9): p. 3537-46.
127. Ohta, T., et al., *Loss of Keap1 function activates Nrf2 and provides advantages for lung cancer cell growth.* Cancer Res, 2008. **68**(5): p. 1303-9.
128. Singh, A., et al., *Dysfunctional KEAP1-NRF2 interaction in non-small-cell lung cancer.* PLoS Med, 2006. **3**(10): p. e420.
129. Wang, R., et al., *Hypermethylation of the Keap1 gene in human lung cancer cell lines and lung cancer tissues.* Biochem Biophys Res Commun, 2008. **373**(1): p. 151-4.
130. Kohno, T., et al., *A catalog of genes homozygously deleted in human lung cancer and the candidacy of PTPRD as a tumor suppressor gene.* Genes Chromosomes Cancer, 2010.
131. Suzuki, T., J. Maher, and M. Yamamoto, *Select Heterozygous Keap1 Mutations Have a Dominant-negative Effect on Wild-type Keap1 In Vivo.* Cancer Research, 2010: p. epub.
132. Kan, Z., et al., *Diverse somatic mutation patterns and pathway alterations in human cancers.* Nature, 2010. **466**(7308): p. 869-73.
133. Bivona, T.G., et al., *FAS and NF-kappaB signalling modulate dependence of lung cancers on mutant EGFR.* Nature, 2011. **471**(7339): p. 523-6.
134. Garnis, C., et al., *High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH.* Int J Cancer, 2006. **118**(6): p. 1556-64.
135. Ishkanian, A.S., et al., *A tiling resolution DNA microarray with complete coverage of the human genome.* Nat Genet, 2004. **36**(3): p. 299-303.
136. Wong, K.K., et al., *A comprehensive analysis of common copy-number variations in the human genome.* Am J Hum Genet, 2007. **80**(1): p. 91-104.
137. McCarroll, S.A., et al., *Integrated detection and population-genetic analysis of SNPs and copy number variation.* Nat Genet, 2008. **40**(10): p. 1166-74.
138. Beroukhim, R., et al., *The landscape of somatic copy-number alteration across human cancers.* Nature, 2010. **463**(7283): p. 899-905.
139. Schmid, J.A. and A. Birbach, *IkappaB kinase beta (IKKbeta/IKK2/IKBKB)--a key molecule in signaling to the transcription factor NF-kappaB.* Cytokine Growth Factor Rev, 2008. **19**(2): p. 157-65.
140. Kohno, T., et al., *KIF5B-RET fusions in lung adenocarcinoma.* Nat Med, 2012. **18**(3): p. 375-7.
141. Takeuchi, K., et al., *RET, ROS1 and ALK fusions in lung cancer.* Nat Med, 2012. **18**(3): p. 378-81.

142. Pao, W. and K.E. Hutchinson, *Chipping away at the lung cancer genome.* Nat Med, 2012. **18**(3): p. 349-51.
143. Hirsch, F.R., et al., *Evaluation of HER-2/neu gene amplification and protein expression in non-small cell lung carcinomas.* Br J Cancer, 2002. **86**(9): p. 1449-56.
144. Okudela, K., et al., *PIK3CA mutation and amplification in human lung cancer.* Pathol Int, 2007. **57**(10): p. 664-71.
145. Onozato, R., et al., *Activation of MET by gene amplification or by splice mutations deleting the juxtamembrane domain in primary resected lung cancers.* J Thorac Oncol, 2009. **4**(1): p. 5-11.
146. 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.
147. Gazdar, A.F., et al., *Lung cancer cell lines as tools for biomedical discovery and research.* J Natl Cancer Inst, 2010. **102**(17): p. 1310-21.
148. Beroukhim, R., et al., *Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma.* Proc Natl Acad Sci U S A, 2007. **104**(50): p. 20007-12.
149. Coe, B.P., et al., *FACADE: a fast and sensitive algorithm for the segmentation and calling of high resolution array CGH data.* Nucleic Acids Res, 2010. **38**(15): p. e157.
150. Chari, R., et al., *A sequence-based approach to identify reference genes for gene expression analysis.* BMC Med Genomics, 2010. **3**: p. 32.
151. Thu, K.L., et al., *Genetic disruption of KEAP1/CUL3 E3 ubiquitin ligase complex components is a key mechanism of NF-kappaB pathway activation in lung cancer.* J Thorac Oncol, 2011. **6**(9): p. 1521-9.
152. Zhao, X., et al., *Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis.* Cancer Res, 2005. **65**(13): p. 5561-70.
153. Job, B., et al., *Genomic aberrations in lung adenocarcinoma in never smokers.* PLoS One, 2010. **5**(12): p. e15145.
154. Fischer, U., et al., *Cloning of a novel transcription factor-like gene amplified in human glioma including astrocytoma grade I.* Hum Mol Genet, 1997. **6**(11): p. 1817-22.
155. Munnia, A., et al., *Expression, cellular distribution and protein binding of the glioma amplified sequence (GAS41), a highly conserved putative transcription factor.* Oncogene, 2001. **20**(35): p. 4853-63.
156. Barretina, J., et al., *Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy.* Nat Genet, 2010. **42**(8): p. 715-21.
157. Shedden, K., et al., *Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study.* Nat Med, 2008. **14**(8): p. 822-7.
158. Petitjean, A., et al., *Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database.* Hum Mutat, 2007. **28**(6): p. 622-9.
159. Park, J.H. and R.G. Roeder, *GAS41 is required for repression of the p53 tumor suppressor pathway during normal cellular proliferation.* Mol Cell Biol, 2006. **26**(11): p. 4006-16.

160. Medjkane, S., et al., *Myocardin-related transcription factors and SRF are required for cytoskeletal dynamics and experimental metastasis.* Nat Cell Biol, 2009. **11**(3): p. 257-68.
161. Prencipe, M., et al., *Identification of transcription factors associated with castration-resistance: is the serum responsive factor a potential therapeutic target?* Prostate, 2013. **73**(7): p. 743-53.
162. Zimmermann, K., et al., *Targeted disruption of the GAS41 gene encoding a putative transcription factor indicates that GAS41 is essential for cell viability.* J Biol Chem, 2002. **277**(21): p. 18626-31.
163. Schulze, J.M., A.Y. Wang, and M.S. Kobor, *YEATS domain proteins: a diverse family with many links to chromatin modification and transcription.* Biochem Cell Biol, 2009. **87**(1): p. 65-75.
164. Cai, Y., et al., *The mammalian YL1 protein is a shared subunit of the TRRAP/TIP60 histone acetyltransferase and SRCAP complexes.* J Biol Chem, 2005. **280**(14): p. 13665-70.
165. Doyon, Y., et al., *Structural and functional conservation of the NuA4 histone acetyltransferase complex from yeast to humans.* Mol Cell Biol, 2004. **24**(5): p. 1884-96.
166. Piccinni, E., et al., *Direct interaction of Gas41 and Myc encoded by amplified genes in nervous system tumours.* Acta Biochim Pol, 2011. **58**(4): p. 529-34.
167. Lauffart, B., et al., *Interaction of the transforming acidic coiled-coil 1 (TACC1) protein with ch-TOG and GAS41/NuBI1 suggests multiple TACC1-containing protein complexes in human cells.* Biochem J, 2002. **363**(Pt 1): p. 195-200.
168. Harborth, J., K. Weber, and M. Osborn, *GAS41, a highly conserved protein in eukaryotic nuclei, binds to NuMA.* J Biol Chem, 2000. **275**(41): p. 31979-85.
169. Debernardi, S., et al., *The MLL fusion partner AF10 binds GAS41, a protein that interacts with the human SWI/SNF complex.* Blood, 2002. **99**(1): p. 275-81.
170. Schmitt, J., et al., *GAS41 amplification results in overexpression of a new spindle pole protein.* Genes Chromosomes Cancer, 2012. **51**(9): p. 868-80.
171. Marine, J.C. and G. Lozano, *Mdm2-mediated ubiquitylation: p53 and beyond.* Cell Death Differ, 2010. **17**(1): p. 93-102.
172. Manfredi, J.J., *The Mdm2-p53 relationship evolves: Mdm2 swings both ways as an oncogene and a tumor suppressor.* Genes Dev, 2010. **24**(15): p. 1580-9.
173. Momand, J., et al., *The MDM2 gene amplification database.* Nucleic Acids Res, 1998. **26**(15): p. 3453-9.
174. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
175. Ponting, C.P. and T.G. Belgard, *Transcribed dark matter: meaning or myth?* Hum Mol Genet, 2010. **19**(R2): p. R162-8.
176. Carninci, P., et al., *The transcriptional landscape of the mammalian genome.* Science, 2005. **309**(5740): p. 1559-63.
177. Hung, T. and H.Y. Chang, *Long noncoding RNA in genome regulation: prospects and mechanisms.* RNA Biol, 2010. **7**(5): p. 582-5.
178. Taft, R.J., et al., *Non-coding RNAs: regulators of disease.* J Pathol, 2010. **220**(2): p. 126-39.
179. Zhang, L., et al., *microRNAs exhibit high frequency genomic alterations in human cancer.* Proc Natl Acad Sci U S A, 2006. **103**(24): p. 9136-41.

180. Calin, G.A., et al., *Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers.* Proc Natl Acad Sci U S A, 2004. **101**(9): p. 2999-3004.
181. Kozomara, A. and S. Griffiths-Jones, *miRBase: annotating high confidence microRNAs using deep sequencing data.* Nucleic Acids Res, 2014. **42**(Database issue): p. D68-73.
182. Iorio, M.V. and C.M. Croce, *MicroRNAs in cancer: small molecules with a huge impact.* J Clin Oncol, 2009. **27**(34): p. 5848-56.
183. Yanaihara, N., et al., *Unique microRNA molecular profiles in lung cancer diagnosis and prognosis.* Cancer Cell, 2006. **9**(3): p. 189-98.
184. Takamizawa, J., et al., *Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival.* Cancer Res, 2004. **64**(11): p. 3753-6.
185. Garofalo, M., et al., *miR-221&222 regulate TRAIL resistance and enhance tumorigenicity through PTEN and TIMP3 downregulation.* Cancer Cell, 2009. **16**(6): p. 498-509.
186. Gallardo, E., et al., *miR-34a as a prognostic marker of relapse in surgically resected non-small-cell lung cancer.* Carcinogenesis, 2009. **30**(11): p. 1903-9.
187. Jiang, L., et al., *Hsa-miR-125a-3p and hsa-miR-125a-5p are downregulated in non-small cell lung cancer and have inverse effects on invasion and migration of lung cancer cells.* BMC Cancer, 2010. **10**: p. 318.
188. Markou, A., et al., *Prognostic value of mature microRNA-21 and microRNA-205 overexpression in non-small cell lung cancer by quantitative real-time RT-PCR.* Clin Chem, 2008. **54**(10): p. 1696-704.
189. Yu, S.L., et al., *MicroRNA signature predicts survival and relapse in lung cancer.* Cancer Cell, 2008. **13**(1): p. 48-57.
190. *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
191. 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.
192. Pio, G., et al., *Integrating microRNA target predictions for the discovery of gene regulatory networks: a semi-supervised ensemble learning approach.* BMC Bioinformatics, 2014. **15 Suppl 1**: p. S4.
193. 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.
194. Brown, K.R., et al., *NAViGaTOR: Network Analysis, Visualization and Graphing Toronto.* Bioinformatics, 2009. **25**(24): p. 3327-9.
195. Lockwood, W.W., et al., *Integrative genomics identified RFC3 as an amplified candidate oncogene in esophageal adenocarcinoma.* Clin Cancer Res, 2012. **18**(7): p. 1936-46.
196. Takahashi, Y., et al., *MiR-107 and MiR-185 can induce cell cycle arrest in human non small cell lung cancer cell lines.* PLoS One, 2009. **4**(8): p. e6677.
197. Zhu, W., et al., *Expression of miR-29c, miR-93, and miR-429 as potential biomarkers for detection of early stage non-small lung cancer.* PLoS One, 2014. **9**(2): p. e87780.

198. Zhu, C.Q., et al., *Understanding prognostic gene expression signatures in lung cancer.* Clin Lung Cancer, 2009. **10**(5): p. 331-40.
199. Landi, M.T., et al., *MicroRNA expression differentiates histology and predicts survival of lung cancer.* Clin Cancer Res, 2010. **16**(2): p. 430-41.
200. 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.
201. Bishop, J.A., et al., *Accurate classification of non-small cell lung carcinoma using a novel microRNA-based approach.* Clin Cancer Res, 2010. **16**(2): p. 610-9.
202. Vosa, U., et al., *Meta-analysis of microRNA expression in lung cancer.* Int J Cancer, 2013. **132**(12): p. 2884-93.
203. Guan, P., et al., *Meta-analysis of human lung cancer microRNA expression profiling studies comparing cancer tissues with normal tissues.* J Exp Clin Cancer Res, 2012. **31**: p. 54.
204. Hamamoto, J., et al., *Identification of microRNAs differentially expressed between lung squamous cell carcinoma and lung adenocarcinoma.* Mol Med Rep, 2013. **8**(2): p. 456-62.
205. 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.
206. Lazar, V., et al., *Integrated molecular portrait of non-small cell lung cancers.* BMC Med Genomics, 2013. **6**: p. 53.
207. Yan, J.W., J.S. Lin, and X.X. He, *The emerging role of miR-375 in cancer.* Int J Cancer, 2013.
208. Jonsdottir, K., et al., *Validation of expression patterns for nine miRNAs in 204 lymph-node negative breast cancers.* PLoS One, 2012. **7**(11): p. e48692.
209. Mathe, E.A., et al., *MicroRNA expression in squamous cell carcinoma and adenocarcinoma of the esophagus: associations with survival.* Clin Cancer Res, 2009. **15**(19): p. 6192-200.
210. Nishikawa, E., et al., *miR-375 is activated by ASH1 and inhibits YAP1 in a lineage-dependent manner in lung cancer.* Cancer Res, 2011. **71**(19): p. 6165-73.
211. Tan, M., et al., *SAG/ROC2/RBX2 is a HIF-1 target gene that promotes HIF-1 alpha ubiquitination and degradation.* Oncogene, 2008. **27**(10): p. 1404-11.
212. Nie, L., et al., *Notch-induced Asb2 expression promotes protein ubiquitination by forming non-canonical E3 ligase complexes.* Cell Res, 2011. **21**(5): p. 754-69.
213. Sato, Y., et al., *Degradation of phosphorylated p53 by viral protein-ECS E3 ligase complex.* PLoS Pathog, 2009. **5**(7): p. e1000530.
214. Ishikawa, K., H. Ishii, and T. Saito, *DNA damage-dependent cell cycle checkpoints and genomic stability.* DNA Cell Biol, 2006. **25**(7): p. 406-11.
215. Sierant, M.L., N.E. Archer, and S.K. Davey, *The Rad9A checkpoint protein is required for nuclear localization of the claspin adaptor protein.* Cell Cycle, 2010. **9**(3): p. 548-56.
216. Maniwa, Y., et al., *His239Arg SNP of HRAD9 is associated with lung adenocarcinoma.* Cancer, 2006. **106**(5): p. 1117-22.
217. Margueron, R. and D. Reinberg, *The Polycomb complex PRC2 and its mark in life.* Nature, 2011. **469**(7330): p. 343-9.

218. Zhou, V.W., A. Goren, and B.E. Bernstein, *Charting histone modifications and the functional organization of mammalian genomes.* Nat Rev Genet, 2011. **12**(1): p. 7-18.
219. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
220. Hall, R.D., J.E. Gray, and A.A. Chiappori, *Beyond the standard of care: a review of novel immunotherapy trials for the treatment of lung cancer.* Cancer Control, 2013. **20**(1): p. 22-31.
221. Zielinski, C., et al., *Rationale for targeting the immune system through checkpoint molecule blockade in the treatment of non-small-cell lung cancer.* Ann Oncol, 2013. **24**(5): p. 1170-9.
222. Decoster, L., I. Wauters, and J.F. Vansteenkiste, *Vaccination therapy for non-small-cell lung cancer: review of agents in phase III development.* Ann Oncol, 2012. **23**(6): p. 1387-93.
223. Akbay, E.A., et al., *Activation of the PD-1 pathway contributes to immune escape in EGFR-driven lung tumors.* Cancer Discov, 2013. **3**(12): p. 1355-63.

Appendix

Appendix A Sample sets and demographic data of cohorts used in this thesis

A.1 Samples sets used in Chapter 2

Sample Set	Cohort	# Samples	# AC	# SqCC	# Non-Neoplasitic Lung Tissue	Assay Type
1	BC Cancer Agency - Vancouver	261	169	92	0	Copy Number - BCCRC Whole Genome Tiling Path Array CGH
2	BC Cancer Agency - Vancouver	92	30	13	48	DNA Methylation - Illumina HumanMethylation27 chip
3	BC Cancer Agency - Vancouver	49	29	20	0	Gene Expression - Custom Affymetrix
4	GEO Duke University – GSE3141	111	58	53	0	Gene Expression - Affymetrix GeneChip Human Genome U133 Plus 2.0 Array
5	BC Cancer Agency – Vancouver	0	0	0	67	Gene Expression - Affymetrix GeneChip Human Genome U133 Plus 2.0 Array
6	GEO Samsung Medical Center - GSE8894	138	62	76	0	Gene Expression - Affymetrix GeneChip Human Genome U133 Plus 2.0 Array

A.2 Demographic data of BCCRC 261 AC and 92 SqCC with copy number profiles

		AC (n=169)	SqCC (n=92)
Stage	I	76 (44.9%)	32 (34.7%)
	II	40 (23.6%)	32 (34.7%)
	III	22 (13%)	14 (15.2%)
	IV	27 (16%)	10 (10.9%)
Sex	Male	63 (37.3%)	66 (71.7%)
	Female	106 (62.7%)	26 (28.3%)
Age	Median	66	69
	Age Range	35-90	48-88
Smoking Status	Current smoker	48 (28.4%)	30 (32.6%)
	Former smoker	80 (47.3%)	61 (66.3%)
	Never smoker	38 (22.5%)	
	N/A	3 (1.8%)	1 (1.1%)

A.3 Datasets used to determine the prevalence and subtype specificity of YEATS4 amplification and overexpression

DataSet	Samples	Platform	Institute	Website
1	169 AC & 92 SqCC tumors	aCGH	BCCRC	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31798
2	35 AC & 13 SqCC with matched non-malignant tissue	Custom Affymetrix expression arrays	BCCRC	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31799
3	59 AC with matched non-malignant tissue	qRT-PCR	BCCRC	N/A
4	18 NSCLC cell lines	qRT-PCR	BCCRC	N/A
5	83 AC & matched non-malignant tissue	SNP 6, Illumina WG-6 v3 BeadChip Arrays	Early Detection Research Network (EDRN)/Canary Foundation	http://edrn.nci.nih.gov/science-data
6	155 SqCC & 77AC with matched non-malignant tissue	SNP 6	Max Plank Institute (GSE25016)	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25016
7	354 AC with matched non-malignant tissue	SNP 6	Broad Institute- Database of genotypes and phenotypes	http://www.broadinstitute.org/cgibin/cancer/publications/pub_paper.cgi?mode=view&paper_id=169
8	199 AC with matched non-malignant tissue	Agilent 44K aCGH	Memorial Sloan Kettering Cancer Center	http://cbio.mskcc.org/Public/lung_array_data/
9	508 cell lines from multiple human cancers	SNP6	Sanger Cell Line Project	http://www.sanger.ac.uk/genetics/CGP/CellLines/
10	277 AC & 201 SqCC; 25 AC & 17 SqCC with non-malignant tissue	SNP6 and RNA seq	The Cancer Genome Atlas	https://tcga-data.nci.nih.gov/tcga/

Appendix B List of publications

This Appendix lists all of the publications I contributed to during my degree that were either published, accepted, currently in submission or prepared for submission. In total there are 26 publications. First or co-first authorships are underlined.

1. Thu KL, **Pikor L**, Kennett J, Alvarez C, Lam WL (2010) Methylation analysis by DNA immunoprecipitation. *Journal of Cellular Physiology* 222: 522-31
2. Raj Chari, Kim M. Lonergan, **Larissa A. Pikor**, Bradley P. Coe, Chang Qi Zhu, Timothy H. W. Chan, Calum MacAulay, Ming-Sound Tsao, Stephen Lam, Raymond T. Ng, Wan L. Lam (2010) A sequenced-based approach to identify reference genes for gene expression analysis. *BMC Medical Genomics* 3:32,1-11.
3. Katey S. S. Enfield, Greg L Stewart, **Larissa Pikor**, Carlos E. Alvarez, Stephen Lam, Wan L. Lam and Raj Chari (2011) MircroRNA gene dosage alterations and drug response in lung cancer. *Journal of Biomedicine and Biotechnology* 2011:474632,1-15
4. **Pikor LA**, Enfield K, Heryet C, Lam WL (2011). DNA extraction from paraffin embedded material. *Journal of Visualized Experiments Mar 26;(49)*:*2763*
5. Thu KL, **Pikor LA**, Chari R, Wilson IM, MacAulay CE, Gazdar AF, Lam S, Lam WL, Lockwood WW (2011) Disruption of KEAP1/CUL3 E3 ubiquitin ligase complex components is a key mechanism for NFkB pathway activation in lung cancer. *Journal of Thoracic Oncology* 6(9):1521-9.
6. WW Lockwood, K Thu, L Lin, **L Pikor**, R Chari, WL Lam, DG Beer (2012) Integrative genomics identified RFC3 as an amplified candidate oncogene in esophageal adenocarcinoma. *Clinical Cancer Research* 1;18(7): 1936-46
7. Enfield KSS, **Pikor LA**, Martinez VD, Lam WL (2012). Mechanistic roles of non-coding RNAs in lung cancer biology and their clinical implications. *Genetics Research International* 2012:737416, 1-16
8. Lockwood WW, Wilson IM, Coe BP, Chari R, **Pikor LA**, Thu KL, Yee J, English J, Murray N, Tsao MS, Minna J, Gazdar AF, MacAulay CE, Lam S, Lam WL (2012). Divergent genomic and epigenomic landscapes of lung cancer subtypes underscore the selection of different oncogenic pathways during tumor development. *PLoS One* 2012;7(5) e37775
9. **Pikor L**, Thu K, Vucic E, Lam W (2013). The detection and implication of genome instability in cancer. *Cancer Metastasis Review;32 (3-4):341-52*
10. Crea F, Sun L, **Pikor L**, Lam WL, Helgason CD (2013). Mutational analysis of Polycomb genes in solid tumors identifies PHC3 amplification as a possible cancer-driving genetic alteration. *British Journal of Cancer 109(6):1699-702*
11. **Pikor LA**, Ramnarine VR, Lam S, Lam WL (2013). Genetic alterations defining NSCLC subtypes and their therapeutic implications. *Lung Cancer 82(2)179-189*
12. Martinez VD, Vucic EA, **Pikor LA**, Thu KL, Hubaux R, Lam WL (2013). Frequent concerted genetic mechanisms disrupt multiple components of the NRF2 inhibtor KEAP1/CUL3/RBX1 E3-ubiquitin ligase complex in thyroid cancer. *Molecular Cancer 12:124,1-6*

13. **Pikor LA**, Lockwood WW, Chari R, Vucic EA, Lam S, Lam WL (2013). YEATS4 is a novel oncogene amplified in non-small cell lung cancer that regulates the p53 pathway. *Cancer Research 73(24):7301-12.*
14. Thu KL, Radulovich N, Becker-Santos DD, **Pikor LA**, Pusic A, Lockwood WW, Tsao MS, Lam WL (2014). SOX15 is a novel tumor suppressor in pancreatic cancer with a role in Wnt/β-catenin signaling. *Oncogene 33(3):379-88*
15. Martinez VD, Vucic EA, Thu KL, **Pikor LA**, Lam S, Lam WL (2013) 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 & Nec*k, In press.
16. Crea F, Watahiki A, Quagliata L, Xue H, **Pikor L**, Parolia A,Wang Y, LinD, Lam WL, Farrar WL, Isogai T, Morant R, Castori-Eppenberger S, Chi KN, Wang Y, Helgason CD (2014) Identification of a long non-coding RNA as a novel biomarker and potential therapeutic targetd for metastatic prostate cancer. *Oncotrarget*, In press.
17. Vucic EA, Thu KL, **Pikor LA**, Enfield KSS, Yee J, English JC, MacAulay CE, Lam S, Jurisica I, Lam WL (2014) Smoking status impacts microRNA mediated prognosis in lung adenocarcinoma biology. *Molecular Cancer*, Under Review.
18. Martinez VD, Vucic EA, Thu KL, **Pikor LA**, Hubaux R, Lam WL (2014) Unique pattern of component gene disruption in the NRF2 inhibitor KEAP1/CUL3/RBX1 E3-ubiquiting ligase complex in serous ovarian cancer. BioMed Research International, Under Review.
19. Hubaux R, Thu KL, Vucic EA, **Pikor LA**, Kung SHY, Martinez VD, Mosslemi M, Becker-Santos DD, Gazdar AF, Lam S, Lam WL (2014) Microtubule affinity-regulating kinase 2 contributes to cisplatin sensitivity through modulation of the DNA damage response in non-small cell lung cancer. *Oncotarget*. Under review.
20. Crea F, Quagliata L, Frumento P, Azad A, Xue H, **Pikor LA**, Watahiki A, Morant R, Castori-Eppenberger S, Wang Y, Parolia A, Lam WL, Gleave M, Chi KN, Wang YZ, Helgason CD (2014) Integrated analysis of the prostate cancer small-nucleolar transcriptome reveals snoRNA55 as a novel diagnostic and prognostic biomarker. *Annals of Oncology*. Under review.

Manuscripts in preparation.

21. **Pikor LA**, Chari R, Kennett JY, Solis LM, Valencia I, Wistuba I, Gazdar AF, Lam S, Lam WL (2014) SIRPA is a candidate tumor suppressor in lung adenocarcinoma that induces senescence mediated by *RB* and is associated with *EGFR* mutation. In preparation.
22. **Pikor LA**, Thu KL, Vucic EA, Lam S, Lam WL (2014). miRNA sequencing identifies miRNAs that differentiate between adenocarcinoma and squamous cell carcinoma and distinguish tumors from non-malignant tissue and benign lesions.
23. Thu KL, Radulovich N, Becker-Santos DD, **Pikor LA**, Lam WL, Tsao MS (2014) SOX15 and other SOX family members are important mediators of tumorigenesis in multiple cancer types. *OncoScience*. In preparation.
24. Becker-Santos DD, Thu KL, **Pikor LA**, Vucic EA, MacAulay CE, Jurisica I, Robinson WP, Lam S, Lam WL (2014). miRNA expression patterns in human non-small cell lung cancer and fetal lung: a comparative study. In preparation for *Journal of Thoracic Oncology*

25. Conway E, **Pikor LA**, Lam S, Lam WL (2014). Macrophages, inflammation and cancer. In preparation.
26. Mosslemi M, Thu KL, Vucic EA, **Pikor LA**, Ng RT, MacAulay CE, Lam WL (2014). Development of a Multi-dimensional Integrative Tumor gene Ranking Algorithm (MITRA) for the identification of candidate genes in cancer. In preparation