Abstract

**Background:** Lung cancer is the world's leading cause of cancer mortality. The main factors contributing to this are the late stage of disease at the time of diagnosis and a lack of effective chemotherapeutic strategies. A better understanding of the molecular origins and basic biology of lung cancer will lead to the development of early detection techniques and novel therapies to address these issues. I applied an integrative genomics approach utilizing high-resolution whole genome profiling technologies to uncover causal gene disruption in lung cancer and identify candidates associated with the development of specific lung cancer subtypes.

**Hypotheses:** (i) Genes key to lung tumorigenesis will be identified in recurrently altered genomic regions. (ii) Lung cancer subtypes require distinct genetic alterations for neoplastic development.

**Materials/Methods:** DNA copy number data from lung cancer specimens were integrated with expression data to identify genes contributing to tumorigenesis. Subsequently, this approach was applied to compare lung cancer subtypes to discover genes and pathways specifically disrupted in each. Quantitative RT-PCR and immunohistochemistry were performed to validate results from microarray experiments and cell models were utilized to confirm the functional significance of identified genes.

**Results:** I identified novel gene candidates frequently deregulated in lung cancer which contribute to tumorigenesis, supporting the first hypothesis. In addition, the comparison of lung cancer subtypes identified subtype-specific genetic events and delineated genes and pathways important in their differential development, supporting the second hypothesis. Significantly, I discovered a novel squamous cell lineage specific oncogene, *BRF2*, which affects polymerase III transcribed genes.

**Conclusions:** Integrative genomic analysis is an effective means for identifying key gene disruptions in lung cancer. Furthermore, these findings suggest that lung cancer subtypes require distinct genetic alterations for tumorigenesis, uncovering the specific targets disrupted by these alterations for the first time. Most importantly, activation of *BRF2* represents a novel mechanism of tumorigenesis through the increase of polymerase III mediated transcription and the targeted activation of this gene in SqCC suggests that it may be an excellent candidate for new treatment strategies tailored to this subtype. Together, this work highlights the need for tailoring therapies to the specific cancer subtypes.
# Table of Contents

Abstract ......................................................................................................................................... ii  
Table of Contents ......................................................................................................................... iii  
List of Tables ................................................................................................................................ ix  
List of Figures ............................................................................................................................... x  
Acknowledgements ..................................................................................................................... xii  
Dedication .................................................................................................................................. xiv  
Co-Authorship Statement ............................................................................................................ xv  
Chapter 1: Introduction ................................................................................................................. 1  
1.1 Introduction to Lung Cancer ................................................................................................. 2  
1.2. Lung Anatomy and Histology ............................................................................................. 2  
1.3 Lung Cancer Subtypes and Clinical Features ..................................................................... 3  
  1.3.1 Small cell lung cancer .................................................................................................. 3  
  1.3.2 Non-small cell lung cancer ........................................................................................... 4  
1.4 Molecular Biology of Lung Cancer ...................................................................................... 6  
  1.4.1 Genetic alterations ........................................................................................................ 6  
  1.4.2 Epigenetic alterations ................................................................................................... 7  
  1.4.3 Signalling pathways ...................................................................................................... 8  
  1.4.4 Targeted therapies ........................................................................................................ 9  
1.5 Cell Lineage Hypothesis of Tumorigenesis ....................................................................... 10  
  1.5.1 Cell lineage model ...................................................................................................... 10  
  1.5.2 Lineage/subtype specificity ......................................................................................... 10  
  1.5.3 Therapeutic implications ............................................................................................. 11  
1.6 Global Profiling Technologies for the Analysis of Cancer Genomes ................................. 12  
  1.6.1 Gene expression microarray analysis ......................................................................... 12  
  1.6.2 Array comparative genomic hybridization ................................................................... 13  
  1.6.3 Integrative analysis ..................................................................................................... 14  
1.7 Thesis Theme and Rationale for Study ............................................................................. 15  
1.8 Objectives and Hypotheses ............................................................................................... 15  
1.9 Specific Aims and Thesis Outline ...................................................................................... 16  
1.10 References ...................................................................................................................... 20
Chapter 2: High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH ................................................................. 26

2.1 Introduction ........................................................................................................... 27

2.2 Results and Discussion .......................................................................................... 28

2.2.1 Whole genome segmental copy number profiling ................................................. 28

2.2.2 Frequent copy number alterations ....................................................................... 28

2.2.3 Regions commonly altered in NSCLC ................................................................. 29

2.2.4 Copy number decrease ....................................................................................... 30

2.2.5 Array CGH vs. LOH ......................................................................................... 30

2.2.6 Loci methylated in lung cancer .......................................................................... 31

2.2.7 Segmental amplifications .................................................................................. 31

2.2.8 Multiple regions on chromosome 7 .................................................................... 32

2.2.9 Squamous cell carcinoma vs. adenocarcinoma ................................................. 33

2.3 Conclusions .......................................................................................................... 34

2.4 Materials and Methods .......................................................................................... 35

2.4.1 Sample collection ............................................................................................... 35

2.4.2 Array construction ............................................................................................. 35

2.4.3 Probe labeling and hybridization ....................................................................... 35

2.4.4 Imaging and analysis ......................................................................................... 35

2.5 References ............................................................................................................ 44

Chapter 3: Integrative genomic and gene expression analysis of chromosome 7 identified novel oncogene loci in non-small cell lung cancer .......................................................... 49

3.1 Introduction .......................................................................................................... 50

3.2 Results and Discussion .......................................................................................... 51

3.2.1 High level amplification events in frequently altered regions ............................... 51

3.2.2 Integration of gene dosage and expression data .................................................... 51

3.2.3 Quantitative RT-PCR validation in cell lines ....................................................... 52

3.2.4 Validation of genes of interest using gene expression data for clinical NSCLC tumours ......................................................................................... 52

3.2.5 Quantitative PCR validation of FTSJ2, NUDT1, TAF6, and POLR2J in clinical samples ................................................................................................. 53

3.3 Materials and Methods .......................................................................................... 54

3.3.1 Cell line samples and DNA extraction ............................................................... 54

3.3.2 Tiling path array comparative genomic hybridization and data analysis ............ 54
Chapter 7: BRF2 is a lineage specific oncogene amplified early in squamous cell lung cancer development ................................................................. 142

7.1 Introduction ............................................................................................................................................................................. 143

7.2 Results and Discussion ................................................................................................................................................................. 143

7.2.1 8p amplification is restricted to the SqCC cancer type .............................................................................................................. 143

7.2.2 BRF2 gene expression drives selection of the 8p amplicon in lung SqCC .................................................................................. 144

7.2.3 BRF2 contributes to SqCC tumorigenesis by regulating cell growth and proliferation .............................................................. 146

7.2.4 BRF2 activation is an early event in SqCC development ............................................................................................................ 148

7.2.5 Increased RNA processing is associated with BRF2 overexpression ..................................................................................... 149

7.3 Conclusions .................................................................................................................................................................................. 150

7.4 Materials and Methods ....................................................................................................................................................................... 151

7.4.1 DNA samples .............................................................................................................................................................................. 151

7.4.2 Tiling path array comparative genomic hybridization .................................................................................................................. 151

7.4.3 Comparison of cell type alteration frequencies .......................................................................................................................... 152

7.4.4 Gene expression microarray analysis of clinical tumor specimens .............................................................................................. 153

7.4.5 Gene expression microarray analysis of normal bronchial epithelial cells ............................................................................... 153

7.4.6 Statistical analysis of gene expression data .............................................................................................................................. 153

7.4.7 Integration of genetic and gene expression data ........................................................................................................................ 153

7.4.8 Reverse transcriptase polymerase chain reaction analysis of transcription levels in clinical tumor samples .......................................................... 154

7.4.9 Cell lines and culture conditions .............................................................................................................................................. 154

7.4.10 TaqMan analysis of transcription levels in cancer cell lines ...................................................................................................... 155

7.4.11 Western blot analysis of protein levels ........................................................................................................................................ 155

7.4.12 RNAi knockdown ........................................................................................................................................................................ 155

7.4.13 3-[4, 5-dimethylthiazol-2-yl]-2, 5-diphenyltetrazolium bromide (MTT) assay ............................................................................. 156

7.4.14 Construction of the BRF2 expression vector ............................................................................................................................ 156

7.4.15 In vitro cell growth assays ......................................................................................................................................................... 157
7.4.16 Immunohistochemistry ................................................................. 157
7.4.17 Significance analysis of microarrays (SAM) .............................. 157
7.4.18 Functional assessment of BRF2 associated genes .................... 158
7.5 References ...................................................................................... 164
Chapter 8: Conclusions ....................................................................... 168
8.1 Summary ....................................................................................... 169
  8.1.1 Development and application of integrative genomic approaches for the study of lung cancer ................................................................. 169
  8.1.2 Comparison of lung cancer subtypes .......................................... 171
8.2 Significance and Conclusions .......................................................... 173
  8.2.1 Novel genetic alterations and candidate genes involved in lung tumorigenesis ................................................................. 173
  8.2.2 Genetic mechanisms involved in the development of lung cancer subtypes ................................................................. 174
  8.2.3 Developing new therapeutic strategies for lung cancer treatment ................................................................. 175
8.3 Future Directions ........................................................................... 177
  8.3.1 Mechanism of BRF2 mediated tumorigenesis and assessment of Pol III as a therapeutic target in lung cancer ................................................................. 177
  8.3.2 Functional and clinical characterization of gene candidates ................................................................. 178
  8.3.3 Validating the lineage specific tumorigenic potential of subtype specific genes ................................................................. 178
  8.3.4 Refinement of lung cancer subtypes ........................................... 179
  8.3.5 Developmental signalling pathways in normal lineage development and cancer ................................................................. 180
  8.3.6 Multidimensional integrative analysis of lung tumor genomes ................................................................. 180
8.4 References ...................................................................................... 182
List of Tables

Table 3.1. Amplified and overexpressed genes within regions of recurrent genomic gain on chromosome 7 in NSCLC cell lines ........................................................................................................60
Table 4.1. Summary and distribution of amplicons by cancer type..................................................82
Table 4.2. Canonical pathways affected by amplification in NSCLC .................................................83
Table 5.1. Differential deregulation of genes in key biochemical pathways between NSCLC and SCLC ........................................................................................................................................106
Table 6.1. Top gene networks associated with AC and SqCC PSCNA-regulated targets...........132
Table 6.2. PSCNA-regulated genes associated with survival in each subtype.........................133
Table 6.3. Genes previously implicated in NSCLC displaying subtype specific disruption.......135
List of Figures

Figure 1.1. Principles of array comparative genomic hybridization .................................................. 19
Figure 2.1. Whole genome profile of squamous cell carcinoma line HCC95. .................................... 37
Figure 2.2. Whole genome evaluation of genetic alteration frequency for NSCLC cell lines ............. 38
Figure 2.3. High level amplification at 8q24. .................................................................................. 39
Figure 2.4. High level amplification at 14q13. .............................................................................. 40
Figure 2.5. Multiple amplifications on chromosome 7. ................................................................. 41
Figure 2.6. Multiple segmental amplifications on 7q. ................................................................. 42
Figure 2.7. Comparison of lung adenocarcinoma and squamous cell carcinoma genomes .............. 43
Figure 3.1. Representative genomic alterations within chromosome 7. ....................................... 57
Figure 3.2. Validation of amplified and over-expressed genes from candidate regions in a separate cohort of 111 NSCLC clinical tumours ................................................................. 58
Figure 3.3. RT-qPCR results of candidate oncogenes in ten matched tumour/normal clinical samples ......................................................................................................................... 59
Figure 4.1. Hotspots of amplification in cancer genomes. ............................................................. 78
Figure 4.2. Impact of amplification on gene transcription levels. .................................................... 79
Figure 4.3. Frequent amplification and overexpression of multiple EGFR family signaling components in NSCLC. ........................................................................................................ 80
Figure 4.4. SHC1 disruption in NSCLC cell lines and clinical tumors ............................................. 81
Figure 5.1. SMRT array profile of the SCLC NCI-H1672 cells. ...................................................... 100
Figure 5.2. Copy number alterations in SCLC and NSCLC .......................................................... 101
Figure 5.3. Differential expression as a result of copy number alteration ..................................... 102
Figure 5.4. Contribution of copy number induced gene expression differences to the SCLC and NSCLC phenotypes .................................................................................................. 104
Figure 5.5. Differential targets of copy number induced expression changes in key biochemical pathways between SCLC and NSCLC ................................................................. 105
Figure 6.1. Copy number alterations in AC and SqCC ................................................................. 127
Figure 6.2. Differential expression as a result of PSCNAs ............................................................. 128
Figure 6.3. Genes deregulated by PSCNAs contribute to AC and SqCC phenotypes ................... 129
Figure 6.4. Gene networks involved in the development of SqCC and AC .................................. 130
Figure 6.5. ELAVL1 is a PSCNA-regulated gene that predicts poor survival in SqCC ........................ 131
Figure 7.1. Chromosome 8p amplification in NSCLC is restricted to the SqCC lineage ............... 159
Figure 7.2. BRG2 is a lineage specific oncogene targeted by amplification in SqCC ............... 160
Figure 7.3. BRF2 activation contributes to cell growth and proliferation................................................... 161
Figure 7.4. Amplification and overexpression of BRF2 in preinvasive SqCC lesions. ....................... 162
Figure 7.5. BRF2 expression in SqCC precancerous stages. ................................................................. 163
Acknowledgements

I would like to acknowledge the contributions of the many members of the Wan Lam Laboratory who contributed to this work, in particular the co-authors of each of the manuscript chapters presented herein. In addition, I would like to thank the grant support and scholarships which supported the research included in this thesis. Specific acknowledgements from the published versions of each chapter are detailed below:

Chapter 2: The authors thank S. K. Watson for array synthesis, J. J. Davies, B. Chi and R. J. De Leeuw for guidance in data analysis and T. P. H. Buys for careful reading of the manuscript. Scholarships to C.G. and W.W.L. from the Michael Smith Foundation of Health Research and the Natural Sciences and Engineering Research Council.

Chapter 3: This work was supported by funds from Genome Canada/BC and Canadian Institutes for Health Research (CIHR) as well as scholarships from the Natural Sciences and Engineering Research Council, the Michael Smith Foundation for Health Research, and CIHR to W.W.L, T.P.H.B., R.C., and B.P.C.

Chapter 4: This work was supported by funds from CIHR, Genome Canada/BC, Lung Cancer SPORE P50CA70907, DOD VITAL, the Gillson Longenbaugh and Anderson Charitable Foundations as well as scholarships from NSERC, CIHR and MSFHR to WWL, RC and BPC.

Chapter 5: We thank SK Watson for array synthesis and JJ Davies for useful discussion. This work was supported by funds from the CIHR, National Cancer Institute of Canada, Genome British Columbia/Genome Canada, Lung Cancer SPORE P50CA70907, NIH (USA) Grant 1U01CA96109 and scholarships to BPC and WWL from the Michael Smith Foundation for Health Research and the Natural Sciences and Engineering Research Council.

Chapter 6: This work was supported by funds from CIHR, Genome Canada, and NCI SPORE Grant P50CA70907.

Chapter 7: The authors would also like to thank K. Niessen for performing the HBEC infections and A. Carraro, J. Korbelik and D. Ceron for scoring the IHC cases. This work was supported by funds from CIHR, Genome Canada, Genome British Columbia, Canadian Cancer Society, Lung Cancer SPORE, DOD VITAL, the Gillson Longenbaugh and Anderson Charitable Foundations as well as scholarships from the Natural Sciences and Engineering Research
Council, CIHR and the Michael Smith Foundation for Health Research to W.W.L., R.C., B.P.C. and T.P.H.B.
Dedication

To my family.
Co-Authorship Statement

Chapters 2 to 7 were co-authored as manuscripts for publication. The following author lists apply for each chapter:


Contribution: I developed the data analysis methods, interpreted the results, co-wrote the manuscript and made all figures and tables.


Contribution: I conceived the study, performed experiments, developed the analysis methodologies, interpreted the results and co-wrote the paper.


Contribution: I conceived the concept of the study, performed experiments, developed and performed all data analyses, interpreted the results and wrote the paper.


Contribution: Brad Coe and I shared responsibility for conceiving the project, experimental work, as well as all data analysis and writing.


Contribution: I conceived the project, analyzed the data and wrote the manuscript.

Contribution: I conceived the project, planned and performed experiments, analyzed all data, interpreted the results and findings and wrote the manuscript.
Chapter 1: Introduction

Portions of this chapter have been published as:


comparative genomic hybridization. *Cancer Informatics* 2:48-58. *[co-first authorship]*


*These authors contributed equally.*
1.1 Introduction to Lung Cancer

Lung cancer is the leading cause of cancer mortality worldwide responsible for over one million deaths each year (Parkin et al., 2005). In Canada alone, 23,900 people were diagnosed with lung cancer in 2008 and 20,200 died from the disease (Canada, 2008). This represents 27% of all cancer related mortalities which is more than prostate, breast and colorectal cancers combined (Canada, 2008). Tobacco smoke represents the major etiological agent for lung cancer with smokers at a 10-20 fold increased risk of developing disease compared to never smokers (Sato et al., 2007). When combined with passive exposure, tobacco smoke is thought to cause approximately 85% of all lung cancer cases (Sun et al., 2007a). Besides smoking, exposure to other carcinogens such as arsenic and asbestos as well as genetic (e.g. familial history) and viral (e.g. human papillomavirus) factors have been suggested to influence lung cancer risk (Sun et al., 2007a).

Despite recent advances in diagnosis and treatment, the prognosis for lung cancer still remains poor with a five-year survival rate of only 16% for all stages combined (Jemal et al., 2008). The dismal outcome is attributed to two main factors. First, the vast majority of lung cancer patients are diagnosed at a late stage of disease when options for treatment are mainly palliative (Wistuba & Gazdar, 2006). In contrast, patients presenting with early stage disease are candidates for potentially curative surgical intervention and show significantly better outcomes (Gomez & Silvestri, 2008) (see below). Secondly, there is a lack of effective drugs available to treat lung tumors as current standard therapies such as platinum-based doublet regimes only provide modest survival benefits (Sato et al., 2007). Clearly, a better understanding of the molecular origins and basic biology of lung cancer is urgently needed in order to develop new early detection techniques and novel therapies to address these issues and improve lung cancer prognosis.

1.2. Lung Anatomy and Histology

The lung is one of the most anatomically and histologically complex organs in the human body. Anatomically, the lungs can be divided into two main regions: the conducting airways and respiratory compartment which play a primary role in gas transport and exchange, respectively. In proximal-distal order, the conducting portion consists of the bronchi, bronchioles and terminal bronchioles whereas the respiratory portion contains the respiratory bronchioles, alveolar ducts and alveoli (Junqueira et al., 1995). Histologically, the lung mirrors this compartmental organization with the epithelium separated into distinct regions of unique cellular composition.
based on airway level. The large bronchial airways, which have muscular walls reinforced with cartilage, are lined with pseudostratified epithelium consisting of basal, ciliated, goblet and serous cells along with less frequent pulmonary neuroendocrine cells (PNECs) (Junqueira et al, 1995; Kannan & Wu, 2006; Otto, 2002; Snyder et al, 2009). Proceeding distally along the conducting airways, simplified columnar epithelium appears which is similar in composition but contains Clara cells and is devoid of goblet cells. This is characteristic of the bronchiolar airways which lack cartilage and have incomplete muscular walls. The respiratory bronchioles are the transition point between the conducting and respiratory portions of the lung. Here, the ciliated cuboidal epithelium of the airways merge with the simple squamous epithelium of the alveoli containing type 1 and 2 pneumocytes, which function in gas transfer and surfactant production, respectively (Junqueira et al, 1995; Kannan & Wu, 2006; Otto, 2002; Snyder et al, 2009).

1.3 Lung Cancer Subtypes and Clinical Features

Matching the complexity of the organ itself, lung cancer is not a homogeneous entity but a collection of phenotypically diverse and regionally distinct neoplasias. Based on clinical and histological criteria, lung cancer is separated into two major subtypes: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) (Travis, 2002). The pathological features, clinical characteristics and anatomical distribution of these subtypes are detailed in the following sections.

1.3.1 Small cell lung cancer.

The distinction between SCLC and NSCLC is important as they exhibit drastically different clinical and biological characteristics. SCLC comprises ~20% of all lung cancer cases and is the most aggressive lung tumour due to its rapid doubling time and poor clinical outcome (Gustafsson et al, 2008). These tumors rapidly and extensively invade the bronchial wall and surrounding parenchyma making lymphatic and blood-borne dissemination a common feature at the time of diagnosis (Gustafsson et al, 2008). As such, SCLC patients are often poor candidates for surgical resection. Standard cytotoxic chemotherapy agents have shown antitumor activity with response rates of 70-90% for both limited (hemithorax restricted with regional lymph node metastases) and extensive (presence of metastases outside the thorax) stage disease (Gustafsson et al, 2008). However, despite this strong initial response to therapy, patients rapidly develop resistance and most relapse. Thus, SCLC has the worst prognosis of all lung cancers with an overall five-year survival rate of only 4.8% (2003) (Gustafsson et al,
The minimal increase in survival over the past 30 years (3.9% in 1973) is clearly indicative of the lack of effective treatment strategies (Gustafsson et al, 2008; Linnoila, 2006). In addition, the increase in long term survivors (20%) among those with limited stage disease suggests that early detection techniques may also be imperative to improve these rates (Gustafsson et al, 2008).

SCLCs typically localize to the central airway compartment, predominately the proximal bronchi (Sun et al, 2007a). PNECs have been proposed as the likely originating cells of this disease as SCLC tumors express a range of neuroendocrine markers normally expressed on these cells (Linnoila, 2006). However, no phenotypically identifiable precursor lesion for SCLC has been identified to date. This has lead to the suggestion that these tumors may arise directly from normal or mildly abnormal (hyperplastic) bronchial epithelium without passing through a stepwise histological sequence as seen in NSCLC development (see below)(Gustafsson et al, 2008; Wistuba & Gazdar, 2006). Histologically, SCLC is readily identifiable by light microscopy as it displays densely packed tumor cells with scant cytoplasm, absent or very small nucleoli and high mitotic rate (Stevens et al, 2002; Travis, 2002).

1.3.2 Non-small cell lung cancer.

NSCLC is the most common form of lung cancer representing 80-85% of all cases and can be further divided into two main tumor types: squamous cell carcinoma (SqCC) and adenocarcinoma (AC) (Travis, 2002). Although additional subdivisions exist (for example large cell carcinoma), SqCC and AC account for the vast majority of all diagnosed NSCLC cases (>80%) and are the primary focus of this thesis. In general, NSCLC is less aggressive than SCLC and displays a slightly more favourable clinical outcome. However, this is highly dependent on the stage of disease at the time of diagnosis. Unlike SCLC, both AC and SqCC are classified by the traditional tumor size, nodal, or distant metastasis (TNM) staging system and which treatment option is chosen is typically defined by stage (Flieder, 2007). Current treatment strategies for NSCLC include surgical resection, platinum-based doublet chemotherapy and radiation therapy alone or in combination (Sato et al, 2007; Sun et al, 2007b). Patients with early stage disease undergo complete tumor resection with or without adjuvant chemotherapy whereas those with locally advanced, unresectable disease often receive concurrent chemoradiotherapy (Sato et al, 2007; Sun et al, 2007b). As with SCLC, chemotherapy is the only treatment option for patients with metastatic NSCLC and is mainly palliative, offering a modest survival benefit. Reflecting the effectiveness of these treatments, long term survival rates of 60 to 75% are observed for early stage disease while higher stage
disease demonstrates increasingly poor survival (Flieder, 2007). There are virtually no long
term survivors for high grade metastatic disease; thus, new techniques facilitating the sensitive
detection of lung cancer are urgently needed to shift the stage of diagnosis to early disease and
improve patient outcomes (Flieder, 2007; Gomez & Silvestri, 2008). Overall, SqCC displays a
slightly more aggressive phenotype than AC and has a worse overall 5-year survival rate (14%
compared to 17%) (Fry et al, 1999). Current treatment modalities have reached their
therapeutic plateau and new strategies are needed to improve these rates.

SqCC comprises approximately 30% of all lung cancer cases and, like SCLC, arises mainly in
the central airways including the main bronchi and their larger branches (Travis, 2002). These
tumors usually demonstrate clear features of squamous differentiation including intracellular
bridges, cellular keratinisation and the formation of keratin pearls (Stevens et al, 2002; Travis,
2002). Since there are no squamous cells in the main airways, the progenitor cells for SqCC
are not definitively known (Snyder et al, 2009). However, the morphology, marker expression
(i.e. keratin 14) and distribution of SqCC has many characteristics in common with basal cells
(Giangreco et al, 2007). In addition, the development of squamous metaplastic epithelium in the
proximal airways is often preceded by basal cell hyperplasia (Wistuba & Gazdar, 2006). These
observations have led to a well established, multistep model of SqCC pathogenesis involving
the transformation of normal lung epithelium through a continuous spectrum of precursor lesions
to invasive cancer. The process initiates with basal cell hyperplasia, progresses though
squamous metaplasia, various degrees of dysplasia (mild, moderate and severe) and
carcinoma in-situ before finally developing into a fully malignant tumour (Wistuba & Gazdar,
2006). Each step is characterised by an increase in morphological and cytological changes and
can be separated with good reproducibility.

AC is the most common NSCLC subtype (~50%) and develops in the small bronchi and
bronchioles of the lung periphery (Travis, 2002). Classically, the main histological feature of
these tumors is the formation of cells into a glandular acinar pattern with the resulting acini often
being filled with mucus (Stevens et al, 2002; Travis, 2002). However, there is great deal of
histological heterogeneity within AC and tumors of this subtype often contain one or more
additional, distinct features such as papillary, bronchioalveolar, solid, clear cell, mucinous, and
signet ring patterns (Stevens et al, 2002; Travis, 2002). AC cells often exhibit mixed airway and
alveolar cell characteristics with the expression of clara cell secretory protein (CCSP) and
surfactant protein C (SPC) (Giangreco et al, 2007). This has lead to the suggestion that they
arise from clara cells (which secrete CCSP) or type 2 pneumocytes (which secrete SPC) near
the bronchoalveolar junction (Giangreco et al., 2007; Otto, 2002). Similar to SqCC, a multistep progression model has been proposed for AC. This process begins with atypical adenomatous hyperplasia (AAH) and progresses through low grade bronchioloalveolar carcinoma (BAC) lesions to invasive cancer (Wistuba & Gazdar, 2006). This model is supported by morphometric, cytofluorometric and molecular studies. However, since the peripheral lung is difficult to monitor and study, it is possible that additional routes to AC also exist.

1.4 Molecular Biology of Lung Cancer

Lung cancer is thought to result from the sequential accumulation of somatic DNA alterations that begin in normal epithelium and increase in severity during cancer progression, consistent with the multistep model of carcinogenesis (Wistuba, 2007). These alterations cause the activation of oncogenes and inactivation of tumor suppressor genes, leading to the deregulation of fundamental cellular processes which confer malignant growth (Sekido et al., 2003). As such, there has been extensive effort to identify and characterize genetic alterations in order to delineate genes and pathways involved in lung tumorigenesis.

1.4.1 Genetic alterations

There are multiple types of DNA alteration that can cause gene disruption including sequence mutations, chromosomal translocations and copy number changes (Albertson et al., 2003). Previous studies have demonstrated that each of these mechanisms may contribute to lung cancer development. For example, sequence analysis of candidate genes has identified constitutive activating mutations in oncogenes such as EGFR and KRAS and inactivating mutations in tumor suppressor genes including Rb and p53 (Fong et al., 2003). Likewise, chromosome translocations, although rare in solid tumors, have also been described and those leading to ALK gene fusions have recently been implicated in a subset of lung tumors (Herbst et al., 2008).

Genomic DNA copy number alterations are a prominent mechanism of gene disruption which contributes to tumor evolution through gene dosage-induced changes in expression levels (Albertson et al., 2003). Amplifications/gain of chromosome segments may lead to the overexpression of oncogenes whereas deletion/loss may lead to silencing the expression of tumor suppressor genes (Albertson et al., 2003). These aberrations are distinctive features of lung cancers and can be detected using both cytogenetic and molecular methods (Balsara & Testa, 2002). Prior to the initiation of this thesis, several published studies aimed to identify
regions which are consistently amplified/deleted in lung cancer genomes and thus, likely to harbour genes critical for tumorigenesis. This has lead to the discovery of multiple chromosome regions implicated in lung cancer development and progression including gains on chromosome arms 1q, 3q, 5p, 7q, 8q, 11q, 16p and losses on 3p, 4q, 5q, 6q, 8p, 9p, 13q, 17p and 19p (Balsara & Testa, 2002; Sato et al, 2007). However, due to limitations of conventional genome profiling technologies (see section 1.6), identifying the target genes of these alterations has proven difficult. Although many regions can be associated with known oncogenes (i.e. MYC for the gain on 8q) and tumor suppressor genes (i.e. CDKN2A for the loss on 9p), candidates for other loci remain ambiguous (Balsara & Testa, 2002). Fine mapping of these regions using PCR or FISH based techniques has aided in the identification of novel candidate genes in lung cancer. For example, these approaches lead to the refinement of the 3p loss to three discrete regions encompassing 600 kb at 3p21.3, 3p14.2 and 3p12 and the discovery of FHIT as the tumor suppressor gene within 3p14.2 in both SCLC and NSCLC (Sekido et al, 2003). However, these methods are very labour intensive making progress slow and many target genes of DNA alteration in lung cancer remain to be uncovered (Balsara & Testa, 2002). It is reasonable to expect that additional genes important in lung cancer pathogenesis will be identified by in-depth assessment of other recurrently abnormal regions.

1.4.2 Epigenetic alterations

Epigenetic changes refer to features such as chromatin and DNA modifications that are stable over rounds of cell division but do not involve changes in underlying DNA sequences (Bird, 2007). DNA methylation is a well-known epigenetic mechanism of gene regulation that involves the reversible chemical modification of cytosine nucleotides (Esteller, 2008). Aberrant methylation can occur at specific promoter CpG islands as well as in broad “neighbourhoods” where more than one gene is affected (Frigola et al, 2006; Shames et al, 2007). Generally, methylation at these sites leads to repression of gene expression whereas demethylation has the opposite effect (Shames et al, 2007). Alterations in DNA methylation levels are associated with many human diseases and are a hallmark of cancer(Esteller, 2007). Transcriptional silencing of genes involved in growth regulation (e.g. p16/CDKN2A) via promoter hypermethylation is a frequent mechanism of tumor suppressor inactivation in lung cancer (Sato et al, 2007; Shames et al, 2007). Likewise, gene activation reportedly occurs through DNA hypomethylation of specific promoters, leading to overexpression of affected genes (Shames et al, 2007). Promoter hypermethylation for specific genes often occurs early in tumor development and the detection of methylated DNA sequences in biological fluids (sputum,
blood) highlights a promising tool for the early detection of lung cancer if robust targets can be identified (Sato et al., 2007; Shames et al., 2007).

To date, the majority of genes shown to be aberrantly methylated in lung cancer have been discovered using candidate gene approaches such as methylation sensitive polymerase chain reaction (MS-PCR) (Esteller, 2007). As with locus specific methods for the discovery of genetic alterations, these techniques are labour intensive, hindering the discovery of aberrantly methylated genes in lung cancer. In a recently study, Shames et al examined lung cancer cell lines to identify changes in promoter methylation by inducing re-expression of genes through exposure to a global de-methylating agent, 5-aza-cytidine, and found 31 novel genes to be hypermethylated in lung cancer, with 8 genes showing changes in methylation across multiple epithelial tumor types, demonstrating the value in using a genome-wide approach to survey DNA methylation (Shames et al., 2006). Additional high-throughput approaches are being developed to assay DNA methylation in a genome-wide manner suitable for the analysis of clinical tumor specimens (Esteller, 2007). These methods promise to lead to a growing list of genes with abnormal methylation patterns in lung cancer. However, since these techniques were in their infancy at the time this work was initiated, the focus of this thesis is on genetic alterations in lung cancer development (see sections 1.6-1.8).

1.4.3 Signalling pathways

It is currently believed that several genetic and epigenetic alterations are required before lung cancers become clinically evident (Sato et al., 2007). Collectively, these alterations affect genes that positively and negatively regulate pathways involved in cell proliferation, apoptosis, immortalization, genome stability, angiogenesis, invasion and metastasis, the so called “Hallmarks of Cancer” (Hanahan & Weinberg, 2000). Cumulative evidence from molecular genetics, gene expression and functional studies have identified key pathways involved in these cellular processes that, when deregulated, may lead to lung cancer formation. These include EGFR family, PI3K/AKT/PTEN, RAS/RAF/MEK/ERK, p53, CDKN2A-CyclinD1-CDK4-RB, Telomerase and VEGF signalling pathways as well as others (Fong et al., 2003; Herbst et al., 2008; Sato et al., 2007; Sekido et al., 2003). What are less evident are the mechanisms causing the disruption of these pathways in malignant cells. Since each pathway contains multiple components, there are multiple routes that can underlie their disruption. For example, the CDKN2A-CyclinD1-CDK4-RB pathway is essential for controlling the G1-S transition of the cell cycle and is functionally altered in lung and other cancers (Sato et al., 2007). This can occur through activating alterations (i.e. amplification) affecting Cyclin D1 or CDK4 or inactivating
alterations (i.e. deletion) of Rb or CDKN2A (Sato et al, 2007). In addition, DNA disruption of genes that directly regulate any of these components may lead to pathway deregulation. However, since studies traditionally focus on the analysis of only one or a few integral pathway components, the mechanisms involved in the disruption of these pathways in lung cancer remain poorly understood.

1.4.4 Targeted therapies

With our increasing understanding of lung cancer biology has come the recent development of rationally designed therapeutic strategies to combat this disease. These therapies target components of the key pathways described above to inhibit the hallmark processes required for tumorigenesis (Auberger et al, 2006). This is based on the observation that cancer cells with activation of these pathways often become addicted to their abnormal function for survival and maintenance of the malignant phenotype, a phenomena known as “oncogene addiction” (Thomas et al, 2006). One of the best characterized examples of this are drugs which target the transmembrane tyrosine kinase (TK) receptor EGFR. When bound to ligand, receptor dimerization occurs which in turn activates the TK domain of EGFR leading to its autophosphorylation and further activation of a cascade of signalling events resulting in cell proliferation, inhibition of apoptosis, invasion and angiogenesis (Sun et al, 2007b). Since EGFR and its ligands are overexpressed in a large portion of NSCLCs, agents targeting the receptor including the TK inhibitors (TKIs) gefitinib and erlotinib and monoclonal antibodies cetuximab and panitumumab were developed for use in NSCLC treatment (Sun et al, 2007b). Subsequent clinical trials showed that a subset of lung tumors that were sensitive to treatment with TKIs harboured a mutation in the TK domain of EGFR leading to constitutively active signalling. These tumors are addicted to this aberration and inhibition leads to a drastic initial response to therapy. As such, gefitinib and erlotinib are now approved for treatment of advanced NSCLC (Auberger et al, 2006).

The EGFR example highlights the success of designing drugs to specifically target the molecular mechanisms driving lung cancer development. Clinical trials are now underway for a wide array of additional drugs including those that target VEGFR, Ras, MEK, and PI3K (Auberger et al, 2006). However, some important considerations need to be addressed. For example, the vast difference in patient response to these therapies underlines the need to better understand the genetic mechanisms responsible for pathway deregulation. This will be essential in order to predict which patients will best respond to therapy, as exemplified by the EGFR TK mutation cases described above. Regardless, these examples prove the concept
that with a better understanding of the mechanisms involved in lung cancer development, we can design more effective methods to aid in its treatment.

1.5 Cell Lineage Hypothesis of Tumorigenesis

As stated before, lung cancer is a heterogeneous disease comprised of multiple histological subtypes. The histological heterogeneity of lung cancer likely reflects differences in cell derivation, genetic alterations and pathogenetic pathways. The following sections will describe how these associations may influence discrepancies in tumor biology and impact the development of targeted therapeutic strategies.

1.5.1 Cell lineage model

Lung cancer subtype classifications, as with cancers of other organs, are based largely on the assumed associations between cell lineages and tumorigenesis (Garraway & Sellers, 2006a; Garraway & Sellers, 2006b). For example, SqCC and AC are thought to arise through the malignant transformation of histologically distinct cell types, basal cells and clara cells or type 2 pneumocytes, respectively (Giangreco et al., 2007). SCLC on the other hand arises from pulmonary neuroendocrine cells (Giangreco et al., 2007). These associations have been made through the phenotypic resemblance of characteristics such as growth and morphology between tumor cells and their normal counterparts. Interestingly, the lung cancer subtypes have a distribution pattern resembling that of the specific cell types from which they are thought to originate (SqCC/basal cells and SCLC/PNECs = central airways, AC/clara/type 2 cells = peripheral airways), adding more credence to this cell lineage based model of carcinogenesis (Giangreco et al., 2007).

1.5.2 Lineage/subtype specificity

The emergence of tumor cells from normal precursors is thought to involve a complex interplay between genetics and cell lineage (Garraway & Sellers, 2006b). Due to the different cell types involved as well as the attributes of an individual cell’s local environment or niche, it is logical to assume different mechanisms are involved in tumorigenesis for each lung cancer subtype. Thus, specific genes and their respective pathways may lead to carcinogenesis only when disrupted in permissive conditions. For example, a gene may have oncogenic properties when aberrated in basal cells in the central compartment because it supports growth under these conditions; however, the same gene may have no effect on clara cells in the lung periphery. Recent studies using transgeneic mouse models have supported this theory. Of particular
interest, murine models have been developed that introduce Kras mutations, which are common in human lung cancer, throughout the entire lung (Giangreco et al., 2007). Remarkably, although all airway epithelial cells contained this mutation, only adenomatous hyperplastic lesions - precursors to AC - localized to the bronchoalveolar region developed in these mice (Giangreco et al., 2007). Likewise, a model deleting both Rb and TP53 throughout the airways only resulted in PNEC hyperplasia and the development of metastatic tumors resembling SCLC (Giangreco et al., 2007).

Cell lineage may also have a dramatic effect on the manifestation of genetic alterations during the development of each lung cancer subtype as only those promoting a malignant phenotype in the specific cellular context will be selected and maintained (Garraway & Sellers, 2006b). This is supported by evidence from conventional cytogenetic studies which revealed distinct copy number profiles for different lung cancer subtypes (Pei et al., 2001; Petersen et al., 1997; Sy et al., 2004). However, due to the low resolution of the techniques and the small number of samples profiled, the specific target genes of these alterations remain unknown. Nonetheless, this data combined with that from the mouse model studies suggests that very particular genetic alterations in specific cell types are necessary for the development of the individual lung cancer subtypes. Furthermore, this implies that the resulting phenotypic heterogeneity of the subtypes is a reflection of these differences.

*Please note that throughout the course of this thesis, lung cancer subtypes may also be referred to interchangeably as lung cancer cell types and lung cancer cell lineages.

1.5.3 Therapeutic implications

Fundamental discrepancies in tumor biology may be a primary factor determining the poor outcomes of lung cancer patients as biological differences that segregate with lineage may also lead to differences in response to therapies (Garraway & Sellers, 2006b). Thus, it is becoming clear that lineage may be an important consideration when selecting and developing therapeutic approaches for lung cancer.

An example of this is already in common practice as SCLC and NSCLC are treated separately due to the observation that cancers of the former lineage tend to be much more responsive to initial treatment with conventional cytotoxic agents. However, with NSCLC, no clinical distinction is made between the different cancer subtypes and stage is the primary factor used in determining which treatment regime is applied. By treating AC and SqCC as a single
disease, we are ignoring the important biological factors underlying their differential
development which may lead to suboptimal response rates to therapy.

The recent advancement of targeted therapeutic strategies has highlighted the clinical
importance of the biological differences associated with lung cancer subtypes. For example, a
clinical trial with the VEGF inhibitor bevacizumab has shown a survival benefit for patients with
AC and was recently approved for use in advanced NSCLC (Auberger et al, 2006). However,
pulmonary haemorrhaging associated with this treatment was observed in a subset of patients,
all of which had SqCC tumors (Sun et al, 2007b). Thus, it appears that dependence on VEGF
signalling for tumor maintenance is AC specific and patients with SqCC were excluded from
more recent clinical trials with this drug (Sun et al, 2007b). An additional example pertains to
EGFR TK inhibitors (TKI). Of late, it has been demonstrated that the mutation in the TK domain
of EGFR occurs exclusively in AC tumors (Gazdar & Minna, 2008). Since this mutation
correlates with response to TKI treatment, lineage should be considered when assessing which
patients should receive these drugs.

Overall, this knowledge suggests that treatment strategies specifically targeted to the different
phenotypes of lung cancer may drastically increase patient response rates and outcome.
Therefore, distinguishing the key molecular mechanisms responsible for the development of
each lung cancer lineage will be essential in order to define appropriate avenues for diagnosis
and therapeutic intervention.

1.6 Global Profiling Technologies for the Analysis of Cancer Genomes

A greater characterization of the alterations involved in lung cancer development is needed in
order to understand the genes and pathways involved in tumorigenesis. Array technologies
have allowed the transition from single gene/locus assays to global profiling and have opened
avenues for a systematic approach to understand the complex biology of lung cancer
(Lockwood et al, 2006). These technologies include gene expression microarrays and CGH
arrays for the analysis of copy number changes.

1.6.1 Gene expression microarray analysis

Prior to the advent of high throughput expression profiling tools, gene expression was measured
a single gene at a time, typically by northern blot or PCR based techniques. Early global
profiling methods, such as differential display and subtractive hybridization, required

One of the first high throughput tools developed for the study of cancer was expression microarray analysis. Expression microarrays are based on the concept of competitively hybridizing reference and sample RNA samples which have been fluorescently labeled with different dyes to a glass slide with immobilized DNA targets representing transcripts of interest. By imaging the slides in each fluorescent channel and computing the ratios with which each sample bound to the cDNA targets, a gene expression ratio could be determined. This allowed rapid quantification of gene expression without the difficulty of multiplexing PCR reactions at a significantly reduced cost per target compared to conventional methods. The technology was rapidly improved by moving to large scale analysis of thousands of human cDNAs and has since evolved towards oligonucleotide (oligo) based technologies which allow consistent, affordable array production with highly specific probes that can be used in one color assays (Bammler et al, 2005; Larkin et al, 2005; Shi et al, 2006).

Gene expression patterns in lung cancer revealed through microarray analysis have yielded some insight into the disease. Initial profiling studies were able to cluster histologic and prognostic groupings (Beer et al, 2002; Bhattacharjee et al, 2001; Bild et al, 2006). Recent studies have attempted to further sub-classify histologic subtypes based on multi-gene models, as well as predict patient survival and chemotherapeutic response (Larsen et al, 2007a; Larsen et al, 2007b; Potti et al, 2006a; Potti et al, 2006b; Raponi et al, 2006). However, as not all gene expression changes are causal to disease development, it is challenging to distinguish critical events from reactive changes through correlative studies using gene expression profiling alone.

**1.6.2 Array comparative genomic hybridization**

Prior to the development of global approaches, the majority of methods used to detect copy number alterations in cancer focused on single locus assays (similar to initial gene expression assays). The advent of conventional CGH allowed researchers to understand the patterns of gene dosage across the entire genome, albeit at a relatively low resolution of ~10 Mbp (Kallioniemi et al, 1992; Lockwood et al, 2006). The capabilities of chromosome-based CGH were improved by the development of array CGH, whereby DNA targets are spotted onto a glass surface to serve as hybridization targets as an alternative to using metaphase chromosome spreads (Pinkel et al, 1998; Solinas-Toldo et al, 1997). In this method, DNA from both reference and test genomes are differentially labelled with fluorescent dyes and
competitively hybridized to DNA targets arrayed on a slide (Figure 1.1). The hybridized slide is then scanned and the resulting signal intensity ratio at each DNA target reflects the copy number status of the DNA segment. By referring the segment to its corresponding position on the human genome map, the genes affected by copy number alteration can be identified (Fiegler et al., 2003; Ishkanian et al., 2004; Snijders et al., 2001).

Numerous advances in array CGH technology have been made since its development in the mid 1990s with increased genome coverage and target density, improving resolution and sensitivity of detection. The majority of array CGH platforms use either oligo or large insert clone (LIC) DNA targets (Davies et al., 2005). Oligos are short DNA fragments of approximately 21-60 nucleotides in length whereas LICs are typically bacterial artificial chromosome (BAC) clones which are ~100 kb in size. Historically, arrays were designed to cover specific chromosomes (Buckley et al., 2002; Buckley et al., 2005), chromosome arms (Coe et al., 2005; Garnis et al., 2003; Henderson et al., 2005) or selected regions of the genome implicated in disease (Albertson et al., 2000; Schwaenen et al., 2004). In contrast, genome wide arrays that sample copy number status of loci at megabase intervals have facilitated rapid survey for regions of loss and gain (Fiegler et al., 2003; Greshock et al., 2004; Snijders et al., 2001).

Alternatively, cDNA microarrays, initially designed for gene expression profiling, have been used to assess copy number status of coding regions (Pollack et al., 1999; Squire et al., 2003). The development of high density arrays consisting of tens of thousands of DNA targets spanning the entire human genome has enabled precision mapping of the boundaries of genetic alterations throughout the genome in a single experiment (Bignell et al., 2004; Ishkanian et al., 2004; Selzer et al., 2005; Zhao et al., 2004).

Each technology allows the high resolution profiling of sample genomes with distinct advantages and disadvantages for each. For example, BAC arrays require far less sample input than oligo arrays allowing the analysis of low yield microdissected specimens while oligo have the potential to offer greater resolving power (Lockwood et al., 2006). Currently, however, the resolution of both assays is sufficient to identify the specific gene targets of copy number alterations, and thus, both technologies occupy a specific place in today’s genomics field (Coe et al., 2007).

1.6.3 Integrative analysis

As stated above, not all of the gene expression changes observed in a tumor are causal to cancer development and global gene expression analysis alone cannot distinguish between
causal and reactive changes. Corresponding alteration at the DNA level is regarded as evidence of causality (Broet et al, 2009). For example, changes in DNA copy number are key to the progression and development of many cancers and lead to the deregulation of genes responsible for carcinogenesis (Hyman et al, 2002; Pollack et al, 2002). Hence, examining genetic events in conjunction with the changes in gene expression pattern should improve the identification of causal changes that lead to disease phenotype. Furthermore, the genes that demonstrate deregulation in expression as a result of these primary genetic events are ideal targets for therapeutic intervention due to the direct nature of their activation/inactivation, as compared to genes that may be activated by complex trans-regulation networks. Thus, genes deregulated as a result of DNA alterations can be viewed as the primary oncogenic targets in a tumor that lead to downstream pathway abrogation. Through the integration of genome-wide expression and copy number data, the identification of such changes is now possible on a global scale. Therefore, the application of this integrative analysis to lung cancer holds great potential to uncover pathologically related genes.

1.7 Thesis Theme and Rationale for Study

The theme of this thesis is understanding the genetic basis of lung cancer subtypes. Since they display distinct phenotypes and clinical characteristics, knowledge of the genetic mechanisms underlying these differences will lead to the development of novel subtype specific strategies for diagnosis and treatment.

1.8 Objectives and Hypotheses

The main objective of this thesis is to comprehensively identify the genetic alterations, genes and pathways responsible for the differential development of lung cancer subtypes using an integrative genomics approach.

This is based on the following main hypotheses:

**Hypothesis 1** – Genes key to lung tumorigenesis will be identified in recurrently altered genomic regions.

**Hypothesis 2** - Lung cancer subtypes require distinct genetic alterations for neoplastic development.
1.9 Specific Aims and Thesis Outline

This thesis consists of several manuscripts assembled in non-chronological order to best address the hypotheses and aims.

**Aim 1: Develop methods for high resolution integrative analysis of lung cancer genomes in order to identify key genetic alterations.**

Chapters 2 to 4 describe the development and application of methods for the high resolution integrative analysis of lung cancer genomes necessary to address hypotheses 1 and 2.

At the initiation of this thesis, microarray platforms for the high resolution, global analysis of cancer genomes were in their infancy. Therefore, there was a great need to develop and optimize methodologies for the generation and analysis of array CGH and expression data. In addition, these methods needed to be assessed in order to determine their usefulness in identifying novel regions of genomic alteration and affected genes in cancer. Since cancer cell lines are easier to obtain than clinical specimens and yield large amounts of high quality DNA and RNA, they offer a valuable resource to generate datasets for this purpose. In addition, the majority of these lines have been extensively characterized by conventional cytogenetic techniques, offering a comparison to determine the effectiveness of high resolution approaches. Thus, chapters 2 to 4 were focused mainly on the analysis of cancer cell lines for these purposes.

Chapter 2 details the examination of NSCLC cell line genomes using a whole genome tiling path array for CGH analysis. This array was developed in our laboratory and facilitated whole genome profiling at an unprecedented resolution. Using this technology, I comprehensively analyzed copy number gains and losses and discovered novel regions of recurrent copy number alteration which contained putative oncogenes and tumor suppressor genes, providing evidence supporting Hypothesis 1. In addition, this study also provided preliminary evidence that NSCLC subtypes contain distinct genetic alterations. The computational strategies developed in this chapter provided the basis for subsequent array CGH analysis in later chapters.

Chapter 3 was a continuation on the findings from chapter 2. In this study, we aimed to uncover the target genes of novel regions of amplification on chromosome 7 in NSCLC through the integration of genomic and gene expression data. The results from this analysis provided more evidence that regions of recurrent alteration harbour genes important to lung cancer
development (Hypothesis 1) and demonstrated the utility of an integrative genomic approach to identify such genes.

In chapter 4, we expanded the integrative genomics approach to a genome-wide scale. Through the analysis of over 100 cancer cell lines from diverse tissue types, we uncovered novel hotspots of genomic amplification that were previously undetectable by conventional genomics approaches. Additionally, this analysis revealed that gene amplification is a common mechanism of oncogene activation in lung cancer and discovered novel candidates involved in NSCLC tumorigenesis. Most importantly, this study further emphasized the contribution of gene dosage changes to the development of lung cancer. Together, chapters 2 to 4 provide crucial evidence supporting hypothesis 1.

Aim 2: Identify DNA alterations specific to lung cancer subtypes.

Although touched upon in chapters 2 and 4, genomic differences between lung cancer subtypes were not comprehensively assessed. Therefore, I next aimed to expand the integrative approaches in order to directly compare the genomes of lung cancer subtypes. Chapter 5 details the initial development of these methods and their application to the comparison of NSCLC and SCLC cell lines. Since SCLC and NSCLC display drastically different phenotypes, they provided the best opportunity to discover differences at the genomic level. This analysis revealed that numerous regions of genomic alteration specific to each cell type exist, addressing hypothesis 2.

Chapter 6 expanded this analysis to clinical tumor specimens. Since they are the most common types of lung cancer and are typically regarded as a single disease entity in terms of therapy, I focused this analysis on the comparison of AC and SqCC tumors. As was the case in chapter 5, this investigation uncovered distinct genomic alterations in each subtype, providing further evidence in support of hypothesis 2.

Aim 3: Discover subtype specific genes and delineate the downstream consequences of gene disruption.

After identifying subtype specific alterations, I next aimed to identify the genes targeted by these alterations as well as the resulting pathways and cellular processes they disrupt in order to provide insight into the differential mechanisms of tumorigenesis. This was accomplished through integration of the genetic findings described in Aim 2 with gene expression data and also performed in chapters 5 and 6. This analysis revealed the key genes disrupted by subtype
specific alterations which may be causally involved in cancer development. Although the same pathway may be disrupted in different tumors, it is improbable that malfunction of the same gene is responsible for pathway disruption in all tumors. Therefore, multiple hits (alterations in multiple pathway components) would alert to the significance of a given pathway. To delineate the downstream consequences of gene disruption on disease phenotypes, subtype specific genes were then analyzed in terms of their involvement in cellular processes and pathways. This showed patterns of pathway disruption specific to each subtype, further addressing hypothesis 2.

Lastly, I aimed to further characterize specific gene targets in order to determine their involvement in subtype initiation and progression and assess their usefulness as potential diagnostic markers and therapeutic targets. For this purpose, I focused on a novel region of amplification uncovered in chapter 6 that was specific to SqCC. Through the analysis of SqCC precursor lesions, I was able to identify the specific target of this alteration and validate its contribution to cancer development in cell model experiments. This work is the focal point of chapter 7.

While these chapters represent separate works, additional information can be obtained though the interpretation of the results as a whole. This is discussed in the Conclusions of this thesis.
Figure 1.1

Figure 1.1. Principles of array comparative genomic hybridization. Tumor and normal reference DNA are differentially labeled with cyanine-5 and cyanine-3 respectively and competitively hybridized to a genomic microarray. The array consists of DNA targets selected to span chromosome regions or the entire genome. These targets are typically spotted in replica. The ratio of the two fluorescence signal intensities reflects the relative copy number at that target. The ratio for each spot is plotted against its corresponding position in the human genome to generate a copy number profile.
1.10 References


Chapter 2: High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH

A version of this chapter has been published as:


*Please see the published version of this chapter for all supplementary materials.*
Lung cancer is the most common cause of cancer mortality in the world. Non-small cell lung carcinoma (NSCLC) accounts for ~80% of all diagnosed lung cancers and consists of two main sub-groups, squamous cell carcinoma (SqCC) and adenocarcinoma (AC) (Travis et al, 2004). It is currently believed that several genetic alterations are required before lung cancers become clinically evident. Typical changes observed in NSCLC include activation of cMYC, RAS, EGFR, and BCL2 as well as inactivation of TP53, CDKN2A, and FHIT (Sekido et al, 2003).

Genetic and biochemical studies have been greatly reliant on a canonical set of cell lines derived from squamous cell and adenocarcinoma tumors (Phelps et al, 1996). These lines have been instrumental in the identification of differentially expressed genes, the screening of chemotherapeutic agents and the modeling of cellular responses to the introduction of engineered genes (Chung et al, 2001; Fan et al, 2002; Shiraga et al, 2002; Whiteside et al, 2004). Since cultured cells, and their parental tumors, often exhibit chromosomal aberrations, and that there have been numerous reports of gross gains and losses of chromosomal regions in these cell lines, it is essential to measure their segmental copy number alterations across their entire genome.

Numerous studies aimed at identifying gross copy number alterations of both NSCLC cell lines and tumors have been reported (Balsara & Testa, 2002). Based on comparative genomic hybridization (CGH) analysis, which detects gross copy number alterations, the most commonly cited amplifications are 3q, 5p, 7q, 8q, 11q, 16p, and 17q while the most common deletions include 3p, 4q, 5q, 6q, 8p, 9p, 13q, 17q, and 19p. In addition to CGH studies, microsatellite analysis, which detect allelic imbalance or loss of heterozygosity, has also been used to identify alterations at specific loci throughout the genome (Chmara et al, 2004; Marsit et al, 2004; Woenckhaus et al, 2003). In a genome wide LOH analysis Girard et al. surveyed 399 loci for allelic imbalance in a panel of NSCLC cell lines (Girard et al, 2000). They reported the most frequent LOH to occur on chromosome arms 1p, 3p, 4p, 4q, 5q, 8p, 9p, 9q, 10p, 10q, 13q, 15q, 17p, 18q, 19p, Xp, and Xq. Hybridization to single nucleotide polymorphism (SNP) arrays has made the detection of allelic imbalance a high throughput analysis (Janne et al, 2004).

Although NSCLC has been profiled with CGH and high resolution LOH assays, a comprehensive whole genome effort has yet to be reported. Development of a submegabase resolution tiling (SMRT) array, consisting of over 32,000 ordered bacterial artificial chromosome (BAC) clones spanning the human genome, has allowed for the complete copy number analysis.
of a tumor genome in a single experiment and is similar to conventional CGH (Ishkanian et al., 2004). However, the resolution is greatly increased by the tiling nature of the BAC clones. Profiling of these NSCLC cell lines will not only allow for the identification of recurrent changes in NSCLC that may be relevant clinically, but will also provide valuable insight for future experiments using these lines as lung cancer models. In this report we profiled 28 commonly used NSCLC cell lines for segmental DNA copy number alterations across the whole genome using the comprehensive SMRT array and compared the alterations measured across the genomes of the SqCC and AC subtypes.

2.2 Results and Discussion

2.2.1 Whole genome segmental copy number profiling

Experimental models for studying lung cancer often use one or more of the available cell lines derived from NSCLC tumors (for example, references (Amann et al., 2005; Engelman et al., 2005; Shigematsu et al., 2005)). Because of their wide use in experimental studies it is imperative that we have a complete understanding of their genomes. Using the whole genome SMRT array we analyzed 28 NSCLC cell lines (1 large cell carcinoma, 9 squamous cell carcinomas, and 18 adenocarcinomas) for segmental gains and losses of genetic material throughout the entire genome. Since this array contains 32,433 overlapping DNA segments, the technology is capable of the detection of small segmental copy number alterations and localizing their exact breakpoints. Signal intensity ratios for each BAC (measured across triplicate spots) were processed and displayed as log2 plots using SeeGH software (Chi et al., 2004a). Figure 2.1 shows the SeeGH karyogram of the HCC95 cell line. This figure demonstrates the identification of both single copy as well as high level gains and losses. For example, within this profile we observe loss of the entire 3p arm in contrast to the gain of the distal half of 3q. Most of the chromosomes in this profile showed multiple segmental alterations. Interestingly, a high level amplification is evident on 11q which contains the known oncogene Cyclin D1 (Figure 2.1 arrow). SeeGH karyograms for all 28 cell lines are presented as supplemental figures at http://www.bccrc.ca/cg/ArrayCGH_Group.html. Also, the raw data has been deposited at the Gene Expression Omnibus (accession number GSE2922). Stringently, regions were only scored as altered if two or more consecutive BAC clones were altered.

2.2.2 Frequent copy number alterations

The vast amount of data obtained from using such a comprehensive array and the overlapping nature of the clone set presents a bioinformatics challenge, especially when comparison among
samples is required. Data alignment and visualization is an important step in the discovery of novel regions recurrently altered. Two commonly used techniques were employed to identify regions of frequent gains and losses. First, we generated a frequency diagram as a summary for all the samples assayed (Figure 2.2). This is achieved by identifying altered regions for each profile by using aCGH-Smooth software (Jong et al, 2004; Jong et al, 2003) and then summarizing the alterations in a frequency plot (see Materials and Methods). The limitation of this strategy is that the computer algorithm may miss or “smooth over” small alterations. Secondly, to focus on a specific chromosome arm we aligned the raw data for each sample using the TreeView program, which is more commonly used for expression data (http://jtreeview.sourceforge.net). This direct approach generated “heat maps” which facilitate the identification of altered regions and their boundary clones. However, the program does not make use of the overlapping information present in the tiling clones.

Each cell line showed numerous copy number changes, both amplifications and deletions with the number of breakpoints identified for each line ranging from 27 to 172 (Supplemental 1). Nearly every chromosome in each sample contained copy number changes. From the frequency analysis we observed the general trends of 3p, 4q, 8p, 9p, 10p, 13q, and 18q loss where amplification was rarely observed (Figure 2.2). The opposite is shown for chromosome 1q, 2p, 5p, 7p, 8q, and 20q where amplification is detected with, in most cases, less than a 10% frequency of loss. This is in contrast to most other chromosomes where both amplification and deletions are observed in these cell lines. The fact that every chromosome exhibits alterations may reflect the instability inherent in cell lines and their potentially heterogeneous populations. However, a number of studies have characterized clinical cases of NSCLC with conventional CGH with comparable results (Goeze et al, 2002; Petersen et al, 1997; Petersen et al, 2000; Sy et al, 2004). Supplemental Table 1 highlights the characteristics of the cell lines profiled in this report.

2.2.3 Regions commonly altered in NSCLC

Regions of amplification and deletion are often associated with the presence of oncogenes and tumor suppressor genes. Within this set of cell lines we have identified a number of altered regions that are highly recurrent, indicating that they may contain genes important for the cancerous phenotype. Interestingly the amplifications appeared to be focal changes whereas the deletions were more widespread, often encompassing an entire chromosome arm. Supplemental Table 2 summarizes the most frequently altered regions.
2.2.4 Copy number decrease

Among the frequently deleted regions are loss of 3p and 9p. Loss of heterozygosity (LOH) at several loci along the 3p arm has been detected in more than 90% of NSCLC (Zabarovsky et al., 2002). There is evidence for at least four separate regions of alteration on chromosome 3p, which undergo homozygous deletions in either lung tumors or cell lines (Braga et al., 2002; Girard et al., 2000; Todd et al., 1997; Zabarovsky et al., 2002). The frequency analysis for chromosome 3p from this whole genome array CGH analysis shows that the entire 3p arm is frequently deleted. However, three distinct regions were observed to be lost at a higher frequency than the rest of the chromosome arm. These regions include: 3p24.3-p24.2 (containing THRB, RARB, TOP2B), 3p24.1 (containing RBMS3), and 3p14.2-3p14.1 (containing FHIT). Deletions on chromosome arm 9p have also been reported to be a frequent and early event in NSCLC (Girard et al., 2000; Kohno et al., 1999; Tomizawa et al., 2002). However, such multiple regions have not been previously reported. From our analysis we define three common regions of deletion on chromosome 9p (9p23, 9p22.1-9p21.1, 9p13-9p11.2), one of which, 9p22.1-9p21.1, contains the known tumor suppressor gene CDKN2A. The region at 9p23 contains no known genes while the 9p13 region contains six genes, including PAX5. The PAX5 gene encodes for the transcription factor, B cell-specific activating protein, that regulates CD19, which is believed to negatively control cell growth. Aberrant promoter methylation of PAX5 has previously been reported in lung cancer (Palmisano et al., 2003).

2.2.5 Array CGH vs. LOH

Microsatellite marker analysis has been useful in identifying regions harboring potential tumor suppressor genes. Of the 28 lines profiled for segmental gains and losses in this study, 13 were previously examined for LOH using microsatellite analysis (Girard et al., 2000). We compared the array CGH results with the LOH data for 42 markers spanning chromosomes 3 and 9. From the LOH data there were 448 informative loci from all the lines, of these 252 (57%) matched the CGH data (copy number loss matching LOH = 29%, copy number gain matching LOH = 7%, normal copy number matching retention = 21%). Of the unmatched loci, 25% of informative loci showed copy number loss but with retention at the microsatellite loci and 18% showed copy number increase with retention of heterozygosity. While the majority of loci showed concordance of microsatellite analysis and CGH data, the number of disagreements supports the use of both techniques to achieve a complete genomic characterization. Microsatellite analysis and array CGH assess different types of changes: array CGH measures differences in
copy number whereas microsatellite analysis assess changes in the alleles (Zabarovsky et al, 2002b).

2.2.6 Loci methylated in lung cancer

According to Knudson's two-hit hypothesis, a tumor suppressor gene loses function only when both alleles are inactivated (Knudson, 1971). Alterations in DNA methylation patterns are widely recognized as a contributing factor in human tumorigenesis and may function to silence genes, typically in concert with deletions or mutations of one of the alleles. To determine if frequently methylated regions correlated to regions of copy number loss, we selected eleven genes that have been previously shown to be methylated in NSCLC and assessed their loci for DNA copy number changes in the 28 cell lines. Of the eleven genes analyzed eight (APC 32%, CDKN2A 89%, DAPK1 82%, GATA4 46%, RARB 82%, RASSF1 32%, Reprimo 35%, ROBO1 71%) showed frequencies of deletion greater than 30%. Intriguingly, CCND2, DAB2, and, GATA5 showed low frequency of deletion but high frequencies of amplification at 46%, 67%, and, 75% respectively. Our findings indicate that the eleven loci known to be methylated in lung cancer are frequently altered in copy number.

2.2.7 Segmental amplifications

A striking feature observed in the whole genome profiles of these cell lines is the presence of high level amplifications. For this analysis we defined a high level amplification as log2 signal intensity ratio reaching +1. Supplemental Table1 summarizes all of the high level amplifications detected in this set of cell lines. The most frequent high level amplification occurred in 10 of the 28 cell lines at 8q24.21, which contains the known oncogene cMYC. The MYC proteins regulate normal cell growth through activation of genes involved in DNA synthesis, RNA metabolism, and cell cycle progression (Grandori et al, 2000). The usual mechanism of cMYC activation is through gene amplification resulting in overexpression. Figure 2.3 shows the alignment of 10 profiles at 8q24 with the high level cMYC amplification. In some of the cases the alterations is large spanning ~26 Mbp (HCC1195) and in the smallest case (H1395) spans only ~1 Mbp.

High level amplifications were also detected in six of the cell lines at 5p15.33 and 14q13.2-q13.3. The alteration on 5p contains hTERT, which has been implicated in numerous cancer types, including lung cancer (Sakura et al, 1999; Tomoda et al, 2002). hTERT is sufficient to induce telomerase activity resulting in immortalization of the cells (Bodnar et al, 1998; Ramirez et al, 2004). Telomerase is active in 70%-90% of malignant tissues whereas most somatic cells
have no detectable activity (Kim et al, 1994). Telomeres are essential for normal cell activity and telomerase is the enzyme required for adding the repeat sequences to the ends of chromosomes. In the absence of telomerase the chromosome ends begin to shorten with each cell division, and at a critical length, growth arrest or senescence is induced. Ectopic expression of hTERT, resulting in telomerase activity, has been shown to facilitate the immortalization of cells (Bodnar et al, 1998; Ramirez et al, 2004).

The high level amplification detected at 14q13 occurs at a frequency similar to that of the hTERT locus. While there have been reports of frequent LOH on this chromosome arm in lung cancer, frequent amplifications have not been reported (Abujiang et al, 1998). As with the high level amplification containing cMYC on 8q, the 14q alteration varies in size, from ~10 Mbp (HCC1195) to 2 Mbps (HCC1833) (Figure 2.4). The minimal region of interest, defined by H2009 (telomere) and HCC1833 (centromere), spans 2Mbp (RP11-138N6 to RP11-356L20) and includes 9 known genes (PSMA6, NFKBIA, INSM2, BRMS1, MBIP, TITF1, NKX2-8, PAX9, and SLC25A21).

2.2.8 Multiple regions on chromosome 7

Polysomy of chromosome 7, as well as regional amplifications on the same chromosome, have been identified as frequent events in NSCLC (Balsara & Testa, 2002; Testa et al, 1994; Zojer et al, 2000). Unexpectedly, we observed multiple regions of amplification across chromosome 7. Chromosome arm 7p contains two regions of amplification, 7p22.1-7p22.3 and 7p15.3-7p11.2, both occurring in over 80% of cases. The telomeric alteration spans ~6 Mbp and contains a number of genes, including RAC1, RBAK, MAD1L1, and MAFK, that may play role in tumorigenesis (Region A in Figure 2.5). The centromeric alteration is large spanning over 39 Mbp.

Copy number profiling of chromosome 7q revealed four recurrent regions of amplification (Figures 2.5 and 2.6). Two of the most frequent of these alterations occurred at the centromere and telomere (Region C and F, Figure 2.6). The centromeric amplification is small spanning only 0.5 Mbp and contains no known genes. The telomeric amplification is slightly larger spanning ~1.7 Mbp and contains a number of predicted genes but only two known genes, PTPRN2 and VIPR2. Region D (7q11.23) and E (7q22.1) are also frequent regions of alterations. Within the 7q11.23 region maps only predicted genes while the 7q22.1 region is very gene rich, containing 56 known genes within the 4.4 Mbp. Numerous Wnt family members have recently been implicated in lung cancer (Garnis et al, 2005). Frizzled 9 (FZD9), a homolog
of frizzled 3 which binds Wnt1, maps within the 7q22.1 amplicon. The presence of multiple regions of amplification on chromosome 7 indicates that at least four potential oncogenes may map to this chromosome. Additional functional studies are required to implicate specific genes in lung cancer pathogenesis.

2.2.9 Squamous cell carcinoma vs. adenocarcinoma

NSCLC is the most frequently occurring lung cancer type, with squamous cell carcinoma and adenocarcinoma as the two main sub-types. SqCC and AC are distinguished by differences in their histopathological and molecular characteristics: SqCC develops more rapidly and is located in the central airways, whereas AC is thought to originate from the epithelium of the lung periphery. The variation in progression and development of SqCC and AC may be due to underlying differences in genetic alterations between the two subtypes. To date, studies using conventional techniques such as LOH analysis and CGH have attempted to distinguish the genetic abnormalities that parallel the development of these different lung cancer subtypes. Previous findings suggest that, although a similar degree of genetic imbalance exists for AC and SqCC, each type is characterized by a unique pattern of imbalance (Pei et al., 2001). Molecular changes characteristic to each sub-group may provide insight into the different disease mechanisms.

We generated frequency diagrams for the two subtypes in order to identify copy number alteration differences between them. Figure 2.7 is an overlay of the frequency plots where the yellow area indicates alteration regions the two subtypes have in common while red indicates regions of alteration present in AC and green represents regions of alteration specific to SqCC. Generally the two subtypes follow the same patterns of alteration although the changes in the adenocarcinomas appeared to be more frequent. This similarity is particularly evident on chromosomes 8 and 11 (Figure 2.7). Regions of frequent amplification shared by the two subtypes include 5p, chromosome 7, 8q, 11q13, 19q, and 20q. Shared regions of deletion are 3p, 4q, 9p, 10p, 10q, chromosome 18, and chromosome 21. Additionally, some striking differences between the two subtypes were observed. Chromosome 6, 8p, 9q,15q, and chromosome 16 appear to have a much higher frequency of deletion in adenocarcinoma than the squamous subtype. There are also smaller regions of amplification present on chromosome 12 and 14 that are specific to the adenocarcinoma cases. Interestingly the SqCC and AC cell lines appear to have an opposite pattern of change at chromosome 2q and 13q. These chromosome arms are frequently deleted in adenocarcinoma but amplified in squamous cell carcinomas lines. While chromosome 17p shows deletion in both subtypes it is more frequent
in the squamous lines while amplification of chromosome 17p is frequent in the adenocarcinomas. Chromosome 3q, which is known to be amplified in both subtypes showed frequent alteration at 3q23-3q26 in the SqCC lines and at 3q22 in the AC lines. Conventional metaphase CGH studies frequently report amplification of 3q as one of the most common amplifications in lung squamous cell carcinoma (Chujo et al, 2002; Pei et al, 2001; Sy et al, 2004). Although the amplification of 3q is a frequent event, fine mapping of this alteration has proved difficult. As a result a number of genes located at 3q25-q27 have been implicated in lung cancer (Brass et al, 1997; Massion et al, 2002; Soder et al, 1997). The comparison of these lines indicates that there may be tumor suppressor genes and oncogenes specific to the different NSCLC subtypes. However, comparison of SqCC and AC tumors are required to verify the significance of these observations.

2.3 Conclusions

In summary, we have generated the first complete whole genome copy number profiles for a set of 28 NSCLC cell lines commonly used as experimental models for the study of lung cancer. The adaptation of microarray technology in the CGH application has greatly improved the detection of segmental DNA copy number changes: in specific chromosomal regions (Albertson et al, 2000; Bruder et al, 2001; Garnis et al, 2004a), whole chromosome arms (Coe et al, 2005; Garnis et al, 2005; Garnis et al, 2004b; van Duin et al, 2005), and at megabase intervals across the whole genome (Fiegler et al, 2003; Snijders et al, 2001; Veltman et al, 2003; Weiss et al, 2003). Furthermore, recent construction of the SMRT array has enabled the entire genome to be analyzed in a tiling path fashion (Ishkanian et al, 2004). The unprecedented tiling resolution of the SMRT array allows for defining breakpoints in an unbalanced chromosomal rearrangement to within a single BAC clone. At this resolution, small sub-megabase changes, previously undetected by conventional CGH and marker based techniques such as LOH, are now being discovered. In this study we have identified regions of frequent copy number gains and losses throughout these genomes. Unexpectedly, the detailed analysis of chromosome 7 identified 6 frequently altered regions across this chromosome arm, thus implicating the possibility of multiple oncogenes on chromosome 7. Additionally, profiles of the squamous cell carcinoma and adenocarcinoma NSCLC subtypes were compared to reveal specific similarities and also differences indicating that the two subtypes may develop through the activation or silencing of different genes. Future work will include the validation of these findings in clinical specimens.
2.4 Materials and Methods

2.4.1 Sample collection

The lung cancer cell lines were established at the National Cancer Institute (NCI-H series) and at the Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center (HCC series) except for SW-900 and SK-MES-1 (Supplemental 1) (Fogh et al., 1977; Phelps et al., 1996). They were grown in RPMI 1640 supplemented with 5% fetal bovine serum. The cell lines were verified by fingerprinting using the Powerplex 1.2 system (Promega) which contains 9 polymorphic markers. These cell lines have been deposited for distribution in the American Type Culture Collection (http://www.atcc.org).

2.4.2 Array construction

Array construction is outlined in a recent report by Ishkanian et al. (Ishkanian et al., 2004). Briefly, 32,433 clones spanning the entire genome at ~1.5 fold coverage were selected from the RPCI-11 BAC clone library. DNA from each BAC was isolated, amplified by linker mediated (LM) PCR, and re-dissolved in an MSP printing solution. PCR products were denatured by boiling prior to robotic spotting. The entire set of PCR products were spotted in triplicate onto two aldehyde-coated slides.

2.4.3 Probe labeling and hybridization

Random prime probe labeling as has been previously described was used (Garnis et al., 2003). 200 ng of test and reference DNA were separately labeled using Cyanine 5 and Cyanine 3 dCTPs fluorescence markers respectively. Probes were combined, denatured and annealed in a solution containing 200 μg human Cot-1 DNA (Invitrogen) and 100 μl DIG Easy hybridization solution (Roche) containing sheared herring sperm DNA (Sigma-Aldrich) and yeast tRNA (Calbiochem). Denaturing of the probe occurred at 85°C for 10 minutes and repetitive sequences were blocked at 45°C for 1 hr prior to hybridization. The probe mixture was applied to the slide surface and hybridized for 36 hours at 42 °C. The arrays were washed 5 times for 5 min in 0.1X SSC, 0.1% SDS at room temperature with agitation. Each array was then rinsed repeatedly in 0.1X SSC and dried by centrifugation.

2.4.4 Imaging and analysis

A Charge Couple Device (CCD) based imaging system (Arrayworx eAuto, API, Issaquah, WA) was used to determine signal intensities of the Cyanine5/Cyanine3 channels and images were
analyzed with Softworx array analysis software. In order to balance the differences between the fluorescent cyanine dye intensities normalization of the hybridized slides occurred by global median normalization. Standard deviations (SD) for the triplicate spots were calculated. All spots with SDs above 0.075 or signal to noise ratios below a ratio of 20 were removed from the analysis. Custom viewing software, SeeGH, was used to visualize all data as Log2 ratio plots where each dot represents one BAC (Chi et al, 2004b). Normalization of the data and variation between slides has been previously reported (de Leeuw et al, 2004). The filtered, normalized data was run through a smoothing algorithm in the aCGH-Smooth program (http://www.few.vu.nl/~vumarray/acghsmooth.html), as previously described (de Leeuw et al, 2004; Jong et al, 2004), using default settings except for lambda = 6.75 and breakpoints = 100. This program identifies breakpoints defining segmental copy number gains and losses using a local search algorithm that calculates the probability that the signal ratio for each clone corresponds with the same copy number status as a window of adjacent clones with a maximum likelihood estimation. The frequency of alteration for each BAC was then determined and plotted in SeeGH to visualize areas of recurrent deletion and amplification. To further visualize regions of interest, Java TreeView was used to create heat maps from the normalized log2 ratio data of selected chromosome arms (Saldanha, 2004).
Figure 2.1. Whole genome profile of squamous cell carcinoma line HCC95. Normalized log2 signal intensity ratios were plotted using SeeGH software, version 2.2.2. A log2 signal ratio of zero represents equivalent copy number between the sample and the reference DNA. Clones with standard deviations among the triplicate spots >0.075 or a signal to noise ratio <10 were disqualified from further analysis. Cytoband pattern for each chromosome is to the left for each plot. Vertical lines denote log2 signal ratios from -1 to 1 with copy number increases to the right and decreases to the left of zero. Each black dot represents a single BAC clone.
Figure 2.2

Figure 2.2. Whole genome evaluation of genetic alteration frequency for NSCLC cell lines. Filtered, normalized array ratio data was subjected to a smoothing algorithm as described in materials and methods which allowed the delineation of our clones into 3 groups: Normal, Gain and Loss. Frequencies of alteration were determined for each subtype by assigning a numerical score to each group and averaging across the number of measurements at each loci. The resulting numbers were plotted using SeeGH software version 2.2.2 to visualize areas of recurrent change. Vertical lines to the right and left of each chromosome represent the frequency of gain (0 to 100%) and of loss respectively. Red lines denote the proportion of samples showing amplification at a locus while green lines represent the proportion showing loss.
Figure 2.3. High level amplification at 8q24. Alignment of ten profiles at the MYC locus showing high level amplifications. Normalized log2 signal intensity ratios were plotted. High level amplifications were defined as log2 signal intensities equal to or greater than 1. Vertical lines denote log2 signal ratios from -1 to 1 with copy number increases to the right (red lines) and decreases to the left (green lines) of zero (purple line). Each black dot represents a single BAC clone. The recurrently amplified region is shaded orange.
Figure 2.4. High level amplification at 14q13. Alignment of six profiles showing a high level amplification at 14q13. Normalized log2 signal intensity ratios were plotted using SeeGH software, version 2.2.2. High level amplifications were defined as log2 signal intensities equal to or greater than 1. Vertical lines denote log2 signal ratios from -1 to 1 with copy number increases to the right (red lines) and decreases to the left (green lines) of zero (purple line). Each black dot represents a single BAC clone. The recurrently amplified region is shaded orange.
Figure 2.5. Multiple amplifications on chromosome 7. To highlight the regions of recurrent amplification on chromosome 7q in NSCLC, a 4x magnification of chromosome 7 frequency of alteration (Figure 2) is shown. The vertical blue bars represent the 6 regions of interest labeled A to F. Examples from individual SMRT array sample profiles are shown to the right of selected regions.
Figure 2.6. Multiple segmental amplifications on 7q. Colorimetric representation of 7q array CGH data viewed by Java TreeView. To compare multiple profiles, we used Java TreeView version 1.0.11-osx to generate a colored gene copy number matrix (http://jtreeview.sourceforge.net). Intensities of red and green coloration indicate an increased or decreased log2 signal ratio for each clone respectively. Each column represents a separate array CGH profile. Each row corresponds to a BAC clone and each column represents a sample (sample ID at top). 7q cytoband pattern is to the left. Recurrent amplifications are denoted by vertical blue lines to the right.
Figure 2.7. Comparision of lung adenocarcinoma and squamous cell carcinoma genomes. Frequencies of alteration were separately determined for AC and SqCC samples and visualized using SeeGH software version 2.2.2 as described in Figure 2. The plots were then overlaid to determine areas of difference between the two subtypes. Yellow represents regions frequently altered in both AC and SqCC while red and green regions are more frequently altered in AC and SqCC respectively.
2.5 References


Chapter 3: Integrative genomic and gene expression analysis of chromosome 7 identified novel oncogene loci in non-small cell lung cancer

A version of this chapter has been published as:


Please see the published version of this chapter for all supplementary materials.
3.1 Introduction

Lung cancer is broadly classified into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) types, with the latter group accounting for approximately 80% of cases and having an overall 5-year survival rate of ~15% (Travis et al, 1999). NSCLC is subclassified into multiple subtypes including adenocarcinomas (AC), squamous cell carcinomas (SqCC), and large cell carcinomas (LCC). While DNA sequence mutations and epigenetic alterations have been demonstrated to drive lung cancer oncogene activation (Wilson et al, 2006), changes in gene dosage have consistently been shown to be a major driving force in lung tumors (Lockwood et al, 2008). Although previous work has investigated gene expression dysregulation in lung cancer cells, there have been very few attempts to incorporate the impact of genomic alterations on gene transcription levels (Coe et al, 2006; Tonon et al, 2005; Zhou et al, 2006). Integrative analysis of recurring genomic and gene expression alterations in NSCLC tumours and cell models may be necessary to yield new insight into lung cancer biology.

Multiple studies have shown that aberrations in DNA copy number on chromosome 7 occur frequently in lung cancer, and that gain of chromosome 7 has been associated with NSCLC aggressiveness (Pei et al, 2001), a finding consistent with observations from other cancer types (Arslantas et al, 2007; Garcia et al, 2003; Waldman et al, 1991). Both focal regions of DNA amplification and whole chromosome number imbalances have been observed (Balsara & Testa, 2002; Panani & Roussos, 2006; Zojer et al, 2000). Alteration at chromosome 7 is generally thought to arise through selection for activation of the EGFR oncogene at 7p12, though others have shown that amplification of the MET oncogene at 7q31 can also occur (Engelman et al, 2007). The role of EGFR and MET activation has been elucidated in a subset of lung tumours (Engelman et al, 2007; Sequist et al, 2007). In a previous study, we observed that several regions of recurring genome alteration on chromosome 7 do not overlap with the EGFR and MET loci, raising the possibility of additional oncogene loci residing on this chromosome (Garnis et al, 2006). Those genes exhibiting over-expression in addition to genomic gain likely represent key “driver” alterations contributing to the cancer phenotype while those without concurrent mRNA increases are likely “passenger” alterations (Albertson, 2006). In this study, through the integration of genomic and gene expression analyses of cell lines and clinical tumour samples, we have uncovered additional lung cancer oncogene candidates situated on chromosome 7.
3.2 Results and Discussion

Segmental gain or amplification of chromosome 7 has typically been attributed to selection for
EGFR or MET oncogene activation (Engelman et al, 2007; Sequist et al, 2007). It has recently
been reported that additional regions of segmental genomic gain recur within chromosome 7 for
NSCLC cell lines (Garnis et al, 2006) (Table 3.1). (See Table S2 for a complete list the genes
located within each region). We have built upon this analysis, providing evidence for the
importance of candidate regions by (i) identifying high level amplification events and (ii)
quantifying concurrent gene expression changes in cell lines and clinical samples.

3.2.1 High level amplification events in frequently altered regions

The amplification of a chromosome segment may highlight its importance in cancer
development as amplicons are thought to represent the selection of genes that facilitate tumour
growth. Therefore, to refine the recurring regions of copy number gain on chromosome 7, we
first evaluated which of these regions showed high level amplification in at least two of the
analyzed cell lines. Table 3.1 summarizes the samples harbouring DNA amplification.
Remarkably, 19 of the 30 cell lines studied showed DNA amplification in Region 2.

3.2.2 Integration of gene dosage and expression data

While combining frequency of gain and the presence of high level amplification events is an
established approach for identifying key gene alterations in lung cancer (Tonon et al, 2005),
these genes are more likely to be involved in cancer biology if they can also be demonstrated to
have elevated expression levels. This is intuitive, given the accepted idea that amplified loci
offer a selective advantage by conferring such over-expression (Albertson, 2006). Expression
levels for genes residing within the five regions displaying high level amplification defined above
were determined from Affymetrix microarray data for the same NSCLC cell lines analyzed for
genomic alterations. Specifically, since we aimed to identify genes with expression driven by
increased gene dosage, the expression level for each of these genes (which were amplified in
two or more lines) was then compared between cell lines with amplification and those that
displayed neutral copy number by using the Mann-Whitney U-test. Of the 101 genes evaluated,
28 (27.7%) showed concurrent amplification and over-expression (Table 3.1). (Over-expressed
genes without concurrent copy number changes were likely activated by processes other than
copy number alteration.) At least eight of these 28 genes have also been described as having
increased copy number in other cancer types. Recently, Yang et al. showed significant
correlation between gene copy change and mRNA expression for GNB2, COPS6, and CCT6A
on chromosome 7 in gastric cancer (Yang, 2007). Additionally, concurrent copy number aberration and dysregulation of expression for MCM7, NUDT1, CCT6A, and GNB2 was noted in transformed follicle centre lymphoma (Martinez-Climent et al, 2003). EGFR amplification has been correlated with over-expression in a variety of cancer types, including NSCLC (Reis-Filho et al, 2006; Reissmann et al, 1999; Rossi et al, 2005).

3.2.3 Quantitative RT-PCR validation in cell lines

To validate the microarray data for gene expression changes within amplicons, real-time quantitative PCR was performed on select genes shown to be amplified and over-expression in the six frequently altered regions on chromosome 7 to validate the importance of these novel regions. Gene-specific probes for NUDT1 (Region 1), EGFR (Region 2), and TAF6 (Region 5) were used to assess expression levels in cell lines with both gene amplification and neutral copy number in comparison to pooled normal lung cDNA. On average, EGFR, NUDT1, and TAF6 gene expression was > 9, > 32, and > 5-fold over-expressed in cell lines with amplification when compared to normal lung cDNA (respectively) and only > 4, > 2, and > 3-fold over-expressed in cell lines with neutral copy number status compared to normal lung cDNA (respectively). Analysis by Mann-Whitney U-test showed that expression of each of these genes was significantly different between cell lines with amplification and lines with neutral copy number ($p \leq 0.05$). Figure 3.1 shows representative genome plots for cell lines with either amplification or neutral copy number status for NUDT1 and EGFR (corresponding expression data for each line is also shown). These data confirmed that several distinct regions contain genes activated in NSCLC.

3.2.4 Validation of genes of interest using gene expression data for clinical NSCLC tumours

To determine if the genes identified in our study using cell lines are in fact disrupted in clinical samples, we compiled publicly available gene expression microarray data for 111 clinical NSCLC samples (GEO Accession number GSE3141) (Figure 3.2). Nine genes from within these genomic loci exhibited $\geq 2$ fold expression compared to normal bronchial epithelial (NHBE) cells in at least 45% of lung tumours (~50 cases). These included FTSJ2 and NUDT1 (Region 1), TNS3, and ECOP (Region 2), and ZNF3, TAF6, TSC22D4, MOSPD, and POLR2J (Region 5).
3.2.5 Quantitative PCR validation of FTSJ2, NUDT1, TAF6, and POLR2J in clinical samples

Those genes previously implicated in cancer development and progression – FTSJ2, NUDT1, TAF6, and POLR2J – were selected for additional analysis. The FtsJ homolog 2 (FTSJ2) encodes a putative RNA methyltransferase previously implicated in cell proliferation and seen to be over-expressed in lung cancer cells (Ching et al., 2002). (Other RNA methyltransferases have recently been implicated in a variety of cancer types, supported the idea that such genes are key in carcinogenesis (Frye & Watt, 2006).) Nudix (nucleoside diphosphate linked moiety X)-type motif 1 (NUDT1) encodes an enzyme involved in maintaining genomic integrity and has previously been demonstrated to have elevated activity in NSCLC (Chong et al., 2006; Speina et al., 2005) and several other cancer types (Hibi et al., 1998; Kennedy et al., 1998; Okamoto et al., 1996; Wani et al., 1998). Our finding is the first report of amplification-driven NUDT1 over-expression. This is evident in both NSCLC cell lines and clinical samples. POLR2J (RNA Polymerase II [DNA-directed] polypeptide J), located in a different region of chromosome 7 observed to harbor recurring alterations, encodes a subunit of RNA Polymerase II and may contribute to transcriptional dysregulation that can drive malignant phenotypes (Fanciulli et al., 2000). The nearby TAF6 gene, similarly activated, encodes a component of the TFIID complex, which is known to initiate transcription by RNA polymerase II (Albright & Tjian, 2000; Green, 2000). This complex can act as a co-activator for upstream DNA binding transcription factors and recognizes the core promoter elements of a number of cancer-associated genes including P53, JUN, CCND1, GADD45, and P21 (Bell & Tora, 1999; Green, 2000; Lu & Levine, 1995; Thut et al., 1995). The capacity of TAF6 to regulate genes governing a variety of key cellular processes also makes it an attractive oncogene candidate.

Further validation of expression results for the FTSJ2, NUDT1, TAF6, and POLR2J genes was undertaken in a panel of ten fresh-frozen NSCLC and matched normal lung clinical samples. FTSJ2 and NUDT1 showed higher expression in all ten tumour samples (compared to matched normal tissue), with six samples having fold changes ≥ 2. TAF6 and POLR2J showed higher expression in nine tumour samples compared to the matched normal tissue, with two samples having at least double the expression in tumours. All genes were significantly over-expressed relative to normal by a one-tailed Wilcoxon sign-rank test (p ≤ 0.01) (Figure 3.3). This clear activation of putative oncogenes other than EGFR and MET supports the conclusion that lung tumourigenesis is driven by several distinct genome loci on chromosome 7. The fact that this
was made clear by integration of genomic and gene expression data demonstrates the utility of moving beyond uni-dimensional analysis to identify key genes driving cancer processes.

3.3 Materials and Methods

3.3.1 Cell line samples and DNA extraction

The 30 NSCLC cell lines used in this study are detailed in Table S1. These cell lines were established at the National Cancer Institute (NCI-H series) and the Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center (HCC series). All cell lines were either acquired from The American Type Culture Collection or supplied by Dr. John Minna from the Hamon Center for Therapeutic Oncology Research and grown according to specifications. DNA was isolated using a standard procedure with proteinase K digestion followed by phenol-chloroform extraction (Lockwood et al, 2007).

3.3.2 Tiling path array comparative genomic hybridization and data analysis

Segmental DNA copy number profiles were generated for each of the 30 NSCLC cell lines by whole genome tiling path aCGH as previously described (Ishkanian et al, 2004; Lockwood et al, 2007; Shadeo & Lam, 2006; Watson et al, 2007). Images of the hybridized arrays were then analyzed using SoftWoRx Tracker Spot Analysis software (Applied Precision), and systematic biases were removed from all array data files using a stepwise normalization procedure as previously described (Khojasteh et al, 2005; Lockwood et al, 2008). SeeGH was used to combine replicates and visualize all data as log2 ratio plots in karyograms (Chi et al, 2004; Chi et al, 2008). All raw array data files have been made publicly available through the System for Integrative Genomic Microarray Analysis (SIGMA) database, which can be accessed at http://sigma.bccrc.ca (Chari et al, 2006).

Regions of high level genomic amplification within each cell line were determined using an algorithm as previously described (Lockwood et al, 2008). Briefly, aCGH data was filtered to exclude clones with standard deviations between replicate values >0.075 and clones were identified as members of high level amplifications if its resulting log2 ratio was ≥0.8 (Choi et al, 2006; Choi et al, 2007; Lockwood et al, 2007). Moving averages of varying window sizes were used then used to identify amplicon boundaries (Lockwood et al, 2008).
3.3.3 Integration of copy number status and gene expression microarray data

Integration of gene expression and copy number data was performed as described previously (Lockwood et al, 2008). First, genes present in each amplicon were listed and those seen to be amplified two or more times in the 30 lines were identified. Copy number status (gain, loss or neutral) for each gene locus was defined using aCGH-Smooth (Jong et al, 2004) with lambda and breakpoint per chromosome settings at 6.75 and 100, respectively as previously described (Lockwood et al, 2008). Using these criteria, samples with neutral copy number (equal number of copies between tumor and normal reference DNA) for each gene of interest were defined. The algorithm described above was then used to determine if a cell line harbored amplification at the given locus. Array CGH measures relative and not absolute copy number and does not take into account changes due to ploidy. As such, alterations in gene dosage attributed to ploidy alone were not considered in this analysis. RNA expression profiles for 30 NSCLC cell lines were obtained from Gene Expression Omnibus (accession number GSE4824) (Zhou et al, 2006). The Affymetrix gene expression microarray probe sets corresponding to these genes were then determined and probes filtered for those demonstrating a present or marginal quality score in at least 50% of the amplified lines. The copy number status for each gene was then dichotomized to neutral vs amplified samples and gene expression data were then compared between the groups using the Mann-Whitney U test to identify those that were over-expressed in the amplified samples with a \( p \)-value \( \leq 0.05 \).

3.3.4 Quantitative real time PCR expression analysis of cell line and clinical tumour samples

cDNA was synthesized from 5 µg of total RNA using an ABI High Capacity cDNA Archive Kit (Applied Biosystems, Foster City, CA, USA). An aliquot of 100 ng of cDNA was used for each real-time PCR reaction. TaqMan gene expression assays were performed using standard TaqMan reagents and protocols on the Applied Biosystems 7500 Fast Real-Time PCR System (Applied Biosystems). Gene expression Assay IDs used include: \textit{NUDT1} (Hs00159343_m1), \textit{EGFR} (Hs00193306_m1) and \textit{TAF6} (Hs00425763_m1). Samples were run in triplicate and normalized against a \textit{eukaryotic 18S rRNA} endogenous control (Hs99999901_s1). The relative fold change of the target gene in each cell line sample compared to a pooled normal lung cDNA reference sample (AM7968, Ambion, Austin, TX, USA) was performed using the \( 2^{-\Delta\Delta Ct} \) method(Coe et al, 2006). The Mann-Whitney U-test was used to determine whether the expression of these genes was significantly different in cell lines with amplification compared to those with neutral copy number of the particular locus in question (p-value \( \leq 0.05 \)).
3.3.5 Analysis of publically available gene expression data for clinical lung tumours

Publically available Affymetrix gene expression data from 111 NSCLC tumours (53 SqCC and 58 AC) (GEO Accession number GSE3141) (Bild et al, 2006) and a normal human bronchial epithelial cell line (NHBE) (GEO Accession number GSE4824) (Zhou et al, 2006) was used to further validate the level of expression of candidate genes in our study shown to be amplified and over-expressed in the frequently altered regions. MAS5 normalized NSCLC data were used to calculate fold changes in expression for each gene compared to the NHBE reference sample. Affymetrix probes showing the highest overall signal intensity in the tumour samples were chosen for the analysis. The number of samples having ≥2 fold over-expression for the gene of interest was calculated. Genes with ≥45% of NSCLC samples having a ≥2 fold change were determined to be significant.

3.3.6 Quantitative real time PCR expression analysis of clinical samples

Ten fresh-frozen lung NSCLC tumours and their corresponding matched normal lung tissue were obtained from Vancouver General Hospital, Vancouver, B.C., Canada. Microdissection of tumour cells was performed under the guidance of two lung pathologists. Total RNA was isolated and 1 µg was converted into cDNA as described above. Gene-specific qPCR was performed for NUDT1, TAF6, POLR2J (Hs00196523_m1), FTSJ2 (Hs00203647_m1) and 18S rRNA as described above. 18S normalized cycle thresholds for each of the genes were compared between the tumours and corresponding normal tissue from the individual patients to determine the fold change in expression using the $2^{-\Delta\Delta Ct}$ method. Because these genes were hypothesized to be over-expressed owing to DNA amplification, a one-tailed Wilcoxon sign-rank test was used to determine whether expression of these genes was significantly different between matched tumour and normal samples (p-value <0.05) (Lockwood et al, 2008).
Figure 3.1. Representative genomic alterations within chromosome 7. (A) and (B) show aCGH profiles. Normalized log2 signal intensity ratios were plotted using SeeGH software. A log2 ratio of 0 represents equivalent copy number between the sample and reference DNA. Vertical lines denote log2 ratios from +1, +0.5, 0, -0.5, and 1, with copy number increases on the left and copy number decreases on the right of the centre line. Each black dot represents a single BAC clone. (A) aCGH profile for two cell lines, one with neutral and amplified copy number status for the EGFR locus. (B) aCGH profile for two cell lines, one with neutral and amplified copy number status for the NUDT1 locus. qPCR results are seen in (C) and (D). Fold-change expression levels were calculated for cell lines with amplified (in red) and neutral copy (in blue) number status of the EGFR locus (C) and NUDT1 locus (D) compared to a pooled normal lung cDNA reference sample. All samples were 18S-normalized and fold change calculations were performed using the 2-ΔddCt method. Significant Mann-Whitney U test p-values (≤ 0.05) indicate that the expression of EGFR and NUDT1 is significantly different in cell lines with amplification compared to those with neutral copy number. Fold-change values are listed on the vertical axis and sample names are listed on the horizontal axis.
Figure 3.2. Validation of amplified and over-expressed genes from candidate regions in a separate cohort of 111 NSCLC clinical tumours. The expression for the 28 genes from the six candidate regions which were amplified and over-expressed in the 30 NSCLC cell lines (Table 1) was assessed in a separate panel of 111 clinical NSCLC tumours to determine their relative expression compared to normal bronchial epithelial (NHBE) cells. Each candidate region is labelled according to its position on chromosome 7 and identified by an individual color. The respective genes from these regions which are amplified and over-expressed are presented as a histogram and color coded to match the respective region in which they are located. The percent of samples (n=111) in which each gene is ≥ 2 fold over-expressed compared to NHBE cells is presented. The red line depicts the cut-off of 45% of samples which was determined to be significant.
Figure 3.3. RT-qPCR results of candidate oncogenes in ten matched tumour/normal clinical samples. 18S-normalized cycle threshold values for the matched normal and tumour sample were used to calculate fold change values using the 2-ΔΔCt method. The results of a one-tailed Wilcoxon sign-rank test show that expression of FTSJ2, NUDT1, TAF6, and POLR2J is significantly different between matched tumour and normal samples (p < 0.05). Fold-change values are listed on the vertical axis. Fold-change values >5 are shown on the top of the graph above a hatch mark. Sample numbers are listed on the horizontal axis.
### Table 3.1. Amplified and overexpressed genes within regions of recurrent genomic gain on chromosome 7 in NSCLC cell lines

<table>
<thead>
<tr>
<th>Region*</th>
<th>Base pair position</th>
<th>Chromosome band location</th>
<th>Size (Mbp)</th>
<th># of genes in region</th>
<th>Cell lines with amplification</th>
<th>Genes amplified and overexpressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>telomere-6,162,500</td>
<td>7p22.3-7p22.1</td>
<td>~6</td>
<td>40</td>
<td>HCC2279, HCC193, HCC1833</td>
<td>FTSJ2, NUDT1</td>
</tr>
<tr>
<td>2</td>
<td>16,296,094-57,693,752</td>
<td>7p15.3-7p11.2</td>
<td>~41</td>
<td>164</td>
<td>HCC461, HCC4006, H2087, H2009, H358, H2122, H1229, H1650, H3255, HCC827, HCC279, HCC193, H1993, HCC78, H1819, H157, HCC95</td>
<td>YKT6, TNS3, FIGNL1, EGFR, LANCL2, ECOP, FKBP9, MRPS17, GBAS, PSPH, CCT6A, SUMF2, CHCHD2</td>
</tr>
<tr>
<td>3</td>
<td>60,748,436-66,207,032</td>
<td>7q11.1-7q11.21</td>
<td>~5.4</td>
<td>18</td>
<td>none</td>
<td>N/A</td>
</tr>
<tr>
<td>4</td>
<td>71,745,312-75,278,128</td>
<td>7q11.23</td>
<td>~3.5</td>
<td>41</td>
<td>HCC193, HCC1195, H358, H1819, H1229</td>
<td>DKFZP434A0131</td>
</tr>
<tr>
<td>5</td>
<td>97,643,752-101,681,248</td>
<td>7q22.1</td>
<td>~4</td>
<td>75</td>
<td>HCC95, HCC1195, H2122, H1993</td>
<td>G10, PTCD1, ZNF38, ZNF3, COP56, MCM7, TAF6, TSC22D4, MOSPD3, GNB2, AP1S1, POLR2J</td>
</tr>
<tr>
<td>6</td>
<td>156,267,184-158,524,992</td>
<td>7q36.3</td>
<td>~2.2</td>
<td>6</td>
<td>HCC336, HCC193</td>
<td>none</td>
</tr>
</tbody>
</table>

* Regions according to Garnis et al 2006.
3.4 References


Chong IW, Chang MY, Chang HC, Yu YP, Sheu CC, Tsai JR, Hung JY, Chou SH, Tsai MS, Hwang JJ, Lin SR (2006) Great potential of a panel of multiple hMTH1, SPD, ITGA11 and


Lockwood WW, Chari R, Coe BP, Girard L, Macaulay C, Lam S, Gazdar AF, Minna JD, Lam WL (2008) DNA amplification is a ubiquitous mechanism of oncogene activation in lung and other cancers. *Oncogene*


Rossi MR, La Duca J, Matsui S, Nowak NJ, Hawthorn L, Cowell JK (2005) Novel amplicons on the short arm of chromosome 7 identified using high resolution array CGH contain over expressed genes in addition to EGFR in glioblastoma multiforme. *Genes Chromosomes Cancer* 44: 392-404


Chapter 4: DNA amplification is a ubiquitous mechanism of oncogene activation in lung and other cancers

A version of this chapter has been published as:


Please see the published version of this chapter for all supplementary materials.
4.1 Introduction

Genetic aberration and the consequential activation of oncogenes are key to cancer development. Chromosomal translocation is known as the major event in oncogene activation (Futreal et al., 2004). However, the prevalence of alternate mechanisms such as DNA amplification have not been extensively quantified, even though oncogenes have been found in (i) cytogenetically visible double minutes (DM) which are circular, extrachromosomal elements a few megabases in size that replicate autonomously, (ii) homogenous staining regions (HSR) which are large regions of tandem repeats within a chromosome thought to be formed by repeated Breakage-Fusion-Bridge cycles and (iii) discrete insertions distributed throughout the genome (Albertson, 2006).

Surprisingly, relatively few oncogenes, when compared to chromosome translocation, have been shown to undergo amplification as mechanism of activation during cancer development. In fact, a recent version (January 22, 2007) of a census of genes causally implicated in cancer (cancer genes) originally described by Futreal et al reported only seven oncogenes meeting their criteria as being recurrently amplified in the development of human cancers: AKT2 in ovarian cancer, ERBB2 in breast and ovarian cancer, MYCL1 in small cell lung cancer (SCLC), MYCN in neuroblastoma, REL in Hodgkin lymphoma, EGFR in glioma and non-small cell lung cancer (NSCLC), and MYC in numerous cancers (Futreal et al., 2004). We propose that the low incidence of oncogenes reported to be activated by amplification may be attributed to the failure of detection rather than governed by tumor biology. Unlike copy number gains which are generated by aneuploidy or unbalanced translocations and affect large chromosomal regions, amplifications are traditionally defined as the increase of chromosome segments 0.5-10 megabases (Mb) in size (Myllykangas et al., 2006). The small size of amplicons may escape detection by conventional cytogenetic methods; consequently, the contribution of DNA amplification to the oncogenic process may be grossly underestimated. With advances in high resolution whole genome profiling technologies (Garnis et al., 2006; Tonon et al., 2005), the complexity of the cancer genome is becoming evident, and the prevalence of DNA amplification as a mechanism in the activation of oncogenes needs to be re-evaluated.

In this study, we determined the precise boundaries of amplified chromosomal segments in 104 cancer cell lines from multiple tissues of origin and deduced novel regions of the genome which are hotspots for genomic amplification. These hotspots were then analyzed for their association with genes involved in tumorigenesis and fragile sites. We assessed the functional impact of a
subset of the identified hotspots in a panel of NSCLC cell lines and tumors to determine their effect on gene transcription levels and their contribution to the activation of cellular pathways potentially involved in lung tumorigenesis.

4.2 Results

4.2.1 Identification of discrete amplicons in cancer genomes

24,892 genomic loci were assessed for each of the 104 cell lines, scanning all autosomes at a resolution of ~50 kb (Coe et al., 2007). Altogether, 3431 amplicons were detected across all samples (see Supplementary Methods) with an average size of 0.68 Mb and a median of 0.33 Mb (Table 4.1, supplementary information (SI) Table 1). The number of amplicons per genome varied from 0 to 199 with an average of 33. Hematological malignancies (leukemia and lymphomas) had ~10 amplicons whereas epithelial cancers had an average of 36.

4.2.2 Unexpected frequent amplification of known oncogenes

The most recent version (January 22, 2007) of the Cancer Gene Consensus of the Cancer Genome Project at the Sanger Institute (Futreal et al., 2004), contains 363 cancer genes whose aberration are causal in the development of specific cancers. Of these, 70 are tumor suppressor genes, 292 are oncogenes and one can act as both. Only seven (2%) of these oncogenes were shown to be predominately activated by amplification compared to 268 (92%) which are activated mainly by chromosomal translocation. Our data showed amplification at these loci: MYC (28/104), ERBB2 (10/104), EGFR (7/104), MYCL1 (6/104), AKT2 (3/104) and MYCN (1/104). REL amplification was not detected in our dataset, as Hodgkin Lymphoma, in which this gene is amplified, is not represented in our study.

Unexpectedly, 145 of the 292 oncogenes (~50%) showed amplification, with 78 oncogenes (27%) at ≥ 2 times (SI Table 2). Of the genes amplified in ≥ 5 cell lines, only MYC, ERBB2, EGFR and MYCL1 have been reported. The frequent amplification of SS18L1, NTRK1 and PRDM16 are novel findings, as translocation was the known mechanism. Indeed, numerous oncogenes which are primarily activated by translocation were commonly amplified in the sample set (SI Table 3). The number of oncogenes amplified per genome also varied with an average of 3.5 genes. Remarkably, the genomes of NSCLC HCC1195 (SI Fig. 1) and SCLC line H526 each harbor 22 amplified oncogenes, whereas 25 of the lines had no known oncogenes amplified (SI Table 4).
4.2.3 Novel hotspots of frequent genomic amplification in cancer genomes

The high incidence of oncogene amplification per genome suggested that this is a common mechanism of gene activation. Therefore, the discovery of genomic regions that undergo frequent copy number amplification may lead to the identification of novel oncogenes. The genomic coordinates of all amplicons were determined and aligned for all 104 samples (Figure 4.1). DNA segments amplified ≥5 times were stringently considered as hotspots; they are found in ~5% of samples (see Supplementary Methods). In total, 135 hotspots covering 3% of genome were identified with an average size of 0.67 Mb. Regions of genomic amplification were distributed on all autosomes except chromosome 4, and in all tumor types analyzed (SI Table 5). Amplicons are most frequently localized to 1q21-23, 5p15, 7p13-11, 8q22-24, 11q13, 14q12-21, 14q32, 17q12-21, and 20q13. A total of 538 unique genes were contained within the hotspots (SI Table 6) (see Supplementary Methods). Interestingly, the majority of these hotspots did not contain the 292 known oncogenes.

There was no association between amplification hotspots and known fragile sites in the human genome based on chi-squared test at the chromosome band level, even though co-localization do exist, for example, the three fragile sites on 8q. Figure 4.1 summaries the location of the 86 common fragile sites assayed relative to hotspots of amplification.

4.2.4 Novel amplification hotspots contain putative oncogenes

Remarkably, 27 of the top 100 most frequently amplified genes have been previously described to be overexpressed in various cancers (SI Table 7), but aside from MYC and ERBB2, the mechanism leading to overexpression of these genes was largely unknown. To further explore the properties of the amplified genes, functional and biological characteristics were evaluated through the use of Ingenuity Pathways Analysis (Ingenuity® Systems, see (see Supplementary Methods). Functional Analysis identified a significant association between the amplified genes with genes involved in cancer \( (p=6.67E-06-3.03E-02) \); the two significance values refer to a range of specific sub-functions) and other diseases (SI Tables 8 and 9). Furthermore, Canonical Pathway Analysis was used to determine the main signaling pathways in which the amplified genes were involved (SI Table 10). Neuregulin Signaling \( (p=1.12E-02) \), also known as EGFR family signaling, was the most significantly affected with GRB7, SHC1, SRC, EGFR, ERBB2, and AKT1 comprising the amplified genes.
4.2.5 Impact of amplification on gene expression levels

To understand the effects of amplification on gene regulation and transcription, we focused on one type of cancer. Parallel gene expression profiles and array CGH data were integrated for 27 NSCLC cell lines. The expression levels for genes within amplification hotspots (displayed in SI Fig. 2) were compared between samples with amplification and those with neutral copy number status using the Mann-Whitney U Test (see Supplementary Methods). In total, 221 out of 442 of the amplified genes were expressed at significantly higher levels ($p \leq 0.05$) with increased gene dosage (Figure 4.2 and SI Table 11). For the majority of these genes, amplification is a novel mechanism for activation, although the expression levels of a subset such as $MYC$, $EGFR$, $CDK4$, $MAFB$, and $MET$ are known to be affected by increase in gene dosage.

4.2.6 Multiple components of the EGFR family signaling pathway are activated by DNA amplification in NSCLC cell lines and clinical tumors

To relate the genes activated by amplification in NSCLC to biological functions, Functional and Canonical Pathway Analysis were performed using IPA software (SI Tables 12 and 13). EGFR family signaling was the most affected canonical pathway ($p=6.03E-03$) with five genes: $AKT1$, $CDK5$, $EGFR$, $MYC$, and $SHC1$ amplified and overexpressed in the 27 NSCLCs (Table 4.2, Figure 4.2). The amplification and subsequent overexpression of $AKT1$, $CDK5$, and $SHC1$ are novel findings in NSCLC. Figure 4.3 displays the interaction of these genes during EGFR family signaling and the resulting downstream effects of the activation of this pathway which includes cell proliferation and survival. Interestingly, nearly 60% of the cell lines analyzed had one or more components of the EGFR family pathway overexpressed as a result of amplification. The alteration of $EGFR$ and $MYC$ alone could not explain pathway disruption in all cases as $\sim 31\%$ (5/16) of samples with activated downstream components harbored amplification of one or more of $CDK5$, $SHC1$ or $AKT1$ independent of $EGFR$ and $MYC$. Furthermore, by breaking the NSCLC lines down into their histological subtypes, it was discovered that 14 out of 20 (70%) adenocarcinoma samples -- whereas no squamous and only one large cell carcinoma samples -- had altered components, suggesting that the disruption of this pathway is prevalent in the adenocarcinoma subtype of lung cancer.

To further validate our results, quantitative real-time PCR was performed on select genes, $AKT1$, $CDK5$, and $SHC1$. First, expression levels for these genes were determined using the $\Delta\Delta$Ct method and compared between cell lines and normal lung tissue to confirm their
overexpression. Relative to the normal lung reference, \( AKT1 \) was 10.49 fold overexpressed in samples with gene amplification compared to 1.94 fold overexpression in NSCLC cells with neutral copy number status for this gene (SI Fig. 3). Likewise, \( CDK5 \) (SI Fig. 4) and \( SHC1 \) (Figure 2.4) also showed higher expression with increase gene dosage, suggesting a strong correlation of gene dosage and transcription levels for these genes. Expression changes held true in clinical specimens as clinical adenocarcinoma samples frequently showed overexpression of \( AKT1, CDK5, \) and \( SHC1 \) compared to their corresponding matched normal lung tissues (SI Fig. 3c, 4c and Fig. 4c, respectively). Since these genes were hypothesized to be overexpressed due to DNA amplification, a one-tailed Wilcoxon sign-rank test was used to determine whether overexpression of these genes was significant in the set of matched tumor and normal samples. Indeed, each gene was significantly overexpressed in the tumors compared to their matched normal \((p<0.01)\) confirming the results from the cell lines (qPCR data is provide in SI Tables 14 and 15).

4.3 Discussion

Oncogene activation is traditionally associated with translocation events. We hypothesized that DNA amplification is a prevalent, but underestimated, mechanism of oncogene activation in cancer genomes. To our knowledge, no studies to date have assembled a large panel of paired high resolution copy number and gene expression data to accurately assess this question. In this study, we examined 104 cancer cell lines comprising various tissues of origin (Table 1) at 24,892 autosomal loci per genome, a resolution, that detected amplicons as small as 0.05 Mb in size (Ishkanian et al., 2004) and discovered that not only is the incidence of oncogene amplification much greater than previously believed, but specific regions of the genome are hotspots for segmental amplification in cancer cells.

4.3.1 Amplification as a major mechanism of oncogene activation

The activation of oncogenes is a hallmark of tumor development. Cancer cells frequently display chromosome rearrangements resulting in the deregulation of gene expression, as well as in the fusion of genes raising oncogenic activity. As such, the majority of known oncogenes, including 92% of those analyzed in this study, have been discovered through their involvement in disease-specific chromosomal translocations (Futreal et al., 2004). Thus, the high incidence of amplification we report suggests that oncogenes may have multiple mechanisms of activation, with the increase in gene dosage being a prominent mechanism of activation. This was particularly evident in the fact that genes which have been shown to be activated primarily
by translocation were frequently amplified (SI Table 3). The majority of these genes have not been shown to be activated by amplification previously and as such, this data represent a novel finding. For example, t(14;20)(q32;q12) translocation is known to juxtapose IgH enhancers to the MAFB gene locus upregulating its expression in multiple myeloma (Wang et al, 1999) (Boersma-Vreugdenhil et al, 2004). We demonstrated that MAFB amplification also occurs in lung (H1395, H1650 and H1666), cervical (SW756), and liver (HepG2) cancer cells. Likewise, NTRK1, is an oncogene frequently activated by translocation (Roccato et al, 2005). Fusion with TPR, TPM3 or TFG results in constitutive tyrosine kinase activity (Pierotti et al, 1996). Remarkably, we detected also NTRK1 copy number increased in lung (HCC366, HCC1833, HCC1195, H82, H526, H2122, and H187) and breast (ZR7530) cancer genomes, suggesting that amplification and subsequent overexpression may be an alternate mechanism of activation.

4.3.2 Existence of an amplifier phenotype

There is evidence suggesting that some cancer cells have a greater propensity to undergo DNA amplification than others and that there is an underlying genetic basis for this “amplifier phenotype” (Albertson, 2006). Our data showed that the number of oncogenes amplified may differ in individual genomes. The variation existed across both general cancer classes and individual tissue types. In addition, amplification of the same genes in the different tumor types suggests that there may be a selective advantage to have certain genes or their related functions elevated in the context of cancer development and that amplifications are not simply by-products of general genomic instability characteristic of late stage tumors. It was also common for a sample to simultaneously harbor multiple amplicons on different chromosomes, highlighting the possibility of an underlying genetic basis for amplification development (SI Fig. 1). It has been proposed that amplifications are mainly related to solid tumors and are seldom involved in hematological malignancies in which oncogene activation is generally associated with translocations (Mitelman, 2000). Indeed, only 11 known oncogenes were amplified in leukemia or lymphoma genomes in our dataset, whereas 144 were amplified in epithelial cases (SI Table 2). Similar results were observed in the number of amplicons in each genome (~10 per hematological, ~36 in epithelial samples). These results suggest that subsets of cancers such as the epithelial cancers are driven by an amplifier phenotype, whereas others, typically hematological malignancies, develop mainly through different genetic mechanisms such as chromosomal translocation.
4.3.3 Integration of copy number status and gene expression microarray data

Figure 4.1 indicates that regions of frequent copy number amplification are preferentially localized in genome. These results are similar to those found in a bibliomics survey which looked across 73 different neoplasms using conventional CGH (Myllykangas et al., 2006). However, in this study we further refined the regions beyond the chromosome band level, determining the exact genes affected by these aberrations (SI Table 6), identifying novel hotspots such as the discrete amplicons on 14q.

There are two potential factors that may determine the localization of amplification hotspots. First, the selective pressure imposed on the tumor may lead to the selection of amplification of regions containing genes advantageous to tumor growth. Consistent with this theory, we observed a significant enrichment for genes involved in cellular functions and canonical pathways commonly involved in tumorigenesis within the amplification hotspots (SI Tables 8-10). Many genes implicated in key biological processes such as cell cycle and cell growth and proliferation contained within these regions that may be considered as novel candidate oncogenes. In addition, 27 of the top 100 genes within the hotspots have been previously described to be overexpressed in cancer further supporting their oncogenic role (SI Table 7). Since the mechanism leading to the overexpression of the majority of these genes was previously unknown, our data suggests that amplification may be a key mechanism of their activation.

Second, the intrinsic features of the chromosome regions themselves may be involved in their preferential amplification (Wahl et al., 1984). Mechanistic models such as breakage-fusion-bridge and episome excision imply that two double stranded DNA breaks are required to initiate amplification generation (Myllykangas & Knuutila, 2006). As such, it has been proposed that regions of the genome which are more susceptible to breakage have a greater propensity to undergo amplification, i.e. fragile sites (Buttel et al., 2004; Hellman et al., 2002). Conventional CGH studies indicated that many amplification hotspots co-localized with fragile sites, however, this association was not statistically significant on a genome-wide scale, presumably due to inadequate resolution (with chromosome band level comparison) as fragile sites and amplification hotspots covered 30% and 45% of the genome respectively (Myllykangas et al., 2006). Our analysis addressed this problem as the increased resolution of the platform used in this study allowed the refinement of amplification hotspots, limiting their coverage to ~3% of the genome. However, although we observed a general trend of the co-localization of amplification hotspots and common fragile sites (Figure 4.1), the global association was still not significant.
We speculate that the cloning of fragile sites to determine their specific sequences is needed to complement our array CGH data in order to accurately assess their association. Nevertheless, there is a strong possibility that the hotspot regions represent damage prone sites in the genome and further investigation in the future is warranted. Notably, in addition to fragile sites, other genomic features such as copy number variations and segmental duplications may also contribute to DNA rearrangement in cancer cells (Squire et al, 2003).

4.3.4 Global impact of amplification on gene expression levels in NSCLC

Although array CGH allows the fine mapping of amplification boundaries at unprecedented resolution, multiple genes may map to an individual amplicon. Therefore, integration of copy number and expression data is needed to distinguish overexpressed genes from bystander genes within amplicons. To accomplish this, we integrated parallel array CGH and expression data for a subset of 27 NSCLC cell lines. Aside from genes known to be activated by amplification in NSCLC such as EGFR and MYC, our data suggest that the expression of oncogenes including CDK4, MAFB, BCL11B and MET are also driven by amplification (Figure 4.2). Since these genes are typically activated by translocation or missense mutation in malignancies other than NSCLC, this data further supports our hypothesis that amplification is an alternate mechanism of oncogene activation in a subset of cancers.

The integration of the datasets identified expressed genes within novel amplification hotspots in lung cancer that are potentially involved in tumorigenesis. For example, Thyroid Transcription Factor 1 (TITF1) in the novel hotspot at 14q12-q13 is known to be overexpressed specifically in lung adenocarcinoma, the predominate subtype of NSCLC from which the 27 cell lines in our study were derived (Fabbro et al, 1996). This gene encodes a homeodomain transcription factor that is involved in regulating pulmonary development and gene expression (Apergis et al, 1998) and has been proposed to be a lineage marker for tumors arising from the peripheral airway (Stenhouse et al, 2004). Adenocarcinomas that express TITF1 are dependent on its persistent expression for survival (Tanaka et al, 2007). Acquired somatic alteration of these genes during differentiation leads to aberrant lineage-survival pathway signaling. The resulting tumors become addicted to persistent expression of these genes for survival, known as “lineage addiction” (Garraway & Sellers, 2006). Our data suggests that amplification may be a mechanism driving the expression of lineage-survival oncogenes in a subset of cancers, and further supports a role for TITF1 in lung adenocarcinoma tumorigenesis (Figure 4.2 and SI Fig. 5).
On a global scale, we found amplification had a strong impact on transcription levels as 50% of the genes from amplification hotspots within these samples showed enhanced expression as a consequence of alteration. This falls within observations of previous studies which reported 19.3% to 62% of amplified genes being overexpressed (Heidenblad et al, 2005; Hyman et al, 2002; Pollack et al, 2002; Wolf et al, 2004). Nevertheless, since this is the first study to integrate high resolution copy number and gene expression profiles in lung cancer on a whole genome scale, future analysis will be needed to confirm the functional impact of the genes in each amplicon.

4.3.5 Novel disruptions of the EGFR family signaling pathway in NSCLC by gene amplification

Our data indicate that multiple components of the EGFR family signaling pathway are frequently amplified and overexpressed in NSCLC (Figure 4.3). Deregulation of this network commonly occurs in cancer and specifically NSCLC. In NSCLC, the mechanism of deregulation is usually attributed to receptor overexpression or point mutations in the catalytic domain resulting in ligand-independent constitutive receptor activation and signaling (Bubil & Yarden, 2007). However, in cases with normal levels of wild-type receptor, aberrant constitutive signaling may also occur. Strikingly, our data suggest that amplification of downstream signaling components may be an alternate mechanism of pathway activation in a subset of tumors. Although alteration of this pathway was detected in 59% of the cell lines analyzed, only 31% of these cases could be explained by EGFR activation. Indeed, the majority of cell lines with pathway perturbation (69%) contained amplification of key signaling components downstream of the receptor level (Fig. 3). The overexpression of these components was confirmed in NSCLC tumors, highlighting their clinical significance (Fig. 4 and SI Figs. 3 and 4). Although MYC amplification has been previously reported, this is the first study to describe the frequent amplification and overexpression of SHC1, AKT1 and CDK5 in NSCLC. Since these genes are involved in the activation of mitogenic signaling pathways involved in EGFR induced transformation and their overexpression has been previously implicated in cancer, these results suggest that their direct genetic activation may play a causal role in NSCLC tumorigenesis and highlights the impact of these novel amplifications on cancer biology (Hennessy et al, 2005).

The direct amplification of downstream components may also have substantial effect on the response to clinical treatment strategies. The high frequency of EGFR family overexpression in cancer has led to the development of targeted therapeutics aimed at inhibiting receptor function. For example, anti ErbB2 antibodies are currently used for breast cancer treatment and EGFR
specific tyrosine kinase inhibitors (TKIs) such as Gefitinib and Erlotinib are used in NSCLC therapy (Bublil & Yarden, 2007). The receptor-independent activation of downstream signaling components would impact the effectiveness of these treatment strategies as constitutive activation of signaling pathways would occur regardless of receptor inhibition. Previous studies have shown that downregulation of the AKT/PI3K signaling pathway is required for EGFR TKIs to induce apoptosis in cancer cells (Hemstrom et al., 2006). In addition, activated AKT/PI3K signaling due to MET amplification has been shown to lead to Gefitinib resistance in NSCLC cells (Engelman et al., 2007). Thus, the direct activation of AKT1 and CDK5 would lead to resistance in a NSCLC due to maintained AKT/PI3K signaling in the presence of inhibitor. Likewise, in a drug-resistant NSCLC cell line, alterations of adaptor-protein mediated signal transduction from EGFR, such as those initiated by SHC1, has been proposed as a possible mechanism of resistance to Gefitinib (Koizumi et al., 2005). Our findings highlight the need to assess the activation status of downstream signaling components and suggest that amplification of AKT1, SHC1, CDK5 and MYC may be used for this process.

4.4 Conclusion

Since alterations at the DNA level potentially represent causal events in the development of cancer, the genes deregulated as a result of amplification in NSCLC can be viewed as the primary oncogenic targets in a tumor that lead to downstream pathway abrogation. The gain of function effect of gene amplifications makes them ideal targets for therapeutic intervention due to the direct nature of their activation and the fact that a tumor can become addicted to their enhanced expression (Weinstein, 2002). Although cancer cell lines were used in this study, the genomic and transcriptional characteristics of such models have been shown to mirror primary tumors and are appropriate systems to identify molecular features that predict or indicate response to targeted therapies (Greshock et al., 2007; Neve et al., 2006). Furthermore, validation of a subset of amplified genes in primary tumors confirmed their clinical importance. Future studies of additional primary tumors will be required to further validate the role of the hotspots in clinical specimens and confirm that they are not artifacts of in vitro culture. Our discovery of high incidence of amplification suggests that it is a major mechanism of oncogene activation in cancer and will provide essential starting points for the discovery of novel oncogenes.
4.5 Materials and Methods

4.5.1 Whole genome profiling

DNA copy number profiles for 104 cancer cell lines of lung, breast, prostate, cervical, skin, ovarian, liver and hematological origins were used in this study (SI Table 16). DNA was isolated by proteinase K digestion followed by phenol-chloroform extraction. Array hybridization was performed as previously described (Lockwood et al., 2007) using SMRT array v.2 (Ishkanian et al., 2004; Watson et al., 2007). Array images were analyzed using SoftWoRx Tracker Spot Analysis software (Applied Precision, Issaquah, WA). Systematic biases were removed using the stepwise normalization procedure CGH Norm (Khojasteh et al., 2005). SeeGH software allowed visualization of log2 ratio plots in karyograms (Chi et al., 2004). All raw array data files have been made publicly available through the System for Integrative Genomic Microarray Analysis (SIGMA), which can be accessed at http://sigma.bccrc.ca (Chari et al., 2006).

4.5.2 Gene expression profiling

RNA samples from 27 NSCLC cell lines and normal human bronchial epithelial (NHBE) cells were analyzed using the Affymetrix Gene Chips HG-U133A and HG-U133B (Henderson et al., 2005) (Zhou et al., 2006) (SI Table 16). These arrays together represent 23,583 unique genes based on Unigene build 173. The identity of these cell lines have been verified by DNA fingerprint using the Powerplex 1.2 system (Promega). Data normalization and microarray analysis was performed using Affymetrix Microarray Suite 5.0 as previously described (Zhou et al., 2006). The microarray data have been uploaded to GEO (Gene Expression Omnibus, accession number GSE-4824).

4.5.3 Statistical analysis of array data

Detailed methods describing the statistical analysis used for the identification of amplicons, amplification hotspots, amplification hotspots and fragile site co-localization, functional assessment of amplified genes and integration of genomic and gene expression data are provided in Supplementary Methods.

4.5.4 Gene specific quantitative real-time reverse transcriptase PCR analysis

TaqMan gene expression assays [AKT1 (Hs00178289_m1), SHC1 (Hs00427539_m1), CDK5 (Hs00358991_g1), and 18s rRNA (Hs99999901_s1)] were performed using 100 ng of cDNA samples in a Applied Biosystems 7500 Fast Real-Time PCR System (Applied Biosystems,
Foster City, CA). The △△Ct method was used for expression quantification using the average cycle threshold of 18S rRNA for normalization (Coe et al, 2006) and Human Lung Total RNA (AM7968, Ambion, Austin, TX) as a reference. For clinical samples, total RNA was isolated from ten microdissected frozen lung adenocarcinoma and matched normal tissue obtained from Vancouver General Hospital using RNeasy Mini Kits (QIAGEN Inc., Mississauga, ON) and 1 µg was converted to cDNA for gene-specific quantitative PCR for AKT1, SHC1, CDK5, and 18S rRNA. Cycle thresholds comparison yielded expression changes in the tumors. Because these genes were hypothesized to be overexpressed owing to DNA amplification, a one-tailed Wilcoxon sign-rank test was used to determine whether overexpression was significant in the set of matched tumor and normal samples.

4.5.5 Fluorescence in situ hybridization (FISH)

FISH was performed as previously described (Watson et al, 2004). Briefly, 100 ng of linker-mediated PCR-amplified BAC DNA was labeled through a random priming reaction with Spectrum Green or Red dUTP (Vysis, Markham, ON). Hybridization was performed in a 50% formamide buffer at 37°C for 18 hours and imaged with Q Capture imaging software (Q Imaging, Burnaby, BC).
Figure 4.1. Hotspots of amplification in cancer genomes. A histogram summarizing the regions of amplification across all 104 samples with the resulting values scaled to the segment with the highest count (28) and plotted against their corresponding genomic position. Hotspots are denoted by the dark blue shading, while the light blue shading represents regions amplified ≤5 times. Triangles mark common fragile sites. Detailed genomic position of hotspots and common fragile sites are provided in SI Tables 5 and 17.
Figure 4.2. Impact of amplification on gene transcription levels. The relative expression values for samples with amplification and those with neutral copy number status are plotted as heatmaps for overexpressed genes from representative hotspots. The expression values for each gene have been normalized and scaled across the samples from 0 to 100.
Figure 4.3

**Figure 4.3. Frequent amplification and overexpression of multiple EGFR family signaling components in NSCLC.** Diagram highlighting the interaction of EGFR, SHC1, CDK5, SHC1 and MYC in the EGFR family signaling pathway. Altered components are shaded grey. The table summarizes the number and specific samples with amplification and overexpression of each pathway component. The total represents the number of samples with at least one pathway component amplified and overexpressed.
Figure 4.4. SHC1 disruption in NSCLC cell lines and clinical tumors. (A) Representative array CGH profiles for samples with and without SHC1 amplification. Vertical lines denote log2 signal ratios from -1 to 1 with copy number increases to the right (red lines) and decreases to the left (green lines) of 0 (purple line), with amplified region shaded orange. Red and green arrows mark clones used in subsequent FISH analysis. (B) SHC1 expression in NSCLC cell lines. The normalized fold change of expression compared to a normal lung reference is plotted for samples with amplification (red) and those with neutral copy number status (black). (C) Overexpression of SHC1 in clinical tumors. The log2 fold change in expression levels of SHC1 relative to their matched normal lung tissue is plotted for each tumor. The p-value from the Wilcoxon sign-rank test is indicated. (D and E) FISH confirmation of SHC1 amplification in H1395. FISH was performed using BAC clones mapping to SHC1 (RP11-624P9) and to an adjacent neutral copy number region (RP11-313J15).
Table 4.1. Summary and distribution of amplicons by cancer type.

<table>
<thead>
<tr>
<th>Tissue Type</th>
<th># Samples</th>
<th>Total # Amplicons</th>
<th>Amplicons/Tumor</th>
<th>Average Size (Mbp)</th>
<th>*Most Frequent Amplification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung</td>
<td>53</td>
<td>1690</td>
<td>31.9</td>
<td>0.68</td>
<td>8q24.21 (28%)</td>
</tr>
<tr>
<td>Breast</td>
<td>17</td>
<td>905</td>
<td>53.2</td>
<td>0.77</td>
<td>8q24.21 (59%)</td>
</tr>
<tr>
<td>Lymphoid</td>
<td>9</td>
<td>101</td>
<td>11.2</td>
<td>0.54</td>
<td>9p13.3, 13q31.3, 18q21.33-22.1 (33%)</td>
</tr>
<tr>
<td>Cervix</td>
<td>8</td>
<td>45</td>
<td>5.6</td>
<td>0.65</td>
<td>5p15.33 (38%)</td>
</tr>
<tr>
<td>Skin</td>
<td>4</td>
<td>395</td>
<td>98.8</td>
<td>0.35</td>
<td>7p13.3 &amp; 7q35 (75%)</td>
</tr>
<tr>
<td>Blood</td>
<td>3</td>
<td>23</td>
<td>7.7</td>
<td>0.80</td>
<td>N/A</td>
</tr>
<tr>
<td>Prostate</td>
<td>3</td>
<td>82</td>
<td>27.3</td>
<td>0.86</td>
<td>14q21.3-22.1 (67%)</td>
</tr>
<tr>
<td>Bone</td>
<td>2</td>
<td>104</td>
<td>52</td>
<td>1.02</td>
<td>N/A</td>
</tr>
<tr>
<td>Colon</td>
<td>2</td>
<td>22</td>
<td>11</td>
<td>0.70</td>
<td>N/A</td>
</tr>
<tr>
<td>Ovary</td>
<td>2</td>
<td>3</td>
<td>1.5</td>
<td>0.33</td>
<td>N/A</td>
</tr>
<tr>
<td>Liver</td>
<td>1</td>
<td>61</td>
<td>61</td>
<td>0.83</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>104</strong></td>
<td><strong>3431</strong></td>
<td><strong>32.8</strong></td>
<td><strong>0.68</strong></td>
<td></td>
</tr>
</tbody>
</table>

* The most frequent regions were only determined if found in >2 samples for the corresponding tissue type.
Table 4.2. Canonical pathways affected by amplification in NSCLC.

<table>
<thead>
<tr>
<th>Signaling Pathway</th>
<th>P-value</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuregulin</td>
<td>6.0E-03</td>
<td>MYC, CDK5, SHC1, EGFR, AKT1</td>
</tr>
<tr>
<td>Huntington's Disease</td>
<td>1.1E-02</td>
<td>SDHA, POLR2H, CDK5, SHC1, POLR2J, EGFR, AKT1</td>
</tr>
<tr>
<td>VEGF</td>
<td>3.1E-02</td>
<td>VEGF, ARNT, SHC1, AKT1</td>
</tr>
<tr>
<td>Insulin Receptor</td>
<td>3.2E-02</td>
<td>PARD3, PPP1R3D, PPP1CA, SHC1, AKT1</td>
</tr>
<tr>
<td>Nitric Oxide</td>
<td>3.9E-02</td>
<td>VEGF, CALM1, AKT1</td>
</tr>
<tr>
<td>Hypoxia</td>
<td>4.3E-02</td>
<td>VEGF, ARNT, AKT1</td>
</tr>
</tbody>
</table>
4.6 References


Chapter 5: Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer

A version of this chapter has been published as:


*These authors contributed equally.

Please see the published version of this chapter for all supplementary materials.
5.1 Introduction

Lung cancer is the leading cause of cancer related deaths worldwide (Parkin et al, 2005). The disease is classified into two major histological groups: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). Tobacco smoke is a major etiological factor, especially in SCLC. SCLC comprises approximately 20% of all lung cancers and exhibits a neuroendocrine phenotype while NSCLC lacks these features and makes up the remaining 80% of cases. SCLC exhibits a more aggressive phenotype that inevitably reoccurs after initial response to chemotherapy while the clinical outcome of NSCLC is often hard to determine (Kurup & Hanna, 2004; Stupp et al, 2004; Zakowski, 2003). Much of our current knowledge of these subtypes has been derived from a canonical set of cell lines derived from primary tumours (Phelps et al, 1996). These lines have been particularly crucial in the understanding of SCLC for which surgical resection is rarely performed (Rostad et al, 2004).

The variation in development and progression of SCLC and NSCLC may be a result of underlying differences in genetic alteration. Although histological classification can separate these two subtypes, previous studies using conventional genome scanning techniques such as loss of heterozygosity analysis and comparative genomic hybridization (CGH) have shown that differences and similarities in genetic aberration exist between SCLC and NSCLC. (Balsara & Testa, 2002; Girard et al, 2000). The limited resolutions of these methods have hampered the ability to identify discrete differences in genetic alterations, which are essential to understanding the biochemical deregulation that lead to the unique phenotypes of NSCLC and SCLC. Furthermore, the lack of a well defined progenitor cell type for SCLC has presented a major challenge in establishing specific gene expression levels (Coe et al, 2005).

Due to these limitations, it has become apparent that combining genomic and gene expression data will be essential for identifying new tumour suppressors and oncogenes (Henderson et al, 2005; Tonon et al, 2005). In addition, many genome wide platforms have proved useful in defining recurrent regions of alteration in lung cancer cells (Tonon et al, 2005; Zhao et al, 2005). With the development of whole genome tiling path array comparative genomic hybridization (aCGH), segmental copy number changes unique to each cell type can be defined at high resolution (Ishkanian et al, 2004). This technology allows the fine mapping of genomic alteration boundaries to within a single bacterial artificial chromosome (BAC) clone, identifying the precise genes potentially affected by a copy number alteration (CNA). Since alterations at
the DNA level are the initial events in cancer development, the gene expression changes that occur as a result of these alterations will be important in tumourigenesis.

To determine novel differences in CNA between the two lung cancer cell types, we profiled the genomes of 41 lung cancer cell lines (14 SCLC and 27 NSCLC) using the whole genome tiling path array for CGH analysis. The integration of expression data for these regions verified our findings and identified the gene expression changes associated with CNA. Furthermore, comparing expression and copy number levels between NSCLC and SCLC without the requirement for normal expression levels circumvented a significant hurdle in the analysis of SCLC. Additionally, difference-based analysis compensates for random cell culturing artefacts, allowing insight into the clinical disease. Grouping the differentially altered genes by biological function revealed cellular pathways that may drive the pathological development of these cell types. The discovery of these genes affected by phenotype specific CNA (PSCNA) may shed light on disease mechanisms and identify novel molecular targets for therapeutics and diagnostics.

5.2 Results and Discussion

5.2.1 Copy number analysis of lung cancer cell genomes

To facilitate the high resolution search for novel genetic alterations unique to each lung cancer cell type, we analyzed 14 SCLC and 27 NSCLC cell lines with the SMRT CGH array. This array allows the accurate assessment of segmental DNA copy number changes at 32,433 overlapping genomic loci in a single experiment, producing copy number maps at 100 kbp resolution across the entire sequenced human genome (Ishkanian et al., 2004). After co-hybridizing differentially labelled sample DNA and a male genomic DNA reference, fluorescence signal intensity ratios for each array element were determined and displayed as log2 plots using SeeGH software. Genetic alterations were identified in all cell lines analyzed. Figure 5.1 shows an example SeeGH karyogram for the SCLC cell line H1672. Upon visual analysis of this profile, areas of segmental gain and loss representing multiple levels of copy number change can be observed. For example, the telomeric end of chromosome arm 13q contains regions showing both single copy gain and high level amplification (Figure 5.1). In addition to the multiple segmental alterations affecting the majority of chromosomes in this sample, discrete micro-amplifications and deletions are also detected such as those highlighted on chromosome arms 18q and 2q respectively. These minute changes may have been missed by marker-based
techniques and highlight the resolution of the tiling path array. Array CGH karyograms for all the cell lines are available online at http://www.bccrc.ca/cg/ArrayCGH_Group.html.

5.2.2 Frequency analysis

Regions of chromosomal alteration key to the development of tumours will be present in multiple samples. By aligning the profiles of multiple genomes, patterns of gain and loss are revealed and minimal regions that potentially contain tumour suppressor genes and oncogenes can be identified. Thus, after generating the whole genome tiling path array CGH profiles of the lung cancer genomes, we then proceeded to identify recurrent regions of aberration within each cell type. To do this we employed a computer algorithm, aCGH-Smooth, to aid in the automated detection of regions of chromosomal gain and loss (Jong et al., 2004). The frequency of alteration of each genomic locus assayed was then calculated individually for the cell types and plotted using SeeGH Frequency Plot software as previously described (Coe et al., 2005). The data used to generate the frequency diagrams is present in Supplementary Material. The frequency plots and a detailed description of the recurrent regions of alteration specific to these SCLC and NSCLC cell lines have been reported (Coe et al., 2005; Garnis et al., 2005).

Genetic alterations unique to each cell type may contain genes responsible for the difference in disease development and clinical behaviour. To identify these regions, we overlaid the frequency plot diagrams of the SCLC and NSCLC samples and then compared the alteration frequencies in the two groups to determine regions that were statistically different by a 3x2 Fishers exact test and exclusion of regions which demonstrated increased gain and loss frequency for a single cell type (Figure 5.2). In this figure, areas indicated in green are more frequently altered in SCLC while those in red are more frequently altered in NSCLC. The yellow represents areas of overlap between the two frequency plots. Regions shaded in blue are those determined to be differentially altered in the cell types.

5.2.3 Regions of similarity

Among the regions that were not statistically different, there were some striking similarities (Figure 5.2). Consistent with previous reports, chromosome 3p loss was present in approximately 75% of both the NSCLC and SCLC samples (Balsara & Testa, 2002). This is consistent with previous results demonstrating that the deletion of putative tumour suppressor genes (TSGs), such as FHIT and RASSF1, contained on this chromosome arm are important genetic events in the development of lung cancers (Zabarovsky et al., 2002). Likewise, copy number loss of chromosome arm 4q was evident in ~50% of samples in each cell type mirroring
results observed using conventional CGH (Petersen et al, 1997a; Petersen et al, 1997b)(Figure 5.2).

The NSCLC and SCLC cell lines also showed similar frequency of copy number gain on chromosomes arm 5p as well as at chromosome bands 7p22.3 and 11q13.1-11q14.1. Over-representation of the entire 5p arm was a recurrent event in both cell types with the telomeric end of 5p15.33 showing the greatest amount of change. This region contains the Telomerase Reverse Transcriptase (hTERT) gene which has been implicated in cell immortalization in numerous cancers (Ramirez et al, 2004; Tomoda et al, 2002). Gain of the 11q13.1-11q14.1 region was present in >50% of the lung cancer cell lines with the highest degree of concordance at 11q13.3 (Figure 5.2). Cyclin D1, which is involved in the inactivation of the retinoblastoma protein and progression of the cell cycle through the G1-S phase, is located at this loci (Muller et al, 1994). This finding supports the theory that amplification of this gene is an important event in tumourigenesis (Fu et al, 2004). The gain of 7p22 was particularly interesting as it was the most common copy number aberration in both cell types. The minimal common alteration within this amplified area in the SCLC cell lines contains only one gene, MAD1L1 (validated by Coe et al.) (Coe et al, 2005). Although this is a checkpoint gene involved in growth inhibition, its gain has been reported in other cancers (de Leeuw et al, 2004; Jin et al, 1999; Tsukasaki et al, 2001). The high frequency of MAD1L1 amplification in the NSCLC samples as well suggests that this gene may play an essential role in the development of lung cancers (Garnis et al, 2005).

It is noteworthy that a subset of the genomic similarities between the SCLC and NSCLC cell lines could be resultant of adaptation to culturing conditions. Due to this, the greatest insight into the biology of the clinical disease will be attainable through analysis of differences (rather than their similarities) in genomic alterations and gene regulation.

5.2.4 Regions of difference

Through our analysis, numerous regions throughout the genome were determined to be differentially altered between the SCLC and NSCLC samples. This difference-based approach compensates for random cell culturing artefacts and should identify the regions most strongly linked to clinical disease. These regions ranged in size from whole chromosomes (chromosome 21) to discrete peaks, kilobases in size (3q27.1). Using our stringent, multi-step criteria (Fisher’s Exact test followed by additional thresholding), we detected several regions that differed strongly in their alteration status between the cell types, we refer to these as phenotype specific copy number alterations (PSCNAs). These included 1p36.33-1p34.2, 2p25.3-2p24.3,
3q26.33-3q28, 5q34-5q35.3, 6q24.2-6q27, 7p13-7p11.2, 8q21.2-8q22.3, 8q24.11-8q24.23, 9p22.3-9p21.1, 10q11.21-10q11.23, 12q24.31-12q24.33, 13q12.11-13q13.1, 13q32.2-13q34, 17q11.2, 18p11.23-18p11.21, 18q21.1-18q22.2, 19p13.2-19p12, and 21q11.2-21q22.3.

Some of these regions showed completely opposite patterns of alteration in the different cell types. 21q11.2-21q22.3 was a striking example as it is very frequently gained in SCLC but deleted in the NSCLC cases. Other regions were altered (gained or lost) in one cell type but remained almost unchanged in the other, for example the 8q21.2-8q22.3 locus that is commonly gained only in NSCLC. In addition, we observed chromosome segments altered in the same manner in both cell types, but to a greater extent in one over the other. 7p13-7p11.2 displays this characteristic as it is gained in ~50% of the SCLC cell lines and ~80% of the NSCLC samples.

The genes within these major regions of disparity may be responsible for the difference in disease development. However, not all genes contained in these regions will be differentially expressed as a consequence of the PSCNAs. To validate these CNAs and identify genes within these regions responsible for the different cell phenotypes, gene expression analyses were required.

5.2.5 Identification of genes differentially expressed between SCLC and NSCLC caused by phenotype specific copy number alteration

Validation of the genomic differences identified between SCLC and NSCLC cell lines was performed by assessment of their impact at the gene expression level. This is achieved by integrating Affymetrix expression profiling data with the array CGH data presented above. Due to the lack of a defined normal cell type for SCLC the definition of specific over and under expression of genes is difficult to establish. To circumvent this limitation we compared Affymetrix absolute expression values for both the NSCLC and SCLC samples to determine differential expression between the cell types.

Genes contained within the regions of peak genomic copy number difference were selected from the expression data and filtered to identify only those genes which exhibited expression differences between the two cell types presumably as a result of the copy number differences (Affymetrix gene expression data for the regions of genomic difference is available in Supplemental Material). A strict Mann-Whitney U test p value threshold of 0.001 as well as a requirement for expression differences to match the direction of copy number difference (i.e.
increased expression in samples with a higher frequency of copy number gain and reduced expression in cells with a high frequency of copy number loss). This analysis identified 243 of 5185 analyzed Affymetrix probe sets, corresponding to 159 unique RefSeq genes, as being differentially regulated between SCLC and NSCLC (Figure 5.3) (Also presented in Supplementary Material). The nature of our approach filters out genes with differential expression due to factors other than copy number such as methylation and the mutation and up/down regulation of upstream genes. As such, these 159 genes most likely represent the expression differences resulting from SCLC and NSCLC PSCNAs. This hypothesis is supported by principal components analysis, which demonstrated the strong contribution of the 159 genes to the differential phenotypes of SCLC and NSCLC (Figure 5.4).

Analysis of the 159 genes not only revealed several expected findings such as an increased level of EGFR expression in NSCLC, but identified novel differentially expressed genes such as MRP5 (Amann et al, 2005; Ritter et al, 2005) which exhibited increased expression in SCLC. This gene encodes an ABC transporter known to clear various chemotherapeutics from the cytoplasm and increased expression in lung cancer has been associated with exposure to platinum drugs (Oguri et al, 2000). Furthermore, another study has correlated MRP5 expression to cisplatin chemoresistant lung cancer cell lines (Weaver et al, 2005). This result suggests a possible mechanism of enhanced chemotherapeutic resistance for the SCLC cells.

5.2.6 Biological pathways differentially altered in SCLC and NSCLC

Further analysis of the differentially expressed genes revealed a strikingly high number of genes, are present in a small set of interconnected pathways. The presence of multiple genes affected by PSCNA in the MAPK and EGFR pathways lead us to examine the known interactors for each of these genes to elucidate a biochemical differentiation between SCLC and NSCLC cells. The results of this analysis are displayed in Figure 5.5. Twenty-two of the genes differentially altered between SCLC and NSCLC are components of the cell cycle, EGFR, MAPK, p38MAPK, and WNT pathways (Table 5.1). Four genes (E2F2, SOX11, MAP3K4, and HSPH1) which represent critical nodes in these pathways were further examined by Real-Time PCR validating differential expression between SCLC and NSCLC. Pathway information was derived from the Signal Transduction Knowledge Environment (stke.sciencemag.org), the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg), and the following references: (Bracken et al, 2003; Campos et al, 2004; Einarson et al, 2004; Hyodo-Miura et al, 2002; Ishitani et al, 1999; Li & Guan, 2004; Lundberg & Weinberg, 1999; Polager & Ginsberg, 2003; Rubin & Atweh, 2004; Sakamuro & Prendergast, 1999; Sasahira et al, 2005; Schneider et
al, 2002; Shaulian & Karin, 2001; Taguchi et al, 2000; Wada & Penninger, 2004; Williams et al, 2003; Wu et al, 2003; Yamagishi et al, 2002; Zebedee & Hara, 2001). Of particular interest was a strong increase in the expression of WNT inhibitors in SCLC cells, namely NLK, SOX11, and TCF4. This remarkable result demonstrates that the WNT pathway may not be a significant player in SCLC.

Additionally we detected a strong difference in the regulatory components of the p38MAPK pathway with the reduced expression of two p38 MAPK activating genes in NSCLC (HMGB1, HSPH1) and contrasting over-expression of two p38 MAPK activating genes in SCLC (MAP3K4, DSCAM). We also observed strong PSCNA-related over-expression of several members of the MAPK and cell cycle pathways in both cell types, albeit through different components. In the NSCLC samples, we observed segmental loss and down regulation of the cell cycle inhibitor CDKN2A as well as copy number gain and up regulation of MAPK9 and EGFR when compared to SCLC. In contrast, the SCLC cells demonstrate comparatively higher expression of many pro-proliferative genes; these are detailed in Figure 5.5. Interestingly, several genes with cell cycle inhibitory functions exhibited PSCNA-induced over-expression in SCLC. Due to likely antagonism of these genes by the many up-regulated cell cycle-activating genes, it is possible that they perform a novel role secondary to their primary functions in cell cycle regulation. These differential patterns of oncogenic disruption to cell cycle pathways highlight the need to examine cell type specific targets for therapeutic pathway intervention. For example, although a recent study has shown EGFR is expressed at low levels in SCLC, (Tanno et al, 2004) our results indicate that the pathway is being activated by over-expression of multiple downstream components, potentially bypassing benefits that may be derived from EGFR targeted therapy.

5.3 Conclusions

Whole genome array CGH in conjunction with global expression profiling analysis has allowed the identification of genes deregulated as a result of PSCNA between SCLC and NSCLC cells. The 159 genes revealed as having strongly divergent expression patterns as a result of copy number alterations identified a remarkable pattern of gene deregulation in several key biological pathways. Cell cycle up-regulation in SCLC and NSCLC occurs through drastically different targets, suggesting a need for differential therapeutic target selection. Additionally the WNT pathway, which has recently received much attention for its involvement in NSCLC, appears to be strongly down regulated in SCLC through PSCNA-induced over expression of inhibitory genes. This work represents the first comprehensive search for the causative genetic
alterations distinguishing SCLC and NSCLC by integrating whole genome expression and copy number analysis platforms.

5.4 Methods and Materials

5.4.1 DNA samples

The 41 lung cancer cell lines described were established at the National Cancer Institute (NCI-H series) and at the Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center (HCC series) except for SW-900 and SK-MES-1 (Fogh et al., 1977; Phelps et al., 1996). These cell lines have been deposited for distribution in the American Type Culture Collection (http://www.atcc.org). DNA was extracted from 27 NSCLC: 18 adenocarcinomas (H1395, H1648, H1819, H1993, H2009, H2087, H2122, H2347, HCC78, HCC193, HCC366, HCC461, HCC1195, HCC1833, HCC3255, HCC4006, HCC827 and HCC2279) and nine squamous cell carcinomas (H157, HCC15, HCC2450, HCC95, H520, H226, SW 900, SK-MES-1 and H2170), and 14 SCLC cell lines: nine classical (H187, H378, H889, H1607, H1672, H2107, H2141, H2171, and HCC33) and five variant (H82, H289, H524, H526, and H841). The identity of all 41 cell lines were verified by fingerprinting using the Powerplex 1.2 system (Promega) which contains nine polymorphic markers.

5.4.2 Tiling path array CGH

Segmental copy number status of the 41 lung cancer cell genomes were deduced in array CGH experiments using Sub-Megabase Resolution Tiling-set (SMRT) arrays. These arrays contain 97,299 elements representing 32,433 BAC derived amplified fragment pools spotted in triplicate on two aldehyde-coated glass slides (Ishkanian et al., 2004; Watson et al., 2004). Array hybridization was performed as previously described (Coe et al., 2005; Garnis et al., 2005). Briefly, 200-400 ng of sample and a common reference male genomic DNA (Novagen, Mississauga ON) were separately labelled by random priming in the presence of cyanine-5 dCTP or cyanine-3 dCTP (PerkinElmer, Woodbridge ON), respectively. Labelled sample and reference DNA probes were combined and purified using ProbeQuant Sephadex G-50 Columns (Amersham, Baie d'Urfe, PQ). The probe mixture was precipitated in a solution containing 100 µg Cot-1 DNA (Invitrogen) with 0.1X volume 3 M sodium acetate and 2.5X volume 100% ethanol. The DNA pellet was resuspended in 45 µl of hybridization solution containing 80% DIG Easy hybridization buffer (Roche, Laval, PQ), 100 µg sheared herring sperm DNA (Sigma-Aldrich), and 50 µg yeast tRNA (Calbiochem) and denatured at 85°C for 10 minutes. Repetitive sequences were blocked at 45°C for 1 hour prior to hybridization. Probes were then added to
array slides and placed in a pre-warmed hybridization chamber (Telechem, Sunnyvale, CA). After hybridization for ~40 hours at 45°C, arrays were washed five times for five minutes each in 0.1X SSC, 0.1% SDS at room temperature in the dark with agitation followed by five rinses in 0.1X SSC and dried by centrifugation.

5.4.3 Imaging and data analysis

Images of the hybridized arrays were captured through cyanine-3 and cyanine-5 channels using a charge-coupled device (CCD) scanner system (Applied Precision, Issaquah, WA). Images were then analyzed using SoftWoRx Tracker analysis software (Applied Precision). Spot signal ratio information was mapped to genomic coordinates and median normalized. Custom software called SeeGH was used to combine replicates and visualize all data as log₂ ratio plots in SeeGH karyograms and exclude replicate data points which exceeded a standard deviation of 0.075 (Chi et al., 2004). In addition, genomic imbalances were identified using aCGH-Smooth which uses a genetic local search algorithm to identify breakpoints defining segmental DNA copy number changes by using a maximum likelihood estimation to optimize breakpoint location (Jong et al., 2004). As previously described, the Lambda and breakpoint per chromosome settings were set to 6.75 and 100, respectively (de Leeuw et al., 2004; Jong et al., 2004). The frequency of alteration for each BAC was then individually determined for each lung cancer cell type as described previously and plotted in SeeGH Frequency Plot to visualize areas of recurrent deletion and amplification (Coe et al., 2005). SeeGH software packages are available upon request at: http://www.flintbox.ca/.

5.4.4 Statistical analysis of array CGH alteration frequencies

Regions of differential copy number alteration between SCLC and NSCLC genomes were identified using a stringent multi-step filtering process. The occurrence of copy number gain, loss, and retention at each locus was compared between SCLC and NSCLC data sets using Fishers exact test. Testing was performed using the R statistical computing environment on a 3x2 contingency table with a p value threshold of 0.05. Loci for which the same cell type exhibited an increased frequency of both gain and loss when compared to the other were then excluded from these results in order to compensate for regions demonstrating higher levels of genomic instability but not true differential patterns of alteration. Finally, regions which passed the first two criteria and demonstrated alteration frequencies differing by at least 20% occurrence in either copy number loss or gain were selected for further analysis.
5.4.5 Affymetrix gene expression analysis

Affymetrix HG-U133A and HG-U133B hybridizations were performed as described in Henderson et al (Henderson et al, 2005). RNA expression profiles were generated for 14 SCLC and 22 NSCLC cell lines, all of which are present in the array CGH data set (H187, H378, H889, H1607, H1672, H2107, H2141, H2171, H82, H289, H524, H526, H841, H1395, H157, H1648, H1819, H1993, H2009, H2087, H2122, H2347, H3255, HCC1195, HCC15, HCC1833, HCC193, HCC2279, HCC2450, HCC366, HCC4006, HCC461, HCC78, HCC827, HCC95). Absolute expression values were log transformed and scaled to a score between 0 and 100 using MAS 5.0 (Affymetrix, Santa Clara, CA), and only probe sets demonstrating a present or marginal quality score in at least 50% of samples were considered for further analysis. Gene expression data for SCLC and NSCLC were then compared using the Mann-Whitney U test to identify genes which differed in expression between the two cell types with a p value of at least 0.001. The resulting gene list was then filtered to select only those genes for which the expression change matched the direction predicted by the copy number analysis.

5.4.6 Real time PCR

Real-time PCR validation of expression differences between NSCLC and SCLC was performed on key genes identified through combination of array CGH and Affymetrix gene expression profiling. Five micrograms of total RNA from each cell line profiled by Affymetrix microarrays was converted to cDNA using an ABI High Capacity cDNA Archive Kit (Applied Biosystems, Foster City, CA). 100 ng of cDNA was used for each real time PCR reaction. TaqMan (Applied Biosystems, Foster City, CA) gene expression assays: E2F2 (Hs00231667_m1), SOX11 (Hs00846583_s1), MAP3K4 (Hs00245958_m1), HSPH1 (Hs00198379_m1), B-actin (Hs99999903_m1), 18S rRNA (Hs99999901_s1) were performed using standard TaqMan reagents and protocols on a Biorad I-cycler (Biorad, Hercules, CA). The ΔΔCt method was used for expression quantification using the average of the cycle thresholds for B-actin and 18s RNA to normalize gene expression levels between samples. Expression levels were compared between NSCLC and SCLC by a Mann-Whiney U test as performed for the Affymetrix microarray data.

5.4.7 Principal components analysis

The 243 Affymetrix probe sets deregulated as a result of copy number differences between SCLC and NSCLC were subjected to Principal Component Analysis. Analysis of the samples
was performed using the Statistics Toolbox (Version 5.1) of MATLAB (Version 7.1) (The MathWorks Inc., Natick, MA).
Figure 5.1. SMRT Array Profile of the SCLC NCI-H1672 Cells. Data is presented as a SeeGH karyogram to demonstrate the resolving power of the SMRT technology. Each BAC clone is displayed as a vertical line representing its genomic coverage. The horizontal shift of each line to the left or right of 0 represents the measured Log2 signal ratio from a competitive hybridization with male genomic DNA. A decreased ratio represents a loss of copy number compared to the reference sample while an increased ratio represents an increase in copy number. Multiple levels of segmental copy number alteration as well as microalterations were readily detected (representative examples are highlighted in red and green). SeeGH karyograms for all cell lines analyzed are available at: http://www.bccrc.ca/cg/ArrayCGH_Group.html.
Figure 5.2. Copy Number Alterations in SCLC and NSCLC. Alteration frequencies for SCLC (green) and NSCLC (red) are displayed as bar plots adjacent to chromosomal ideograms. Bars extending to the right of each chromosome represent the frequency of copy number gain; conversely, bars extending to the left represent the frequency of copy number loss. Yellow regions represent overlapping portions of the SCLC and NSCLC alteration frequencies. Blue bars indicate regions demonstrating significantly different alteration frequencies. Vertical brown lines on the left of each frequency diagram indicate regions selected for further analysis.
Figure 5.3. Differential expression as a result of copy number alteration. Affymetrix log transformed absolute expression data for the 243 probe sets exhibiting strong differential expression between SCLC and NSCLC associated with copy number differences are displayed. High level expression is indicated by white/yellow while blue/black indicates progressively lower levels of expression. The SCLC samples are indicated by green highlighting of each column, while NSCLC samples are indicated by red highlighting. Each probe set is sorted according to its chromosomal position and cell lines are sorted alphabetically, according to their cell type. Probe set with annotated gene IDs are labelled with their RefSeq name while probe sets with less reliable mapping are indicated by their probe ID alone. Average expression values were calculated for genes with multiple Affymetrix probe sets, which passed the filtering conditions. These are indicated in blue text (The number of probe sets averaged is indicated in brackets). The primary genomic alterations observed for both SCLC and NSCLC are indicated to the right of each set of expression values (G= “gain”, L= “loss”, no value = “gained and lost” or “no change”).
Figure 5.4. Contribution of Copy Number Induced Gene Expression Differences to the SCLC and NSCLC Phenotypes. Principal components analysis was performed utilizing all 243 Affymetrix probe sets demonstrating expression differences as a result of copy number alterations. The SCLC samples are indicated by solid circles, while the NSCLC samples are indicated by open circles. Strong separation of the SCLC and NSCLC cell lines along principal component 1 demonstrates the contribution of these genes to the differential phenotypes.
Figure 5.5. Differential Targets of Copy Number Induced Expression Changes in Key Biochemical Pathways between SCLC and NSCLC. Strong PSCNA induced expression differences were identified between SCLC and NSCLC in several key pro-proliferate pathways. Genes with increased expression in SCLC when compared to NSCLC are indicated in green, while genes with increased expression in NSCLC are indicated in red. Genes exhibiting a tumour suppressor like pattern of reduced expression as a result of frequent copy number loss in NSCLC are indicated in yellow. Genes added to the pathways for context but for which no expression differences were detected, are indicated in grey. Critical pathway nodes validated by real-time PCR are indicated with a *.
### Table 5.1. Differential deregulation of genes in key biochemical pathways between NSCLC and SCLC

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>Locus</th>
<th>Regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>STMN1</td>
<td>stathmin 1</td>
<td>1p36.11</td>
<td>SCLC +</td>
</tr>
<tr>
<td>E2F2</td>
<td>E2F Transcription Factor 2</td>
<td>1p36.12</td>
<td>SCLC +</td>
</tr>
<tr>
<td>ZNF151 (MIZ1)</td>
<td>Zinc finger protein 151 (Myc-interacting zinc finger protein)</td>
<td>1p36.13</td>
<td>SCLC +</td>
</tr>
<tr>
<td>PRDM2 (RIZ1)</td>
<td>PR Domain-Containing Protein 3 (Rb Protein-Binding Zinc Finger Protein)</td>
<td>1p36.21</td>
<td>SCLC +</td>
</tr>
<tr>
<td>ID2</td>
<td>Inhibitor of DNA binding 2</td>
<td>2p25.1</td>
<td>SCLC +</td>
</tr>
<tr>
<td>SOX11</td>
<td>SRY-Related HMG-Box Gene 11</td>
<td>2p25.2</td>
<td>SCLC +</td>
</tr>
<tr>
<td>MAPK9 (JNK2)</td>
<td>Mitogen-activated protein kinase 9 (C-JUN Kinase 2)</td>
<td>5q35.3</td>
<td>NSCLC +</td>
</tr>
<tr>
<td>MAP3K4</td>
<td>Mitogen-activated protein kinase kinase kinase 4</td>
<td>6q26</td>
<td>SCLC +</td>
</tr>
<tr>
<td>EGFR</td>
<td>Epidermal Growth Factor Receptor</td>
<td>7p11.2</td>
<td>NSCLC +</td>
</tr>
<tr>
<td>CDKN2A (p16INK4A)</td>
<td>cyclin-dependent kinase inhibitor 2A</td>
<td>9p21.3</td>
<td>NSCLC -</td>
</tr>
<tr>
<td>KNTC1</td>
<td>Kinetochore-associated protein 1</td>
<td>12q24.31</td>
<td>SCLC +</td>
</tr>
<tr>
<td>HMGB1</td>
<td>High Mobility Group Box 1 (Amphoterin)</td>
<td>13q12.3</td>
<td>NSCLC -</td>
</tr>
<tr>
<td>HSPH1</td>
<td>Heat Shock 105kD</td>
<td>13q12.3</td>
<td>NSCLC -</td>
</tr>
<tr>
<td>ING1 (p33ING1)</td>
<td>Inhibitor of growth family member 1</td>
<td>13q34</td>
<td>SCLC +</td>
</tr>
<tr>
<td>JAZ1 (SUZ12)</td>
<td>Joined to JAZF1 (Suppressor of ZESTE 12)</td>
<td>17q11.2</td>
<td>SCLC +</td>
</tr>
<tr>
<td>NLK</td>
<td>Nemo-like kinase</td>
<td>17q11.2</td>
<td>SCLC +</td>
</tr>
<tr>
<td>SMAD4</td>
<td>Mothers against decapentaplegic homolog 4</td>
<td>18q21.1</td>
<td>SCLC +</td>
</tr>
<tr>
<td>CCDC5</td>
<td>Coiled-coil domain containing 5</td>
<td>18q21.1</td>
<td>SCLC +</td>
</tr>
<tr>
<td>TCF4</td>
<td>Transcription Factor 4</td>
<td>18q21.2</td>
<td>SCLC +</td>
</tr>
<tr>
<td>JUNB</td>
<td>oncogene jun-B</td>
<td>19p13.13</td>
<td>SCLC +</td>
</tr>
<tr>
<td>TIAM1</td>
<td>T-cell lymphoma invasion and metastasis 1</td>
<td>21q22.11</td>
<td>SCLC +</td>
</tr>
<tr>
<td>DSCAM</td>
<td>Down syndrome cell adhesion molecule</td>
<td>21q22.2</td>
<td>SCLC +</td>
</tr>
</tbody>
</table>

1. Regulation SCLC = Small Cell Lung Cancer; NSCLC = Non-small cell lung cancer; + = Increased expression in the indicated cell type; - = Decreased expression in the indicated cell type
5.5 References


Chapter 6: Genetic pathways involved in the development of non-Small cell lung cancer subtypes

A version of this chapter will be submitted with the following author list:


Supplementary tables for this chapter have been omitted due to size limitations and can be provided upon request.
6.1 Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide and despite current treatments, prognosis remains poor with a five year survival rate of <15% (Parkin et al, 2005; Sato et al, 2007). Squamous cell carcinoma (SqCC) and adenocarcinoma (AC) are the predominant lung cancer subtypes and are traditionally regarded as a single disease entity in terms of systemic therapy (Sato et al, 2007; Travis, 2002). However, these subtypes display distinct phenotypic characteristics probably related to differences in cell derivation, genetic alterations and pathogenetic pathways involved in their development (Giangreco et al, 2007). These fundamental discrepancies in tumor biology may play a primary factor in determining the poor outcomes of lung cancer patients as biological differences that segregate with each subtype may also lead to variations in response to therapies (Broet et al, 2009; Garraway & Sellers, 2006). Thus, distinguishing the key molecular mechanisms responsible for the development of each lung cancer subtype will be essential in order to define appropriate avenues for therapeutic intervention.

The specific genes and cellular pathways responsible for the different phenotypes of SqCC and AC remain largely unknown. Initial gene expression profiling studies have yielded some insight into the tumor subtypes and are able to segregate tumors into histologic groupings based on multi-gene models (Bhattacharjee et al, 2001; Thomas et al, 2006). However, since not all gene expression changes are causal to disease development, it is challenging to distinguish critical events from reactive changes through global gene expression profiles alone (Coe et al, 2008). Gene expression changes corresponding with alteration at the DNA level is often regarded as evidence of causality. For example, changes in DNA copy number are key to the progression and development of many cancers and lead to the deregulation of genes responsible for carcinogenesis (Hyman et al, 2002; Pollack et al, 2002). Hence, examining genetic events in conjunction with the changes in gene expression pattern should improve the identification of causal changes that lead to disease phenotype.

Although genetic disparities have been described, most genome profiling studies performed to date have suffered from low resolution and small sample sizes, limiting the ability to identify specific disruptions unique to each subtype (Luk et al, 2001; Pei et al, 2001; Petersen et al, 1997; Sy et al, 2004; Tonon et al, 2005). Recent advancements in microarray technologies have substantially increased our ability to understand the genomic mechanisms influencing tumorigenesis (Lockwood et al, 2006). In this study, we performed a large-scale integrative
analysis of 271 NSCLC primary tumors (179 AC and 92 SqCC) using high resolution array comparative genomic hybridization (CGH) coupled with gene expression microarray analysis. Our objective was to directly compare the alteration patterns between the tumor types in order to comprehensively identify the causal genetic alterations and deregulated genes responsible for their differential development and clinical characteristics.

6.2 Results

6.2.1 Identification of genomic differences between AC and SqCC

If specific genetic pathways are involved in the development of SqCC and AC, we would expect to find differences in their genome profiles. In order to determine if specific genetic alterations unique to each NSCLC subtype exist, we aimed to identify recurrent, non-random regions of aberration in each group. To do this, we generated and compared whole genome copy number profiles for 271 NSCLC tumors – 179 AC and 92 SqCC (sample set 1a, Supplemental Table 1) – by tiling resolution array comparative genomic hybridization (array CGH) (Ishkanian et al., 2004). After hybridization experiments, genomic profiles were normalized to remove systematic experimental biases and subjected to a smoothing algorithm in order to computationally define regions of copy number gain and loss (Coe et al., 2006b). Individual samples were then grouped by their corresponding subtype and probes were aggregated into regions based on similar copy number status. The resulting frequency of alteration across all autosomes was determined and compared between subtypes using the Fisher’s exact test to identify regions of copy number disparity and the resulting \( p \)-values were corrected for multiple comparisons with a cut-off of \( \leq 0.01 \) considered significant. In addition, regions had to be altered in >20% of samples in a group and the difference between groups >10% to be considered. Figure 6.1 displays the frequency of both gain and loss across the entire genome for AC and SqCC and highlights the corresponding regions of difference identified.

This analysis revealed 259 regions of significant copy number disparity between SqCC and AC, supporting our hypothesis that they develop through different genetic pathways (Figure 6.1c). Of the regions, 167 were SqCC specific in their alteration pattern whereas 86 where specific to AC. Although some of these regions overlapped between the two subtypes, their method of alteration (i.e gain/loss) was specific to an individual group. Since these regions differed strongly in their alteration status between the subtypes, we refer to these as phenotype-specific copy number alterations (PSCNAs). In total, the PSCNA covered approximately 600 Mbp of the genome, mapping to 34 of 39 autosomal chromosome arms, and ranged in size from large
segments of chromosome arms (95.1 Mbp on 4q) to discrete peaks, kilobases in size (0.05 Mbp in multiple places). The specific base pair boundaries of all PSCNAs are present in Supplemental Tables 2 and 3.

6.2.2 Integrative analysis reveals genes targeted by phenotype specific genetic alterations in AC and SqCC

The discovery of regions of copy number disparity between the subtypes suggested that the genes within these might be responsible for the differential development and pathological characteristics of the subtypes. To validate the changes and identify the specific target genes of these alterations, integration of copy number and gene expression analyses were performed. For this purpose, gene expression profiles were generated for a subset of 20 SqCC and 29 AC tumors that were analyzed by array CGH and had sufficient remaining material (sample set 1b, Supplemental Table 1). To our knowledge, this represents the largest integration of genome wide copy number and expression data in clinical lung cancer do date.

We hypothesized that subtype specific genes targeted by the PSCNAs would display three distinct expression characteristics. First, expression should be significantly different between AC and SqCC, reflecting the changes at the DNA level (criterion 1). Second, the expression pattern for a target gene should match the direction predicted by its corresponding PSCNA (criterion 2). For example, if a gene is gained in SqCC it should be expressed at higher levels in SqCC compared to AC. Lastly, if a gene is involved in tumorigenesis, it should be deregulated (over/underexpressed) only in cancerous and not normal tissue, with the direction again mirroring that of the copy number change (criterion 3).

To investigate the first criterion, the specific genes located within each PSCNA were determined and the expression levels compared between the SqCC and AC samples to determine those that were differentially expressed (P<0.01, corrected for multiple comparisons). For SqCC, 5328 unique genes mapped to the PSCNAs representing an average of ~32 genes per region and 2455 (46%) of these were differentially expressed. In AC, 1805 unique genes were located in the PSCNAs (~21 per region) and 343 (19%) were differentially expressed between the subtypes. These genes were then filtered for those that matched the expression direction predicted from copy number status in order to determine those meeting criterion 2. Using these strict statistical criteria, 737 genes (14% of total) were uncovered as candidate genes deregulated as a direct result of PSCNAs in SqCC while 223 (12% of total) were identified for AC. Although some genes overlapped, their disruption status was specific to the individual
cancer subtype and therefore referred to as subtype specific targets. When combined, the SqCC and AC PSCNA regulated candidates represented 776 unique genes and showed a clear distinction in expression levels between the two subtypes.

In addition to demonstrating a linkage between expression and copy number alteration, a candidate PSCNA regulated gene should only be deregulated in cancerous, and not normal, tissues (Croce, 2008). Therefore, to identify genes meeting the 3rd criterion, we analyzed the expression levels of the candidates in an independent panel of 53 SqCC and 58 AC lung tumors and 67 samples of exfoliated bronchial cells from cancer-free individuals generated using the Affymetrix U133 Plus 2 platform (sample sets 2 and 3, Supplemental Table 1). In total, 648 of the 737 SqCC specific and 206 of the 223 AC specific genes had corresponding probes on this array platform. These genes were compared between the respective cancer subtype and the normal bronchial cells in order to determine those that were significantly differentially expressed (P<0.01) in the direction predicted by the corresponding PSCNA in which they were located. This analysis revealed that 378 (51%) of the SqCC specific and 76 (37%) of the AC specific genes were deregulated in cancerous tissues (442 unique). Since these genes met all three criteria of PSCNA regulated targets described above, we determined that they are causal gene expression changes driving the development of each subtype (Figure 6.2).

6.2.3 Genes deregulated by PSCNAs contribute to AC and SqCC phenotypes

Next, we aimed to confirm that these genes were responsible for the different NSCLC phenotypes. 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 442 unique PSCNA-linked genes clearly delineated distinct subtype specific clusters (Figure 6.3a). A receiver operating characteristic (ROC) area under the curve (AUC) value of 0.98 confirmed that principle component 1 was strong discriminator of the subtypes. This was not surprising as the genes were uncovered as being different between the subtypes using this same set of samples. Therefore, to further confirm the role of the genes in subtype development, we applied the same analysis to two independent sample sets generated on different expression array platforms. The first consisted of 111 (58 AC and 53 SqCC, described above) and the second of 99 (48 AC and 51 SqCC) clinical lung tumors (sample set 4, Supplemental Table 1). Strikingly, this analysis was also able to separate the AC and SqCC samples with a great deal of accuracy (ROC AUC values of 0.90 and 0.98, respectively) (Figure 6.3b and 6.3c). The validation in
these large panels of completely independent NSCLC tumors from separate institutions provides further evidence that the genes regulated by PSCNAs are responsible for driving the differential development of AC and SqCC. Furthermore, our results highlight the impact of this novel integrative genome and transcriptome analysis in identifying robust target genes regulated by copy number changes.

6.2.4 Different gene networks are associated with the development of AC and SqCC

Cellular pathways and processes specifically disrupted in individual subtypes may reveal key oncogenic mechanisms driving the differential development of AC and SqCC. Thus, after identifying and validating the genes responsible for the differences between the subtypes, we next wanted to investigate their biological functions. To discover subtype-related networks of biologically related genes we performed Ingenuity Pathway Analysis (IPA) of the 76 AC and 378 SqCC specific target genes. This analysis revealed two main gene networks with overlapping biological functions for each individual subtype (Figure 6.4, Table 6.1). SqCC exhibited disruptions in gene networks that both function in regulating DNA replication, recombination and repair with additional roles in cell cycle, cellular organization and cell death. Genes involved in SqCC network 1 were mainly associated with the binding and modification of histone proteins H3 and H4 that regulate chromosome stability, cell division and gene transcription (Figure 6.4a). SqCC network 2 was focused on genes that control DNA replication and cell cycle progression (cyclins, RB, and E2F) and influence cell death (ERK, caspase) (Figure 6.4b). Meanwhile, the primary networks in AC displayed overlapping functions associated with controlling cell cycle and cancer development. The main AC specific gene network was composed primarily of genes regulated by the transcription factor HNF4α whereas AC network 2 contained numerous genes controlled by TGFβ and TP53 (Figures 6.4c and 6.4d, respectively). The independent nature of the gene networks implicated in AC and SqCC was further suggestive of distinct methods of tumorigenesis for the subtypes.

6.2.5 Subtype specific genes are associated with distinct clinical characteristics in AC and SqCC

Lastly, we aimed to determine the influence of PSCNA targeted genes on the clinical characteristics of AC and SqCC. Since these genes are disrupted during tumor development in a specific subtype alone, we reasoned that their expression should only correlate with specific clinical features in the corresponding subtype and not the other subtype or NSCLC (AC + SqCC) in general. To test this, we determined the survival associations using Kaplan-Meier
analysis for each subtype specific gene in AC, SqCC and NSCLC in two independent expression datasets (sample sets 2 and 4, Supplemental Table 1). Genes with expression that correlated with survival in either dataset for the subtype that they were disrupted were then identified. Overall, this analysis revealed 10 AC and 17 SqCC specific genes that had significant (P<0.05) associations with survival (Table 6.2). The associations were relatively specific to an individual subtype as no AC genes were correlated with survival in SqCC and only two SqCC genes (SLC1A5 and TP73L) in AC. However, this overlap could be explained in the case of SLC1A5. Interestingly, this gene showed an opposite pattern in terms of survival with low expression associated with poor survival in AC and high expression with poor survival in SqCC. Thus, although survival related in both subtypes, SLC1A5 is still specific in its association pattern. The only gene significantly correlated with survival in both subtypes that did not show specificity was the SqCC activated gene TP73L as high expression was indicative of poor survival in AC as well. This was also one of the two subtype specific genes - the other being the AC specific gene ITIH4 - that were associated with survival in NSCLC as a whole. Therefore, in total, all but two of the 27 (>90%) subtype specific genes that had expression levels associated with survival were restricted in their associations to an individual subtype.

6.3 Discussion

Previous studies suggest that distinct patterns of genomic alteration exist for AC and SqCC; however, the specific genes responsible for the different tumor phenotypes remain largely unknown. In this study, we provide the first comprehensive investigation for the causative genetic alterations distinguishing AC and SqCC by integrating whole-genome expression and copy number data. Our analysis revealed that the NSCLC subtypes are drastically different at the genomic level and allowed for the identification of genes targeted by subtype specific genetic alterations during cancer development. In addition, we discovered distinct gene networks associated with each subtype, signifying unique pathways to tumorigenesis.

The 259 PSCNAs detected in this study gave a general picture of the genetic pathways involved in the development of AC and SqCC. SqCC develops mainly though the acquisition of major gains on chromosome arms 2p, 3q, 8p, 12p, 19p, 20p, and 22q and losses on 2q, 4p, 4q, 5q, and 11q whereas gains on 1q and losses on 3q, 6q, 8p, 12p, 15q, 17p, 19p, 19q, and 22q are important in AC (Figure 6.1, Supplemental Tables 2 and 3). The majority of these regions were more frequently altered (gained or lost) in either SqCC or AC specifically, suggesting that they are important in the development of the individual subtype only. Interestingly however,
numerous regions showed a completely opposite pattern of alteration in the different subtypes with one having frequent gain and the other frequent loss or vice versa. For example, a discrete PSCNA spanning 1.91 Mbp on chromosome bands 8p12-p11.23 is commonly gained in SqCC while lost in AC, implying that the genes in these regions may play completely different roles during the development of the individual NSCLC subtypes, acting as tumor suppressor genes in AC and an oncogenes in SqCC.

The PSCNAs are consistent with those indentified by previous conventional and array CGH studies which compared the two subtypes (Garnis et al, 2006; Luk et al, 2001; Pei et al, 2001; Petersen et al, 1997; Sy et al, 2004; Tonon et al, 2005; Yakut et al, 2006); however, numerous additional regions were also found. Commonly, the gain of 1q has been reported to be characteristic of AC whereas gain of 3q and 12p along with loss of 4q are specific to SqCC. All of these regions were detected in our study, confirming our results. However, due to the increased resolution of our array platform, the boundaries of these previously known regions were fine mapped beyond chromosome bands to the kilobase level. For example, chromosome 12p gain in SqCC was refined to five discrete regions of 2.46, 2.52, 0.69, 2.41 and 0.11 Mbps in size. Interestingly, the only subtype specific region that overlaps between all previous studies is the gain of 3q in SqCC. In fact, a recent high resolution array CGH study determined the subtypes have almost completely overlapping genomic profiles with SqCC 3q gain as the only subtype specific alteration (Tonon et al, 2005). However, it is likely that the small sample size of this (18 AC and 26 SqCC tumors) and other previous studies only allowed extremely large differences in alteration frequency between the subtypes to be detected. Our greater than four-fold increase in sample size allows lower frequency yet significant PSCNAs to be discovered, explaining the increase in subtype specific differences seen in our study compared to previous reports. Thus, in addition to fine-mapping known regions, our discovery of these novel PSCNAs suggest that AC and SqCC are more different at the molecular level than previously believed and provides further evidence that they develop through distinct genetic pathways.

The identification of the genes deregulated as a result of PSCNA between AC and SqCC was a major finding which was facilitated by the integration of the copy number with expression data. We discovered 378 genes which had disruptions specific to SqCC and 76 with disruptions specific to AC. Importantly, these genes were able to accurately delineate the disease subtypes in two completely independent datasets, confirming their contribution to the different tumor phenotypes. Investigation of the specific genes altered in AC and SqCC uncovered many important insights. For instance, the identification of genes from multiple chromosomes further
confirms that changes other than 3q gain in SqCC underlie the differential development of the subtypes. This was particularly evident as many prominent regions of genetic disparity, such as those on chromosome 8p, 12p and 19p, contained numerous genes specifically deregulated in each subtype, suggesting these alterations play a key role in the divergent phenotypes.

Another important finding was that numerous genes which have previously been implicated in NSCLC tumorigenesis, prognosis and response to chemotherapy are preferentially disrupted in a specific subtype (Table 6.3 for a subset of these genes). For example, previously implicated oncogenes such as NOTCH3 and FOXM1 are overexpressed through increased gene dosage specifically in SqCC while the putative tumor suppressors DUOX1 and PRDM2 are deleted and underexpressed specifically in AC and SqCC, respectively (Dang et al, 2000; Koon et al, 2007; Luxen et al, 2008; Wang et al, 2008). Our data suggests for the first time that these genes may be involved in tumorigenesis exclusively in a particular subtype. This information will become particularly important when designing targeted therapeutic strategies based around these genes. The development of MEK inhibitors highlights this point (Sun et al, 2007). Since activated MEK1 and MEK2 phosphorylate and activate ERK (MAPK1), the deregulation of MAPK1 in a subset of ACs may be an important consideration in determining the efficacy of this treatment (Gollob et al, 2006). Likewise, numerous studies have aimed to identify genes associated with prognosis in NSCLC in order to better determine patient outcome (Guo et al, 2008); however, our data suggests that these relationships may be subtype specific as well. For example, low CD9 expression has previously been linked with poor prognosis in NSCLC (Higashiyama et al, 1995); thus, the specific inactivation of this gene in AC may signify that the association may be more relevant to this subtype.

Strikingly, genes which are known to influence NSCLC response to conventional chemotherapy were also deregulated in a subtype specific manner. For example, previous studies have shown that the increased expression of the multidrug resistance protein ABCC5 is associated with resistance to gemcitabine and cisplatin (Oguri et al, 2006; Weaver et al, 2005). Thus, the activation of ABCC5 in SqCC may lead to increased resistance of these tumors to treatment regimes based on these drugs. Similarly, the finding that ERCC1 disruption was subtype specific is particularly significant. ERCC1 is a nucleotide excision repair gene which functions in repairing DNA adducts and lesions induced by smoking-related carcinogens (Olaussen et al, 2007). As such, low expression levels of ERCC1 have been implicated in lung cancer susceptibility (Cheng et al, 2000) and tumorigenesis whereas high expression levels are associated with favorable overall prognosis (Olaussen et al, 2007). However, since ERCC1 is
also involved in the repair mechanism of cisplatin-induced DNA adducts in cancer cells, high expression levels lead to increased resistance to platinum-based chemotherapies (Herbst et al., 2008) (Vilmar & Sorensen, 2009). Low expression, on the other hand, leads to sensitivity to these drugs (Felip & Rosell, 2007). Recent clinical trials have subsequently described a significantly better outcome for patients who received adjuvant cisplatin-based combination chemotherapy if their resected tumors expressed low levels of ERCC1 (Herbst et al., 2008; Olaussen et al., 2007). Therefore, our finding that this gene is inactivated specifically in AC has major clinical consequences in terms of guiding disease management and treatment strategies in order to define appropriate treatment regimens for patients. To our knowledge, this is the first report demonstrating the subtype specificity of ERCC1 expression levels in NSCLC and further highlights how biological differences between AC and SqCC may influence patient response to therapy.

As stated above, great effort has been employed to define genes associated with poor prognosis in NSCLC, mainly to identify those patients at high risk of recurrence. In order to demonstrate the impact of genetic differences between the lung cancer subtypes on clinical characteristics such as prognosis, we determined the correlation with survival for all PSCNA-regulated genes. Remarkably, this analysis revealed that the AC and SqCC genes predictive of survival displayed these relationships in a subtype restricted manner; PSCNA genes predictive of poor prognosis in SqCC were not associated with survival in AC and vice versa. For instance, ELAVL1 was a PSCNA-regulated gene in both subtypes, with inactivation common in AC and activation prevalent in SqCC. Interestingly, overexpression in SqCC was also significantly correlated with poor survival (Figure 6.5). ELAVL1 (also known as HuR) is a RNA-binding protein involved in post-transcriptional mRNA stabilization (Levy et al., 1998). Overexpression of this gene increases the stability and half-life of VEGF mRNA, contributing to tumor angiogenesis and increasing tumor size (Yoo et al., 2006). As such, further investigation is warranted to determine the role this gene plays in tumorigenesis and patient outcome in SqCC. Furthermore, we discovered that specific genes may be indicative of totally different clinical outcomes depending on which subtype they are disrupted. For example, SLC1A5 was activated specifically in SqCC and high expression of this gene correlated with poor survival in this subtype as well. However, the opposite was true in AC; high expression was associated with favorable survival and low expression with poor survival. 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. In addition, this underlines the importance of separating AC and SqCC when assessing genes involved in predicting patient prognosis and other clinical outcomes.

Network based analysis of the PSCNA-regulated genes revealed insights into the oncogenic mechanisms driving the differential development of AC and SqCC. Two major, non-overlapping gene networks were uncovered for each subtype based on significant cumulative disruptions in biologically related genes (Figure 6.4, Table 6.1). The SqCC gene networks were both associated with DNA replication, recombination and repair and were also individually involved in controlling cellular organization and assembly (Network 1), cell cycle (Network 1) cancer (Network 2) and cell death (Network 2). The first network was composed mainly of genes that bind and modify the histone proteins H3 and H4. These proteins are fundamental building blocks of eukaryotic chromatin and are involved in a myriad of DNA-templated processes such as replication, repair, recombination and chromosome segregation (Strahl & Allis, 2000). In addition, they are part of the essential machinery responsible for regulating gene transcription by altering chromatin structure to control DNA accessibility (Strahl & Allis, 2000). A wide-array of post-translational covalent modifications including acetylation, methylation, phosphorylation, ubiquitylation and sumoylation in the N-terminal tail domains of histones control these processes (Esteller, 2008). Changes in these modifications have the potential to affect the structure and integrity of the genome and to disrupt normal patterns of gene expression during tumorigenesis (Esteller, 2008). Recently, global alterations of histone modification patterns have been reported in human cancers (Barlesi et al, 2007). Our data suggest that direct deregulation of histone modification enzymes including PRMT1, SAE1, SET8, CHAF1A and UHRF1 may cause these effects and play a key role during the development of lung SqCC. Interestingly, histone modification alterations have been observed to occur more frequently in lung SqCC than AC, supporting our findings (Van Den Broeck et al, 2008).

The second SqCC gene network was composed mainly of genes regulating cell cycle at the G1/S phase transition. This checkpoint is tightly controlled in normal cells as it determines the final decision to enter DNA replication and proliferate or undergo growth arrest, a function mediated by the retinoblastoma (RB) protein pathway (Dobashi, 2005). Deregulation of this pathway is common during malignant transformation though the activation of cell cycle progression promoting (cyclins, cyclin dependant kinases (CDKs), E2F transcription factors) or inactivation of progression inhibiting (RB, CDK inhibitors) components (Moroy & Geisen, 2004). Strikingly, we found that direct activation of cyclins E1 and D2 (CCNE1 and CCND2, respectively) as well as genes that lead to their increased expression (EIF4EBP1, FOXM1,
HSBP1 via ERK) is frequent in SqCC. In addition, we found that genes essential for the initiation of DNA replication during S phase including MCM2, MCM5 and CDC45L are also activated in this subtype. Lastly, as deregulation of the G1/S checkpoint often stimulates apoptosis, we saw complementary activation of pro-survival genes (AKT1S1, FOSL2 and HSPB1) as well as the inactivation of pro-apoptotic genes (CASP6) that may bypass this event. Together, this data suggests that deregulation of the G1/S checkpoint and subsequent initiation of DNA replication plays an essential role in stimulating increased cell proliferation during SqCC tumorigenesis. Interestingly, since histone modifications also play an essential role in DNA replication (Esteller, 2007), there may be a synergistic effect between the two SqCC gene networks that contributes to tumor development.

The AC gene networks both contained genes involved in regulating cell cycle and cancer development. AC network 1 was focused mainly on genes known to be targeted by the transcription factor HNF4α. HNF4α regulates a large set of genes in a cell-specific manner and is necessary for cell differentiation during normal embryonic development and maintenance of a differentiated epithelial phenotype in adults (Lazarevich & Fleishman, 2008). Deregulation of HNF4α has been documented for hepatocellular and renal cell carcinoma where loss of expression leads to increased cellular proliferation, progression and dedifferentiation (Grigo et al, 2008; Lazarevich et al, 2004; Lucas et al, 2005; Sel et al, 1996; Watt et al, 2003). This data suggests that HNF4α may act as a tumor suppressor in epithelial carcinogenesis (Lazarevich & Fleishman, 2008). Interestingly, although HNF4α was not affected, we found that numerous downstream targets of this gene are downregulated specifically in AC. Thus, this may have the same net affect as inactivation of HNF4α itself and lead to increased cellular proliferation during AC tumorigenesis.

The second AC gene network was mostly composed of genes which are controlled by TGF-β and p53 signaling. TGF-β inhibits cell cycle progression and induces apoptosis and the loss of these growth inhibitory effects can occur through several different mechanisms in NSCLC (Sato et al, 2007). We provide evidence that the downregulation of TGF-β signaling target genes may be a prominent mechanism in bypassing these effects in AC, contributing to increased cell proliferation during cancer development. Likewise, p53 signaling leads to cell cycle arrest in order to permit DNA repair or apoptosis and maintain genomic stability (Sato et al, 2007). p53 is the most frequently mutated gene in human cancers, occurring in ~50% of all NSCLC cases (Sato et al, 2007). Thus, the loss of downstream target genes such as ERCC1, which functions in both DNA repair and apoptosis, may be an alternate mechanism to bypass the growth
inhibitory effects of activated p53 signaling in a subset of AC tumors. Remarkably, TGF-β has been shown to activate both p53 and HNF4α, providing a link between the genes in both the major AC gene networks.

In conclusion, high resolution integrative analysis of NSCLC genomes delineated novel tumor subtype-specific genetic alterations responsible for driving the differential development and resulting phenotypes of AC and SqCC. The specific genes and networks identified in this study provide essential starting points for clarifying 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 should be studied separately when designing treatment strategies and testing new drugs in clinical trials.

6.4 Materials and Methods

6.4.1 DNA samples

Formalin-fixed, paraffin embedded and fresh-frozen tissues were collected from St. Paul’s Hospital, Vancouver General Hospital and Princess Margret 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 manual microdissection to ensure >70% tumor cell content. DNA was isolated using standard procedure with proteinase K digestion followed by phenol-chloroform extraction as previously described (Garnis et al, 2005).

6.4.2 Tiling path array comparative genomic hybridization

Array hybridization was performed as previously described (Baldwin et al, 2005; Coe et al, 2006a; Lockwood et al, 2007). Briefly, equal amounts (200-400 ng) of sample and single male reference genomic DNA were differentially labeled and hybridized to SMRT array v.2 (BCCRC Array Laboratory, Vancouver, BC) previously described to give optimal genome coverage (Ishkanian et al, 2004; Watson et al, 2007).

Hybridized arrays were imaged using a charge-coupled device (CCD) camera system and analyzed using SoftWoRx Tracker Spot Analysis software (Applied Precision, Issaquah, WA). Systematic biases were removed from all array data files using a stepwise normalization procedure as previously described (Khojasteh et al, 2005; Lockwood et al, 2008). SeeGH software was used to combine replicates and visualize all data as log2 ratio plots (Chi et al,
Stringently, all replicate spots with a standard deviation above 0.075 or signal to noise ratios below three were removed from further analysis. The clones were then positioned based on the human March 2006 (hg18) genome assembly. Genomic imbalances (gains and losses) within each sample were identified using aCGH-Smooth (Jong et al, 2004) with lambda and breakpoint per chromosome settings at 6.75 and 100, respectively as previously described (Coe et al, 2006b). The resulting frequency of alteration was then determined for each lung cancer cell type as described previously (Coe et al, 2006b).

6.4.3 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 clones within 10 Mb. Probes were then aggregated into genomic regions if the similarity in copy number status between adjacent clones 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 Fishers 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 clone (Coe et al, 2006b). A Benjamini-Hochberg multiple hypothesis testing correction based on the number of distinct regions was applied and resulting p-values \( \leq 0.01 \) 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.

6.4.4 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., Mississauga, ON). Samples were labeled and hybridized to a custom Affymetrix microarray, containing 43,737 probes mapping to ~23,000 unique genes, according to the manufacture’s protocols (Affymetrix Inc., Santa Clara, CA.) In addition, RNA was obtained from exfoliated bronchial cells of lung cancer free individuals obtained during fluorescence bronchoscopy (Chari et al, 2007). All individuals were either current or former smokers. Expression profiles were generated for all these cases using the Affymetrix U133
Plus 2 platform (Affymetrix Inc., Santa Clara, CA.). All data was normalized using the Robust Multichip Average (RMA) algorithm in R (Irizarry et al, 2003). In addition, two publically available datasets downloaded from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) were used: Affymetrix U133 Plus 2 expression data (accession number GSE3141) (Bild et al, 2006) and PC Human Operon v2 21k expression data (accession numbers GSE5843 and GSE5123) (Choi et al, 2006; Choi et al, 2007) for NSCLC tumors.

6.4.5 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 and computed with the ranksum function in MATLAB (version R2007b). 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 \( \leq 0.01 \) 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 above) against normal bronchial epithelial cells. If multiple probes were present for a gene, the one with the maximum intensity across all samples was used. All comparisons were performed using a one-tailed t-test with unequal variances in Excel and genes with a \( p < 0.01 \) were considered significant. The fold change for tumors vs normals 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) 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. Briefly, ROC analysis is based on comparison of true positive and false positive rates at various cut-offs. An ROC AUC value of
0.5 would indicate that the marker is no better than random chance at separating two groups, while a score of 1 would indicate that the marker is perfect at separating the two groups. Generally a marker with an AUC of 0.8 to 0.9 is considered good, while a AUC of 0.7 to 0.8 would represent a "fair" marker. Calculations were performed using the calculator at: http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html.

6.4.6 Survival analysis

Survival analysis was performed using the statistical toolbox in MATLAB. Expression data for each gene were sorted and survival times were compared between the top 1/3 and bottom 1/3 in expression using publicly available gene expression microarray datasets with survival data. Two tailed P-values were generated using a Mann-Whitney U test and those < 0.05 were considered significant. Kaplan-Meier plots were then generated for each gene of interest.

6.4.7 Network identification

Functional identification of gene networks was performed using Ingenuity Pathway Analysis program version 7.4 (Ingenuity® Systems, www.ingenuity.com). 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.
Figure 6.1. Copy number alterations in AC and SqCC. Alteration frequencies for AC (red) and SqCC (blue) are displayed across the entire human genome. Solid vertical black lines represent chromosome boundaries whereas the dotted black lines represent chromosome arm boundaries. A) Frequency of copy number gain. B) Frequency of copy number loss. C) The significance of copy number disparity (inverse p-value) between AC and SqCC subtypes is depicted. Solid black lines represent regions considered statistically different (p ≤ 0.01) whereas grey lines are not.
Figure 6.2. Differential expression as a result of PSCNAs. Transformed absolute expression data for the 442 unique genes exhibiting disruption in expression levels as a result of copy number differences are displayed. In addition, these genes are up or down-regulated in the subtype which they are disrupted compared to normal lung tissue (see results). High-level expression is indicated by red while black indicates progressively lower levels of expression. The AC samples are indicated by red highlighting on the top of each column, while SqCC samples are indicated by blue highlighting. Each gene is sorted according to its chromosomal position. There is a clear distinction in the expression of these genes indicating their specific involvement in the subtypes.
Figure 6.3. Genes deregulated by PSCNAs contribute to AC and SqCC phenotypes. Principal components analysis was performed utilising all genes demonstrating expression differences as a result of copy number alterations using: A) Data generated for 49 NSCLC tumors (29 AC, 20 SqCC) as part of this study which was used in gene discovery; B) Publically available data from 111 NSCLC tumors (58 AC, 53 SqCC) used as test set #1; C) Publically available data from 99 NSCLC tumors (48 AC, 51 SqCC) used as test set #2. AC samples are indicated by red circles, while the SqCC samples are indicated by blue circles. Strong separation of the AC and SqCC tumors along principal component 1 in all sets demonstrates the contribution of these genes to the differential phenotypes.
Figure 6.4. Gene networks involved in the development of SqCC and AC. Ingenuity Pathway Analysis was used to identify biologically related networks from the subtype specific genes deregulated by PSCNAs (see Methods). The top two resultant gene networks for each subtype are displayed. Solid lines denote direct interactions while dotted lines represent indirect interactions between the genes. Network components highlighted in red are upregulated in the corresponding subtype whereas those highlighted green are downregulated. Those not highlighted are used by the software to display relationships. Additional information about the genes and their interactions can be found at www.ingenuity.com. A) SqCC network #1 displaying potential interactions between multiple histone regulating genes. B) SqCC network #2 of genes related to cyclin/RB signalling. C) AC network #1 of genes related to HNF4 signalling. D) AC network #2 displaying multiple genes regulated by TGF-beta signaling.
Figure 6.5.  *ELAVL1* is a PSCNA-regulated gene that predicts poor survival in SqCC.  
*ELAVL1* was a PSCNA-regulated gene in each subtype with loss and underexpression in AC and gain and overexpression in SqCC.  A) Expression of *ELAVL1* between 29 AC and 20 SqCC tumors generated as part of this study.  The expression values are presented as normalized log2 intensity values.  B) Kaplan-Meier analysis of *ELAVL1* in SqCC.  Association of high expression with poor survival was statistically significant (p < 0.05) in this subtype only (see Methods section).
Table 6.1. Top gene networks associated with AC and SqCC PSCNA-regulated targets

<table>
<thead>
<tr>
<th>Network</th>
<th>Molecules in Network</th>
<th>Score</th>
<th>Focus Molecules</th>
<th>Top Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>SqCC1</td>
<td>ASF1B, Basc, BLM, BRD4, CHAF1A, CLDND1, DHRS12, FBL, FCHO2, Histone h3, Histone h4, LARP2 (includes EG:55132), LIMCH1, LRBA, Mapk, MIRN124, NCAPD2, NCAPH2, NHP2L1, NOL5A, PNO1, PRMT1, RFC2, RFC4, SAE1, SETD8 (includes EG:387893), SMC4, SNRPB, SNRPD2, STRAP, SUCLG2, TBL1XR1, TCF3, UBC, UHRF1</td>
<td>53</td>
<td>30</td>
<td>Cellular Assembly and Organization, DNA Replication, Recombination, and Repair, Cell Cycle</td>
</tr>
<tr>
<td>SqCC2</td>
<td>AKT1S1, BRF2, Calpain, CASP6, Caspase, CCND2, CCNE1, CDC45L, CDKN2D, CENPO, COP57A, Cyclin D, Cyclin E, E2f, EIF4EBP1, EIF4G1, ERK, FOSL2, FOXM1, HSPB1, MAML3, MCM2, MCM5, NCBP2, NOTCH3, OTUD4, PCYT1A, PHB2, PRNP, RAP2B, RAPGEF2, Rb, SRPRB, TOM1L1, Ubiquitin</td>
<td>45</td>
<td>27</td>
<td>DNA Replication, Recombination, and Repair, Cancer, Cell Death</td>
</tr>
<tr>
<td>AC1</td>
<td>AMBP, ARD1A, ARL1, BLOC1S1, C18orf6, C22orf28, CCDC53, CDC42EP3, CNBP, CTDSPL2, ERCC5, GTF2H3, GTF2H4, HNF4A, HNRNPR, ITH4, LCMT2, MAGOH, MIRN20A (includes EG:406982), NAT13, NCBP1 (includes EG:4686), OPA1, PITPNB, PLDN, POLRMT, QTPT, RAB18 (includes EG:61876), RIOK3, RPL31, RPL35, RPL23A (includes EG:6147), SPATA5L1 (includes EG:79029), TFB1M, UMPS, VHL</td>
<td>33</td>
<td>16</td>
<td>Cancer, Cardiovascular System Development and Function, Cell Cycle</td>
</tr>
<tr>
<td>AC2</td>
<td>ALPL, ASH2L, ATP13A3 (includes EG:79572), BMP4, BRCC3, BRIP1, DENR (includes EG:8562), DLL3, EIF4A2, ERCC1, ERCC5, ERL12, IKKKB, KEAP1, MAPK14, MARCH5, MIB1, MPHOSPH9, MYCN, PROSC, retinoic acid, RPS15, SMCR7L, SMG1, SP1, SPAG9, TFAM, TGFBI, TOP2B, TOPBP1, TP53, UBA52, UBE2V1, UFD1L, ZFP36</td>
<td>27</td>
<td>14</td>
<td>Cell Cycle, Gene Expression, Cancer</td>
</tr>
</tbody>
</table>

1. Score = score is a numerical value used to rank networks according to their degree of relevance to the Network Eligible molecules in the dataset. Score is based on the hypergeometric distribution and is calculated with the right-tailed Fisher’s Exact Test. The score is the negative log2 of this p-value. 2. Focus Molecules = the number of genes involved in the particular network. 3. Bolded functions represent those that overlap with between the networks from the same subtype.
Table 6.2. PSCNA-regulated genes associated with survival in each subtype

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>Locus</th>
<th>Regulation</th>
<th>AC P-Value</th>
<th>NSCLC P-Value</th>
<th>SqCC P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENPA</td>
<td>centromere protein A</td>
<td>2p23.3</td>
<td>SqCC +</td>
<td>-</td>
<td>-</td>
<td>7.82E-03</td>
</tr>
<tr>
<td>ITIH4</td>
<td>inter-alpha (globulin) inhibitor H4 (plasma Kallikrein-sensitive glycoprotein)</td>
<td>3p21.1</td>
<td>AC +</td>
<td>5.53E-03</td>
<td>0.02</td>
<td>-</td>
</tr>
<tr>
<td>CNBP</td>
<td>CCHC-type zinc finger, nucleic acid binding protein</td>
<td>3q21.3</td>
<td>AC -</td>
<td>0.01</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>TP73L</td>
<td>tumor protein p63</td>
<td>3q26</td>
<td>AC - , SqCC +</td>
<td>0.03</td>
<td>2.30E-03</td>
<td>0.03</td>
</tr>
<tr>
<td>KPNA4</td>
<td>karyopherin alpha 4 (importin alpha 3)</td>
<td>3q26.1</td>
<td>SqCC +</td>
<td>-</td>
<td>-</td>
<td>0.03</td>
</tr>
<tr>
<td>GOLPH4</td>
<td>golgi integral membrane protein 4</td>
<td>3q26.2</td>
<td>SqCC +</td>
<td>-</td>
<td>-</td>
<td>0.04</td>
</tr>
<tr>
<td>OPA1</td>
<td>optic atrophy 1 (autosomal dominant)</td>
<td>3q29</td>
<td>AC -</td>
<td>4.31E-03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>WDR53</td>
<td>WD repeat domain 53</td>
<td>3q29</td>
<td>SqCC +</td>
<td>-</td>
<td>-</td>
<td>0.04</td>
</tr>
<tr>
<td>KIAA1109</td>
<td>N/A</td>
<td>4q27</td>
<td>SqCC -</td>
<td>-</td>
<td>-</td>
<td>0.03</td>
</tr>
<tr>
<td>RAPGEF2</td>
<td>Rap guanine nucleotide exchange factor (GEF) 2</td>
<td>4q32.1</td>
<td>SqCC -</td>
<td>-</td>
<td>-</td>
<td>0.04</td>
</tr>
<tr>
<td>KLHL2</td>
<td>kelch-like 2, Mayven (Drosophila)</td>
<td>4q32.3</td>
<td>SqCC -</td>
<td>-</td>
<td>-</td>
<td>9.01E-03</td>
</tr>
<tr>
<td>TMEM192</td>
<td>transmembrane protein 192</td>
<td>4q32.3</td>
<td>SqCC -</td>
<td>-</td>
<td>-</td>
<td>0.04</td>
</tr>
<tr>
<td>CDC40</td>
<td>cell division cycle 40 homolog (S. cerevisiae)</td>
<td>6q21</td>
<td>AC -</td>
<td>0.03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CSDA</td>
<td>cold shock domain protein A</td>
<td>12p13.2</td>
<td>AC - , SqCC +</td>
<td>9.71E-03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PARP11</td>
<td>poly (ADP-ribose) polymerase family, member 11</td>
<td>12p13.32</td>
<td>AC -</td>
<td>0.03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ADIPOR2</td>
<td>adiponectin receptor 2</td>
<td>12p13.33</td>
<td>AC - , SqCC +</td>
<td>0.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ATP6V0A2</td>
<td>ATPase, H+ transporting, lysosomal V0 subunit a2</td>
<td>12q24.31</td>
<td>AC -</td>
<td>8.17E-03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CDK2AP1</td>
<td>CDK2-associated protein 1</td>
<td>12q24.31</td>
<td>SqCC +</td>
<td>-</td>
<td>-</td>
<td>0.03</td>
</tr>
<tr>
<td>QPRT</td>
<td>quinolinate phosphoribosyltransferase (nicotinate-nucleotide pyrophosphorylase (carboxylating))</td>
<td>16.11.2</td>
<td>AC +</td>
<td>9.71E-03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ELAV1</td>
<td>ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 (Hu antigen R)</td>
<td>19p13.2</td>
<td>AC - , SqCC +</td>
<td>-</td>
<td>-</td>
<td>0.03</td>
</tr>
<tr>
<td>LMNB2</td>
<td>lamin B2</td>
<td>19p13.3</td>
<td>SqCC +</td>
<td>-</td>
<td>-</td>
<td>0.03</td>
</tr>
<tr>
<td>SLC1A5</td>
<td>solute carrier family 1 (neutral amino acid transporter), member 5</td>
<td>19q13.32</td>
<td>SqCC +</td>
<td>0.01</td>
<td>-</td>
<td>0.04</td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Chromosome</td>
<td>Subtype</td>
<td>Reg.</td>
<td>Exp.</td>
<td>P-Value</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------------------------------</td>
<td>------------</td>
<td>---------------</td>
<td>------</td>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>AKT1S1</td>
<td>AKT1 substrate 1 (proline-rich)</td>
<td>19q13.33</td>
<td>SqCC +</td>
<td></td>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td>NUP62</td>
<td>nucleoporin 62kDa</td>
<td>19q13.33</td>
<td>SqCC +</td>
<td></td>
<td></td>
<td>0.03</td>
</tr>
<tr>
<td>MCM3AP</td>
<td>minichromosome maintenance complex component 3 associated protein</td>
<td>21q22.3</td>
<td>SqCC -</td>
<td></td>
<td></td>
<td>1.92E-03</td>
</tr>
<tr>
<td>PITPNB</td>
<td>phosphatidylinositol transfer protein, beta</td>
<td>22q12.1</td>
<td>AC - , SqCC +</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C22orf9</td>
<td>chromosome 22 open reading frame 9</td>
<td>22q13.31</td>
<td>SqCC +</td>
<td></td>
<td></td>
<td>4.52E-03</td>
</tr>
</tbody>
</table>

1. Regulation AC = Adenocarcinoma; SqCC = Squamous Cell Carcinoma; + = Increased copy number and expression in the indicated subtype; - = Decreased copy number and expression in the indicated subtype. 2. P-Value calculated from Mann-Whitney U-Test (see Methods).
Table 6.3. Genes previously implicated in NSCLC displaying subtype specific disruption

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>Locus</th>
<th>Regulation(^1)</th>
<th>Role in NSCLC</th>
<th>Representative References(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRDM2</td>
<td>PR domain containing 2, with ZNF domain</td>
<td>1p36.21</td>
<td>SqCC -</td>
<td>tumorigenesis</td>
<td>PMID: 17693662</td>
</tr>
<tr>
<td>FOXP1</td>
<td>forkhead box P1</td>
<td>3p14.1</td>
<td>SqCC -</td>
<td>tumorigenesis</td>
<td>PMID: 11751404</td>
</tr>
<tr>
<td>FHIT</td>
<td>fragile histidine triad gene</td>
<td>3p14.2</td>
<td>SqCC -</td>
<td>tumorigenesis, prognosis, response</td>
<td>PMIDS: 8620533, 18690840, 14976524</td>
</tr>
<tr>
<td>ABCC5</td>
<td>ATP-binding cassette, sub-family C (CFTR/MRP), member 5</td>
<td>3q27.1</td>
<td>SqCC +</td>
<td>response</td>
<td>PMID: 10728601</td>
</tr>
<tr>
<td>HSPB1</td>
<td>heat shock 27kDa protein 1</td>
<td>7q11.23</td>
<td>SqCC +</td>
<td>prognosis, response</td>
<td>PMIDS: 18383892, 16328069</td>
</tr>
<tr>
<td>DLX5</td>
<td>distal-less homeobox 5</td>
<td>7q21.3</td>
<td>SqCC +</td>
<td>prognosis</td>
<td>PMID: 18413826</td>
</tr>
<tr>
<td>EIF4EBP1</td>
<td>eukaryotic translation initiation factor 4E binding protein 1</td>
<td>8p12</td>
<td>SqCC +</td>
<td>tumorigenesis</td>
<td>PMID: 16618749</td>
</tr>
<tr>
<td>STRAP</td>
<td>serine/threonine kinase receptor associated protein</td>
<td>12p12.3</td>
<td>SqCC +</td>
<td>tumorigenesis</td>
<td>PMID: 16778189</td>
</tr>
<tr>
<td>CD9</td>
<td>CD9 molecule</td>
<td>12p13.31</td>
<td>AC -</td>
<td>prognosis</td>
<td>PMID: 8521390</td>
</tr>
<tr>
<td>CCND2</td>
<td>cyclin D2</td>
<td>12p13.32</td>
<td>SqCC +</td>
<td>tumorigenesis</td>
<td>PMID: 14506731</td>
</tr>
<tr>
<td>FOXM1</td>
<td>forkhead box M1</td>
<td>12p13.33</td>
<td>SqCC +</td>
<td>tumorigenesis</td>
<td>PMID: 16489016</td>
</tr>
<tr>
<td>DUOX1</td>
<td>dual oxidase 1</td>
<td>15q21.1</td>
<td>AC -</td>
<td>tumorigenesis</td>
<td>PMID: 18281478</td>
</tr>
<tr>
<td>NOTCH3</td>
<td>Notch homolog 3 (Drosophila)</td>
<td>19p13.12</td>
<td>SqCC +</td>
<td>tumorigenesis</td>
<td>PMID: 10944559</td>
</tr>
<tr>
<td>KEAP1</td>
<td>kelch-like ECH-associated protein 1</td>
<td>19p13.2</td>
<td>AC -</td>
<td>tumorigenesis, prognosis</td>
<td>PMIDS: 17020408, 1831659</td>
</tr>
<tr>
<td>ILF3</td>
<td>interleukin enhancer binding factor 3, 90kDa</td>
<td>19q13.2</td>
<td>AC -</td>
<td>prognosis</td>
<td>PMID: 19088038</td>
</tr>
<tr>
<td>ELAVL1</td>
<td>ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 (Hu antigen R)</td>
<td>19q13.2</td>
<td>AC -, SqCC +</td>
<td>tumorigenesis</td>
<td>PMID: 10900464</td>
</tr>
<tr>
<td>ERCC1</td>
<td>excision repair cross-complementing rodent repair deficiency, complementation group 1</td>
<td>19q13.32</td>
<td>AC -</td>
<td>tumorigenesis, prognosis, response</td>
<td>PMID: 10910954</td>
</tr>
<tr>
<td>AKT1S1</td>
<td>AKT1 substrate 1 (proline-rich)</td>
<td>19q13.33</td>
<td>SqCC +</td>
<td>tumorigenesis</td>
<td>PMID: 16174443</td>
</tr>
<tr>
<td>CRKL</td>
<td>v-crk sarcoma virus CT10 oncogene homolog (avian)-like</td>
<td>22q11.21</td>
<td>SqCC +</td>
<td>tumorigenesis</td>
<td>PMID: 16391854</td>
</tr>
<tr>
<td>MAPK1</td>
<td>mitogen-activated protein kinase 1</td>
<td>22q11.21-22q11.22</td>
<td>AC -</td>
<td>tumorigenesis, prognosis</td>
<td>PMID: 14997206</td>
</tr>
</tbody>
</table>

1. Regulation AC = Adenocarcinoma; SqCC = Squamous Cell Carcinoma; + = Increased copy number and expression in the indicated subtype; - = Decreased copy number and expression in the indicated subtype. 2. PMID = Pubmed Identification
6.5 References


136
Coe BP, Lee EH, Chi B, Girard L, Minna JD, Gazdar AF, Lam S, MacAulay C, Lam WL (2006a) Gain of a region on 7p22.3, containing MAD1L1, is the most frequent event in small-cell lung cancer cell lines. *Genes Chromosomes Cancer* 45: 11-9


Sel S, Ebert T, Ryffel GU, Drewes T (1996) Human renal cell carcinogenesis is accompanied by a coordinate loss of the tissue specific transcription factors HNF4 alpha and HNF1 alpha. 
*Cancer Lett* 101: 205-10


Chapter 7: BRF2 is a lineage specific oncogene amplified early in squamous cell lung cancer development

A version of this chapter has been submitted for publication as:


Supplementary tables for this chapter have been omitted due to size limitations and can be provided upon request.
7.1 Introduction

Lung cancer is the leading cause of cancer deaths worldwide with non-small cell lung cancer (NSCLC) accounting for the majority of cases (Jemal et al., 2008). Despite advances in treatment, the prognosis for lung cancer patients remains poor with an overall five year survival rate of ~15%\(^1\). Squamous cell carcinoma (SqCC) and adenocarcinoma (AC) are the predominant NSCLC cell types. Currently, they are regarded as a single disease entity in terms of systemic therapy, which only provides a modest improvement in survival compared to best supportive care (Travis, 2002). However, SqCC and AC display distinct development and progression characteristics probably related to the specific cell lineages from which they develop, which in turn affect the range of genetic alterations required for tumor initiation in a lineage restricted manner (Garraway & Sellers, 2006). The identification of the molecular differences between the tumor types will have a significant impact in designing novel therapies that can improve the treatment outcome.

The subsistence of a DNA amplicon is thought to result from selection of genes within the amplified region that promote tumor growth (Albertson, 2006). The specific requirements for tumorigenesis in different cell lineages may therefore be associated with selection of different amplicons. Copy number increase of 8p12-8p11.21 is one of the most frequent focal changes in NSCLC occurring in ~9-35% of cases, with amplification present in ~3-8% of cases in the literature, a frequency rivaling those of established NSCLC oncogenes such as MYC (~6%) and EGFR (~3%) (Kendall et al., 2007; Tonon et al., 2005). In this study, we determine the lineage specificity of the 8p amplicon and discovered a novel oncogene restricted to tumorigenesis in the squamous cell lineage.

7.2 Results and Discussion

7.2.1 8p amplification is restricted to the SqCC cancer type

We compared the 8p chromosome arm of 161 NSCLC tumors – 103 AC and 58 SqCC (sample set 1a, supporting information (SI) Table S1) – by tiling resolution array comparative genomic hybridization (array CGH) (Ishkanian et al., 2004). After hybridization experiments, genomic profiles were normalized to remove systematic experimental biases and subjected to a smoothing algorithm in order to computationally define regions of copy number gain and loss along the entire length of chromosome arm 8p (Coe et al., 2006). Individual samples were then grouped by their corresponding cell type, and probes were aggregated into regions based on
similar copy number status. The resulting frequency of alteration for each region along the arm was compared between cell types using the Fisher’s exact test to identify regions of copy number disparity, and the resulting p-values were corrected for multiple comparisons with a cut-off of ≤ 0.01 considered significant. In addition, regions had to be altered in >20% of samples in a group and the difference between groups of >10% to be considered (Methods). Although the telomeric portion of 8p was frequently lost in both AC and SqCC cell types, two regions spanning a total of 5.65 Mbp at 8p12-8p11.21 were found to be frequently gained specifically in SqCC (Figure 7.1a and 7.1b, Table S2). Copy number increase of focal regions at 8p12-p11.21 was found in up to 40% of SqCC tumors, while DNA loss was the most prevalent event in AC (~39%). In addition, high level amplification was present in ~12% of SqCC samples (7/58) demonstrating the preferential selection for this alteration in tumors of this cell lineage. 

The increased incidence of 8p amplification in comparison to previous reports is attributed to analyzing the cell types as distinct groups, as opposed to combining all the NSCLC cell types as a single entity. These results indicate that gain or amplification of 8p12-8p11.21 is restricted to SqCC and occurs far more frequently than previously thought, highlighting the importance of considering cell lineage in genomic studies of malignancies from the same tissue site.

7.2.2 BRF2 gene expression drives selection of the 8p amplicon in lung SqCC

The cell type dependent pattern of 8p amplification raised the possibility that a lineage-specific oncogene may be driving the preferential selection of this amplicon in SqCC. Such a gene should display five fundamental properties each translating into its own testable hypothesis. First, increased expression would be restricted to SqCC tumors mirroring the specificity of DNA amplification (hypothesis 1). Second, as the target of the amplicon, expression would be higher in SqCC tumors with gain/amplification than those without (hypothesis 2). Third, expression should be significantly higher in SqCC tumors than normal bronchial epithelial cells; that is, the gene should be activated in cancerous and not normal tissue (hypothesis 3). Fourth, the gene should have oncogenic potential and provide a growth and/or survival advantage to cells when overexpressed (hypothesis 4). Lastly, if necessary for initiating tumorigenesis, amplification should occur early in tumor development and therefore be present in lung SqCC precursor lesions (hypothesis 5).

To test the first hypothesis, we generated gene expression microarray profiles for a subset of 47 tumors (34 AC, 13 SqCC) which were also analyzed by array CGH (and had sufficient remaining material) in order to integrate genetic and gene expression information (sample set 1b, Table S1). In total, 62 probes corresponding to 45 unique genes mapped to within the
alteration boundaries (Supplemental Table 2). To identify lineage-restricted genes, we compared the expression levels for all probes between the AC and SqCC samples. Since we predicted candidate genes to be over-expressed in SqCC, a one-tailed Mann Whitney U-Test was used with Benjamini-Hochberg corrected p-values ≤0.01 considered significant. Ten unique genes meeting these criteria were uncovered from this analysis which showed a clear distinction in expression levels between the AC and SqCC tumors (Figure 7.1c, Table S3).

After identifying these SqCC specific genes, we next aimed to ensure that amplification is responsible for their differential expression, as these will be candidate targets driving amplicon selection (hypothesis 2). For this purpose, we utilized two complementary approaches. First, a non-parametric Spearman correlation co-efficient was calculated for each gene using Z-transformed copy number ratios and log_{10} gene expression ratios (Methods). Five of the ten genes (LSM1, BRF2, ASH2L, TM2D2 and WHSC1L1) had a correlation co-efficient of >0.75 and a corrected p-value (representing the statistical significance of a positive correlation) of <0.01 and were further considered as candidates (Table S3 for all values). The second approach involved the comparison of expression levels between SqCC tumors with gene dosage increase (gain/amplification) and those with neutral copy number status (Methods). Of the five genes with a positive association between copy number and expression, only three (LSM1, BRF2, and ASH2L) also showed significantly elevated transcript levels specifically in SqCC samples with gain or amplification and were therefore determined to be regulated by copy number (Table S3).

In addition to demonstrating a linkage between expression and amplification, a candidate oncogene should only be expressed at elevated levels in cancerous, and not normal, tissues (Croce, 2008). Therefore, to test the 3rd hypothesis, we analyzed the RNA levels of these three genes in an independent panel of 53 SqCC lung tumors and 67 samples of exfoliated bronchial cells from cancer-free individuals generated using the Affymetrix U133 Plus 2 platform (sample sets 2 and 3, Table S1). Strikingly, only BRF2 was aberrantly expressed (>2 fold, p<1.0 x 10^{-8}) in cancerous tissues identifying it as the sole gene passing the three main criteria of a candidate lineage-specific oncogene described above (Figure 7.2, Table S3). To further confirm these observations, a third, independent sample set consisting of 118 NSCLC tumors and 39 non-neoplastic lung tissues (sample set 4, Supplemental Table 1) was analyzed for BRF2 expression by quantitative-polymerase chain reaction (qPCR) (Methods). Consistent with the microarray results, expression of BRF2 in primary tumors was significantly higher than that in
the non-neoplasia tissues \((p<0.001)\) with overexpression more common in SqCC than AC \((p=0.03)\), supporting our findings.

Taken together, results from testing the first three hypotheses clearly demonstrate that \(BRF2\) is the driver gene of the 8p amplicon and identify it as a candidate lineage-specific oncogene in SqCC. Previous studies investigating this amplified region in NSCLC have proposed \(FGFR1\) and \(WHSC1L1\) as potential oncogenes (Tonon \textit{et al}, 2005; Zhao \textit{et al}, 2005). However, we ruled out \(FGFR1\) as a possible target as it was not differentially expressed between AC and SqCC, and as such, was excluded from further analysis. This is in agreement with a study by Tonon \textit{et al} that suggested \(WHSC1L1\) as the more likely amplification target in NSCLC (Tonon \textit{et al}, 2005). Although we demonstrated that \(WHSC1L1\) expression was restricted to SqCC and correlated with increased gene dosage, it was not significantly higher in samples with gain/amplification or different between normal and cancerous cells \((p=0.12, \text{fold change}=1.3)\) and therefore, also discounted. It is worth noting that amplification of 8p12-11 is also prevalent in breast cancer and has been widely examined in order to delineate potential driver genes (Bernard-Pierrot, 2008; Garcia \textit{et al}, 2005; Gelsi-Boyer \textit{et al}, 2005; Ray \textit{et al}, 2004; Streicher \textit{et al}, 2007; Yang \textit{et al}, 2006). While \(BRF2\) is overexpressed when amplified in breast tumors, other targets including \(LSM1, BAG4, RAB11FIP1\) and \(PPAPDC1B\) have recently been described as more likely candidates. Thus, although amplification of this region occurs in both breast and lung SqCC tumors, it is likely that different driver genes are involved. As such, although 8p amplification is not restricted to lung SqCC, the targeting of \(BRF2\) as an oncogene appears to be specific to lung SqCC and this is the first study to implicate it in cancer lineage to date.

### 7.2.3 \(BRF2\) contributes to SqCC tumorigenesis by regulating cell growth and proliferation

\(BRF2\) encodes a subunit of a transcription initiation complex responsible for RNA polymerase III (Pol III) mediated transcription (Cabart & Murphy, 2001; Schramm \textit{et al}, 2000). Pol III transcribes a limited set of genes that encode nontranslated RNAs including 5S rRNA, tRNA, 7SL RNA and U6 RNA that are essential for protein synthesis and RNA processing (White, 2004). Since these processes are fundamental determinates of the capacity of a cell to grow, increased activity of Pol III is often observed during cancer development (White, 2005). Indeed, transformed cells express elevated levels of Pol III transcripts and inhibition of these transcripts limits cell growth and proliferation (Goodfellow & White, 2007). It has been proposed that deregulation of Pol III in transformed cells can occur through three different mechanisms: release from cellular repressors, direct activation by oncogenes, and overexpression of
transcription factors (White, 2004). In normal cells where growth is tightly controlled, tumor suppressors including RB, p53 and PTEN repress Pol III transcription (Felton-Edkins et al, 2003; Woiwode et al, 2008). Inactivation of these genes or activation of oncogenes such as MYC and ERK reverse this process (Felton-Edkins et al, 2003; Gomez-Roman et al, 2006; Goodfellow & White, 2007). Interestingly, the majority of these genes are mutated in lung cancer representing a potential mechanism of increasing Pol III activity and subsequently, cell growth potential during tumorigenesis. However, transcription factors are often the limiting components of Pol III mediated transcription and elevated levels of these components have been observed in numerous cancer types (White, 2005). Recently, the overexpression of another Pol III transcription factor BRF1 has been shown to increase Pol III mediated transcription, resulting in the transformation of cells in vitro and tumor formation in vivo (Johnson et al, 2008; Marshall et al, 2008). A study by Marshall et al was the first to implicate Pol III deregulation as a causative factor in cancer formation (Marshall et al, 2008); however, there have been no studies reported to date of activating mutations in Pol III subunits or associated transcription factors in tumors. Therefore, we hypothesized that the amplification and overexpression of BRF2 may contribute to lung SqCC tumorigenesis by contributing to increased cell growth and proliferation, representing a novel alternative mechanism of increasing Pol III transcription in cancer.

To test this hypothesis (hypothesis 4), we performed complementary loss and gain of function in vitro experiments using lung cancer cell lines and immortalized human bronchial epithelial cell (HBEC) line respectively. Twenty NSCLC cell lines (16 AC and 4 SqCC) previously analyzed by array CGH were assayed for BRF2 expression by qPCR (Methods). Mirroring the findings from the clinical tumor specimens, BRF2 expression was strongly correlated with gene dosage with the two cell lines with amplification (HCC95 and H520) displaying the highest transcript levels (data not shown). In addition, both these lines were derived from SqCC samples and no AC contained amplification, re-enforcing the lineage specificity of BRF2 activation. To determine the effect of BRF2 overexpression on BRF2 protein levels, three cell lines were selected for western-blot analysis: a SqCC with amplification (H520), an AC with neutral copy number (H1395) and an AC with loss (H2347) (Figure 7.3a). Consistent with an oncogenic role, high protein levels were only found in H520. To assess the functional significance of BRF2 activation in this line, siRNA mediated knockdown was performed. BRF2 siRNA pool transfection resulted in ~90% knockdown of expression relative to the non-targeting control siRNA pool as verified by qPCR ($p=0.045$, Figure 7.3b) and significantly reduced cell proliferation as measured by the 3-
These results demonstrate a crucial role for BRF2 in contributing to the sustained cellular proliferation and survival of this SqCC line. To further validate its oncogenic potential, we overexpressed BRF2 by stable transduction of immortalized HBEC lines and measured cell growth compared to vector-expressing controls. HBEC lines are immortalized without the use of viral oncoproteins, have minimal genetic changes and do not exhibit a transformed phenotype (Ramirez et al., 2004; Sato et al., 2006). In addition, since they express epithelial markers and morphology and can differentiate into mature airway cells, they represent an attractive model for testing the importance of specific gene alteration found in the initiation of epithelium derived lung cancer (Ramirez et al., 2004; Sato et al., 2006). Strikingly, the introduction of BRF2 alone resulted in a modest but significant increase in cellular growth and saturation density \((p=0.035)\), further supporting a tumorigenic role for this gene (Figure 7.3d). Furthermore, as p53 is inactivated in ~50% of NSCLC tumors and is known to repress Pol III mediated transcription, we sought to investigate the impact of BRF2 overexpression in conjunction with p53 silencing on HBEC growth. Interestingly, the combination of these two alterations enhanced cell growth greater than each alteration alone \((p=2.36 \times 10^{-5})\) suggesting a synergistic role for these alterations in promoting proliferation (Figure 3d). Taken together, our results demonstrate that BRF2 overexpression contributes to a tumorigenic phenotype by regulating cell growth and proliferation, confirming the functional significance of BRF2 gene amplification in SqCC.

7.2.4 BRF2 activation is an early event in SqCC development

The cell type restricted pattern of activation coupled with its transformation potential strongly implicates BRF2 as a lineage-specific oncogene in lung SqCC. SqCC carcinogenesis is thought to be a multistep process that involves the transformation of normal mucosa though a continuous range of precursor lesions up to carcinoma in situ (CIS) before invasive cancer and finally metastasis (Wistuba & Gazdar, 2006). However, since most studies focus on clinically evident tumors, little is known about the molecular events preceding the development of lung cancer and the underlying basis of carcinogenesis. Unlike dysplastic lung lesions that rarely progress, the majority of CIS cases will become invasive cancer (Wistuba & Gazdar, 2006). Therefore, we hypothesized that critical alterations necessary for disease progression would be evident in preinvasive CIS lesions and persist in invasive tumors. To determine if BRF2 activation occurs early in SqCC development (hypothesis 5) we analyzed gene dosage in a panel of 20 CIS lesions (sample set 5, Supplemental Table 1) obtained by autofluorescence bronchoscopy (Methods). Remarkably, array CGH revealed BRF2 copy number increases in
the majority of CIS cases (Figure 7.4a) with 35% (7/20) demonstrating high-level amplification (log2 ratio >0.8, Figure 4b). WHSC1L1 and FGFR1 were only amplified five times and once respectively, excluding these genes as primary driver genes of the amplicon (Figure 7.4b and 7.4c). To confirm that amplification results in increased expression of BRF2 in preinvasive lesions, we performed immunohistochemistry (IHC) on a CIS sample (CIS2) with amplification (Figure 7.4c). As expected, BRF2 expression was elevated in CIS epithelia in this sample in comparison to normal epithelia from the same patient (Figure 7.4d). Strong BRF2 expression was also observed in additional CIS cases with lower levels in earlier stages of neoplastic progression (mild, moderate and severe dysplasia) and little or no staining in benign lesions (hyperplasia and metaplasia), confirming that gene activation is an early event in SqCC development (Figure 7.5). Interestingly, the only benign lesion in which BRF2 expression was observed was obtained from a patient that had also developed CIS (Figure 7.5). The high frequency of activation in pre-invasive lesions suggests that BRF2 plays a critical role in the initiation and progression of SqCC through the increase of cell growth potential. Since patient survival can be significantly improved if the lesions are detected and treated at their pre-invasive stage, the identification of genes involved in the development of CIS and invasive SqCC is of vital clinical importance (Sato et al, 2007; Wistuba & Gazdar, 2006). Our finding that BRF2 is a lineage-specific oncogene amplified early inSqCC development, and not expressed in normal lung tissue, represents a critical step in understanding the progression of SqCC, and represents a promising target for therapeutic intervention.

7.2.5 Increased RNA processing is associated with BRF2 overexpression

To identify other genes and functions that may be associated with BRF2 mediated initiation of tumorigenesis, we performed Significance Analysis of Microarrays (SAM) on a panel of 111 NSCLC tumors (sample set 2, Supplemental Table 1) followed by gene enrichment analysis using Ingenuity Pathway Assist (IPA) (SI Methods). This analysis revealed 86 genes which were significantly increased (78) or decreased (8) (False discovery rate, FDR, <5%) in tumors with the highest BRF2 expression (Table S4). IPA analysis revealed enrichment for genes with diverse biological functions including RNA post-transcriptional modification, gene expression, cell cycle and cancer (Table S5). The identification of RNA post-transcriptional modification as the most significantly affected function (p= 1.7 x 10^{-6} - 4.73 x 10^{-2}, the two significance values refer to a range of specific sub-functions) was significant as this is one of the main roles of Pol III related transcripts as stated above. The genes related to this function which are increased in expression include FBL, CPSF6, RRP9, SNRPA, SFRS10, CSTF2T, LSM1, and CPSF3 and...
are involved in the modification, polyadenylation, and processing of both mRNA and rRNA. Since these are fundamental processes necessary for proper protein production and therefore cell growth, upregulation of these components may be associated with the increased proliferative capacity of SqCC cells upon BRF2 activation. However, the exact nature of this association is currently unknown and future studies will be needed to understand the mechanism responsible for BRF2 induced cell growth in SqCC.

Interestingly, BRF2 is also specifically involved in transcription from type 3 Pol III promoters which are responsible for the transcription of small nuclear RNA (snRNA) genes (Dieci et al, 2007; Saxena et al, 2005). snRNAs are responsible for a range of regulatory functions including the alteration of gene expression and a potential role for snRNAs in the genomic instability of cancer has been proposed (Rew, 2003). One particular snRNA transcribed by type III promoters is U6 snRNA which forms the catalytic core of the spliceosome (Butcher & Brow, 2005). The spliceosome performs the splicing of precursor mRNA in eukaryotic cells, removing introns and joining exons. This process is tightly regulated during growth and development and aberrant splicing has been linked to numerous human diseases including cancer (Faustino & Cooper, 2003). In fact, many oncogenes demonstrate alternative splicing patterns associated with neoplasia, and splicing regulatory factor expression levels have been shown to increase during cancer progression. Strikingly, many of the genes we identified as being associated with increased BRF2 expression, including SNRPA and SFRS10, are known to interact with snRNAs including U6 in the spliceosome complex. In addition, SNAPC5, which encodes a member of the snRNA activating complex that is required in conjunction with BRF2 to initiate transcription from snRNA promoters (Henry et al, 1998), was also found to be increased in samples with high BRF2 expression. Taken together, our data suggests that BRF2 mediated increase of U6 as well as other splicing regulatory factors may contribute to oncogenesis in SqCC with 8p amplification. Future studies of the role BRF2 overexpression plays in spliceosome function will yield insight into this potential function, and its role in the neoplastic transformation of lung epithelium to SqCC.

7.3 Conclusions

In summary, here we show that the focal amplification of chromosome 8p12, one of the most frequent genetic events in non-small cell lung cancer (NSCLC), plays a key role in squamous cell lineage specificity of the disease. Through the integration of genetic and gene expression data for >330 clinical tumor specimens in conjunction with function cell model studies, we
identified \textit{BRF2} as the target of this amplification and a lineage specific oncogene, the only such oncogene described for lung SqCC to date. In addition, we highlight the oncogenic potential of \textit{BRF2} for the first time and associate its activation with increased RNA processing and resultant cell growth potential. The lineage dependence model suggests that cancer cells rely on the constitutive activation of lineage regulating genes involved in normal development for their continued survival and proliferation (Garraway \& Sellers, 2006). \textit{BRF2} is unique in that it is not a prototypical lineage-specific oncogene as no role in normal lineage development has been established. These data suggest that lineage-specific oncogenes may span numerous biological functions, and they are not limited only to the established class of transcription factors (lineage survival oncogenes) discovered to date but also a class of genes selected in tumorigenesis in a cell lineage specific manner. These results combined with the recent discovery of \textit{TITF1} (thyroid transcription factor 1) as a lineage survival oncogene amplified in lung AC (Kwei \textit{et al}, 2008; Tanaka \textit{et al}, 2007; Weir \textit{et al}, 2007) suggests that the genes required to initiate tumorigenesis in distinct biological contexts may shape the preferential selection of amplifications and resulting phenotypes specific to different cancers, highlighting the opportunity for treatment design targeting specific cell type.

\section*{7.4 Materials and Methods}

\subsection*{7.4.1 DNA samples}

Formalin-fixed, paraffin embedded and fresh-frozen tissues were collected from St. Paul’s Hospital, Vancouver General Hospital and Princess Margret Hospital following approval by the Research Ethics Boards. Formalin fixed paraffin embedded lung CIS samples were collected by fluorescence bronchoscopy directed biopsies at the British Columbia Cancer Agency. 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 (Garnis \textit{et al}, 2005).

\subsection*{7.4.2 Tiling path array comparative genomic hybridization}

Array hybridization was performed as previously described (Baldwin \textit{et al}, 2005; Coe \textit{et al}, 2006; Lockwood \textit{et al}, 2007). Briefly, equal amounts (200-400 ng) of sample and single male reference genomic DNA were differentially labeled and hybridized to SMRT array v.2 (BCCRC Array Laboratory, Vancouver, BC) previously described to give optimal genome coverage (Ishkanian \textit{et al}, 2004; Watson \textit{et al}, 2007).
Hybridized arrays were imaged using a charge-coupled device (CCD) camera system and analyzed using SoftWoRx Tracker Spot Analysis software (Applied Precision, Issaquah, WA). Systematic biases were removed from all array data files using a stepwise normalization procedure as previously described (Khojasteh et al., 2005; Lockwood et al., 2008). SeeGH software was used to combine replicates and visualize all data as log2 ratio plots (Chi et al., 2004; Chi et al., 2008). Stringently, all replicate spots with a standard deviation above 0.075 or signal to noise ratios below three were removed from further analysis. The clones were then positioned based on the human March 2006 (hg18) genome assembly. Genomic imbalances (gains and losses) within each sample were identified using aCGH-Smooth (Jong et al., 2004) with lambda and breakpoint per chromosome settings at 6.75 and 100, respectively as previously described (Coe et al., 2006). The resulting frequency of alteration was then determined for each lung cancer cell type as described previously (Coe et al., 2006). High level amplifications were determined using an algorithm previously described with the log2 threshold set at >0.6 for tumors and >0.8 for CIS cases (due to different levels of cell heterogeneity) (Lockwood et al., 2008). Stringently, regions were only scored as amplified if 2 or more consecutive array elements met this criteria.

### 7.4.3 Comparison of cell type 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 clones within 10 Mb. Probes were then aggregated into genomic regions if the similarity in copy number status between adjacent clones was at least 90% across all samples from the same cell type. 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 clone (Coe et al., 2006). A 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.
7.4.4 Gene expression microarray analysis of clinical tumor specimens

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., Mississauga, ON). Samples along with universal reference RNA were labeled and hybridized to a custom Agilent Whole Genome Oligonucleotide microarray, containing 39,909 probes mapping to ~22,000 unique genes, according to the manufacture’s protocols (Agilent Technologies, Santa Clara, CA). The resulting expression data were processed and normalized using Rosetta Resolver software (Rosetta Inpharmatics, Seattle, WA). Affymetrix U133 Plus 2 expression data for NSCLC tumors were downloaded from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/, accession number GSE3141) and normalized using Microarray Suite (MAS) 5.0 (Bild et al., 2006). A summary containing the number of samples analyzed and corresponding platform is presented in Supplemental Table 1.

7.4.5 Gene expression microarray analysis of normal bronchial epithelial cells

RNA was obtained from exfoliated bronchial cells of lung cancer free individuals obtained during fluorescence bronchoscopy (Chari et al., 2007). All individuals were either current or former smokers. Expression profiles were generated for all cases using the Affymetrix U133 Plus 2 platform and normalized using MAS 5.0.

7.4.6 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 cell types on chromosome arm 8p were determined. Comparisons between expression levels for AC and SqCC tumors as well as SqCC tumors and normal bronchial cells were performed using the Mann-Whitney U test and computed with the ranksum function in Matlab. 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. Probes with a corrected p-value ≤ 0.01 were considered significant. If multiple probes mapped to the same gene, the one with the lowest p-value (Agilent data) or with maximum intensity across the data (Affymetrix) was used.

7.4.7 Integration of genetic and gene expression data

To integrate gene expression with copy number data, two methods were used. First, a 10 kb moving average was generated using the normalized log2 array CGH ratios for each sample
with copy number and expression. These values were subsequently standardized using a Z-transformation for each sample throughout the whole genome in order to facilitate better comparisons across the sample set. An average Z-score was then calculated using the values corresponding to the genomic intervals spanning each of the genes of interest on chromosome 8p. Finally, a non-parametric Spearman correlation coefficient was calculated using the Z-scores for copy number and log_{10} ratios for gene expression across all samples of interest. The corresponding \( p \)-value representing the statistical significance of a positive correlation was calculated and a Benjamini-Hochberg multiple hypothesis testing correction applied as described above. For the second method, the copy number status was determined by aCGH-Smooth as described above and mapped to genes of interest from clones using genomic coordinates from the UCSC Genome Browser (hg18). The gene expression levels for all genes were then compared between samples with copy number gain/amplification against samples which were copy number neutral using the Mann Whitney U test (Lockwood \textit{et al}, 2008). An association was deemed significant if the Benjamini-Hochberg corrected \( p \)-value \( \leq 0.05 \) and the median and mean gene expression in the samples with gain/amplification were higher than those samples which were copy number neutral.

7.4.8 Reverse transcriptase polymerase chain reaction analysis of transcription levels in clinical tumor samples

qPCR was performed on SDS7900HT (Applied Biosystems, Foster City, CA) using SYBR Green and the \( \Delta \Delta C_t \) method with \textit{RPS13} expression levels used as reference for normalization. Primers used were: BRF2\_F: GTGAAGCTCCTGGGACTGGAT, BRF2\_R: GTATTTGGCTGGCACAGAAGG, RPS13F: GTTGCTGTTCGAAAGCATCTTG, and RPS13R: AATATCGAGCCAAACGTTGAA. Associations between \textit{BRF2} expression and clinicopathological features were evaluated by the Wilcoxon test. Breakdown of samples used are provided in Table S1.

7.4.9 Cell lines and culture conditions

NSCLC cell lines H520, H1395 and H2347 were purchased from American Type Culture Collection (ATCC). Cells were maintained in RPMI-1640 Medium (Invitrogen, Carlsbad, CA, USA) supplemented with 10% fetal bovine serum (Invitrogen). The HBEC3-KT immortalized normal HBEC line was established by introducing mouse Cdk4 and hTERT into normal HBECs obtained from a 65 year old woman without cancer (Ramirez \textit{et al}, 2004). The HBEC3-KT53 line was established by stably knocking down p53 in the original cell line, HBEC3-KT (Sato \textit{et al},
These two parental lines were used to overexpress BRF2 via the pMSCV vector (see below). All HBEC3 cell lines were cultured in K-SFM (Invitrogen) medium containing 50 ug/ul bovine pituitary extract (Invitrogen) and 5 ng/ul EGF (Invitrogen).

7.4.10 TaqMan analysis of transcription levels in cancer cell lines

Five micrograms of total RNA isolated from cultured cells was converted to cDNA using an ABI High Capacity cDNA Archive Kit (Applied Biosystems). An aliquot of 100 ng of cDNA was used for each real-time PCR reaction. TaqMan (Applied Biosystems) gene expression assays: BRF2 (Hs00217757_m1) and 18s rRNA (Hs99999901_s1) were performed using standard TaqMan reagents and protocols on a Applied Biosystems 7500 Fast Real-Time PCR System (Applied Biosystems). The \( \Delta\Delta Ct \) method was used for expression quantification using the average cycle threshold for 18S rRNA to normalize gene expression levels between samples (Coe et al., 2006). Cycle thresholds for the primers were then compared between the individual cell lines and a pooled normal lung cDNA reference sample generated from Human Lung Total RNA (AM7968, Ambion, Austin, TX) to identify the fold change represented.

7.4.11 Western blot analysis of protein levels

Cells were washed twice with cold PBS and and lysed in the presence of protease inhibitors. Each cleared lysate was diluted and boiled for electrophoresis and transferred to polyvinylidene (PVDF) membrane (Sato et al., 2006). Membranes pre-blocked with 3% bovine serum albumin in PBS with 0.05% Tween-20 (PBST) were incubated with primary antibodies against BRF2 (Abcam, 1:500 dilution) for 1 hour at room temperature. After three washes in PBST, the membranes were incubated with horseradish-peroxidase (HRP) conjugated donkey anti-goat polyclonal antibody (Abcam, 1:2000 dilution) for 45 minutes at room temperature. After three PBST washes, antibody binding was visualized by enhanced chemiluminescence (GE Healthcare, Piscataway, NJ). Subsequently, the bound antibodies were stripped from the membranes with a buffer containing 62.5 mM Tris-HCl, pH 6.7, 2% SDS and \( \beta \)-mercaptoethanol and reprobed with monoclonal antibody to beta-actin (Abcam, 1:6000) to confirm equal sample loading.

7.4.12 RNAi knockdown

ON-TARGETplus SMARTpool siRNAs targeting BRF2 along with a negative control (ON-TARGETplus siCONTROL Non-Targeting siRNA pool) were obtained from Dharmacon (Lafayette, CO, USA). All experiments were performed in triplicate. Cells were subcultured at a
ratio of 1:3 or 1:6 using 0.25% trypsin-EDTA (Gibco). Transfection efficiency was optimized using siGLO Green Transfection Indicator (Dharmacon). For the transfection experiments, cells were seeded in 24-well culture plates at 60,000 cells/mL 24 hours before transfection. Cells were transfected at a final concentration of 100nM siRNA using Lipofectamine RNAiMAX (Invitrogen) according to the manufacturer’s instructions. The cells were then incubated at 37°C for 24 hours before RNA analysis, 48 hours for protein and 72 hours for MTT assays. BRF2 expression levels for multiple independent knockdowns were determined by qPCR as described above and scaled relative to the average of the control siRNA treated cells (±SEM). Statistical significance was determined using the Student’s T-test in Matlab and p ≤0.05 was considered significant.

7.4.13 3-[4, 5-dimethylthiazol-2-yl]-2, 5-diphenyltetrazolium bromide (MTT) assay

MTT assay (Trevigen, Gaithersburg, MD) was used to determine the status of cell proliferation in siRNA experiments according to the manufacturer’s instructions. Briefly, exponentially growing cells were diluted to a concentration of 313,000 cells/mL in RPMI-1640 with 10% FBS, seeded in triplicate in 96-well plates and incubated at 37°C for 4 hours. The cells were then treated with 10 μL of MTT reagent for 4 hours before adding 100 μL of detergent reagent to solubilize the formazan precipitate. The reaction product was then quantified by measuring absorbance at 570 nm with reference to 650 nm. The mean ±SEM absorbance for experiments from three independent transfections were normalized to the average of the respective controls and analyzed for statistical significance using the Student’s t-Test in Matlab.

7.4.14 Construction of the BRF2 expression vector

The BRF2 sequence from the pBRF2-HORF construct (Invitrogen) was cloned into the retroviral vector pMSCV-hygro (Clontech, Mountain View, CA) and sequenced. This construct was named pMSCV-BRF2. The pMSCV-BRF2 construct and the vector alone (pMSCV) were then transfected into the PheonixTM Ampho retroviral packaging cell line (Orbigen, San Diego, CA) according to manufacturer’s protocols. Subsequent infections into HBEC3-KT and HBEC3-KT53 were performed and plasmid containing cells were selected by treating with 20 μg/ul of hygromycin for 10 days. This resulted in the generation of four stable cell lines: HBEC3-KT-pMSCV-BRF2, HBEC3-KT-pMSCV (vector control), HBEC3-KT53-pMSCV-BRF2, and HBEC3-KT53-pMSCV (vector control).
7.4.15 In vitro cell growth assays

Growth curves were determined for each of the six HBEC cell lines by culturing 1000 cells in triplicate in 12 well plates and counting on the 3rd, 6th, 8th and 10th day. The average ±SEM for each line is reported. Experiments were performed in two or more times and a representative experiment is shown. $P$-values were calculated using the Student’s t-Test when comparing two conditions and ANOVA when comparing three. All calculations were performed in Matlab.

7.4.16 Immunohistochemistry

Slides were deparaffinized using xylene and rehydrated through an ethanol series to water. Antigen retrieval was performed using a decloaking chamber exposure at 15 psi for 20 minutes in sodium citrate buffer (pH 6.0). Endogenous peroxidase enzyme activity was blocked using 3% H2O2 in methanol for 30 minutes at room temperature. Slides were washed in 1% PBS and then blocked using 10% skim milk for 6 hours at room temperature. Slides were incubated for 16 hours at 4°C with a 1:200 dilution of goat polyclonal anti-BRF2 primary antibody (Abcam) followed by incubation with a donkey anti-goat biotinylated secondary antibody (Santa Cruz Biotechnology, Santa Cruz, CA). Normal goat IgG was used as negative control (Santa Cruz Biotechnology). Detection was accomplished using DAB (ImmunoCruz staining system, Santa Cruz Biotechnology). Slides were then counterstained using hematoxylin and the area within the diagnostic area was scored by three independent observers based on the following criteria: 0 = no positive staining, 1 = 25% positive cells, 2 = 50% positive cells, 3 = 75% positive cells and 4 = 100% positive cells. Conflicting scores were resolved by choosing the value consistent between two observers or the average of all three varying scores.

7.4.17 Significance analysis of microarrays (SAM)

Using the 111 NSCLC samples in the dataset by Bild et al. (Table S1), samples were sorted by highest to lowest expression for BRF2 based on the probe with the highest average intensity across the dataset (Bild et al, 2006). Differential gene expression analysis using SAM (Tusher et al, 2001) was performed using the ten samples with highest BRF2 expression against the ten samples with lowest BRF2 expression. A $q$-value threshold of ≤ 0.05 was used to identify differentially expressed genes associated with high BRF2 expression.
7.4.18 Functional assessment of BRF2 associated genes

Functional analyses were generated through the use of Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com) as previously described (Lockwood et al, 2008). Functional Analysis identified the biological functions that were most significant to the data set. Fisher’s exact test was used to calculate a p-value determining the probability that each biological function assigned to the data set is due to chance alone.
Figure 7.1. Chromosome 8p amplification in NSCLC is restricted to the SqCC lineage.

a, Frequency of gain/amplification along chromosome arm 8p is depicted for 103 AC (red) and 58 SqCC clinical tumor specimens (blue). b, The significance of copy number disparity (inverse p-value) between AC and SqCC cell type groups is depicted for 8p. Solid black lines represent regions considered statistically different (p ≤0.01) whereas dashed lines are not. c, Relative expression for genes within regions of copy number difference which were also expressed at significantly higher levels in SqCC (n=13) compared to AC (n=34) tumors (p ≤0.01). The color scale ranges from black (low expression) to white (high expression).
Figure 7.2. BRF2 is a lineage specific oncogene targeted by amplification in SqCC. 

a. Comparison of BRF2 mRNA expression values for AC (n=34) and SqCC (n=13) tumors (p = 0.0056). Box-plots depict the median group expression (red line), the 25th and 75th percentiles (blue box) and the limits of 95% of samples for each group (outside lines) with values for all other samples represented by red crosses. Expression values for all plots are in arbitrary Log10 units.

b. Spearman’s correlation of Z-transformed array CGH copy number ratios and expression values for BRF2 in 13 SqCC tumors (correlation coefficient = 0.87). Each diamond represents an individual sample.

c. Comparison of BRF2 expression between SqCC tumors with neutral copy number status (n=4) and SqCC tumors with gain/amplification (n=6) (p=0.048).

d. Difference in BRF2 expression levels between 67 exfoliated bronchial cell samples from cancer-free patients and 53 SqCC tumors from an independent sample set (p<1.0 x 10^-8).
Figure 7.3. BRF2 activation contributes to cell growth and proliferation. a, Concordance between BRF2 copy number (array CGH), expression (qPCR) and protein (immunoblot) levels in H520, H1395 and H2347 NSCLC cell lines. b, Decrease in BRF2 mRNA levels in H520 cells transfected with BRF2 siRNA relative to those treated with non-targeting control siRNA (mean ± standard error measure (SEM) of two independent experiments; p=0.045). c, Decrease in proliferation of H520 cells with BRF2 siRNA transfection relative to non-targeting siRNA transfected controls as determined by the MTT assay (mean ± SEM of three independent experiments; p=0.024). Increased saturation density in both d, BRF2 expressing HBEC and e, BRF2 and p53RNAi expressing HBEC compared to their respective controls (mean ± SEM of triplicate samples; p=0.035 and 7.67 x 10^{-4} at day 10 respectively). * p<0.05, ** p<0.001.
Figure 7.4. Amplification and overexpression of BRF2 in preinvasive SqCC lesions. a, Frequent copy number increase of chromosome arm 8p in 20 bronchial CIS lesions. Samples are ordered in columns and ordered by genomic position along 8p. The color scale ranges from white (neutral copy number, N) to black (amplification, Amp). Data from representative normal lung (N) and SqCC tumor samples (T) are displayed to the left and right of the CIS cases respectively. b, Amplification score along chromosome 8p for the 20 CIS cases. Regions of amplification were defined for each case and summarized across the group to determine the incidence of occurrence. Dashed lines represent the positions of BRF2, WHSC1L1 and FGFR1 from top to bottom respectively. c, Array CGH copy number profiles for two individual CIS cases with 8p amplification. Each black dot represents an array element ordered by genomic position. Those shifted to the left of the middle line (N) have decreased copy number (Del) whereas those shifted to the right have increased copy number (Amp). Dashed lines represent the positions of the three genes as in c. The region highlighted in orange represents the region of high level amplification in each sample. The amplicon in CIS1 includes only BRF2 with WHSC1L1 and FGFR1 outside or spanning the boundaries while the amplicon in CIS2 contains all three genes. d, Immunostaining of CIS2 with anti-BRF2 polyclonal antibody revealed elevated staining in CIS epithelia compared with normal from the same tissue section.
Figure 7.5. BRF2 expression in SqCC precancerous stages. Immunostaining of 21 lung SqCC precursor lesions with anti-BRF2 polyclonal antibody revealed a monotonic increase in BRF2 expression with increasing histopathology grade. The area within the diagnostic area was scored as follows: 0 = no positive staining, 1 = 25% positive cells, 2 = 50% positive cells, 3 = 75% positive cells and 4 = 100% positive cells. Each sample is represented by a single dot above its corresponding grade with the horizontal black lines representing the median IHC score for each grade. Red samples highlight multiple grades taken from the same individual patient (see text).
7.5 References


Chapter 8: Conclusions

Portions of this chapter are excerpts from the abstracts of the manuscripts detailed in chapters 2 to 7.
8.1 Summary

Lung cancer is a devastating disease and leading cause of cancer related mortality worldwide (Parkin et al, 2005). The dismal outcome for lung cancer patients is mainly attributed to the late-stage of disease at the time of diagnosis as well as the lack of effective treatment strategies (Sato et al, 2007). Understanding the molecular mechanisms driving lung cancer tumorigenesis will lead to rational development of diagnostics and therapeutics based on the biology of disease.

Lung cancer is thought to result from the sequential accumulation of somatic DNA alterations that begin in normal epithelium and increase in severity during cancer progression, consistent with the multistep model of carcinogenesis (Wistuba, 2007). These alterations cause the activation of oncogenes and inactivation of tumor suppressor genes, leading to the deregulation of fundamental cellular processes which confer malignant growth (Sekido et al, 2003). However, the discovery of such genetic alterations has been limited due to the technical limitations of conventional cytogenetic analysis approaches (Balsara & Testa, 2002). In addition, as not all gene expression changes observed in a tumor are causal to cancer development, global gene expression analysis alone cannot distinguish between causal and reactive changes (Coe et al, 2008). Thus, an integrative genomics approach examining genetic events in conjunction with the changes in gene expression pattern should improve the identification of causal changes that lead to disease phenotype. For this reason, the initial chapters of this thesis describe the development and application of analysis methods for the integration of genome-wide expression and copy number data using newly developed high resolution microarray platforms. Using these approaches, I demonstrated that integrative analysis can uncover pathologically related genes in lung cancer, validating hypothesis 1.

8.1.1 Development and application of integrative genomic approaches for the study of lung cancer

Chromosomal regions harbouring tumor suppressors and oncogenes are often deleted or amplified (Albertson et al, 2003). Array comparative genomic hybridization detects segmental DNA copy number alterations in tumor DNA relative to a normal control (Lockwood et al, 2006). The recent development of a bacterial artificial chromosome array, which spans the human genome in a tiling path manner with >32,000 clones, facilitated whole genome profiling at an unprecedented resolution (Ishkanian et al, 2004). In chapter 2, I comprehensively described and compared the genomes of 28 commonly used non-small cell lung carcinoma (NSCLC) cell
models, derived from 18 adenocarcinomas (AC), 9 squamous cell carcinomas and 1 large cell carcinoma using this technology (Garnis et al, 2006). Analysis at such resolution not only provided a detailed genomic alteration template for each of these model cell lines, but revealed novel regions of frequent duplication and deletion. Significantly, a detailed analysis of chromosome 7 identified 6 distinct regions of alterations across this chromosome, implicating the presence of multiple novel oncogene loci on this chromosome. As well, a comparison between the squamous and AC cells revealed alterations common to both subtypes, such as the loss of 3p and gain of 5p, in addition to multiple hotspots more frequently associated with only 1 subtype. Interestingly, chromosome 3q, which is known to be amplified in both subtypes, showed 2 distinct regions of alteration, 1 frequently altered in squamous and 1 more frequently altered in AC. This data demonstrated the unique information generated by high resolution analysis of NSCLC genomes and uncovered the presence of genetic alterations prevalent in the different NSCLC subtypes.

Subsequent genome studies by other groups were undertaken in both cell models of NSCLC and clinical samples to identify alterations underlying disease behaviour, and many also identified recurring aberrations of chromosome 7, supporting the findings from chapter 2 (Weir et al, 2007). The presence of recurring chromosome 7 alterations that did not span the well-studied oncogenes \textit{EGFR} (at 7p11.2) and \textit{MET} (at 7q31.2) raised the hypothesis of additional genes on this chromosome that contribute to tumourigenesis. In chapter 3, I demonstrated that multiple loci on chromosome 7 are indeed amplified in NSCLC, and through integrative analysis of gene dosage alterations and parallel gene expression changes, identified new lung cancer oncogene candidates, including \textit{FTSJ2}, \textit{NUDT1}, \textit{TAF6}, and \textit{POLR2J} (Campbell et al, 2008). Activation of these key genes was confirmed in panels of clinical lung tumour tissue as compared with matched normal lung tissue. The detection of gene activation in multiple cohorts of samples strongly supported the presence of key genes involved in lung cancer that are distinct from the \textit{EGFR} and \textit{MET} loci on chromosome 7.

The identification of novel regions of DNA amplification suggested that additional oncogene candidates that may be involved in lung cancer development could be uncovered using an integrative genomic approach. Therefore, in chapter 4, I answered the fundamental question of the contribution of DNA amplification as a molecular mechanism driving oncogenesis (Lockwood et al, 2008). Comparing 104 cancer lines representing diverse tissue origins identified genes residing in amplification 'hotspots' and discovered an unexpected frequency of genes activated by this mechanism. The 3431 amplicons identified represent approximately 10
per hematological and approximately 36 per epithelial cancer genome. Many recurrently amplified oncogenes were previously known to be activated by disease-specific translocations only. The 135 hotspots identified contained 538 unique genes and were enriched for proliferation, apoptosis and lineage-dependency genes, reflecting functions advantageous to tumor growth. Integrating gene dosage with expression data validated the downstream impact of the novel amplification events in both cell lines and clinical samples. Importantly, this analysis discovered that multiple downstream components of the EGFR-family-signaling pathway, including CDK5, AKT1 and SHC1, are overexpressed as a direct result of gene amplification in lung cancer. These findings suggested that amplification is a far more common mechanism of oncogene activation than previously believed and that specific regions of the lung cancer genome are hotspots of amplification.

8.1.2 Comparison of lung cancer subtypes

Lung cancer is not a homogeneous entity but a collection of phenotypically diverse and regionally distinct neoplasias (Giangreco et al., 2007). The histological heterogeneity of lung cancer likely reflects differences in cell derivation, genetic alterations and pathogenetic pathways. These fundamental discrepancies in tumor biology may be a primary factor determining the poor outcomes of lung cancer patients as biological differences that segregate with the subtypes may also lead to differences in response to therapies (Garraway & Sellers, 2006). Therefore, after developing methods to discover causal genetic alterations in lung cancer, I next aimed to utilize these approaches to uncover the genetic alterations and corresponding genes responsible for the differential development of lung cancer subtypes, addressing hypothesis 2. This was the focus of chapters 5 to 7.

SCLC and NSCLC comprise the two major cell types of lung cancer. Although these cell types can be distinguished readily at the histological level, knowledge of their underlying molecular differences is very limited. In chapter 5, I compared 14 SCLC cell lines against 27 NSCLC cell lines using an integrated array comparative genomic hybridization and gene expression profiling approach to identify subtype-specific disruptions (Coe et al., 2006). Using stringent criteria, I identified 159 of the genes that are responsible for the different biology of these cell types. Sorting of these genes by their biological functions revealed the differential disruption of key components involved in cell cycle pathways. This novel comparative combined genome and transcriptome analysis not only identified differentially altered genes, but also revealed that certain shared pathways are preferentially disrupted at different steps in these cell types. SCLC exhibited increased expression of MRP5, activation of Wnt pathway inhibitors, and upregulation
of p38 MAPK activating genes, while NSCLC showed downregulation of \textit{CDKN2A}, and upregulation of \textit{MAPK9} and \textit{EGFR}. This information suggested that cell cycle upregulation in SCLC and NSCLC occurs through drastically different mechanisms, and highlighted the need for differential molecular targets selection in the treatment of these cancers.

The identification of molecular differences between SCLC and NSCLC demonstrated that distinct patterns of genetic alteration are involved in the development of lung cancer subtypes. Thus, I next aimed to adapt this approach to the analysis of clinical tumor samples. I focused this analysis on the comparison of AC and SqCC tumors since they are the most common types of lung cancer and are typically regarded as a single disease entity in terms of therapy.

Although previous studies suggested that distinct patterns of genomic alteration exist for AC and SqCC, the specific genes responsible for the different tumor phenotypes remained largely unknown (Balsara & Testa, 2002). Initial gene expression profiling studies yielded some insight into the tumor subtypes and were able to segregate tumors into histologic groupings based on multi-gene models (Bhattacharjee \textit{et al}, 2001; Thomas \textit{et al}, 2006). However, as stated above, not all gene expression changes are causal to disease development. Thus, in chapter 6, I performed a large-scale integrative analysis of 271 NSCLC primary tumors (179 AC and 92 SqCC) using high resolution array comparative genomic hybridization (CGH) coupled with gene expression microarray analysis. This analysis uncovered 259 regions of significant copy number disparity between the subtypes, 167 which were SqCC specific and 86 that were AC specific. Analysis of expression data for the genes in these regions identified 378 SqCC specific and 76 AC specific genes differentially deregulated as a result of these copy number changes. Principle component analysis in three independent datasets confirmed that these gene signatures were able to accurately delineate the disease subtypes demonstrating their contribution to the different tumor phenotypes. Furthermore, grouping of these genes by biological interactions identified gene networks associated with the development of AC and SqCC. Strikingly, SqCC exhibited frequent disruption of histone modification enzymes and genes involved in the RB-Cyclin pathway whereas AC displayed the downregulation of genes targeted by TGFβ, p53 and HNFα transcription factors. In addition, numerous genes which had previously been implicated in NSCLC development, prognosis and response to therapy were found to be restricted in their disruption patterns to individual cancer subtypes. For example, \textit{ERCC1}, which is known to influence lung cancer risk and response to cisplatin based therapies, was shown to be inactivated preferentially in AC. Together, this data demonstrated that AC and
SqCC develop through different genetic pathways and discovered the specific target genes of these alterations for the first time.

The results from chapters 5 and 6 suggested that the preferential selection of genetic alterations drive tumorigenesis in specific cell lineages giving rise to lung cancer subtypes. In chapter 7, I further characterized one of the most prominent alterations and discovered a cell lineage specific genetic event that may provide novel target for new treatment strategies. Through integrative genetic analyses of multiple independent cohorts of clinical tumor samples, I identified the overexpression of \textit{BRF2}, a RNA polymerase III (Pol III) transcription initiation factor, as the result of increased gene dosage specifically in the squamous cell lineage leading to SqCC. Ectopic expression of \textit{BRF2} in human bronchial epithelial cells induced a transformed phenotype and demonstrated downstream oncogenic effects, while siRNA mediated knockdown suppressed growth of cells overexpressing \textit{BRF2}. Frequent activation of \textit{BRF2} in pre-invasive bronchial carcinoma in situ and dysplastic lesions provided evidence that \textit{BRF2} expression is an early event in cancer development of this cell lineage. In addition, this gene was part of the top SqCC gene network identified in chapter 6, further highlighting its importance. This data suggested that genetic alteration of \textit{BRF2} represents a novel mechanism of lung tumorigenesis through the increase of Pol III mediated transcription in SqCC.

\textbf{8.2 Significance and Conclusions}

I have demonstrated that an integrative genomics approach utilizing high resolution microarray platforms is a powerful tool to uncover key gene disruptions in lung cancer. By combining expression and copy number data, the complexity of the cancer genome can be reduced, allowing the discovery of causal gene alterations which have escaped detection with conventional genomics technologies. Application of these approaches has yielded numerous insights into lung cancer biology which will have significant impact on treatment strategies and the clinical management of this disease.

\textbf{8.2.1 Novel genetic alterations and candidate genes involved in lung tumorigenesis}

Although a subset of genes involved in lung cancer development are well characterized (e.g. \textit{EGFR} and \textit{KRAS}), the vast majority responsible for tumorigenesis remain largely unknown. Thus, one of the most important findings from this thesis is the discovery of novel candidate gene and pathway disruptions which are potentially involved in lung cancer development. For example, through the comprehensive profiling of amplifications in lung cancer, I (1) showed that
DNA amplification is a common mechanism in cancer gene activation, and (2) discovered novel oncogenes amplified in NSCLC including multiple downstream components of the EGFR-family-signalling pathway (chapter 4). Since response to EGFR based therapies cannot be fully explained by receptor status alone, the activation of alternative genes in this pathway may not only represent additional routes to tumor development, but also have a drastic impact on patient response to therapy.

Importantly, I discovered novel target genes with less defined roles in tumorigenesis which are frequently deregulated in lung cancer. For instance, multiple candidate oncogenes aside from the known targets *EGFR* and *MET* were identified in recurrently altered regions on chromosome 7 in NSCLC (chapters 2 and 3). While some of these genes have previously characterized biological functions which implicate a potential role in cancer development (including *FTSJ2, NUDT1, TAF6*, and *POLR2J*), the cellular function of others such as *ECOP, TSC22D4* and *MOSPD* have not been determined. Since these genes are disrupted in almost 100% of NSCLCs, elucidating their specific biological functions may reveal unique mechanisms of oncogenesis, opening exciting new avenues for cancer treatment.

Significantly, seeing as the genes identified in this thesis are deregulated as a direct result of changes in gene dosage, they are likely the primary genetic events responsible for tumor formation (Coe *et al.*, 2008). This makes them ideal targets for therapeutic intervention due to the direct nature of their activation/inactivation, as compared to genes that may be deregulated by complex trans-regulation networks. The previous success of drugs targeting the alterations driving the malignant phenotype of lung cancer such as TKIs for EGFR suggests that these discoveries may be of great clinical application. Thus, the specific genes identified in this study may provide the basis for the subsequent development of tailored therapeutic approaches to combat this devastating disease (see section 8.3.2).

**8.2.2 Genetic mechanisms involved in the development of lung cancer subtypes**

Similar to the beginning chapters of this thesis, the majority of high throughput studies on lung cancer have been primarily focused on the identification of frequently disrupted genes. As such, little emphasis was placed on delineating the molecular mechanisms underlying the development of different lung cancer subtypes. Therefore, the most important finding of this thesis was that lung cancer subtypes require distinct genetic alterations for tumorigenesis (chapters 5 to 7).
Although the clinical distinction between SCLC and NSCLC was well established, little was known about the underlying molecular differences responsible for the different NSCLC subtypes, mainly AC and SqCC, which represent the greatest proportion of all lung cancer cases. The findings from chapters 2-4 suggested that AC and SqCC have common genes involved in neoplastic development; for example, the analysis of chromosome 7 identified genes including \( \text{EGFR} \) which are frequently activated in both subtypes. However, through the development of approaches to directly compare subtypes of lung cancer, I found that the AC and SqCC also contain distinct gene and pathway disruptions responsible for their different phenotypes. For instance, SqCC exhibited frequent disruption of histone modification enzymes and genes involved in the RB-Cyclin pathway whereas AC displayed the downregulation of genes targeted by TGF\( \beta \), p53 and HNF\( \alpha \) transcription factors. Thus, although the subtypes may have similar signalling status for some pathways, others are deregulated specifically in each, indicative of significantly different mechanisms of tumor initiation and progression.

Importantly, further characterization of one of the most prominent genetic differences between these tumor types (gain of chromosome arm 8p) identified a novel SqCC lineage specific oncogene, called \( \text{BRF2} \), which affects polymerase III transcribed genes. It has recently been shown that increased Pol III activity and corresponding increase in its transcribed products is a common event in cellular transformation. However, the manner in which Pol III becomes deregulated and whether or not this is causal to cancer development remains unclear. Previous studies have shown that activation of oncogenes such as MYC and RAS and inactivation of tumor suppressor genes such as p53 and RB can lead to an increase in Pol III activity in cancer (White, 2008). Interestingly, these genes are commonly disrupted in lung cancer (Sato et al, 2007). The findings of chapter 7 suggest for the first time that the direct genetic activation of Pol III transcription factor subunits may be an alternate mechanism leading to increased Pol III activity (Marshall & White, 2008). In addition, the activation of \( \text{BRF2} \) in precursor lesions and its ability to increase cell proliferation in \textit{vitro} indicate that this phenomena may be causal in SqCC development. Since SqCC grows more rapidly than AC, activation of this gene may play a major role in distinguishing the different disease phenotypes. This data indicates that activation of \( \text{BRF2} \) represents a novel mechanism of lung SqCC tumorigenesis through the increase of Pol III mediated transcription.

### 8.2.3 Developing new therapeutic strategies for lung cancer treatment

Since AC and SqCC are typically treated as a single disease entity, the discovery that different genes and genetic alterations are involved in tumor development will have a significant impact
on disease management. In terms of the immediate future, the demonstration that genes involved in response to conventional chemotherapeutics are differentially dysregulated between AC and SqCC may guide the selection of treatment strategies that lead to better patient response to therapy and improve survival rates. For example, the cancer cell lineages displayed differential deregulation of the cisplatin resistance associated genes \textit{ABCC5} and \textit{ERCC1} which may influence response to treatment with this drug. Activation of \textit{ABCC5} in SqCC suggests that these tumors may be more resistant to cisplatin based therapies whereas the frequent inactivation of \textit{ERCC1} in AC implies the opposite scenario for this subtype. Since cisplatin forms the backbone of standard NSCLC chemotherapeutic strategies, the subtype specific disruption of these well established response genes means a rethinking of current treatment approaches is necessary (Sato \textit{et al}, 2007). In addition, this information suggests that better gene signatures for the prediction of disease recurrence and markers for early detection may be achieved though the analysis of AC and SqCC separately. Thus, this information holds the potential to drastically influence future clinical trials and research studies on lung cancer.

Furthermore, my findings will be particularly important in the concept of developing new targeted therapies for lung cancer patients as the distinct mechanisms of tumorigenesis in the different cell lineages suggests that the conventional "one size fits all" approach to therapeutic intervention may not succeed. Although some genes and pathways may be commonly disrupted across the subtypes, the exclusive deregulation of others may lead to major variations in response to therapies. For example, the disruption of multiple histone modification enzymes specifically in SqCC suggests that these tumors may benefit from new targeted therapies aimed at inhibiting these molecules. The recent development of several histone deacetylase (HDAC) inhibitors, which reverse gene silencing and exert antiproliferative effects by upregulating the expression of tumor suppressor genes, highlight the potential effectiveness of such treatments (Sun \textit{et al}, 2007b). Since these and other histone modifying therapies are currently being considered for clinical trials in lung cancer, subtype specificity should be a vital consideration in study design.

Taken together, this work highlights the need for tailoring therapies to the specific lung cancer subtypes and offers candidate genes and pathways for this purpose. For example, targeted activation of \textit{BRF2} in lung SqCC suggests that it may be an excellent candidate for new treatment strategies and early diagnostic methods tailored to this NSCLC subtype (see section 8.3.1 for more detail). Although the subject of future work, the approaches developed will also be essential to further characterize a comparable AC specific oncogene for treatment of this
Overall, these results confirm at the molecular level that these lung cancer subtypes are distinct disease entities and should be studied separately when designing treatment strategies and testing new drugs in clinical trials.

8.3 Future Directions

Through the course of addressing the hypotheses of this thesis, numerous additional questions and avenues of research have been generated. Future work investigating these issues will be imperative to further define mechanisms of lung cancer subtype development and assess the use of specific genes and pathways for therapeutic intervention.

8.3.1 Mechanism of BRF2 mediated tumorigenesis and assessment of Pol III as a therapeutic target in lung cancer

The immediate future plans are to further characterize the role of BRF2 in SqCC tumorigenesis. A recent study has demonstrated that overexpression of another Pol III transcription factor subunit, BRF1, can increase cellular proliferation and induce oncogenic transformation in vitro and form tumors in vivo (Marshall & White, 2008). Interestingly, this study showed that the tumorigenic effects of BRF1 could be partially attributed to the increase in one of its Pol III-transcribed targets, tRNA<sub>MET</sub> (the tRNA required for initiating polypeptide synthesis) (Marshall & White, 2008). Thus, it will be interesting to assess the specific BRF2-mediated Pol III transcript levels in tumors and cell lines and determine how these products may act oncogenically. One specific transcript of interest is U6 snRNA. As described in chapter 7, this snRNA plays a vital role in the splicing of mRNA, a fundamental process involved in protein synthesis and thus, the capacity of a cell to grow (White, 2005). However, if the increase in BRF2 drives higher levels of U6, an additional scenario could arise which may also contribute to tumorigenesis. Through alternate splicing, a single gene can generate multiple transcripts from a common mRNA precursor, leading to protein diversity (Pajares et al., 2007). This process is tightly regulated to ensure appropriate splice forms are expressed according to cell type, developmental stage and response to external stimuli (Pajares et al., 2007). Aberrant splicing has been associated with various diseases including cancer where inappropriate splicing generates protein isoforms offering a selective advantage that contribute to tumor initiation. It has previously been reported that alterations in the cellular-splicing-regulatory machinery results in changes in the splicing patterns of many cancer related genes (Pajares et al., 2007). Since U6 is a key catalytic component of the spliceosome, increased levels of this snRNA through the activation of BRF2 could potentially result in hyperactive splicing and the production of oncogenic proteins not
usually present in normal precursor cells. This is particularly interesting as activation of *BRF2* is an early event in SqCC tumorigenesis, suggesting this may be a mechanism of SqCC initiation. Therefore, although I have demonstrated that BRF2 activation can lead to an increase in cell growth, it will be interesting to investigate additional roles this gene may play in cancer development. After this is determined, methods for therapeutic intervention involving this gene can be surveyed.

More fundamentally, it will also be interesting to further characterize the extent of Pol III deregulation in lung cancer. Determining additional mechanisms of activation in tumors without *BRF2* amplification may be one avenue of investigation. This will likely be facilitated by examining the interaction of *BRF2* with known tumor suppressor genes and oncogenes and assessing the alteration status of other components involved in Pol III mediated transcription. In addition, establishing whether or not Pol III deregulation is a SqCC specific phenomena or involved in additional subtypes such as AC is an important consideration. Together, this information will be essential in order to determine how Pol III can be exploited for patient benefit.

### 8.3.2 Functional and clinical characterization of gene candidates

On a similar note, it will also be a priority to explore the role of other genes and pathways that were uncovered during the course of this thesis. For example, what selective advantages do amplified genes confer when activated? What is the relationship of gene disruption with clinical characteristics such as response to therapy? As with *BRF2*, questions such as these need to be answered before specific gene targets can be considered for translation to the clinic and has already been initiated for some candidates. For example, the finding that numerous downstream components of the EGFR-pathway were activated by genomic amplification spurred a follow-up study by our group involving the comprehensive investigation of alterations in EGFR-pathway components and their relationship to EGFR tyrosine kinase inhibitor sensitivity (Gandhi *et al.*, 2009). Studies such as these are currently being considered for other candidates particularly those displaying subtype specific disruption patterns.

### 8.3.3 Validating the lineage specific tumorigenic potential of subtype specific genes

Longer term work would focus on further examining the hypothesis that the subtype specific genes identified in this study are involved in the malignant transformation of specific cell lineages. This will require numerous functional studies in vitro and in vivo involving the disruption of gene candidates in different cell types and assaying the results. Additionally, it will be essential to further implicate the specific targets responsible for the phenotypes of the
different cancer histologies. This may be aided by the comparison of the subtypes to the same cancer lineages from different organs. For example, recent results have shown that chromosomal amplification at 3q is common to multiple human cancers, but has a specific predilection for SqCCs of mucosal origin (Sarkaria et al, 2006). Furthermore, fine mapping this alteration has identified candidate genes which appear to be involved in tumorigenesis of squamous cell lineage (Sarkaria et al, 2006). Thus, comparison across different tissue types may allow the further refinement of phenotype specific genes by cancelling out tissue specific patterns of alteration. Conversely, it may allow the determination of how similar the genomes of cancer lineages from different organs are and whether unique therapeutic strategies may be required for each tissue site.

8.3.4 Refinement of lung cancer subtypes

In addition to identifying alterations specific to the main histological subtypes of lung cancer, further refinement of subgroups may reveal other important findings. For example, recent evidence has emerged that cancers from smokers are different at the molecular level than those from never smokers (Sun et al, 2007a). The predominant form of lung cancer among never smokers is AC and studies have revealed that EGFR mutations are common in these cancers whereas KRAS mutations are common in AC from ever smokers (Sun et al, 2007a). Furthermore, AC contains numerous histological sub-features which may indicate different molecular mechanisms underlying development. This has been supported by gene expression studies which showed that AC can be classified into distinct subclasses based on expression signatures (Thomas et al, 2006). Interpretation of the findings from chapters 2 and 6 also suggest that molecular heterogeneity exists with AC. Analysis of recurrent alterations on chromosome 7 (figure 2.6) identified two groups of AC; one with multiple regions of copy number increase and one without these features. Likewise, the analysis in chapter 6 revealed less AC specific targets than SqCC specific targets, suggesting that heterogeneity within this subtype may confound the discovery of subtype specific alterations. Interestingly, different rates of genetic imbalance where not responsible for these observations as ACs were found to contain a similar amount of genetic alterations as SqCCs on average (data not shown). Therefore, although the major findings of this thesis - particularly the discovery of BRKF2 - were mainly related to SqCC, further refining the AC subtype into more specific subgroups may lead to a clearer picture of the genes and pathways involved in the development of this subtype. Nonetheless, the methods devised in this thesis will be integral in this process.
Another important avenue of investigation will be the relationship of developmental signalling pathways to specific lung cancer subtypes. It has been shown that pathways important in determining cell fate decisions including Hedgehog, Notch and Wnt are aberrantly activated in human cancers, including lung (Daniel et al., 2006). Recently, the relationship of these pathways to the development of the specific lung epithelial cell lineages has begun to unravel. For example, Hedgehog signalling has been shown to be essential for the normal development of neuroendocrine cells and is re-activated in SCLC tumorigenesis (Daniel et al., 2006). Likewise, although not conclusively proven, Notch and Wnt signalling appear to be important in other airway types and NSCLCs. Interestingly, these pathways appear to show specificity in their deregulation as activation of Notch signalling is growth inhibitory in SCLC (Daniel et al., 2006). Remarkably, the results from this thesis demonstrated that numerous genes involved in these signalling pathways as well as transcription factors known to be involved in lung development show subtype specificity in their disruption patterns in tumors (Chapter 6). For example, NOTCH3 was activated specifically in SqCC suggesting that the Notch pathway may not only be involved in tumorigenesis of this cell lineage but also the normal development of the precursor cells for this disease (basal cells). Determining the relationship of these pathways with the specific cancer cell lineages will be particularly important, especially considering that recent preclinical studies indicate that pharmacologically targeting these pathways is feasible (Sun et al., 2007b).

Lastly, although this thesis focused on copy number, other mechanisms of gene disruption also exist. Epigenetic factors such as DNA methylation of tumor suppressor genes (covalent modification of DNA affecting transcription factor binding and chromatin structure and accessibility), histone modifications (affecting chromatin structure and accessibility) and micro RNA (direct regulation of specific gene expression) play a significant role in the deregulation of gene expression in tumors (Esteller, 2007; Esteller, 2008). New technology platforms are enabling the detection of changes in miRNA levels, DNA methylation profiles, and histone modifications (Boyle et al., 2008; Esteller, 2008; Lakshmipathy et al., 2007). Multidimensional integrative analysis will provide a more complete picture of causal genetic events, leading to a better understanding of the disruption of gene regulation in cancer. For example, high-throughput approaches have recently been developed to assay DNA methylation in a genome-wide manner (Esteller, 2008). Previous studies have shown variations in the methylation
patterns between lung cancer subtypes, suggesting they may play a vital role in their differential
development and clinical phenotypes (Toyooka et al, 2003). Thus, integration of high
throughput methylation data with the complementary copy number and gene expression
findings from this thesis may further clarify mechanisms of tumorigenesis in each cancer cell
lineage.
8.4 References


