THE ROLE OF SPECIFIC GENOMIC ALTERATIONS IN SMALL CELL LUNG CANCER AGGRESSIVENESS

By

BRADLEY P. COE

B.Sc., The University of British Columbia, 2001

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

DOCTOR OF PHILOSOPY

in

THE FACULTY OF GRADUATE STUDIES
(Pathology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2008

Abstract

Small Cell Lung Cancer (SCLC) is a very aggressive neuroendocrine tumour of the lung, which demonstrates a 5 year survival of only 10% for extensive stage disease (20-30% for limited stage), with only modest improvement over the last few decades. Identification of new molecular diagnostic and therapeutic targets is thus imperative. Previous efforts in identifying molecular changes in SCLC by gene expression profiling using microarrays have facilitated disease classification but yielded very limited information on SCLC biology. Previous DNA studies have been successful in identifying several loci important to SCLC. However the low resolution of conventional chromosomal Comparative Genomic Hybridization (CGH) has limited the findings to large chromosomal regions with only a few specific candidate genes discovered to date. Thus, to further understand the biological behaviour of SCLC, better methods for studying the genomic alterations in SCLC are necessary.

This thesis highlights the development of array CGH technology for the high resolution dissection of aneuploidy in cancer genomes and the application of this new technology to the study of SCLC. I present the development of the first whole genome CGH array which offered unprecedented resolution in the profiling of cancer genomes allowing fine mapping of genes in a single experiment. Through application of DNA based analysis in conjunction with integrated expression analysis and comparison of SCLC to less aggressive non-small cell lung tumours I have identified novel patterns of pathway disruption specific to SCLC. This included alteration to Wnt pathway members and striking patterns of cell cycle activation through predominantly downstream disruption of signalling pathways including direct activation of the E2F transcription factors, which are normally repressed by the Rb gene.

Analysis of targets of the E2F/Rb pathway identified EZH2 as being specifically hyper-activated in SCLC, compared to NSCLC. EZH2 is a polycomb group gene involved in the control of many cellular functions including targeted DNA methylation and escape from senescence in hematopoietic stem cells.

Taken together these results suggest that in SCLC, downstream disruption may replace multiple upstream alterations leading to activation independent of a specific mitogenic pathway, and that EZH2 represents a potentially important therapeutic target.

Table of Contents

Abstract ii

Table of Contents iii

List of Tables viii

List of Figures ix

List of Abbreviations xi

Acknowledgements xiii

Co-Authorship Statement xv

Dedication xvii

Chapter 1: Introduction 1

1.1 Introduction to Lung Cancer 2

1.2 Non Small Cell Lung Cancer 2

1.3 Small Cell Lung Cancer 3

1.3.1 Clinical Features 3

1.3.2 Cell Lines 4

1.3.3 Molecular Biology of SCLC 5

1.3.4 The Need for New Tools to Study SCLC 7

1.4Tools for the Molecular Profiling of Cancer 7

1.4.1 Expression Analysis 7

1.4.2 Array CGH 8

1.5 Thesis Theme 9

1.6 Objectives and Hypothesis 10

1.7 Specific Aims and Thesis Outline 10

1.8 References 15

Chapter 2: A Tiling Resolution DNA Microarray with Complete Coverage of the Human Genome 20

2.1 Introduction 21

2.2 Results 22

2.2.1 Array sensitivity 22

2.2.2 Array resolution compared to conventional CGH 23

2.2.3 Comparison to previous array CGH 23

2.2.5 Identification of minute regions of alteration 24

2.3 Discussion 25
2.4 Methods 25
2.4.1 BAC Clone Selection, Preparation and Validation 25
2.4.2 Array Production from BAC DNA 25
2.4.3 DNA labelling and hybridization 26
2.4.4 Array Imaging and Analysis 26
2.5 References 33
Chapter 3: SeeGH – A software tool for visualization of whole genome array comparative genomic hybridization data. 36
3.1 Introduction 37
3.2 Software Environment and Information Sources 38
3.3 Results and Discussion: 39
3.3.1 Overview of Data Flow 39
3.3.2 Input Requirements 39
3.3.3 Data Filtering and Storage 40
3.3.4 Data Presentation 41
3.3.4.1Genomic View 41
3.3.4.2 Chromosome View 43
3.3.4.3 Accessing Previously Entered Data 44
3.4 Conclusions 44
3.5 Availability and Requirements 45
3.6 References 52
Chapter 4: High-resolution array CGH increases heterogeneity tolerance in the analysis of clinical samples. 53
4.1 Introduction 54
4.2 Results and Discussion 54
4.3 Conclusions 57
4.4 References 62
Chapter 5: Resolving the resolution of array CGH 63
5.1 Introduction 64
5.2 Results and Discussion 64
5.2.1 Alteration Detection is Dependent on Array Element Distribution 64
5.2.2 Evaluation of Practical Performance 66
5.2.3 Mapping of Breakpoints is Dependent on Local Resolution 68

5.2.4 Defining Functional Resolution 69

5.2.5 Sample Considerations 69

5.2.6 Selecting a Platform 70

5.3 Conclusions 71

5.4 Materials and Methods 71

5.4.1 Array Platform Data Sources 71

5.4.2 Implementation of ResCalc 72

5.4.3 Calculation of Optimal Detection Sensitivity 72

5.4.4 Calculation of Practical Detection Sensitivity 73

5.4.5 Calculation of Breakpoint Precision 73

5.5 References 81

Chapter 6: High resolution Chromosome 5p Array CGH Analysis of Small Cell Lung Carcinoma Cell Lines 83

6.1 Introduction 84

6.2 Results and Discussion 85

6.3 References 94

Chapter 7: Gain of a region on 7p22.3, containing MAD1L1, is the most frequent event in small-cell lung cancer cell lines 96

7.1 Introduction 97

7.2 Results and Discussion 98

7.2.1 Generation of High Resolution SMRT array CGH profiles of SCLC Cell Lines 98

7.2.2 Identification of Frequent Alterations 99

7.2.3 Profiling of Matched Normal Cell Lines 101

7.2.4 7p22.3 is the most frequently gained region in SCLC Lines 102

7.3 Conclusions 103

7.4 Materials and Methods 104

7.4.1 Cell Lines 104

7.4.2 SMRT Array CGH 104

7.4.3 Automated Alteration Identification 105

7.4.4 Frequency Plot Application 105

7.4.5 Fluorescence In-Situ Hybridization 105

7.5 References 115

Chapter 8: Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. 118

8.1 Introduction 119

8.2 Results and Discussion 120

8.2.1 Copy Number Analysis of Lung Cancer Cell Genomes. 120

8.2.2 Frequency Analysis. 121

8.2.3 Regions of Similarity. 122

8.2.4 Regions of Difference. 123

8.2.5 Identification of Genes Differentially Expressed Between SCLC and NSCLC Caused by Phenotype Specific Copy Number Alteration. 124

8.2.6 Biological Pathways Differentially Altered in SCLC and NSCLC. 126

8.3 Conclusions. 127

8.4 Methods and Materials: 127

8.4.1 DNA Samples. 127

8.4.2 Tiling Path Array CGH. 128

8.4.3 Imaging and Data Analysis. 129

8.4.4 Statistical Analysis of Array CGH Alteration Frequencies. 129

8.4.5 Affymetrix Gene Expression Analysis. 130

8.4.6 Real Time PCR. 130

8.4.7 Principal Components Analysis. 131

8.5 References 140

Chapter 9: EZH2 is over-expressed in SCLC as a result of genomic deregulation of the Rb/E2F pathway 144

9.1 Introduction 145

9.2 Results and Discussion 146

9.2.1 Genomic Profiling of SCLC Tumours 146

9.2.2 Comparison of Genomic Profiles for SCLC Tumours and Cell Lines 147

9.2.4 The E2F/Rb pathway is specifically deregulated in SCLC 148

9.2.5 Over-expression of EZH2 150

9.3 Conclusions 152

9.4 Methods 153

9.4.1 Tissue Samples 153

9.4.2 Array CGH 153

9.4.3 Expression Analysis 154

9.5 References 162

Chapter 10: Conclusions 166

10.1 Summary 167

10.1.1 Technical Development of array CGH 167

10.1.2 Profiling of SCLC 169

10.2 Conclusions 172

10.3 Future Directions 174

10.4 References 176

List of Tables

Table 5.1. Comparison of array CGH technologies. 80

Table 6.1. Summary of novel alterations 93

Table 7.1. Alterations with Common Boundaries in Blood Lymphocytes and SCLC Cell Lines113

Table 8.1. Differential Deregulation of Genes in Key Biochemical Pathways between NSCLC and SCLC 138

Table 9.1. Clinical Data for SCLC tumours in this study. 159

Table 9.2. Real Time PCR Summary....................161

List of Figures

Figure 1.1. Conventional genomic profiling tools. 13

Figure 1.2. Principle of Array CGH. 14

Figure 2.1. Detection of two-fold copy number changes in TAT-1 lymphoma cell line on chromosome arms 8q and 18q. 28

Figure 2.2. Whole genome SMRT array CGH of lung cancer cell line H526. 29

Figure 2.3. Amplification of chromosome 8q24.12–.13 in colorectal cancer cell line COLO320. 30

Figure 2.4. Identification of a novel microamplification by tiling resolution array CGH in COLO320. 31

Figure 2.5. Identification of microdeletions. 32

Figure 3.1. Overall View of SeeGH Data Flow. 46

Figure 3.2. SeeGH "New Data" Window. 47

Figure 3.3. SeeGH "Chromosome View" Window. 48

Figure 3.4. SeeGH "Search" Window. 49

Figure 3.5. SeeGH "Chromosome View" Window. 50

Figure 3.6. SeeGH "Existing Data" Window. 51

Figure 4.1. Tissue Heterogeneity simulation. 58

Figure 4.2. Minimal detectable alteration sizes. 60

Figure 4.3. SMRT array CGH profile of an archival squamous cell lung tumor containing ~30% tumor cells. 61

Figure 5.1. Schematic Overview of Array CGH Platform Designs. 75

Figure 5.2. Theoretical Detection Sensitivity. 76

Figure 5.3. Single Copy Detection Sensitivity. 77

Figure 5.4. Breakpoint Precision. 79

Figure 6.1. 5p Array CGH Profile of H526. 90

Figure 6.2. Summary of the 15 SCLC Lines Profiled. 91

Figure 7.1. SMRT Array Profile of the SCLC Cell Line NCI-H289. 107

Figure 7.2. Overview of Weighted Frequency Scoring. 108

Figure 7.3. Weighted Frequency Analysis of Genomic Alterations in SCLC Cell Lines. 109

Figure 7.4. Identification of Copy Number Polymorphisms and Other Non-Somatic Alterations. 110

Figure 7.5. The 7p22.3 Locus is Highly Frequently Gained in SCLC Cell Lines, and Present in Other Tumor Types. 111

Figure 8.1. SMRT Array Profile of the SCLC NCI-H1672 Cells. 132

Figure 8.2. Copy Number Alterations in SCLC and NSCLC. 133

Figure 8.3. Differential Expression as a Result of Copy Number Alteration.......... 134

Figure 8.4. Contribution of Copy Number Induced Gene Expression Differences to the SCLC and NSCLC Phenotypes. 136

Figure 8.5. Differential Targets of Copy Number Induced Expression Changes in Key Biochemical Pathways between SCLC and NSCLC. 137

Figure 9.1. Comparison of SCLC Cell Lines and Tumours. 155

Figure 9.2. Deregulation of the E2F transcription factors. 157

Figure 9.3. Hyper-activation of EZH2 in SCLC. 158

List of Abbreviations

Abbreviation	Definition
aCGH	Array Comparative Genomic Hybridization
BAC	Bacterial Artificial Chromosome
bp	base pairs
CAN	Copy Number Alteration
CCD	Charge Coupled Device
cDNA	complimentary DNA
CEP	Centromere Probe
CGH	Comparative Genomic Hybridization
DNA	Deoxyribose Nucleic Acid
EGFR	Epidermal Growth Factor
EZH2	Enhancer of Zeste, Drosophila, Homolog 2
FFPE	Formalin-Fixed, Paraffin Embeded
FISH	Fluorescence In-Situ Hybridization
kbp	kilobase pairs
LMPCR	Linker Mediated Polymerase Chain Reaction
LOH	Loss Of Heterozygosity
LUCA	LUng CAncer
MAPK	Mitogen-Activated Protein Kinase
Mbp	Megabase pairs

mRNA	messenger RNA
NSCLC	Non-Small Cell Lung Cancer
PcG	Polycomb Group
PCR	Polymerase Chain Reaction
PMT	Photo-Multiplies Tube
PSCNA	Phenotype Specific Copy Number Alteration
Rb	Retinoblastoma
RNA	RiboNucleic Acid
RT-PCR	Real Time PCR
SCLC	Small Cell Lung Cancer
SMRT	Sub-Megabase Resolution Tiling-set
SNR	Signal to Noise Ratio
STMN	Stathmin
TNM	Tumour, Node, Metastisis
TRIO	Triple Function Domain
tRNA	transfer RNA
WNT	Wingless-Type

Acknowledgements

I would like to acknowledge the contributions of the many members of the Wan Lam Lab who contributed to this work, in particular the co-authors of each of the manuscript chapters presented herein. Specific acknowledgements from the published versions of each chapter are detailed below:

Chapter 2: The authors wish to thank the following for their contribution to this publication: J. Vielkind, S. Lam, D. Horsman, M. Rosin, S. Herst, K. Lonergan, S. Ralph, J. Rathmann, R. Seagraves, M. Krzywinski, P. Lansdorp, G. Bebb, J. Schein, I. Bosdet, D. Smailus, Z. Xu, C. Brown, J. Minna and A. Gazdar.

Chapter 3: We would like to thank Spencer Watson, Chad Malloff, and Adrian Ishkanian for their input on features to incorporate into SeeGH.

Chapter 4: We thank Dr. Jean LeRiche for evaluation of lung specimens and Dr. Carolyn Brown for X chromosome discussions.

Chapter 5: The authors would like to thank Timon Buys, Raj Chari, Jonathan Davies, Ron deLeeuw, Will Lockwood and Ian Wilson (BCCRC) for careful proofreading of the manuscript, and Eric Lin of Agilent Technologies for providing the unpublished BT474 hybridization.

Chapter 6: We thank Adrian Ishkanian, Chad Malloff, and Spencer Watson for their assistance in the construction of the array and Marco Marra for providing BAC clones.

Chapter 7: The authors would like to thank Ron deLeeuw, Spencer Watson and Laura-Jane Henderson for array production and FISH protocol.

Chapter 8: The authors wish to thank S. K. Watson for array synthesis and J.J. Davies for useful discussion.

I would also like to acknowledge scholarship support from the National Sciences and Engineering Research Council and Michael Smith Foundation for Health Research.

The research presented herein was funded but he following granting agencies: Genome Canada/ Genome British Columbia (All Chapters), National Cancer Institute of Canada (Chapters 5,6,7,8) Canadian Institute for Health Research (Chapters 4,5,6,7,8), NIH Lung SPORE (P50CA70907) (Chapters 6,7,8), National Institutes of Health (Chapter 8)

Co-Authorship Statement

Chapters 2 to 10 were co-authored as manuscripts for publication. The following author lists apply for each chapter:

Chapter 2: Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, Ling V, MacAulay C, Lam WL. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet.* **36**(3):299-303.

Contribution: I developed the imaging and printing procedures used to produce the array. I also assisted in the development of data analysis methods and many experiments in the preliminary development of array CGH in out laboratory, as well as assisting in the writing of he manuscript.

Chapter 3: Chi, B, deLeeuw, RJ, Coe BP, MacAulay, C, Lam WL (2004). SeeGH – A software tool for visualization of whole genome array comparative genomic hybridization data BMC Bioinformatics **9**(5):13

Contribution: I designed the look and feel of the application and contributed to feature suggestions and the writing of the manuscript.

Chapter 4: Garnis C*, Coe BP*, Lam SL, MacAulay C, Lam WL. (2005) High-resolution array CGH increases heterogeneity tolerance in the analysis of clinical samples. *Genomics.* **85**(6):790-3 *** These authors contributed equally**

Contribution: I performed and co-designed all statistical analysis of the data and assisted in the hybridizations and writing.

Chapter 5: Coe BP, Ylstra B, Carvalho B, Meijer GA, MacAualy C, Lam WL (2007) Resolving the resolution of array CGH *Genomics* **89**(5):647-653

Contribution: I conceived the concept for the manuscript, developed the algorithm for comparing array platforms, acquired public array data and wrote the manuscript.

Chapter 6: Coe BP, Henderson LJ, Garnis C, Tsao MS, Gazdar AF, Minna J, Lam S, Macaulay C, Lam WL. (2005) High-resolution chromosome arm 5p array CGH analysis of small cell lung carcinoma cell lines. *Genes Chromosomes Cancer* **94**(3):308-313

Contribution: I co-constructed the 5p CGH array with C Garnis. Hybridization work was split with LJ Henderson. I performed all data analysis and writing.

Chapter 7: Coe BP, Lee EH, Chi B, Girard L, Minna JD, Gazdar AF, Lam S, MacAulay C, Lam WL. (2006) 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**(1):11-19

Contribution: I co-performed the array CGH hybridizations and performed the data analysis and the majority of the writing.

Chapter 8: Coe BP*, Lockwood WW*, Girard L, Chari R, MacAualy C, Lam S, Gazdar AF, Minna JD, Lam WL (2006) Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. *BJC* **94**:1927-1935 *** These authors contributed equally**

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

Chapter 9: Coe BP, Aviel-Ronen S, Andrea Pusic, Gazdar AF, Minna JD, Lam S, Tsao MS, Lam WL

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

Dedication

To my family.

Chapter 1: Introduction

1.1 Introduction to Lung Cancer

Lung cancer is the leading cause of cancer mortality worldwide and is predominantly linked to smoking which increases cancer risk by up to 30 times that of non-smokers. The disease is classified by clinical and histological criteria into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). The division between SCLC and NSCLC is important, as they represent a drastically different biology and clinical course. SCLC accounts for ~20% of all lung cancer cases and is biologically the most aggressive lung tumour due to its short doubling time and poor clinical outcome. NSCLC is the most common type of lung cancer accounting for ~80% of cases. (Ward et al., 2006)

1.2 Non Small Cell Lung Cancer

NSCLC is the most common lung neoplasm and can be further subdivided into squamous cell carcinoma (~30%), adenocarcinoma (~40-50%), large cell carcinoma (~10%) and several other less common subclasses including unclassifiable tumours. Squamous cell carcinoma is a primarily central tumour and demonstrate squamous differentiation with intracellular bridges, keratinisation and squamous pearls. Squamous carcinoma is believed to develop through a multi stage progression model initiating with basal cell hyperplasia, progressing though squamous metaplasia, dysplasia and carcinoma in-situ to a fully malignant tumour. Adenocarcinoma is the most common NSCLC subtype (~40-50%) and develops in the peripheral lung. Adenocarcinoma is typically highly heterogeneous and tends to display multiple sub histologies. Similarly to squamous cell carcinoma a multistep progression model has been proposed starting from atypical adenomous hyperplasia and progressing though low grade bronchial alveolar carcinoma lesions to an invasive phenotype. Large Cell carcinoma is the least common major subtype of NSCLC and a specific subset of these tumours display a neuroendocrine phenotype with survival similar to SCLC. (Giangreco et al., 2007; Travis, 2002; Ward et al., 2006)

Both adenocarcinoma and squamous carcinomas of the lung are classified by the traditional TNM (tumour, node, metastasis) staging system and stage is highly predicative of survival. Long term survival rates of 60 to 75% are observed for early stage disease, while higher stage disease demonstrates increasingly poor survival with virtually no long term survivors for high grade metastatic disease. (Flieder, 2007; Gomez & Silvestri, 2008)

1.3 Small Cell Lung Cancer

1.3.1 Clinical Features

A predominantly central tumour, SCLC is unique in its clinical and histological presentation. SCLC demonstrates the strongest linkage to smoking with greater than 90% of cases being attributable to a history of smoking. Unlike NSCLC, SCLC is often diagnosed at a very late stage with metastasis often presenting at initial diagnosis. Due to the broad spread of disease by initial diagnosis, surgical resection is rarely an option and chemotherapy combined with radiation is the only recourse. SCLC is a very chemo and radiosensitive disease which responds very strongly to initial therapy. However the disease relapses as a chemo-insensitive metastatic tumour in virtually all cases. For these reasons SCLC is not characterized by the traditional TNM (tumour, node, metastasis) staging system. Instead a two stage system is used which consists of limited and extensive stage disease. Limited stage disease represents 40% of cases and is defined as being confined to a single region of the lung which can be covered by a tolerable radiation field. Limited stage disease demonstrates a median survival of 18 to 24 months and long term survival is observed in 20-25% of cases with combined chemo and radiotherapy. Extensive stage disease is characterized by spread outside the thorax and a median survival of only 8-11 months with virtually no long term survivors (1-2%). Unlike limited stage disease, extensive disease is primarily treated with palliative intent and radiotherapy is not common, as sufficient palliation is achieved with chemotherapy alone. (De Ruysscher et al., 2006; Krug & Miller, 2003; Lally et al., 2007; Lewinski & Zulawski, 2003; Murray et al., 1993; Murray & Turrisi, 2006; Pignon et al., 1992; Rostad et al., 2004; Roth et al., 1992; Slotman et al., 2007; Socinski & Bogart, 2007; Souhami et al., 1997; Sundstrom et al., 2002; Warde & Payne, 1992; Zakowski, 2003)

Like some large cell and carcinoid tumours SCLC is classified as demonstrating a neuroendocrine phenotype due to the expression of various cell markers associated with neuroendocrine cells such as neural cell adhesion molecule (*NCAM*), however SCLC demonstrates by far the most aggressive clinical course (Kraus et al., 2002; Krug & Miller, 2003; Lewinski & Zulawski, 2003; Rozengurt, 1999; Sattler & Salgia, 2003; Travis, 2002).

Currently two major subtypes of SCLC are recognized. Classic SCLC appears as very small cells with scant cytoplasm, a high mitotic index, and extensive necrosis. Cell diameters are approximately that of 2-3 small lymphocytes with a predominant growth pattern of diffuse sheets. Combined SCLC represents fewer than 10% of cases and appears as a tumour containing features of SCLC in combination with features of other NSCLC cell types (Socinski & Bogart, 2007; Travis, 2002; Zakowski, 2003). 1% to 3% of cases are combined with adenocarcinoma or squamous cell carcinoma, and 4% to 6% of cases appear as a mixture with large cell carcinoma (Travis, 2002). Cases where combined SCLC demonstrates a large cell phenotype are sometimes classified as variant SCLC and may demonstrate loss of certain neuroendocrine cell markers such as chromogranin A, and gastrin releasing peptide. (Broers et al., 1988; Gazdar et al., 1985; Kraus et al., 2002; Rozengurt, 1999; Zakowski, 2003)

* Due to the requirements of various journals, reviewer requests, and dates of publication the statistics described in the body of this thesis may differ from those presented here. These data supersede those presented elsewhere in this thesis.

1.3.2 Cell Lines

Due to the aggressive clinical course of SCLC, surgery is rarely a treatment option (Lally et al., 2007; Rostad et al., 2004). Although samples are occasionally surgically resected the majority of these are archived as formalin fixed paraffin embedded (FFPE) samples, limiting the application of many molecular techniques which depend on high quality DNA or RNA. For this reason, fresh clinical samples are very rare, and thus much research has depended on the availability of cell lines, such as those characterized by Phelps et al (Phelps et al., 1996), which are the focus of this thesis (as they are readily available from the American Type Culture Collection and well characterized), or paraffin embedded archival material. Most SCLC lines grow as suspension cultures of spheroids of multiple cells limiting the use of many transfection protocols.

1.3.3 Molecular Biology of SCLC

Prior to the initiation of this thesis, several published studies have attempted to identify genomic regions of DNA loss or gain in SCLC, leading to the discovery of multiple genomic regions such as frequent gains on chromosome arms 1p, 2p, 5p, 8q and losses of 3p, 13q and 17p. However the low resolution of conventional whole genome profiling leads to difficulties in identifying specific genes of interest and many of these studies associated copy number alterations with previously known oncogenes such as *Myc* family genes (1p,2p,8q) and tumour suppressors such as *p53* (17p) and *Rb* (13q). (Ashman et al., 2002; Levin et al., 1994; Levin et al., 1995; Lindblad-Toh et al., 2000; Petersen et al., 1997; Testa et al., 1997)

In order to identify novel genes associated with lung cancer, labour intensive fine mapping techniques were often required including loss of heterozygosity and FISH based methods (see section 1.4.2 of this chapter). These approaches have identified novel minimal regions such as 5p13 (containing *SKP2*) , and the lung cancer (LUCA) cluster on 3p21.3 (Senchenko et al., 2004; Yokoi et al., 2002). However progress has been slow and many target genes of DNA alteration remain to be found ((Balsara & Testa, 2002).

In addition to genomic studies many groups have attempted to understand the biology of SCLC through the application of gene expression microarrays. This powerful technology allows the analysis of thousands of genes simultaneously, however the complexity of the data has relegated most studies to disease sub-classification and prognosis, rather than specific gene discovery (Bangur et al., 2002; Difilippantonio et al., 2003; Jones et al., 2004; Pedersen et al., 2003; Virtanen et al., 2002).

Evidence from the application of conventional techniques, gene expression profiling, existing knowledge from other cancers, and both modern and conventional profiling technologies suggests that SCLC is the result of the disruption of multiple signalling cascades leading to uncontrolled growth, loss of apoptosis, disruption of telomerase, changes in DNA repair pathways and drug efflux/resistance. Many studies have identified multiple paracrine and autocrine signalling pathways, as well as downstream effectors leading to the pro-proliferative and neuroendocrine phenotype. These include alteration of *c-MET* and paracrine activation by HGF potentially leading to an invasive phenotype; autocrine/paracrine activation and mutation of *c-kit* leading to growth; disruption of the AKT and MAPK pathways leading to increased growth and drug resistance; and disruption of Notch and hedgehog regulation of the neuroendocrine phenotype and lung development (Boldrini et al., 2004; Kraus et al., 2002; Osada & Takahashi, 2002; Rozengurt, 1999; Sattler & Salgia, 2003; Shan et al., 2007; Vestergaard et al., 2006; Watkins et al., 2003a; Watkins et al., 2003b). Additionally a large degree of research has focused on the loss of Rb and its unusually high frequency of disruption (90% in SCLC) in neuroendocrine tumours (Beasley et al., 2003; Shimizu et al., 1997).

In spite of our increasing understanding of SCLC biology, unfortunately there are still no targeted therapeutics available. Attempts to target deregulated receptors such as c-kit, or cell surface markers such as NCAM have failed to achieve a significant therapeutic effect, and thus new therapeutic targets are desperately needed (Jensen & Berthold, 2007; Kraus et al., 2002; Rossi et al., 2004; Tiseo & Ardizzoni, 2007).

1.3.4 The Need for New Tools to Study SCLC

Due to the rarity of surgical resection, fresh frozen tumour material is very rare and most specimens will be acquired as formalin fixed paraffin embedded (FFPE) tissue samples which are not applicable to expression analysis and contain low quality DNA which is not applicable to all profiling approaches. Additionally expression changes are often the result of complex interactions between multiple pathway disruptions and may not reflect changes key to carcinogenesis. DNA analysis can help identify those genes that are primarily deregulated as a result of genomic alterations, however whole genome approaches are essential if we are to understand the overall role of genetics in the phenotype of a disease (Lockwood et al., 2006). Thus genome wide copy number based tools are critical for the understanding of SCLC, both for use in combination with expression on cell lines samples and the analysis of FFPE clinical specimens. For this reason development of array comparative genomic hybridization, a new technique which allows unprecedented detail in the analysis of aneuploidy in cancer genomes, is likely to yield great insight into SCLC by allowing fine mapping of genomic alterations in a single experiment, and rapid identification of novel target genes.

1.4Tools for the Molecular Profiling of Cancer

Arrays technologies have led to transition from single gene to whole genome assays. This was initially accomplished by transition from single loci mapping (loss of heterozygosity (LOH) etc) to chromosomal comparative genomic hybridization (CGH) and the move from expression analysis of a single gene to tens of thousands of genes in a single experiment (**Figure 1.1**). The benefit of microarray technology has since been applied to copy number. Through utilization of these tools, there has been an exponential increase in the amount of biological data available.

1.4.1 Expression Analysis

One of the first high throughput tools developed for the study of model systems and disease was expression microarray analysis. Traditionally gene expression was measured by Northern blot or PCR based techniques which allowed the interrogation of a single transcript at a time.

The advent of DNA microarrays allowed researchers to deposit tens of thousands of cDNA targets on a single glass slide (Staudt & Brown, 2000). The slide then acted as a hybridization target for fluorescently labelled cDNA from the sample of interest, and allowed quantitative assessment of gene transcription at a scale never before possible. Using these arrays researchers were for the first time able to examine the underlying biology of the transcriptome and understand the alterations that occurred during disease pathogenesis.

Since the development of early expression microarrays, the technique has advanced to utilize oligonucleotide targets which assay the transcriptome with multiple probes for each gene, offering coverage as high as one probe per exon (Coe et al., 2007).

However as understanding of the transcriptome has increased, so has the apparent complexity of expression data. Many genes are capable of regulating the expression of each other and only a subset of genes are controlled by the initial causal processes driving tumourigenesis such as aberrations in DNA methylation levels, chromatin restructuring and modified gene dosage (Coe et al., 2006; Pollack et al., 2002; Snijders et al., 2001). Thus, the study of gene expression in conjunction with other genomic metrics is important to fully understand the regulation of genes during disease progression.

1.4.2 Array CGH

Somatic DNA copy number alterations are hallmarks of cancer, leading to disruptions in the expression of oncogenes and/or tumour suppressor genes, whilst constitutional DNA copy number variations have been associated with developmental disorders. Prior to the development of array CGH the majority of research focused on single locus assays (similar to initial gene expression assays).

The development of conventional CGH (utilizing metaphase chromosome spreads as a hybridization target) allowed researchers to understand the patterns of gene dosage across the entire genome, albeit at a relatively low resolution of ~10Mbp. The development of cDNA expression microarrays provided the first ability to move CGH from metaphase chromosome into a gene resolution assay; however, cDNA targets lack introns that are present in genomic probe mixtures, resulting in relatively low signal-to-noise ratios and limited estimation of copy number. The technology was then improved by utilizing large insert clones containing segments of ~150kbp of human DNA (BAC clones) as the hybridization target. This technology was initially applied as a 3,000 clone array and later adapted into the 32,000 element SMRT array detailed in chapter 2 (**Figure 1.2**) (de Leeuw et al., 2004; Ishkanian et al., 2004; Snijders et al., 2001). Additionally, this progression in array density has required the development of new informatics tools to decipher the resultant data; this is discussed in chapter 3 and a detailed review of informatics strategies can be found in Chari et al (Chari et al., 2006; Chi et al., 2004). During the same timeline, oligonucleotide (25-80bp nucleotide probes) arrays were also developed with the goal of improving the maximal resolution of CGH beyond the size of a BAC clone. Both technologies have matured to allow high resolution profiling of the genome with distinct advantages and disadvantages for each. For example, BAC arrays require far less sample input than oligonucleotide arrays allowing the analysis of low yield microdissected specimens. Conversely, BAC arrays are more difficult to produce and have reached their maximum resolution with the SMRT array (Coe et al., 2007; Ylstra 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 niche in today's genomics field (For a more detailed discussion of array platforms please see Chapter 4).

1.5 Thesis Theme

The theme of this thesis is the determination of the genomic alterations, and patterns of pathway disruption that lead to the aggressive nature of SCLC. This is accomplished through the integration of novel genomic profiling tools with gene expression analysis and comparisons between SCLC and the less aggressive NSCLC subtypes.

1.6 Objectives and Hypothesis

The objective of this work is to determine the genetic alterations key to the aggressive nature of SCLC, by comparison with NSCLC through high resolution whole genome DNA copy number profiling.

The major objectives of this work are thus to demonstrate that:

1) Oncogenes and tumour suppressors key to SCLC tumourigenesis, will be identified in recurrent regions of gain and loss detectable by high resolution genomic profiling.

2) Genomic alterations key to the aggressive phenotype of SCLC can be identified by comparison of genomic profiles of SCLC and the less aggressive NSCLC.

3) Patterns of biochemical pathway regulation unique to SCLC, can be identified by integration of copy number and gene expression changes in SCLC and NSCLC.

1.7 Specific Aims and Thesis Outline

This thesis consists of several manuscripts assembled into a non-chronological order to best address the hypothesis of this thesis.

Aim 1: Development of high resolution array CGH profiling tools for the analysis of SCLC genomes.

Chapters 2 to 5 describe the development of the high resolution genome analysis tools necessary to address hypothesis 1 to 3.

Chapter 2 details the development of a novel high resolution array CGH platform utilizing BAC clones covering the human genome in a tiling path. To reflect this novel design strategy we refer to this platform as the Sub-Megabase Resolution Tiling-set (SMRT) array. This platform allows rapid fine mapping of disease specific genomic alterations, essential to addressing hypothesis 1. In parallel to the development of the SMRT array it became apparent that no existing tools were capable of adequately displaying the data generated by a high resolution array CGH experiment. Thus Chapter 3 describes the development of SeeGH, a software application which allows simple visualization of array CGH data in the context of traditional chromosome ideograms.

At the initiation of this thesis and following the initial publication of the SMRT array, the established protocols for array CGH were optimized for use with large quantities of DNA that are not attainable from archival FFPE material. Additionally a primary concern in the profiling of archival specimens is that of sample purity, as tumours represent mixtures of normal and cancer cells. Thus Chapter 4 focuses on the characterization of the SMRT arrays applicability to these samples, and the determination of tolerable limits for normal cell percentages in analyzed samples.

During the progress of this project many technologies have been designed to supplant conventional metaphase CGH technology. However these new technologies have not been comprehensively compared and separation of manufacturer claims from real world performance is essential in determining the ideal platform for a particular project. Thus chapter 5 focuses on the development of a new definition of resolution for array CGH platforms and the comparison of real world performance between these platforms.

Aim 2: Array CGH profiling of SCLC Cell Lines.

Small Cell Lung Cancer cell lines are far easier to obtain than clinical specimens, and RNA can be easily obtained for the identification of downstream gene expression changes. Thus the majority of this thesis focuses on the analysis of cell lines. Chapters 6 and 7 highlight the genomic profiling of SCLC cell lines both using the SMRT array developed in chapter 2 and a specific array covering chromosome arm 5p, The chromosome 5p array was developed before the whole genome SMRT array and utilized during its optimization as we originally planned on producing primarily chromosome arm specific arrays. This work identified multiple novel regions of copy number alteration in SCLC thus addressing hypothesis 1.

Aim 3: Comparison of SCLC and NSCLC Cell Line genomes.

After the initial genomic profiling of SCLC cell lines in Aim 2, we compared the SCLC cell lines profiles to NSCLC cell lines using an integrative combination of copy number and gene expression data. This data was then used to identify both genomic regions specific to SCLC and specific patterns of pathway alterations which may explain the SCLC phenotype, thus testing hypotheses 2 and 3.

Aim 4: Validation of genomic alterations in SCLC tumours.

In addition to the work performed on cell lines in Aims 2 and 3 we profiled a set of FFPE primary SCLC tumours in Chapter 9. This work was utilized to separate the genomic alterations that may be specific to cell culture transformation from those relevant to clinical disease. This work is presented in further support of hypotheses 2 and 3, by relating the findings of aims 2 and 3 to clinical SCLC cases, and expanding the analysis of the pathways identified in the cell lines.

While these papers represent separate works, additional information can be gained by integrating the results from these manuscripts. This is discussed in the Conclusions of this thesis.

Figure 1.1. Conventional genomic profiling tools.

PLACEHOLDER INSERT FIGURE 1.1

Figure 1.2. Principle of Array CGH..

PLACEHOLDER INSERT FIGURE 1.2

1.8 References

Ashman, J.N., Brigham, J., Cowen, M.E., Bahia, H., Greenman, J., Lind, M. & Cawkwell, L. (2002). Chromosomal alterations in small cell lung cancer revealed by multicolour fluorescence in situ hybridization. *Int J Cancer*, 102, 230-6.

Balsara, B.R. & Testa, J.R. (2002). Chromosomal imbalances in human lung cancer. *Oncogene*, 21, 6877-83.

Bangur, C.S., Switzer, A., Fan, L., Marton, M.J., Meyer, M.R. & Wang, T. (2002). Identification of genes over-expressed in small cell lung carcinoma using suppression subtractive hybridization and cDNA microarray expression analysis. *Oncogene*, 21, 3814-25.

Beasley, M.B., Lantuejoul, S., Abbondanzo, S., Chu, W.S., Hasleton, P.S., Travis, W.D. & Brambilla, E. (2003). The P16/cyclin D1/Rb pathway in neuroendocrine tumors of the lung. *Hum Pathol*, 34, 136-42.

Boldrini, L., Ursino, S., Gisfredi, S., Faviana, P., Donati, V., Camacci, T., Lucchi, M., Mussi, A., Basolo, F., Pingitore, R. & Fontanini, G. (2004). Expression and mutational status of c-kit in small-cell lung cancer: prognostic relevance. *Clin Cancer Res*, 10, 4101-8.

Broers, J.L., Pahlplatz, M.M., Katzko, M.W., Oud, P.S., Ramaekers, F.C., Carney, D.N. & Vooijs, G.P. (1988). Quantitative description of classic and variant small cell lung cancer cell lines by nuclear image cytometry. *Cytometry*, 9, 426-31.

Chari, R., Lockwood, W.W. & Lam, W.L. (2006). Computational methods for the analysis of array comparative genomic hybridization. *Cancer Informatics*, 48-58.

Chi, B., DeLeeuw, R.J., Coe, B.P., MacAulay, C. & Lam, W.L. (2004). SeeGH--a software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics*, 5, 13.

Coe, B.P., Lockwood, W.W., Girard, L., Chari, R., Macaulay, C., Lam, S., Gazdar, A.F., Minna, J.D. & Lam, W.L. (2006). Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. *Br J Cancer*, 94, 1927-35.

Coe, B.P., Ylstra, B., Carvalho, C., Meijer, G.A., Macaulay, C. & Lam, W.L. (2007). Resolving the resolution of array CGH. *Genomics*, In Press.

de Leeuw, R.J., Davies, J.J., Rosenwald, A., Bebb, G., Gascoyne, R.D., Dyer, M.J., Staudt, L.M., Martinez-Climent, J.A. & Lam, W.L. (2004). Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum Mol Genet*, 13, 1827-37. Epub 2004 Jun 30.

De Ruysscher, D., Pijls-Johannesma, M., Bentzen, S.M., Minken, A., Wanders, R., Lutgens, L., Hochstenbag, M., Boersma, L., Wouters, B., Lammering, G., Vansteenkiste, J. & Lambin, P. (2006). Time between the first day of chemotherapy and the last day of chest radiation is the most important predictor of survival in limited-disease small-cell lung cancer. *J Clin Oncol*, 24, 1057-63.

Difilippantonio, S., Chen, Y., Pietas, A., Schluns, K., Pacyna-Gengelbach, M., Deutschmann, N., Padilla-Nash, H.M., Ried, T. & Petersen, I. (2003). Gene expression profiles in human non-small and small-cell lung cancers. *Eur J Cancer*, 39, 1936-47.

Flieder, D.B. (2007). Commonly encountered difficulties in pathologic staging of lung cancer. *Arch Pathol Lab Med*, 131, 1016-26.

Gazdar, A.F., Bunn, P.A., Jr., Minna, J.D. & Baylin, S.B. (1985). Origin of human small cell lung cancer. *Science*, 229, 679-80.

Giangreco, A., Groot, K.R. & Janes, S.M. (2007). Lung cancer and lung stem cells: strange bedfellows? *Am J Respir Crit Care Med*, 175, 547-53.

Gomez, M. & Silvestri, G.A. (2008). Lung cancer screening. *Am J Med Sci*, 335, 46-50.

Ishkanian, A.S., Malloff, C.A., Watson, S.K., DeLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M.A., Ling, V., MacAulay, C. & Lam, W.L. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, 36, 299-303. Epub 2004 Feb 15.

Jensen, M. & Berthold, F. (2007). Targeting the neural cell adhesion molecule in cancer. *Cancer Lett*, 258, 9-21.

Jones, M.H., Virtanen, C., Honjoh, D., Miyoshi, T., Satoh, Y., Okumura, S., Nakagawa, K., Nomura, H. & Ishikawa, Y. (2004). Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet*, 363, 775-81.

Kraus, A.C., Ferber, I., Bachmann, S.O., Specht, H., Wimmel, A., Gross, M.W., Schlegel, J., Suske, G. & Schuermann, M. (2002). In vitro chemo- and radio-resistance in small cell lung cancer correlates with cell adhesion and constitutive activation of AKT and MAP kinase pathways. *Oncogene*, 21, 8683-95.

Krug, L. & Miller, V. (2003). Introduction: Small Cell Lung Cancer - A Frustrating Disease. *Semin Oncol*, 30, 1-2.

Lally, B.E., Urbanic, J.J., Blackstock, A.W., Miller, A.A. & Perry, M.C. (2007). Small cell lung cancer: have we made any progress over the last 25 years? *Oncologist*, 12, 1096-104.

Levin, N.A., Brzoska, P., Gupta, N., Minna, J.D., Gray, J.W. & Christman, M.F. (1994). Identification of frequent novel genetic alterations in small cell lung carcinoma. *Cancer Res*, 54, 5086-91.

Levin, N.A., Brzoska, P.M., Warnock, M.L., Gray, J.W. & Christman, M.F. (1995). Identification of novel regions of altered DNA copy number in small cell lung tumors. *Genes Chromosomes Cancer*, 13, 175-85.

Lewinski, T. & Zulawski, M. (2003). Small cell lung cancer survival: 3 years as a minimum for predicting a favorable outcome. *Lung Cancer*, 40, 203-13.

Lindblad-Toh, K., Tanenbaum, D.M., Daly, M.J., Winchester, E., Lui, W.O., Villapakkam, A., Stanton, S.E., Larsson, C., Hudson, T.J., Johnson, B.E., Lander, E.S. & Meyerson, M. (2000). Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat Biotechnol*, 18, 1001-5.

Lockwood, W.W., Chari, R., Chi, B. & Lam, W.L. (2006). Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *Eur J Hum Genet*, 14, 139-48.

Murray, N., Coy, P., Pater, J.L., Hodson, I., Arnold, A., Zee, B.C., Payne, D., Kostashuk, E.C., Evans, W.K., Dixon, P. & et al. (1993). Importance of timing for thoracic irradiation in the combined modality treatment of limited-stage small-cell lung cancer. The National Cancer Institute of Canada Clinical Trials Group. *J Clin Oncol*, 11, 336-44.

Murray, N. & Turrisi, A.T., 3rd. (2006). A review of first-line treatment for small-cell lung cancer. *J Thorac Oncol*, 1, 270-8.

Osada, H. & Takahashi, T. (2002). Genetic alterations of multiple tumor suppressors and oncogenes in the carcinogenesis and progression of lung cancer. *Oncogene*, 21, 7421-34.

Pedersen, N., Mortensen, S., Sorensen, S.B., Pedersen, M.W., Rieneck, K., Bovin, L.F. & Poulsen, H.S. (2003). Transcriptional gene expression profiling of small cell lung cancer cells. *Cancer Res*, 63, 1943-53.

Petersen, I., Langreck, H., Wolf, G., Schwendel, A., Psille, R., Vogt, P., Reichel, M.B., Ried, T. & Dietel, M. (1997). Small-cell lung cancer is characterized by a high incidence of deletions on chromosomes 3p, 4q, 5q, 10q, 13q and 17p. *Br J Cancer*, 75, 79-86.

Phelps, R.M., Johnson, B.E., Ihde, D.C., Gazdar, A.F., Carbone, D.P., McClintock, P.R., Linnoila, R.I., Matthews, M.J., Bunn, P.A., Jr., Carney, D., Minna, J.D. & Mulshine, J.L. (1996). NCI-Navy Medical Oncology Branch cell line data base. *J Cell Biochem Suppl*, 24, 32-91.

Pignon, J.P., Arriagada, R., Ihde, D.C., Johnson, D.H., Perry, M.C., Souhami, R.L., Brodin, O., Joss, R.A., Kies, M.S., Lebeau, B. & et al. (1992). A meta-analysis of thoracic radiotherapy for small-cell lung cancer. *N Engl J Med*, 327, 1618-24.

Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Borresen-Dale, A.L. & Brown, P.O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99, 12963-8.

Rossi, A., Maione, P., Colantuoni, G., Guerriero, C. & Gridelli, C. (2004). The role of new targeted therapies in small-cell lung cancer. *Crit Rev Oncol Hematol*, 51, 45-53.

Rostad, H., Naalsund, A., Jacobsen, R., Strand, T.E., Scott, H., Heyerdahl Strom, E. & Norstein, J. (2004). Small cell lung cancer in Norway. Should more patients have been offered surgical therapy? *Eur J Cardiothorac Surg*, 26, 782-6.

Roth, B.J., Johnson, D.H., Einhorn, L.H., Schacter, L.P., Cherng, N.C., Cohen, H.J., Crawford, J., Randolph, J.A., Goodlow, J.L., Broun, G.O. & et al. (1992). Randomized study of cyclophosphamide, doxorubicin, and vincristine versus etoposide and cisplatin versus alternation of these two regimens in extensive small-cell lung cancer: a phase III trial of the Southeastern Cancer Study Group. *J Clin Oncol*, 10, 282-91.

Rozengurt, E. (1999). Autocrine loops, signal transduction, and cell cycle abnormalities in the molecular biology of lung cancer. *Curr Opin Oncol*, 11, 116-22.

Sattler, M. & Salgia, R. (2003). Molecular and cellular biology of small cell lung cancer. *Semin Oncol*, 30, 57-71.

Senchenko, V.N., Liu, J., Loginov, W., Bazov, I., Angeloni, D., Seryogin, Y., Ermilova, V., Kazubskaya, T., Garkavtseva, R., Zabarovska, V.I., Kashuba, V.I., Kisselev, L.L., Minna, J.D., Lerman, M.I., Klein, G., Braga, E.A. & Zabarovsky, E.R. (2004). Discovery of frequent homozygous deletions in chromosome 3p21.3 LUCA and AP20 regions in renal, lung and breast carcinomas. *Oncogene*, 23, 5719-28.

Shan, L., Aster, J.C., Sklar, J. & Sunday, M.E. (2007). Notch-1 regulates pulmonary neuroendocrine cell differentiation in cell lines and in transgenic mice. *Am J Physiol Lung Cell Mol Physiol*, 292, L500-9.

Shimizu, E., Zhao, M., Shinohara, A., Namikawa, O., Ogura, T., Masuda, N., Takada, M., Fukuoka, M. & Sone, S. (1997). Differential expressions of cyclin A and the retinoblastoma gene product in histological subtypes of lung cancer cell lines. *J Cancer Res Clin Oncol*, 123, 533-8.

Slotman, B., Faivre-Finn, C., Kramer, G., Rankin, E., Snee, M., Hatton, M., Postmus, P., Collette, L., Musat, E. & Senan, S. (2007). Prophylactic cranial irradiation in extensive small-cell lung cancer. *N Engl J Med*, 357, 664-72.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. & Albertson, D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet*, 29, 263-4.

Socinski, M.A. & Bogart, J.A. (2007). Limited-stage small-cell lung cancer: the current status of combined-modality therapy. *J Clin Oncol*, 25, 4137-45.

Souhami, R.L., Spiro, S.G., Rudd, R.M., Ruiz de Elvira, M.C., James, L.E., Gower, N.H., Lamont, A. & Harper, P.G. (1997). Five-day oral etoposide treatment for advanced small-cell lung cancer: randomized comparison with intravenous chemotherapy. *J Natl Cancer Inst*, 89, 577-80.

Staudt, L.M. & Brown, P.O. (2000). Genomic views of the immune system*. *Annu Rev Immunol*, 18, 829-59.

Sundstrom, S., Bremnes, R.M., Kaasa, S., Aasebo, U., Hatlevoll, R., Dahle, R., Boye, N., Wang, M., Vigander, T., Vilsvik, J., Skovlund, E., Hannisdal, E. & Aamdal, S. (2002). Cisplatin and etoposide regimen is superior to cyclophosphamide, epirubicin, and vincristine regimen in small-cell lung cancer: results from a randomized phase III trial with 5 years' follow-up. *J Clin Oncol*, 20, 4665-72.

Testa, J.R., Liu, Z., Feder, M., Bell, D.W., Balsara, B., Cheng, J.Q. & Taguchi, T. (1997). Advances in the analysis of chromosome alterations in human lung carcinomas. *Cancer Genet Cytogenet*, 95, 20-32.

Tiseo, M. & Ardizzoni, A. (2007). Current status of second-line treatment and novel therapies for small cell lung cancer. *J Thorac Oncol*, 2, 764-72.

Travis, W.D. (2002). Pathology of lung cancer. *Clin Chest Med*, 23, 65-81, viii.

Vestergaard, J., Pedersen, M.W., Pedersen, N., Ensinger, C., Tumer, Z., Tommerup, N., Poulsen, H.S. & Larsen, L.A. (2006). Hedgehog signaling in small-cell lung cancer: frequent in vivo but a rare event in vitro. *Lung Cancer*, 52, 281-90.

Virtanen, C., Ishikawa, Y., Honjoh, D., Kimura, M., Shimane, M., Miyoshi, T., Nomura, H. & Jones, M.H. (2002). Integrated classification of lung tumors and cell lines by expression profiling. *Proc Natl Acad Sci U S A*, 99, 12357-62. Epub 2002 Sep 6.

Ward, J.P.T., Ward, J., Leach, R.M. & Wiener, C.M. (2006). *The Respiratory System at a Glance*. Blackwell Publishing: Massachusetts.

Warde, P. & Payne, D. (1992). Does thoracic irradiation improve survival and local control in limited-stage small-cell carcinoma of the lung? A meta-analysis. *J Clin Oncol*, 10, 890-5.

Watkins, D.N., Berman, D.M. & Baylin, S.B. (2003a). Hedgehog signaling: progenitor phenotype in small-cell lung cancer. *Cell Cycle*, 2, 196-8.

Watkins, D.N., Berman, D.M., Burkholder, S.G., Wang, B., Beachy, P.A. & Baylin, S.B. (2003b). Hedgehog signalling within airway epithelial progenitors and in small-cell lung cancer. *Nature*, 422, 313-7.

Ylstra, B., van den Ijssel, P., Carvalho, B., Brakenhoff, R.H. & Meijer, G.A. (2006). BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res*, 34, 445-50.

Yokoi, S., Yasui, K., Saito-Ohara, F., Koshikawa, K., Iizasa, T., Fujisawa, T., Terasaki, T., Horii, A., Takahashi, T., Hirohashi, S. & Inazawa, J. (2002). A novel target gene, SKP2, within the 5p13 amplicon that is frequently detected in small cell lung cancers. *Am J Pathol*, 161, 207-16.

Zakowski, M.F. (2003). Pathology of small cell carcinoma of the lung. *Semin Oncol*, 30, 3-8.

Chapter 2: A Tiling Resolution DNA Microarray with Complete Coverage of the Human Genome

A version of this chapter has been previously published as:

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, Ling V, MacAulay C, Lam WL. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet.* **36**(3):299-303. doi:10.1038/ng1307

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

2.1 Introduction

Identification of chromosomal imbalances and variation in DNA copy number is essential to our understanding of disease mechanisms and pathogenesis. Array CGH (Pinkel et al., 1998) or matrix CGH (Solinas-Toldo et al., 1997) offers the highest resolution for a practical genome-wide detection of chromosomal alterations. This technique is derived from the concept of conventional CGH (Kallioniemi et al., 1992), which has contributed greatly to the molecular characterization of both somatic and constitutional genomic DNA mutations over the last decade (Forozan et al., 1997; Knuutila et al., 2000; Wells & Levy, 2003). The primary limitation of conventional CGH is in resolution (~20 Mb) as this method detects segmental copy number changes on metaphase chromosomes (Kallioniemi et al., 1992). In array CGH, the metaphase chromosome spread is replaced by BACs, PACs, or YACs containing human DNA as targets, increasing the resolution to the distance between the selected marker DNA clones (Pinkel et al., 1998; Solinas-Toldo et al., 1997). Genome screening using array CGH has great potential in the characterization of numerous chromosomal disorders.

Efforts to construct DNA arrays spanning the human genome consisted of spotting 2460 (Snijders et al., 2001) or 3500 (Fiegler et al., 2003) marker BAC clones, representing the sequenced genome at an average interval of approximately 1 Mb. These studies showed that sufficient target-DNA printing solution could be generated from individual BACs using PCR based protocols. Since the target product was PCR derived, it is easily replenishable, obviating the need for multiple rounds of laborious large scale BAC DNA preparations. These arrays are sensitive enough for detection of single copy changes, but the technique is limited by the small number of BAC markers representing the genome on the slide, rather than the methodology. Even at this resolution, array CGH proves to be useful for detecting chromosomal aberrations associated with congenital abnormalities and somatic malignancies (Kraus et al., 2003; Veltman et al., 2003; Veltman et al., 2002; Weiss et al., 2003).

Recent studies have focused on higher-density regional arrays for fine mapping and identifying new genes in specific chromosomal regions(Albertson et al., 2000; Bruder et al., 2001; Garnis et al., 2003; Garnis et al., 2004a; Garnis et al., 2004b; Wilhelm et al., 2002). For example, a candidate oncogene for association with breast cancer (*CYP24*) was identified on 20q13.2 using an array of 29 overlapping clones spanning this region (Albertson et al., 2000). The need for a tiling resolution array to map these amplification or deletion boundaries is indicated by the fact that two separate regions of amplification within 20q13.2 contained two separate putative oncogenes, which would not have been detected by a lower resolution array. These studies show that the resolving power of array CGH is maximized when the detection of single copy number changes is combined with a tiling or overlapping set of BAC clones.

We created the first tiling resolution BAC array with complete coverage of the human genome using 32,433 fingerprint-verified individually amplified BAC clones. Here we show that such a complete genome comparison is capable of identifying micro-amplifications and micro-deletions, which may contain genes involved in disease pathogenesis. We call this Sub-Megabase Resolution Tiling-set for array CGH (SMRT array).

2.2 Results

2.2.1 Array sensitivity

To assess the sensitivity of the SMRT array, we hybridized the well characterized EBV-transformed lymphoma cell line TAT-1 to normal male genomic DNA. Genomic regions containing *BCL2* (18q21) and *MYC* (8q24) in TAT-1 were previously shown to have a twofold copy-number increase by FISH analysis (Denyssevych et al., 2002). We detected these previously reported amplifications at both loci, and we delineated their boundaries (**Figure 2.1**). Boundaries of amplification on chromosome 8 were between BAC clone RP11-143H8 at 8q22.2 and RP11-263C20 at 8q24.13. Boundaries of amplification on chromosome 18 were between

BAC clone RP11-159K14 at 18q21.32 and RP11-565D23 at 18q23. These data illustrate the detection sensitivity of array CGH.

2.2.2 Array resolution compared to conventional CGH

To demonstrate the resolving power of the SMRT array, we compared the $\log_2$ ratio profile of lung cancer cell line H526 (Girard et al., 2000; Levin et al., 1994) (**Figure 2.2a**) to the previously published conventional chromosomal CGH data (http://amba.charite.de/~ksch/cghdatabase/index.htm). All patterns of gains and losses were matched, including large changes (*e.g.* the amplification of 7q and 8q and loss of the entire chromosome 10), as well as complex changes (*e.g.* the multiple amplifications on chromosome 1 and the multiple deletions on chromosome 4). Notably, conventional chromosomal CGH identified a highly amplified region on the telomeric end of chromosome arm 2p, apparently covering approximately one fourth of the whole chromosome. However, the SMRT array analysis showed this amplification to be precisely localized to a 1.3 Mb fragment at 2p24.3, bordered by BAC clones RP11-351F4 and RP11-701O10, which contains the *MYCN* oncogene. The resolving power of this whole genome array enables us to define breakpoints to within single BAC clones. For example, the deletion breakpoint on chromosome arm 3p was localized to between BAC clones RP11-632O5 and RP11-594F16 at 3p21.1 (**Figure 2.2b**). This finding was subsequently confirmed by FISH analysis (**Figure 2.2c**).

2.2.3 Comparison to previous array CGH

To compare our tiling resolution array against current array CGH technology, we profiled colorectal cancer cell line COLO320 (ref. 22) which has been characterized in two previous array CGH studies (Snijders et al., 2001; Wessendorf et al., 2002). We confirmed the amplification at 8q24 *MYC* region identified by these studies. Furthermore, the SMRT system further defined this segmental copy number increase precisely to a 1.9 Mb region bordered by BAC clones RP11-810D23 and RP11-294P7 (**Figure 2.3**).

A detailed analysis of our COLO320 profile identified new microamplifications on chromosome arms 13q, 15q, 16p, and 22q (Supplemental Figure 1) which were not detected by the two prior high resolution CGH studies (Snijders et al., 2001; Wessendorf et al., 2002). For example, we identified a 300 Kb micro-amplification at 13q12.2 containing only three genes (according to UCSC Genome Browser April 2003 Freeze): caudal type homeobox transcription factor 2 (*CDX2*), insulin promoter factor 1 (*IPF-1*) and GS homeobox 1 (*GSH1*) (**Figure 2.4a**). *CDX2* is a transcription factor expressed in the intestine and altered in colorectal cancers (Kim et al., 2002). FISH analysis verified this microamplification and showed that it was within a homogeneously staining region (**Figure 2.4b**). These findings illustrate the usefulness of a tiling resolution BAC array for comprehensive assessment of genomic integrity.

2.2.5 Identification of minute regions of alteration

In addition to micro-amplifications, we also detected small deletions in a number of tumor cell lines. For example, we detected a 1.25 Mb deletion containing the gene *CDKN2A* (also called *p16*) in lymphoma cell line Z138C at 9p21.3 (**Figure 2.5a**). Deletion of *CDKN2A* occurs in approximately one-half of mantle cell lymphoma tumors as detected by FISH (Dreyling et al., 1997). This deletion is bordered by RP11-328C2 and RP11-275H17 (**Figure 2.5a**). Sub-megabase size micro-deletions can be accurately mapped in a single whole genome array CGH experiment. This is made possible by the overlapping clone coverage and their distribution on the array. A notable example is a 240 Kb deletion at 7q22.3 in the breast cancer cell line BT474, containing *PRKAR2B*, a regulatory kinase, and *HBP1*, a G1 inhibitory kinase regulated by p38 MAP kinase (Xiu et al., 2003) (**Figure 2.5b**). Such micro-deletions have not been reported previously. The mechanism(s) by which such deletions are effected are not known. Whether this micro-deletion affects the expression of *PRKAR2B* or the neighboring gene, *PIK3CG*, remains to be determined. The two experiments described here show how small, previously unidentified alterations that have the potential to contribute to disease may easily be identified in a single SMRT array experiment.

2.3 Discussion

Array CGH is a proven method for accurate, robust and rapid genome-wide assessment of DNA copy number variation. Current users of array CGH technology consider BAC DNA markers positioned at 1–2 Mb intervals to be "high-resolution" coverage. This view has been perpetuated by conventional whole genome analysis tools, such as microsatellite marker analysis of loss of heterozygosity, in which small interspaced "sequence tagged sites" are assayed for genomic imbalance, and the genomic integrity between these sites must be inferred. In contrast, tiling resolution array CGH has the potential to identify minute genomic changes. In this study, we constructed a Sub-Megabase Resolution Tiling-set for array CGH (SMRT array), comprising 32,433 overlapping BAC clones covering the entire human genome. This tiling resolution, combined with the proven sensitivity of array CGH, makes the technique ideal for identifying new genes and will prove useful for unraveling the genetic basis of numerous diseases.

2.4 Methods

2.4.1 BAC Clone Selection, Preparation and Validation

Selection and the map position of the 32,433 clones has been described previously and is available at The Children's Hospital Oakland Research Institute (http://bacpac.chori.org/genomicRearrays.php). We Validated clone identity by comparing *Hind*III fingerprints to the FPC BAC fingerprint database (McPherson et al., 2001) (http://genome.wustl.edu/projects/human/index.php?fpc=1). These clones provide ~1.5 fold coverage of the human genome, giving an approximate resolution of 80 Kb (*i.e.*, 2/3 of an average BAC clone).

2.4.2 Array Production from BAC DNA

We prepared The DNA samples to be spotted on the array by PCR using linkers (primer sequences available upon request). The protocol for linker mediated PCR was previously described (Watson et al., 2004). We precipitated The PCR products with ethanol and redissolved in an MSP printing solution (Telechem), denatured them by boiling and re-arrayed them for robotic printing in triplicate using a VersArray ChipWriter Pro (BioRad). This arrayer uses a 12 x 4 array of SMP2.5 Stealth Micro Spotting Pins (Telechem/ArrayIT) depositing DNA spots of 0.8 nl at ~ 1 µg/µl at 133 micrometer distances. We spotted the entire set of 32,433 solutions in triplicate onto 2 aldehyde-coated slides. Limited numbers of SMRT arrays are available on a cost recovery basis.

2.4.3 DNA labelling and hybridization

We labeled 400 ng of test and reference DNA separately using Cyanine-3 and Cyanine-5 dCTPs according to a random priming protocol previously described (Garnis et al., 2003). Before hybridization, we combined the DNA probes and purified them using ProbeQuant Sephadex G-50 Columns (Amersham) to remove unincorporated nucleotides. We then added 200 µg human Cot-1 DNA (Invitrogen), precipitated the mixture and re-suspended in 100 µl DIG Easy hybridization solution (Roche) containing sheared herring sperm DNA (Sigma-Aldrich) and yeast tRNA (Calbiochem). The probe was denatured at 85°C for 10 minutes and repetitive sequences were blocked at 45°C for 1 hr before hybridization. We carried out Prehybridization in the same buffer. We applied the probe mixture to the slide surface, fixed the coverslips and incubated them at 42°C for 36 hours. We washed the arrays five times for 5 min each in 0.1X saline sodium citrate, 0.1% SDS at room temperature with agitation. We then rinsed each array repeatedly in 0.1X saline sodium citrate and dried by centrifugation.

2.4.4 Array Imaging and Analysis

We imaged hybridized slides using a CCD based imaging system (Arrayworx eAuto, Applied Precision) and analyzed with SoftWoRx Tracker Spot Analysis software. We averaged The ratios of the triplicate spots and calculated standard deviations (SD). All spots with SDs >0.075 or signal to noise ratios <20 were removed from the analysis. We used Custom viewing software (SeeGH) to visualize all data as Log2 ratio plots where each dot represents one BAC. This software is available upon request.

Reference male versus reference female hybridization detected no unexpected gains or losses and random variability of $\log_2$ ratios are not observed (Supplementary Figure 2). Furthermore, owing to overlapping clone coverage, a single clone with aberrant signal ratio would not be considered an amplification or deletion. Finally, since the clones are not spotted in the order of their map position, adjacent clones are distributed throughout our array.

Figure 2.1. Detection of two-fold copy number changes in TAT-1 lymphoma cell line on chromosome arms 8q and 18q.

PLACEHOLDER INSERT FIGURE 2.1

Figure 2.2. Whole genome SMRT array CGH of lung cancer cell line H526.

PLACEHOLDER INSERT FIGURE 2.2

Figure 2.3. Amplification of chromosome 8q24.12–.13 in colorectal cancer cell line COLO320.

PLACEHOLDER INSERT FIGURE 2.3

Figure 2.4. Identification of a novel microamplification by tiling resolution array CGH in COLO320.

PLACEHOLDER INSERT FIGURE 2.4

Figure 2.5. Identification of microdeletions.

PLACEHOLDER INSERT FIGURE 2.5

2.5 References

Albertson, D.G., Ylstra, B., Segraves, R., Collins, C., Dairkee, S.H., Kowbel, D., Kuo, W.L., Gray, J.W. & Pinkel, D. (2000). Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat Genet*, **25,** 144-6.

Bruder, C.E., Hirvela, C., Tapia-Paez, I., Fransson, I., Segraves, R., Hamilton, G., Zhang, X.X., Evans, D.G., Wallace, A.J., Baser, M.E., Zucman-Rossi, J., Hergersberg, M., Boltshauser, E., Papi, L., Rouleau, G.A., Poptodorov, G., Jordanova, A., Rask-Andersen, H., Kluwe, L., Mautner, V., Sainio, M., Hung, G., Mathiesen, T., Moller, C., Pulst, S.M., Harder, H., Heiberg, A., Honda, M., Niimura, M., Sahlen, S., Blennow, E., Albertson, D.G., Pinkel, D. & Dumanski, J.P. (2001). High resolution deletion analysis of constitutional DNA from neurofibromatosis type 2 (NF2) patients using microarray-CGH. *Hum Mol Genet*, **10,** 271-82.

Denyssevych, T., Lestou, V.S., Knesevich, S., Robichaud, M., Salski, C., Tan, R., Gascoyne, R.D., Horsman, D.E. & Mayer, L.D. (2002). Establishment and comprehensive analysis of a new human transformed follicular lymphoma B cell line, Tat-1. *Leukemia*, **16,** 276-83.

Dreyling, M.H., Bullinger, L., Ott, G., Stilgenbauer, S., Muller-Hermelink, H.K., Bentz, M., Hiddemann, W. & Dohner, H. (1997). Alterations of the cyclin D1/p16-pRB pathway in mantle cell lymphoma. *Cancer Res*, **57,** 4608-14.

Fiegler, H., Carr, P., Douglas, E.J., Burford, D.C., Hunt, S., Scott, C.E., Smith, J., Vetrie, D., Gorman, P., Tomlinson, I.P. & Carter, N.P. (2003). DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer*, **36,** 361-74.

Forozan, F., Karhu, R., Kononen, J., Kallioniemi, A. & Kallioniemi, O.P. (1997). Genome screening by comparative genomic hybridization. *Trends Genet*, **13,** 405-9.

Garnis, C., Baldwin, C., Zhang, L., Rosin, M.P. & Lam, W.L. (2003). Use of complete coverage array comparative genomic hybridization to define copy number alterations on chromosome 3p in oral squamous cell carcinomas. *Cancer Res*, **63,** 8582-5.

Garnis, C., Campbell, J., Zhang, L., Rosin, M.P. & Lam, W.L. (2004a). OCGR array: an oral cancer genomic regional array for comparative genomic hybridization analysis. *Oral Oncol*, **40,** 511-9.

Garnis, C., Coe, B.P., Ishkanian, A., Zhang, L., Rosin, M.P. & Lam, W.L. (2004b). Novel regions of amplification on 8q distinct from the MYC locus and frequently altered in oral dysplasia and cancer. *Genes Chromosomes Cancer*, **39,** 93-8.

Girard, L., Zochbauer-Muller, S., Virmani, A.K., Gazdar, A.F. & Minna, J.D. (2000). Genome-wide allelotyping of lung cancer identifies new regions of allelic loss, differences between small cell lung cancer and non-small cell lung cancer, and loci clustering. *Cancer Res*, **60,** 4894-906.

Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F. & Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258,** 818-21.

Kim, S., Domon-Dell, C., Wang, Q., Chung, D.H., Di Cristofano, A., Pandolfi, P.P., Freund, J.N. & Evers, B.M. (2002). PTEN and TNF-alpha regulation of the intestinal-specific Cdx-2 homeobox gene through a PI3K, PKB/Akt, and NF-kappaB-dependent pathway. *Gastroenterology*, **123,** 1163-78.

Knuutila, S., Autio, K. & Aalto, Y. (2000). Online access to CGH data of DNA sequence copy number changes. *Am J Pathol*, **157,** 689.

Kraus, J., Pantel, K., Pinkel, D., Albertson, D.G. & Speicher, M.R. (2003). High-resolution genomic profiling of occult micrometastatic tumor cells. *Genes Chromosomes Cancer*, **36,** 159-66.

Levin, N.A., Brzoska, P., Gupta, N., Minna, J.D., Gray, J.W. & Christman, M.F. (1994). Identification of frequent novel genetic alterations in small cell lung carcinoma. *Cancer Res*, **54,** 5086-91.

McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K., Fulton, R., Kucaba, T.A., Wagner-McPherson, C., Barbazuk, W.B., Gregory, S.G., Humphray, S.J., French, L., Evans, R.S., Bethel, G., Whittaker, A., Holden, J.L., McCann, O.T., Dunham, A., Soderlund, C., Scott, C.E., Bentley, D.R., Schuler, G., Chen, H.C., Jang, W., Green, E.D., Idol, J.R., Maduro, V.V., Montgomery, K.T., Lee, E., Miller, A., Emerling, S., Kucherlapati, Gibbs, R., Scherer, S., Gorrell, J.H., Sodergren, E., Clerc-Blankenburg, K., Tabor, P., Naylor, S., Garcia, D., de Jong, P.J., Catanese, J.J., Nowak, N., Osoegawa, K., Qin, S., Rowen, L., Madan, A., Dors, M., Hood, L., Trask, B., Friedman, C., Massa, H., Cheung, V.G., Kirsch, I.R., Reid, T., Yonescu, R., Weissenbach, J., Bruls, T., Heilig, R., Branscomb, E., Olsen, A., Doggett, N., Cheng, J.F., Hawkins, T., Myers, R.M., Shang, J., Ramirez, L., Schmutz, J., Velasquez, O., Dixon, K., Stone, N.E., Cox, D.R., Haussler, D., Kent, W.J., Furey, T., Rogic, S., Kennedy, S., Jones, S., Rosenthal, A., Wen, G., Schilhabel, M., Gloeckner, G., Nyakatura, G., Siebert, R., Schlegelberger, B., Korenberg, J., Chen, X.N., Fujiyama, A., Hattori, M., Toyoda, A., Yada, T., Park, H.S., Sakaki, Y., Shimizu, N., Asakawa, S., et al. (2001). A physical map of the human genome. *Nature*, **409,** 934-41.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B.M., Gray, J.W. & Albertson, D.G. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, **20,** 207-11.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. & Albertson, D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet*, **29,** 263-4.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T. & Lichter, P. (1997). Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20,** 399-407.

Veltman, J.A., Fridlyand, J., Pejavar, S., Olshen, A.B., Korkola, J.E., DeVries, S., Carroll, P., Kuo, W.L., Pinkel, D., Albertson, D., Cordon-Cardo, C., Jain, A.N. & Waldman, F.M. (2003). Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Res*, **63,** 2872-80.

Veltman, J.A., Schoenmakers, E.F., Eussen, B.H., Janssen, I., Merkx, G., van Cleef, B., van Ravenswaaij, C.M., Brunner, H.G., Smeets, D. & van Kessel, A.G. (2002). High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am J Hum Genet*, **70,** 1269-76.

Watson, S.K., deLeeuw, R.J., Ishkanian, A.S., Malloff, C.A. & Lam, W.L. (2004). Methods for high throughput validation of amplified fragment pools of BAC DNA for constructing high resolution CGH arrays. *BMC Genomics*, **5,** 6.

Weiss, M.M., Kuipers, E.J., Postma, C., Snijders, A.M., Siccama, I., Pinkel, D., Westerga, J., Meuwissen, S.G., Albertson, D.G. & Meijer, G.A. (2003). Genomic profiling of gastric cancer predicts lymph node status and survival. *Oncogene*, **22,** 1872-9.

Wells, D. & Levy, B. (2003). Cytogenetics in reproductive medicine: the contribution of comparative genomic hybridization (CGH). *Bioessays*, **25,** 289-300.

Wessendorf, S., Fritz, B., Wrobel, G., Nessling, M., Lampel, S., Goettel, D., Kuepper, M., Joos, S., Hopman, T., Kokocinski, F., Dohner, H., Bentz, M., Schwaenen, C. & Lichter, P. (2002). Automated screening for genomic imbalances using matrix-based comparative genomic hybridization. *Lab Invest*, **82,** 47-60.

Wilhelm, M., Veltman, J.A., Olshen, A.B., Jain, A.N., Moore, D.H., Presti, J.C., Jr., Kovacs, G. & Waldman, F.M. (2002). Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer. *Cancer Res*, **62,** 957-60.

Xiu, M., Kim, J., Sampson, E., Huang, C.Y., Davis, R.J., Paulson, K.E. & Yee, A.S. (2003). The transcriptional repressor HBP1 is a target of the p38 mitogen-activated protein kinase pathway in cell cycle regulation. *Mol Cell Biol*, **23,** 8890-901.

Chapter 3: SeeGH – A software tool for visualization of whole genome array comparative genomic hybridization data.

A version of this chapter has been previously published as:

Chi, B, deLeeuw, RJ, Coe BP, MacAulay, C, Lam WL (2004). SeeGH – A software tool for visualization of whole genome array comparative genomic hybridization data BMC Bioinformatics **9**(5):13 doi:10.1186/1471-2105-5-13

3.1 Introduction

Metaphase comparative genomic hybridization (CGH) is a molecular cytogenetic technique used to detect segmental DNA copy number differences between two samples of DNA (Kallioniemi et al., 1992). This is accomplished by a competitive hybridization of two differentially labeled samples to normal metaphase chromosomes, allowing the detection of single copy number changes at a resolution of 10-20Mb(Kallioniemi et al., 1992). Array CGH improves on the resolution of copy number profiling by utilizing discrete genomic loci spotted onto glass microscope slides as opposed to metaphase chromosomes as the hybridization target (Snijders et al., 2001). In array CGH the resolution in detecting segmental copy number changes is limited only by the distance between and size of the genomic DNA segments spotted on the array. With the completion of the human and mouse genome sequence (Lander et al., 2001; Waterston et al., 2002) it is possible to construct arrays consisting of a tiling set of DNA segments spanning the entire genome. Currently this approach allows the screening of tens of thousands of genomic segments for copy number alterations in a single experiment. After co-hybridization of differentially labeled DNA samples to an array, two high resolution fluorescence images, one for each labeled probe, are generated. Signal ratios for each clone which act as a proxy for copy number are obtained from these images using one of the many available array analysis software packages. However, map visualization of tens of thousands of spot data points is a daunting task. Many groups simply use Microsoft Excel to display individual plots of each region, however the failure of excel to display multiple plots in an interactive fashion as well as the limitation to 65535 rows of data limits its functionality in high resolution aCGH analysis. Here we present a visualization tool called SeeGH that translates spot signal ratio data from array CGH experiments to displays of high resolution, segmentally annotated chromosome profiles resembling a conventional CGH karyotype diagram facilitating the detection of genetic alterations.

3.2 Software Environment and Information Sources

SeeGH was created using Borland's C++Builder6 development platform and programmed using the language C++. Structured Query Language (SQL) was embedded in the C++ code to make queries to the backend database, MySQL version 4.0 (http://mysql.com/downloas/mysql-4.0.html). MySQL was chosen as the database server since it is publicly available and capable of handling large data files with high data throughput. The software was developed on Microsoft Windows 2000 (service pack 2) and tested for compatibility with Windows XP. Therefore, SeeGH should function on any windows based machine running windows 2000 or later operating system.

Human physical map information used in the example presented here was obtained from the April 2003 assembly on the UCSC Genome Browser Gateway website (http://genome.ucsc.edu). The SeeGH software, source code, and documentation are publicly available upon request (http://bccrc.ca/cg/ArrayCGH_Group.html).

We demonstrate the use of SeeGH by viewing array CGH data obtained by co-hybridizing tumor cell line DNA, labeled with cyanine-5, and normal male DNA, labeled with cyanine-3, to an array constructed from a Human "32k" BAC Re-array Clone set (http://bccrc.ca/cg/ArrayCGH_Group.html). This array contains 32,433 BAC clone derived DNA segments spotted in triplicate on two microarray slides. To facilitate explanations of data processing, in our description below we will follow a single BAC clone (RP11-6J2) from array production to final display in SeeGH. Amplified DNA product from the BAC RP11-6J2 was spotted in triplicate from well D06 of a 384 well plate in the same manner as the remaining 32,433 BAC clones which make up the array. Experimental details for the construction and use of our 32,433 loci CGH array are described elsewhere (Ishkanian et al., 2004). Briefly, array CGH is based on, homologous sequences from each probe competitively hybridizing to the three spots representing a single clone. Post hybridization, two high resolution 16 bit TIFF images, one derived from each fluorescently labeled probe, were obtained using an Arrayworx eAuto CCD based scanner (Applied Precision Instruments). These two images were then transferred to SoftWorx Tracker analysis software (Applied Precision Instruments) and paired for spot segmentation and feature extraction. Spot annotation information (e.g. signal ratio, and signal to noise ratio) for each image pair were then exported to a tab delimited text file. RP11-6J2 is represented in this output file by three unique rows describing the features for each of the triplicate spots.

3.3 Results and Discussion:

3.3.1 Overview of Data Flow

Data from the tab-delimited output file are filtered to remove unnecessary information output by the SoftWorx Tracker software before converting replicate spot data into single data records of standard deviations and averaged spot ratios. These filtered records and experiment identifiers are then filed in the database. One of two routes are utilized for displaying information from the database, direct for further annotation and via a data converter as positional and ratio data. In addition, chromosome specific information such as base pair position of each chromosome band is routed through the data converter for presentation (**Figure 3.1**)

3.3.2 Input Requirements

To accommodate output from various scanner/analyzer software packages, the only input requirement of SeeGH is a tab delimited text file with the following six fields for each array spot: a unique identifier, the base pair starting position of the clone on the chromosome, chromosome number, channel 1 signal to noise ratio (Ch1 SNR), channel 2 signal to noise ratio (Ch2 SNR), and $\log_2$ spot ratio (**Figure 3.2** buttons 1-6). Two additional fields, clone name and accession number may contain further text information (**Figure 3.2** buttons 7-8). Additional fields of miscellaneous data may be included in the tab delimited text file as the user is required to enter the total number of columns and the specific column number for each of the required data fields

(**Figure 3.2** buttons 1-9). For example, the text file exported from SoftWorx Tracker contains a total of 72 fields for each spot imaged from the array.

Input files can be located and opened by using the Browse button or by manually entering their file path (**Figure 3.2** button 12). Because array CGH experiments contain replicate spots to ensure high confidence in spot ratios SeeGH was designed with the capability of accepting up to five replicate spots (**Figure 3.2** button 10). Replicate spot ratio records are identified by their use of a common unique identifier and these spots are averaged and their standard deviations calculated. In a mantle cell lymphoma versus normal male hybridization, our example clone RP11-6J2 demonstrated triplicate spot ratios of -0.02690442, 0.009741764, and 0.04698608 respectively. Averaging these spots resulted in an average spot ratio of 0.0099414 and a standard deviation of 0.0369457. If replicate spots have been previously averaged then SeeGH requires that the 'Number of Replicates' field should be set to one and the spot standard deviations must be included in the records of the input file (**Figure 3.2** buttons 10,11).

SeeGH also requires the user to enter a basic description for each data file. The required fields are bar code/unique identifier, disease type, experimenter, and date (**Figure 3.2** buttons 13-16). Additional information may be entered into the "Comments" field but is not required (**Figure 3.2** button 17).

3.3.3 Data Filtering and Storage

Once all the required information has been entered, pressing the 'Load File' button will create a record in the 'Existing Data' table containing the five file description fields (BarCode, Disease_Type, Date, Experimenter, and Comments). The BarCode field is used as a key to generate 25 new tables which consist of a filtered input data table and one table per uniquely identified chromosome (for human material 1-22, X and Y). For our example experiment BarCode 10300047 points to these 25 new tables and the information for all three replicates of RP11-6J2 are located in the filtered input data table. The calculated average ratio and standard deviation as well as the lowest signal to noise ratio (SNR) for the three spots for each channel are placed into the appropriate chromosome table along with the required annotation information reducing the three replicate records to a single chromosome record. For example, the data for RP11-6J2 from our experiment, which is a clone derived from chromosome 6, would be stored in chromosome table 10300047_chr6.

3.3.4 Data Presentation

3.3.4.1Genomic View

The Genomic View window appears automatically after new data has been loaded into the database (**Figure 3.3**). The Genomic View consists of 24 tiles (one for each unique chromosome) each measuring 100 by 150 pixels with the origin pixel position (0, 0) at the bottom left corner for each tile. In order to graphically plot chromosomes and spot ratios, SeeGH takes the base pair information for each chromosome and spot ratio, converts them to pixel position coordinates, and draws the image of each chromosome and spot ratio into a tile using the pixel position coordinates.

The chromosomal information used to draw the chromosomes is contained in 49 text files. For each chromosome arm there is a corresponding file that contains band names and base pair positions. The p and q arms of the 22 auotsomes and 2 sex chromosomes are represented in a total of 48 files. The 49th file contains information about total chromosome lengths and individual arm lengths for each chromosome. In the example presented in this paper we used information from the UCSC April 2003 assembly to create these files. These files are included with the software and can be updated with new chromosomal mapping information as it becomes available. Using this information, the total base pair length of each chromosome arm is converted into pixel position y-coordinates using a base pair to pixel conversion formula (pixel position y-coordinate = base pair position / 1,700,000). This same formula is used to calculate each chromosome band's start and end pixel position y-coordinate from the 48 band information files. Chromosomes are drawn in the Genomic View with the x-coordinate starting at pixel 10 and having a width of 20 pixels.

The base pair start information for spot ratios is retrieved from the 24 chromosome tables created in the database for each experiment and converted into pixel position y-coordinates using the same formula. The x-coordinate for each spot ratio is calculated using a similar pixel conversion formula (pixel position of x-coordinate = X_Axis + spot ratio * One_Ratio). One_Ratio is given a default value of 10 pixels and X_Axis is set to a constant of 50. Therefore the y and x co-ordinates of our example clone (RP11-6J2) are 68, 60 (y-coordinate = 115712602 / 1700000, x-coordinate = 60 + 0.00994114 * 10).

Chromosomes and corresponding spot ratios are plotted on each tile using the calculated x and y coordinates. The 24 resulting tiles are displayed in the Genomic View as an 8 by 3 grid (**Figure 3.3** button 1). The Genomic View allows manipulation of several display parameters: ratio lines, ratio width, standard deviation filters, and signal to noise filters.

Ratio lines can be displayed at +/- 0.5, 1.0, 1.5 and 2.0, with a default display of +/- 1.0 (**Figure 3.3** buttons 2-5). Ratio width can be increased or decreased by inputting a numerical modifier that expands or contracts the x-coordinates of the spot ratios relative to the X_Axis (pixel position of x-coordinate = X_Axis + spot ratio * (One_Ratio + modifier)) (**Figure 3.3** button 6). Another feature available in SeeGH is the ability to display only those spots that meet user defined criteria. These criteria include a standard deviation cutoff and/or a minimum signal to noise ratio for either Ch1 SNR or Ch2 SNR (**Figure 3.3** buttons 7-9). The 8 by 3 tiled image can be saved as a bitmap which can be viewed or printed using any image viewing software (**Figure 3.3** button 10).

While in the Genomic View, the user can also search for a specific spot based on unique identifier, clone name, or accession number. An example search is shown in **Figure 3.3**: button 11 and **Figure 3.4**: buttons 1-2. Once located, the appropriate Chromosome View is automatically opened with a line through the chromosome image at the appropriate spot loci and the spot is highlighted. A Chromosome View can also be opened without the need for inputting a search term by selecting a chromosome with the left mouse button and choosing a magnification from the pop-up menu (**Figure 3.3** button 12).

3.3.4.2 Chromosome View

The Chromosome View displays the selected chromosome tile as a 649 by 673 pixel image with a zoom factor incorporated into the base pair to pixel conversion formula (pixel position y-coordinate = base pair position * zoom factor / 1,700,000) which increases or decreases the total pixel length for the chromosome image. The x-coordinates for displaying the chromosome now start at pixel 100 and have a width of 40 pixels. The x-coordinates for spot ratios are calculated using the same formula (X_Axis + spot ratio * Ratio_One) with Ratio_One equal to 50 pixels and X_Axis set to a constant of 375. For our demonstration clone the coordinates become 272,375 in the tile.

In the Chromosome View, the user is given many of the same features available in the Genomic View: hiding spots based on standard deviation criteria or signal to noise ratios, changing ratio widths of the spot image, adding or deleting ratio lines of 0.5, 1.0, 1.5 and 2.0, and saving the image as a bitmap (**Figure 3.5** buttons 1-5). However, the Chromosome View provides many additional features that are unavailable in the Genomic View: the display of standard deviations for replicate spots, flagging of high standard deviations, mouse-over activated spot information, continuous zoom, the ability to scroll along the chromosome, display UCSC regional information, and clear search results (**Figure 3.5** buttons 6-12).

Spot standard deviations, are displayed as a line through each spot and can be turned on or off simply by checking or unchecking a box in the Chromosome View (**Figure 3.5 & 3.6**). In addition, standard deviation lines which exceed a user defined value (**Figure 3.5** button 7) can be flagged in red. One key feature added in the Chromosomal View is the 'mouse-over' functionality which displays specific spot information when the mouse cursor is positioned over a spot. The spot information displayed consists of the clone name, accession number, unique id, base pair starting position, ratio, standard deviation, and signal to noise ratio for both channel 1, and channel 2 (**Figure 3.5** button 8). The zoom feature in Chromosome View functions the same as in the Genomic View, and can be accessed multiple times for limitless magnification (**Figure 3.5** button 9). The Chromosome View can be scrolled up or down at a rate set by the user (**Figure 3.5** button 10). UCSC base pair positions are given for the displayed image (**Figure 3.5** button 11). The final feature clears the highlighted results of the Search function (**Figure 3.5** button 12).

3.3.4.3 Accessing Previously Entered Data

The Existing Data window contains a list of all the files that have been loaded into the program (**Figure 3.6** buttons 1-3). The displayed list can be limited by searching for data sets with specific search criteria (**Figure 3.6** buttons 1-2). Alternately, the list can be ordered by selecting a field from the drop down menu and performing a search function without entering any search criteria. A data set can be selected by highlighting a row in the list of existing data (**Figure 3.6** button 3). Once selected, the data set can either be viewed or deleted (**Figure 3.6** buttons 4-5). Deleting a data set removes all tables from the database, whereas, viewing opens a Genomic View for that data.

3.4 Conclusions

We have developed an array CGH data viewing tool which improves upon conventional viewing methods by displaying data in dynamically explorable conventional karyotype diagrams. This holistic genome view allows the user to easily recognize patterns in a genome wide data set while quickly identifying the chromosome bands implicated, a feature lacking in excel based approaches which display data as linear plots which are not directly correlated to chromosomal regions. In SeeGH, a user has the ability to quickly access data point information such as clone name, NCBI sequence accession number, and base pair starting position which allows for precise localization of genetic alteration boundaries. In addition, a user can easily filter data for quality assurance by removing data points which do not meet signal to noise or standard deviation criteria.

SeeGH is simple to set up, requiring only MySQL version 4.0 and runs under Microsoft Windows 2000 or later operating systems. The open design of SeeGH allows easy for specific needs and future plans to include the incorporation of features for multiple experiment comparisons.

3.5 Availability and Requirements

Project Name: SeeGH

Project Homepage: http://www.bccrc.ca/ArrayCGH

Operating System: Microsoft Windows 2000 or later

Programming Language: C++, SQL

Other Requirements: MySQL database

License: Academic Software License

Any Restrictions to use by non-academics: Yes

Figure 3.1

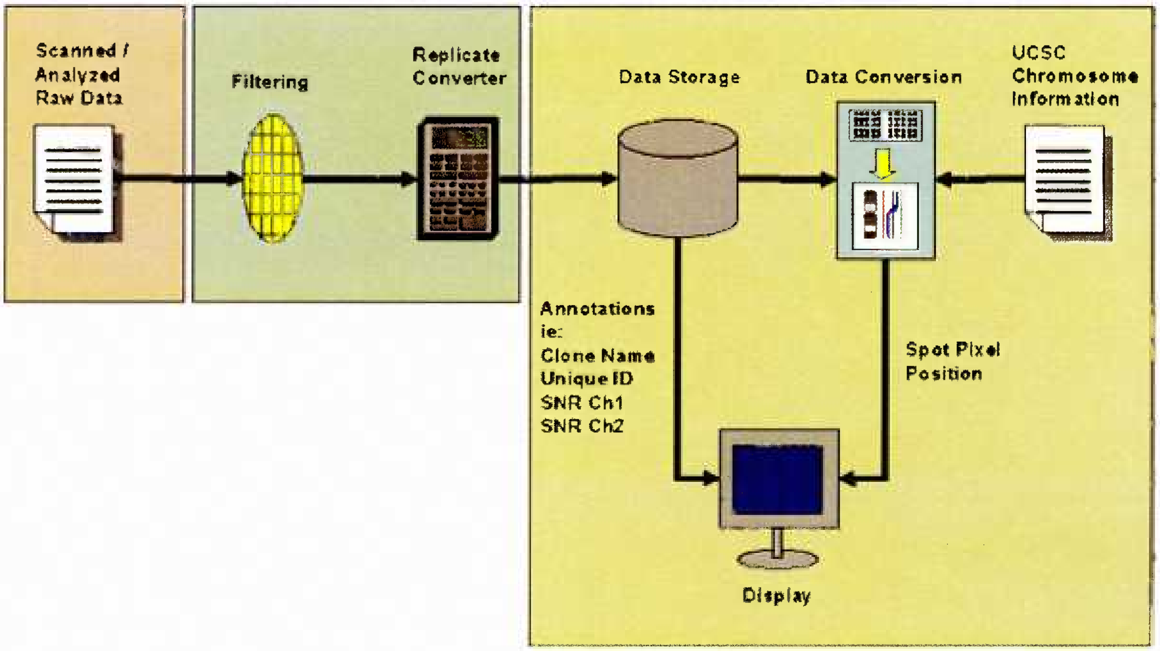
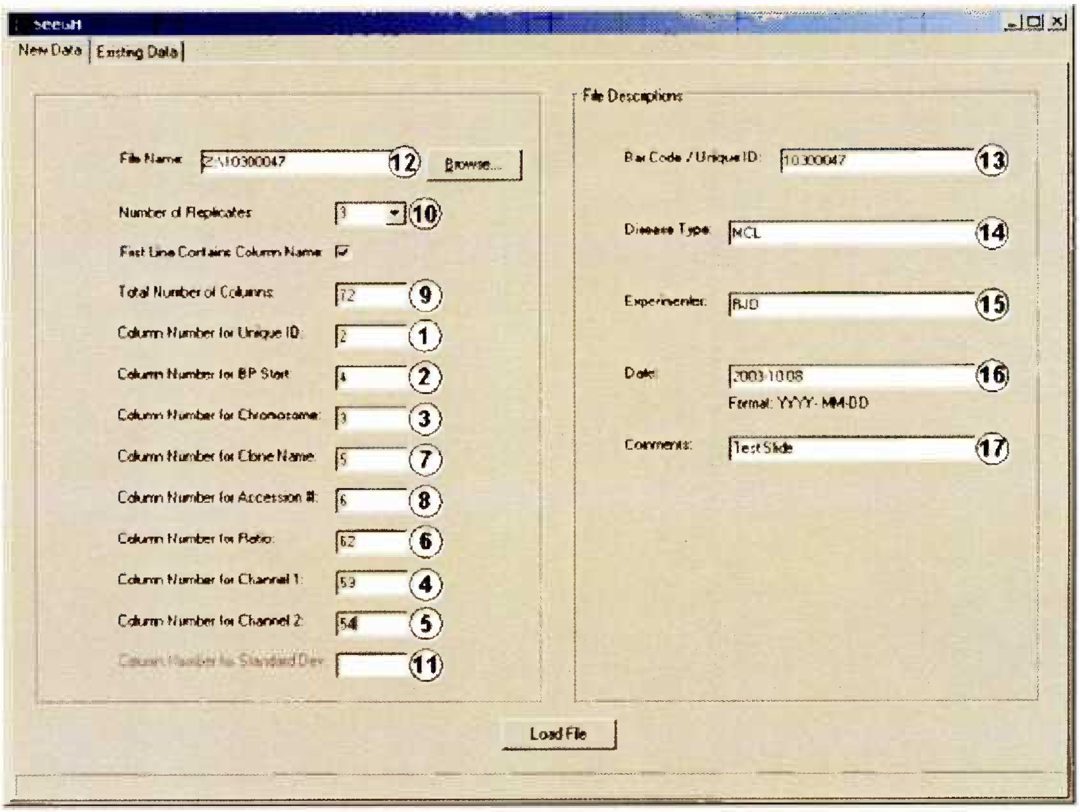


Figure 3.1. Overall View of SeeGH Data Flow. The user inputs data formatted as a tab delimited text file. The relevant data is then extracted from the text file via a filtering algorithm and replicate ratios and features are averaged before being stored in an SQL database. Ratio data is displayed via a data converter which converts ratio data to x, y plot coordinates, whereas annotation information is read directly from the SQL database.

Figure 3.2



File Name: Z:\10300047

Number of Replicates: 3

First Line Contains Column Name: ☑

Total Number of Columns: 72

Column Number for Unique ID: 2

Column Number for BP Start: 4

Column Number for Chromosome: 3

Column Number for Clone Name: 5

Column Number for Accession #: 6

Column Number for Ratio: 62

Column Number for Channel 1: 53

Column Number for Channel 2: 54

Column Number for Standard Dev:

File Descriptions

Bar Code / Unique ID: 10300047

Disease Type: MCL

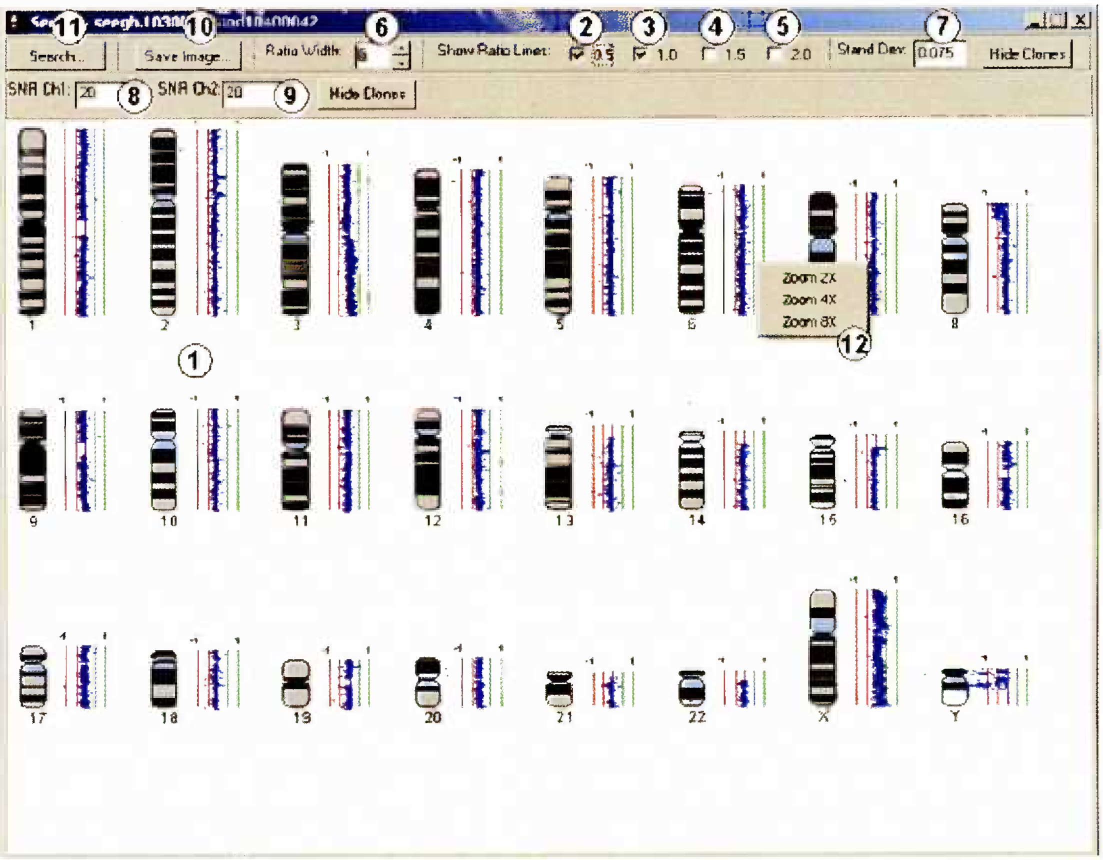
Experimenter: RJD

Date: 2003-10-08
Format: YYYY- MM-DD

Comments: Test Slide

Figure 3.2. SeeGH "New Data" Window. Buttons correspond to descriptions in text.

Figure 3.3



Ratio Width: 5

Show Ratio Lines: ☑ 0.5 ☑ 1.0 □ 1.5 □ 2.0

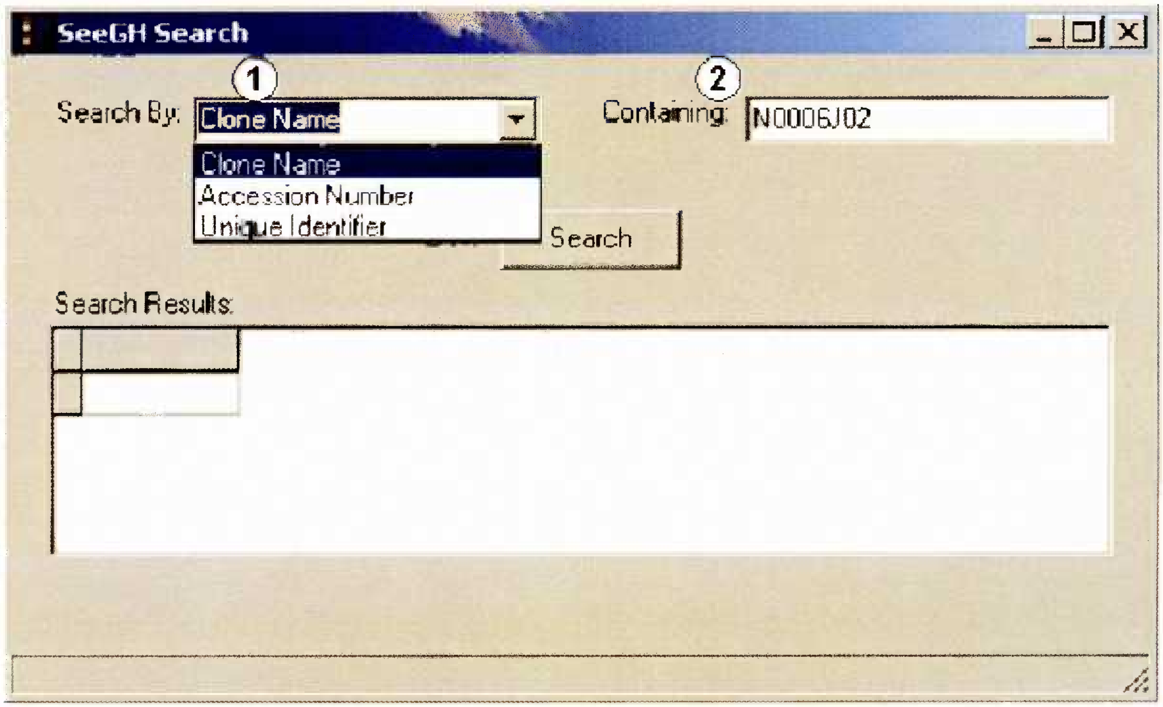
Stand Dev: 0.075

SNR Ch1: 20

SNR Ch2: 20

Figure 3.3. SeeGH "Chromosome View" Window. Reconstructed chromosome 6 array CGH profile from 97,299 array elements. Mantle cell lymphoma DNA (labeled with Cye5) was competitively hybridized with normal male (labeled with Cye3) to an array of 32,433 DNA segments spotted in triplicate (97,299 elements). The information from the 97,299 elements was imported into SeeGH and is displayed. Buttons correspond to descriptions in the text.

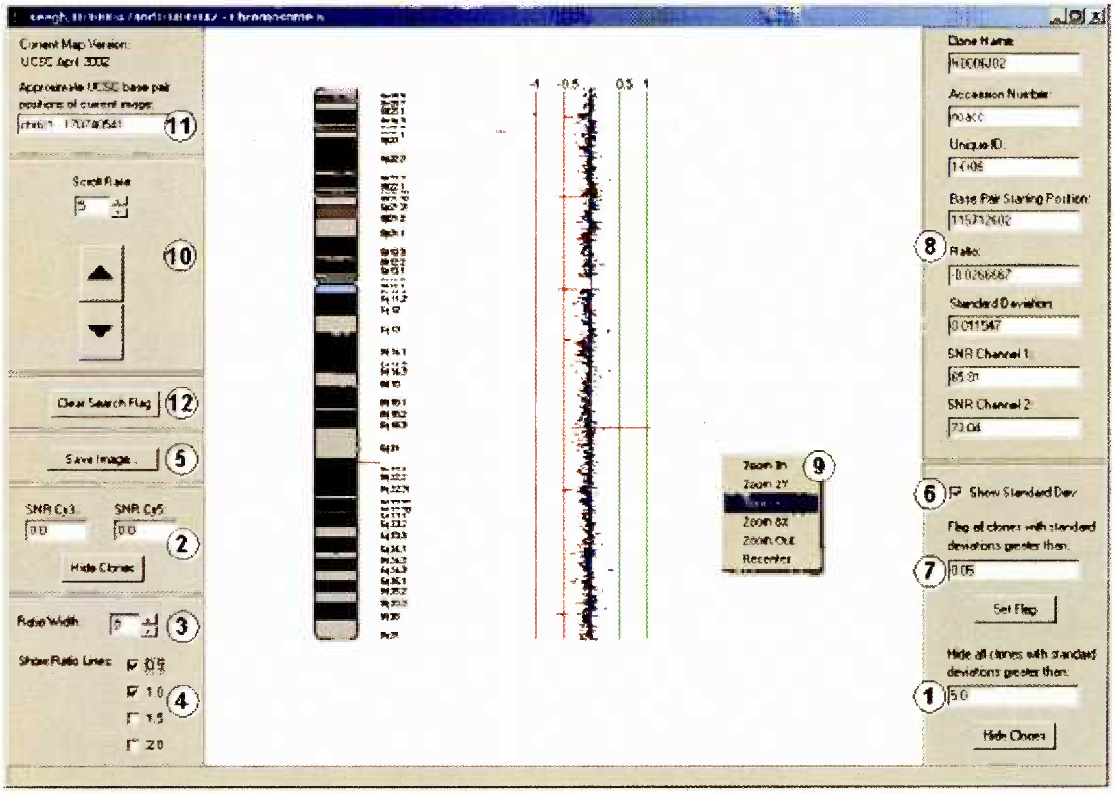
Figure 3.4



N0006J02

Figure 3.4. SeeGH "Search" Window. Buttons correspond to descriptions in the text.

Figure 3.5



Current Map Version:
UCSC April 2002

Approximate UCSC base pair positions of current image:
chr6:1 - 170740541

Scroll Rate: 5

Clear Search Flag

Save Image...

SNR Cy3: 0.0
SNR Cy5: 0.0
Hide Clones

Ratio Width: 5

Show Ratio Lines: ☑ 0.5 ☑ 1.0 □ 1.5 □ 2.0

Clone Name: N0C06J02

Accession Number: no acc

Unique ID: 14006

Base Pair Starting Position: 115712602

Ratio: -0.0266667

Standard Deviation: 0.011547

SNR Channel 1: 65.01

SNR Channel 2: 73.04

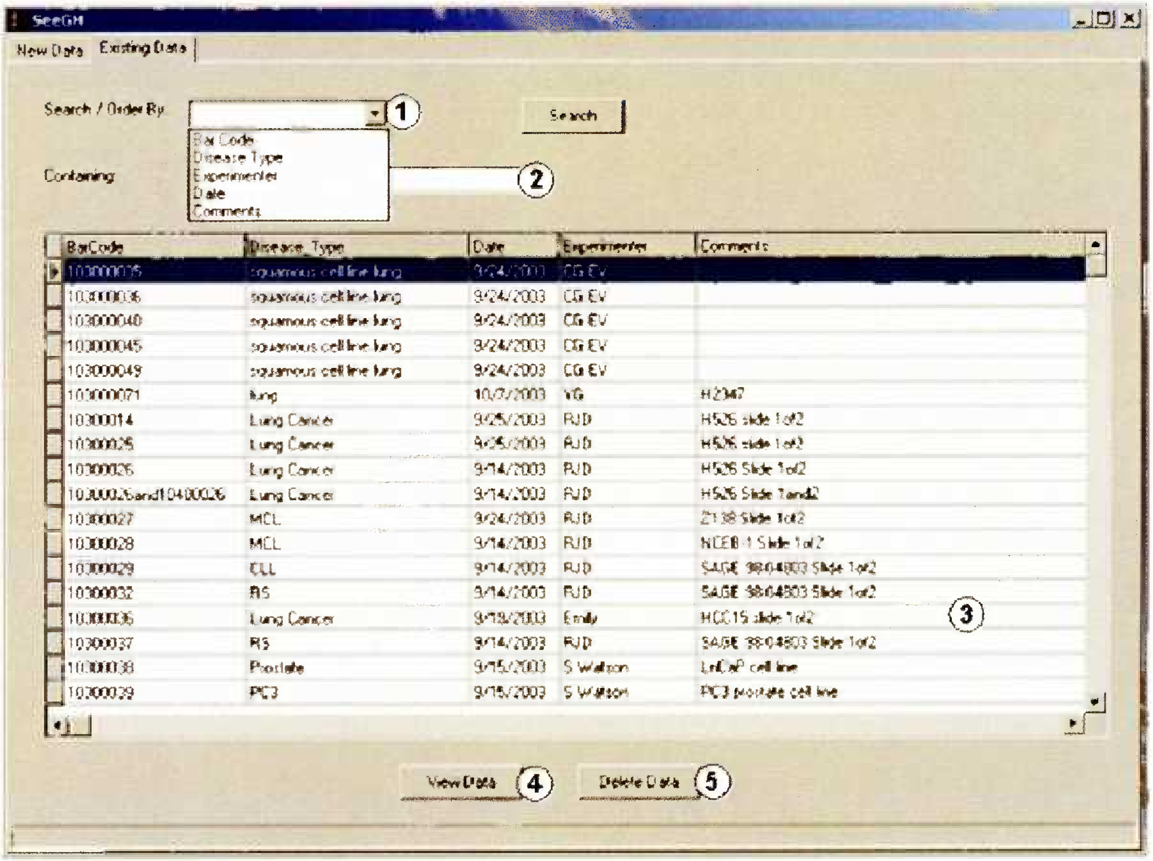
☑ Show Standard Dev

Flag all clones with standard deviations greater than: 0.05
Set Flag

Hide all clones with standard deviations greater than: 5.0
Hide Clones

Figure 3.5. SeeGH "Chromosome View" Window. 1,972 DNA segments are displayed for chromosome 6. The red line through the chromosome denotes the location of the search DNA segment which is highlighted. Horizontal lines through each data point represent standard deviations of the triplicate elements. Buttons correspond to descriptions in the text.

Figure 3.6



BarCode	Disease_Type	Date	Experimenter	Comments
103000035	squamous cell line lung	9/24/2003	CG EV	
103000036	squamous cell line lung	9/24/2003	CG EV	
103000040	squamous cell line lung	9/24/2003	CG EV	
103000045	squamous cell line lung	9/24/2003	CG EV	
103000049	squamous cell line lung	9/24/2003	CG EV	
103000071	lung	10/7/2003	YG	H2347
10300014	Lung Cancer	9/25/2003	RJD	H526 slide 1of2
10300025	Lung Cancer	9/25/2003	RJD	H526 slide 1of2
10300026	Lung Cancer	9/14/2003	RJD	H526 Slide 1of2
10300026and10400026	Lung Cancer	9/14/2003	RJD	H526 Slide 1and2
10300027	MCL	9/24/2003	RJD	2138 Slide 1of2
10300028	MCL	9/14/2003	RJD	NCEB-1 Slide 1of2
10300029	CLL	9/14/2003	RJD	SAGE 38-04803 Slide 1of2
10300032	RS	9/14/2003	RJD	SAGE 38-04803 Slide 1of2
10300036	Lung Cancer	9/19/2003	Emily	HCC15 slide 1of2
10300037	RS	9/14/2003	RJD	SAGE 38-04803 Slide 1of2
10300038	Prostate	9/15/2003	S Watson	LnCaP cell line
10300039	PC3	9/15/2003	S Watson	PC3 prostate cell line

Figure 3.6. SeeGH "Existing Data" Window. Buttons correspond to descriptions in the text.

3.6 References

Ishkanian, A.S., Malloff, C.A., Watson, S.K., DeLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M.A., Ling, V., MacAulay, C. & Lam, W.L. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, **36,** 299-303.

Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F. & Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258,** 818-21.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409,** 860-921.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. & Albertson, D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet*, **29,** 263-4.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigo, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420,** 520-62.

Chapter 4: High-resolution array CGH increases heterogeneity tolerance in the analysis of clinical samples.

A version of this chapter has been previously published as:

Garnis C*, Coe BP*, Lam SL, MacAulay C, Lam WL. (2005) High-resolution array CGH increases heterogeneity tolerance in the analysis of clinical samples. *Genomics*. **85**(6):790-3 © 2005 Elselvier. doi:10.1016/j.ygeno.2005.02.015

*** These authors contributed equally**

4.1 Introduction

Array Comparative genomic hybridization (array CGH) detects segmental DNA copy number gains and losses in tumor genomes. This is achieved by the competitive hybridization of differentially labeled reference and sample genomic DNA to specific genomic loci spotted in an array format, facilitating high resolution scanning for genetic alterations (Carvalho et al., 2004; Garnis et al., 2003; Ishkanian et al., 2004; Lucito et al., 2003; Pollack et al., 1999; Snijders et al., 2001). In order to adapt whole genome array CGH for high through-put analysis of tumor genomes, especially in a clinical setting, this technology would have to not only be applicable to formalin-fixed paraffin-embedded tissue specimens, but also be able to tolerate tissue heterogeneity. Tumor samples are typically highly heterogeneous, containing multiple normal cell types in addition to the cancer cells of interest. Since array CGH identifies genetic alterations by comparing DNA copy number of the cancer cells against those of normal diploid cells, normal cell contamination in a tumor specimen would compromise detection sensitivity.

Microdissection of specific cell populations is a common approach to overcoming tissue heterogeneity. The utility of microdissected archival material in array CGH studies is well documented, however this approach often requires genomic DNA amplification to yield sufficient material for hybridization (Daigo et al., 2001; Massion et al., 2002; Wang et al., 2004). The time consuming microdissection requirement hampers the broad application of this technique and its utility as a high throughput technology. In this report we investigated the heterogeneity tolerance of array CGH and showed that increasing array density directly improves detection sensitivity.

4.2 Results and Discussion

To determine the maximum amount of contaminating normal cells tolerable while allowing detection of single copy amplifications and deletions, we simulated heterogeneity by mixing precise proportions of male (X) and female (XX) DNA samples (**Figure 4.1a**), and then validated our model in clinical specimens. The SMRT array was used for assaying detection sensitivity (de Leeuw et al., 2004; Ishkanian et al., 2004). The Submegabase Resolution Tiling (SMRT) Array was used for assaying detection sensitivity (de Leeuw et al., 2004; Ishkanian et al., 2004). The SMRT array consists of 32,433 bacterial artificial chromosomes arranged in a tiling path that spans the entire genome. Array hybridization protocols were performed as previously described [4]. Briefly, for the heterogeneity titration series 400 ng of test and reference DNA were separately labeled through a random priming reaction with cyanine 3 and cyanine 5 respectively. The probes were precipitated and then combined, denatured, and blocked in a solution containing 100 μg Cot-1 DNA in 45 μl DIG Easy hybridization solution (Roche, Laval, Que.), sheared herring sperm DNA (Invitrogen, Burlington, Ont.), and yeast tRNA (Calbiochem, Mississaga, Ont.). Probe hybridization to the SMRT array occurred over 36 hours at 45°C.

In our simulation experiments, first we observed the expected linear approach to a 1:1 average signal ratio for the X loci as the level of contaminating normal cells increased. Due to the increase in overlap between the ratio distributions between the X chromosome and autosome it became apparent that thresholds would not be appropriate for identifying alterations due to the large percentage of clones which would be falsely classified, and that small alterations would be more difficult to detect than alterations spanning a large number of clones. (**Figure. 4.1b-d**).

Secondly, to model the impact of heterogeneity on the probability of detecting an alteration of a given size, we utilized segments of the X chromosome from the contamination hybridizations to model single copy gains and losses of varying sizes within the autosome. This model was based on using Welch's approximate t-test to compare every 1, 2, 4, 8, 16, 32, 64, 128, 256, and 512 clone window over the entire autosome to the X chromosome and vice versa. Remarkably, we distinguished a 4 clone single copy loss (~0.4 Mb) on the SMRT array under 50% contamination and a 64 clone deletion (~6.4 Mb) under 75% contamination (**Figure 4.2a**). As our results are based solely on the number of clones altered and not the genomic size of an alteration we can infer that a 1 Mbp resolution CGH array when compared to the 0.1Mbp resolution SMRT CGH array, would follow the same detection probabilities and be limited to detecting a 4Mbp single copy loss under 50% contamination and 64Mbp single copy loss under 75% contamination. It appears that this model applies to both single copy loss and single copy gain scenarios (**Figure 4.2**). The more measurements performed over a sequence improved the confidence in detecting alterations, supporting the concept that increasing array resolution reduces the need for microdissection.

We verified the modeled effect of heterogeneity on detection sensitivity using a clinical lung cancer specimen. **Figure 4.3a** shows an H&E stained section from a tumor from a male patient. Histological assessment suggested a mix of 30% tumor cells infiltrated with 70% stromal cells, lymphocytes as well as carbon deposits. Enumeration of tumor and normal cell nuclei in the displayed area counted 28% ±15% cancer cells, in agreement with the initial histological assessment. DNA extracted from this mixed cell population was co-hybridized against differentially labeled normal female DNA (as a reference) onto the SMRT CGH array. Analysis of X chromosome loci yielded the expected average 0.5 log2 ratio (**Figure 4.3b**). Even with the compromised DNA quality from formalin-fixed paraffin-embedded archival material and the high level of normal cell contamination, copy number changes in the tumor DNA were apparent. Large alterations, such as gain of 5p and loss of 5q, as well as high level amplifications (for example, the 2.5 Mb *CCND1* region) were readily detected (**Figure 4.3b,c**). Presumably multi-copy amplifications are easily detectable regardless of clone density.

Given that only ~30% of the cells in our sample are cancerous, the signal ratio for a single copy loss within these cells is expected to be 1.7:2 (0.23 log2 ratio). The signal ratio associated with the 5q loss, which contains the well studied *APC* gene, fits this expectation. More interestingly, we also observed a small 1.49 Mb single copy deletion (across 22 clones) at 2p22 with the same ratio (**Figure 4.3d**). Significantly, this observation fits well into our model of detection sensitivity in mixed tissue which predicted the ability to identify a loss spanning between 4 clones (the threshold for 50% contamination) and the predicted 64 clone threshold for 75%

contamination (**Figure 4.2a**). With the ability to detect copy number changes with only 30% tumor cell content, tedious microdissection is not required when using the SMRT array. This is crucial for the development of array CGH as a clinical screening tool for use in a high throughput setting.

4.3 Conclusions

In conclusion, we have determined the sensitivity of array CGH to single copy changes under heterogeneous conditions. Our results show that increasing array resolution directly improves detection sensitivity to smaller alterations in heterogeneous tissue. Clinical tumor samples are typically highly heterogeneous containing mixed cell types in addition to the malignant cells. As a result microdissection is often required before genome wide analysis with current techniques, posing a barrier to high throughput profiling of tumor genomes and the introduction of array CGH to diagnostic cancer cytogenetics. The extremely high resolution of the SMRT array, and hence its sensitivity to small alterations under high levels of contamination, greatly reduces the need for time consuming microdissection, removing a major hurdle for the introduction of array CGH into the clinical setting.

Figure 4.1. Tissue Heterogeneity simulation. (**a**) Summarizes the composition of the DNA mixture used to mimic normal cell contamination. In order to determine our array's sensitivity to single copy number changes, we set up a titration experiment comparing X chromosome loci to autosomal loci in comparisons of male and female DNA. In this model system we simulated a single copy deletion by hybridizing normal male versus normal female DNA, generating a 1:2 ratio of X chromosomes. Contamination from normal cells was then simulated by spiking varying amounts of female DNA into the male DNA sample. Single copy amplifications were modeled by comparing a 50/50 mixture of male and female DNA against a male DNA reference. In this model, contamination from normal cells was simulated by spiking varying amounts of male DNA into the male/female DNA mixture. We performed SMRT array CGH hybridizations simulating 15, 30, 50, and 75% contamination for both the amplification and deletion models. (**b**) Hybridization data mimicking single copy deletion experiment. Data is shown as the normalized hybridization signal ratio for all of chromosome 1 (yellow) and X chromosome (green) loci plotted versus genomic position. Standard deviations are indicated by vertical bars at each data point. (**c-d**) Summary of the average signal ratios observed for the single copy loss and gain contamination titration hybridizations. The autosomal loci are summarized by yellow bars and the X chromosome loci are summarized by green bars, each at the average observed signal ratios for these regions. One standard deviation from the mean for each region is indicated on the plot. With increasing contamination we observed a linear decrease in ratio separation between the autosome and X chromosome as well as an increase in the overlap between the distributions.

Figure 4.1

Simulated normal cell contamination level	Probe DNA mixture of Male DNA and Female DNA		Reference DNA	X dosage in probe vs reference
Series for simulating single copy loss				
0%	100%	0%	Female	1:2
15%	85%	15%	Female	1.15:2
30%	70%	30%	Female	1.3:2
50%	50%	50%	Female	1.5:2
75%	25%	75%	Female	1.75:2
Series for simulating single copy gain				
0%	50%%	50%	Male	1.5:1
15%	57.5%	42.5%	Male	1.43:1
30%	65%	35%	Male	1.35:1
50%	75%	25	Male	1.25:1
75%	87.5%	12.5	Male	1.13:1

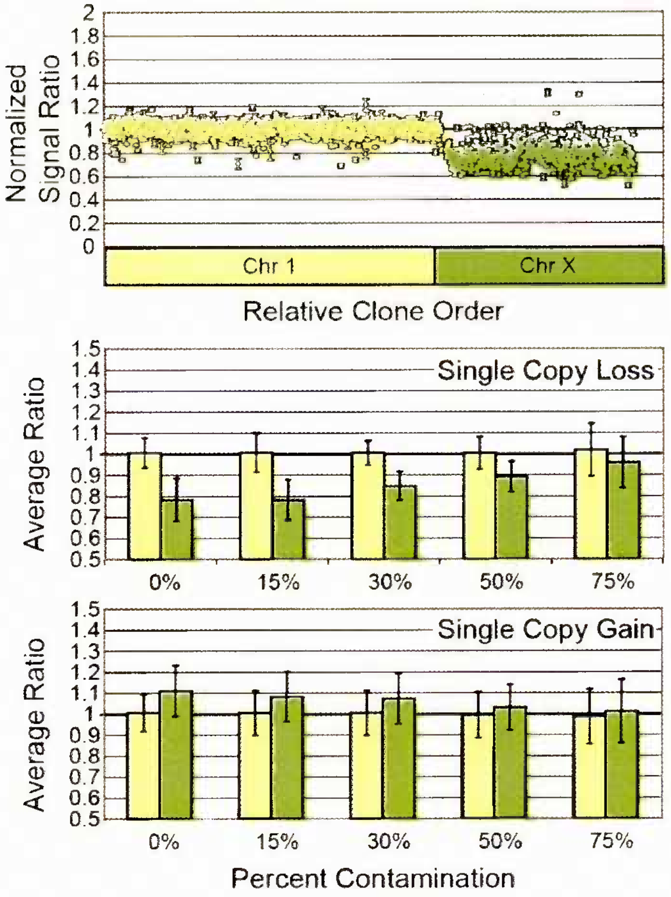


Figure 4.2

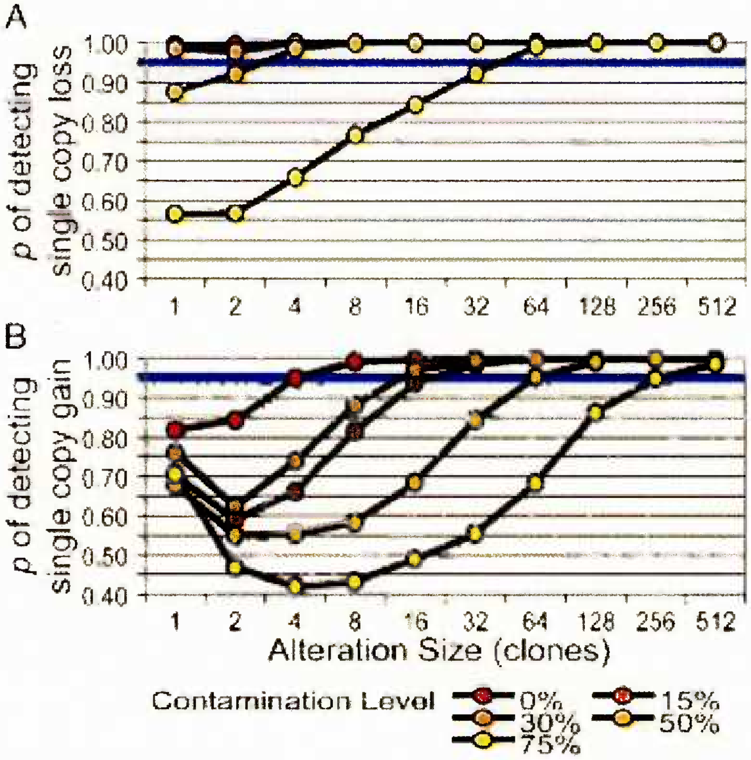


Figure 4.2. Minimal detectable alteration sizes. In order to calculate the probability of detecting a segmental alteration of a particular size at a particular normal cell contamination level we first divided the autosome and X chromosome into segments of 1, 2, 4, 8, 16, 32, 64, 128, 256 and 512 clones (each clone representing three measurements). The autosomal segments were used to simulate areas of retention within the X chromosome and the X chromosome segments were used to simulate alterations within the autosome. This was accomplished by first using Welch's approximate t-test to determine if a particular segment from the autosome could be identified as distinct from the X chromosome with a p value of 0.05. The frequency with which the autosomal segments were not identified as being distinct from the X chromosome defined the percent of segments incorrectly detected as altered. The inverse test of segments from the X chromosome being compared to the autosome was used to determine the percentage of segments correctly detected as altered. Due to the different variances between the X chromosome and autosome distributions as well as the nature of the t-test, we cannot assume that the fraction of clones correctly and incorrectly identified as altered will sum to exactly 100%. As such, we calculate the probability of detecting an alteration at a particular contamination level and alteration size as: Probability of detection = Fraction Correctly Identified as Altered / (Fraction Correctly Identified as Altered + Fraction Incorrectly Identified as Altered). Due to the shift from an inter to intra-clone measure of variance associated with alterations spanning only a small number of clones the p values reported for the smallest alterations exhibit a slight overestimate in detection probability, particularly in the highest contamination levels. This is most apparent in the single copy gain scenario as single copy gains (3:2 allele ratio) exhibit less ratio separation from normal than single copy losses(1:2 allele ratio). However since these probabilities are well below the threshold for reliable detection the overestimate does not affect our results.

Figure 4.3

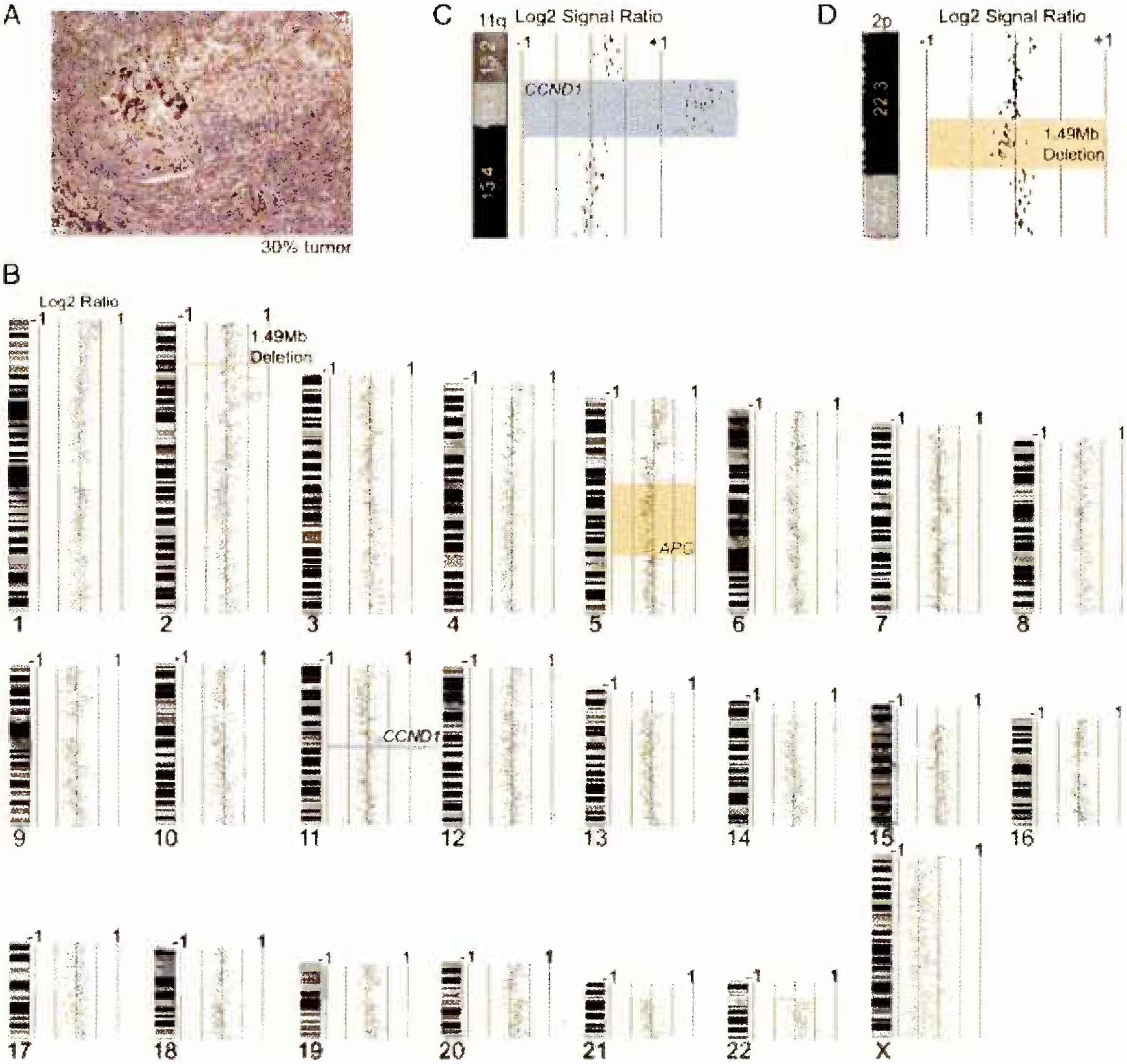


Figure 4.3. SMRT array CGH profile of an archival squamous cell lung tumor containing ~30% tumor cells. (**a**) H&E stained squamous cell lung carcinoma tissue section containing ~30% tumor cells by histological evaluation. (**b**) SeeGH karyogram (Chi et al., 2004) of the DNA extract from the tissue in (**a**). Hybridization was performed as described by Ishkanian et al. except we used 100 ng of tumor DNA against 100 ng of female reference DNA (Ishkanian et al., 2004). Each black dot/line represents a single BAC clone marking its genomic position and log2 signal ratio value. Regions of interest are highlighted. (**c**) Magnified SeeGH view of the highly amplified region on chromosome 11q containing CCND1. (**d**) Magnified SeeGH view of a 1.49 Mb single copy number deletion on chromosome 2p.

4.4 References

Carvalho, B., Ouwerkerk, E., Meijer, G.A. & Ylstra, B. (2004). High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J Clin Pathol*, **57,** 644-6.

Chi, B., DeLeeuw, R.J., Coe, B.P., MacAulay, C. & Lam, W.L. (2004). SeeGH - A software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics*, **5,** 13.

Daigo, Y., Chin, S.F., Gorringe, K.L., Bobrow, L.G., Ponder, B.A., Pharoah, P.D. & Caldas, C. (2001). Degenerate oligonucleotide primed-polymerase chain reaction-based array comparative genomic hybridization for extensive amplicon profiling of breast cancers : a new approach for the molecular analysis of paraffin-embedded cancer tissue. *Am J Pathol*, **158,** 1623-31.

de Leeuw, R.J., Davies, J.J., Rosenwald, A., Bebb, G., Gascoyne, R.D., Dyer, M.J., Staudt, L.M., Martinez-Climent, J.A. & Lam, W.L. (2004). Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum Mol Genet*, **13,** 1827-37.

Garnis, C., Baldwin, C., Zhang, L., Rosin, M.P. & Lam, W.L. (2003). Use of complete coverage array comparative genomic hybridization to define copy number alterations on chromosome 3p in oral squamous cell carcinomas. *Cancer Res*, **63,** 8582-5.

Ishkanian, A.S., Malloff, C.A., Watson, S.K., DeLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M.A., Ling, V., MacAulay, C. & Lam, W.L. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, **36,** 299-303.

Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., West, J.A., Rostan, S., Nguyen, K.C., Powers, S., Ye, K.Q., Olshen, A., Venkatraman, E., Norton, L. & Wigler, M. (2003). Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res*, **13,** 2291-305.

Massion, P.P., Kuo, W.L., Stokoe, D., Olshen, A.B., Treseler, P.A., Chin, K., Chen, C., Polikoff, D., Jain, A.N., Pinkel, D., Albertson, D.G., Jablons, D.M. & Gray, J.W. (2002). Genomic copy number analysis of non-small cell lung cancer using array comparative genomic hybridization: implications of the phosphatidylinositol 3-kinase pathway. *Cancer Res*, **62,** 3636-40.

Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D. & Brown, P.O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*, **23,** 41-6.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. & Albertson, D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet*, **29,** 263-4.

Wang, G., Brennan, C., Rook, M., Wolfe, J.L., Leo, C., Chin, L., Pan, H., Liu, W.H., Price, B. & Makrigiorgos, G.M. (2004). Balanced-PCR amplification allows unbiased identification of genomic copy changes in minute cell and tissue samples. *Nucleic Acids Res*, **32,** e76.

Chapter 5: Resolving the resolution of array CGH

A version of this chapter has been previously published as:

Coe BP, Ylstra B, Carvalho B, Meijer GA, MacAualy C, Lam WL (2007) Resolving the resolution of array CGH *Genomics* **89**(5):647-653 © 2007 Elselvier. doi:10.1016/j.ygeno.2006.12.012

5.1 Introduction

Array comparative genomic hybridization (aCGH) has rapidly supplanted conventional metaphase CGH as the standard protocol for identifying segmental copy number alterations in disease state genomes (Albertson & Pinkel, 2003; de Leeuw et al., 2004). Currently, many genome-wide aCGH platforms are available that span the human genome at specific intervals to facilitate mapping of genetic alterations; however, these platforms are often only described by the number of elements present on the array or the average element spacing, which may not accurately reflect the relative performance of one platform to another, especially given the potential for highly variable element distribution throughout the genome being interrogated (Table I) (Barrett et al., 2004; Braude et al., 2006; de Leeuw et al., 2004; Greshock et al., 2004; Ishkanian et al., 2004; Lips et al., 2005; Snijders et al., 2001; Snijders et al., 2005; van den Ijssel et al., 2005; Weiss et al., 2004; Zhao et al., 2005).

The primary concerns a user may have in selecting an aCGH platform for gene discovery are: "What is the minimal alteration size that can be reliably detected?"; "How precisely will the alteration boundaries be defined?"; and "What are the sample requirements?" In this study we derive new performance definitions through the introduction of "functional resolution", a new metric that incorporates the distribution of array data points, and describe a Java based application "ResCalc" which automates the calculation of performance metrics for any aCGH platform (including any species, and arrays covering only specific chromosomal segments). In addition, we discuss the practical performance characteristics and sample requirements of the major human aCGH platforms.

5.2 Results and Discussion

5.2.1 Alteration Detection is Dependent on Array Element Distribution

It is important to take into account the distribution and length of array elements in order to accurately determine the detection sensitivity to various alteration sizes (**Figure 5.1**). A common practice is to utilize average or median element genomic spacing as a definition of resolution even when the distribution of array elements is non-uniform. However, it is an oversimplification to assume a uniform distribution of array elements and calculate resolution as the number of elements divided by the genome size (Nannya et al., 2005). Another misstatement is to define the resolution of an array by the length of the array elements. For example, it would be erroneous to claim that an array consisting of 100 bp elements offers a resolution of 100 bp – unless every element was tiled contiguously. In this case the concept being relayed is the increased sensitivity of a single array element to a small alteration, not the overall resolution of the array. To cover the entire genome at this resolution would require approximately 30 million contiguous array elements. As this example demonstrates, a simple report of the number of measurements performed and the size of each element is an unreliable method for determining the actual utility of a platform in detecting an alteration of a given size.

We propose that the detection sensitivity of an array is best described by the probability of detecting any alteration of a given size. As discussed by Davies et al. (de Leeuw et al., 2004) the probability of detecting an alteration can be calculated for all possible alterations sizes by determining the percentage of alterations of a given size that would reside between adjacent array elements (**Figure 5.2a**). **Figure 5.2b** demonstrates the result of applying this definition to several key array platforms using the ResCalc algorithm. Tiling arrays using large insert clones such as bacterial artificial chromosomes (BAC) demonstrate very robust performance due to the uniform distribution of elements across the genome and the presence of very few gaps in the genomic coverage (Garnis et al., 2003; Ishkanian et al., 2004) (**Figure 5.2b**). Due to the reduced sensitivity of large insert clones to alterations smaller than 50 kbp, oligonucleotide platforms offer better theoretical performance in detecting small alterations (Barrett et al., 2004; Selzer et al., 2005; Zhao et al., 2005). The Nimblegen 385,000 oligonucleotide array, offers the highest theoretical performance of all platforms detecting 95% of 15 kbp alterations (**Figure 8.2b**). The probabilities of detecting alterations drop drastically for the lower density platforms at small alteration sizes with the Affymetrix 100K platform detecting less than 55% of 27 kbp alterations and 25% of 10 kbp alterations (**Figure 5.2b**).

5.2.2 Evaluation of Practical Performance

Due to various sources of experimental noise, optimal performance is rarely attainable for any array CGH platform. Although high level amplification may be readily detectable with all platforms regardless of noise, the ability to detect single copy gains will be dependent on both the noise of the platform and ratio response of each element. In order to detect a single copy alteration with confidence, the average ratio for a region of copy number gain must differ from the average ratio for a normal portion of the genome by at least 1 standard deviation. When the intrinsic noise of a platform does not allow separation of single copy alterations this can be compensated for through pooling multiple array elements by averaging to reduce the overall noise of the profile (Ylstra et al., 2006).

For cross platform comparison of noise and ratio response, we use the human breast cancer cell line BT474 which has previously been characterized by high resolution FISH mapping. It contains a near tetraploid genome with 104 chromosomes per nuclei (an average of 4.5 copies of each chromosome) (Venter et al., 2005). The ratio separation produced by a single copy gain can be inferred by comparing chromosome bands 8p11-12 and 8q22 which are present at four and six copies respectively, in each BT474 nuclei (2:3 copy number ratio) (**Figure 5.3a**). **Figure 5.3b** shows the copy number profiles for BT474 generated by expert groups using their preferred platforms. By calculating the average $\log_2$ ratio and standard deviation for 8p11-12 and 8q22 and pooling various numbers of array elements, we can determine the performance of each platform. **Figure 5.3c** demonstrates the results of comparing average $\log_2$ signal ratios based on individual and pooled elements for each platform. The SMRT array did not require pooling of elements and thus a single copy change may be reliably detected by a single array element. This criterion is also applicable to the UPenn, Spectral Chip 2600 and HumArray v3.2 platforms which use BACs as array elements. The Agilent 244A platform demonstrated the highest sensitivity of the oligonucleotide platforms with a single element being sufficient to detect a single copy alteration, while the Affymetrix and VUMC platforms required pooling three and two elements respectively to allow separation of single copy differences (**Figure 5.3c**). It is worth noting that the Agilent 244A data represents the result of a dye flip array pair. To determine the effect of this transformation on functional resolution we compared the noise levels in the individual hybridizations to the averaged ratios (data not shown). Both of the hybridizations had noise levels sufficient to detect a single copy change with a single array element (data for 8p11-12 demonstrated between 0.9 and 1.1 times the standard deviation of the same region in the pooled data set). However, despite the minimal improvement to overall experimental noise in this example, it is worth noting that the single spot per loci design of the Agilent platform makes it difficult to determine if a single clone is a true positive or the result of a hybridization artifact without a replicate hybridization.

Data was not available for Nimblegen 385,000 element platform, as a result we utilized the definition provided by Selzer et al. to determine that at least 5 elements must be affected to detect a single copy alteration (Selzer et al., 2005).

Similarly, the Illumina Linkage IV platform performance is expected to be comparable to the Affymetrix platforms due to the use of short oligonucleotides and similar sample labeling technology (Lips et al., 2005).

Figure 5.3d demonstrates the output of ResCalc for several human array platforms, repeating the computation described for theoretical detection sensitivity and incorporating the need to pool various number of array elements to allow detection of single copy alterations. Although high level changes such as homozygous deletions and amplifications below 50 kbp may be detectable with large insert clone arrays such as the SMRT, UPenn, Spectral Chip 2600 and HumArray v3.2 platforms, sensitivity to single copy alterations is greatly reduced in this size range and this sensitivity is reflected by not calculating performance metrics below 50kbp.

Using this metric the Agilent 244A platform demonstrates the highest performance for single copy alterations between 1 and 49kbp (8.7% to 97.5%). The SMRT array demonstrates the highest performance above 50kbp (98.2% to 99.9%). As alteration sizes approach 500 kbp the Agilent 244A, SMRT, Nimblegen 385K and Affymetrix 500K platforms demonstrate very similar performance (**Figure 5.3d**).

5.2.3 Mapping of Breakpoints is Dependent on Local Resolution

In addition to concerns regarding the minimum alteration size that can be reliably detected, the user requires information regarding the precision with which the boundaries of an alteration can be defined. An optimal measurement of edge precision takes into account the fact that the mapping of an alteration boundary is dependent on the distance to the nearest unaffected array element (**Figure 5.4a**). In the case of overlapping array elements (for example overlapping large insert clones), breakpoints can be mapped to within a single array element and thus the intra and inter-element spacing should also be taken into account. Although the probability of detecting a breakpoint within an array element in an oligonucleotide platform or interval based large insert clone array element is lower, the reduced ratio response of a partially gained/lost clone may also be utilized in positioning a breakpoint. Thus, by incorporating the end to end spacing between each array element end, we can determine the proportion of the genome represented by intra/inter-element intervals smaller than a given size. We can then determine the proportion of breakpoints (one potential breakpoint per nucleotide position in the genome) that can be defined with at least the threshold level of precision. This becomes our measurement of edge precision (**Figure 5.4a**). **Figure 5.4b** demonstrates the precision output of ResCalc for several key human aCGH platforms. We observe that increasing the number of array elements drastically changes the slope of the edge precision curve, resulting in a large proportion of edges being detectable at higher levels of precision. The current maximal theoretical performance is demonstrated by the use of the 385,000 oligonucleotide Nimblegen array, followed by the Agilent 244A and Affymetrix 500K arrays, with the relatively uniformly distributed clone ends on the SMRT array demonstrating the fourth best precision.

5.2.4 Defining Functional Resolution

It is apparent that increasing the number of array elements does not result in a linear increase in performance (**Figures 5.2 to 5.4**). Factors including element size and uniformity of element distribution are key contributors to the theoretical performance of an array platform. In defining the functional resolution of an array platform, we propose integrating these metrics into the description of each technology. If we are analyzing samples in the context of mapping genetic alterations, it is prudent to assume that resolution may be best defined by the level of performance (sensitivity and precision) that is applicable to describing 95% of genomic alterations. Thus the alteration size at which only 1 in 20 single copy genomic alterations escape detection will define the practical sensitivity of a platform, while the alteration size at which 1 in 20 high level copy number alterations escape detection defines the theoretical sensitivity. Incorporating these measurements as well as the precision demonstrated for 19 in 20 breakpoints will define the functional resolution of the platform (**Figures 5.2 to 5.4**). Table I lists the functional resolutions (as determined by the ResCalc application) of all platforms discussed in this study. It is noteworthy that no one platform demonstrates the highest performance for all metrics at this time. The Nimblegen 385,000 element platform demonstrates the current maximum precision of 24 kbp and theoretical resolution of 15 kbp however single copy alteration sensitivity is limited to 54 kbp alterations. Similarly the Agilent platform demonstrates the highest single copy number alteration sensitivity of 36 kbp, however precision is limited to 56 kbp (**Table 5.1**). It is obvious that oligonucleotide platforms demonstrate improved sensitivity to single copy alterations as they increase their density; however, this is currently only practical for specific loci as whole human genome arrays with very high densities currently span more that two chips (Selzer et al., 2005).

5.2.5 Sample Considerations

A key consideration in selection of an aCGH platform is whether it is suitable for analyzing the type of samples at hand. Formalin fixed paraffin embedded (FFPE) samples are currently restricted to platforms which do not require probe complexity reduction steps (such as the Illumina and Affymetrix platforms). Currently low yield FFPE samples are most applicable to large insert clone platforms such as the SMRT array, while oligonucleotide platforms which lack genome-complexity reduction steps in probe generation may be capable of analyzing these samples as well depending on attainable DNA yield (Garnis et al., 2003; Ylstra et al., 2006). Sample DNA amplification can drastically reduce the amount of primary material required, however noise and bias is introduced by non-linear amplification of sequences, limiting utility in the analysis of limited yield clinical specimens. Currently several platforms are capable of analyzing un-amplified samples with limited yield (less than 1 μg) including all large insert clone platforms (The Spectral Chip 2600 uses 1ug of DNA if the dye flip experiment is excluded), and the VUMC and Affymetrix oligonucleotide platforms (including the Agilent 244A platform if the dye flip experiment is excluded).

5.2.6 Selecting a Platform

It is important to note that attaining the highest possible resolution is not the only factor in determining the platform best suited to a particular analysis. High resolution arrays demonstrate a significant cost increase over low resolution platforms which may be more appropriate depending on the hypothesis of the study at hand (Ylstra et al., 2006). Another important consideration is the utility offered by combined LOH/CGH platforms which can increase our understanding of cryptic copy number alterations (an important consideration is the percentage of heterozygous calls obtained in an average reference sample, which will determine the probability of generating a usable LOH call in a specific alteration) (Lips et al., 2005; Zhao et al., 2004; Zhao et al., 2005). Taking these concerns into account as well as the theoretical and practical sensitivity, breakpoint precision, and sample requirements (both quality and DNA yield) will help determine the platform best suited to approach each biological hypothesis.

5.3 Conclusions

In this cautionary note, we highlight that the extrapolation of local resolution could misrepresent "functional-resolution" of an aCGH platform across the genome. Our proposed metrics incorporate the distribution of array elements allowing a more objective comparison of array platforms. We envision that standard calculations of performance such as "functional resolution" as defined by ResCalc will prove invaluable in the future description/comparison of aCGH platforms.

5.4 Materials and Methods

5.4.1 Array Platform Data Sources

Array Mapping files were obtained from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/projects/geo/) for the Agilent (GEO Accession: GPL4091), VUMC (GEO Accession: GPL2827), and Spectral Chip 2600 (GEO Accession: GPL3780) platforms. Mappings were acquired from manufacturer web sites for the Affymetrix 100K/500K (www.affymetrix.com), Illumina IV (www.illumina.com), HumArray 3.2 (cancer.ucsf.edu/array/services.php#humanBAC), SMRT (www.bccrc.ca/cg/ArrayCGH_Group.html), and Upenn (www.genomics.upenn.edu/people/faculty/weberb/CGH/html/downloads.htm) platforms. The Nimblegen 385K mapping was acquired from an internal hybridization results file.

For oligonucleotide platforms often only one mapping position is provided for an oligo, in this case the position of the second end is derived by adding 1 oligo size to the provided bp position. In the case of the Nimblegen platform which uses isothermic oligos of varying sizes we based our calculations on an average 60bp oligo length applied to each element. Similarly for the HumArray data, several BAC clones only had one base pair position associated and the second end was assumed to be 150kbp distal.

BT474 data files were acquired from the following sources: SMRT (GEO Accession: GSM69198), VUMC (GEO Accession: GSM73557), Affymetrix (http://research.dfci.harvard.edu/meyersonlab/snp/snp.htm). The Agilent 244A data was generated from an averaged dye flip experiment performed by Agilent Technologies using their standard protocols (www.opengenomics.com) and has been submitted to GEO (GEO Accession GSE6415). Array data is also available from the System for Integrative Genomic Microarray Analysis (SIGMA) interactive web database (http://sigma.bccrc.ca), which was used to generate the image in Figure **5.3b** (Chari et al., 2006).

5.4.2 Implementation of ResCalc

ResCalc is implemented as a command line executable java application. The application requires JRE 5.0 Update 7.0 or better. Briefly the application requires a tab delimited text file describing each element present on the arrays chromosomal position and the base pair co-ordinates of the start and end of each array element. Additionally a file is required which annotates the location of the centromere on each chromosome described in the platform file. There is no restriction on the number or names of chromosomes in both the platform and centromere description files, thus the algorithm can be run on arrays covering any portion of the any genome. The executable, centromere description files for several human genome sequence builds, and documentation are available at http://sigma.bccrc.ca/ResCalc.html

5.4.3 Calculation of Optimal Detection Sensitivity

The optimal detection sensitivity for a platform is calculated as follows.

Firstly the set of all inter-element gaps are defined as the set of all positive differences between the end base pair position of one array element and the start base pair position of the next array element, excluding differences which include the centromere of the chromosome being examined. The number of potential alterations of a given size is defined such that one alteration may start at each unique base pair position of the genome being interrogated.

The number of alterations of a given size that will be contained completely within a gap represent the alterations that will escape detection with the current array platform. This is calculated by subtracting the current alteration size from each inter-element gap and summing the positive residuals.

The total size of the interrogated genome is then calculated as the sum of all center to center element intervals and is used to define the percentage of possible alterations which will be detected under optimal conditions as 1 – (number of alterations missed / genome size).

5.4.4 Calculation of Practical Detection Sensitivity

The first step in calculating the practical sensitivity for an array platform is to determine the number of elements that must be averaged to reduce the variation in the data enough to allow detection of a single copy alteration (defined as separating a region of normal copy number from a region of single copy gain by one standard deviation).

The calculation is then performed similarly to the calculation for optimal detection sensitivity with the following modifications. The number of clones that must be averaged to detect a single copy change defines the penalty. The inter-element gaps are defined as the set of positive differences between the end base pair position of one element and the start base pair position of the element exactly penalty -1 elements away. If this spacing is greater than the alteration size being interrogated the number of alterations missed is defined as follows:

If the base pair start position of the next element outside of the penalty window – alteration size is less than the base pair position of the end of the next clone the number of alterations missed is defined as the interval between the end of the current clone and the position defined above. Else the number of alterations missed is defined as the interval between the current elements end base pair position and the end base pair position of the next array element.

5.4.5 Calculation of Breakpoint Precision

The percentage of breakpoints which may be mapped with a precision of at least n base pairs is defined as the cumulative distribution of inter element end base pair intervals (excluding intervals which span a centromere) > the currently interrogated level of precision.

Figure 5.1

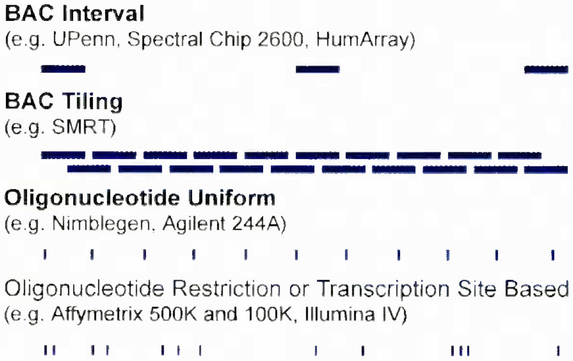

BAC Interval
(e.g. UPenn, Spectral Chip 2600, HumArray)
BAC Tiling
(e.g. SMRT)
Oligonucleotide Uniform
(e.g. Nimblegen, Agilent 244A)
Oligonucleotide Restriction or Transcription Site Based
(e.g. Affymetrix 500K and 100K, Illumina IV)

Figure 5.1. Schematic Overview of Array CGH Platform Designs. BAC arrays are typically produced with uniform genomic distribution or with overlapping/tiling clones. Oligonucleotide arrays may also be produced with uniform genomic distribution, however for some platforms the need for genome reduction labeling steps or design biased towards transcriptional sites leads to non-uniform element distribution causing local resolution to vary drastically.

Figure 5.2

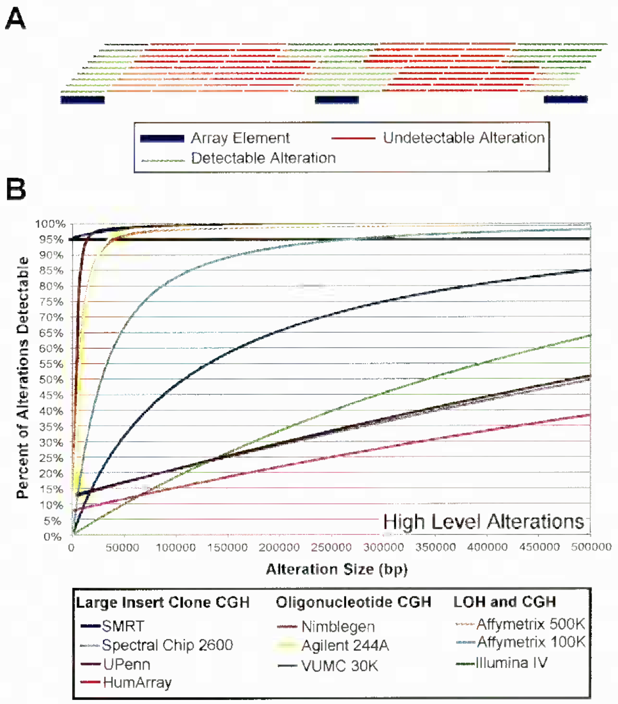


Figure 5.2. Theoretical Detection Sensitivity. (**a**) Detection sensitivity for each array platform was calculated based on the percent of possible alterations of a given size that interact with at least one array element (blue bars). To determine the proportion of alterations of size n bp detectable by an array platform we first defined the set of all possible alterations (possible alterations are represented by red and green bars) of size n bp for all genomic regions covered by the array (excluding centromeres and acrocentric regions). We then calculated the percentage of alterations not detectable as those that are completely contained within each coverage gap. (**b**) Detection sensitivities for each platform are plotted for alteration sizes from 1 kbp to 500 kbp, the alteration size at which platform exceeds a 95% detection rate defines the optimal sensitivity of that platform.

Figure 5.3. Single Copy Detection Sensitivity. (**a**) The BT474 cell line contains an average of 4.5 copies of each chromosome. Previous FISH studies characterized chromosome 8 into segments with 4, 5 and 6 copies. By comparing the ratios observed for 6 and 4 copies we can simulate the performance of a 3:2 copy number ratio. (**b**) Comparison of copy number profiles of chromosome 8 across 3 platforms. (**c**) Determination of the number of elements which must be pooled to allow detection of single copy alterations. BT474 profiles were used to determine the number of elements that must be pooled to separate the average ratios for 4 and 6 copies by at least one standard deviation (indicated by *) for the SMRT, Agilent, VUMC and Affymetrix (The noise of the Affymetrix Mapping 10K is projected to be equivalent to the 100K and 500K set due to the use of genomic reduction steps and identical oligonucleotide design strategy) platforms. (**d**) Single copy alteration detection sensitivities for each platform are plotted for alteration sizes from 1 kbp to 500 kbp, the alteration size at which platform exceeds a 95% detection rate defines the optimal sensitivity of that platform. Each platform was penalized based on the number of elements that must be pooled according to the calculation described in part c. Data was not available for the Nimblegen platform, data is adjusted to incorporate pooling of 5 elements as described in Selzer et al (Selzer et al., 2005). Similarly, data was not available for the Illumina platform, due to the similar probe length and labeling technology to the Affymetrix platform a 3 clone requirement was assumed.

Figure 5.3

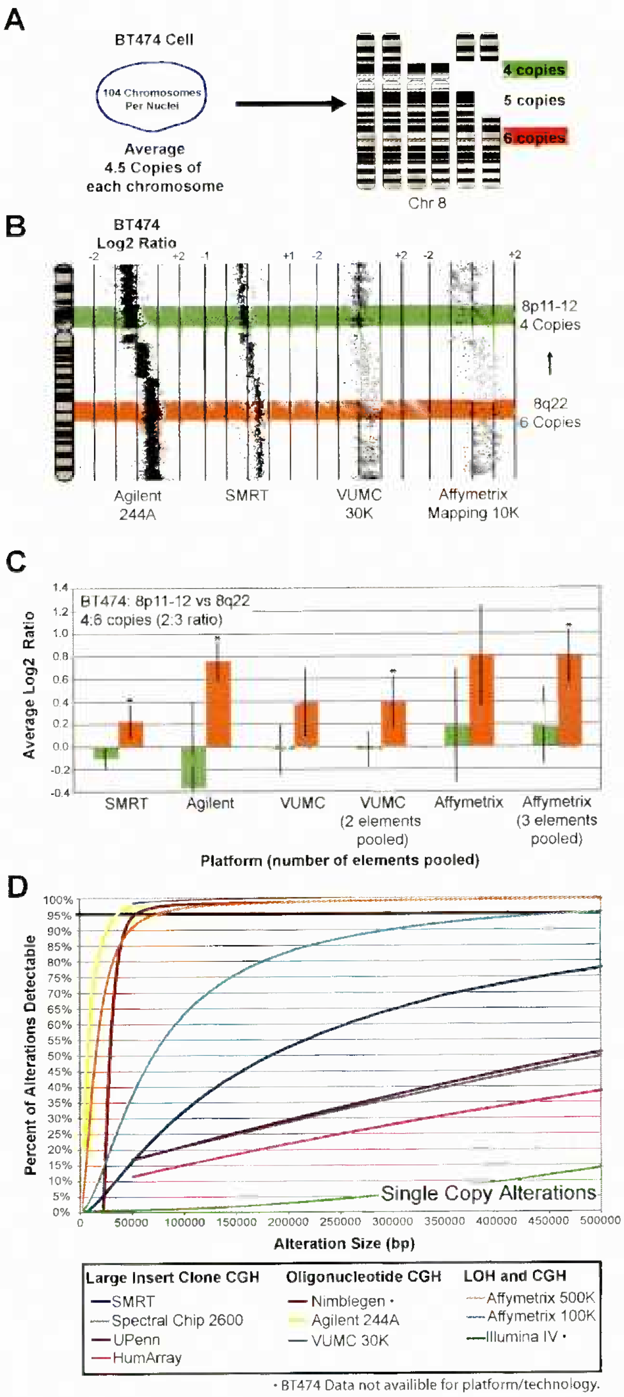

A
BT474 Cell
104 Chromosomes Per Nuclei
Average
4.5 Copies of
each chromosome
4 copies
5 copies
6 copies
Chr 8
B
BT474
Log2 Ratio
-2
+2
-1
+1
-2
+2
-2
+2
8p11-12
4 Copies
8q22
6 Copies
Agilent
244A
SMRT
VUMC
30K
Affymetrix
Mapping 10K
C
BT474: 8p11-12 vs 8q22
4:6 copies (2:3 ratio)
Average Log2 Ratio
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
SMRT
Agilent
VUMC
VUMC
(2 elements
pooled)
Affymetrix
Affymetrix
(3 elements
pooled)
Platform (number of elements pooled)
D
Percent of Alterations Detectable
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
0
50000
100000
150000
200000
250000
300000
350000
400000
450000
500000
Single Copy Alterations
Alteration Size (bp)
Large Insert Clone CGH
SMRT
Spectral Chip 2600
UPenn
HumArray
Oligonucleotide CGH
Nimblegen •
Agilent 244A
VUMC 30K
LOH and CGH
Affymetrix 500K
Affymetrix 100K
Illumina IV •
• BT474 Data not availible for platform/technology.

Figure 5.4

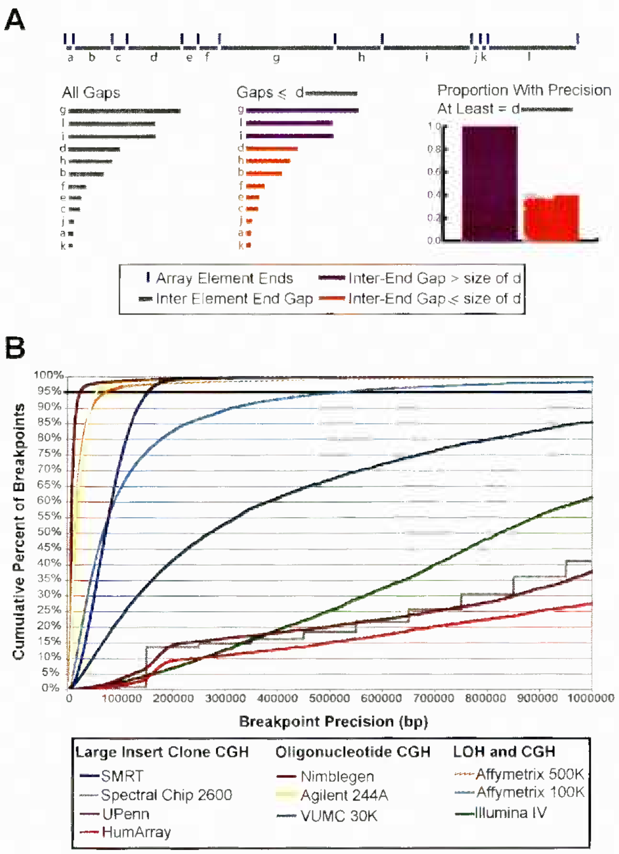


Figure 5.4. Breakpoint Precision. (**a**) The precision with which a breakpoint can be defined is derived from the genomic distance between each element end (as alteration boundaries can be defined to reside within an array element). The set of all inter-element end gaps in the genome can then be determined (a to l) and sorted by increasing size. The percentage of the genome covered by inter-element end gaps less than n bp in width (example size of gap "d") defines the proportion of breakpoints which demonstrate a precision of at least n bp (assuming 1 possible breakpoint per base pair). (**b**) Breakpoint precisions for each platform are plotted for alteration sizes ranging from 1 kbp to 1 Mbp, the precision level at which platform exceeds 95% defines the optimal breakpoint precision of that platform.

Table 5.1. Comparison of array CGH technologies.

Platform	Technology	Functional Resolution			Sample Labeling	Sample Requirements	Notes
		Theoretical Sensitivity	Single Copy Sensitivity	Breakpoint Precision			
Nimblegen 385K	Oligonucleotide (45-85 nt)	15 kbp	54 kbp*	24 kbp	Whole Genome	1 – 3 µg	*Single copy sensitivity is estimated based on analysis parameters described in Selzer et al.
Agilent 244A	Oligonucleotide (60 nt)	36 kbp	36kbp	56 kbp	Whole Genome	0.5 µg (1 µg with dye flip)	DNA amplification reduces DNA requirements to 0.1 µg of DNA per slide (0.2 µg with dye flip). (not tested in this manuscript)
Affymetrix GeneChip® Human Mapping 500K Set	Oligonucleotide (25 nt)	41 kbp	75 kbp	74 kbp	PCR reduction	0.5 µg	Platform is also used for LOH analysis.
Sub-Megabase Resolution Tiling-set (SMRT)	Large Insert Clone (BAC)	50 kbp	50 kbp	152 kbp	Whole Genome	0.1 µg	High level amplifications below 50kbp may be detectible, this is not indicated.
Affymetrix GeneChip® Human Mapping 100K Set	Oligonucleotide (25 nt)	271 kbp	476 kbp	528 kbp	PCR reduction	0.5 µg	Platform is also used for LOH analysis.
VUMC MACF Human 30K	Oligonucletide (60nt)	1.05 Mbp	1.32 Mbp	1.94 Mbp	Whole Genome	0.3 µg	Invitrogen has recently released a 50K oligonucleotide library suitable for array CGH including intragenic oligonucleotides.
Illumina Linkage IV	Oligonucleotide (40nt)	1.35 Mbp	2.66 Mbp	2.06 Mbp	PCR reduction	1 µg	Illumina has recently released a 100K (Infinium) assay. Both platforms are also used for LOH analysis.
UPenn	Large Insert Clone (BAC)	1.99 Mbp	1.99 Mbp	3.15 Mbp	Whole Genome	1 µg	Sample Requirement is likely 100ng due to use of BAC clones.
Spectral Chip 2600	Large Insert Clones (BAC)	2.65 Mbp	2.65 Mbp	4.55 Mbp	Whole Genome	1 ug (2 ug with dye flip)	Sample Requirement is likely 100ng due to use of BAC clones.
HumArray 3.2	Large Insert Clone (BAC)	5.07 Mbp	5.07 Mbp	8.75 Mbp	Whole Genome	0.6 µg	Sample Requirement is likely 100ng due to use of BAC clones.

5.5 References

Albertson, D.G. & Pinkel, D. (2003). Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet*, **12 Spec No 2,** R145-52.

Barrett, M.T., Scheffer, A., Ben-Dor, A., Sampas, N., Lipson, D., Kincaid, R., Tsang, P., Curry, B., Baird, K., Meltzer, P.S., Yakhini, Z., Bruhn, L. & Laderman, S. (2004). Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc Natl Acad Sci U S A*, **101,** 17765-70.

Braude, I., Vukovic, B., Prasad, M., Marrano, P., Turley, S., Barber, D., Zielenska, M. & Squire, J.A. (2006). Large scale copy number variation (CNV) at 14q12 is associated with the presence of genomic abnormalities in neoplasia. *BMC Genomics*, **7,** 138.

Chari, R., Lockwood, W.W., Coe, B.P., Chu, A., Macey, D., Thomson, A., Davies, J.J., Macaulay, C. & Lam, W.L. (2006). SIGMA: A System for Integrative Genomic Microarray Analysis of Cancer Genomes. *BMC Genomics*, Submitted.

de Leeuw, R.J., Davies, J.J., Rosenwald, A., Bebb, G., Gascoyne, R.D., Dyer, M.J., Staudt, L.M., Martinez-Climent, J.A. & Lam, W.L. (2004). Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum Mol Genet*, **13,** 1827-37.

Garnis, C., Baldwin, C., Zhang, L., Rosin, M.P. & Lam, W.L. (2003). Use of complete coverage array comparative genomic hybridization to define copy number alterations on chromosome 3p in oral squamous cell carcinomas. *Cancer Res*, **63,** 8582-5.

Greshock, J., Naylor, T.L., Margolin, A., Diskin, S., Cleaver, S.H., Futreal, P.A., deJong, P.J., Zhao, S., Liebman, M. & Weber, B.L. (2004). 1-Mb resolution array-based comparative genomic hybridization using a BAC clone set optimized for cancer gene analysis. *Genome Res*, **14,** 179-87.

Ishkanian, A.S., Malloff, C.A., Watson, S.K., DeLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M.A., Ling, V., MacAulay, C. & Lam, W.L. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, **36,** 299-303.

Lips, E.H., Dierssen, J.W., van Eijk, R., Oosting, J., Eilers, P.H., Tollenaar, R.A., de Graaf, E.J., van't Slot, R., Wijmenga, C., Morreau, H. & van Wezel, T. (2005). Reliable high-throughput genotyping and loss-of-heterozygosity detection in formalin-fixed, paraffin-embedded tumors using single nucleotide polymorphism arrays. *Cancer Res*, **65,** 10188-91.

Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D.K., Kennedy, G.C. & Ogawa, S. (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*, **65,** 6071-9.

Selzer, R.R., Richmond, T.A., Pofahl, N.J., Green, R.D., Eis, P.S., Nair, P., Brothman, A.R. & Stallings, R.L. (2005). Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer*, **44,** 305-19.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. & Albertson, D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet*, **29,** 263-4.

Snijders, A.M., Schmidt, B.L., Fridlyand, J., Dekker, N., Pinkel, D., Jordan, R.C. & Albertson, D.G. (2005). Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene*, **24,** 4232-42.

van den Ijssel, P., Tijssen, M., Chin, S.F., Eijk, P., Carvalho, B., Hopmans, E., Holstege, H., Bangarusamy, D.K., Jonkers, J., Meijer, G.A., Caldas, C. & Ylstra, B. (2005). Human and mouse oligonucleotide-based array CGH. *Nucleic Acids Res*, **33,** e192.

Venter, D.J., Ramus, S.J., Hammet, F.M., de Silva, M., Hutchins, A.M., Petrovic, V., Price, G. & Armes, J.E. (2005). Complex CGH alterations on chromosome arm 8p at candidate tumor suppressor gene loci in breast cancer cell lines. *Cancer Genet Cytogenet*, **160,** 134-40.

Weiss, M.M., Kuipers, E.J., Postma, C., Snijders, A.M., Pinkel, D., Meuwissen, S.G., Albertson, D. & Meijer, G.A. (2004). Genomic alterations in primary gastric adenocarcinomas correlate with clinicopathological characteristics and survival. *Cell Oncol*, **26,** 307-17.

Ylstra, B., van de IJssel, P., Carvalho, B., R.H., B. & Meijer, G.A. (2006). BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH) *Nucleic Acids Res.*, **32,** 445-450.

Zhao, X., Li, C., Paez, J.G., Chin, K., Janne, P.A., Chen, T.H., Girard, L., Minna, J., Christiani, D., Leo, C., Gray, J.W., Sellers, W.R. & Meyerson, M. (2004). An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res*, **64,** 3060-71.

Zhao, X., Weir, B.A., LaFramboise, T., Lin, M., Beroukhim, R., Garraway, L., Beheshti, J., Lee, J.C., Naoki, K., Richards, W.G., Sugarbaker, D., Chen, F., Rubin, M.A., Janne, P.A., Girard, L., Minna, J., Christiani, D., Li, C., Sellers, W.R. & Meyerson, M. (2005). Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Res*, **65,** 5561-70.

Chapter 6: High resolution Chromosome 5p Array CGH Analysis of Small Cell Lung Carcinoma Cell Lines

A version of this chapter has been previously published as:

Coe BP, Henderson LJ, Garnis C, Tsao MS, Gazdar AF, Minna J, Lam S, Macaulay C, Lam WL. (2005) High-resolution chromosome arm 5p array CGH analysis of small cell lung carcinoma cell lines. *Genes Chromosomes Cancer* **94**(3):308-313 © 2005 Wiley-Liss, Inc. doi:10.1002/gcc.20137

6.1 Introduction

Small cell lung cancer (SCLC) accounts for 20% of the yearly cases of lung cancer in the United States. Survival rates for this disease have seen little improvement over the last 20 years, with median survival times of only 7-13 months. This may be due in part to the high frequency of relapse with resistant micrometastatic disease after initial chemotherapy. Identification of prognostic and therapeutic molecular targets may contribute to improve survival rates of patients with SCLC (Bremnes et al., 2003; Krug and Miller, 2003).

Genomic amplifications of chromosome arm 5p have been observed frequently in small cell lung cancer. However, conventional comparative genomic hybridization (CGH) studies have been unable to define regions more precisely than a few chromosome bands in SCLC (Levin et al., 1994; Reid et al., 1994; Ashman et al., 2002; Balsara and Testa, 2002; Reid et al 1994; Yokoi et al., 2003). The high frequency of aberrations on this arm in SCLC, combined with the fact that both amplification and deletion events are present on 5p, suggests that there may be genes involved in SCLC transformation other than the well-characterized *TERT* (telomerase reverse transcriptase) (Sattler and Salgia, 2003) and *SKP2* (S-phase-associated kinase protein 2) (Yokoi et al., 2002, 2003) genes.

Recent technological advances in array CGH facilitate the high-resolution mapping of regional copy number aberrations. Arrays that span the genome at 1-2 Mbp intervals (Snijders et al., 2001) as well as more specific arrays consisting of higher resolution coverage of specific loci have been widely used for detecting genetic alterations in human disease (Bruder et al., 2001; Albertson and Pinkel, 2003). However, the analysis of entire chromosome arms at a tiling resolution has currently been applied only to chromosome 22, with a recent report of a CGH array containing overlapping clones at approximately 0.075 Mbp resolution (Buckley et al., 2002). Here, we describe the construction of a 0.1 Mbp resolution array spanning chromosome arm 5p and the application of this sub-megabase resolution tiling set (SMRT) CGH array in the analysis of copy number aberrations on chromosome arm 5p in SCLC cell lines.

6.2 Results and Discussion

The 5p CGH array consists of 491 fingerprint-verified, LMPCR-amplified BAC DNA samples spanning the 50 Mbp p arm of chromosome 5 from 5p11 to 5p15.33. The average resolution is 10 clones per megabase, with an average clone size of 150 kb. This represents a 10 fold increase in resolution over currently available array coverage of the 5p region.

We utilized the physical map of the human genome (The International human genome mapping consortium, 2001) and the UCSC genome browser (Kent et al., 2002) to facilitate the selection of this subset of clones from the whole genome bacterial artificial chromosome (BAC) re-array set (http://bacpac.chori.org/pHumanMinSet.htm) which represents minimal overlapping coverage of chromosome arm 5p. The conversion of BAC DNA samples to linker-mediated PCR (LMPCR) products at a sufficient concentration for spotting was performed as previously described (Garnis et al., 2004).

Ninety-six random loci scattered throughout the human genome were included on the array as internal controls. Spots representing X and Y chromosome loci allow the verification of the sensitivity of the arrays to single-copy-number alterations. Additionally, LMPCR-amplified human genomic DNA spots which are used in the normalization of our signal intensities due to the even nature of their hybridization to each probe were added to the array.

The 5p SMRT CGH array was tested for its ability to detect single-copy-number changes by hybridization with normal male and female DNAs. This resulted in the detection of the single copy change at the X chromosome loci as a signal ratio increase (female/male) on the appropriate spots, whereas autosomal loci showed an equivalent copy number (represented by a log2 signal ratio of 0). Additionally, normal versus normal DNA hybridizations were used for verifying successful amplification of clones. Clones that deviated by greater than 3 standard deviations (+/- 0.13) from 0 on a Log2 signal ratio plot of the 5p tiling set were discarded (Veltman et al., 2003). These thresholds were also used to identify altered clones in the experiments.

Furthermore, the array CGH data were verified by comparison against conventional CGH data. For example, the SCLC cell line NCI-H526 revealed a centromeric region of copy number increase which is consistent with the description present at http://amba.charite.de/~ksch/cghdatabase/index.htm. Our array CGH profile not only matched the conventional CGH data, but it also defined the breakpoint to a region of 0.2 Mbp between CTD-2335O24 and RP11-420H15 in a single experiment (**Figure 6.1**).

In this study, we profiled 15 SCLC cell lines and identified multiple discontinuous regions of amplification and deletion on 5p. This sample set consists of 12 classical (NCI-H187, NCI-H378, NCI-H889, NCI-H1184, NCI-H1607, NCI-H1672, NCI-H1963, NCI-H2141, NCI-H2171, NCI-H2195, NCI-H2227, HCC33) and 3 variant (NCI-H82, NCI-H289, NCI-H526) phenotype cell lines.

Eight of the 15 samples showed whole arm amplification of 5p with several displaying additional small regional aberrations. Others displayed more localized regions of alteration, with the exception of NCI-H82, which appeared to be unaltered (**Figure 6.2e**).

The variant cell line NCI-H526 clearly demonstrates a 12.7 Mbp centromeric amplification (**Figure 6.1a**) spanning from 5p12 to a precise breakpoint in a region of approximately 0.2 Mbp between clones CTD-2335O24 and RP11-420H15 on 5p13.3 (**Figure 6.1e**). Fluorescence in-situ hybridization (FISH) was used to verify this breakpoint (**Figure 6.1c**). This segmental alteration represents the smallest centromeric amplification observed in our sample set (**Figure 6.2e**, blue bar) and contains 62 genes annotated from the RefSeq database, including the *SKP2* oncogene (Yokoi et al., 2002).

Sample NCI-H187 revealed the smallest amplification on the telomeric end of 5p in our sample set (**Figure 6.2a, e**). This amplification spans 5.8 Mbp from RP11-15J3 to RP11-753F19 and contains the well characterized *TERT* gene (Sattler and Salgia 2003) as well as 15 others.

In addition to regional gains, microamplifications involving 2-3 clones were detected. The detection of such changes is attributed to the high resolution of the 5p array. It is possible that minute moderate level amplifications such as the ones we observe are common in tumors but have escaped detection by conventional cytogenetic analysis. Also, it has been suggested that low copy gains may be as important as or possibly more important than high level gains in gene dysregulation, as they affect more genes (Hyman et al., 2002). Several of these minute amplifications were recurrent across multiple lines and are discussed below. For example, sample NCI-H2171 clearly demonstrates amplification of the adjacent clones RP11-104O20 and RP11-99I1 (**Figure 6.2b**, E red arrow) as well as RP11-816H3 and RP11-597N1 (**Figure 6.2b, e** pink arrow) on 5p15.2. This pair of microamplifications is recurrent in five additional cell lines. These regions each contain a single gene, *TRIO* (triple functional domain, *PTPRF* interacting, GenBank Acc #: NM_007118; **Figure 6.2** red arrow), and *ANKH* (ankylosis, progressive homolog, GenBank Acc #: NM_054027; **Figure 6.2** pink arrow) respectively. Neither gene has been implicated in lung cancer; therefore, their biological relevance to SCLC remains to be explored. *TRIO* however has recently been shown to be a putative oncogene in bladder cancer (Zheng et al, 2004), FISH was used to verify the amplification of this gene in the SCLC line H2171 and showed a low level gain of the *TRIO* locus (RP11- 20B15) when compared to a neighboring 5p clone (RP11- 466j3), which demonstrated a lower level of copy number gain (data not shown).

Whereas amplification of chromosome arm 5p is commonly observed in lung cancer, deletions have also been detected by conventional CGH (Balsara and Testa, 2002; Yokoi et al., 2003). However, these deleted regions have not been defined precisely. Using the 5p SMRT CGH array, we have observed several distinct deletions. For example, the SCLC line NCI-H1963 demonstrates two distinct regions of deletion: a ~0.6 Mbp deletion at 5p13.1 and a 6.6 Mbp deletion from RP11-571G9 at 5p1.2 to RP11-711H13 on 5p15.32. However, these discrete deletions are found only in this cell line. Nonetheless, the detection of a submegabase deletion on 5p13.1 further demonstrates the resolution of our array.

More interestingly, SMRT array CGH profiling has allowed us to identify many microdeletions on chromosome arm 5p, several of which recurred in multiple samples. Microdeletions are an interesting discovery because they have rarely been discussed. Although small deletions have been seen in developmental disorders (Stewart et al., 2004; Kriek et al., 2004), deletions of this scale have simply been undetectable without very tedious profiling before high resolution array CGH. Remarkably, RP11-447B16 on 5p13.2 is frequently (13/15 samples) seen at a significantly lower signal ratio than are its neighboring clones (**Figure 6.2a, b, c, e**, blue arrow). The identification of a recurrent microdeletion is further supported by the observation of a decreased signal ratio on the clone's overlapping neighbors (BAC clones mapped adjacent to each other are physically separated on the array). These neighboring clones are likely partially deleted and therefore display a signal ratio decrease proportional to the amount of probe absent. Annotation of this region through alignment with the draft sequence of the human genome by use of the UCSC Genome Browser shows that this deletion may disrupt a hypothetical gene, *FLJ10233*.

Several other putative microdeletions were observed across multiple cell lines. A microdeletion centered at clone RP11-107C3 was observed in 9/15 SCLC cell lines (**Figure 6.2e**, green arrow). Finally a minute deletion was observed at clone RP11-29E11 in 10 cases. This deletion spans a ~0.5 Mbp region to RP11-230I5 in H289 (**Figure 6.2c, e**, purple arrow). In four samples, RP11-230I5 appeared as a distinct single-clone deletion separated from the microdeletion at RP11-29E11 by ~0.3 Mbp. None of these deletions contain known genes and as such their significance in SCLC will require future study for verification.

In conclusion, we have demonstrated that 5p SMRT array CGH can identify known aberrations at the *TERT* and *SKP2* loci as well as fine map novel copy number changes on this arm, such as the microamplifications of *TRIO* and *ANKH* (**Table 6.1**). This high resolution approach has allowed the identification of novel microdeletions, which have escaped detection by conventional screening methods such as microsatellite analysis and CGH and may play a role in SCLC, verification of these alterations in clinical samples will further support their role in SCLC tumorogenesis, and future application of this array to other disease types will greatly facilitate cancer gene discovery on this arm.

Figure 6.1

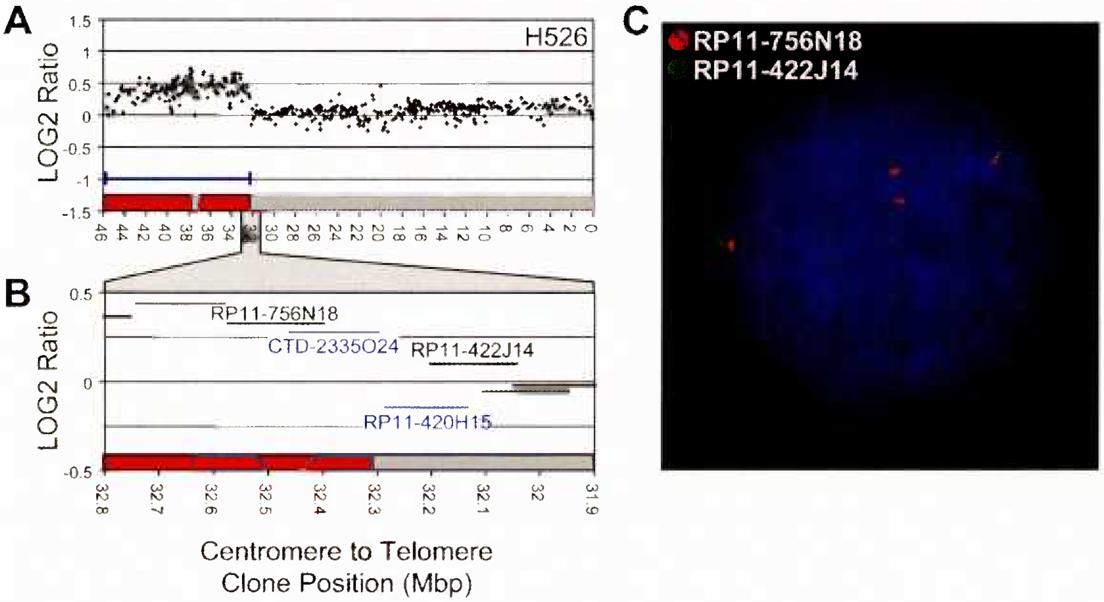


Figure 6.1. 5p Array CGH Profile of H526. (**a**) Array CGH Profile of the NCI-H526 cell line. Array CGH profiling was performed by co-hybridizing a reference and sample DNA labeled with Cyanine 5 and Cyanine 3 dNTPs, respectively. The labeling, hybridization and imaging protocols were described previously (Garnis et al., 2004). Data is displayed as a plot of the log2 H526 versus normal signal ratio for each clone on the array versus genomic position in mega base pairs. The red bars on the bottom of the plot indicate the amplified regions. The blue double ended line highlights the 12.7Mb SKP2 amplification region discussed in the text and detailed in Fig 2E. (**b**) Detail of the NCI-H526 5p13.3 Breakpoint. Clones are displayed as horizontal lines at their observed log2 signal ratio representing their relative positions and sizes. Clones drawn in blue represent those nearest the breakpoint. (**c**) FISH validation of the 5p NCI-H526 breakpoint. Two clones adjacent to the breakpoint were chosen for FISH validation, RP11-756N18 and RP11-422J14 were labeled with Spectrum Red and Green respectively through a random priming reaction and FISH was performed as described in (Henderson et al., 2004).

Figure 6.2. Summary of the 15 SCLC Lines Profiled. **(a-d)**. Profiles of SCLC cell lines NCI-H187, NCI-H2171, NCI-H289 and NCI-H1963. Profiles are displayed as the log2 signal ratio for each clone on the array versus genomic position in mega base pairs. The color coding on the bottom of each plots indicates the type of aberration defined; Green represents a deletion, Gray represents retention of normal copy number, dark and light red indicate low and high level amplifications respectively. Colored arrows and double ended bars are described below. **(e)** Copy number matrix displaying the SMRT aCGH results for each SCLC line profiled, and a representative normal male versus normal female control hybridization (labelled MvF). Intensities of red and green coloration indicate an increased or decreased log2 signal ratio for each clone respectively. Gray coloration indicates clones discarded due to high standard deviations (>0.075). Each column represents a separate aCGH profile, with the sample name indicated. Green Blue and Purple Arrows represent regions of microdeletion. The double ended bars on the right of the copy number matrix represent the *SKP2* (blue line) and *TERT* (black line) minimal regions of alteration.

Figure 6.2

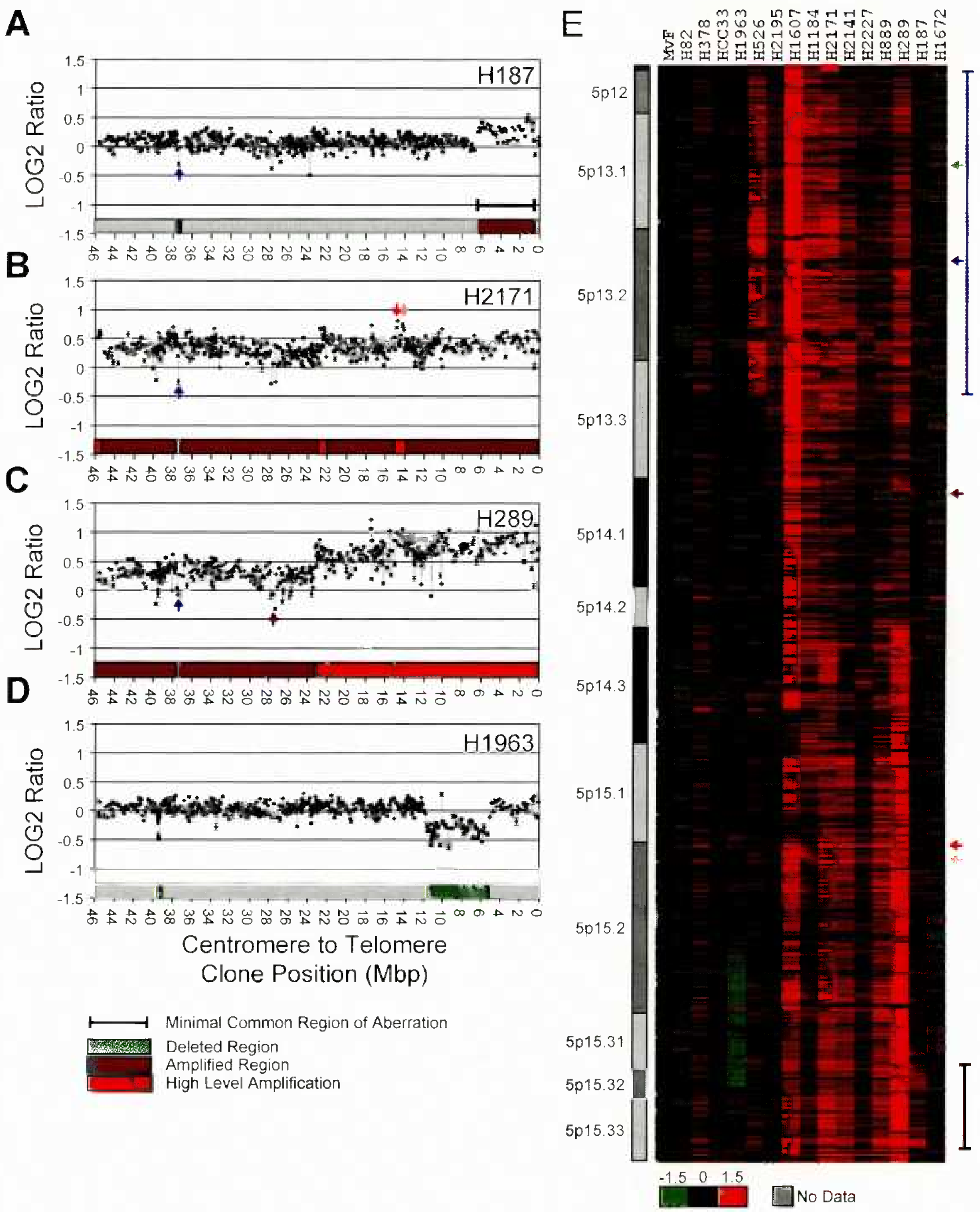

A
B
C
D
E
LOG2 Ratio
H187
H2171
H289
H1963
1.5
1
0.5
0
-0.5
-1
-1.5
46 44 42 40 38 36 34 32 30 28 26 24 22 20 18 16 14 12 10 8 6 4 2 0
Centromere to Telomere
Clone Position (Mbp)
Minimal Common Region of Aberration
Deleted Region
Amplified Region
High Level Amplification
MvF
H82
H378
HCC33
H1963
H526
H2195
H1607
H1184
H2171
H2141
H2227
H889
H289
H187
H1672
5p12
5p13.1
5p13.2
5p13.3
5p14.1
5p14.2
5p14.3
5p15.1
5p15.2
5p15.31
5p15.32
5p15.33
-1.5 0 1.5
No Data

Table 6.1. Summary of novel alterations

Alteration locus	Genes in region	Type of alteration	Frequency
RP11-104O20 - RP11-99I1	*TRIO*	Copy number gain	6/15 cell lines.
RP11-816H3 - RP11-597N1	*ANKH*	Copy number gain	6/15 cell lines.
RP11-447B16	*FLJ10233*	Copy number decrease	13/15 cell lines.
RP11-107C3	None known	Copy number decrease	9/15 cell lines
RP11-29E11	None known	Copy number decrease	10/15 cell lines

6.3 References

Albertson DG, Pinkel D. 2003. Genomic microarrays in human genetic disease and cancer. Hum Mol Genet 12:R145-R152.

Ashman JN, Brigham J, Cowen ME, Bahia H, Greenman J, Lind M, Cawkwell L. 2002. Chromosomal alterations in small cell lung cancer revealed by multicolour fluorescence in situ hybridization. Int J Cancer 102:230-236.

Balsara BR, Testa JR. 2002. Chromosomal imbalances in human lung cancer. Oncogene 21:6877-6883.

Bremnes RM, Sundstrom S, Aasebo U, Kaasa S, Hatlevoll R, Aamdal S. 2003. The value of prognostic factors in small cell lung cancer: results from a randomised multicenter study with minimum 5 year follow-up. Lung Cancer 39:303-313.

Bruder CE, Hirvela C, Tapia-Paez I, Fransson I, Segraves R, Hamilton G, Zhang XX, Evans DG, Wallace AJ, Baser ME, Zucman-Rossi J, Hergersberg M, Boltshauser E, Papi L, Rouleau GA, Poptodorov G, Jordanova A, Rask-Anderson H, Kluwe L, Mautner V, Sainio M, Hung G, Mathiesen T, Moller C, Pulst SM, Harder H, Heiberg A, Honda M, Niimura M, Sahlen S, Blennow E, Albertson DG, Pinkel D, Dumanski JP. 2001. High resolution deletion analysis of constitutional DNA from neurofibromatosis type 2 (NF2) patients using microarray-CGH. Hum Mol Genet 10:271-282.

Buckley PG, Mantripragada KK, Benetkiewicz M, Tapia-Paez I, Diaz De Stahl T, Rosenquist M, Ali H, Jarbo C, De Bustos C, Hirvela C, Sinder Wilen B, Fransson I, Wedell A, Beare DM, Collins JE, Dunham I, Albertson D, Pinkel D, Bastian BC, Faruqi AF, Lasken RS, Ichimura K, Collins VP, Dumankski JP. 2002. A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. Hum Mol Genet 11:3221-3229.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863-14868.

Garnis C, Coe BP, Ishkanian A, Zhang L, Rosin MP, Lam WL. 2004. Novel regions of amplification on 8q distinct from the MYC locus and frequently altered in oral dysplasia and cancer. Genes Chromosomes Cancer 39:93-98.

Henderson LJ, Okamoto I, Lestou VS, Ludkovski O, Robichaud M, Chhanabhai M, Gascoyne RD, Klasa RJ, Connors JM, Marra MA, Horsman DE, Lam WL. 2004. Delineation of a minimal region of deletion at 6q16.3 in follicular lymphoma and construction of a bacterial artificial chromosome contig spanning a 6-megabase region of 6q16-q21. Genes Chromosomes Cancer 40:60-65.

Hyman E, Kauraniemi P, Hautaniemi S, Wolf M, Mousses S, Rozenblum E, Ringer M, Sauter G, Monni O, Elkahloun A, Kallioniemi OP, Kallioniemi A. 2002. Impact of DNA amplification on gene expression patterns in breast cancer. Cancer Res 62:6240-6245

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12:996-1006.

Kriek M, White SJ, Bouma MC, Dauwerse HG, Hansson KB, Nijhuis JV, Bakker B, van Ommen GJ, den Dunnen JT, Breuning MH. 2004. Genomic imbalances in mental retardation. J Med Genet 41:249-255

Krug L, Miller V. 2003. Introduction: Small cell lung cancer - a frustrating disease. Semin Oncol 30:1-2.

Levin NA, Brzoska P, Gupta N, Minna JD, Gray JW, Christman MF. 1994. Identification of frequent novel genetic alterations in small cell lung carcinoma. Cancer Res 54:5086-5091.

Ried T, Peterson I, Holtgreve-Grez H, Speicher MR, Schrock E, du Manoir S, Cremer T. 1994. Mapping of multiple DNA gains and losses in primary small cell lung carcinomas by comparative genomic hybridization. Cancer Res 54:1801-1806

Sattler M, Salgia R. 2003. Molecular and cellular biology of small cell lung cancer. Semin Oncol 30:57-71.

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG. 2001. Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet 29:263-264.

Stewart DR, Huang A, Faravelli F, Anderlid BM, Medne L, Ciprero K, Kaur M, Rossi E, Tenconi R, Nordenskjold M, Gripp KW, Nicholson L, Meschino WS, Capua E, Quarrell OW, Flint J, Irons M, Giampietro PF, Schowalter DB, Zaleski CA, Malacarne M, Zackai EH, Spinner NB, Krantz ID. 2004. Subtelomeric deletions of chromosome 9q: a novel microdeletion syndrome. Am J Med Genet 128A:340-351

The international human genome mapping consortium. 2001. A physical map of the human genome. Nature 409:934-941.

Veltman JA, Fridlyand J, Pejavar S, Olshen AB, Korkola JE, DeVries S, Carroll P, Kuo WL, Pinkel D, Albertson D, Cordon-Cardo C, Jain AN, Waldman FM. 2003. Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. Cancer Res 63:2872-2880.

Yokoi S, Yasui K, Saito-Ohara F, Koshikawa K, Iizasa T, Fujisawa T, Terasaki T, Horii A, Takahashi T, Hirohashi S, Inazawa J. 2002. A novel target gene, SKP2, within the 5p13 amplicon that is frequently detected in small cell lung cancers. Am J Pathol 161:207-216.

Yokoi S, Yasui K, Iizasa T, Takahashi T, Fujisawa T, Inazawa J. 2003. Down-regulation of SKP2 induces apoptosis in lung-cancer cells. Cancer Sci 94:344-349.

Chapter 7: Gain of a region on 7p22.3, containing MAD1L1, is the most frequent event in small-cell lung cancer cell lines

A version of this chapter has been previously published as:

Coe BP, Lee EH, Chi B, Girard L, Minna JD, Gazdar AF, Lam S, MacAulay C, Lam WL. (2006) 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**(1):11-19 © 2006 Wiley-Liss, Inc doi:10.1002/gcc.20260

7.1 Introduction

Small cell lung cancer (SCLC) is a highly aggressive neoplasia, which represents 15 to 20% of yearly lung cancer cases. (Al-Ajam et al., 2005)Two stages are used to describe SCLC clinically: Limited and Extensive. Patients presenting with limited stage disease (33% of cases) exhibit median survival times of 18 months, with long-term survival in 4-5% of cases, while cases of extensive disease exhibit median survival times of 9 months with virtually no long-term survivors. (Simon & Wagner, 2003; Weinmann et al., 2003)The poor survival of this disease is partially due to the fast growth of the lesions and tendency for early widespread metastasis. As most cases present with metastasis at the time of diagnosis, chemotherapy is the major treatment regime, however relapse is almost inevitable. (Walker, 2003)

Many approaches have been applied to identify genes disrupted in SCLC for the purpose of putative marker and therapeutic target discovery. Chromosomal Comparative Genomic Hybridization (CGH) studies, and Loss of Heterozygosity (LOH) analysis aimed to identify genetic alterations, however these studies are limited by the low resolution or limited coverage of these technologies. (Girard et al., 2000; Levin et al., 1994; Ried et al., 1994)Additionally, gene expression profiling studies aimed to identify genes specifically over or under-expressed in SCLC, but the lack of knowledge of a progenitor cell type precludes the definition of "normal" gene expression levels. As a result, expression profiling efforts contribute mainly to disease sub-classification. (Bhattacharjee et al., 2001; Virtanen et al., 2002)

The advent of array CGH technology allows us to examine the genome at resolutions much higher than conventional CGH and much greater coverage than LOH. By replacing the metaphase spreads, which serve as the hybridization target in conventional CGH, with ordered sets of bacterial artificial chromosome (BAC) clones, each representing ~150 kbp of unique human sequence, resolution can be improved by 10 to 100 fold. (Ishkanian et al., 2004; Snijders et al., 2001)

Previously we produced high-resolution (10 overlapping clones per megabase) CGH arrays spanning chromosome arms 1p and 5p and utilized these arrays to fine map novel alterations in SCLC cell lines. (Coe et al., 2005; Henderson et al., 2005)The recently developed Sub-Megabase Resolution Tiling-set (SMRT) CGH array improves on these technologies by allowing simultaneous measurement of 32,433 overlapping chromosomal loci, spanning the sequenced human genome. (Ishkanian et al., 2004)

Due to the rarity of surgical resection, SCLC samples are difficult to acquire. As such cell lines have proven invaluable in the research of this disease. In this study, we profiled 14 SCLC cell lines and 6 matched normal B cell lines utilizing the Sub-Megabase Resolution Tiling-set (SMRT) CGH array, generating the highest resolution copy number maps produced to date of small cell lung cancer genomes. Here we report that the analysis of SCLC genomes identified many genomic alterations, including a novel high frequency gain at the distal end of chromosome 7p.

7.2 Results and Discussion

7.2.1 Generation of High Resolution SMRT array CGH profiles of SCLC Cell Lines

The SMRT array represents the first tool allowing tiling analysis of copy number changes across the entire human genome in a single experiment. Co-hybridization of differentially labeled sample and male reference DNA to the SMRT array has allowed simultaneous analysis of DNA copy number at 32,433 overlapping genomic loci producing the highest-resolution copy number maps of SCLC cell lines to date. Male reference DNA was used for all hybridizations, regardless of the sample sex to limit the influence of reference-based variations in our results. For this reason as well as the effect of the pseudoautosomal content of chromosome X on hybridization results, we have excluded analysis of chromosome X in our results. Profiles of the 14 SCLC cell lines and 6 matched normal lines are available at: http://www.bccrc.ca/cg/ArrayCGH_Group.html

The copy number maps generated for these SCLC cell lines were first visualized using our in-house developed SeeGH viewing software. By visual analysis multiple discontinuous regions of gain and loss, representing multiple levels of copy number in each sample, are readily identified. The tiling path nature of the SMRT array enables the fine-mapping of breakpoints to within a single BAC clone. In addition to detecting gains and losses of chromosomal regions, we have also detected micro-alterations as small as 200 kbp in these samples, which would have escaped detection by previous whole genome screens (**Figure 7.1**).

To supplement our visual analysis we also utilized an automated alteration identification technique. The aCGH-Smooth application developed by Jong et al "simplifies" aCGH data by representing each unique copy number level it detects with a single Log2 ratio value. (Jong et al., 2004)The use of this aCGH-Smooth application allows us to easily identify the alterations present in our samples without tedious manual assessment of every locus.

In concordance with previous studies, the SCLC cell lines appear to be highly genomically unstable. (Engelholm et al., 1985; Girard et al., 2000; Levin et al., 1994; Ried et al., 1994)On average 56.0% (range 11.3% to 87.8%) of the genomic loci are altered in each line, with 58% of these alterations representing gains and 42% representing regions of loss.

7.2.2 Identification of Frequent Alterations

Due to their highly altered genomes and the large volume of data generated by 14 SMRT array analyses of the SCLC cell lines, simple visual inspection of altered loci proved inefficient to identify minimal recurrent regions of alteration. Additionally line summaries, which are commonly applied to CGH studies (example in Reid et al.), (Ried et al., 1994) fail to accurately represent the level of detail presented in high-resolution copy number data sets.

To analyze this high-resolution data set we combined automatic alteration identification (as described above) with a modified frequency plot procedure. To generate a measure of alteration frequency we first assigned scores of -1/+1 to each clone representing first level loss or gain respectively.

In order to preserve the valuable fine mapping information that can be gained from detection of second level gains and losses which are found within or adjacent to a background of single copy gain or loss we added a weighting of +/- (1/(n+1) to the score assigned to these loci, where n is the number of samples analyzed. The denominator of n+1 was chosen to make obvious the contribution of second level gains or losses. Summing the amplification and deletion scores for each locus across all experiments and dividing the total scores by n+1 (the +1 is added to account for the weighting factor) generated a "weighted" frequency of alteration for each locus. The largest value possible for n samples is (n+(n/(n/+1)))/(n+1) which for large n approaches 1. This data was displayed utilizing a modified version of the SeeGH software package, which is designed to display frequency data (**Figure 7.2**). The frequency information for each clone is downloadable from: http://www.bccrc.ca/cg/ArrayCGH_Group.html.

Application of weighted frequency analysis to the autosomes of SCLC cell lines reveals multiple regions of highly recurrent change, many of which are much smaller than a single chromosome band. The pattern of gains (chromosome 1,3q,5p,8q,17,18,19,20) and losses (chromosome 3p,4q,5q,8p,10,13q,17p) discussed by Balsara and Testa in their review of chromosomal imbalance in human lung cancer as detected by conventional techniques is also present in our data set representing good concordance between our data set and those generated by conventional methods. (Balsara & Testa, 2002) In support of our method's validity, we detected and fine-mapped minimal regions of alteration containing genes previously linked to SCLC such as *hTERT*, *MYC*, *MYCL*, *CDK4*, *TRIO*, and others (Coe et al., 2005; Johnson et al., 1988; Sattler & Salgia, 2003)(**Figure 7.3a**). FISH was used to validate the *MYC* amplification in the SCLC cell line NCI-H524 (**Figure 7.3b**).

In addition to fine mapping these well known regions of alteration we detected many novel minimal regions of alteration. Setting of strict thresholds to detect only the peak alterations detected 2.6% of genomic loci as gained with a weighted frequency score of > 0.75 and 3.8% of loci lost with a score of < -0.65.

7.2.3 Profiling of Matched Normal Cell Lines

Recently a study demonstrated that the normal human genome to be more variable than previously thought. (Sebat et al., 2004) The discovery of wide spread DNA copy number polymorphisms prompted us to profile DNA from the7 available normal B-lymphocyte (BL) cell lines, which were derived from the same patients as the SCLC cell lines. These copy number profiles were then used to validate the somatic nature of the alterations we detected. Initial analysis of these matched normal cell lines showed, as expected, stable genomes with few large copy number alterations.

Analysis of these copy number profiles confirmed the presence of several known copy number polymorphisms such as the *SMA3* gene on 5q, which is seen as part of a 500 kbp duplicon or deletion in many normal individuals (**Figure 7.4a**). (Lefebvre et al., 1995; Sebat et al., 2004)Interestingly we also detected non-somatic alterations, which may be linked to tumorigenesis. An interesting example of this is seen as a deletion of ~3 Mbp on chromosome 11q22.3 in both NCI-H1672 and its matched normal BL1672 (**Figure 7.4b**). This deletion encompasses several apoptosis relates genes (*CASP1*, *CASP4 CASP5* and *ICEBERG*). (Druilhe et al., 2001) Due to the tumor suppressing nature of apoptotic genes such as the caspase gene family this raises the possibility of genetic susceptibility in this individual. This case clearly demonstrates the benefit of separately analyzing matched normal and tumor samples against a common DNA reference. Had the tumor cell line been hybridized directly against the matched normal we would fail to detect non-somatic alterations. Future profiling of a large set of normal individuals will be necessary to clarify the true nature of these non-somatic changes, and what role (if any) they may play in SCLC development.

In addition to detection of alterations shared with the tumor lines we detected alterations unique to the normal cell lines. The alterations, which are not B cell specific, such as the Immunoglobulin rearrangement on 22q (**Figure 7.4c**) can most likely be attributed to culturing artifacts from the many passages of these lines since their creation. Details of the subset of alterations, which were present in both the BL and SCLC cell lines with identical boundaries, are present in **Table 7.1**.

7.2.4 7p22.3 is the most frequently gained region in SCLC Lines

The most highly recurrent copy number alteration detected in the SCLC cell lines is a gain present on 7p22.3. This gain is present in 13/14 SCLC cell lines and 0/6 matched normal B cell lines. The spectrum of alterations, we detect at this locus are of varying sizes and levels of copy number increase. Examples are demonstrated in **Figure 7.5a&b**. The minimal common region of gain is represented by a microalteration of ~350 kbp present in 4 of the SCLC cell lines, while 9 cell lines demonstrate wider regions of alteration (**Figure 7.5a**). FISH validation of the microalteration was performed on NCI-H2107 and demonstrated an increased signal count between the region of copy number gain (RP11-414M15) and a neighboring retention (RP11-436P19) (**Figure 7.5c**).

To determine if this alteration may also be relevant to non-small cell lung cancers we probed a lung cancer tissue microarray for specific aneuploidy at the 7p22.3 microalteration loci. Microalterations are often observed as tandem duplications, which may escape detection by FISH analysis. (Christian et al., 1999; Gervasini et al., 2002; Lefebvre et al., 1995) Additionally the tissue microarray contained very few cores for which adequate digestion conditions could be established. Despite this, FISH analysis of squamous lung cancer cores detected specific gain of RP11-414M15 when compared to a centromere probe (172 to 114 signals in 50 nuclei) in one of four enumerable cases. (**Figure 7.5d**) Two of the enumerable cases exhibited an increase in per nuclei counts for both the centromere and RP11-414M15, and one case failed to detect an increase at RP11-414M15.

Alignment of the minimal common alteration to the UCSC April 2003 assembly identified a single gene centered at the minimal amplified region (**Figure 7.5e**). Mitotic arrest deficient-like protein 1 like 1 (*MAD1L1*) is the human homologue of a yeast gene (*MAD1*) involved in cell growth control and inhibition of entry into S phase until late G1. (Rottmann et al., 2005) Studies have demonstrated that over-expression of *MAD1L1* inhibits cell proliferation, and prevents resting cells from re-entering cell cycle. (Gehring et al., 2000)Recently a report demonstrated that Cyclin E and CDK2, which are both gained with moderate frequency in our sample set (Weighted Frequency Scores of 0.48 and 0.54 respectively), antagonize the MAD1L1 dependent inhibition of proliferation through an unknown mechanism. (Rottmann et al., 2005)Additional studies have shown that MAD1L1 interacts with MYC both as an antagonist preventing MYC/MAX heterodimerization, which regulates the transcription of genes involved in both proliferation and apoptosis, and by associating with MYC to enforce inhibition of *hTERT* transcription. (Ohta et al., 2002; Zou et al., 2005)

The proliferation inhibiting nature of MAD1L1 is seemingly contradictory to our results, which identify frequent genomic gain at this locus in small cell lung cancer as well as a gain in a squamous cell lung tumour. A recent report of copy number gain at 7p22.3 in mantle cell lymphoma supports our observation. (de Leeuw et al., 2004)Interestingly another recent study has discovered that when the *hTERT* promoter E-Box is mutated MAD1 reverses its usual role and enforces expression of the mutant hTERT. (Zou et al., 2005)These contradictory observations suggest a potentially complex role for *MAD1L1* in the development of small cell lung cancer as well as other tumor types, however future functional studies will be required to understand the implication of this alteration in SCLC biology.

7.3 Conclusions

In this brief article, we have shown SMRT array CGH to be a very powerful tool in the analysis of SCLC genomes. The high resolution copy number maps produced for these lines have enabled the rapid fine mapping of multiple regions of highly recurrent copy number alteration. In addition to detecting and fine mapping regions containing previously characterized oncogenes we have identified regions of polymorphism and other non-somatic alterations. Strikingly, we identified a novel somatic copy number gain present at 7p22.3 in 14/15 cell lines. The minimal common region of alteration on 7p22.3 implicates a single gene in SCLC, which demonstrates characteristics that would imply a tumor suppressing nature, though recent accounts have shown that this gene may play an oncogenic role.

7.4 Materials and Methods

7.4.1 Cell Lines

The cell lines describe in this manuscript were established at the National Cancer Institute (NCI-H and BL series) and at the Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center (HCC series). These cell lines have been deposited for distribution in the American Type Culture Collection (http://www.atcc.org). DNA was extracted from 14 SCLC lines, 9 Classical (H187, H378, H889, H1607, H1672, H2107, H2141, H2171, and HCC33) and 5 Variant (H82, H289, H524, H526, and H841). Additionally DNA was extracted from 6 normal lymphocyte cell lines (BL289, BL1607, BL1672, BL2107, BL2141, and BL2171) derived from the same patients as the corresponding NCI-H cell lines. The cell lines were fingerprint verified using the Powerplex 1.2 system (Promega, Nepean ON) which contains 9 polymorphic markers.

7.4.2 SMRT Array CGH

Array CGH was performed as previously described. (Coe et al., 2005; Henderson et al., 2005; Ishkanian et al., 2004)Briefly, in each hybridization experiment 400 ng of sample DNA and a common reference male genomic DNA (Novagen, Mississauga ON) were labeled with cyanine-5 dCTP and cyanine-3 dCTP (PerkinElmer, Woodbridge ON), respectively, through a random priming reaction and were co-hybridized to the SMRT array. Post-hybridization images were captured using an ArrayWorx CCD based imaging system and analyzed using the SoftWorx array analysis platform (Applied Precision, Issaquah WA).

7.4.3 Automated Alteration Identification

Automated identification of segmental gains and losses prior to frequency scoring was performed using the aCGH-Smooth application described by Jong et al. which utilizes a breakpoint detection system to first identify alteration boundaries within array CGH data and then smoothes the ratios between breakpoints to a single value via a clustering algorithm. (Jong et al., 2004)The default program parameters were designed for 1 Mbp CGH arrays. We empirically optimized the following non-default parameters for the breakpoint detection algorithm to better analyze out high-resolution data: Lambda set to 6.75 and the maximum number of breakpoints in initial pool to 100. Since the SMRT array data file sizes are being beyond the scope of the aCGH-Smooth application to analyze as a whole, we smoothed chromosomes 1-12 and 13-Y independently and then combined the results.

7.4.4 Frequency Plot Application

The Frequency Plot program was developed in the Borland C++ environment, and was tested on Microsoft Windows 2000 and Windows XP. The framework is similar to that of our previous SeeGH application, (Chi et al., 2004)which is freely available online and utilizes a MySQL backend database for all storage and retrieval of data. The application will accept as input any tab delimited text file with the following required inputs for each data point: Unique ID, UCSC base pair position, Chromosome number, Clone Name, Accession Number, Amplification score (out of 1), Deletion Score (out of 1). For ease of data retrieval the following descriptors can be added to each experiment imported: BarCode / Unique ID, Disease Type, Experimenter, Date, and Comments. The Frequency Plot software package is downloadable at: http://www.flintbox.com/technology.asp?tech=FB706FB.

7.4.5 Fluorescence In-Situ Hybridization

Cell line Fluorescence In-Situ hybridization (FISH) experiments were performed as previously described. (Coe et al., 2005; Henderson et al., 2005; Watson et al., 2004)Briefly 300 ng of linker mediated PCR amplified BAC DNA was labeled through a random priming reaction with either Spectrum Red or Spectrum Green dUTP (Vysis, Markham ON). Hybridizations were performed in a 50% formamide buffer at 37°C for 18 hours and images were acquired using a Zeiss Axioplan fluorescence microscope (Zeiss, Toronto ON) and Northern Eclipse Microscopy Software Package (Empix Imaging, Mississauga ON).

Human Lung Cancer (IMH-305) tissue microarrays (Imgenex, San Diego CA) were processed by conventional cytogenetic methods (Xylene deparaffinization followed by NaSCN treatment and pepsin digestion) prior to probe hybridization. Vysis CEP probe was used to validate chromosome 7 copy number and served as an internal control while random prime labeled RP11-414M15 measured copy number at 7p22.3. Images were acquired as a z-stack of ten one-micron spaced sections, deconvolved, and flattened using the XY Maximum Projection feature of the Northern Eclipse Software (Empix imaging). Signals were manually counted to detect copy number alterations.

Figure 7.1

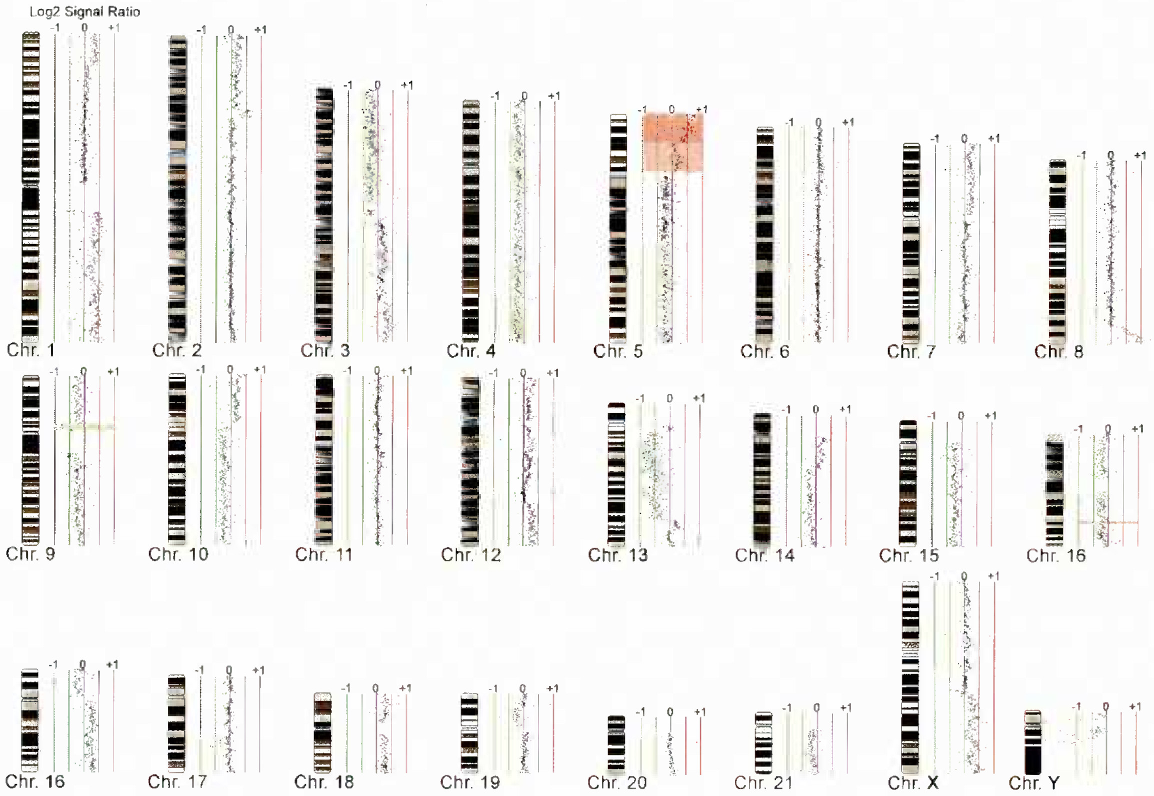


Figure 7.1. SMRT Array Profile of the SCLC Cell Line NCI-H289. Data is presented as a SeeGH karyogram. Each BAC clone on the array is displayed as a line representing the segment of the genome covered, at the measured Log2 Signal Ratio in a competitive hybridization with normal male genomic DNA. The shift of each data point to the left of 0 represents a decrease of copy number while a shift to the right represents an increase in copy number. Multiple levels of segmental copy number gain and loss can readily be detected (examples in dark red, light red and green highlighting). Additionally we detect microalterations which would have escaped detection by conventional screens (example highlighted in red on 16q).

Figure 7.2

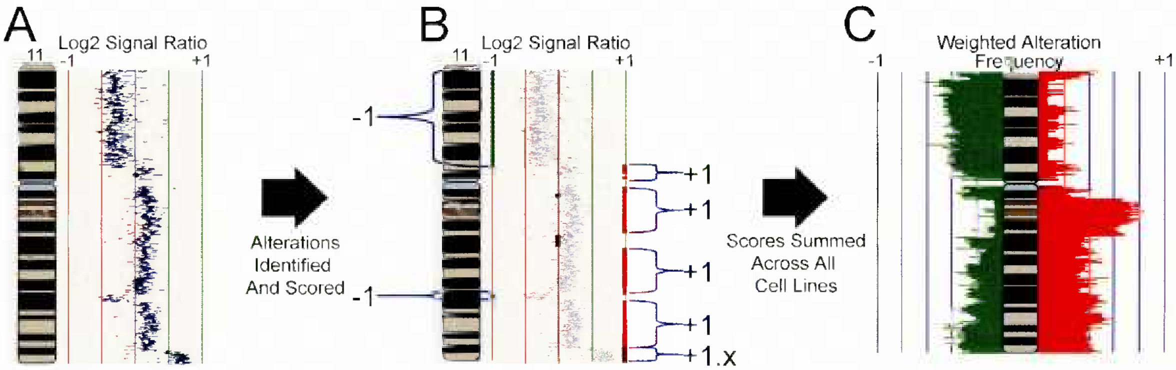


Figure 7.2. Overview of Weighted Frequency Scoring. (**a**) SMRT array CGH profile of chromosome 11 in the SCLC cell line NCI-H1672. (**b**) Scoring of alterations. Vertical lines represent regions defined as deletion (green), amplification (red), and retention (black). The weighting applied to second level alterations is indicated by x, where x is 1/(n+1). (**c**) Example of frequency diagram generated after combing the scores for all of the SCLC experiments. Amplification weighted frequency is represented by red bars to the right of the chromosome ideogram while deletion weighted frequency is indicated as green bars to the left of the ideogram.

Figure 7.3

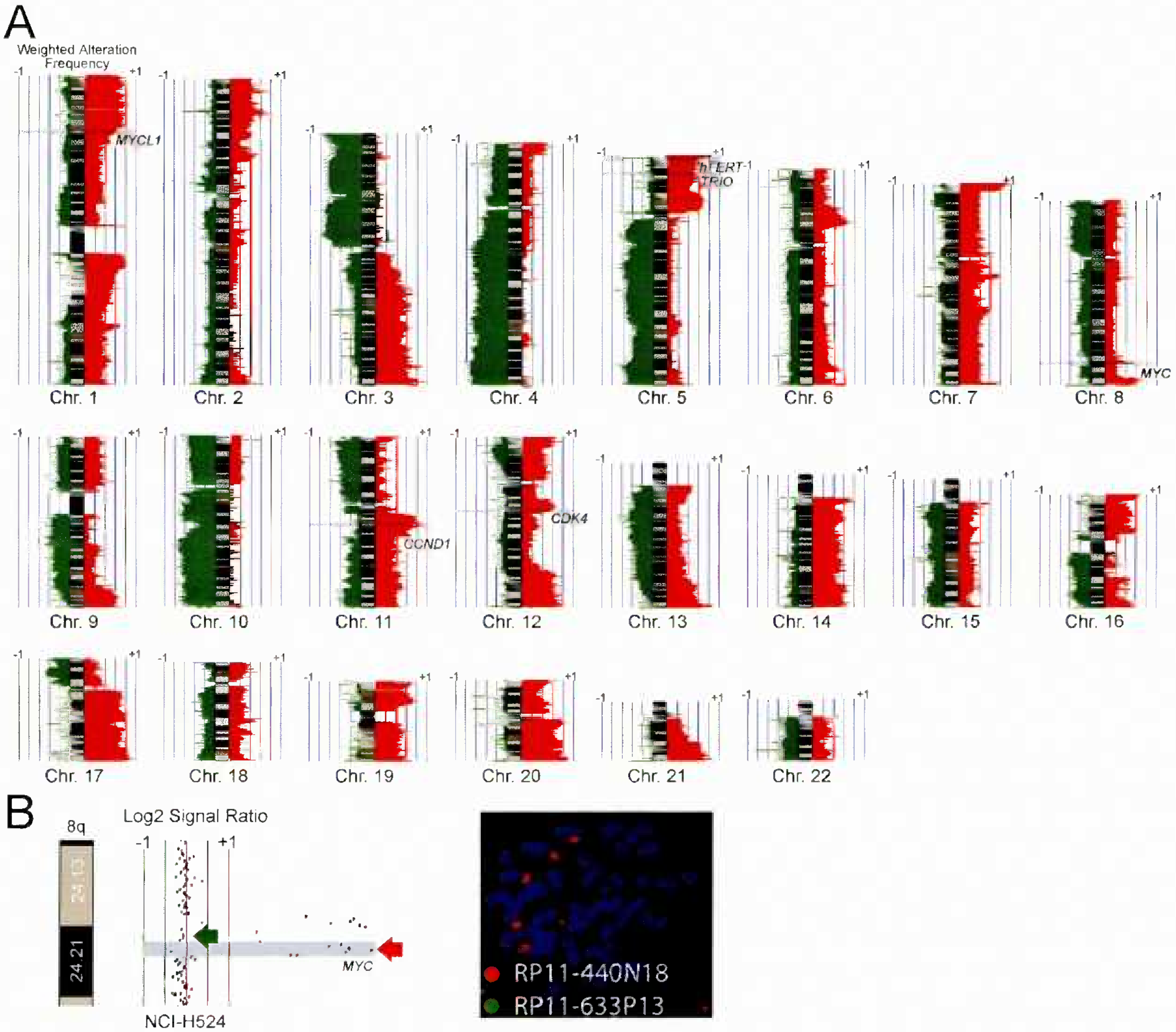


Figure 7.3. Weighted Frequency Analysis of Genomic Alterations in SCLC Cell Lines. (**a**) Weighted frequency profiles generated for the SCLC cell line SMRT array data. Blue highlighting indicates selected fine mapped regions of alteration, which contain genes previously linked to SCLC. (**b**) Amplification of MYC in SCLC. The SMRT array CGH profile of chromosome 8q24 in NCI-H524 demonstrates the minimal ~830kbp high-level amplification at the MYC loci. FISH validation with clones RP11-440N18 and RP11-633P13 confirmed the fine mapped alteration (right panel).

Figure 7.4

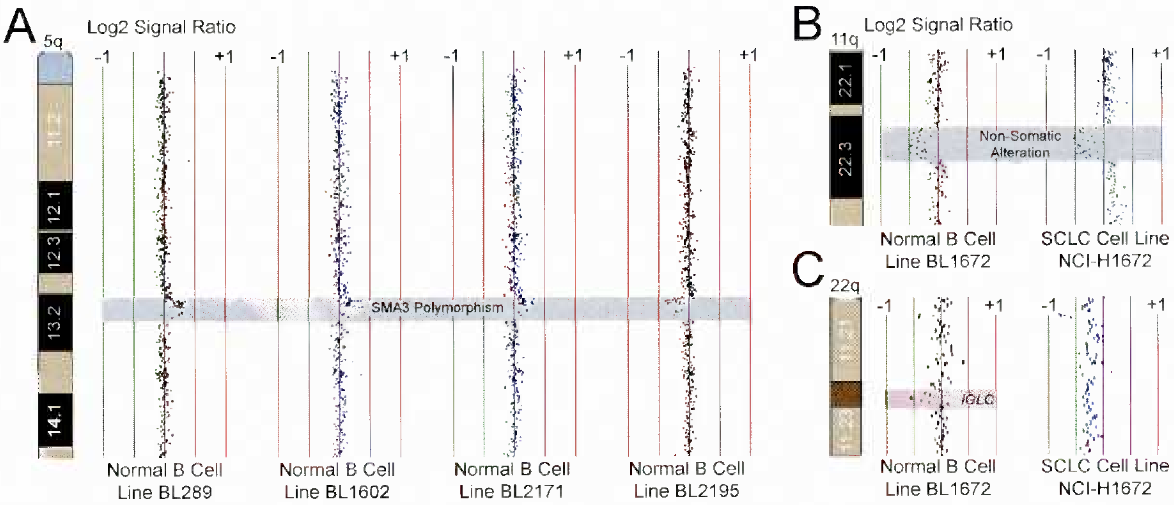


Figure 7.4. Identification of Copy Number Polymorphisms and Other Non-Somatic Alterations. (**a**) Detection of the SMA3 Polymorphism on Chromosome 5q. The SMA3 locus on chromosome 5q is well known to demonstrate variable copy number in the normal human population. Here we observe a ~2 Mbp region of alteration (highlighted in blue) demonstrating both single copy loss and gain in multiple samples. (**b**) Detection of non-somatic alterations. Deletion of a ~3 Mbp segment of chromosome 11q was detected in both the normal B-Cell line BL1672 and the SCLC cell line NCI-H1672 (blue Highlighting). This alteration contains several apoptosis related genes (CASP1, CASP4 CASP5 and ICEBERG) hinting at a cancer predisposing nature for this particular copy number alteration. (**c**) Deletion was detected specific to BL1672 at the IGLC (Immunoglobulin Lambda Constant 1) loci (pink highlighting).

Figure 7.5. The 7p22.3 Locus is Highly Frequently Gained in SCLC Cell Lines, and Present in Other Tumor Types. (**a**) Frequency plot delineating a highly frequent (13/14) gain at 7p22.3 and example profiles of cell lines demonstrating gain at 7p22.3. Copy number gain at 7p22.3 is the most frequent numerical aberration in SCLC cell lines as demonstrated by a sharp peak in the weighted frequency plot for this region. Alterations detected at this loci range from multi-megabase alterations (examples NCI-H289 and NCI-H107) to specific microalterations (examples NCI-H526 and NCI-H2141). (**b**) The minimal region of alteration is a somatic copy number gain. A segmental gain was observed spanning ~350 kbp in the cell line NCI-H2107 and was not present in the matched normal cell line BL2107. (**c**) RP11-414M15 (red) which is centered on the 7p22.3 microalteration was compared with the adjacent retained clone RP11-436P19 (green) by FISH analysis in NCI-H2107. Specific copy number gain was observed, and the displayed nuclei, presents a 7 to 5 allele ratio between these loci. (**d**) Detection of 7p22.3 gain in a squamous cell lung cancer tumor. RP11-414M15 (red) was compared against Vysis chromosome 7 enumeration probe (CEP7 green) by FISH analysis of a tissue microarray. Displayed is a representative section of a squamous cell lung cancer tissue core which displayed specific gain of 7p22.3 (172 to 114 signals in 50 nuclei). (**e**) Alignment of the SMRT array clones contained in the 7p22.3 region to the UCSC April 2003 genome browser freeze reveals a single gene (*MAD1L1*) at the microalteration locus. Array clones are color coded according to their respective Log2 Ratios in the NCI-H2107 cell line. Gray clones represent non-informative loci; while shades of red and green represent varying degrees of gain and loss (lighter shades represent higher levels of alteration).

Figure 7.5

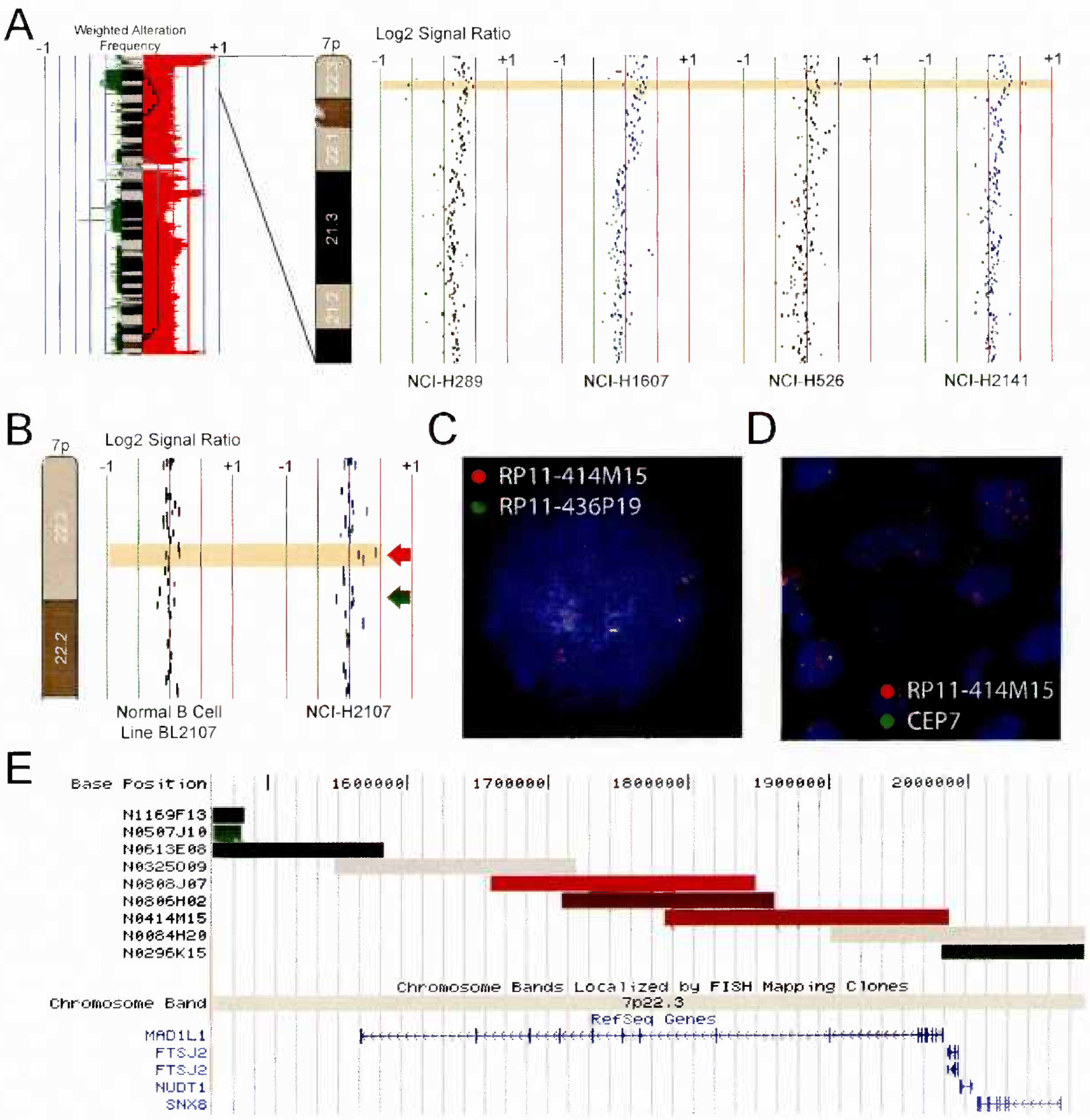

A
Weighted Alteration Frequency
-1
+1
7p
22.3
22.1
21.3
21.2
Log2 Signal Ratio
-1
+1
-1
+1
-1
+1
-1
+1
NCI-H289
NCI-H1607
NCI-H526
NCI-H2141
B
7p
22.3
22.2
Log2 Signal Ratio
-1
+1
-1
+1
Normal B Cell
Line BL2107
NCI-H2107
C
RP11-414M15
RP11-436P19
D
RP11-414M15
CEP7
E
Base Position
1600000
1700000
1800000
1900000
2000000
N1169F13
N0507J10
N0613E08
N0325O09
N0808J07
N0806H02
N0414M15
N0084H20
N0296K15
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
7p22.3
RefSeq Genes
MAD1L1
FTSJ2
FTSJ2
NUDT1
SNX8

Table 7.1. Alterations with Common Boundaries in Blood Lymphocytes and SCLC Cell Lines

Sample	Alteration Type	Chromosome	Start (Mbp)	End (Mbp)	Size (Mbp)
H289/BL289	Gain	5q13.2	69.0	70.3	1.3
	Gain	12q21.31	81.5	81.7	0.2
	Loss	15q13.2	27.8	28.6	0.7
	Loss	15q13.3	30.0	30.5	0.6
	Loss	15q11.2	18.5	19.8	1.3
H1607/BL1607	Gain	1q42.12	223.1	223.3	0.2
	Gain	5q13.2	68.9	69.8	0.9
H1672/BL1672	Loss	1p33	48.4	48.7	0.3
	Loss	2p16.2	50.6	51.0	0.5
	Gain	5q23.3	130.5	130.7	0.2
	Loss	11p15.3	11.4	11.7	0.4
	Loss	11q22.3	104.0	106.9	2.9
	Gain	18q21.1	44.3	44.5	0.2
	Loss	22q11.21	17.0	17.5	0.5
	Loss	22q11.21	19.7	19.9	0.2
H2107/BL2107	Gain	1p36.13	15.9	16.2	0.4
	Loss	2p11.2	88.9	89.2	0.3
	Loss	2p11.2	89.7	90.0	0.3
	Loss	2q32.3	193.9	194.2	0.3
	Loss	3q26.3-3q29	177.5	199.2	21.7

	Loss	Chr6	Whole arm Loss		
	Loss	8p23.1	7.1	7.9	0.8
	Gain	8q11.21	49.5	49.8	0.3
	Loss	11q22.3-11q23.2	107.7	114.3	6.6
	Gain	13q31.1	82.0	82.4	0.4
	Loss	14q32.33	104.3	105.3	1.0
	Loss	17p13.3	2.0	2.3	0.3
	Loss	22q11.21	17.0	17.3	0.3
	Loss	22q11.22	21.1	21.6	0.5
H2141/BL2141	Gain	1q25.2-1q25.3	176.6	176.9	0.3
	Loss	5q35.3	178.6	178.8	0.2
	Loss	9q13	62.9	63.0	0.1
	Loss	10p15.1	4.0	4.2	0.2
	Gain	16p13.11	16.2	16.8	0.6
H2171/BL2171	Gain	5p11	45.5	45.9	0.4
	Gain	14q11.2	18.1	18.4	0.3
	Gain	22q11.1	14.4	14.9	0.5

* Start and End Positions are Based on UCSC April 2003 Assembly

7.5 References

Al-Ajam, M., Seymour, A., Mooty, M. & Leaf, A. (2005). Ten Years of Disease-Free Survival between Two Diagnoses of Small-Cell Lung Cancer: A Case Report and a Literature Review. *Med Oncol*, **22,** 89-98.

Balsara, B.R. & Testa, J.R. (2002). Chromosomal imbalances in human lung cancer. *Oncogene*, **21,** 6877-83.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J. & Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*, **98,** 13790-5. Epub 2001 Nov 13.

Chi, B., DeLeeuw, R.J., Coe, B.P., MacAulay, C. & Lam, W.L. (2004). SeeGH--a software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics*, **5,** 13.

Christian, S.L., Fantes, J.A., Mewborn, S.K., Huang, B. & Ledbetter, D.H. (1999). Large genomic duplicons map to sites of instability in the Prader-Willi/Angelman syndrome chromosome region (15q11-q13). *Hum Mol Genet*, **8,** 1025-37.

Coe, B.P., Henderson, L.J., Garnis, C., Tsao, M.S., Gazdar, A.F., Minna, J., Lam, S., Macaulay, C. & Lam, W.L. (2005). High-resolution chromosome arm 5p array CGH analysis of small cell lung carcinoma cell lines. *Genes Chromosomes Cancer*, **42,** 308-13.

de Leeuw, R.J., Davies, J.J., Rosenwald, A., Bebb, G., Gascoyne, R.D., Dyer, M.J., Staudt, L.M., Martinez-Climent, J.A. & Lam, W.L. (2004). Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum Mol Genet*, **13,** 1827-37. Epub 2004 Jun 30.

Druilhe, A., Srinivasula, S.M., Razmara, M., Ahmad, M. & Alnemri, E.S. (2001). Regulation of IL-1beta generation by Pseudo-ICE and ICEBERG, two dominant negative caspase recruitment domain proteins. *Cell Death Differ*, **8,** 649-57.

Engelholm, S.A., Vindelov, L.L., Spang-Thomsen, M., Brunner, N., Tommerup, N., Nielsen, M.H. & Hansen, H.H. (1985). Genetic instability of cell lines derived from a single human small cell carcinoma of the lung. *Eur J Cancer Clin Oncol*, **21,** 815-24.

Gehring, S., Rottmann, S., Menkel, A.R., Mertsching, J., Krippner-Heidenreich, A. & Luscher, B. (2000). Inhibition of proliferation and apoptosis by the transcriptional repressor Mad1. Repression of Fas-induced caspase-8 activation. *J Biol Chem*, **275,** 10413-20.

Gervasini, C., Bentivegna, A., Venturin, M., Corrado, L., Larizza, L. & Riva, P. (2002). Tandem duplication of the NF1 gene detected by high-resolution FISH in the 17q11.2 region. *Hum Genet*, **110,** 314-21. Epub 2002 Mar 20.

Girard, L., Zochbauer-Muller, S., Virmani, A.K., Gazdar, A.F. & Minna, J.D. (2000). Genome-wide allelotyping of lung cancer identifies new regions of allelic loss, differences between small cell lung cancer and non-small cell lung cancer, and loci clustering. *Cancer Res*, **60,** 4894-906.

Henderson, L.J., Coe, B.P., Lee, E.H.L., Girard, L., Gazdar, A.F., Minna, J.D., Lam, S., Macaulay, C. & Lam, W.L. (2005). Genomic and gene expression profiling of minute alterations of chromosome arm 1p in small-cell lung carcinoma cells. *Br J Cancer*, **In Press**.

Ishkanian, A.S., Malloff, C.A., Watson, S.K., DeLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M.A., Ling, V., MacAulay, C. & Lam, W.L. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, **36,** 299-303. Epub 2004 Feb 15.

Johnson, B.E., Makuch, R.W., Simmons, A.D., Gazdar, A.F., Burch, D. & Cashell, A.W. (1988). myc family DNA amplification in small cell lung cancer patients' tumors and corresponding cell lines. *Cancer Res*, **48,** 5163-6.

Jong, K., Marchiori, E., Meijer, G., Vaart, A.V. & Ylstra, B. (2004). Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20,** 3636-7. Epub 2004 Jun 16.

Lefebvre, S., Burglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., Benichou, B., Cruaud, C., Millasseau, P., Zeviani, M. & et al. (1995). Identification and characterization of a spinal muscular atrophy-determining gene. *Cell*, **80,** 155-65.

Levin, N.A., Brzoska, P., Gupta, N., Minna, J.D., Gray, J.W. & Christman, M.F. (1994). Identification of frequent novel genetic alterations in small cell lung carcinoma. *Cancer Res*, **54,** 5086-91.

Ohta, Y., Hamada, Y., Saitoh, N. & Katsuoka, K. (2002). Effect of the transcriptional repressor Mad1 on proliferation of human melanoma cells. *Exp Dermatol*, **11,** 439-47.

Ried, T., Petersen, I., Holtgreve-Grez, H., Speicher, M.R., Schrock, E., du Manoir, S. & Cremer, T. (1994). Mapping of multiple DNA gains and losses in primary small cell lung carcinomas by comparative genomic hybridization. *Cancer Res*, **54,** 1801-6.

Rottmann, S., Menkel, A.R., Bouchard, C., Mertsching, J., Loidl, P., Kremmer, E., Eilers, M., Luscher-Firzlaff, J., Lilischkis, R. & Luscher, B. (2005). Mad1 function in cell proliferation and transcriptional repression is antagonized by cyclin E/CDK2. *J Biol Chem*.

Sattler, M. & Salgia, R. (2003). Molecular and cellular biology of small cell lung cancer. *Semin Oncol*, **30,** 57-71.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T.C., Trask, B., Patterson, N., Zetterberg, A. & Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, **305,** 525-8.

Simon, G.R. & Wagner, H. (2003). Small cell lung cancer. *Chest*, **123,** 259S-271S.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. & Albertson, D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet*, **29,** 263-4.

Virtanen, C., Ishikawa, Y., Honjoh, D., Kimura, M., Shimane, M., Miyoshi, T., Nomura, H. & Jones, M.H. (2002). Integrated classification of lung tumors and cell lines by expression profiling. *Proc Natl Acad Sci U S A*, **99,** 12357-62. Epub 2002 Sep 6.

Walker, S. (2003). Updates in small cell lung cancer treatment. *Clin J Oncol Nurs*, **7,** 563-8.

Watson, S.K., deLeeuw, R.J., Ishkanian, A.S., Malloff, C.A. & Lam, W.L. (2004). Methods for high throughput validation of amplified fragment pools of BAC DNA for constructing high resolution CGH arrays. *BMC Genomics*, **5,** 6.

Weinmann, M., Jeremic, B., Bamberg, M. & Bokemeyer, C. (2003). Treatment of lung cancer in elderly part II: small cell lung cancer. *Lung Cancer*, **40,** 1-16.

Zou, L., Zhang, P.H., Luo, C.L. & Tu, Z.G. (2005). Transcript regulation of human telomerase reverse transcriptase by c-myc and mad1. *Acta Biochim Biophys Sin (Shanghai)*, **37,** 32-8.

Chapter 8: Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer.

A version of this chapter has been previously published as:

Coe BP*, Lockwood WW*, Girard L, Chari R, MacAualy C, Lam S, Gazdar AF, Minna JD, Lam WL (2006) Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. *BJC* **94**:1927-1935 © 2006 Cancer Research UK doi:10.1038/sj.bjc.6603167

***These authors contributed equally.**

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

8.1 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.

8.2 Results and Discussion

8.2.1 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 $\log_2$ plots using SeeGH software. Genetic alterations were identified in all cell lines analyzed. **Figure 8.1** 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 8.1**). 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 on line at http://www.bccrc.ca/cg/ArrayCGH_Group.html.

8.2.2 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 8.2**). 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.

8.2.3 Regions of Similarity.

Among the regions that were not statistically different, there were some striking similarities (**Figure 8.2**). 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 8.2**).

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 8.2**). *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.

8.2.4 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 theses CNAs and identify genes within these regions responsible for the different cell phenotypes, gene expression analyses were required.

8.2.5 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 8.3**) (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 8.4**).

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.

8.2.6 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 8.5**. Twenty-two of the genes differentially altered between SCLC and NSCLC are components of the cell cycle, EGFR, MAPK, p38MAPK, and WNT pathways (**Table 8.1**). 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 8.5**. 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.

8.3 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.

8.4 Methods and Materials:

8.4.1 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.

8.4.2 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

8.4.3 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 $\log_2$ 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/.

8.4.4 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.

8.4.5 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.

8.4.6 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. Data is presented in supplementary material.

8.4.7 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 8.1

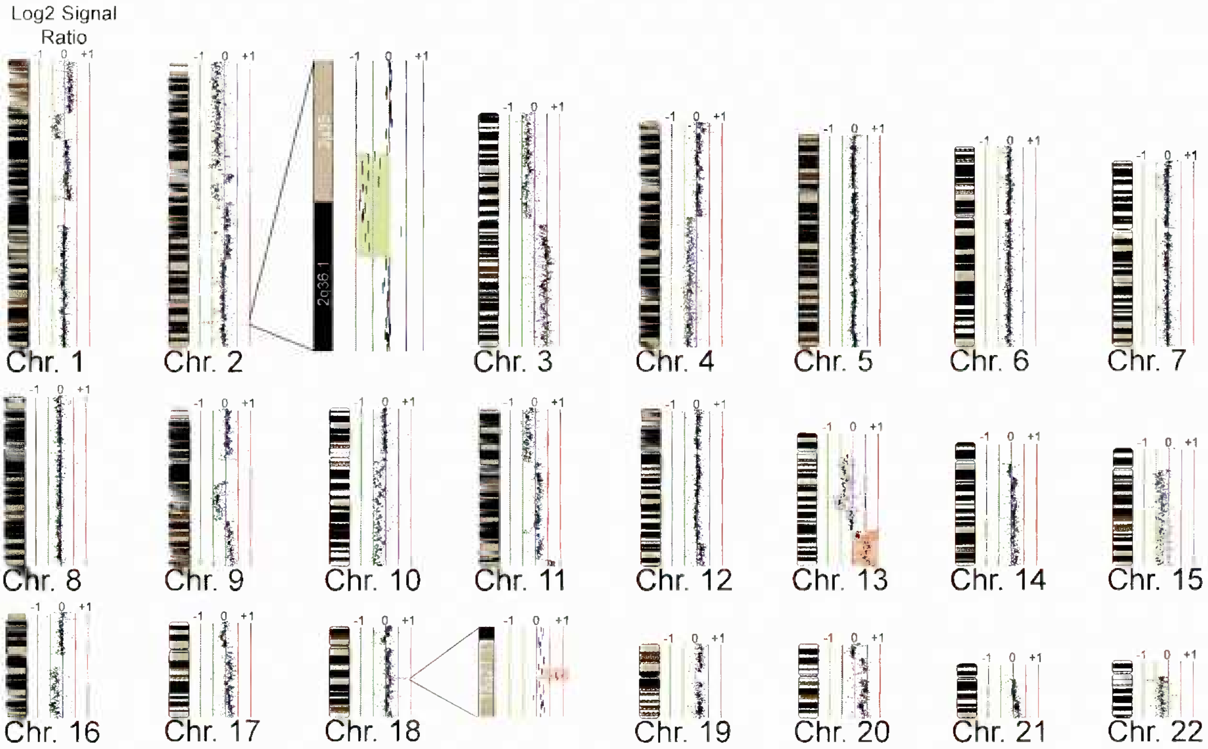


Figure 8.1. SMRT Array Profile of the SCLC NCI-H1672 Cells. Data is presented as a SeeGH karyogram to demonstrate the resolving power of the SMRT technology. Each BAC clone is displayed as a vertical line representing its genomic coverage. The horizontal shift of each line to the left or right of 0 represents 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 and 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 8.2

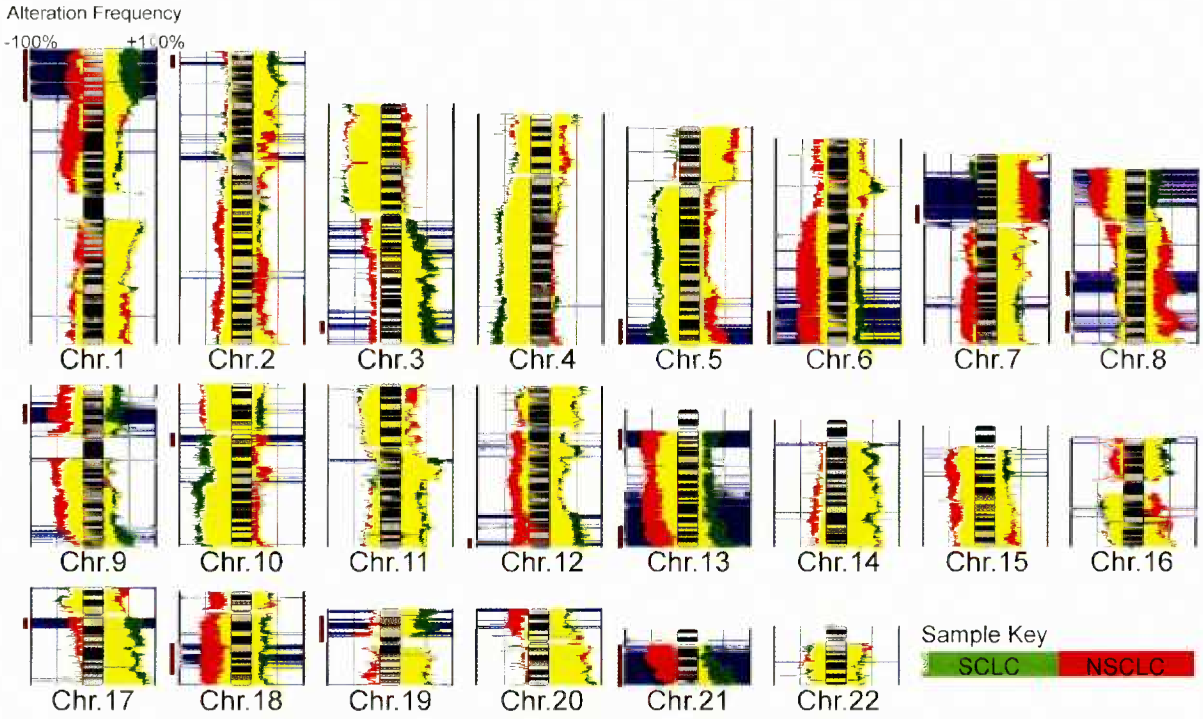


Figure 8.2. Copy Number Alterations in SCLC and NSCLC. Alteration frequencies for SCLC (green) and NSCLC (red) are displayed as bar plots adjacent to chromosomal ideograms. Bars extending to the right of each chromosome represent the frequency of copy number gain; conversely, bars extending to the left 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 8.3. 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 alteration 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 8.3

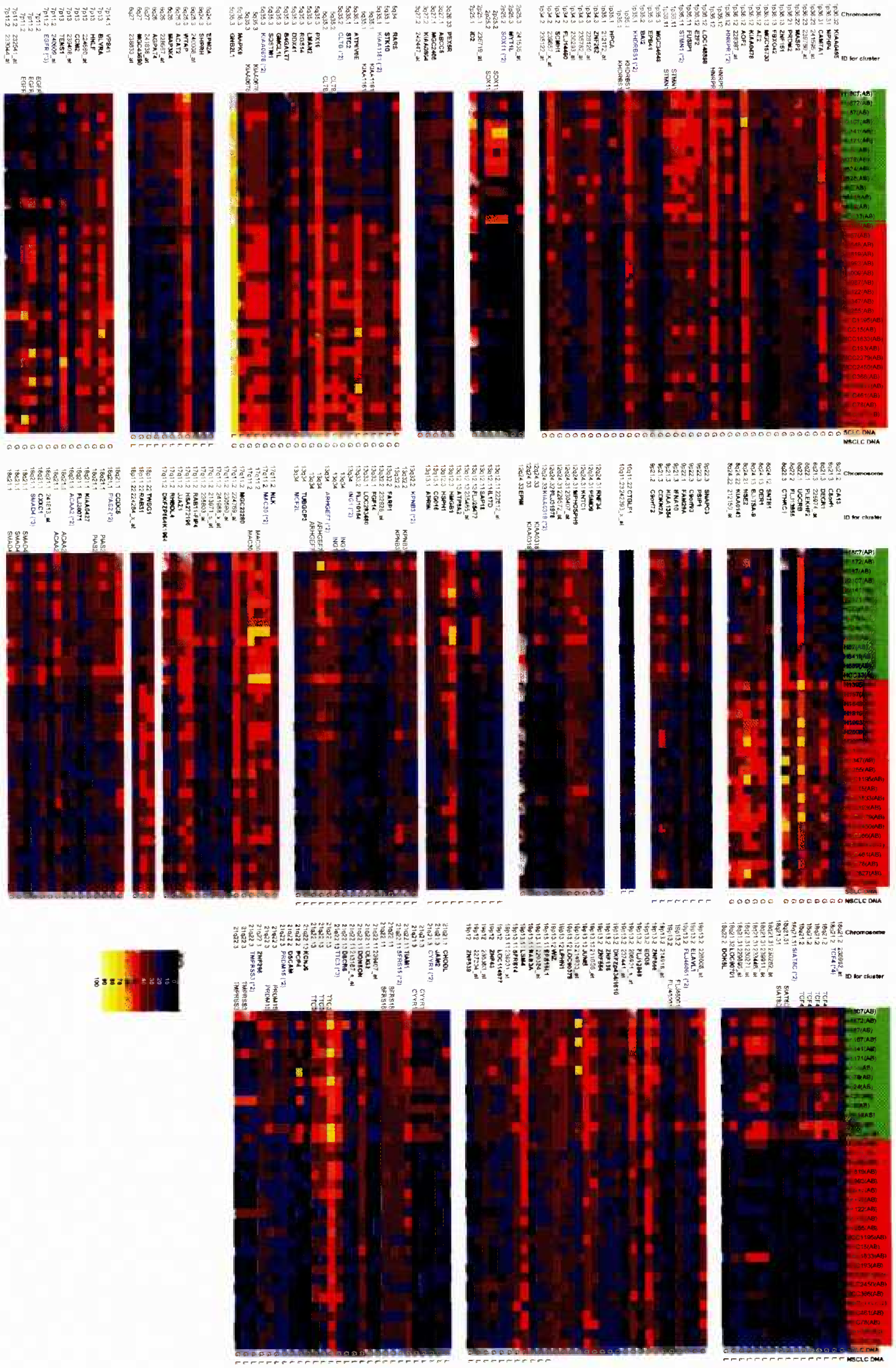

Chromosome
ID for cluster
SCLC DNA
NSCLC DNA
Chromosome
ID for cluster
NSCLC DNA
Chromosome
ID for cluster
SCLC DNA
NSCLC DNA

Figure 8.4

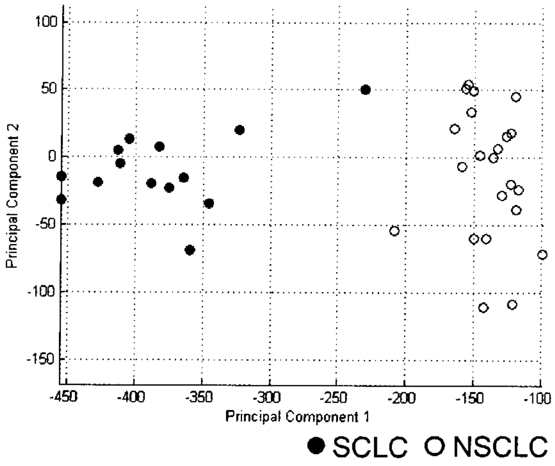


Figure 8.4. Contribution of Copy Number Induced Gene Expression Differences to the SCLC and NSCLC Phenotypes. Principal components analysis was performed utilizing all 243 Affymetrix probe sets demonstrating expression differences as a result of copy number alterations. The SCLC samples are indicated 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 8.5

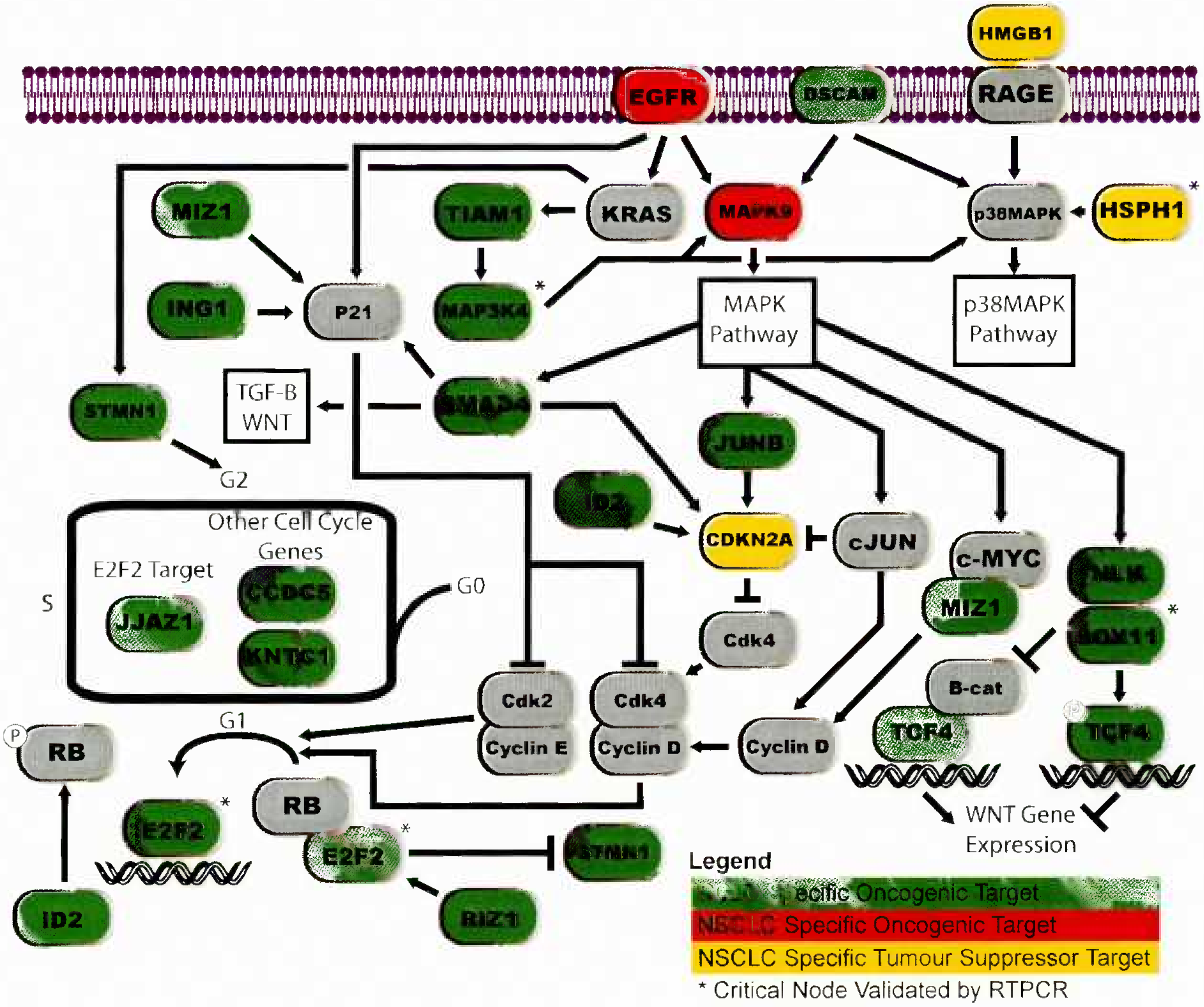


Figure 8.5. Differential Targets of Copy Number Induced Expression Changes in Key Biochemical Pathways between SCLC and NSCLC. Strong PSCNA induced expression differences were identified between SCLC and NSCLC in several key pro-proliferate pathways. Genes with increased expression in SCLC when compared 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 8.1. Differential Deregulation of Genes in Key Biochemical Pathways between NSCLC and SCLC

Gene Symbol	Gene Name	Locus	Regulation[1]
STMN1	stathmin 1	1p36.11	SCLC +
E2F2	E2F Transcription Factor 2	1p36.12	SCLC +
ZNF151 (MIZ1)	Zinc finger protein 151 (Myc-interacting zinc finger protein)	1p36.13	SCLC +
PRDM2 (RIZ1)	PR Domain-Containing Protein 3 (Rb Protein-Binding Zinc Finger Protein)	1p36.21	SCLC +
ID2	Inhibitor of DNA binding 2	2p25.1	SCLC +
SOX11	SRY-Related HMG-Box Gene 11	2p25.2	SCLC +
MAPK9 (JNK2)	Mitogen-activated protein kinase 9 (C-JUN Kinase 2)	5q35.3	NSCLC +
MAP3K4	Mitogen-activated protein kinase kinase kinase 4	6q26	SCLC +
EGFR	Epidermal Growth Factor Receptor	7p11.2	NSCLC +
CDKN2A (p16INK4A)	cyclin-dependent kinase inhibitor 2A	9p21.3	NSCLC -
KNTC1	Kinetochore-associated protein 1	12q24.31	SCLC +
HMGB1	High Mobility Group Box 1 (Amphoterin)	13q12.3	NSCLC -
HSPH1	Heat Shock 105kD	13q12.3	NSCLC -
ING1 (p33ING1)	Inhibitor of growth family member 1	13q34	SCLC +
JJAZ1 (SUZ12)	Joined to JAZF1 (Suppressor of ZESTE 12)	17q11.2	SCLC +
NLK	Nemo-like kinase	17q11.2	SCLC +
SMAD4	Mothers against decapentaplegic homolog 4	18q21.1	SCLC +
CCDC5	Coiled-coil domain containing 5	18q21.1	SCLC +
TCF4	Transcription Factor 4	18q21.2	SCLC +
JUNB	oncogene jun-B	19p13.1	SCLC +

		3	
TIAM1	T-cell lymphoma invasion and metastasis 1	21q22.1 1	SCLC +
DSCAM	Down syndrome cell adhesion molecule	21q22.2	SCLC +

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

8.5 References

Amann, J., Kalyankrishna, S., Massion, P.P., Ohm, J.E., Girard, L., Shigematsu, H., Peyton, M., Juroske, D., Huang, Y., Stuart Salmon, J., Kim, Y.H., Pollack, J.R., Yanagisawa, K., Gazdar, A., Minna, J.D., Kurie, J.M. & Carbone, D.P. (2005). Aberrant epidermal growth factor receptor signaling and enhanced sensitivity to EGFR inhibitors in lung cancer. *Cancer Res*, **65,** 226-35.

Balsara, B.R. & Testa, J.R. (2002). Chromosomal imbalances in human lung cancer. *Oncogene*, **21,** 6877-83.

Bracken, A.P., Pasini, D., Capra, M., Prosperini, E., Colli, E. & Helin, K. (2003). EZH2 is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer. *EMBO J*, **22,** 5323-35.

Campos, E.I., Chin, M.Y., Kuo, W.H. & Li, G. (2004). Biological functions of the ING family tumor suppressors. *Cell Mol Life Sci*, **61,** 2597-613.

Chi, B., DeLeeuw, R.J., Coe, B.P., MacAulay, C. & Lam, W.L. (2004). SeeGH--a software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics*, **5,** 13.

Coe, B.P., Lee, H.L., Chi, B., Girard, L., Minna, J.D., Gazdar, A.F., Lam, S., MacAulay, C. & Lam, W.L. (2005). Gain of a region on 7p22.3, containing MAD1L1, is the Most Frequent Event in Small Cell Lung Cancer Cell Lines. *Genes Chromosomes Cancer*, **In Press**.

de Leeuw, R.J., Davies, J.J., Rosenwald, A., Bebb, G., Gascoyne, R.D., Dyer, M.J., Staudt, L.M., Martinez-Climent, J.A. & Lam, W.L. (2004). Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum Mol Genet*, **13,** 1827-37.

Einarson, M.B., Cukierman, E., Compton, D.A. & Golemis, E.A. (2004). Human enhancer of invasion-cluster, a coiled-coil protein required for passage through mitosis. *Mol Cell Biol*, **24,** 3957-71.

Fogh, J., Wright, W.C. & Loveless, J.D. (1977). Absence of HeLa cell contamination in 169 cell lines derived from human tumors. *J Natl Cancer Inst*, **58,** 209-14.

Fu, M., Wang, C., Li, Z., Sakamaki, T. & Pestell, R.G. (2004). Minireview: Cyclin D1: normal and abnormal functions. *Endocrinology*, **145,** 5439-47.

Garnis, C., Lockwood, W.W., Vucic, E., Ge, Y., Girard, L., Minna, J.D., Gazdar, A.F., Lam, S., MacAulay, C. & Lam, W.L. (2005). High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH. *International Journal of Cancer*, **In Press**.

Girard, L., Zochbauer-Muller, S., Virmani, A.K., Gazdar, A.F. & Minna, J.D. (2000). Genome-wide allelotyping of lung cancer identifies new regions of allelic loss, differences between small cell lung cancer and non-small cell lung cancer, and loci clustering. *Cancer Res*, **60,** 4894-906.

Henderson, L.J., Coe, B.P., Lee, E.H., Girard, L., Gazdar, A.F., Minna, J.D., Lam, S., MacAulay, C. & Lam, W.L. (2005). Genomic and gene expression profiling of minute alterations of chromosome arm 1p in small-cell lung carcinoma cells. *Br J Cancer*, **92,** 1553-60.

Hyodo-Miura, J., Urushiyama, S., Nagai, S., Nishita, M., Ueno, N. & Shibuya, H. (2002). Involvement of NLK and Sox11 in neural induction in Xenopus development. *Genes Cells*, **7,** 487-96.

Ishitani, T., Ninomiya-Tsuji, J., Nagai, S., Nishita, M., Meneghini, M., Barker, N., Waterman, M., Bowerman, B., Clevers, H., Shibuya, H. & Matsumoto, K. (1999). The TAK1-NLK-MAPK-related pathway antagonizes signalling between beta-catenin and transcription factor TCF. *Nature*, **399,** 798-802.

Ishkanian, A.S., Malloff, C.A., Watson, S.K., DeLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M.A., Ling, V., MacAulay, C. & Lam, W.L. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, **36,** 299-303.

Jin, D.Y., Kozak, C.A., Pangilinan, F., Spencer, F., Green, E.D. & Jeang, K.T. (1999). Mitotic checkpoint locus MAD1L1 maps to human chromosome 7p22 and mouse chromosome 5. *Genomics*, **55,** 363-4.

Jong, K., Marchiori, E., Meijer, G., Vaart, A.V. & Ylstra, B. (2004). Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20,** 3636-7. Epub 2004 Jun 16.

Kurup, A. & Hanna, N.H. (2004). Treatment of small cell lung cancer. *Crit Rev Oncol Hematol*, **52,** 117-26.

Li, W. & Guan, K.L. (2004). The Down syndrome cell adhesion molecule (DSCAM) interacts with and activates Pak. *J Biol Chem*, **279,** 32824-31. Epub 2004 May 28.

Lundberg, A.S. & Weinberg, R.A. (1999). Control of the cell cycle and apoptosis. *Eur J Cancer*, **35,** 1886-94.

Muller, H., Lukas, J., Schneider, A., Warthoe, P., Bartek, J., Eilers, M. & Strauss, M. (1994). Cyclin D1 expression is regulated by the retinoblastoma protein. *Proc Natl Acad Sci U S A*, **91,** 2945-9.

Oguri, T., Isobe, T., Suzuki, T., Nishio, K., Fujiwara, Y., Katoh, O. & Yamakido, M. (2000). Increased expression of the MRP5 gene is associated with exposure to platinum drugs in lung cancer. *Int J Cancer*, **86,** 95-100.

Parkin, D.M., Bray, F., Ferlay, J. & Pisani, P. (2005). Global cancer statistics, 2002. *CA Cancer J Clin*, **55,** 74-108.

Petersen, I., Bujard, M., Petersen, S., Wolf, G., Goeze, A., Schwendel, A., Langreck, H., Gellert, K., Reichel, M., Just, K., du Manoir, S., Cremer, T., Dietel, M. & Ried, T. (1997a). Patterns of chromosomal imbalances in adenocarcinoma and squamous cell carcinoma of the lung. *Cancer Res*, **57,** 2331-5.

Petersen, I., Langreck, H., Wolf, G., Schwendel, A., Psille, R., Vogt, P., Reichel, M.B., Ried, T. & Dietel, M. (1997b). Small-cell lung cancer is characterized by a high incidence of deletions on chromosomes 3p, 4q, 5q, 10q, 13q and 17p. *Br J Cancer*, **75,** 79-86.

Phelps, R.M., Johnson, B.E., Ihde, D.C., Gazdar, A.F., Carbone, D.P., McClintock, P.R., Linnoila, R.I., Matthews, M.J., Bunn, P.A., Jr., Carney, D., Minna, J.D. & Mulshine, J.L. (1996). NCI-Navy Medical Oncology Branch cell line data base. *J Cell Biochem Suppl*, **24,** 32-91.

Polager, S. & Ginsberg, D. (2003). E2F mediates sustained G2 arrest and down-regulation of Stathmin and AIM-1 expression in response to genotoxic stress. *J Biol Chem*, **278,** 1443-9. Epub 2002 Nov 21.

Ramirez, R.D., Sheridan, S., Girard, L., Sato, M., Kim, Y., Pollack, J., Peyton, M., Zou, Y., Kurie, J.M., Dimaio, J.M., Milchgrub, S., Smith, A.L., Souza, R.F., Gilbey, L., Zhang, X., Gandia, K., Vaughan, M.B., Wright, W.E., Gazdar, A.F., Shay, J.W. & Minna, J.D. (2004). Immortalization of human bronchial epithelial cells in the absence of viral oncoproteins. *Cancer Res*, **64,** 9027-34.

Ritter, C.A., Jedlitschky, G., Meyer zu Schwabedissen, H., Grube, M., Kock, K. & Kroemer, H.K. (2005). Cellular export of drugs and signaling molecules by the ATP-binding cassette transporters MRP4 (ABCC4) and MRP5 (ABCC5). *Drug Metab Rev*, **37,** 253-78.

Rostad, H., Naalsund, A., Jacobsen, R., Eirik Strand, T., Scott, H., Heyerdahl Strom, E. & Norstein, J. (2004). Small cell lung cancer in Norway. Should more patients have been offered surgical therapy? *Eur J Cardiothorac Surg*, **26,** 782-6.

Rubin, C.I. & Atweh, G.F. (2004). The role of stathmin in the regulation of the cell cycle. *J Cell Biochem*, **93,** 242-50.

Sakamuro, D. & Prendergast, G.C. (1999). New Myc-interacting proteins: a second Myc network emerges. *Oncogene*, **18,** 2942-54.

Sasahira, T., Akama, Y., Fujii, K. & Kuniyasu, H. (2005). Expression of receptor for advanced glycation end products and HMGB1/amphoterin in colorectal adenomas. *Virchows Arch*, **446,** 411-5. Epub 2005 Mar 24.

Schneider, R., Bannister, A.J. & Kouzarides, T. (2002). Unsafe SETs: histone lysine methyltransferases and cancer. *Trends Biochem Sci*, **27,** 396-402.

Shaulian, E. & Karin, M. (2001). AP-1 in cell proliferation and survival. *Oncogene*, **20,** 2390-400.

Stupp, R., Monnerat, C., Turrisi, A.T., 3rd, Perry, M.C. & Leyvraz, S. (2004). Small cell lung cancer: state of the art and future perspectives. *Lung Cancer*, **45,** 105-17.

Taguchi, A., Blood, D.C., del Toro, G., Canet, A., Lee, D.C., Qu, W., Tanji, N., Lu, Y., Lalla, E., Fu, C., Hofmann, M.A., Kislinger, T., Ingram, M., Lu, A., Tanaka, H., Hori, O., Ogawa, S., Stern, D.M. & Schmidt, A.M. (2000). Blockade of RAGE-amphoterin signalling suppresses tumour growth and metastases. *Nature*, **405,** 354-60.

Tanno, S., Ohsaki, Y., Nakanishi, K., Toyoshima, E. & Kikuchi, K. (2004). Small cell lung cancer cells express EGFR and tyrosine phosphorylation of EGFR is inhibited by gefitinib ("Iressa", ZD1839). *Oncol Rep*, **12,** 1053-7.

Tomoda, R., Seto, M., Tsumuki, H., Iida, K., Yamazaki, T., Sonoda, J., Matsumine, A. & Uchida, A. (2002). Telomerase activity and human telomerase reverse transcriptase mRNA expression are correlated with clinical aggressiveness in soft tissue tumors. *Cancer*, **95,** 1127-33.

Tonon, G., Wong, K.K., Maulik, G., Brennan, C., Feng, B., Zhang, Y., Khatry, D.B., Protopopov, A., You, M.J., Aguirre, A.J., Martin, E.S., Yang, Z., Ji, H., Chin, L. & Depinho, R.A. (2005). High-resolution genomic profiles of human lung cancer. *Proc Natl Acad Sci U S A*, **102,** 9625-30.

Tsukasaki, K., Miller, C.W., Greenspun, E., Eshaghian, S., Kawabata, H., Fujimoto, T., Tomonaga, M., Sawyers, C., Said, J.W. & Koeffler, H.P. (2001). Mutations in the mitotic check point gene, MAD1L1, in human cancers. *Oncogene*, **20,** 3301-5.

Wada, T. & Penninger, J.M. (2004). Mitogen-activated protein kinases in apoptosis regulation. *Oncogene*, **23,** 2838-49.

Watson, S.K., deLeeuw, R.J., Ishkanian, A.S., Malloff, C.A. & Lam, W.L. (2004). Methods for high throughput validation of amplified fragment pools of BAC DNA for constructing high resolution CGH arrays. *BMC Genomics*, **5,** 6.

Weaver, D.A., Crawford, E.L., Warner, K.A., Elkhairi, F., Khuder, S.A. & Willey, J.C. (2005). ABCC5, ERCC2, XPA and XRCC1 transcript abundance levels correlate with cisplatin chemoresistance in non-small cell lung cancer cell lines. *Mol Cancer*, **4,** 18.

Williams, B.C., Li, Z., Liu, S., Williams, E.V., Leung, G., Yen, T.J. & Goldberg, M.L. (2003). Zwilch, a new component of the ZW10/ROD complex required for kinetochore functions. *Mol Biol Cell*, **14,** 1379-91.

Wu, S., Cetinkaya, C., Munoz-Alonso, M.J., von der Lehr, N., Bahram, F., Beuger, V., Eilers, M., Leon, J. & Larsson, L.G. (2003). Myc represses differentiation-induced p21CIP1 expression via Miz-1-dependent interaction with the p21 core promoter. *Oncogene*, **22,** 351-60.

Yamagishi, N., Saito, Y., Ishihara, K. & Hatayama, T. (2002). Enhancement of oxidative stress-induced apoptosis by Hsp105alpha in mouse embryonal F9 cells. *Eur J Biochem*, **269,** 4143-51.

Zabarovsky, E.R., Lerman, M.I. & Minna, J.D. (2002). Tumor suppressor genes on chromosome 3p involved in the pathogenesis of lung and other cancers. *Oncogene*, **21,** 6915-35.

Zakowski, M.F. (2003). Pathology of small cell carcinoma of the lung. *Semin Oncol*, **30,** 3-8.

Zebedee, Z. & Hara, E. (2001). Id proteins in cell cycle control and cellular senescence. *Oncogene*, **20,** 8317-25.

Zhao, X., Weir, B.A., LaFramboise, T., Lin, M., Beroukhim, R., Garraway, L., Beheshti, J., Lee, J.C., Naoki, K., Richards, W.G., Sugarbaker, D., Chen, F., Rubin, M.A., Janne, P.A., Girard, L., Minna, J., Christiani, D., Li, C., Sellers, W.R. & Meyerson, M. (2005). Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Res*, **65,** 5561-70.

Chapter 9: EZH2 is over-expressed in SCLC as a result of genomic deregulation of the Rb/E2F pathway.

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

Coe BP, Aviel-Ronen S, Andrea Pusic, Gazdar AF, Minna JD, Lam S, Tsao MS, Lam WL

9.1 Introduction

Small cell lung Cancer (SCLC) is a highly aggressive lung neoplasm which demonstrates very poor clinical outcome compared to other lung cancers, and has seen little improvement over the past 25 years (Lally et al., 2007). Due to its unique clinical course SCLC has a separate staging system from the standard TNM system used for most cancers including all non-small cell lung cancers (NSCLC). SCLC is classified into either limited (33%) or extensive (67%) stage disease based on the degree of tumour spread, with limited stage disease presenting in a single region of the lung with a median survival of 18 months and extensive disease spread throughout the thorax sometimes presenting with distant metastasis at diagnosis and a median survival of 9 months(Simon & Wagner, 2003; Socinski & Bogart, 2007; Weinmann et al., 2003). Due to its highly aggressive nature including rapid growth and often wide spread at the time of initial diagnosis surgery is only rarely offered and chemotherapy is the only recourse. However despite SCLC initially presenting as a chemo-sensitive disease, the majority of patients will relapse after initial treatment, and no targeted therapeutics have yet been approved for SCLC (Rossi et al., 2004; Rostad et al., 2004; Walker, 2003). Due to its highly aggressive nature, new targets for therapeutic intervention are desperately needed.

To date many studies have attempted to understand the biology of SCLC through either expression array based analysis, which has yielded insight into the disease yet the complexity of the data has hampered the application of such studies beyond clustering and disease sub-classification(Bhattacharjee et al., 2001; Jones et al., 2004; Pedersen et al., 2003; Virtanen et al., 2002). Additionally genomic data has contributed greatly to the identification of oncogenes and tumours suppressors but the low resolution of conventional techniques has limited the ability to identify specific disruptions(Girard et al., 2000; Levin et al., 1994; Ried et al., 1994).

Previously we generated high resolution array CGH profiles of SCLC cell lines and identified regions specific to the SCLC phenotype which result in expression disruption (Coe et al., 2005b;

Coe et al., 2006). In this study we report the first high resolution array CGH profiles of SCLC tumours and perform integrated analysis with our previous cell model data to identify alterations important to both clinical disease and cell models.

9.2 Results and Discussion

9.2.1 Genomic Profiling of SCLC Tumours

Due to the rarity of surgical resection in cases of SCLC, fresh frozen material is rarely available thus we have thus acquired a panel of 14, formalin fixed paraffin embedded SCLC tumours. Samples were obtained from the University Health Network and reviewed by Dr. Ming Sound-Tsao. Tissue cores were used in lieu of microdissection due to the high purity of the SCLC samples. Clinical details of the samples are summarized in **Table 9.1**.

In order to identify genomic regions involved in the tumourigenesis of SCLC we analyzed these 14 tumour samples using the SMRT CGH array. This platform allows unbiased detection of copy number alterations at 27,000 overlapping genomic loci (large insert clones) producing copy number profiles at a resolution similar to that of oligonucleotide and SNP platforms with greater number of elements but with much higher tolerance to degraded DNA samples (Coe et al., 2007).

Following co-hybridization with a universal male reference DNA, fluorescence signal ratios for each array element were determined and aligned to the genome. Regions of consistently increased or decreased signal were detected and interpreted as gains or losses of DNA respectively by application of the aCGH smooth algorithm (Jong et al., 2004).

Initial analysis of the DNA profiles for the SCLC tumours revealed the presence of many expected regions of copy number alteration such as loss of 3p and gain of 5p which are common regions of disruption in lung cancer (Balsara & Testa, 2002; Coe et al., 2005a; Yokoi et al., 2002). In order to determine which regions may be relevant to the disease we next examined these patterns of copy number alterations in the context of previously generated cell line data.

9.2.2 Comparison of Genomic Profiles for SCLC Tumours and Cell Lines

In a previous study we profiled a panel of 14 SCLC cell lines and identified specific regions of alterations (Coe et al., 2005b). Initial comparison of the tumour profiles described above with our panel of cell line profiles revealed many regions of similarity as well as difference (**Figure 9.1a**). In general the SCLC tumours tend to display fewer regions of frequent alteration compared to the cell lines and several regions demonstrate different patterns of alteration. It is likely that the reason for observing more genomic alterations in the cell lines is related to the differences in the sample populations. The cell lines mostly reflect very advanced disease, and have likely acquired alterations due to growth in culture; while the tumours reflect mostly limited disease lesions (9/14) thus we expect that patterns of alteration will differ at some regions of the genome related to cell culture specific alterations or markers of advanced disease (Jones et al., 2004; Phelps et al., 1996; Virtanen et al., 2002). A comparison of the limited stage and extensive stage tumours yielded limited differences (<2% of the genome at an uncorrected $p<0.05$ and no changes an uncorrected $p<0.01$). These changed were located mostly on chromosomes 10 and 11, and future analysis with a larger sample set may yield insight into the differences between these stages.

Due to the complex nature of cancer cell genomes, it is often complex to determine which alterations hold genes of critical importance to the disease. Traditional approaches, such as focusing on genes known to undergo homozygous deletion or high level DNA amplification have been applied as screens; however this offers only a limited view of the disease as a whole. To better understand the patterns of genomic alteration important to the specific phenotype of SCLC, we referenced our previous comparison of genomic alteration between SCLC and NSCLC cell lines to identify phenotype specific regions of copy number alteration in cell lines (Coe et al., 2006, Chapter 8 of this thesis). By analyzing our tumour data in the context of only those regions determined to be SCLC specific in the cell line study we should be able to identify those regions most relevant to aggressive phenotype in clinical disease. Analysis of the tumour genomes in the context of only these regions showed relatively good retention of copy number alteration frequencies highlighting the importance of the genes contained within (**Figure 9.1a**). In particular we observed similar alterations frequencies with the cell lines at 1p, 3q, 5q, 10q, and 18q (18q21 gains).

As a result of integrating gene expression data into our previous cell model comparison we identified specific patterns of cell cycle disruption in SCLC cell lines through DNA copy number alterations which induced activation of primarily downstream components of the cell cycle pathway including the E2F2 transcription factor, whereas NSCLC activated more upstream factors such as the epidermal growth factor receptor (EGFR) (Coe et al., 2006).

Re-analysis of this previous gene set taking into account only the genomic regions validated to be present in tumours in this study retained our previous finding of downstream signalling disruption by copy number driven events in SCLC, with retention of regions accounting for TCF4 (18q21), STMN1 (18q21) and E2F2 (1p36) (**Figure 9.1b**). In particular the preservation of E2F2 copy number gain in SCLC prompted further investigation of the Rb pathway in SCLC.

9.2.4 The E2F/Rb pathway is specifically deregulated in SCLC

Our previous observations of downstream members of cell cycle initiating pathways being specifically altered in SCLC in combination with the validation of many of these targets in tumour data has lead us to hypothesize that the primary cell cycle activation in SCLC may be occurring as far downstream as the transcription factors that regulate cell cycle progression

The retinoblastoma pathway has long been known to be a key target of deregulation in SCLC. However the majority of study has been focused on the retinoblastoma gene itself. Rb is frequently lost or mutated (90% in SCLC) and demonstrates reduced expression in the majority of cases(Cooper et al., 2006; Jones et al., 2004; Sattler & Salgia, 2003). The primary function of Rb is its normal inhibition of the E2F transcription factors. When Rb1 is phosphorylated the E2F proteins are free to function as transcription factors activating a collection of cell cycle progression factors. The E2F family of genes is divided into activating (E2F1, E2F2, E2F3) and inhibitory (E2F4, E2F5) members which physically interact with Rb1 (Caputi et al., 2005; Du & Pogoriler, 2006; Lees et al., 1993; Lomazzi et al., 2002; Sattler & Salgia, 2003; Xu et al., 1995).

Recently evidence has demonstrated the E2F genes which interact with Rb may also be primary targets of deregulation. Other studies have detected high level expression of E2F1 and E2F3 in various tumours, including reports of elevated E2F3 transcript and protein levels in SCLC (Cooper et al., 2006; Jones et al., 2004; Lee et al., 2008; Xu et al., 1995). Taking this into account with our observations of specific E2F2 over expression in SCLC we decided to further analyze this pathway in SCLC.

Analysis of copy number in cell lines (n=14) and tumours (n=14) demonstrated a pattern of copy number alterations where at least one of the activating E2Fs or RB1 was altered in every sample, this prompted a detailed analysis of the activating E2Fs and RB1 expression levels in SCLC (**Figure 9.2a**). Comparison of RB1 loss and E2F gain frequencies between limited stage and extensive stage tumours yielded no statistical significance.

Real time PCR analysis of the activating E2Fs and Rb1 in the SCLC cell lines demonstrated a striking pattern of activation. At least two activating E2F members are over expressed by 10X their normal levels in every case of SCLC, additionally Rb1 mRNA is highly reduced in most SCLC samples (**Figure 9.2b and Table 9.2**). These levels of deregulation are far greater than those observed for a panel of NSCLC cell lines suggesting that the regulation is in fact SCLC specific.

Thus it appears that the Rb pathway is deregulated not only through loss of Rb but also gains of the E2F transcription factors in SCLC. This is further supported by observation of E2F over-expression in the Rb positive line H841. This pattern is drastically different from that observed in NSCLC where receptor level or upstream pathway components such as the EGFR gene are often specifically deregulated. This result suggests that the Rb/E2F pathway is activated in 100% of SCLC cell lines, by either loss of Rb or gain of an E2F member, and likely similarly deregulated in tumours.

9.2.5 Over-expression of EZH2

The striking pattern of Rb/E2F deregulation in the SCLC cell lines and retained copy number events in the panel of SCLC tumours profiled in this study prompted us to examine genes downstream of the E2F transcription factors to confirm if the pathway is in fact hyperactive in SCLC.

One target of the E2F/Rb pathway which has recently been described in multiple cancer types is the EZH2 gene. EZH2 is a polycomb group (PcG) gene with a role in embryonic development and differentiation through the epigenetic regulation of the expression of various downstream genes(Grimaud et al., 2006; Kamminga et al., 2006; Vire et al., 2006). It directly controls DNA methylation for several target genes including WNT1, matching well with our previous results seeing multiple hits shutting down the WNT pathway in SCLC cell lines(Coe et al., 2006; Vire et al., 2006).

The additional detection of STMN1 as a 1p SCLC specific oncogene in SCLC highlighted the EZH2 oncogene which is known to function through STMN1, and its expression is directly controlled by the E2F pathways(Chen et al., 2007; Coe et al., 2006).

Expression analysis of EZH2 in NSCLC and SCLC cell lines detected a striking state of hyper-activation in SCLC cells (**Figure 9.3a**). Our results suggest that EZH2 is on average 42 fold over-expressed in SCLC lines compared to only 13 fold over-expression in NSCLC cell lines.

To confirm if the over-expression of EZH2 is also present in tumours we analyzed the data from an independent cDNA expression array study (Jones MH et al.) which profiled a separate set of

9 cell lines and a panel of 15 primary SCLC tumours, in addition to 12 adenocarcinoma samples (NSCLC subtype). Although the exact fold change levels were not directly equivalent (perhaps due to differences between RT-PCR and cDNA microarray dynamic ranges), the trend in expression levels is strikingly similarity to our study with significantly higher expression of EZH2 in SCLC cell lines and tumours compared to NSCLC samples (**Figure 9.3.b**).

Bracken et al. suggest that both DNA amplification and Rb disruption can lead to disruption in EZH2 expression. The significant expression described here is unlikely due to the low level copy number gains observed in the SCLC cell lines and tumours as previous studies have suggested that a 1.5 fold copy number gain only leads to 2 fold changes in mRNA levels for EZH2. Thus high level amplification would be required to lead to such high expression levels. Also of note is the significantly lower expression of EZH2 in NSLC which demonstrate lower frequency of Rb disruption (15%) and far lower levels of the activating E2Fs.

These results suggest that in SCLC, over-expression of EZH2 is strictly controlled by E2F/Rb disruption. Additionally, this suggests that the levels of EZH2 expression are much higher than those of NSCLC, due to the strong activation of E2F/Rb in SCLC.

The PcG has been of great interest in many tumour types recently due to its role in chromatin remodelling and direct control of specific genes by direct DNA methylation. Studies have detected expression specific to metastatic prostate cancer suggesting a role in aggressive behaviour (Yu et al., 2007). Additionally over-expression has been observed in breast, bladder, squamous cell lung cancer and hepatocellular carcinomas (Arisan et al., 2005; Bracken et al., 2003; Breuer et al., 2004; Chen et al., 2007; Hinz et al., 2007). In the case of squamous cell lung cancer expression has been seen in dysplastic squamous cells and tumours but not in normal bronchial epithelial cells, suggesting EZH2 could be an early event in NSCLC (Breuer et al., 2004).

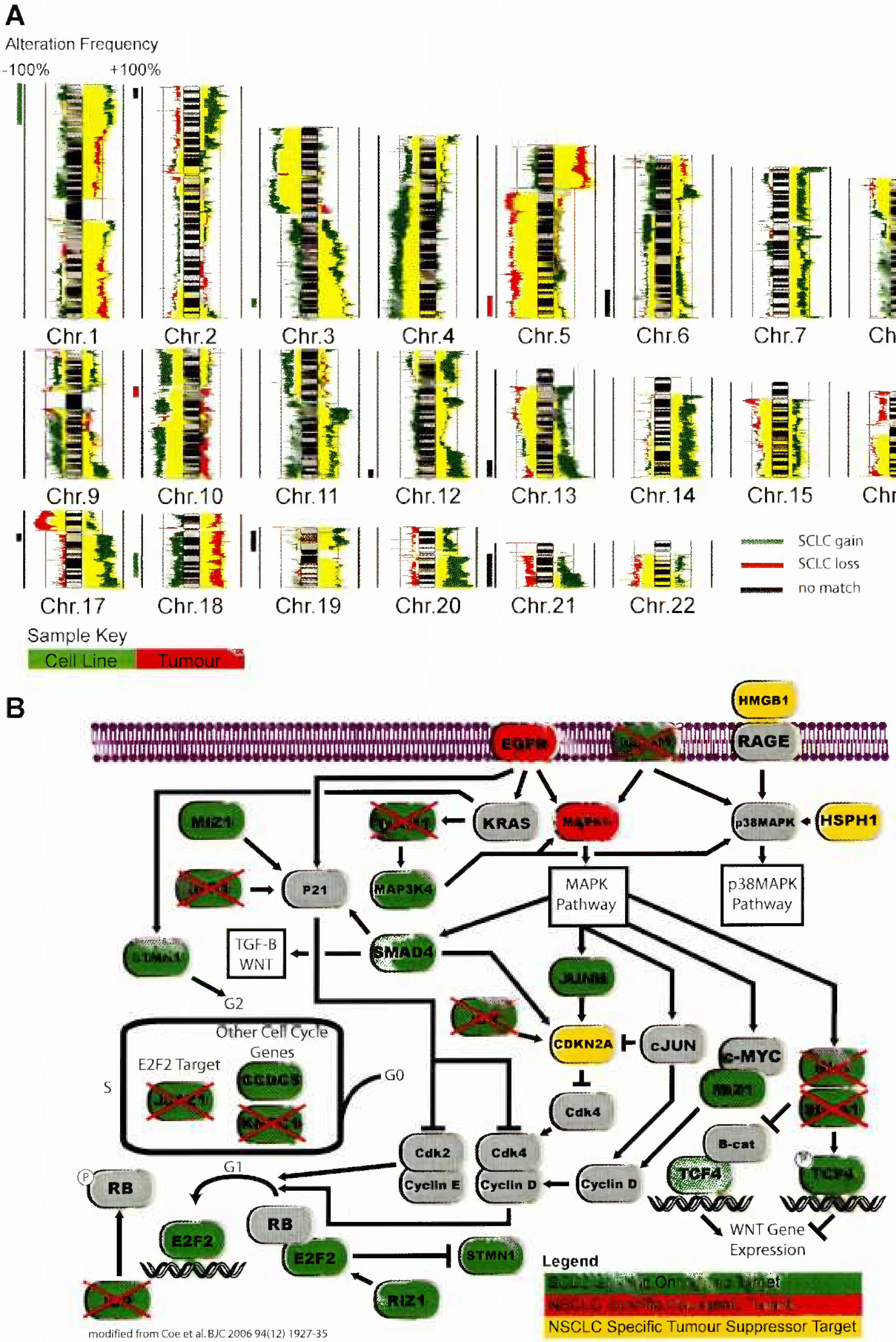
9.4.3 Expression Analysis.

Real Time PCR was performed as previously described using TaqMan gene expression assays with standard protocols on an ABI 7500 Fast thermo =cycler (Applied Biosystems, Foster City, CA, USA). The TaqMan assays used were E2F1 (Hs00153451_m1), E2F2 (Hs00231667_m1), E2F3 (Hs00605457_m1), Rb (Hs00153108_m1), EZH2 (Hs00172783_m1), 18S RNA (HS99999901_s1). Absolute expression values were calculated for E2F1,E2F2,E2F3 and Rb by scaling the delta Ct values (gene-18s) to a value between 0 and 1000.

Validation of EZH2 expression in the independent data set by Jones et al. was performed on data downloaded from the Gene Expression Omnibus (GSE 1037), expression values for the panel of normal lung samples were averaged and used for generating fold change values.

Figure 9.1. Comparison of SCLC Cell Lines and Tumours. (**a**) Comparison of SCLC tumour and cell lines genomes. Alteration frequencies for SCLC tumours (red) and cell lines (green) are displayed as bar plots adjacent to chromosomal ideograms. Bars extending to the right and left of the chromosome represent regions of gain and loss respectively, with yellow representing regions of overlap. Vertical bars on the left of the frequency diagrams represent SCLC specific regions identified in a previous cell line study, with green and red shading representing SCLC specific loss and gain. Grey bars represent regions not retained in the tumour samples. (**b**) Retention of SCLC specific pathway perturbation. Displayed is a modified version of the Figure 8.5 (Coe et al., 2006). Genes specifically over-expressed in SCLC cell lines due to copy number alterations are displayed in green, while those not retained in tumour are indicated by a red X.

Figure 9.1



modified from Coe et al. BJC 2006 94(12) 1927-35

Figure 9.2

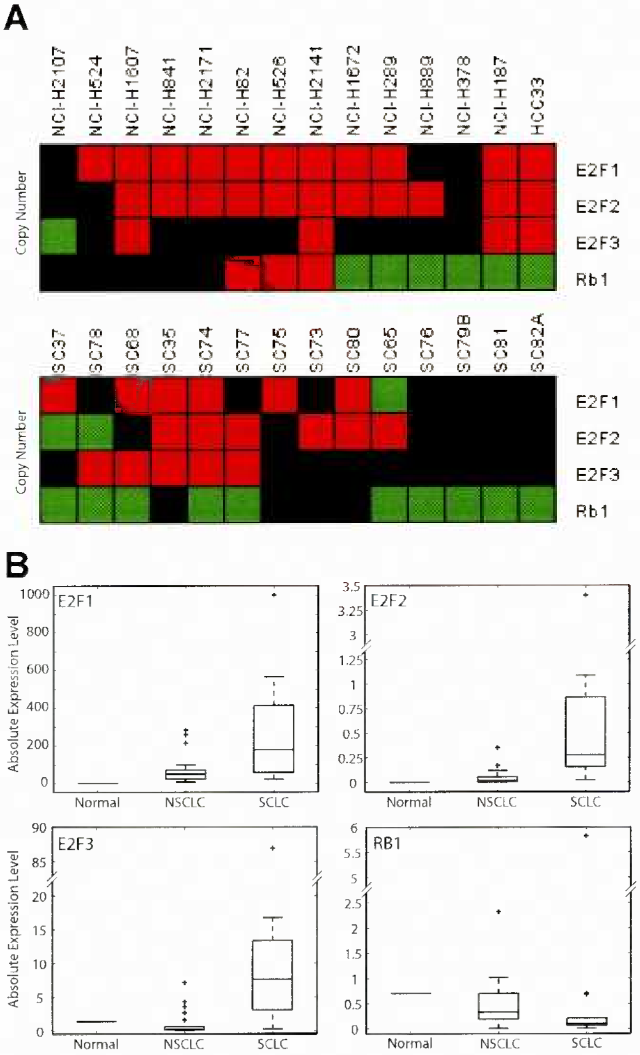


Figure 9.2. Deregulation of the E2F transcription factors. (**a**) Copy number alterations of specific E2F/Rb pathway members in SCLC cell lines (top) and tumours (bottom). Green shading represents loss, while Red represents gain and Black reprents no change. (**b**) Expression of E2F/Rb pathway members by RT-PCR. Data are presented as box-plots of absolute expression levels derived from scale normalized PCR data. The centre line in each box represents the median level while the box represents the interquartile range, with whiskers extending to the last non-outlier data point (defined as 1.5x the interquartile range). Outliers are represented as crosses.

Figure 9.3

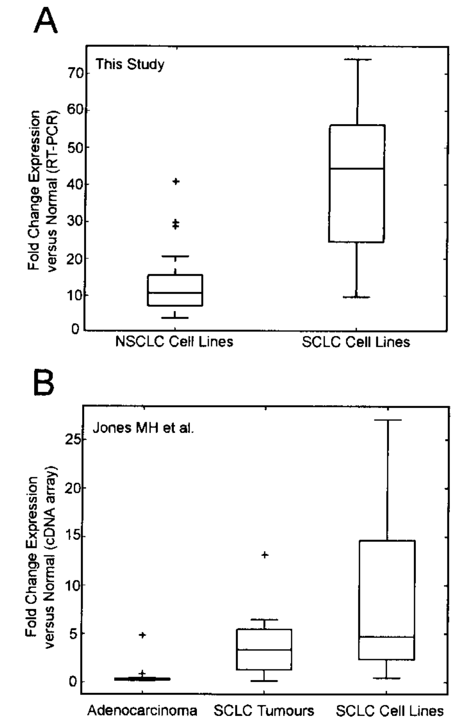


Figure 9.3. Hyper-activation of EZH2 in SCLC. (**a**) Significant overexpression of EZH2 was observed in SCLC cell lines compared to NSCLC cell lines, coinciding with excess activation of the E2F/Rb pathway. (**b**) Similar trends of SCLC hyperactication of EZH2 were also observed in an independent data set consisting of SCLC tumours and cell lines with adenocarcinma representing NSCLC (Jones et al., 2004).

Table 9.1. Clinical Data for SCLC tumours in this study.

ID	Diagnosis	Sex	Age	Limited vs. Extensive Disease (metastisis)	Comment
sc-35	SCLC	M	59	E (other)	
sc-37	SCLC	F	63	E (bone)	
sc-65	SCLC	F	74	L	
sc-68	SCLC	M	69	L	
sc-73	SCLC	F	42	L	The sample is after 2nd course of Chemotherapy. History of other neoplasms.
sc-74	SCLC	M	77	L	
sc-75	SCLC	F	78	L	RLL wedge resection SQCC + SCLC. Sample just from SCLC. History of other neoplasms
sc-76	SCLC	M	70	E	History of other neoplasms
sc-77	SCLC	F	62	L	
sc-78	SCLC	F	75	L	
sc-79B	SCLC (combined tumor)	F	57	E	

sc-80	SCLC	F	64	L	
sc-81	SCLC	F	71	E	
sc-82A	SCLC (combined tumor)	M	80	L	

Table 9.2. Real Time PCR Summary

		Absolute Expression (0-1000)				Relative Expresion (Sample/Normal Lung)				
Sample	RB1 Protein[1]	E2F1	E2F2	E2F3	RB1	E2F1	E2F2	E2F3	RB1	EZH2
Normal		1.223	0.001	1.405	0.703	1.000	1.000	1.000	1.000	1.000
H187	-ve	508.740	0.285	1.405	0.002	415.873	445.722	1.000	0.003	50.4769
H841	+ve	57.313	0.064	0.306	0.703	46.851	100.427	0.218	1.000	14.9249
H378	nd	179.867	0.266	12.913	0.001	147.033	415.873	9.190	0.002	73.8973
H1607	-ve	173.740	0.306	86.870	5.819	142.025	477.713	61.820	8.282	51.86
H889	-ve	34.674	0.156	3.906	0.001	28.345	243.032	2.780	0.001	33.6912
H289	-ve	83.911	0.865	3.118	0.079	68.594	1351.176	2.219	0.113	53.5346
HCC33	nd	37.163	0.236	4.487	0.106	30.379	368.367	3.193	0.151	56.0969
H2171	-ve	20.978	0.023	2.715	0.085	17.148	35.506	1.932	0.121	34.964
H82	-ve	138.696	0.220	15.625	0.093	113.378	343.699	11.119	0.132	9.66621
H2141	-ve	413.225	0.960	13.369	0.209	337.794	1499.224	9.514	0.297	23.6907
H1672	-ve	564.482	0.069	8.229	0.143	461.440	107.635	5.856	0.203	69.72
H526	-ve	385.553	0.753	10.489	0.679	315.173	1176.267	7.464	0.966	38.2119
H524	-ve	1000.000	3.401	16.746	0.059	817.458	5311.855	11.917	0.084	24.5576
H2107	nd	366.021	1.084	7.041	0.150	299.207	1692.570	5.011	0.214	63.2788

[1] Phelps et al. Journal of Cellular Biochemistry Supplement 24:32-91 (1996)

9.5 References

Arisan, S., Buyuktuncer, E.D., Palavan-Unsal, N., Caskurlu, T., Cakir, O.O. & Ergenekon, E. (2005). Increased expression of EZH2, a polycomb group protein, in bladder carcinoma. *Urol Int*, **75,** 252-7.

Balsara, B.R. & Testa, J.R. (2002). Chromosomal imbalances in human lung cancer. *Oncogene*, **21,** 6877-83.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J. & Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*, **98,** 13790-5.

Bondzi, C., Litz, J., Dent, P. & Krystal, G.W. (2000). Src family kinase activity is required for Kit-mediated mitogen-activated protein (MAP) kinase activation, however loss of functional retinoblastoma protein makes MAP kinase activation unnecessary for growth of small cell lung cancer cells. *Cell Growth Differ*, **11,** 305-14.

Bracken, A.P., Pasini, D., Capra, M., Prosperini, E., Colli, E. & Helin, K. (2003). EZH2 is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer. *Embo J*, **22,** 5323-35.

Breuer, R.H., Snijders, P.J., Smit, E.F., Sutedja, T.G., Sewalt, R.G., Otte, A.P., van Kemenade, F.J., Postmus, P.E., Meijer, C.J. & Raaphorst, F.M. (2004). Increased expression of the EZH2 polycomb group gene in BMI-1-positive neoplastic cells during bronchial carcinogenesis. *Neoplasia*, **6,** 736-43.

Caputi, M., Russo, G., Esposito, V., Mancini, A. & Giordano, A. (2005). Role of cell-cycle regulators in lung cancer. *J Cell Physiol*, **205,** 319-27.

Chen, Y., Lin, M.C., Yao, H., Wang, H., Zhang, A.Q., Yu, J., Hui, C.K., Lau, G.K., He, M.L., Sung, J. & Kung, H.F. (2007). Lentivirus-mediated RNA interference targeting enhancer of zeste homolog 2 inhibits hepatocellular carcinoma growth through down-regulation of stathmin. *Hepatology*, **46,** 200-8.

Coe, B.P., Henderson, L.J., Garnis, C., Tsao, M.S., Gazdar, A.F., Minna, J., Lam, S., Macaulay, C. & Lam, W.L. (2005a). High-resolution chromosome arm 5p array CGH analysis of small cell lung carcinoma cell lines. *Genes Chromosomes Cancer*, **42,** 308-13.

Coe, B.P., Lee, H.L., Chi, B., Girard, L., Minna, J.D., Gazdar, A.F., Lam, S., MacAulay, C. & Lam, W.L. (2005b). Gain of a region on 7p22.3, containing MAD1L1, is the Most Frequent Event in Small Cell Lung Cancer Cell Lines. *Genes Chromosomes Cancer*, **In Press**.

Coe, B.P., Lockwood, W.W., Girard, L., Chari, R., Macaulay, C., Lam, S., Gazdar, A.F., Minna, J.D. & Lam, W.L. (2006). Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. *Br J Cancer*, **94,** 1927-35.

Coe, B.P., Ylstra, B., Carvalho, B., Meijer, G.A., Macaulay, C. & Lam, W.L. (2007). Resolving the resolution of array CGH. *Genomics*, **89,** 647-53.

Cooper, C.S., Nicholson, A.G., Foster, C., Dodson, A., Edwards, S., Fletcher, A., Roe, T., Clark, J., Joshi, A., Norman, A., Feber, A., Lin, D., Gao, Y., Shipley, J. & Cheng, S.J. (2006). Nuclear overexpression of the E2F3 transcription factor in human lung cancer. *Lung Cancer*, **54,** 155-62.

Du, W. & Pogoriler, J. (2006). Retinoblastoma family genes. *Oncogene*, **25,** 5190-200.

Girard, L., Zochbauer-Muller, S., Virmani, A.K., Gazdar, A.F. & Minna, J.D. (2000). Genome-wide allelotyping of lung cancer identifies new regions of allelic loss, differences between small cell lung cancer and non-small cell lung cancer, and loci clustering. *Cancer Res*, **60,** 4894-906.

Grimaud, C., Negre, N. & Cavalli, G. (2006). From genetics to epigenetics: the tale of Polycomb group and trithorax group genes. *Chromosome Res*, **14,** 363-75.

Hinz, S., Kempkensteffen, C., Christoph, F., Hoffmann, M., Krause, H., Schrader, M., Schostak, M., Miller, K. & Weikert, S. (2007). Expression of the polycomb group protein EZH2 and its relation to outcome in patients with urothelial carcinoma of the bladder. *J Cancer Res Clin Oncol.*

Ishkanian, A.S., Malloff, C.A., Watson, S.K., DeLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M.A., Ling, V., MacAulay, C. & Lam, W.L. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, **36,** 299-303.

Jones, M.H., Virtanen, C., Honjoh, D., Miyoshi, T., Satoh, Y., Okumura, S., Nakagawa, K., Nomura, H. & Ishikawa, Y. (2004). Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet*, **363,** 775-81.

Jong, K., Marchiori, E., Meijer, G., Vaart, A.V. & Ylstra, B. (2004). Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20,** 3636-7.

Kamminga, L.M., Bystrykh, L.V., de Boer, A., Houwer, S., Douma, J., Weersing, E., Dontje, B. & de Haan, G. (2006). The Polycomb group gene Ezh2 prevents hematopoietic stem cell exhaustion. *Blood*, **107,** 2170-9.

Lally, B.E., Urbanic, J.J., Blackstock, A.W., Miller, A.A. & Perry, M.C. (2007). Small cell lung cancer: have we made any progress over the last 25 years? *Oncologist*, **12,** 1096-104.

Lee, J., Park, C.K., Park, J.O., Lim, T., Park, Y.S., Lim, H.Y., Lee, I., Sohn, T.S., Noh, J.H., Heo, J.S., Kim, S., Lim do, H., Kim, K.M. & Kang, W.K. (2008). Impact of E2F-1 Expression on Clinical Outcome of Gastric Adenocarcinoma Patients with Adjuvant Chemoradiation Therapy. *Clin Cancer Res*, **14,** 82-8.

Lees, J.A., Saito, M., Vidal, M., Valentine, M., Look, T., Harlow, E., Dyson, N. & Helin, K. (1993). The retinoblastoma protein binds to a family of E2F transcription factors. *Mol Cell Biol*, **13,** 7813-25.

Levin, N.A., Brzoska, P., Gupta, N., Minna, J.D., Gray, J.W. & Christman, M.F. (1994). Identification of frequent novel genetic alterations in small cell lung carcinoma. *Cancer Res*, **54,** 5086-91.

Lomazzi, M., Moroni, M.C., Jensen, M.R., Frittoli, E. & Helin, K. (2002). Suppression of the p53- or pRB-mediated G1 checkpoint is required for E2F-induced S-phase entry. *Nat Genet*, **31,** 190-4.

Ma, P.C., Tretiakova, M.S., Nallasura, V., Jagadeeswaran, R., Husain, A.N. & Salgia, R. (2007). Downstream signalling and specific inhibition of c-MET/HGF pathway in small cell lung cancer: implications for tumour invasion. *Br J Cancer*, **97,** 368-77.

Pedersen, N., Mortensen, S., Sorensen, S.B., Pedersen, M.W., Rieneck, K., Bovin, L.F. & Poulsen, H.S. (2003). Transcriptional gene expression profiling of small cell lung cancer cells. *Cancer Res*, **63,** 1943-53.

Phelps, R.M., Johnson, B.E., Ihde, D.C., Gazdar, A.F., Carbone, D.P., McClintock, P.R., Linnoila, R.I., Matthews, M.J., Bunn, P.A., Jr., Carney, D., Minna, J.D. & Mulshine, J.L. (1996). NCI-Navy Medical Oncology Branch cell line data base. *J Cell Biochem Suppl*, **24,** 32-91.

Ried, T., Petersen, I., Holtgreve-Grez, H., Speicher, M.R., Schrock, E., du Manoir, S. & Cremer, T. (1994). Mapping of multiple DNA gains and losses in primary small cell lung carcinomas by comparative genomic hybridization. *Cancer Res*, **54,** 1801-6.

Rossi, A., Maione, P., Colantuoni, G., Guerriero, C. & Gridelli, C. (2004). The role of new targeted therapies in small-cell lung cancer. *Crit Rev Oncol Hematol*, **51,** 45-53.

Rostad, H., Naalsund, A., Jacobsen, R., Strand, T.E., Scott, H., Heyerdahl Strom, E. & Norstein, J. (2004). Small cell lung cancer in Norway. Should more patients have been offered surgical therapy? *Eur J Cardiothorac Surg*, **26,** 782-6.

Rozengurt, E. (1999). Autocrine loops, signal transduction, and cell cycle abnormalities in the molecular biology of lung cancer. *Curr Opin Oncol*, **11,** 116-22.

Sattler, M. & Salgia, R. (2003). Molecular and cellular biology of small cell lung cancer. *Semin Oncol*, **30,** 57-71.

Shan, L., Aster, J.C., Sklar, J. & Sunday, M.E. (2007). Notch-1 regulates pulmonary neuroendocrine cell differentiation in cell lines and in transgenic mice. *Am J Physiol Lung Cell Mol Physiol*, **292,** L500-9.

Simon, G.R. & Wagner, H. (2003). Small cell lung cancer. *Chest*, **123,** 259S-271S.

Socinski, M.A. & Bogart, J.A. (2007). Limited-stage small-cell lung cancer: the current status of combined-modality therapy. *J Clin Oncol*, **25,** 4137-45.

Vestergaard, J., Pedersen, M.W., Pedersen, N., Ensinger, C., Tumer, Z., Tommerup, N., Poulsen, H.S. & Larsen, L.A. (2006). Hedgehog signaling in small-cell lung cancer: frequent in vivo but a rare event in vitro. *Lung Cancer*, **52,** 281-90.

Vire, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.M., Bollen, M., Esteller, M., Di Croce, L., de Launoit, Y. & Fuks, F. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. *Nature*, **439,** 871-4.

Virtanen, C., Ishikawa, Y., Honjoh, D., Kimura, M., Shimane, M., Miyoshi, T., Nomura, H. & Jones, M.H. (2002). Integrated classification of lung tumors and cell lines by expression profiling. *Proc Natl Acad Sci U S A*, **99,** 12357-62.

Walker, S. (2003). Updates in small cell lung cancer treatment. *Clin J Oncol Nurs*, **7,** 563-8.

Weinmann, M., Jeremic, B., Bamberg, M. & Bokemeyer, C. (2003). Treatment of lung cancer in elderly part II: small cell lung cancer. *Lung Cancer*, **40,** 1-16.

Xu, G., Livingston, D.M. & Krek, W. (1995). Multiple members of the E2F transcription factor family are the products of oncogenes. *Proc Natl Acad Sci U S A*, **92,** 1357-61.

Yokoi, S., Yasui, K., Saito-Ohara, F., Koshikawa, K., Iizasa, T., Fujisawa, T., Terasaki, T., Horii, A., Takahashi, T., Hirohashi, S. & Inazawa, J. (2002). A novel target gene, SKP2, within the 5p13 amplicon that is frequently detected in small cell lung cancers. *Am J Pathol*, **161,** 207-16.

Yu, J., Yu, J., Rhodes, D.R., Tomlins, S.A., Cao, X., Chen, G., Mehra, R., Wang, X., Ghosh, D., Shah, R.B., Varambally, S., Pienta, K.J. & Chinnaiyan, A.M. (2007). A polycomb repression signature in metastatic prostate cancer predicts cancer outcome. *Cancer Res*, **67,** 10657-63.

Chapter 10: Conclusions

Portions of this chapter are excerpts from the abstracts of the manuscripts detailed in chapters 2 to 9.

10.1 Summary

SCLC is a difficult disease. Due to its highly aggressive clinical course surgery is rarely curative so chemotherapy is often the only treatment option. Initially patients respond well to chemotherapy but relapse with chemoinsensitive disease is inevitable. Due to the rarity of surgical resection fresh frozen tumour material is very difficult to acquire and most tissue specimens will be limited to FFPE material for which expression analysis is not reliable.

Additionally there is little knowledge of the development of SCLC so the source cell is not precisely known thus pure expression based analysis may detect markers associated with cell type and the tumour itself. As cancer is a disease which involves genomic alterations, and many oncogenic expression changes are due to altered gene dosage, copy number based tools are critical for the understanding of SCLC, both for use in combination with expression in analysis of cell models, and the analysis of clinical specimens. For this reason the initial chapters of this thesis describe the technical development of array comparative genomic hybridization, a new technique which allows unprecedented detail in the analysis of aneuploidy in cancer genomes.

10.1.1 Technical Development of array CGH

Chapter 2 details the construction of a tiling resolution array consisting of 32,433 overlapping BAC clones covering the entire human genome. This array represents a drastic increase in our ability to identify genetic alterations and their boundaries throughout the genome in a single comparative genomic hybridization (CGH) experiment. Through profiling of cancer samples with this platform we identified minute DNA alterations which had escaped previous detection. These alterations include microamplifications and deletions containing known oncogenes and tumour-suppressor genes, as well as novel genes which may be associated with various tumours, demonstrating the need to move beyond conventional marker-based genome analysis techniques which infer status between measured loci. This submegabase resolution tiling set

(SMRT) array CGH platform allows comprehensive assessment of genomic alterations at a level never before possible (Ishkanian et al., 2004).

Once we constructed the SMRT array it became apparent that new tools would be necessary to allow us to visualize and interpret the data. Since array CGH provides copy number data for tens of thousands of DNA segments, optimal visualization requires the reassembly of individual data points into karyogram style chromosome profiles. Thus, in Chapter 3 we developed a visualization tool for displaying whole genome array CGH data in the context of chromosomal location. SeeGH generates high resolution chromosome profiles from standard array ratio data files, data is then displayed in a high resolution display representative of conventional CGH karyotype diagrams with the ability to zoom in on regions of interest and view annotation information such as gene mapping. To generate these diagrams SeeGH imports the data into a database, calculates the average ratio and standard deviation for each replicate spot, and links them to chromosome regions. Once the data is displayed, users have the option of filtering data based on user defined QC criteria, and retrieve annotation information such as clone name, NCBI sequence accession number, ratio, base pair position on the chromosome, and standard deviation. This represents a novel software tool used to view and analyze array CGH data (Chi et al., 2004). The software gives users the ability to view the data in an overall genomic view as well as magnify specific chromosomal regions facilitating the precise localization of genetic alterations. This software was later expanded in Chapter 8 to include plotting of alteration frequency data (Coe et al., 2006a).

Another significant problem that needed to be addressed prior to analysis of cancer specimens was that of sample purity. Tumour biopsies are typically small and contain infiltrating stromal cells, requiring tedious microdissection. This tissue heterogeneity is a major barrier to high-throughput profiling of tumour genomes and is also an important consideration for the introduction of array CGH to clinical settings. In Chapter 4 we demonstrate that increasing array resolution enhances detection sensitivity in mixed tissues and as a result significantly reduces microdissection requirements. In this study, we first simulated normal cell contamination to determine the heterogeneity tolerance of array CGH and then validated this detection sensitivity model on cancer specimens using the newly developed submegabase resolution tiling-set (SMRT) array. As a result we determined that normal cell levels as high as 75% can be tolerated in detection of large scale alterations, while sensitive detection of small alterations is still possible in samples with ~50% purity. (Garnis et al., 2005)

During the progress of this project many technologies have been designed to supplant conventional metaphase CGH technology with the goal of refining the description of segmental copy number status throughout the genome. However, the emergence of new technologies has led to confusion as to how to adequately describe the capabilities of each array platform. The design of a CGH array can incorporate a uniform or a highly variable element distribution. This can lead to bias in the reporting of average or median resolutions, making it difficult to provide a fair comparison of platforms. In Chapter 5, we propose a new definition of resolution for array CGH technology, termed "functional resolution," that incorporates the uniformity of element spacing on the array, as well as the sensitivity of each platform to single-copy alterations. Calculation of these metrics is automated through the development of a Java-based application, "ResCalc," which is applicable to any array CGH platform. As a result of this study we determined that due to its uniform coverage, the SMRT array generates array data at a resolution competitive with newer oligonucleotide platforms with many times more elements. Additionally the SMRT array accomplishes this resolution with a DNA requirement 5x less than most oligonucleotide platforms. (Coe et al., 2007)

10.1.2 Profiling of SCLC

Prior to completing the whole genome SMRT array we performed initial studies of SCLC genomes using an array covering chromosome 5p. Genomic amplification of regions on chromosome arm 5p has been observed frequently in small cell lung cancer (SCLC), implying the presence of multiple oncogenes on this arm. Thus to identify candidate genes on this chromosome arm, we developed a high-resolution, 10-clone-per-megabase bacterial artificial chromosome CGH array for 5p and examined a panel of 15 SCLC cell lines in chapter 6. Utilization of this CGH array has allowed the fine-mapping of breakpoints to regions as small as 200 kb in a single experiment. In addition to reporting our observations of aberrations at the well-characterized SKP2 and TERT loci, we describe the identification of microdeletions that have escaped detection by conventional screens and the identification TRIO and ANKH as novel putative oncogenes (Coe et al., 2005). In addition to SCLC, TRIO has been identified in multiple cancer tissue including bladder and oral cancers and soft tissue sarcomas (Adamowicz et al., 2006; Baldwin et al., 2005; Zheng et al., 2004), highlighting the potential importance of this gene.

After completion of the initial study detailed in chapter 6 we completed the SMRT array (Chapter 2). This enabled us to profile segmental DNA copy number gains and losses across the entire genome at a resolution 100 times that of conventional methods. In chapter 7, we report the analysis of 14 SCLC cell lines and six matched normal B-lymphocyte lines. We detected 7p22.3 copy number gain in 13 of the 14 SCLC lines and 0 of the 6 matched normal lines. In 4 of the 14 cell lines, this gain is present as a 350 kbp gene specific copy number gain centered at MAD1L1 (the human homologue of the yeast gene MAD1). Fluorescence in situ hybridization validated the array CGH finding. Intriguingly, MAD1L1 has been implicated to have tumour-suppressing functions. Our data suggest a more complex role for this gene, as MAD1L1 is the most frequent copy number gain in SCLC cell lines. (Coe et al., 2006a)

Lung cancer is comprised of two major cell types: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). Although these cell types can be distinguished readily at the histological level, knowledge of their underlying molecular differences is very limited. Thus knowledge of the underlying molecular differences between these subtypes may yield insight in to the aggressive nature of SCLC. Thus in chapter 8, we 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. Using stringent criteria, we have 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. Our 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. Small cell lung cancer exhibited increased expression of MRP5, activation of Wnt pathway inhibitors, and up-regulation of p38 MAPK activating genes, while NSCLC showed down-regulation of CDKN2A, and up-regulation of MAPK9 and EGFR. This information suggests that cell cycle up-regulation in SCLC and NSCLC occurs through drastically different mechanisms, highlighting the need for differential molecular target selection in the treatment of these cancers. (Coe et al., 2006b)

After the identification of SCLC specific alterations we profiled a panel of 14 SCLC tumours to identify regions of genomic alteration which are retained in clinical disease (Chapter 9). In general tumour samples displayed less complex genomic profiles compared to cell lines however striking similarities were observed for several of the regions observed in chapter 8. Investigation of these regions highlighted a striking pattern of copy number alterations affecting the E2F transcription factors which interact with Rb. Expression profiling further highlighted the E2F/Rb pathway as all of the E2F transcription factors were significantly over-expressed in all SCLC cell lines examine including the Rb positive line H841. This suggests that the Rb pathway is disrupted not only through disruption of the Rb tumour suppressor but also deregulation of the E2F transcription factors. Examination of downstream genes highlighted EZH2, a polycomb group gene involved in many cellular function including escape from senescence in hematopoietic stem cells(Grimaud et al., 2006; Kamminga et al., 2006; Vire et al., 2006), and has been identified in several cancer types including metastatic prostate cancer, and early stage NSCLC. EZH2 is known to function through STMN1, which was identified in Chapter 9 as a SCLC specific oncogene (Chen et al., 2007; Coe et al., 2006b). Expression analysis identified hyper-activated in SCLC demonstrating 43 fold over-expression, compared to 13 fold over-expression in NSCLC. This suggests that EZH2 may be a potential novel therapeutic target in SCLC.

10.2 Conclusions

We have demonstrated that array CGH is a powerful tool to dissect the genomes of cancer specimens, allowing unprecedented detail in discovering minute alterations which have escaped previous detection and a better understanding of cancer genomics. During the course of this thesis newer array CGH technologies have been developed by multiple laboratories and companies, however the CGH array technology developed in this thesis remains competitive and retains benefits over oligonucleotide based platforms in the ability to profile samples with significantly limited material.

Application of this technology in combination with gene expression data has provided numerous insights into the biology of SCLC. Through analysis of SCLC cell lines with genomic arrays we identified multiple target genes which may be relevant to tumourigenesis in SCLC and many other cancer types, validating hypothesis 1 of this thesis; however the greatest benefit was gained through the comparison of SCLC to other less aggressive lung tumours. Comparison of these data sets yields significant complexity reduction, by highlighting disease specific patterns which would have otherwise escaped detection. Detection of phenotype specific copy number alterations (PSCNA) suggests that SCLC regulates multiple targets by copy number that may partially explain its phenotype, validating hypothesis 2 and 3 of this thesis. For example we identified amplification of the ABCC5 transporter which is involved in resistance to chemotherapy in multiple samples. Strikingly we observed that downstream hits to the cell cycle controlling pathways are characteristic of SCLC. Although the original analysis in chapter

8 was based on literature resources, reanalysis of the data using Ingenuity Pathway Analysis software (www.ingenuity.com) determined that cell cycle is in fact statistically overrepresented in our gene list ($p < 0.05$). This p-value is calculated by determining how many genes would be expected to associate with cell cycle pathways if we selected an equal sized set of random genes form the Affymetrix array platform and comparing the expected frequency to that observed in our gene list. Demonstrating that the methodology developed does indeed identify oncogenes and tumour suppressor genes involved in the aggressiveness of SCLC (further validating hypothesis 2 and 3). As discussed by Bondzi et al. such hits in the case of Rb can bypass the need for upstream hits (Bondzi et al., 2000), perhaps explaining the discovery of multiple mitogenic pathways in the SCLC literature. As activation of downstream pathway members may allow activation by multiple upstream pathways to induce growth. In particular we observed deregulation of the Rb/E2F pathways through not only loss of Rb but also gain of the E2F transcription factors. Analysis of the effects of this regulation identified the E2F target gene EZH2, which is not significantly affected by copy number alterations in our data set. The ability of this gene to control escape from senescence and lead to a metastatic phenotype fits very well with the biology of SCLC, and the observed co-activation of a co-operating gene STMN1 further supports a role for this gene in SCLC (Chen et al., 2007; Coe et al., 2006b; Kamminga et al., 2006; Vire et al., 2006).

Taken together these data suggest that all SCLC cases may be disrupting mitogenic pathways at downstream nodes. This raises an important hypothesis, that if downstream regulation of a pathway occurs, perhaps therapeutic intervention should be best applied to the most downstream target possible. For example EGFR inhibitors have demonstrate little role in SCLC which activates cell cycle much further downstream. This demonstrates that targeting a single mitogenic signalling pathway may be inefficient if deregulation of additional targets can bypass the effects. Thus it is likely that strategies such as attempting to knock down the most frequently amplified gene in a particular cancer may be a naive approach. Although frequent gene disruption events may serve as excellent diagnostic markers for tumour sub-classification and clinical stratification, it is likely that the design of effective targeted therapeutics may require a much more complete understanding of pathway dynamics.

10.3 Future Directions

The immediate future plans to expand this work to include validation of EZH2 protein levels in clinical tumours through immunohistochemistry. If the protein levels are indeed elevated then I hope to pursue a lentivirus based knockdown of EZH2 in cell models of SCLC, using a strategy similar to that of Chen et al. who demonstrated significant growth disruption of hepatocellular carcinoma through lentiviral knockdown (Chen et al., 2007).

Longer term work would focus on examining the hypothesis that downstream cell cycle targets can explain the aggressive nature of the small cell phenotype. Small cell tumours are found in multiple tissue types, albeit at a far lower frequency that the lung (Frazier et al., 2007). It is likely that comparison of SCLC to the other small cell cancers may allow further refinement of the small cell phenotype specific genes by cancelling out tissue specific patterns of alteration. This could provide two benefits, firstly a better understanding of the small cell phenotype, and secondly it would allow us to determine how similar the genomes of various small cell cancers are, and whether unique therapeutic strategies may be required for each tissue site.

A similar approach would be to identify similarities and differences with the other neuroendocrine tumours of the lung (carcinoids and large cell carcinoma) in order to better understand the role of genomics in the neuroendocrine phenotype and identify the SCLC specific alterations which may explain its aggressive nature.

In addition to copy number and expression analysis, DNA methylation analysis allows the identification of epigenetic gene regulation. DNA methylation is an important feature of many cancer types including SCLC and is linked to silencing of tumour suppressor genes and potentially to gene activation as well (Holliday, 2006; Toyooka et al., 2001). The recently development of whole genome approaches to study DNA methylation, thus offers a new dimension with which to study SCLC (Wilson et al., 2006). Given the fact that methylation is a DNA dependent assay it can be applied to clinical FFPE specimens and may offer further insight into identifying genes important to SCLC which may be missed by genetic analysis alone. This brings up an additional interesting feature of EZH2, that it directly regulates DNA methylation (Vire et al., 2006), an important feature in many cancers, with target genes involved in many cellular functions, thus DNA methylation analysis may allow further validation of the function of EZH2 in SCLC, and combination with genomic data may allow a better understanding of the regulation of the SCLC phenotype.

10.4 References

Adamowicz, M., Radlwimmer, B., Rieker, R.J., Mertens, D., Schwarzbach, M., Schraml, P., Benner, A., Lichter, P., Mechtersheimer, G. & Joos, S. (2006). Frequent amplifications and abundant expression of TRIO, NKD2, and IRX2 in soft tissue sarcomas. *Genes Chromosomes Cancer*, **45,** 829-38.

Baldwin, C., Garnis, C., Zhang, L., Rosin, M.P. & Lam, W.L. (2005). Multiple microalterations detected at high frequency in oral cancer. *Cancer Res*, **65,** 7561-7.

Bondzi, C., Litz, J., Dent, P. & Krystal, G.W. (2000). Src family kinase activity is required for Kit-mediated mitogen-activated protein (MAP) kinase activation, however loss of functional retinoblastoma protein makes MAP kinase activation unnecessary for growth of small cell lung cancer cells. *Cell Growth Differ*, **11,** 305-14.

Chen, Y., Lin, M.C., Yao, H., Wang, H., Zhang, A.Q., Yu, J., Hui, C.K., Lau, G.K., He, M.L., Sung, J. & Kung, H.F. (2007). Lentivirus-mediated RNA interference targeting enhancer of zeste homolog 2 inhibits hepatocellular carcinoma growth through down-regulation of stathmin. *Hepatology*, **46,** 200-8.

Chi, B., DeLeeuw, R.J., Coe, B.P., MacAulay, C. & Lam, W.L. (2004). SeeGH--a software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics*, **5,** 13.

Coe, B.P., Henderson, L.J., Garnis, C., Tsao, M.S., Gazdar, A.F., Minna, J., Lam, S., Macaulay, C. & Lam, W.L. (2005). High-resolution chromosome arm 5p array CGH analysis of small cell lung carcinoma cell lines. *Genes Chromosomes Cancer*, **42,** 308-13.

Coe, B.P., Lee, E.H., Chi, B., Girard, L., Minna, J.D., Gazdar, A.F., Lam, S., MacAulay, C. & Lam, W.L. (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.

Coe, B.P., Lockwood, W.W., Girard, L., Chari, R., Macaulay, C., Lam, S., Gazdar, A.F., Minna, J.D. & Lam, W.L. (2006b). Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. *Br J Cancer*, **94,** 1927-35.

Coe, B.P., Ylstra, B., Carvalho, B., Meijer, G.A., Macaulay, C. & Lam, W.L. (2007). Resolving the resolution of array CGH. *Genomics*, **89,** 647-53.

Frazier, S.R., Kaplan, P.A. & Loy, T.S. (2007). The pathology of extrapulmonary small cell carcinoma. *Semin Oncol*, **34,** 30-8.

Garnis, C., Coe, B.P., Lam, S.L., MacAulay, C. & Lam, W.L. (2005). High-resolution array CGH increases heterogeneity tolerance in the analysis of clinical samples. *Genomics*, **85,** 790-3.

Grimaud, C., Negre, N. & Cavalli, G. (2006). From genetics to epigenetics: the tale of Polycomb group and trithorax group genes. *Chromosome Res*, **14,** 363-75.

Henderson, L.J., Coe, B.P., Lee, E.H., Girard, L., Gazdar, A.F., Minna, J.D., Lam, S., MacAulay, C. & Lam, W.L. (2005). Genomic and gene expression profiling of minute alterations of chromosome arm 1p in small-cell lung carcinoma cells. *Br J Cancer*, **92,** 1553-60.

Holliday, R. (2006). Epigenetics: a historical overview. *Epigenetics*, **1,** 76-80.

Ishkanian, A.S., Malloff, C.A., Watson, S.K., DeLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M.A., Ling, V., MacAulay, C. & Lam, W.L. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, **36,** 299-303.

Kamminga, L.M., Bystrykh, L.V., de Boer, A., Houwer, S., Douma, J., Weersing, E., Dontje, B. & de Haan, G. (2006). The Polycomb group gene Ezh2 prevents hematopoietic stem cell exhaustion. *Blood*, **107,** 2170-9.

Toyooka, S., Toyooka, K.O., Maruyama, R., Virmani, A.K., Girard, L., Miyajima, K., Harada, K., Ariyoshi, Y., Takahashi, T., Sugio, K., Brambilla, E., Gilcrease, M., Minna, J.D. & Gazdar, A.F. (2001). DNA methylation profiles of lung tumors. *Mol Cancer Ther*, **1,** 61-7.

Vire, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.M., Bollen, M., Esteller, M., Di Croce, L., de Launoit, Y. & Fuks, F. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. *Nature*, **439,** 871-4.

Wilson, I.M., Davies, J.J., Weber, M., Brown, C.J., Alvarez, C.E., MacAulay, C., Schubeler, D. & Lam, W.L. (2006). Epigenomics: mapping the methylome. *Cell Cycle*, **5,** 155-8.

Zheng, M., Simon, R., Mirlacher, M., Maurer, R., Gasser, T., Forster, T., Diener, P.A., Mihatsch, M.J., Sauter, G. & Schraml, P. (2004). TRIO amplification and abundant mRNA expression is associated with invasive tumor growth and rapid tumor cell proliferation in urinary bladder cancer. *Am J Pathol*, **165,** 63-9.