GENOMIC BIOMARKERS OF RECURRENCE IN STAGE I NON-SMALL CELL LUNG CANCER

Peter Alexandrov Tcherveniakov

Submitted in accordance with the requirements for the degree of
Doctor of Medicine

The University of Leeds
School of Medicine

February 2015

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

“Stratifying tumour subtypes based on copy number alteration profiles using next-generation sequence data”. Gusnanto A, *Tcherveniakov P*, Shuweihdi F, Samman M, Rabbitts P, Wood HM. Bioinformatics. 2015 Apr 5

“A computational index derived from whole-genome copy number analysis is a novel tool for prognosis in early stage lung squamous cell carcinoma. Belvedere O, Berri S, Chalkley R, Conway C, Barbone F, Pisa F, MacLennan K, Daly C, Alsop M, Morgan J, Menis J, *Tcherveniakov P*, Papagiannopoulos K, Rabbitts P, Wood HM.” Genomics. 2012 Jan;99(1):18-24. doi: 10.1016/j.ygeno.2011.10.006. Epub 2011 Oct 25.

The candidate confirms that he is a co-author on both of these publications and was responsible for gathering the clinical, pathological and survival data of the cases.

Acknowledgements

I would like to thank the following people for their help and support:

Ornella Belvedere and Phil Egan – for helping me find my way in a laboratory environment

Caroline Conway and Rebecca Chalkley - for the tutoring in the methodology of DNA extraction, quality control and slide preparation

Leslie Davison, Kenneth McLennan and Burcu Senguven – for marking the slides and determining the tumour area

Melisa Bickerdike, Catherine Daly and Rajni Bhardwaj – for their help with the DNA libraries

Henry Wood and Stefano Berri – for their work with analysing the sequencing results and GH index

Arief Gusnatnto – for developing the logistic regression model

Vlady Vladimirov – for the technical help

Mr. Richard Milton and Mr. David Jayne – my surgical supervisors for their support and insight

Prof. Pamela Rabbits – for her continued guidance, patience and support

Abstract

Objective

Lung cancer is the leading cause of cancer-related mortality worldwide. Disease stage still remains the best prognostic factor for patients with localized non-small cell lung cancer. The TNM staging system, however, does not address the heterogeneity of this disease. Sub-classification and identification of distinct prognostic sub-groups within each stage may allow the optimization of clinical trial design and potentially improve outcome. This is a retrospective pilot study, in which we attempt to identify genomic biomarkers predictive of recurrence in stage I lung cancer by analysing copy number (CN) data obtained by next-generation sequencing.

Materials and Methods

Ninety eight patients with stage I NSCLC, who underwent elective radical surgery were identified from a tissue bank of 323 tumour samples. Their demographic and surgical data, including their recurrence status were collected and an extensive database compiled. The cases were split into two cohorts depending on their histology (adenocarcinoma vs. squamous cell carcinoma). Formalin-fixed paraffin-embedded blocks were retrieved from the local pathology archive and DNA was extracted from macrodissected tumour tissue using the QiAmp DNA microkit. DNA libraries were prepared and samples were sequenced using Illumina Genome Analyzer II. The frequency of CN gain and loss along the entire genome was compared between the recurrent and non-recurrent cancers.

Results

Comparative whole genome maps of the recurrent and non-recurrent cohort did not show any significant differences. Attempts to distinguish the recurrent from the non-recurrent cohorts with previously published algorithms, based on whole genome CN variation were also unsuccessful. However, a newly devised logistic regression model based on pan-genomic assessment of CN variation was able to differentiate recurrent from non-recurrent cancers in both histological subtypes.

Conclusion

Although no single chromosomal region was associated with cancer recurrence, the two groups were distinguishable with an algorithm that assesses total genomic change. Analysis of a larger cohort will be required for validation.

Table of Contents

Preface

This work is a result of the extensive collaboration between the Thoracic Surgical Division at St. James's University Hospital, Leeds and the Pre-Cancer Genomics group based at the Leeds Institute of Cancer and Pathology and attempts to take advantage of the growing understanding of the complex underlying biology of lung cancer and translate it into actual clinical benefit. It targets a specific group of patients and proposes a concise, practical algorithm, which could have a significant impact on their survival.

As a thoracic surgeon, initially I found myself very much out of my depth as I tried to grasp the sheer volume of work, which Prof. Rabbitts and her team had put into the genetic abnormalities of lung cancer. Luckily, I ended up working with a fantastic group of people, who not only showed me tremendous support (and vast amounts of patience), but were eager to see the results of their work find its way into clinical practice. I benefited greatly from my time spent in the lab, as I found a whole new perspective into a disease that had been the corner stone of my surgical training for almost a decade. I would like to think that I managed to reciprocate this in some way and provide some insight into the clinical course and practical issues surrounding lung cancer management, which not only led to the encouraging results of this study, but will hopefully be useful to my colleagues in their future projects.

This thesis introduces a novel method of analysing tumour copy number data, obtained by next generation sequencing, which could potentially stratify early stage lung cancer into specific subgroups. Thus it could act as an aid to the current TNM staging system and help guide targeted treatment and surveillance for a cohort of patients, which is currently managed in a somewhat uniform fashion. The emergence and continuing advances in next generation high throughput sequencing technologies could make the implementing of the described methodology both technically feasible and affordable on a large scale.

Such complex work can never be the result of the efforts of one person. Fortunately, the end product has benefited significantly from the contribution of numerous colleagues, friends and supervisors. A particular acknowledgment must be made to Prof. Pamela Rabbitts for creating the excellent environment in which the team worked, and for her continuing support and encouragement.

Hopefully, the results presented in this thesis will act as a platform for new projects, which will validate the findings and ultimately lead to measurable clinical benefit for lung cancer patients.

List of abbreviations

AC - Adenocarcinoma

AF - Autofluorescence

CN – Copy number

CNV – Copy number variation

COPD – Chronic obstructive pulmonary disease

CT – Computed Tomography

DNA - Deoxyribonucleic acid

FFPE – Fresh frozen paraffin embedded tissue

H&E – Haematoxylin and eosin

MDT - Multi-disciplinary team meeting

μl – Micro litre

μm – Micro metre

KBP – kilo-base pairs

Ng-seq – Next generation sequencing

NSCLC – Non-small cell lung cancer

PPM - Patient Pathway Management

PET – Positron emission tomography

RNA - Ribonucleic acid

SCLC – Small cell lung cancer

SCC – Squamous cell lung cancer

WHO – World health organization

Chapter 1
Introduction

1.1 Lung cancer - background

Lung cancer is the leading cause of cancer related mortality worldwide, causing more than 1 million deaths annually. Despite years of research, numerous diagnostic and therapeutic advances and many awareness campaigns, the long-term prognosis of patients who are diagnosed with the disease is dismal, with a five-year survival rate of 14-15% (Spira *et al,* 2004, Siegel *et al,* 2013). This is in stark contrast to the 5-year survival rates of other leading causes of cancer deaths, such as colonic cancer (64%), breast cancer (89%) and prostate cancer (98%) (Dela Cruz *et al,* 2009). The number of newly diagnosed lung cancer cases continues to be high, with a rising incidence in women (Dela Cruz *et al,* 2009). The role of tobacco as a causative factor has been firmly established. Smoking can cause the entire spectrum of lung cancers, but is most strongly associated with SCLC and SCC. However, a number of other factors, such as familial predisposition, chronic obstructive pulmonary disease (COPD) and toxic exposure (radon) can also contribute to the development of lung cancer. The World Health Organization estimates that lung cancer deaths worldwide will continue to rise, mainly due to the increasing global tobacco use, especially in China and India (Dela Cruz *et al,* 2009). In the United States and the United Kingdom, the decline in lung cancer rates is projected to level off in two decades because of the slow progress in smoking cessation at present (Molina *et al,* 2008). Approximately 89% of patients who contract lung cancer, will die of the disease. This is largely due to the fact that the majority of the cases present with advanced or metastatic disease, at a stage when radical therapy can no longer be offered (Stiles *et al,* 2009). Timely detection of NSCLC in high-risk individuals could help lower the mortality rates by allowing treatment at an earlier stage. However, currently no guidelines which recommend mass screening exist. The results from the US National Lung Screening Trial, which compared CT scan with chest x-ray as a screening tool, demonstrated a mortality advantage of 20% to participants in the CT group. Despite these figures, the question remains whether sufficient evidence exists to implement a screening programme based on CT and its cost effectiveness (Field *et al,* 2013). A number of novel technologies have potential for screening application and are currently undergoing evaluation

(check bronchoscopy with auto fluorescence, molecular markers in sputum and blood samples). Early diagnosis of lung cancer is vital, as survival of treated patients with stage I disease is significantly better than for those with stage II to IV (Mountain and Dresler, 1997, Spira *et al,* 2004). Surgery remains the most effective treatment for early stage NSCLC and should comprise an anatomical resection with lymph node dissection/sampling (Spira *et al,* 2004). This suggests that improvements in early diagnosis could result in improved survival. A particular focus is placed on molecular methods and their potential role in discovery and targeting treatment in NSCLC.

Histopathological heterogeneity is a major factor in lung cancer diagnosis and treatment (Travis *et al,* 2010). Lung cancer is comprised of three primary histological subtypes: carcinoid, small cell, and non–small cell, which account for about 2%, 13%, and 85% of cases, respectively. Small cell lung cancer (SCLC) is the most aggressive form of lung cancer. Non–small cell lung cancer (NSCLC) can be further subdivided into at least three histologic subtypes: adenocarcinoma, squamous cell carcinoma and large cell carcinoma. Tumours such as adenosquamous and neuroendocrine carcinomas possess histological characteristics of more than one subtype, whereas tumours from the same histopathological subtype may have dissimilar clinical outcomes and biological behaviour, such as different response to chemotherapeutic agents. The differential histopathology between lung cancer subtypes is not always obvious or objective, and proper classification is a critical component of pre-treatment evaluation. This heterogeneity has motivated efforts to classify lung cancers by their molecular profiles (Liu *et al,* 2006, D'Amico 2008).

NSCLC consists of several subtypes (adenocarcinoma, squamous cell cancer, large cell cancer), which share a similar clinical course (Spira *et al,* 2004). The Tumour (T), lymph Node (N) and Metastasis (M) system (TNM) has been the standard staging system for NSCLC, and also the established tool for determining prognosis of the disease (Mountain, 2007) (Figure 1.1).

	Size (Diameter)	Bronchoscopy	Invasion	Nodules
T1	T1a < 2 cm T1b between 2 and 3 cm			
T2	T2a between 3 and 5 cm T2b between 5 and 7 cm or	Involving main bronchus, but >2 cm to carina	Visceral pleura	
T3	Larger than 7 cm or	< 2 cm to carina	Chest wall Pericardium Diaphragm Phrenic nerve Mediastinal pleura	Other nodules in same lobe
T4		Tumour at carina	Heart and great vessels Trachea Oesophagus Spine	Nodules in other ipsilateral lobe

Regional lymph nodes (N)	
N1	In ipsilateral peribronchial and/or ipsilateral hilar lymph nodes and intrapulmonary nodes
N2	In ipsilateral mediastinal and/or subcarinal lymph nodes
N3	In contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene or supraclavicular lymph nodes

	T1a	T1b	T2a	T2b	T3	T4
N0	IA		IB	IIA	IIB	IIIA
N1	IIA		IIA	IIB	IIIA	IIIA
N2	IIIA		IIIA		IIIA	IIIB
N3	IIIB		IIIB		IIIB	IIIB

Figure 1.1. 8th edition of the TNM staging system for NSCLC. Adopted by the WHO

Appropriate staging is of paramount importance in patients diagnosed with lung cancer. It determines the treatment plan, defines prognostic groups and allows comparison of data from different research trials. With the emergence of multi-disciplinary team meetings (MDT's), during which surgeons, physicians, radiologists and pathologists discuss the most appropriate approach for managing cases with NSCLC, the TNM staging system provides a common language of communication between the different specialties. Over the last two decades this classification has undergone significant changes in an attempt to minimize the variability of prognosis within each group and correlate different treatment strategies according to stage. Mediastinal lymph node involvement in particular provides significant prognostic information and plays a central role in determining appropriate management (Little, 2009). As a consequence, establishing the correct stage of the cancer prior to initiating treatment has become particularly relevant. Several minimally invasive techniques have developed to aid accurate staging by obtaining samples of lymph node tissue from the mediastinum including endoscopic ultrasound guided biopsy and video-assisted mediastinoscopy.

Despite technological advances the survival rates for NSCLC remain poor. With its propensity for early spread, the lack of effective tools for its screening and early diagnosis and the inability of systemic therapy to cure metastatic disease the need for new strategies in screening, early detection and targeted therapy in NSCLC is evident. Focus is shifting towards understanding the biological and molecular basis for development of lung cancer with the hope that they will provide new approaches and therapeutic insights (Bunn, 2002).

1.2 Lung cancer – issues with surveillance and treatment

The established standard for treatment of stage I NSCLC is surgical resection. Currently, it remains the most consistently successful option for cure of the disease. However, the recurrence rate in these patients is approximately 35% within the first 5 years, despite them receiving what is considered to be radical therapy (Hoffman *et al,* 2000). The survival rates and disease-free intervals can vary even in patients with very similar clinical staging and pathologic features of the tumour. This suggests that NSCLC, even in its early stage, is a heterogeneous disease. The TNM classification is not able to predict which patients are likely to have recurrence of their disease. This suggests that current methodology for outcome prediction is

inadequate and identification of markers of prognosis is required to pinpoint patients who might benefit from additional therapy (radio- or chemotherapy) or targeted follow up. TNM also fails to adequately reflect the heterogeneous nature of NSCLC and failure to take this into account could lead clinicians to offer identical management plans to subgroups with different long or even short term prognosis. Until recently, the different subtypes of NSCLC such as squamous, large cell and adenocarcinoma were treated similarly. It can be speculated that the poor survival and treatment response rates of NSCLC can partly be attributed to a unified approach in treating a heterogeneous disease. A major focus of lung cancer research has centred on identifying clinically relevant biological markers, by studying lung cancer genomes and plasma protein signatures, that will allow lung cancer treatment to be individualized (Borczuk *et al,* 2010, Taguchi *et al,* 2011).

Currently, there is no consensus among medical professionals on what is the optimal surveillance for patients who undergo radical surgical resection for NSCLC. While all surveillance programs have the same principle at heart - early recognition of asymptomatic cancer recurrence, which will allow more effective therapy and in some cases even control of the disease, the actual protocols differ greatly from centre to centre, and sometimes even from individual to individual. Among the most debated issues are: what routine examinations and clinical tests are appropriate and for how long should the periodic follow-up be carried out.

In an attempt to develop an evidence-based approach to follow-up of patients after curative intent therapy for NSCLC, in 2007 Rubins *et al* performed a systematic literature review of guidelines on lung cancer diagnosis and management published between 2002 and December 2005. They produced the following recommendations:

1. In lung cancer patients treated with curative intent therapy, follow-up for complications related to therapy should be managed by the appropriate specialist and should probably last at least 3 to 6 months. At that point, the patient should be re-evaluated by the multidisciplinary team for entry into an appropriate surveillance program for detecting recurrences and/or metachronous tumours. **Grade of recommendation, 2C**
2. In lung cancer patients treated with curative intent therapy and those having adequate performance and pulmonary functions, surveillance with a history, physical examination and imaging study (either chest

radiograph (CXR) or computed tomography (CT) is recommended every 6 months for 2 years and then annually. All patients should be counselled on symptom recognition and be advised to contact their physician if worrisome symptoms are observed. **Grade of recommendation, 1C**

3. Ideally, surveillance for recognition of a recurrence of the original lung cancer and/or development of a metachronous tumour should be coordinated through a multidisciplinary team approach. If possible, the physician who diagnosed the primary lung cancer and initiated the curative intent therapy should remain as the health-care provider overseeing the surveillance process. **Grade of recommendation, 2C**
4. In lung cancer patients following curative intent therapy, use of blood tests, positron emission tomography (PET) scanning, sputum cytology, tumour markers, and fluorescence bronchoscopy is not currently recommended for surveillance. **Grade of recommendation, 2C**
5. Lung cancer patients who smoke should be strongly encouraged to stop smoking, and offered pharmacotherapeutic and behaviour therapy, including follow-up. **Grade of recommendation, 1A**

These guidelines demonstrate that current methodology might be inadequate for optimal surveillance of patients, who have undergone radical treatment of stage I NSCLC for two key reasons:

1. CT is being widely studied as a method for early detection of lung cancer recurrence. However, no established guidelines for distinguishing nonspecific post-treatment changes related to surgery, radiation therapy, and/or chemotherapy from a recurrence and/or metachronous lung cancer have been defined (Rubins *et al,* 2007). Studies report a high incidence of nodules in groups followed up with chest CT (Lamont *et al,* 2002), and the appropriate protocols for differentiating benign from malignant nodules without excess morbidity and cost from diagnostic procedures have yet to be defined.
2. Tumour molecular heterogeneity is a major reason why NSCLC patients with similar clinical staging and histology can have radically different outcomes (Herbst *et al,* 2008). This is not incorporated in the

current staging systems and surveillance protocols. Pairolero (Pairolero *et al* 1984) implemented a more rigorous follow up protocol for their stage I NSCLC patients - every 4 months for the first 2 years and then every 4 to 6 months thereafter following curative intent surgery. A history, physical examination, CXR, blood tests, urine analysis, and pooled sputum cytology are performed at each visit. However, a substantial number of recurrences were detected at unscheduled visits (41%). Most patients with recurrences were symptomatic (53%), and symptom assessment was the most sensitive method for detecting recurrences. The blood tests, urine analysis, physical examination, and sputum cytology added little benefit to detecting recurrences.

Currently, survival advantage and improved quality of life have not been clearly demonstrated with intensive surveillance programs compared to a less rigorous regimen. In addition the former are more expensive (Rubins *et al,* 2007).

While CT scan undoubtedly has a role to play in this process and can be vital for early detection of cancer recurrence, its value as a surveillance tool can be greatly augmented by the emergence of biomarkers (Bigbee *et al,* 2012). By identifying patients with high risk for cancer recurrence, a “tailor-made” follow up programme can be implemented. High risk groups could have more frequent medical reviews with CT scans scheduled at regular intervals. Any abnormal findings should be treated with a greater degree of suspicion and investigated further. Early re-discussion at a MDT meeting could be beneficial.

Anatomical surgical resection with clear margins and radical lymph node dissection offers the best chance of cure for stage I NSCLC. These patients are currently not considered for any additional therapy, such as adjuvant chemotherapy according to established treatment protocols. However, a significant number will relapse within the first five years after surgery. Adjuvant chemotherapy can offer a survival advantage in patients with radically resected lung cancer. Several trials have demonstrated that this can be as high as 15% at 5 years (Pignon *et al,* 2008, Arriagada *et al,* 2009, Douillard, 2010) and as a result adjuvant chemotherapy has become the standard post-operative management in cases of stage II-III NSCLC (Pisters *et al,* 2007). In a review published in 2005, Visbal *et al,* showed that the role

of adjuvant chemotherapy for the treatment of stage IA NSCLC or for other stages of disease has not been universally established, despite results of several trials suggesting that adjuvant chemotherapy is beneficial in the management of patients with early-stage. In their conclusion, they considered the role that molecularly targeted therapy may play in the future.

There is an evolving beneficial trend in favour of platinum-based adjuvant chemotherapy. Some authors have recommended adjuvant chemotherapy for fit patients with NSCLC lung cancer in stage I (Visbal *et al,* 2005). However, the toxicity of systemic treatment for NSCLC was considered high, and the potential benefits were minimal. In the last decade, clinicians have increased the proportion of lung cancer patients to whom they offer systemic therapy. Novel combinations of third-generation agents have demonstrated better efficacy with response rates of ≥ 30%, better tolerability, and evidence to support second-line and even third-line systemic treatment (*ie*, Docetaxel or Pemetrexed, then Erlotinib) to prolong survival and to improve symptoms and quality of life in patients with advanced NSCLC (Visbal *et al,* 2005).

However, given the morbidity associated with adjuvant chemotherapy and the variable response rate, several authors have underlined the importance of accurately identifying those cases of NSCLC, who will get an optimal result from adjuvant treatment (Arriagada *et al,* 2009, Chen DT *et al,* 2011). Having the ability to identify those patients as subgroups within the established treatment algorithm could be enormously beneficial. Biomarkers could serve as prognostic tools to pinpoint those cancer cases that have a high probability of relapse. This can have significant implications on post-surgical therapy and follow up of patients with Stage I NSCLC. It could lead to the routine implementation of adjuvant chemotherapy in such cases and the formulation of a more structured follow up within the first 5 years.

1.3 Biomarkers

Recent interest in cancer medicine has focused on identifying biomarkers and finding appropriate pathways for their practical application in routine clinical practice. All methods that can serve to quantify changes in biological homeostasis, thus distinguishing what is abnormal from what is normal can be considered biomarkers (Dalton and Friend, 2006). Biomarkers can be classified in numerous categories - from oncogenes, oncogenic protein products, growth factors, receptors, single nucleoid polymorphisms (SNP's) to genomic signatures. They may provide information on many aspects of the malignant process including the primary cancer, lymph node

involvement, likelihood of recurrence, survival prognosis and can be beneficial when selecting therapy or predicting response. Cancer biomarkers may help to overcome limitations in the TNM system and avoid a somewhat uniform approach to the heterogeneous nature of malignant disease. The benefits of integrating a molecular approach into cancer medicine are expected to manifest in two ways (Sung and Cho 2008, Dalton and Friend 2006, D'Amico 2008).

1. Early detection of malignant disease. Biomarkers could help identify people who have a predisposition to develop cancer and also help to diagnose patients at an earlier stage. This could lead to more timely treatment and improved survival.
2. Biomarkers can act as a guide in cancer therapy. Some biomarkers respond to treatment regimens with changes in their expression levels and thus serve as indicators for therapeutic response. Tissue derived biomarkers could be used for potential drug and diagnostic imaging targets.

Currently, a number of cancer therapies are based on specific cancer biomarkers (Sung and Cho, 2008). A prime example is the use of the monoclonal antibody Trastuzumab in breast cancer patients positive for HER2/neu receptor (Arteaga *et al,* 2002). The HER receptors are cell membrane proteins, which stimulate cell proliferation. In certain types of breast cancer, HER2 is over-expressed, and causes cancer cells to reproduce uncontrollably. The combination of trastuzumab with chemotherapy has been shown to increase both survival and response rate, in comparison to trastuzumab alone in patients with breast cancer, who are HER2 positive (Nahta *et al,* 2003).

In a review published in 2008, Sung and Cho suggest that biomarkers can be broadly classified in two categories – nucleic acid based and protein biomarkers.

A) Nucleic acid based. Uncontrolled cell growth is derived from either oncogene activation or tumour suppressor gene inactivation. It is therefore reasonable to assume that genetic biomarkers would be closely related to these genes. They can be further broken down in several subcategories:

1. **Chromosomal changes**. Inactivation of tumour suppressor genes during cell division is one of the key factors that drive clonal cells of

cancer into uncontrolled growth, migration and metastasis (Wistuba *et al,* 1997). Frequently this is induced by loss of DNA or chromosomal rearrangement accidentally happening during cellular division. A well-documented and frequently occurring abnormality is deletion of the short arm of chromosome 3 (3p) where several tumour suppressor genes are present.

2. **Gene hypermethylation**. Altered hypermethylation, methylation of the cytosine phosphate guanosine rich regions (CpG islands) of various promoter regions, is a representative epigenetic change in the cell and may cause gene silencing, particularly of tumour suppressor genes.
3. **Genetic change of oncogenes**. In an opposite action to previous gene silencing, activation of genes involving growth factors, their receptors, their messengers or cell cycle activators by mutations also play key roles in carcinogenesis. Mutation of *ras*, a second messenger delivering proliferation signal to the cell nucleus, is known to be involved in lung cancer. Most *ras* mutations discovered in lung cancer patients appear on codon 12 of *KRAS.* Thus, mutations or alterations of protooncogenes, which cause hyperactivation of the cell cycle, can be good biomarkers in lung cancers.

B) Protein biomarkers. They can be classified as serum biomarkers, tissue biomarkers, and sputum biomarkers (Strauss *et al,* 1994). Many protein fragments circulating in the blood stream are generated in the malignant tissues or originate from circulating proteins and cells derived from the involved tissue. Because the ultimate goal is to use biomarkers for specific, early and non-invasive diagnosis and post-therapy monitoring of cancer, blood would be an extremely appropriate biological material.

Lung cancer is one of the most prevalent malignant diseases worldwide, accounting for approximately one-quarter of all cancer deaths (Siegel *et al,* 2013). This is attributed to the late stage of the establishing of the diagnosis of the disease. The conventionally available screening tools such as X-rays, CT scans, bronchoscopy have not been shown to be effective in early detection of NSCLC. This seems to have a profound impact on survival. The

5-year survival rates for patients with stage IA can be as high as 80%, comparing quite favourably to the 15% overall survival rates for NSCLC (Mulshine *et al,* 2005). Therefore, the discovery of novel lung cancer specific biomarkers, applicable to clinical practice has become an important focus for many researchers.

Several potential biomarkers have been discovered in NSCLC such as hypermethylations of the promoters and mutations in KRAS, and p53, carcinoembryonic antigen and plasma kallikrein B1, but this has failed to translate into significant clinical benefit (Sung HJ and Cho JY 2008). The major obstacles for developing effective markers include tumour heterogeneity, the highly complex interplay between the environment and host and the complexity, multiplicity, and redundancy of tumour-cell signalling networks involving genetic, epigenetic, and microenvironmental effects. Emerging high-throughput techniques for assessing genomic DNA, messenger RNA (mRNA), microRNA, methylation, and protein or phosphoprotein signalling networks should help address these obstacles. The Cancer Genome Atlas is a large-scale project designed to provide a comprehensive profile of human tumours according to their gene mutations, alterations in gene copy number, and epigenetic changes. Lung SCC was one of the first tumours profiled by this atlas.

In pulmonary AC the discovery of multiple molecular abnormalities which are responsible for both the initiation and progression of the disease have resulted in marked changed of the established treatment protocols. The epidermal growth factor receptor (EGFR) regulates important processes involved in carcinogenesis, such as proliferation, apoptosis, angiogenesis and invasion by activating several major downstream signalling pathways (Herbst *et al,* 2008). It is frequently overexpressed in the development and progression of AC (Tang *et al,* 2005, Sato *et al,* 2007, Weihua *et al,* 2008). Clinical trials have shown that Erlotinib (a tyrosine kinase inhibitor) and Cetuximab (a monoclonal antibody against EGFR) (Shepherd *et al,* 2005, Pirker *et al,* 2008) can improve survival and quality of life in selected groups of patients, thus allowing EGFR to be considered a molecular target for therapy.

Molecular profiling of NSCLC using biomarkers could enhance the management of the disease in many different aspects (early diagnosis, surveillance, treatment). Currently, the progress made in the identification of markers, mutations, and genomic signatures is not reflected in the modest

improvement in treatments that are based on these molecular advances (Herbst *et al,* 2008).

1.4 Lung cancer – molecular background

The molecular origins of lung cancer lie in a series of complex interactions between the environment and genetic susceptibility of the host organism. They result in genetic and epigenetic changes, resulting in deregulated signalling pathways (Herbst *et al,* 2008). Emerging techniques for genomic, gene-expression, epigenetic, and proteomic profiling could revolutionize the clinical approach to the disease by helping to identify practical molecular markers of risk stratification (in pre-cancer and recurrence), early detection and prognosis, and treatment sensitivity. In recent years, new methods for high-throughput molecular analysis have been developed in an attempt to identify specific tumour markers. Genomic studies have provided information on lung cancer biology and have shown that carcinogenesis is driven by both genetic and epigenetic changes (Chanin *et al,* 2010).

The Cancer Genome Atlas is a large-scale project designed to provide a comprehensive profile of human tumours according to their gene mutations, alterations in gene copy number and epigenetic changes. Squamous cell carcinoma of the lung was one of the first tumours selected to be profiled by this atlas (Herbst *et al,* 2008). Given the tremendous potential for relatively low-cost genomic sequencing to reveal clinically useful information, cancer genomes of patients could be sequenced routinely as part of their clinical evaluation and continuing clinical management in the not-too-distant future.

The genomes of all cancers accumulate somatic mutations. These include nucleotide substitutions, small insertions and deletions, chromosomal rearrangements and copy number changes that can affect protein-coding or regulatory components of genes. In addition, cancer genomes usually acquire somatic epigenetic ‘marks’ compared to non-neoplastic tissues from the same organ. A subset of the somatic mutations in cancer cells confers oncogenic properties such as growth advantage, tissue invasion and metastasis, angiogenesis, and evasion of apoptosis. These are termed ‘driver’ mutations. The identification of driver mutations will provide insights into cancer biology and highlight new drug targets and diagnostic tests. Knowledge of cancer mutations has already led to the development of specific therapies, such as Trastuzumab for HER2 (also known as NEU or

ERBB2)-positive breast cancers (The Cancer Genome Atlas Network, 2010).

The majority of the key 'driver' mutations in NSCLC have so far been discovered in genes that encode signalling proteins, such as protein kinases and guanosine triphosphate (GTP)-binding proteins. Protein kinases regulate cellular proliferation and survival by transferring phosphate groups from ATP to specific target proteins, while GTP-binding proteins regulate cell growth, differentiation and apoptosis by interacting with multiple downstream effectors (Pao *et al,* 2011).

Two main groups of oncogenes have been investigated in NSCLC in an attempt not only to unlock the mechanisms that lie behind carcinogenesis, but to identify potential therapeutic and screening targets: dominant oncogenes and tumour-suppressor genes. The former (such as the *RAS* and *MYC* families) exert their effect by "overriding" the normal regulatory mechanisms of cellular growth. The latter on the other hand play key roles in cellular growth control, which becomes disturbed if they are deleted or mutated (p53, RB, genes on chromosome 3p) (Kijima *et al,* 2003). They can be responsible for a variety of functions such as inhibition of carcinogenic processes or be involved in repair of damaged DNA (Fong *et al,* 2003).

The *ERBB* family is a group of transmembrane receptor tyrosine kinases which are involved in cell growth regulation in NSCLC. Two members that have key roles in the development of lung cancer are the epidermal growth factor receptor (*EGFR, ERBB1*) and *HER2/neu* (*ERBB2*) (Fong *et al,* 2003). EGFR regulates important processes involved in carcinogenesis, including epithelial proliferation, apoptosis, angiogenesis and invasion, and is frequently overexpressed in NSCLC (Fong *et al,* 2003, Sato *et al,* 2007, Herbst *et al,* 2008).

The *RAS* family of proto-oncogenes (*KRAS, HRAS* and *NRAS*) encode 21-kDa plasma membrane–associated GTP-binding proteins that regulate key signal-transduction pathways involved in normal cellular differentiation, proliferation, and survival. Its members, particularly *KRAS*, can be activated in some lung cancers by point mutations, leading to inappropriate signalling for cell proliferation (Downward *et al,* 2003, Sato *et al,* 2007). *KRAS* mutations are most commonly observed in AC (Richardson *et al,* 1993, Sato *et al,* 2007) and appear to be an early event in carcinogenesis, generally marking a poor prognosis (Herbst *et al,* 2008).

The *MYC* proto-oncogene family encodes nuclear products which function as transcription factors for genes in a variety of cellular processes, including cell growth, cell proliferation and apoptosis (Adhikary *et al,* 2005, Sato *et al,* 2007). The most frequently involved family member in NSCLC, whose activation is usually caused by gene amplification, is *c-MYC* (Fong *et al,* 2003). Studies have shown that *c-MYC* amplification is associated with the development of lymph node metastasis in NSCLC (Kubokura *et al,* 2001).

The p53 gene is a key tumour-suppressor gene. It is located at the chromosome 17p13.1 and encodes a protein, which plays an important role in maintaining integrity when genomic DNA is damaged (e.g. radiation) (Fong *et al,* 2003). The most common genetic changes associated with cancer in humans involve mutations of the p53 gene (in approximately 50% of all cancers), which in turn cause a loss of tumour-suppression function and promote cellular proliferation. (Kijima *et al,* 2003, Sato *et al,* 2007). p53 has played the role of a prototypic model for gene replacement therapy in NSCLC. Clinical trials of p53 gene replacement using a retrovirus p53-expression vector in patients with NSCLC have shown evidence of antitumour activity, as well as the feasibility and safety of gene therapy. Gene replacement therapy using adenoviral p53 has emerged as a novel treatment option. A replication-impaired adenoviral vector, carrying the p53 gene, has been evaluated in both preclinical and clinical trials and results show that it is well-tolerated and can be effective in treatment for numerous cancers including NSCLC, squamous cell carcinoma of the head and neck, hepatocellular carcinoma, glioma, and breast, prostate and colorectal cancers, both as monotherapy and in combination with radiation and/or chemotherapy agents. (Gabrilovich 2006, Senzer *et al,* 2009). None of the genes have been universally implicated in the aetiology of all lung cancer.

Genomic studies have provided evidence that genetic and epigenetic alterations are driving lung cancer genesis. This strongly suggests that cancer genomics could help identify markers of prognosis and predictors of response to treatment. D'Amico and associates assessed a panel of 10 markers associated with oncologic progression in resected stage I NSCLC, reflecting all phases of tumour growth and spread using immunohistochemical analysis (rapid, reproducible, relatively inexpensive, and generally available in most hospitals). A multivariate analysis showed that five of the markers could be independent predictors of recurrence. These included p53 mutation and overexpression of the proto-oncogene *ERBB2* (D'Amico *et al,* 1999). Although each of the individual markers

carried independently significant prognostic information, patients in which the panel of all five was identified were shown to be at a significantly increased risk of cancer related death, despite receiving radical surgical resection for stage I disease. This model attempted to predict the course of the disease by focusing on the oncogenic mechanisms that define cancer biology.

In another recent analysis of 672 invasion-associated genes from 125 frozen specimens of early-stage tumours, microarray and reverse-transcriptase–polymerase-chain-reaction (RT-PCR) analyses identified a molecular signature of five genes as an independent predictor of relapse-free and overall survival (Chen *et al,* 2007). These were *DUSP6*, *MMD*, *STAT1*, *ERBB3* and *LCK*. All of them can play important roles in the biological development of NSCLC. For example, *ERBB3* is a member of the epidermal growth factor family and can lead to shortening of cell survival and is closely linked to metastasis in NSCLC (Muller-Tidow *et al,* 2005), while LCK plays an important role in the differentiation and activation of T cells, as well as in the induction of apoptosis and is expressed in many cancers (Zamoyska *et al,* 2003, Mahabeleshwar *et al,* 2004). The high risk gene signature developed by Chen *et al,* showed particular accuracy in predicting survival in patients with early NSCLC (TNM stage I and II), although the authors did not make a clear distinction between cancer subtypes (AC vs. SCC). This and similar models could prove to be extremely useful when it comes to identifying patients, who would benefit from adjuvant chemotherapy after surgical resection. However, for those benefits to be translated into clinical practice the methodology has to be readily reproducible and widely available.

These studies have described the development of gene-expression, protein, and messenger RNA profiles that are associated in some cases with the outcome of lung cancer (D'Amico *et al,* 1999, Chen *et al,* 2007). However, the extent to which these profiles can be used to refine the clinical prognosis and the context in which improved prognostic capability could be used to alter a clinical treatment decision are not clear.

1.5 Lung cancer – role of CNV and differences between histological subtypes

Appropriate characterization of the complex somatic DNA changes in NSCLC is paramount to the development of targeted therapies. Multiple studies using microarray analysis of gene expression profiles have been

performed in an attempt to improve our understanding of the aetiology of NSCLC and identify prognostic gene sets that can function as biomarkers. Systematic understanding of the molecular basis of a particular type of cancer will require at least three steps: comprehensive characterization of recurrent genomic aberrations (including CNV, nucleotide sequence changes, chromosomal rearrangements and epigenetic alterations); determining their biological role in cancer pathogenesis; and evaluation of their utility for diagnostics, prognostics and therapeutics (Weir *et al,* 2007).

DNA sequence copy number is the number of copies of DNA at a region of a genome. Cancer progression often involves structural abnormality alterations in DNA copy number. Newly developed microarray technologies enable simultaneous measurement of copy number at thousands of sites in a genome. CNV has played an important role in recent cancer studies, particularly in breast cancer (Hicks *et al,* 2006, Pollack *et al,* 2002). Furthermore, analyses of CNV in NSCLC has shown an association with both survival (Go *et al,* 2010) and therapeutic sensitivity (Hirsch *et al,* 2009).

Numerous studies have tried to ”chart” the genomic changes in the different subtypes of NSCLC in an attempt to better understand the correlation with carcinogenesis. However, just as in clinical practice, the different subtypes of NSCLC have for a long time been regarded as a single biological entity in genomic studies (Kim *et al,* 2005, Chen *et al, 2011*). Two recent studies have reported the relationship between genomic changes and disease outcome in NSCLC. Kim *et al,*. have identified several chromosomal regions as negative independent prognostic factors (Huang *et al,* 2009) and Huang *et al* have discovered single nucleotide polymorphisms (SNPs) that may be prognostic for overall survival. However, neither of these studies differentiated between lung tumour histological subtypes in their analysis. In the last few years researchers have become increasingly aware that histological subtypes of NSCLC respond differently to both targeted drugs and newly developed chemotherapies and this is likely related to differences in cell derivation and pathogenetic origins (Sy *et al,* 2004, Mok *et al,* 2009, Broet *et al,* 2009, Lockwood *et al,* 2012). For example - studies have associated a higher response rate in treatment of AC with the EGFR tyrosine kinase inhibitors, reflecting the higher prevalence of EGFR mutations in this subtype (Langer *et at,* 2010) thus highlighting the role of histology and immunohistochemistry in individualizing NSCLC treatment.

Lung cancer originates from bronchial epithelial cells. It is widely believed that the process of carcinogenesis from a normal cell to an invasive

carcinoma is a multistep process involving a number of genetic events including alterations of oncogenes and tumour suppressor genes and must have occurred before lung cancer becomes clinically evident (Panani *et al,* 2006, Herbst *et al,* 2008).

In contrast to many haematological malignancies, which are often characterized by simple and balanced chromosomal changes, epithelial tumours such as NSCLC have multiple complex and unbalanced abnormalities, which for years has complicated the identification of recurrent changes (Testa *et al,* 1997, Panani *et al,* 2006). Cytogenetic analysis demonstrated the numerous somatic genetic events that take place in the development of NSCLC. In 1997, Testa *et al* outlined the emerging patterns of recurrent chromosomal alteration and their biologic significance. They correlated the location of these changes (3p, 9p, 17p) with known tumour-suppressor genes and speculated that their loss/inactivation may play a fundamental role in carcinogenesis.

In 2001, Pei *et al* published the results of a comparative genomic hybridisation (CGH) analysis on 35 AC and 32 SCC, whose goal was to identify differences in the patterns of genomic imbalance between these histological subtypes. Many imbalances, such as gains of 1q, 5p and 8q, were shown to occur at a high frequency in AC as well as in SCC. However, several statistically significant differences were noted. The most prominent of them was gain of 3q, which was seen in 80% of SCC but in only 30% of AC. Another prominent difference was gain in 20p, which was seen in 30% of SCC versus 6% of AC. Furthermore, loss of 4q was seen at a significantly higher rate in SCC than AC while gain of 6p was more common in AC.

Five genes are known to be mutated at high frequency in lung adenocarcinoma—*TP53*, *KRAS*, *STK11*, *EGFR* and *CDKN2A*—as well as several known genes with lower mutation frequencies—*PTEN*, *NRAS*, *ERBB2*, *BRAF* and *PIK3CA* (Weir *et al,* 2007). A study by Weir *et al,*, in which 623 genes from a cohort of 188 tumours were sequenced, identified further significantly mutated genes, more than doubling the list. The newly identified genes included tumour suppressor genes (*NF1*, *RB1*, *ATM* and *APC*) along with tyrosine kinase genes (ephrin receptor genes, *ERBB4*, *KDR*, *FGFR4* and NTRK genes) that could function as proto-oncogenes. They demonstrated that many of these genes were also targeted by copy number variations and/or gene expression changes. However, few genes have shown to consistently have mutations in AC. The incidence of the most frequent mutation (*TP53)* is around 35% (Ding *et al,* 2008). This lack of a

universal mutation pattern suggests that the molecular pathogenesis in the development of lung adenocarcinoma is quite variable and different subtypes probably exist.

In an analysis of aCGH data, Chitale *et al* showed that lung adenocarcinomas display non-random patterns of co-occurring gains and losses, one of which is characterized by 7p gains (including the *EGFR* locus) and 8p losses. Previous studies also noted 8p losses, but also failed to narrow the target region (Weir *et al,* 2007). Allelic losses on 8p are well described in other cancers, including breast, prostate, and bladder, with most studies finding a complex pattern that cannot be reduced to a single minimally deleted region (Chitale *et al,* 2007).

Previously, aCGH based studies reported common aberrations in SCC including gains of chromosomal arms 3q, 5p and 8q and losses of 3p, 5q and 8p (Sy *et al,* 2004, Yakut *et al,* 2006, Tai *et al,* 2004, Pei *et al,* 2001, Chujo *et al,* 2002). Several common high copy number amplifications are 2p15-p16, 3q24-q29, 8p11-p12, 8q23-q24, and 12p12 (Pei *et al,* 2001, Boelens *et al,* 2009). Studies have identified candidate (onco) genes located on these sites. They include *BCL11A* (2p), *REL* (2p), epithelial cell transforming sequence 2 oncogene (*ECT2*) (3q), *PIK3CA* (3q), *ADAM9* (8p), *MYC* (8q), and *KRAS* (12p) (Boelens *et al,* 2009). Gains on 7q have been previously described and associated to positive lymph nodes in NSCLC in general and gain of 7q and loss of 4q have been reported to be related to general metastatic behaviour of SCC (Pei *et al,* 2001, Yan *et al,* 2005). Gains on 8q have been described in several cancer types in relation to metastasis, progression, poor prognosis, or survival and have been identified in SCC (Boelens *et al,* 2009). In 2009, Bass *et al,* showed that a peak of genomic amplification on chromosome 3q26.33, found in lung and oesophageal SCC, contains the transcription factor gene *SOX2. SOX2* expression is required for proliferation and anchorage-independent growth of lung and oesophageal cell lines and was identified as a lineage survival oncogene in lung and oesophageal SCC.

In May 2012 Lockwood *et al* published a large-scale analysis of 261 primary NSCLC tumours (making a clear distinction between SCC and AC), integrating genome-wide DNA copy number, methylation and gene expression profiles in an attempt to identify subtype-specific molecular alterations relevant to new agent design and choice of therapy. Comparison of AC and SCC genomic and epigenomic landscapes revealed 778 altered genes with corresponding expression changes. The study identified key oncogenic pathways disrupted in each subtype that are likely to serve as the

basis for their differential tumour biology and clinical outcomes. Downregulation of HNF4a target genes was the most common pathway specific to AC, while SCC demonstrated disruption of numerous histone modifying enzymes as well as the transcription factor E2F1. Overall, the findings of the study suggested that AC and SCC develop through distinct pathogenetic pathways that should have significant implication in the approach to the clinical management of NSCLC (Lockwood *et al,* 2012).

In 2013 Staff *et al* published their results of an extensive analysis and comparison of the different features of genomic alteration in lung cancer including alteration frequency and amplification patterns aiming to identify important CNV in lung cancer, both NSCLC (SCC, AC and LCC) and SCLC. They analysed 2141 lung cancers and cell lines and observed 89 regions of CNV (55 gains and 34 losses) distributed across all autosomes and an analysis of a random subset of 1606 cancers showed that 62% and 80% of the 89 regions were detected in >90% or >70% of permutations, respectively. They came to the conclusion that AC exhibits a generally lower frequency of CNV compared to other histology's, while several CNV are markedly shared between different histology's. As an example, characteristic CNV in AC (with frequency of occurrence of 40% or higher) were amplification in 1q, 5p, 7p and deletions in 8p, 9p, 13q, 17p, 18q and 19p. Similarly for SCC their findings were amplification in 1q, 3q, 5p, 7p, 7q, 8p, 8q, 12p and deletions in 1p, 3p, 4p, 4q, 5q, 8p, 9p, 10q, 13q, 17p, 18q, and 21q.

Their analysis, perhaps surprisingly, concluded that genomic instability affects AC to a lesser extent compared to other histology groups and speculated that the observed heterogeneity of genomic abnormalities in the different histological groups of lung cancer supports the existence of further molecular subtypes, which might have clinical relevance, such as targeted therapy.

Whilst the vast majority of copy number studies have examined cancer genomes in a *locus* by *locus* manner, one study by Hicks *et al* correlated survival in patients with breast cancer not to individual *loci* but to a pan-genomic index that measured the type and extent of genomic damage. The group examined 243 breast tumours and identified three distinct patterns of genomic CNV, naming them according to the appearance of their karyograms (graphical representation of the chromosomes in a karyotype). They observed an association between certain types of karyograms and disease aggressiveness and speculated that CN profiling might provide useful information in guiding clinical decisions (Hicks *et al,* 2006).

The results from the aforementioned studies clearly show that CNV analysis, on its own or as part of an integrated approach, can be used to great effect

not only in the genomic studies of NSCLC, but also to differentiate the specific genetic changes responsible for the process of carcinogenesis in the different histological subtypes (AC and SCC).

1.6 Next generation sequencing

In 1977, Sanger and associates published an article describing a new approach for determining nucleotide sequences in DNA. It was based on using chain-terminating dideoxynucleotide analogs that caused base-specific termination of primed DNA synthesis and utilizing gel electrophoresis to separate the products of the reaction. This method, after some additional refinement, became the main tool of the research community in the attempt to decipher the code of human DNA and translate those findings into clinical practice.

The first complete sequencing of a human genome (as part of the Human Genome Project) was accomplished in 2003, using a modified version of the platform developed by Sanger *et al,*. This undertaking took 13 years and cost an estimated $2.7 billion (Voelkerding *et al,* 2009). In 2008, Wheeler and associates published their findings, after sequencing a complete individual human genome by using massive parallel DNA sequencing. Their project took approximately 5 months and cost around $1.5 million (Wheeler *et al,* 2008) and demonstrated the advantages of the "next-generation" sequencing platforms, which have emerged in the last 10 years.

The NG-seq approach offers a number of advantages over traditional methods, including the ability to fully sequence large numbers of genes (hundreds to thousands) in a single test run and simultaneously detect deletions, insertions, CNV and translocations in cancer genomes (Ross *et al,* 2011).

All NG-seq platforms perform massively parallel sequencing of clonally amplified or single DNA molecules that are spatially separated in a flow cell (Voelkerding *et al,*. 2009). This is in contrast with the traditional Sanger sequencing, which is based on the electrophoretic separation of chain-termination products produced in individual sequencing reactions (Sanger *et al,*, 1977). As a massively parallel process, NG-seq generates hundreds of megabases to gigabases of nucleotide-sequence output in a single instrument run, depending on the platform (Voelkerding *et al,*. 2009). Several platforms are commercially available. The Illumina platform was utilized in this study, as it was already in use in the facility and experience with its performance and analyzing the data was readily available.

The Illumina platform utilizes a sequencing-by-synthesis approach coupled with bridge amplification on the surface of a flow cell. Each flow cell is divided into eight separate lanes. The interior surfaces of the flow cells have covalently attached oligos complementary to specific adapters that are ligated onto the library fragments. DNA fragment-to-oligo hybridization on the flow cell occurs by active heating and cooling steps. This is followed by a subsequent incubation with the amplification reactants and an isothermal polymerase that generates millions of clusters of the library fragments. In the sequencing step, each cluster is supplied with polymerase and four differentially labelled fluorescent nucleotides that have their 3-OH chemically inactivated. This blocking modification ensures that only a single base will be incorporated per flow cycle. After each nucleotide is incorporated, an excitation followed by an imaging step takes place to identify the incorporated nucleotide in each cluster. A chemical deblocking treatment removes the fluorescent group and allows the incorporation of the following nucleotide during the next flow cycle (Shokralla *et al,* 2012). The sequence of each cluster is computed and subjected to quality filtering to eliminate low-quality reads (Shendure *et al,* 2008).

The production of large numbers of low-cost reads makes the NG-seq platforms useful for many applications. Furthermore, in comparison to automated Sanger sequencing they have dramatically increased throughput and lowered expenditure. This has provided a challenge to the existing IT facilities in terms of data storage and computational analysis to align read data (Metzker 2010).

A useful utilization of NG-seq in clinical practice will place significant demands on laboratory infrastructure, will require extensive computational resources and a thorough knowledge of cancer medicine and biology. It is anticipated that continuing advances in this technology will lower the overall cost, speed the turnaround time, increase the breadth of genome sequencing, detect epigenetic markers and other important genomic parameters, and become applicable to smaller and smaller specimens. (Ross *et al,* 2011).

1.7 Choice of methodology

The majority of the discovery efforts are based on the collection, storage and processing of tissue specimens obtained at the time of surgery, bronchoscopy or other diagnostic procedures. After informed consent, all biological specimens need to be collected under a standard operating

procedure and quality control must be in place to guarantee adequacy of the samples. This requires a concerted effort between clinicians, pathologists and researchers. Whilst profiling using high-throughput technologies is best served by the use of fresh frozen materials, and tumour-derived markers are likely to be present at lower levels in blood and other biological specimens, for the purpose of this study we have used formalin-fixed, paraffin-embedded tissue. This has several advantages. It allows the use of large collections of available tissue, which are appropriately catalogued and easier to handle. Information on long-term survival, natural history and cancer recurrence can be obtained from medical or pathologist databases. Methods of extraction allow the recovery of usable DNA and RNA for high-throughput discovery and validation strategies. Promising single-molecule sequencing and high-throughput oncogene mutation profiling represent strategies that may be applicable to small clinical samples in the future to address personalized medicine. However, before the next generation of sequencing enters clinical use, issues of cost, data analysis and interpretation will have to be resolved (Ocak *et al,* 2009).

The ability to detect CNV of cancer cells is a crucial step to access the severity of chromosomal rearrangements and to find chromosomal regions where breakpoints are located. Furthermore, comparison of CNV across tumours from different patients makes it possible to find regions commonly duplicated or lost, which highlights the locations of cancer-related genes. Several methodologies are available to detect CNV, such as Comparative Genomic Hybridization, array CGH, single nucleotide polymorphism array (SNP arrays) and, more recently, a new generation of sequencing machines enabled massively parallel sequencing (Roche 454, Illumina GAII, HiSeq, MiSeq, ABI SOLiD, Ion Torrent PGM), making it possible to sequence full genomes at affordable cost.

NG-seq is one of the most significant recent technological advances in cancer research and could potentially bridge the gap between the scientific and clinical setting. This is largely defined by its ability to analyze entire human genomes in a matter of days, while at the same time allowing for massive parallel sequencing of multiple DNA samples. The rapid development in informational technologies and bioinformatics has allowed for the generation of large amounts of cheap data and, perhaps more importantly - its analysis (Ulahannan *et al,* 2013). Technological advancement of sequencing platforms has improved not only data quality and throughput, but has led to a decrease in cost, with the price of

sequencing of a whole genome estimated to be around 1000 USD (Meldrum *et al,* 2011).

With the Illumina NG-seq platform, genomic DNA is sheared (either mechanically or using enzymes) into fragments of 75–150 base pairs. Adapters are ligated to the fragments and bind them to the surface of a flow cell channel. The fragments are then amplified and sequencing commences by adding four labelled reversible terminating nucleotides. The fluorescent signal, which is emitted after the addition of the terminating nucleotide is captured and denotes the type of base incorporated in the sequence. The cycle is repeated one base at a time generating a series of images with each image representing a single base in the priming sequence (Ulahannan *et al,* 2013).

It has been shown how it is possible to multiplex several samples in one Illumina GAII lane making copy number analysis by sequencing affordable and competitive with aCGH or SNP arrays (Wood *et al,* 2010). As we expect sequencing technologies to become more widespread, affordable and accurate, copy number analysis by low coverage sequencing will become even more convenient and informative. Furthermore sequencing is possible even with low amounts of DNA extracted from formalin-fixed paraffin-embedded specimens (Wood *et al,* 2010).

1.8 Study design

This is a retrospective pilot study, the aim of which is to identify genomic patterns of recurrence in patients with stage I NSCLC using CNV analysis of cancer DNA from FFPE, obtained by Ng-seq. While several studies have already identified molecular profiles, which are associated with poor outcomes and higher risk of recurrence, their discoveries have failed to influence established clinical practice. The reasons for this are complex. They range from complicated methodology, which is not easily reproducible in routine clinical practice and/or time consuming to lack of adequate validation studies. Thomas D'Amico and associates presented their biological risk model for NSCLC as far back as 1998, using immunohistochemical analysis of molecular markers associated with different oncogenic pathways (D'Amico *et al*, 1998). Despite convincing results and large number of patients, this study has had very limited clinical impact. Chen *et al,* published in 2007 in the NEJM a reverse-transcriptase PCR–based five gene signature, the presence of which was associated with an increased risk of recurrence and decreased overall survival (Chen *et al*,

2007). The results of this study, which involved a Chinese patient population, were also validated with the use of a set of published NSCLC microarray data from patients from a Western population. The authors strongly felt that their model could successfully be used to guide post-operative management in patients with stage I NSCLC, by further sub-stratifying them according to risk. However, the methodology that they used is somewhat time consuming and difficult to apply on a large scale. With the recent advances of high-throughput genotyping, screening for specific disease loci on a genome-wide scale is now becoming not only possible, but ever more practical and affordable. In 2004, Paris *et al* published a study, which looked into the relationship between CNV and recurrence in prostate cancer. In a cohort of 64 patients, their analysis revealed numerous recurrent copy number aberrations. The authors noted that gain at 11q13.1 seemed to be predictive of postoperative recurrence, independent of stage and grade and suggested this could be an important finding on the road to developing personalized care for patients with prostate cancer.

The findings of the above-mentioned studies suggest that what is currently defined as stage I NSCLC is in fact a heterogeneous disease group, to which currently there is a unified therapeutic approach.

In an article published in the Lancet in August 2013, Rossel *et al* discussed the current role of genetics and biomarkers in the personalisation of treatment of NSCLC. They commented on the growing evidence for substantial genetic heterogeneity among individual non-small cell lung cancers and underlined how this phenomenon is likely to be responsible for cancer resistance to chemotherapeutic agents (Rossel *et al*, 2013). The group concluded that the key to developing effective targeted therapy would require an unbiased, systematic, and genome-wide analysis of individual tumours in every patient undergoing treatment for NSCLC.

The aim of this project is to determine if there are detectable differences between the genomic signatures of recurrent and non-recurrent stage I NSCLC, by using a methodology that can be implemented in routine clinical practice. By taking advantage of the increasing availability of high throughput sequencing technologies, which allow multiple samples to be processed simultaneously, the discovery and validation of such signatures could readily be implemented into the clinical practice. A tumour sample, obtained at the time of surgery could be sequenced and analysed whilst the patient is recovering in the early post-operative period. Multiple samples could be processed at the same time and their genomic signatures made available for the initial patient follow up appointment and/or multi-disciplinary team (MDT) discussion. If a high-risk profile is discovered, the patient could be offered adjuvant chemotherapy and/or a specific surveillance program aimed at

detecting early recurrence. This would offer a “tailored” approach to patient care.

1.9 Hypothesis and Objectives

1.9.1 Hypothesis

The hypothesis of this thesis is that patients with stage I NSCLC can be further stratified into clinically significant subgroups, based on their likelihood of recurrence after radical surgery and that these subgroups can be identified by genomic signatures based on CNV data obtained by Ng-seq.

1.9.2 Objectives

The main objectives of the thesis are:

1. To differentiate the genomic signatures of recurrent vs. non recurrent stage I NSCLC tumour samples, who underwent radical surgery using CNV analysis of Ng-seq data
2. To develop a practical algorithm for this, which could be introduced into clinical practice.
3. To suggest a specific role for this algorithm in clinical practice, which would improve patient care.

Chapter 2
Materials and Methods

2.1 Sample collection

A cohort of 323 formalin fixed paraffin embedded (FFPE) wax blocks of tumour samples, stored in metal containers at room temperature in a dedicated area, was available in our laboratory. They were obtained from patients who underwent elective surgery for NSCLC over a 5-year period (1997-2002) in the Thoracic Surgery Department at Leeds Teaching Hospitals NHS Trust. They were previously acquired from the archives of the Department of Pathology in the Leeds General Infirmary. All of the tumours had undergone routine histological evaluation postoperatively and had been confirmed primary lung cancers. One hundred and seventy two were SCC and 151 were AC. Due to the nature of disease progression in NSCLC and the extent of radical surgery (complete excision with clear resection margins), the size of the tumour blocks allowed for multiple sampling.

2.2 Study criteria

The criteria for inclusion in the study were:

1. Stage I cancers according to the 8th edition of the TNM classification
2. Sample recovered from patients who underwent radical surgery (lobectomy or pneumonectomy)
3. Patients who had not received any adjuvant or neo-adjuvant treatment (chemotherapy or radiotherapy)
4. Confirmed evidence of recurrence (either local or systemic) or confirmed disease-free 5- year survival.

The following exclusion criteria were defined:

1. Cases of early peri-operative death (within 3 months from surgery).
2. Patients with history of another cancer.
3. Patients who underwent sublobar resections (wedge, segmentectomy) and patients with positive resection margins who required post-operative radio and/or chemotherapy.
4. Patients with insufficient lymph node sampling to obtain a formal TNM stage.

2.3 Data gathering

Ethics approval for the project had already obtained by the department (reference number 07/Q1206/30). Its conditions were revised in order to ensure that the study was compliant. No conflicting issues were identified. An anonymized, secure database was compiled, including relevant demographics, clinical and outcome data (donor's age at diagnosis, gender, histology report, stage of disease, type of surgical procedure and recurrence details). Due to the fact that most of the cases underwent surgery more than a decade prior to the commencement of this study, I used several sources to obtain the relevant data, in particular the outcome and long-term survival data. All cases, for which survival and recurrence status could not be conclusively verified, were excluded from the study. The Patient Pathway Management (PPM) database, PACS radiology system and histopathology database (Co-path) used in the Leeds Teaching Hospitals Trust as well as the Yorkshire Cancer Registry were utilized. PPM incorporates a large amount of data, including pathology reports, survival, clinician letters, data on trial participation, additional operations and procedures (cancer related or not). The PACS system contains images and reports from radiology studies performed in Leeds Teaching Hospitals. The Yorkshire cancer registry was referred to for outcome data, in particular, cause of death. This was necessitated by the fact that a few of the patients were referred from outside of the Leeds area and the data on PPM was incomplete. Several cases required additional discussion with a consultant pathologist in order to confirm the definitive staging of the cancer. The cases that complied to the selection criteria were grouped into two cohorts as per the study design – A) Patients who underwent radical surgery for stage I NSCLC and had recurrent cancer within the first five years from the operation and B) Patients who underwent radical surgery for stage I NSCLC and did not have recurrent cancer within the first five years from the operation. Recurrence was confirmed by radiological findings (CT scan) (Figure 2.1. Breakdown of sample cohort by histology and recurrence statusFigure).

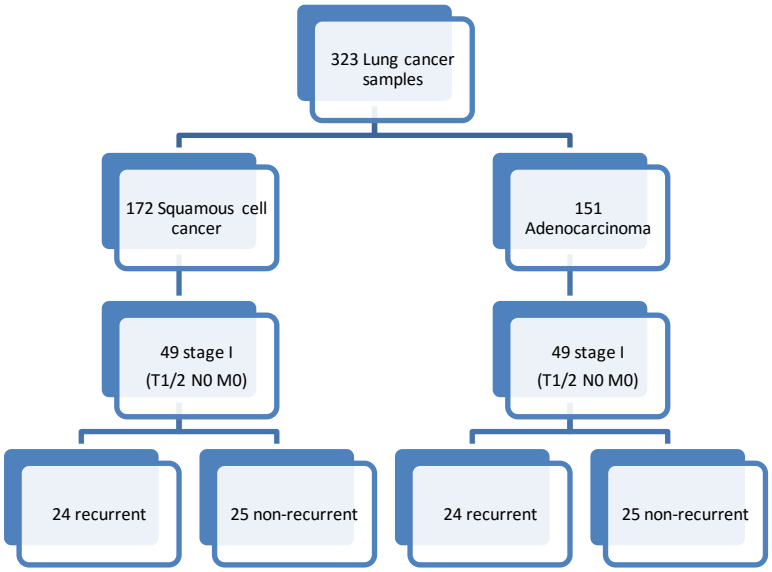


Figure 2.1. Breakdown of sample cohort by histology and recurrence status

2.4 FFPE tumour block sectioning

The FFPE tumour blocks were cut into sections using a microtome. From each sample, deemed suitable for the study, seven consecutive 7-μm-thick sections were cut and mounted on a slide, which was labelled with the relevant coded number. A further 4μm-thick section was cut from each block and stained with haematoxylin and eosin (H&E). An independent pathologist, blind to patient identity and diagnosis, reviewed all of the H&E stained slides in order to (i) confirm the diagnosis and histology reported in the original pathology report; (ii) mark the most representative tumour areas in each slide using a fine-tipped permanent marker and (iii) evaluate the percentage of tumour cells in the marked area, corresponding to the macrodissected tissue used for DNA extraction. The data was then entered into the secure database.

2.5 DNA isolation

Tumour genomic DNA from macrodissected FFPE tissue was extracted using a commercially available QIAmp DNA mini kit (Qiagen, Crawley, West Sussex, UK). The slides, corresponding to the actual tumour block were initially heated on a hot plate at 65°C for 3 min. Following this, the slides were loaded onto racks and de-waxed and rehydrated by immersion into glass baths in the following succession: xylene for 5 min, 100% ethanol for 3 min, 90% ethanol for 3 min, 70% ethanol for 3 min and finally ddH_2O, where they remained until further processing. Sections were then immediately

macrodissected using sterile disposable scalpels (Swann-Morton Ltd, Sheffield, England) to harvest the tumour tissue; the corresponding H&E-stained and marked-by-pathologist slide was used as a guide (Figure 2). All seven slides containing the 7-μm-thick sections from each tumour block were macrodissected for every case. DNA extraction was performed using the QIAamp DNA Mini Kit according to the manufacturer's instructions (Qiagen). All the macrodissected tissue from each case was placed in a separate microfuge tube, labelled with the unique patient study ID. Following this, 180-µl of Buffer ATL and 20-µl of Proteinase K were added to the tube. The samples were mixed by pulse-vortexing for 10 s. and placed in a water bath for incubation at 56 ^{0}C for 48 hours to obtain complete lysis of the tissue. They were reviewed after 24 hours and an additional 20 μl of Proteinase K was added to those samples that still had free-floating tissue. The tubes were vortexed daily for 15-s to enable mixing. After 48 hours the samples were removed from the water bath and 200 μl of Buffer AL was added and mixed by pulse-vortexing for 15-s The samples were then incubated at 70 °C for 10 minutes and briefly centrifuged after cooling down. 200 μl of 100% ethanol were added to each tube as a next step. The samples were pulse-vortexed for 10-s and left to incubate at room temperature for 5 minutes. The mixture was then carefully transferred to a spin column, using a pipette, and centrifuged at 8000 rpm for 1 minute. The spin column was transferred to a clean centrifuge tube and the tube containing the filtrate was discarded. 500-μl of buffer AW1 was added to each tube and the samples were centrifuged again at the same speed. The spin column was transferred to a clean centrifuge tube and the tube containing the filtrate was discarded. 500-μl of buffer AW2 were added to each tube and the samples were centrifuged at 14 000 rpm for 3 minutes. The spin column was placed in a clean 1.5 ml tube, labelled with the corresponding study ID. The remaining filtrate was discarded. 100-μl of buffer AE added to each column and the samples were incubated for 5 min at room temperature. Following this, they were centrifuged at 8000 rpm for 1 minute. The filtrate was clearly labelled as Elution 1. The spin columns were placed in clean tubes (clearly labelled as Elution 2) and buffer AE was added to obtain a second sample from each block. The final result was two DNA elutions of each sample, prepared in 100-µl of buffer and stored at 4 °C.

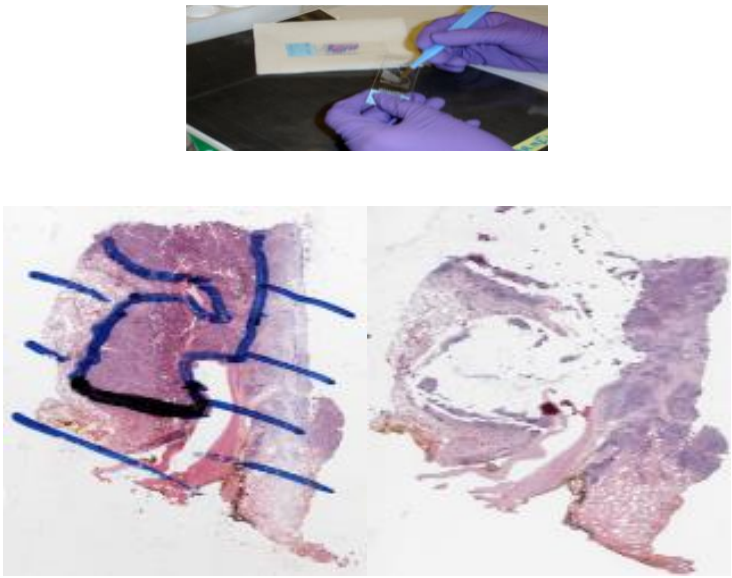


Figure 2.2. FFPE tumour tissue was macro-dissected (top) using an H&E slide, previously marked by a pathologist (bottom).

2.6 Quality control

Quality control of the samples, assessing DNA concentration and purity was performed using both spectrophotometry (Nanodrop-8000) (Figure 2.3) and fluorescent nucleic acid staining (Quant-iT PicoGreen dsDNA BR assay, Invitrogen, Paisley, UK).

2.6.1 Spectrophotometry with Nanodrop

The Nanodrop is a cuvette-free full spectrum (220-750 nm) spectrophotometer. It can measure samples of just 1-μl with high accuracy and reproducibility and is accurate for concentrations from 5-ng/μl up to 3,000-ng/μl. It measures DNA, RNA and protein concentration.

The concentration and quality of DNA were measure using a ND- 1000 spectrophotometer (NanoDrop, Wilmington, DE, USA). The procedure was performed immediately after extraction. The nanodrop was first normalised with 1μl of diH_2O and then blanked with 1μl buffer AE (the elutant in all samples). The sensor of the ND-1000 was wiped dry with a clean tissue before adding 1μl of the selected DNA sample. Measurement of the concentration by UV spectrophotometry was initiated using the associated software package (NanoDrop 1000 v3.7.1). The nucleic acid concentration in ng/μl was calculated and recorded automatically by the software in a

spreadsheet along with 260/280nm (DNA). The samples were processed in batches of three, thus allowing for two measurements of each elution. After each run, the sensor was wiped dry again with a clean tissue and measurement of the next batch took place. All data from the programme spreadsheet were then exported into Microsoft Excel™ and saved for future reference.

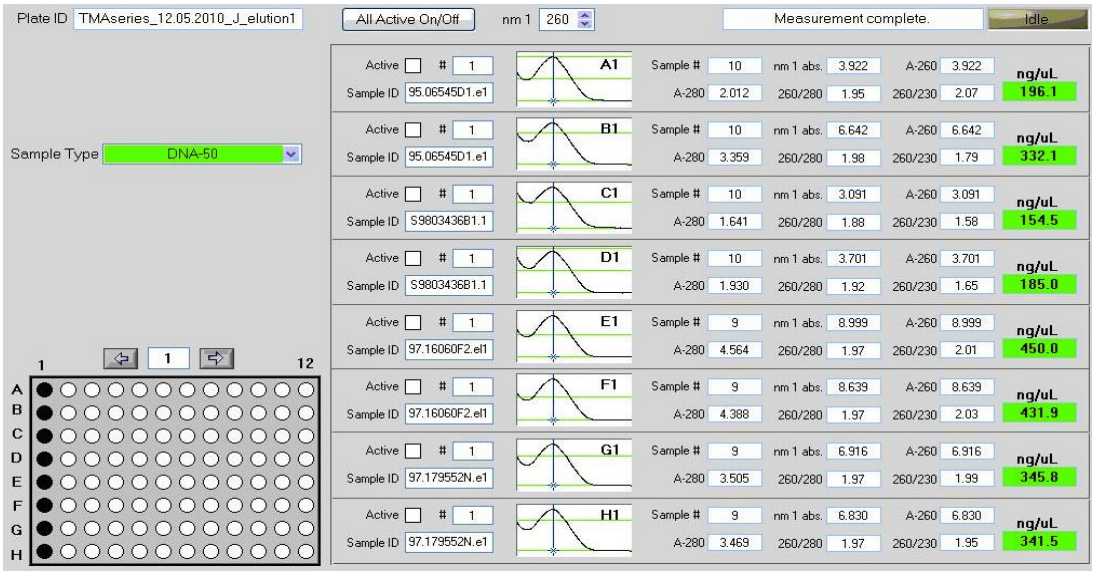


Figure 2.3. Example of Nanodrop run. Measured nucleic acid concentrations are highlighted in green.

2.6.2 PicoGreen

Picogreen is an ultra-sensitive fluorescent nucleic acid stain for quantifying double-stranded DNA. The major disadvantage of the spectrophotometry is the contribution of nucleotides, single-stranded nucleic acids and proteins to the signal, the interference caused by contaminants commonly found in nucleic acid preparations and the inability to distinguish between DNA and RNA.

The Quant-iT™ dsDNA Broad-Range Assay Kit was used for DNA quantification.

The kit provides a concentrated assay reagent, dilution buffer, and pre-diluted DNA standards.

The assay is highly selective for double-stranded DNA over RNA, and in the range of 2–1000 ng the fluorescence signal is linear with DNA. The assay is performed at room temperature. Common contaminants, such as salts, solvents, detergents, or protein are well tolerated in the assay.

A working solution was prepared by diluting Quant-iT™ dsDNA BR reagent 1:200 in Quant-iT™ dsDNA BR buffer. 200 µl of the working solution were loaded into each microplate well. 10 µl of each DNA standard were added to separate wells and mixed. 10 µl of each investigated DNA sample were added to separate wells and mixed. The plate was loaded in a reader and the fluorescence was measured. A standard curve was used to determine the DNA amounts.

2.7 DNA library preparation

The principle underpinning Ng-seq is the use of small fragments of DNA, the base sequence of which are sequentially identified from emitted auto fluorescent signals as each fragment is re-synthesized from a DNA template strand. The sample DNA is first fragmented into numerous small segments (using enzymatic or mechanical shearing), which are “sequenced” in millions of parallel reactions. The small segments (or strings) of DNA are tagged with known adapters and are called reads. The position of each read is established by aligning them to a known reference genome

DNA libraries were prepared and sequenced using methods previously described by our group [Wood *et al*, 2010].

1. DNA was first sheared into a random library of 100-300 base-pair long fragments. This was performed on a Covaris S2 Sample Preparation System (Covaris Inc., Woburn, MA, USA) and checked for appropriate size distribution on an Agilent Bioanalyser DNA 1000 LabChip (Figure 2.4).
2. After fragmentation the ends of the DNA-fragments were repaired. End repair was performed by using the End-It DNA End Repair Kit (Epicentre Biotechnologies, Madison, WI, USA).
3. A-Addition. An A-overhang was added at the 3'-end of each strand using Klenow DNA polymerase.
4. Ligation. DNA ligases catalyze the formation of a phosphodiester bond between the 3' hydroxyl and 5' phosphate of adjacent DNA residues. This reaction is used to add bar-coded adapters to fragmented DNA. Adaptors, which are necessary for amplification and sequencing, are ligated to both ends of the DNA-fragments. Libraries were prepared for sequencing with a unique 6bp adapter ligated to enable multiplexing.

5. Enrichment. PCR was used to enrich/amplify final adapter modified fragment sample to increase the overall amount of library prep. A 15 cycle enrichment regime was used for all samples.
6. Quality control. DNA quantification was performed using PicoGreen as well as microchip electrophoresis using Agilent Bioanalyser analysis. An example is shown in Figure 2.5.

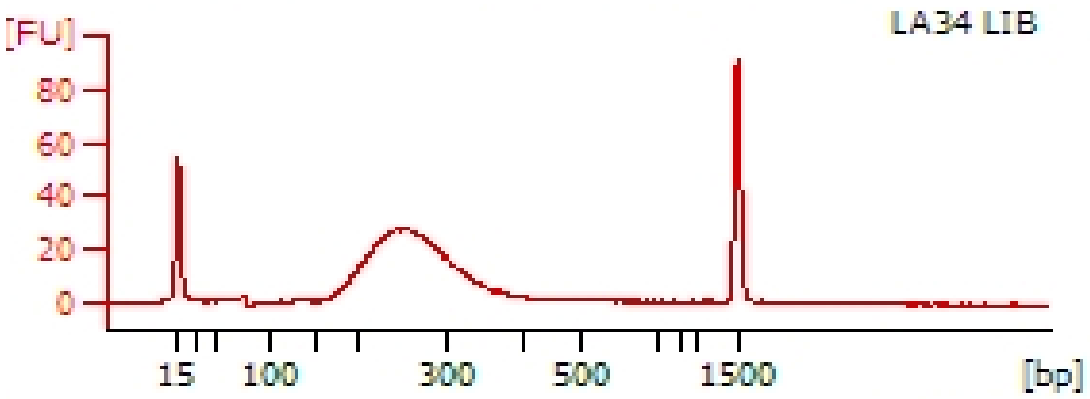


Figure 2.4. Example of quality control of library preparation using Agilent Bioanalyser. The two tall, narrow peaks represent size selection markers. The tented curve shows the amount of DNA in the prep.

Twenty samples were pooled per lane on an Illumina GAII sequencer for 76 cycles of single end sequencing resulting in 70bp of genomic sequence and 6bp of adapter. Files were split according to adapter sequence and the remaining 70bp fragments were “mapped” to a known human genome (USCS hg19) using the Burrows-Wheeler Alignment tool (BWA), thus determining their genomic site (Li and Durbin, 2009). A key advantage of Burrows–Wheeler algorithm-based software programs is their relatively low memory requirement. The process of mapping generates a BAM file, in which the reads from the sequencing have been assigned a position relative to the reference genome while retaining information regarding unmapped reads. In this study sequences were aligned using a bwa suite (version 0.5.9-r16). A software package written in R, called CNAnorm (Gusnanto *et al,* 2012) and designed at the University of Leeds was used to normalize the data. CNAnorm has been used to analyse Ng-seq data in several previous projects (Wood *et al,* 2010, Belevedere *et al,* 2011) and one of its main advantages is normalization of sequencing data obtained using low coverage (one read every 100-10,000 bp). It provides an algorithm that corrects sample contamination with normal cells and adjusts for genomes of

different sizes so that the actual copy number of each region can be estimated. Copy number was calculated by splitting the genome into windows averaging 300 tumour reads per window. The windows were then aligned to a control sample in order to determine the copy number variations. The control sample was constructed from a pool of 20 normal British individuals downloaded from the 1000 genomes project (Durbit *et al,* 2010). The ratio for number of tumour and normal reads in each window was calculated.

Copy number karyograms were generated for each sample using the CNAnorm software package.

A statistical analysis was then undertaken using the Bioconductor package KC-SMARTR (Venkatraman *et al*, 2007, Clijn *et al,* 2007), which can detect significantly altered regions and compare two groups of samples. The latter is the major advantage of this particular software package and was the main reason it was chosen over other similar packages.

2.8 Models of global genomic patterns associated with recurrence

2.8.1 Genomic signature based on karyogram patterns

In 2006 Hicks *et al* examined 243 breast tumours and identified three distinct patterns of genomic CN variation, naming them according to the appearance of their karyograms: ‘simplex’ - few aberrations, mostly involving whole chromosome arms, ‘sawtooth’ - many aberrations spread throughout the genome and ‘firestorm’ - like simplex, with local regions of complex damage. They observed an association between “firestorm” and disease aggressiveness and speculated that CN profiling might provide information useful in making clinical decisions. The approach of Hicks *et al*. was adapted and applied to the investigated cohort in an attempt to identify an association between a specific global pattern of CNV and cancer recurrence.

2.8.2 Pangenomic index (GH)

The method of generating a novel pangenomic index was developed by our group as a part of a larger project looking into SCC, and has been published in 2011 (Belvedere *et al,* 2011). As a first step to calculating the GH index the distribution of copy number along the entire genome is calculated and presented in density plots. The density plots are generated after the copy number data from the sequencing is smoothed using the CNAnorm software

package. The areas where the CNV most commonly occur along the genome are represented as “peaks” on the plot and the three most prominent are labelled A, B and C (figure 6). The “highest” peak represents the most common copy number state in that particular sequenced genome. In an attempt to quantify and compare global patterns of copy number change three mathematical measures, namely G-stat, H-stat and the combined GH index are introduced. The G-stat and H-stat are calculated from the relative heights and positions of the peaks.

G represents the part of the cancer genome that sits below the baseline, and is essentially a measure of genomic loss. Intuitively, high G values should be associated with worse prognosis. Higher peaks indicate occurrence of more genomic loss and this could be associated with more significant loss of tumour suppressor genes and therefore their role in inhibiting carcinogenesis. However, G on its own can’t be considered as an adequate prediction tool as it fails to incorporate a number of events along the genome, which could be critical to the biological behaviour of the cancers. To address this issue the H stat was introduced. It serves to reflect the complexity of copy number changes along the entire genome and attempts to quantify tumour homo/heterogeneity. It is calculated from the relative heights of the two tallest peaks on the density plots. The height of the second tallest peak is divided by the height of the tallest peak to give its numerical value. Therefore genomes with one prominent peak and several smaller ones on the density plot will have a low absolute value of H, while smaller height differences between the peaks with yield a higher H value. From a biological point of view, tumours which are more heterogeneous (lower H values) are likely to carry greater malignant potential i.e. be more likely to recur or metastasize. Heterogenous tumours tend to have multiple loci with intermediate CNV, which on a density plot will be represented by one prominent peak and multiple lower ones.

The GH index attempts to combine the two stats and provide a more precise account of the copy number variations occurring in the cancer genome and correlate them with clinical features. This is a novel index, which has so far been applied in a very limited setting. However, our previous study (Belvedere *et al,* 2011) showed that combining the G and H stat in the GH index [G x (1-H)], led to an improved p-value in predicting survival in SCC ($p=0.003$), significantly exceeding that of G ($p=0.18$) and H ($p=0.09$) on their own.

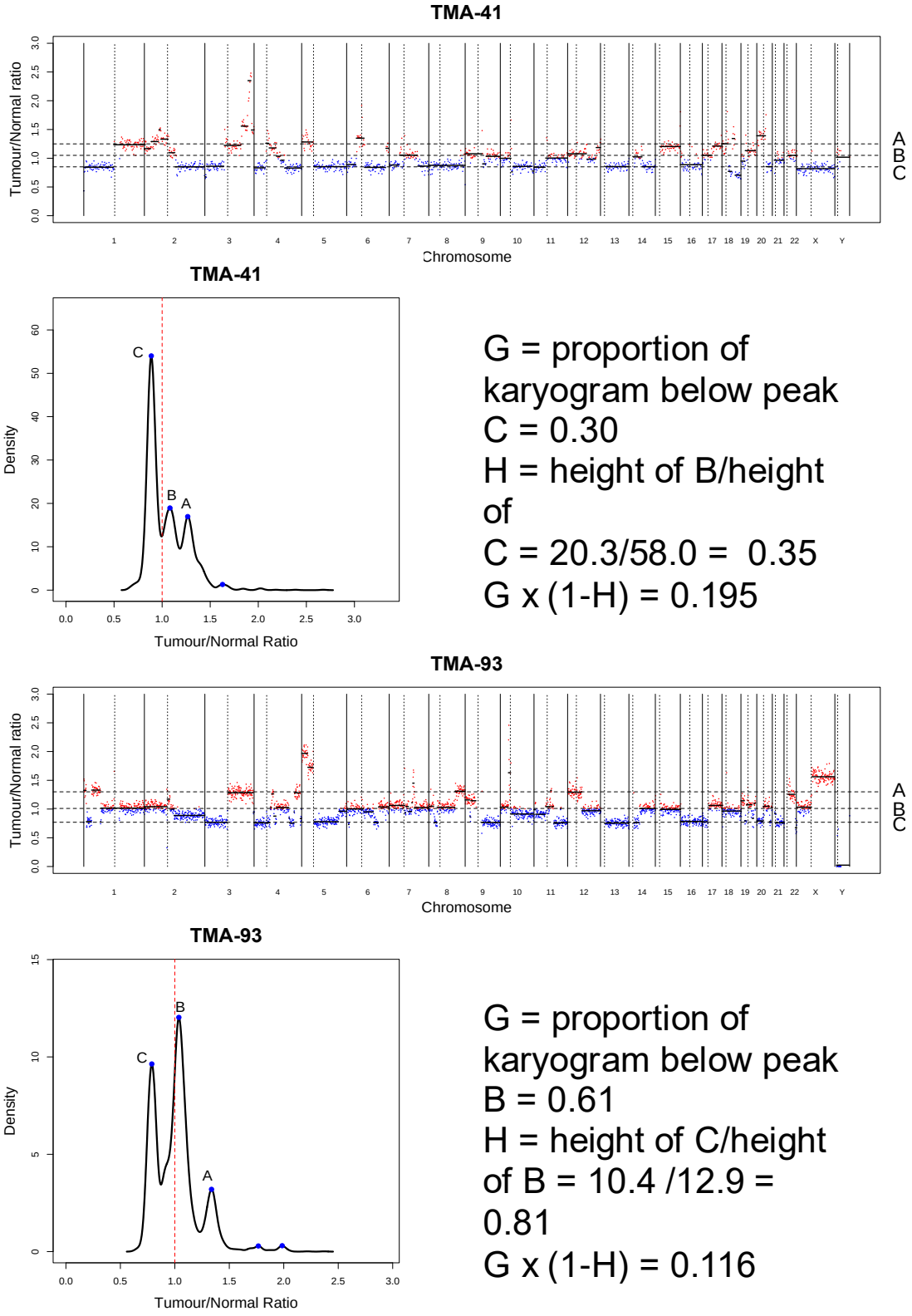


Figure 2.5. Defining G-stat and H-stat based on patterns of genomic gain and loss. For each genome, a density plot was drawn and peaks, reflecting the level of activity (gain and/or loss) along the genome were identified. To calculate the G-stat, the proportion of the genome with copy number less than the highest peak was measured. It corresponds to the number of dots that fall to the left of the red dotted line (density plots) as a fraction of the total. To measure the H-stat, the ratio of the heights of the two highest peaks (second over first) was calculated. For sample TMA-41 G = 0.30, H = 0.35. GH = G(1 − H) = 0.195.

2.8.3 Logistic regression model

For the purpose of this study we wanted to model an essentially binary outcome (recurrent versus non-recurrent cancer) as a function of an independent variable, namely the genomic signature of the cancer, based on its CNV. Regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). Logistic regression models the relationship between a dependent and one or more independent variables, and the significance of the relationships (between dependent and independent variables) that are modelled. Logistic regression estimates the probability of an event occurring (for the purposes of this study - the probability of stage I NSCLC recurring within 5 years from radical surgery). The aim is to predict from a knowledge of relevant independent variables the probability (*p*) that it is 1 (event occurring) rather than 0 (event not occurring). The binary response for this study was defined as: Recurrent (value=1) and Non-recurrent (value=0). The relationship between the dependent and the independent variables is non-linear. The aim of this analysis in the context of the study is to predict whether a tumour will recur or not, based on the patients' CNV profiles and/or their clinical characteristics (Age and Gender). The probability (p) estimates the likelihood of tumour recurrence AND non-recurrence at the same time. For example, if the probability of recurrence for a tumour based on its CNV is 0.7, then the probability for it not to recur is just 0.3.

Following the alignment, an average window size of 150 kbp was selected. After excluding the X and Y chromosomes and the centromere regions, we ended up with approximately 17 000 windows per sample. The windows were aligned with a common control (16 normal genomes from the 100 genome project: 8 male and 8 female) and to identify the copy number a ratio of the number of reads (between cancer reads and control reads) is calculated for each window.

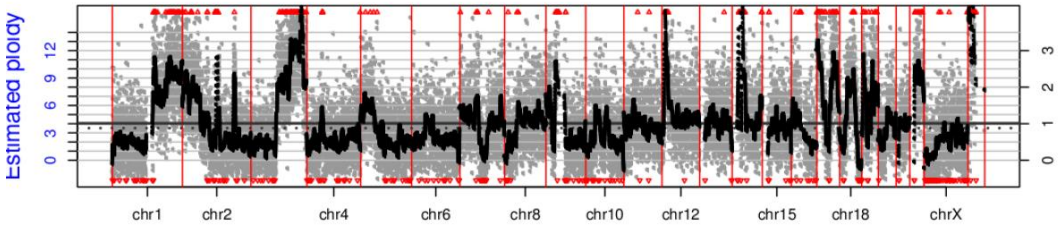
Determining the size of the window has been a somewhat arbitrary issue in the past. If a window is too small (e.g. 1-5 reads), a significant number of genomic regions will have a "zero" read count, while a larger window will have the tendency to "smooth out" discrete pattern of alterations, which could be significant (Gusnanto *et al,* 2012). However in a recent analysis Gusnanto *et al,* 2014) identified an algorithm for determining an optimal window size for CNV analysis of data from high-throughput sequencing. This

is done with a specifically designed software package, written in R, which was used in this study.

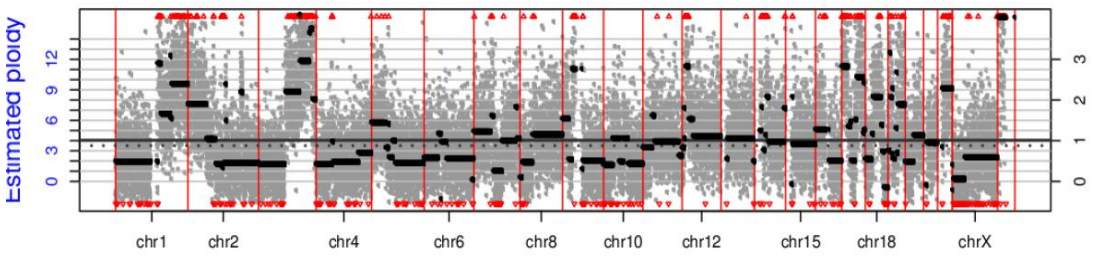
Each window was incorporated as a separate variable in the analysis of the data.

The sequence data from all of the samples could not be compared directly. The reason for this was the varying degree of “contamination” of each tumour with normal cells. To tackle this issue the CNV profile for each tumour sample was calculated using CNAnorm and the data was then normalized using two different forms of segmentation. Segmentation is a process of determining the positions in the cancer genome, where the copy number jumps from one value to another. It splits the genome into different segments where the CNV are distinctly different from its neighbours. It essentially serves to resolve the problem of the background “noise” created by sample contamination and provide a way to normalize and provide a graphic representation of the vast amount of sequencing data. The two different types of segmentation utilized were smooth segmentation and circular binary segmentation.

1. Smooth segmentation – CNV were estimated as a smooth segmented line, which follows the jumps (amplifications) and drops (deletions) in copy number profile



2. DNAcopy – CNV were estimated as circular binary segmented lines. The distinguishing feature of this method is the formation of relatively long, constant segments.



Segmentation allows for the data to be visualized in a graphic manner. The calculated lines serve to identify significant “clusters” of genomic changes, essentially allowing us to “connect the dots” in a way thatgives us an image.

In this analysis, the logistic regression model was fit to the SCC and AC data. Two covariates in the data (age and gender) were taken into account in the analysis.

The analyses of the CNV estimates across 17,571 genomic windows were as follows:

1. Regressing the Recurrence status (Recurrent or Non-Recurrent) as a function of the (fixed) covariates, i.e. Age and Gender only.
2. Regressing the Recurrence status as a function of the CNV profiles of the patients, excluding the fixed covariates (Age and Gender).
3. Regressing the Recurrence status as a function of the fixed covariates and the copy number profiles.

To obtain the prediction of probability of recurrence in this cohort, based on the CNV profiles, a cross validation was performed. The data was randomly split into an estimation set and validation set. Half of the observations (recurrence vs non-recurrence) for each cohort were included in the estimation set and the other half in the validation set. The process of cross validation was performed 100 times with different random selection of samples from the estimation and validation sets.

For a validation of the results of the logistic regression analysis the model was tested in two ways:

1. On tumour data from Ng-seq of 76 patients with NSCLC. Out of them 38 were patients with SCC and 38 with AC. The logistic regression model was used to differentiate histological subtypes (SCC from AC) based on their CN profiles.
2. Each sample of the study was randomly assigned a consecutive number (1-38 for the SCC and 1-48 for the AC cohort). The logistic regression model was fitted in an attempt to differentiate samples with odd numbers from samples with even numbers based on their CN profiles.

Several statistical methods were considered before settling on logistic regression, namely diagonal quadratic discriminant analysis, diagonal linear discriminant analysis and parietal least squares analysis. The data was

tested in both histological cohorts (AC and SCC) with both modes of segmentation (smooth and DNAcopy). The logistic regression model showed best “fit” with most homogenous data distribution in virtually all of the combinations.

While the author acknowledges that sample size determination is an important aspect in the design of a research study and a major step in defining the statistical power, it was felt that it is not appropriate for this work. This main reason for this decision was the limited sample cohort. Three hundred and twenty three tumour samples were available and the original surgery was performed in the same department by a particular group of surgeons, conforming to a certain clinical pathway. A significant amount of time had passed from the time of the original operations, which was invaluable in accurately determining recurrence status, but made it difficult to add new samples to the cohort. After several consultations with a statistician and bio-informatician a decision was reached that a power calculation would not be suitable due to the fact that regardless of the result all samples, which were identified as suitable for the study, would be used in the data analysis. This is essentially a pilot retrospective study, designed to serve a platform for future work. Working within the confines of available resources made it prudent to use all possible samples within the investigated cohort.

Chapter 3
Data collection and DNA extraction

3.1 Assembling patient cohort

Multiple databases were used to obtain complete clinical data and outcome information for the cases. This was necessitated by the retrospective nature of the study and the relatively long time that had passed since the original operation. The patients, from whom the samples were obtained, underwent the primary surgery for NSCLC between 1999 and 2003. Whilst this allowed adequate amount of time to have passed since the initial treatment to assess survival and recurrence, it created difficulties with data gathering. Currently, the PACS radiology system and histopathology database (Co-path) used in the Leeds Teaching Hospitals Trust are synchronized with the Patient Pathway Management (PPM) database to facilitate the management of cancer patients. This is valid for all data accumulated after 2009 and will likely prove very valuable for future studies by greatly facilitating access to a wide array of data. For the purpose of this study PPM was used to obtain survival data and evidence of recurrence. The latter was confirmed by gathering information from the PACS system (radiological evidence for recurrence) and Yorkshire Cancer Registry (cause of death). The original histology reports were carefully examined on Co-path to ensure that an adequate number of lymph nodes was submitted to allow for precise staging. Samples without submitted N2 nodes, despite being considered as stage I, were excluded from the study. The information was compiled in a detailed database including date and type of surgery, histological staging, evidence of recurrence, concomitant cancer treatments and survival data. The cases suitable for the study were then identified with the defined criteria.

Data summarizing the demographics of the two cohorts and the type of surgical procedures performed, is presented in table 3.1.

Demographical and clinical characteristics of patients in SCC cohort

Parameter	n	%
Age at surgery, years		
Median	67.63	
Range	40 – 79	
Gender		
Male	21	52.63
Female	17	47.37
Type of surgery		
Lobectomy/bilobectomy	32	84.21
Pneumonectomy	6	15.79

Demographical and clinical characteristics of patients in AC cohort

Parameter	n	%
Age at surgery, years		
Median	68.94	
Range	52 – 83	
Gender		
Male	19	39.58
Female	29	60.42
Type of surgery		
Lobectomy/bilobectomy	45	93.75
Pneumonectomy	3	6.25

Table 3.1. Demographical and clinical characteristics of the patients

3.2 Squamous cell cohort

3.2.1 Sample selection and demographics

Out of the cohort of 172 squamous cell cancers, 77 samples were classified as stage I. However, only 49 (28.48%) were identified as suitable for the study after applying the entry criteria. In some cases a significant amount of time had passed from the original surgery and discrepancies were noted in the clinical data in different databases. All such samples, for which staging, surgery and/or survival data could be deemed uncertain, were excluded from the study. For example two cases were initially staged as T1N0, although the final histology report mentioned direct invasion by proximity of the tumour into an adjacent N1 lymph node. The reports were discussed with two independent pathologists, who both considered that the staging should be upgraded to T1N1, making the cancers stage II. The two samples were excluded from the study, despite clearing all of the entry criteria on the initial screening. Another case showed a suspected recurrence more than 5 years from the original surgery. It was also excluded from the study, due to the fact that no histological confirmation was obtained and the “recurrence” could be an independent new primary. One case, which underwent the initial lung resection in 1999 was classified in the final pathology report as adenosquamous and was also deemed unsuitable for inclusion. Four of the patients, whose samples were classified as stage I had undergone sublobar resections (wedge resection) and were excluded.

Out of the remaining 49 stage I squamous cell lung samples, 24 were cases of recurrent cancer and 25 of non-recurrent. Tumour DNA was successfully sequenced in 38 (77.56%) of the cases. Eighteen were in the recurrent arm (12 male, 6 female mean age 75.06 years) and 20 in the non-recurrent arm (9 male, 11 female, mean age 64.55 years). The cancer recurrence occurred at a mean time of 16.33 (6 - 39) months from the date of surgery. In 10 of the cases the recurrence was local, while in the remaining 8 it presented as metastasis in more distant locations such as bone (4 cases), brain (1 case), liver (1 case) and abdominal wall (1 case). The 11 cases that failed the sequencing process had passed the quality control with Nanodrop and Picogreen, but failed to produce usable libraries. A repeat DNA extraction was performed, although this did not yield a different outcome.

3.2.2 Tumour area and tumour cell content

The mean tumour cell area was estimated at 62.63% (20-80%). The mean tumour cell content of the samples was 69.73% (30-90%). Two of the samples had sizable zones of inflammation and five of extensive cell necrosis within or adjacent to the tumour area. These were also marked out to facilitate the dissection of the actual cancer cells.

3.2.3 Quality control

Prior to preparation of DNA libraries from the samples, quality control was performed with Nanodrop and Picogreen to confirm the extraction. The Nanodrop analysis was performed immediately upon completion of the DNA extraction (Figure 2.3) via the Qiagen protocol, whilst the Picogreen was carried out upon the completion of DNA extraction of the entire cohort, due to the nature of the technique.

Two measurements for each sample were made using the Nanodrop.

None of the samples in either SCC or AC cohort were excluded based on their Nanodrop or Picogreen readings.

Figure 3.1. Example of Nanodrop run in the SCC cohort. The nucleic acid concentration in each sample is highlighted in green (far right). The graphs representing the 260/280 ratio can be seen next to the sample number and are clearly demonstrating a peak at the appropriate wavelength.

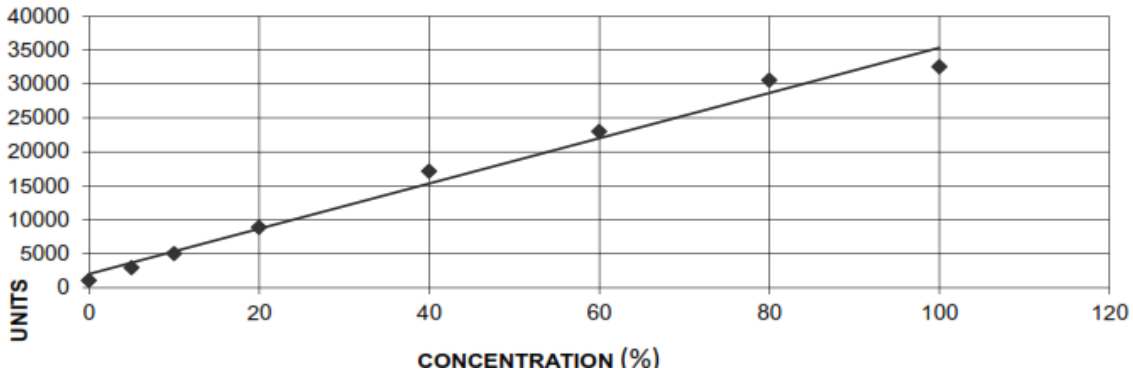


Figure 3.2. Example of Picogreen run in the AC cohort. The straight line represents the established standard and the black dots are the investigated samples. The closer they are to the standard, the more pure is the DNA in the sample.

A summary of the quality control of the cohort is presented in Table 3.2.

The mean value for nucleic acid content measured by Nanodrop in elution 1 in the SCC cohort was 435.73-ng/μl (112.65 - 814.4-ng/μl). The 260/280 ratio was also calculated for the samples. DNA and RNA absorb at 260nm, while proteins absorb at 280nm and the ratio can be used as a measure of purity. The 260/280 ratio was 1.98 for elution 1 and 1.92 for elution 2.

The mean value for nucleic acid content measured by Picogreen in elution 1 in the SCC cohort was 87.53-ng/μl (18.62 – 149.18-ng/μl).

Patient Id	Tumour Area (%)	Tumour cell content (%)	Nanodrop elution 1 concentration ng/µl - read 1	Nanodrop elution 1 concentration ng/µl - read 2	Mean elution 1 (ng/µl)	Picogreen elution 1 ng/µl	Picogreen standard deviation
LS4	60	70	391.6	441.8	416.7	39.63	0.73
LS6	90	95	184	189.6	186.8	18.61	11.03
LS34	40	50	505.2	419.9	462.55	98.85	0.41
LS39	70	80	382.5	346.8	364.65	61.68	0.67
LS60	50	60	521.8	634.4	578.1	102.69	1.03
LS64	20	30	304.9	312.9	308.9	57.14	4.51
LS80	75	80	719.3	743.8	731.6	102.47	0.88
LS88	40	60	533.2	581.7	557.5	98.09	0.15
LS91	40	60	638.5	631.0	634.8	149.18	1.33
LS95	80	90	362.7	472.0	417.4	98.83	3.93
LS97	65	70	282.5	286.3	284.4	114.33	2.77
LS98	50	60	361.9	486.2	424.1	81.11	1.57
LS113	80	70	265.9	298.2	282.1	77.16	3.25
LS121	50	70	369.3	458.6	414.0	93.89	1.74
LS127	70	85	315.1	315.3	315.2	85.67	3.71
LS129	75	85	541.3	545.6	543.5	77.86	7.44
LS130	75	85	309.0	220.0	264.5	38.82	3.17
LS143	70	70	553.0	482.4	517.7	54.27	1.92
LS146	85	80	243.6	124.0	183.8	60.00	9.40
LS147	80	90	357.0	355.0	356.0	73.83	0.69
LS20	70	80	248.8	199.1	223.95	68.00	0.98
LS25	50	50	808.4	779.8	794.1	69.89	0.61
LS37	80	70	112.6	112.7	112.65	78.51	0.62
LS61	80	90	824.4	804.4	814.4	128.85	10.19
LS63	40	50	659.3	639.3	649.3	131.28	5.99

LS86	70	80	364.9	345.4	355.2	98.69	3.98
LS93	80	80	436.5	431.3	433.9	113.97	2.25
LS96	65	70	550.0	559.9	555.0	103.31	0.81
LS122	60	70	513.4	484.4	498.9	53.39	0.89
LS162	40	55	270.4	382.1	326.3	94.40	0.15
LS171	65	85	422.4	466.2	444.3	76.84	22.06
LS40	85	95	466.7	373.5	420.1	130.15	0.56
LS41	70	85	643.8	632.7	638.25	97.59	4.75
LS74	60	60	512.1	495.3	503.7	98.54	1.54
LS160	50	30	159.4	249.1	204.3	57.89	0.99
LS172	40	30	420.6	426.7	423.7	80.04	4.65
LS33	60	70	485.1	484.4	484.75	150.95	4.46
LS84	50	60	402.2	459.1	430.7	109.79	1.66

Table 3.2 Summary of tumour area, Nanodrop and Picogreen values for the SCC samples. The tumour cell area and cell content were determined by a pathologist

3.3 Adenocarcinoma cohort (151)

3.3.1 Sample selection and demographics

Out of the cohort of 151 lung adenocarcinomas, 50 (33.11%) were identified as suitable for the study after applying the criteria. An occasional non-conformity between the databases as observed in the squamous cell cohort was noted in this group as well. Five samples had to be excluded because, although initially classified as adenocarcinomas (by pre-operative biopsy), the final histology report classified them as squamous cell (2 cases), large cell (2 cases) and NSCLC (1 case). The final histology of another interesting case, originally operated on in 1999, showed two independent small adenocarcinomas in the removed lobe, both staged at T1N0. This sample was also excluded from the final cohorts as it can no longer be considered stage I with the current edition of the TNM system. A further sample had to excluded, because the patient from whom it was obtained died very close to the five year margin and the cause of death was too ambiguous (pneumonia unspecified).

Out of the remaining 50 stage I lung adenocarcinomas, 25 were cases of recurrent cancer and 25 of non-recurrent. The cancer recurrence occurred at

a mean time of 26.16 (4 - 58) months from the date of surgery. Tumour DNA was successfully sequenced in 48 (31.79 %) of the cases. Twenty three of them were in the recurrent arm (9 male, 14 female mean age 68.61 years) and twenty five in the non-recurrent arm (10 male, 15 female, mean age 69.24 years). In 17 of the cases the recurrence was local, while in the remaining 6 it presented as metastasis in more distant locations such as brain (3 cases), contralateral lung (2 cases) and spine (1 case). Initially 11 cases (22%) failed the sequencing process, despite passing the quality control. A repeat DNA extraction was undertaken and out of the 11 only two failed to produce libraries.

3.3.2 Tumour area and tumour cell content

The mean tumour cell area was estimated at 65% (25-95%). The mean tumour cell content of the samples was 63.51% (20-90%). Areas of chronic inflammation were more common in this group (in 11 of the cases). Zones of extensive necrosis were present in 2 of the samples.

3.3.3 Quality control

Two measurements for each sample were made using the Nanodrop. Four samples in the AC cohort showed significant discrepancies between their two respective readings and an extremely abnormal appearance of the graph representing the 260/280 ratio. A repeat DNA extraction was performed from the four tumour blocks and the Nanodrop was repeated. Readings uniform with the values obtained in the rest of the cohort were recorded. The aberrant results were attributed to a technical error in the sequence of buffer application in the Qiagen protocol.

A summary of the quality control of the cohort is presented in Table 3.3.

The mean value for nucleic acid content measured by Nanodrop in elution 1 in the AC cohort was 212.12-ng/μl (32.08 – 571.06-ng/μl). This was significantly lower than the SCC cohort. The 260/280 ratio was also calculated for the samples.

The mean value for nucleic acid content measured by Picogreen in elution 1 in the SCC cohort was 36.9-ng/μl (2.7 – 134-ng/μl).

Patient Id	Tumour Area (%)	Tumour cell content (%)	Nanodrop elution 1 concentration ng/μl - read 1	Nanodrop elution 1 concentration ng/μl - read 2	Mean elution 1 (ng/μl)	Picogreen elution 1 ng/μl
LA 170	30	60	32.51	32.08	32.295	17.5
LA 87	85	75	126.11	128.24	127.175	19
LA 95	60	45	100.8	102.88	101.84	20
LA 134	40	50	121.59	122.33	121.96	24
LA 99	70	60	54.16	56.55	55.355	2.5
LA 81	70	70	140.42	95.81	118.115	2.7
LA 172	90	80	491.88	571.06	531.47	127
LA 50	60	70	142.55	142.9	142.725	19
LA 84	25	80	126.54	188.7	157.62	17
LA 78	60	65	28.31	32.54	30.425	5
LA 158	60	70	251.83	254.04	252.935	43
LA 115	85	40	202.39	196.19	199.29	53
LA 11	80	75	335.23	370.2	352.715	105.5
LA 137	80	50	109.24	105.11	107.175	31
LA 25	50	50	65.88	65.7	65.79	5.4
LA 28	80	70	188.9	184.18	186.54	33
LA 1	85	60	141.63	142.01	141.82	15
LA 65	100	60	46.97	46.94	46.955	4
LA 3	25	50	49.08	46.66	47.87	5.8
LA 80	90	70	366.11	358.56	362.335	34
LA 14	95	80	884.06	302.65	593.355	73
LA 59	90	60	336.39	396.17	366.28	68
LA 178	70	85	159.09	156.41	157.75	43
LA 169	70	50	243.24	239.82	241.53	53
LA 5	75	70	465.71	452.18	458.945	46

LA 4	40	80	192.94	186.5	189.72	33
LA 10	45	80	114.24	113.25	113.745	16
LA 146	70	70	420.08	415.15	417.615	55
LA 135	75	60	461.64	484.29	472.965	62
LA 34	40	20	186.59	191.41	189	28
LA 33	75	90	247.26	248.64	247.95	65
LA 37	30	60	60.59	57	58.795	7.3
LA 74	95	75	213.66	212.11	212.885	46
LA 68	50	70	383.42	363.39	373.405	44
LA 61	45	40	23.49	25.12	24.305	4.5
LA 83	95	65	437.94	467.92	452.93	81
LA 56	70	30	255.74	Error	255.74	55
LA 57	50	40	209.59	209.46	209.525	134
LA 127	40	80	66.5	61.28	63.89	10
LA 104	40	70	106.64	101.95	104.295	15
LA 121	85	60	121.57	120.7	121.135	34
LA 122	45	60	99.45	94.7	97.075	31
LA 149	70	80	165.4	161.81	163.605	23
LA 152	20	50	139.27	132.86	136.065	19
LA 153	65	50	218.31	217.54	217.925	41
LA 160	75	80	323.31	317.94	320.625	35
LA 69	80	70	534.56	536.27	535.415	19
LA 73	90	75	216.86	197.63	207.245	47

Table 3.3 Summary of tumour area, Nanodrop and Picogreen values for the AC samples. The tumour cell area and cell content were determined by a pathologist

3.4 DNA Library preparation

Libraries were prepared and no samples were discarded from sequencing based on their Nanodrop or Picogreen results.

Between 200-ng and 1-µg genomic DNA were used to prepare the DNA libraries for sequencing.

A total of 15 DNA samples failed the library prep (13 SCC and 2 AC) and were not sequenced, as they did not show enough library between 150-350bp as measured on an Agilent bioanalyser. In the past experience of our group, it was a common finding for samples, which failed their library prep to have more than sufficient amount of DNA according to Nanodrop and Picogreen readings. We believe that the failure of such samples to generate libraries is due to DNA damage caused by the fixing procedure i.e. the DNA has either physical or chemical damage which stops the various reagents from binding/annealing/ligating. In effect the DNA in the sample is simply not available.

CNV maps were generated for each sample. There was no significant change in preparation protocols between the two cohorts (AC and SCC) and no obvious reason for the disproportionally high failure rate to generate DNA libraries in the SCC group was established, such as longer time elapsed from the original surgery, higher necrosis area or lower tumour cell content. Three of the failed samples in the SCC cohort were deemed to have extensive areas of necrosis/haemorrhage by the pathologist. However, several samples in both cohorts with similar pathological findings went on to generate DNA libraries, which did not have any distinguishable features. Several of the failed samples were obtained prior to the year 2000, suggesting DNA degradation over time, but once again this was not a single distinguishing feature.

3.5 Discussion

One of the most difficult obstacles encountered during this project was collecting the appropriate clinical information for the samples and compiling a credible database. Whilst the results of this study are encouraging, their validation in a larger, independent cohort will be necessary before a real impact on clinical practice can be made. Obtaining such a cohort, comprising a large number of early stage NSCLC, with adequate representation of the major histological subtypes (AC, SCC, LCC) and reliable survival data might prove to be beyond the capabilities of a single thoracic surgery unit. Ideally, any validation of the results of this thesis should be performed in a multicentre setting in an aim to minimize bias. This is largely necessitated by two major factors:

Firstly - the nature of cancer recurrence. A retrospective review focusing on the subject requires a suitably long follow up period, during which a number of significant events relating to the patients' health can occur. The follow up has to be regular and well documented in order to be able to establish an adequate disease timeline. The absence of a unified approach to surveillance of NSCLC and the concerns stemming from this have already been discussed in this manuscript. As a result, studies can lose a significant number of otherwise suitable candidates to a wide range of issues – from inability to establish the time frame of the actual recurrence to "blank" gaps in the follow up (patients failing to attend, permanently moving address etc.). Co-morbidities and/or other malignancies can have a significant impact on survival and make subjects unsuitable for such a specific study, making recruitment of large numbers difficult. An additional issue is obtaining the definitive diagnosis of a recurrence. Histological confirmation is not always possible. Frequently, obtaining a tissue sample can be difficult and the final result can be ambiguous. For example, in a patient undergoing fine needle aspiration biopsy of a suspected tumour recurrence, if the histological result reveals adenocarcinoma, due to the small sample size differentiation between breast, colonic and lung origin can be difficult. If the past medical history includes more than one of these malignancies, such data might be insufficient. Invasive procedures in patients with limited life expectancy can be difficult to justify, particularly if little clinical benefit can be achieved.

Secondly – absence of a unified database. As the tumour samples, investigated in the study, were obtained more than 10 years ago (original operations were performed between 1999 and 2003) no universal record of all the relevant information exists. The final histology, cause of death, extent

of the operation, administration of adjuvant therapy, participation in trials were gathered from different sources. Radiology, pathology, oncology and surgery departments in the Leeds Teaching Hospitals at the time had completely separate databases. Since 2009 the Patient Pathway Manager (PPM) system has been installed as the default database in the Trust. It combines many key elements of other databases such as clinician letters, histology reports, radiology reports, surgery records, adjuvant therapy protocols, participation in clinical trials etc. A lot of information has been uploaded retrospectively, such as clinician letters from the late 90's. PPM and similar databases aim to improve cancer care by acting as a comprehensive repository of the diverse information, that accumulates prospectively. Whilst the system had still not reached its full potential for the purpose of this study, it promises to significantly decrease the workload for future researchers and data collectors, whilst at the same time providing very accurate information on disease progression and survival.

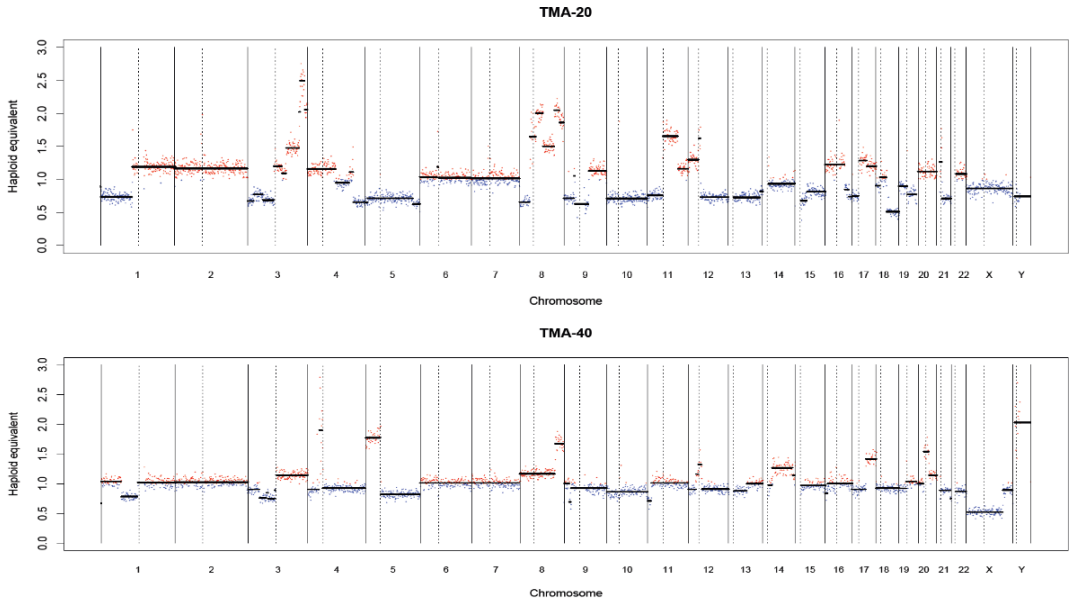
Perhaps there is a case to be made for a unified national cancer database, which will gather accurate information on extent of disease, staging and adjuvant therapy. Whilst for larger countries with more diverse healthcare systems such as the USA and China a task of this calibre can seem daunting, bearing in mind the significant advances in IT technology, such a project could be conceivable for the UK. This would allow researchers to increase the power of their work by providing access to larger study groups and significantly facilitating the identification of independent cohorts for result validation.

Chapter 4
Results: Comparative copy number maps and GH index

4.1 Models of global genomic patterns associated with recurrence

DNA sequence was obtained from 38 SCC and 48 AC. The demographics and surgical procedures are shown in Table 3. The mean read number was 1,030,660 per sample, ranging from 200,000 to 3,000,000. Using 300 reads per window for copy number analysis provided a resolution of approximately 900Kb. The number of breakpoints per sample ranged from 4 to 205.

Karyograms showing regions of gain and loss along the whole genome were generated for each sample. Karyograms exhibited several different types of copy number patterns, in terms of both the proportion of the genomes involved and the complexity of the damage. This ranged from whole chromosome gain and loss to very small but highly amplified regions (Figure 4.1, Appendix B).



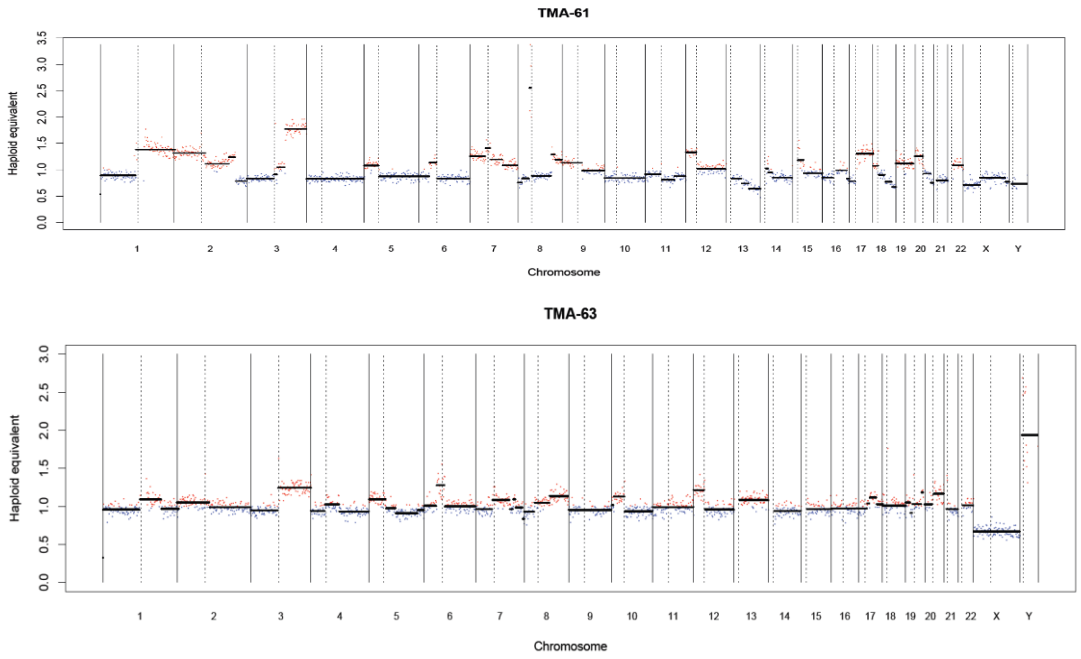


Figure 4.1. Karyograms showing copy number gain (red) and loss (blue) along the genome. 3q amplification is well demonstrated in samples TMA-20 and TMA-61. Samples are from the SCC cohort

4.1.1 Comparative CN maps

The frequency of CN gain and loss along the entire genome was compared between the non-recurrent and recurrent cancers using comparative CNV maps generated by CNAnorm.

4.1.1.1 Squamous cell cancer (figure 4.2)

The KC-SMARTR algorithm showed that no regions were significantly different for any comparison made. Most aCGH and NG seq analyses are performed on samples derived from tissue that contains sub-populations of different cells. This implies that an aCGH measurement will measure the average of CNV of different sub-populations within the sample. KC-SMARTR makes use of the continuous signal to preserve all the information contained in the data. It not only demonstrates aberrant areas along the genome, but it can identify abnormalities that are specific to subgroups within an investigated cohort (de Ronde *et al*, 2010). This made it particularly suitable to the purpose of this study.

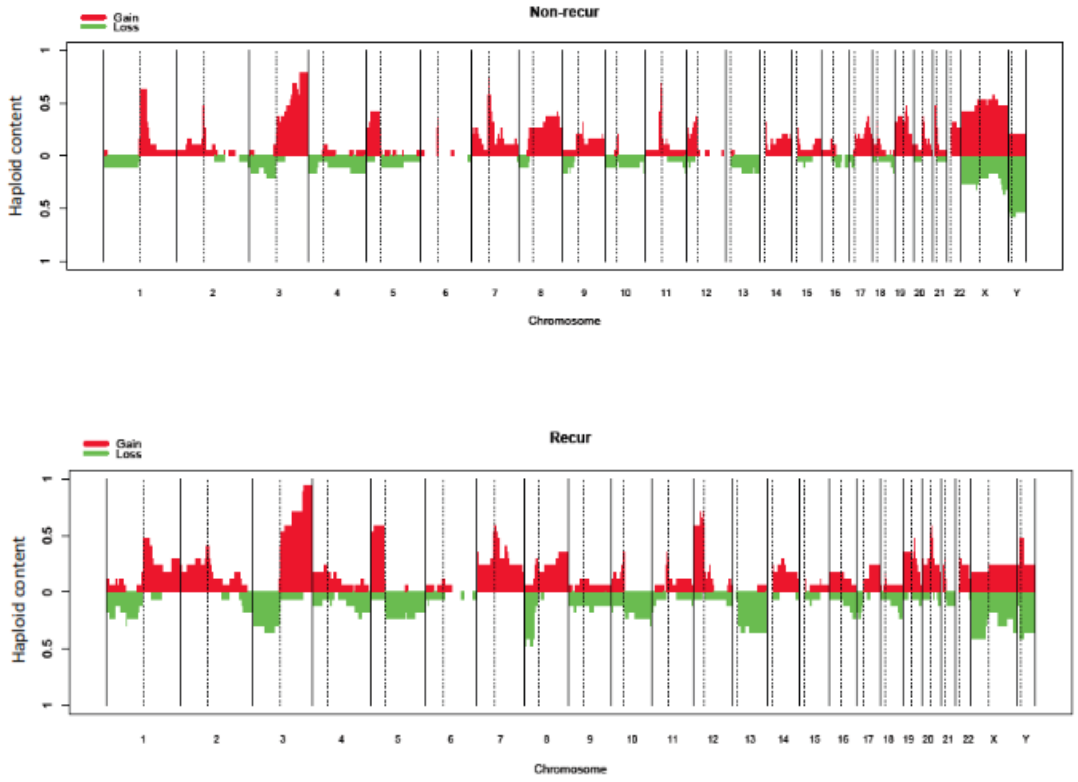


Figure 4.2. Comparative CN maps of the non-recurrent and recurrent cohort in the SCC group

4.1.1.2 Adenocarcinoma cohort (figure 4.3)

The comparative CN maps of the two cohorts were once again very similar. The only difference between the two sets was on the short arm of chromosome 6, where 1/3 of the non-recurrent had a gain but none of the recurrent had CN gain. Taken by itself this would have a significant p value of 0.02, but considering that the analysis comprised around 6000 data points along the genome, finding 120 points that are significantly altered (120/6000 = 0.02) could easily be attributed to random chance. When put through the KC-SMARTR program algorithm, no regions showed statistically significant difference in CNV between the recurrent and non-recurrent cohort.

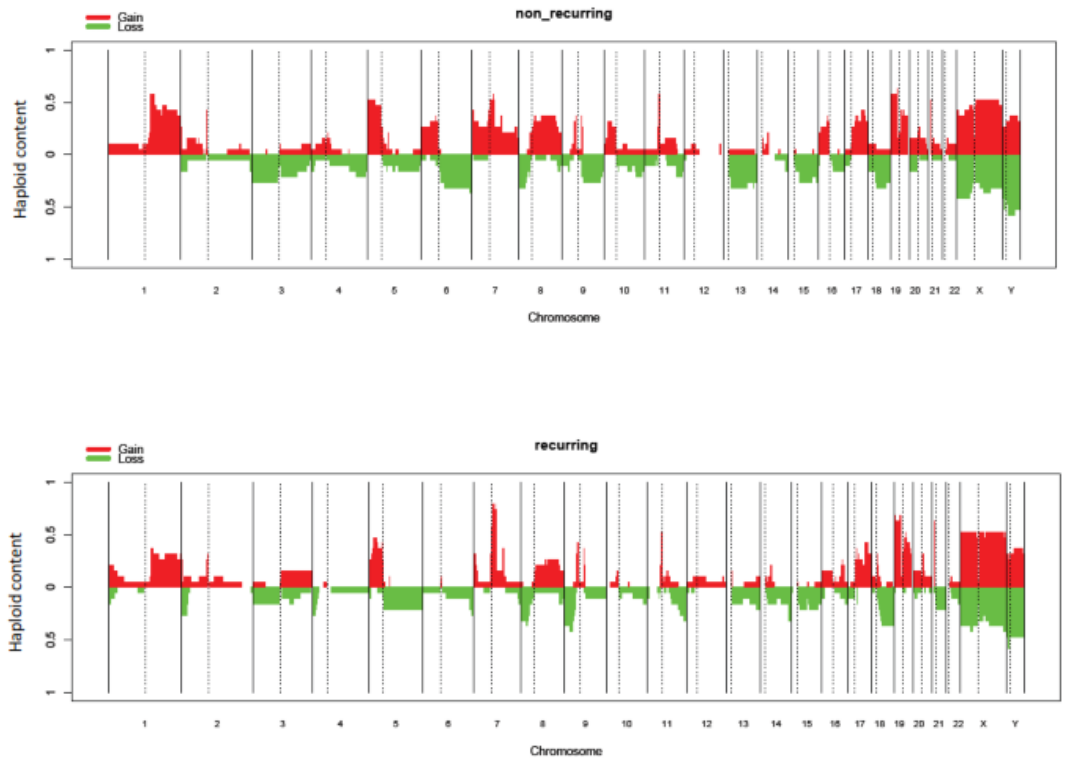


Figure 4.3. Comparative CN maps of the non-recurrent and recurrent cohort in the AC group.

4.1.2 CN patterns along the entire genome

In view of our inability to identify copy number changes at individual genomic *loci* associated with recurrence, I decided to apply the approach of Hicks *et al* and look for global patterns of copy number variation that might be associated with recurrence. This approach classified breast cancer genomes by patterns of damage named ‘simplex’ (few aberrations, mostly involving whole chromosome arms), ‘sawtooth’ (many aberrations spread throughout the genome) and ‘firestorm’ (local regions of intense, complex damage), and generated an algorithm for calculating an index of genomic damage, named F-stat, which was associated with survival in breast cancer. The cancers from our series did not easily fit into the Hicks method of classification, mostly being in a continuous spectrum of genomic damage somewhere between the simplex and sawtooth (Figure 4.4). However, there was no correlation between the CNV patterns and cancer recurrence in the investigated cohorts.

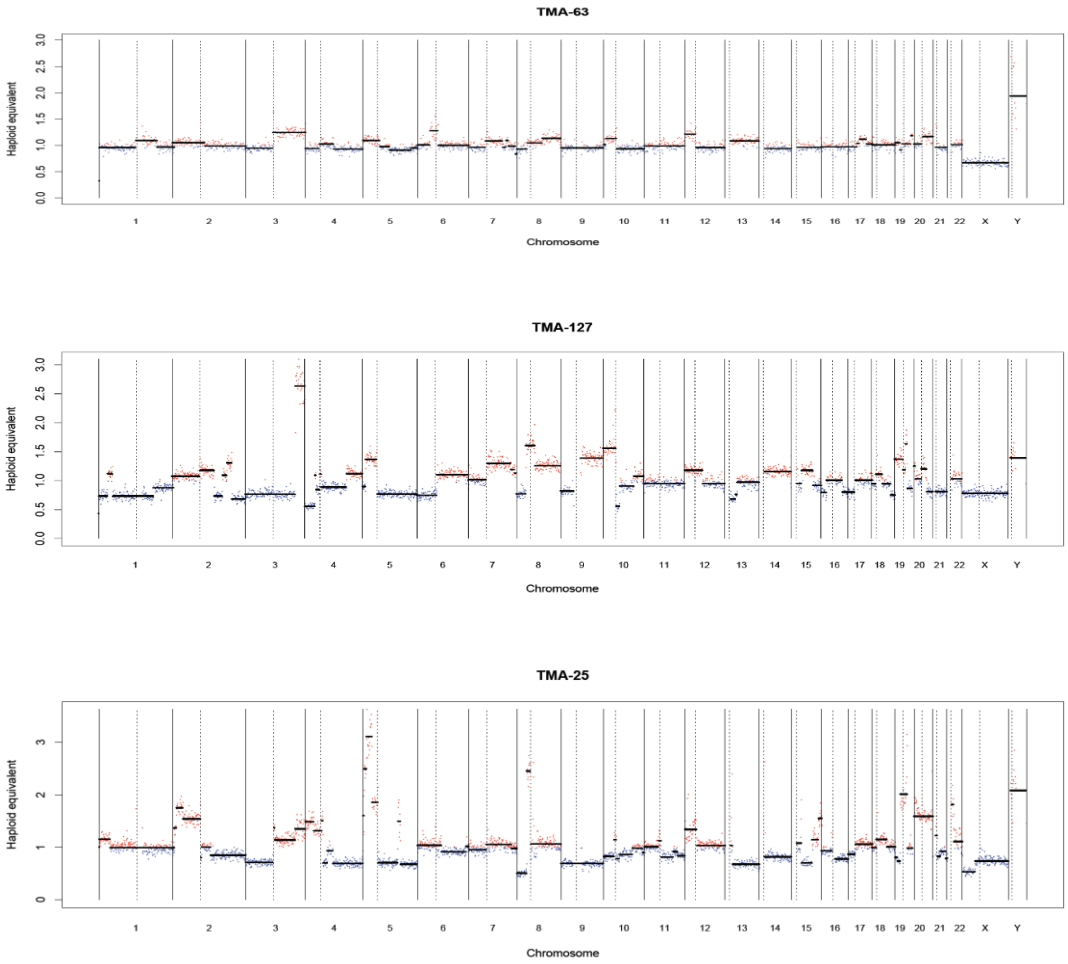


Figure 4.4. Hicks method of classification of genomic signature. Showing “flat” variant (TMA 63), “sawtooth” (TMA 127) and “firestorm” (TMA 25)

4.1.3 Pangenomic index (GH)

The GH index was developed as part of a parallel project questioning whether there is a correlation between CNV in SCC and survival (Belvedere *et al,* 2012). It was devised as an attempt to avoid the constraints of a difficult question – what exactly is the normal state for a cancer genome. The group considered that the traditional approach of regarding the median copy number ratio as ‘normal’ was not necessarily the most appropriate, as it assumes that each cancer has precisely the same amount of gain and loss. Density plots were drawn for the copy number distributions of each sample, with the relative heights of each peak representing the proportion of the genome at that copy number state (Figure 2.5 in Materials and Methods). The three mathematical measures, which had previously been used by our group to develop a prediction model for survival in SCC (Belvedere *et al,* 2011) were calculated for each sample (Table 4.1 and 4.2). An example of

the derivation of the GH index for an individual SCC case is shown in Figure 2.5 (Chapter 2 - Materials and Methods).

Study number	Recurrence status	G stat	H stat	GH index
LS4	R	0.71083071	0.50589573	0.35122449
LS6	N	0.24766098	0.97019642	0.00738118
LS20	R	0.24482925	0.73961243	0.06375049
LS25	R	0.51706308	0.98169198	0.0094664
LS34	R	0.57573943	0.61162551	0.22360251
LS37	R	0.50363714	0.69039788	0.15592713
LS39	R	0.57428571	0.50918262	0.28186941
LS40	R	0.63365222	0.6545467	0.21889725
LS41	N	0.30330189	0.35040911	0.19702214
LS60	N	0.35525071	0.67711092	0.11470658
LS61	R	0.31716418	0.31296719	0.2179022
LS63	R	0.29966076	0.5241449	0.1425951
LS64	N	0.70962099	0.98190515	0.01284049
LS74	N	0.57587391	0.84036645	0.0919288
LS80	R	0.51923077	0.95680084	0.02243033
LS84	R	0.62354189	0.68308377	0.19761055
LS86	N	0.6151743	0.4393707	0.34488474
LS88	R	0.5511811	0.05484348	0.52095241
LS91	N	0.24220167	0.97334977	0.00645473
LS93	R	0.6113114	0.80101635	0.12164097
LS95	N	0.60273973	0.58597544	0.24954905
LS96	N	0.54691689	0.96669411	0.01821555
LS97	N	0.65127701	0.9477977	0.03399816
LS98	R	0.38738128	0.52180489	0.18524383
LS113	N	0.47318148	0.08195194	0.43440334
LS121	N	0.22264265	0.79951929	0.04463556
LS122	N	0.51978573	0.85532959	0.07519762
LS127	R	0.20044114	0.93743444	0.01254071
LS129	N	0.24874698	0.89128248	0.02704316
LS130	N	0.20997709	0.83535179	0.03457235
LS143	N	0.47017319	0.96463618	0.01662712
LS146	N	0.22379644	0.7606187	0.05357268
LS147	N	0.50419776	0.8882907	0.05632358
LS160	N	0.32041999	0.96195226	0.01219126
LS162	R	0.61835245	0.99836312	0.00101217
LS171	R	0.38965517	0.39308955	0.2364858
LS172	N	0.41659312	0.19960276	0.33343998

Table 4.1. G-stat, H-stat and GH index values for SCC cohort

Study number	Recurrence status	G-stat	H-stat	GH index
LA 1	N	0.2691	0.6153	0.1035
LA 3	N	0.523346	0.13497	0.45271
LA 4	N	0.389552	0.187567	0.316485
LA 5	N	0.481481	0.580918	0.20178
LA 11	N	0.159197	0.975981	0.003824
LA 14	N	0.585366	0.227811	0.452013
LA 28	N	0.226863	0.406685	0.134602
LA 33	N	0.348148	0.023127	0.340097
LA 34	R	0.318445	0.956051	0.013995
LA 37	R	0.510345	0.103673	0.457436
LA 56	N	0.462598	0.84724	0.070666
LA 57	N	0.482645	0.664825	0.16177
LA 59	R	0.430776	0.739001	0.112432
LA 61	N	0.537375	0.086478	0.490904
LA 65	N	0.361337	0.133182	0.313213
LA 68	R	0.526112	0.158084	0.442942
LA 73	N	0.546914	0.791156	0.11422
LA 74	N	0.328652	0.1007	0.295556
LA 78	R	0.342501	0.891011	0.037329
LA 80	N	0.20403	0.896907	0.021034
LA 83	N	0.597855	0.845031	0.092649
LA 84	R	0.230961	0.972744	0.006295
LA 87	R	0.2584	0.1229	0.2266
LA 95	R	0.386915	0.111646	0.343717
LA 99	R	0.581197	0.958276	0.02425
LA 104	R	0.270566	0.960498	0.010688
LA 115	N	0.2366	0.7586	0.05712
LA 121	R	0.430343	0.264972	0.316314
LA 122	R	0.496228	0.98713	0.006386
LA 127	N	0.342367	0.781828	0.074695
LA 135	R	0.5489	0.9996	0.000238
LA 137	N	0.4992	0.9264	0.03673
LA 146	R	0.551819	0.826366	0.095815
LA 149	R	0.429405	0.264923	0.315646
LA 152	R	0.3811	0.3146	0.2612
LA 153	N	0.55814	0.884736	0.064334
LA 158	N	0.246243	0.227499	0.190223
LA 160	R	0.541126	0.101062	0.486438
LA 169	N	0.338575	0.130718	0.294317
LA 170	R	0.529412	0.106123	0.473229
LA 172	R	0.590847	0.785761	0.126583

Table 4.2. G-stat, H-stat and GH index values for AC cohort

In the SCC cohort, based on the computational index [G x (1-H)] the scattering of the two groups demonstrated an association between genomic signature and cancer recurrence (p=0.07 Chi-squared test) (4.5). Thirteen cases from the recurrent arm (18 cases) lie below the curved line, which represents the median GH value for all samples and 5 lie above it. In the non-recurrent arm 14 samples are scattered above the curved line and 6 below it. A high computational index, determined by a greater value of the H-stat seems to suggest that a cancer is less likely to recur. A higher absolute value of the G-stat in cases with similar values of the GH index hints at an increased recurrence risk. These results further suggest that patients who undergo radical surgery for stage I SCC could potentially be sub-stratified further, based on the CN variation along the entire genome, rather than in a single locus. Due to the fact that recurrence of the cancer in this particular patient group can differ significantly from the overall survival (several of the cases survived for more than two years after the original relapse) a survival analysis with Kaplan-Meier curves was deemed of little benefit to this particular study and was not performed.

In the AC cohort any obvious tendency towards clustering of the samples was not observed. The scattering, according to the values of the GH index seems random and not following any obvious pattern. Samples from both arms (recurrent vs. non-recurrent) were scattered virtually evenly above and below the curved line with a Chi-squared test for association between genomic signature and cancer recurrence 0.91. These findings support the concept that in lung cancer AC and SCC are in fact two very different biological entities, each with their own sets of genomic changes. The difference between the scatter graphs of the two histological subtypes highlights the difference of their genomic signatures and suggests that clinical differentiation in the management of these cancers could be important.

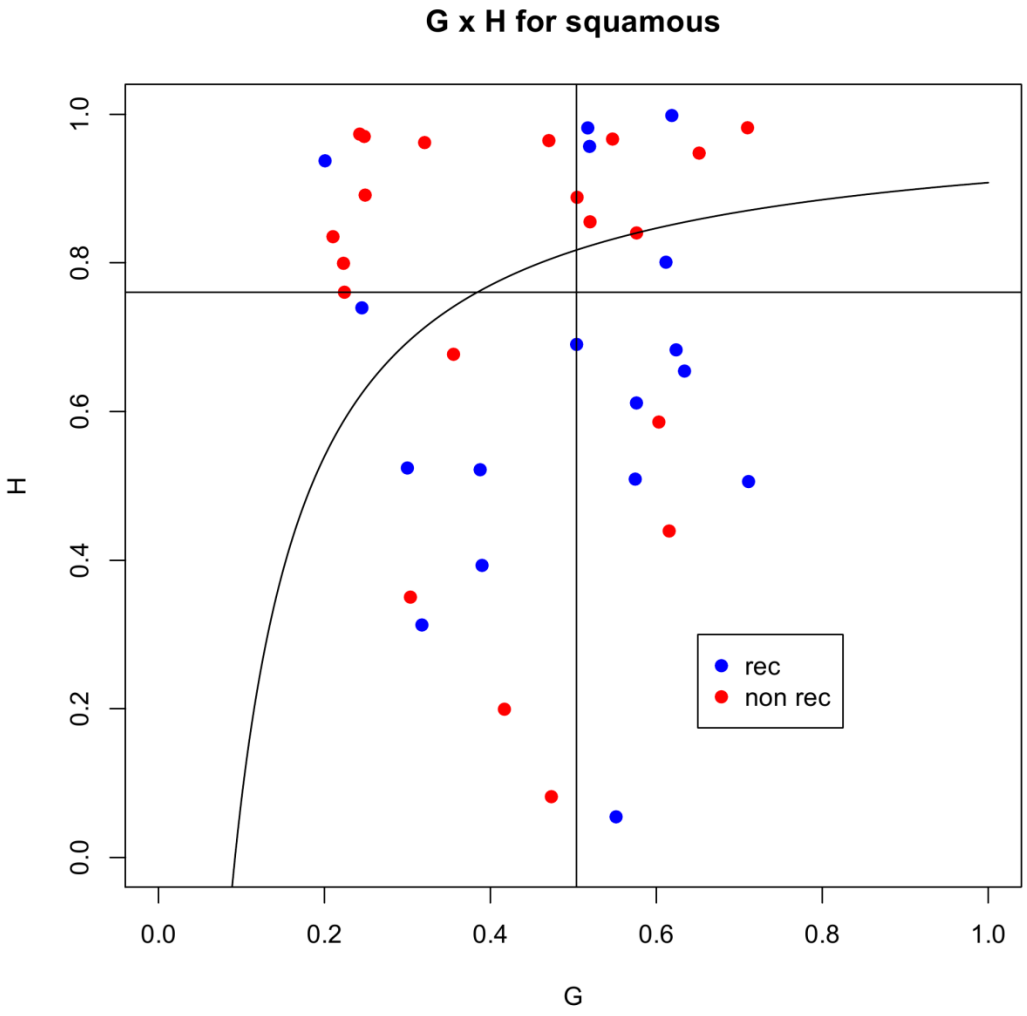


Figure 4.5. Scattering of recurrent (blue) and non-recurrent (red) cancers based on novel pangenomic computational index in the SCC cohort. The vertical line represents the median value for the G-stat and the horizontal for the H-stat. The curve represents the median value for the GH index. The non-recurrent tumours display a tendency to group above the curve, while the recurrent tend to group below the curve.

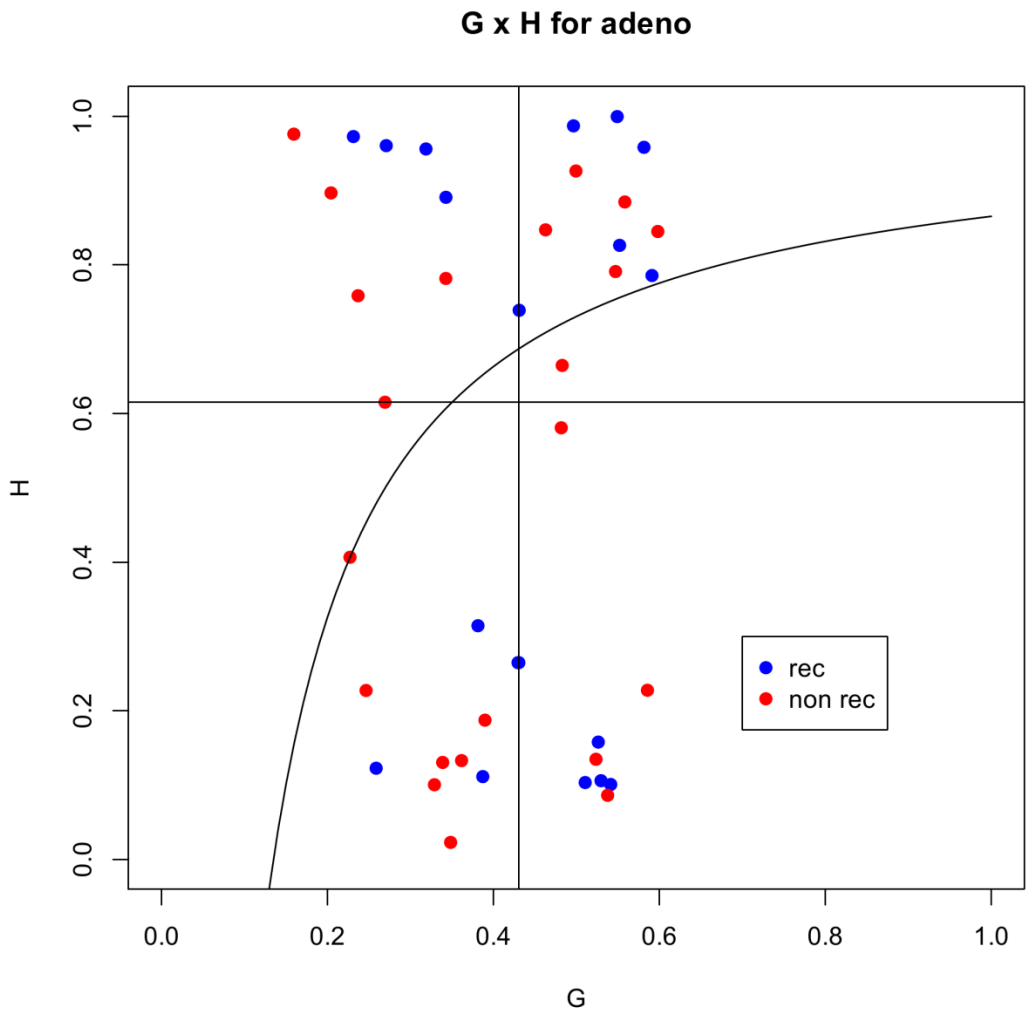


Figure 4.6. Scattering of recurrent (red) and non-recurrent (black) cancers base on novel pangenomic computational index in the AC cohort. There seems to be no obvious clustering according to recurrence based on the GH index.

4.2 Discussion

Over the past few years we have witnessed significant advances in the understanding of cancer genetics and genome biology. This has coincided with a revolution in sequencing technologies, which have become widely implemented in research settings.

The introduction of NG-seq has helped overcome the inherent deficiencies of Sanger-based sequencing platforms such as low throughput, speed and resolution, while at the same time improving cost-effectiveness. Whilst in principle the concept of the two platforms is very similar (the bases of small fragments of DNA are sequentially identified from emitted signals and each fragment is re-synthesized), NG-seq extends this process into millions of

massively parallel reactions, rather than being limited to a small number of DNA fragments. From a clinical perspective there is great potential for NGS in the management and treatment of human health and particularly cancer (Meldrum *et al,* 2011). Perhaps in the not too distant future all patients undergoing cancer treatment with have tumour genomes sequenced in order to individualize therapy and surveillance.

CNV, a form of structural variation, are alterations of the DNA of a genome that results in the cell having an abnormal or, for certain genes, a normal variation in the number of copies of one or more sections of the DNA.

CNVs can be caused by structural rearrangements of the genome such as deletions, duplications, inversions, and translocations. While a lot of CNV occur naturally in the genome as part of the genetic heritage and are naturally stable, they can also arise *de novo* at various stages of the development. Since CNV can correspond with gene expression changes, they may have important roles both in cancer development and drug response. NG-seq has further enabled the identification of CNV in a large scale, cost effective fashion. Several recent studies have shown that genome abnormalities in CN are likely to exert an influence in determining patient prognosis in NSCLC (Broet *et al,* 2009, Staaf *et al,* 2012). Broet *et al* described an integrative genomic prediction model for survival in stage IB NSCLC strategy by combining information about recurrent CNV with genes exhibiting copy number–dependent expression. These reports and previous prediction models for recurrence based on gene expression (D'Amico *et al,* 1999, Chen *et al,* 2007) strongly suggest that patients with stage I NSCLC can be further sub-stratified in prognostic and/or therapeutic groups based on the likelihood of recurrence. Based on these results, this study attempted to produce a model capable of differentiating between the genomic signatures of recurrent and non-recurrent stage I NSCLC, focusing on the two most common histological subtypes, AC and SCC.

Initially, high-resolution karyograms showing the CNV along the entire genome of each tumour sample were generated. By drawing on previous experience with gene expressions models, we speculated that by creating a "cumulative" karyogram for the recurrent and non-recurrent cohorts, individual regional differences along the genome would be exposed, thus allowing a correlation with recurrence. However, no single such region of genomic change was identified in either the AC or the SCC cohort. The reason for this initial failure probably lies in the complex nature of cancer recurrence and this could well explain why previously suggested prediction models have failed to make an impact on clinical practice. Based on the ever increasing insight into cancer genomics, it is reasonable to assume that

recurrence is defined by numerous genetic events, which occur along the entire genome (activation of proto-oncogenes and/or inactivation of tumour-suppressor genes) in different combinations. A single event (gain or loss) is not enough to trigger recurrence and perhaps every cancer recurs in a unique way, activating a number of possible pathological pathways in no predefined order. These genomic changes could occur at different levels, from mutations in single or multiple nucleotides to gains or losses of entire chromosomes. With this in mind, adopting a pan-genome approach seemed like a sensible next step. Applying a technique, recently described by Hicks *et al,* whole genome analysis, which associated specific appearance of the karyograms with survival in breast cancer, also failed to produce a differentiation between the recurrent and non-recurrent cancers. As a next step, an algorithm that relates to total genomic damage and specifically the relative ratios of CN states across the genome, previously presented by our group (Belvedere *et al,* 2011) was applied. This algorithm generated two variables, G stat, which is a measure of genomic loss and H stat, which is a measure of relative homogeneity and complexity of genomic damage. Combining these variables, a novel index was derived (GH), which was demonstrated to be an independent prognostic indicator for survival in early stage SCC. One of the problems that CNV analysis in cancers presents is how to establish a “normal” baseline, according to which the genomic changes (both gain and loss) will be evaluated. As seen on the karyograms produced for the samples in this study, there is a significant number of copy number abnormalities occurring in multiple places along the entire genome. Deviations from the “baseline”, which is traditionally established by calculating the median copy number for the genome, occur so frequently they put into doubt its significance as a reference point. This issue is made more complicated by the fact that taking a median value as a standard for CNV can only be truly justified if genomic gain and loss occur in equal measure. Bearing in mind the complex nature of cancer biology this is unlikely.

The GH index was developed as an attempt to resolve these problems. Its aim was to break away from the concept of a baseline derived from a median value and take into consideration the fact that there could be little balance between gain and loss in a cancer genome. In fact, one of them could prevail and that could be related to its malignancy, e.g. more gain could be associated with greater amplification of proto-oncogenes. Applying this algorithm to the sequencing data did not yield convincing results. Although the scatter graph for the SCC cohort hinted at a possible differentiation between recurrent and non-recurrent tumours, the results in the ACC cohort showed a virtually random distribution.

Perhaps, the failure of the Hicks and GH algorithms to clearly differentiate between recurrent and non-recurrent tumours suggests that a difference must be made between cancer recurrence and survival. This study targeted specifically early stage NSCLC due to its potential as a specific therapeutic group, while the GH index showed predictive value when applied on tumour samples with different TNM stages.

Chapter 5
Logistic regression model

For the purpose of this study there were only two possible clinical outcomes of interest for each sample. The cancer either recurred within the first five years after radical surgery or there was no recurrence. However, due to the significant number of CNV in the genomes of the investigated samples, an assumption was made that the changes in each genomic window (either gain or loss) could be considered as independent variables, which may determine one of the outcomes. To test this hypothesis, a logistic regression model was fitted on the sequencing data from the SCC and AC cohorts. The events in each genomic window (copy number gain and/or loss) for each tumour sample were considered an independent variable in the analysis. Essentially each genomic window was given a "score", which was determined by how closely the CNV occurring in that window was associated with the investigated clinical outcome – recurrence vs. non-recurrence. For example, if copy number gain (amplification), which was observed in a particular window and was present only in recurrent cancer cases, that window would register a very high score for predicting recurrence. If the events of this window were only seen in half of the cases with recurrence, then its score would be high to moderate, while if they also occurred in a number of the non-recurrent cases, then its score would be low. Two covariates in the data (Age and Gender) were also taken into account. Logistic regression was used to examine whether a correlation between cancer recurrence and the observed characteristics (variables) existed in the dataset. Several analyses were performed in order to test the model with other independent variables (Age and Gender) and minimize bias. The analyses that were carried out were as follows:

1. Regressing the Recurrence status (Recurrent vs. Non-Recurrent) as a function of the covariates Age and Gender only.
2. Regressing the Recurrence status as a function of the copy number alteration profiles of the patients, excluding the fixed covariates (Age and Sex)
3. Regressing the Recurrence status as a function of the fixed covariates and the copy number profiles.

Each of those analyses were carried out separately for both the SCC and AC data.

The CNV profiles were estimated using two different segmentation methods: smooth segmentation and discrete segmentation (DNAcopy). Segmentation was used to analyse CNV data by breaking up the windows into separate “segments” that differ from their neighbours based on the distribution of CNV. The sequencing data was analysed using each of the segmentation methods.

5.1 Regressing the Recurrence status (Recurrent vs. Non-Recurrent) as a function of the covariates Age and Gender only.

Cancer recurrence was analysed with the logistic regression model with recurrence status considered as a function of the variables Age and Gender (with no interaction between the two) (figure 5.1). There was no correlation between the two variables and the recurrence status. In Figure 5.1 each investigated cancer sample is represented by a single dot. The recurrent samples are given a value of 1 and the non-recurrent are given a value of zero. The vertical dotted line down the centre represents the probability of recurrence. Therefore, an optimal “fit” for the model would be a grouping of all dots with a value of 1 as far right of the dotted line as possible and all the dots with a value of 0 as far left as possible. The model showed no evidence of correlation between age and gender and cancer recurrence. There was no distinct pattern of grouping for the recurrent and non-recurrent cancers in each of the two investigated cohorts (SCC and AC). These findings point to the conclusion that age and gender as variables are unsuitable predictors of recurrence in NSCLC.

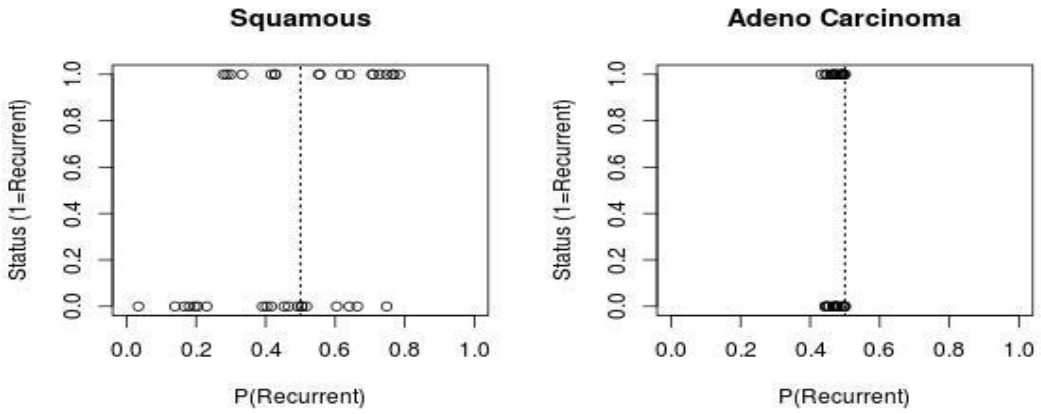


Figure 5.1. Classification of the patients' recurrence status when the variables Age and Gender are used as predictors. The horizontal axis demonstrates the estimated recurrence status, while the vertical – the actual observed recurrence status (1=recurrent, 0=non recurrent). The vertical dotted line represents a 50% probability of recurrence.

5.2 Regressing the Recurrence status as a function of the copy number variation profiles of the patients, excluding the fixed covariates (Age and Gender)

5.2.1 CNV profiles based on Smooth Segmentation

In this analysis, the recurrence status was analysed using CNV profiles only. Using the smooth segmentation, the fitted recurrence status is given in the following figure.

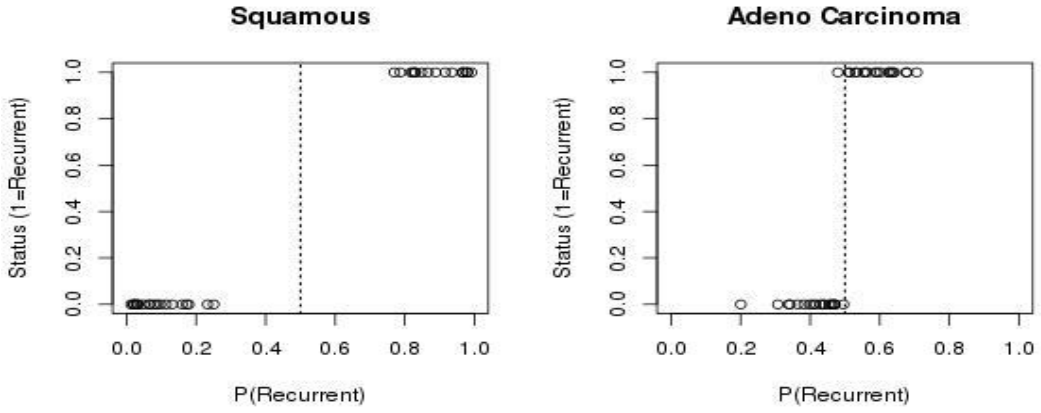


Figure 5.2. Classification for SCC and AC based on the CNV profiles only, using the smooth segmentation data

Figure 5.2 shows clearly that by fitting the logistic regression model to this data, recurrent and non-recurrent cancers can be separated in both cohorts. The recurrent cancers grouped to the right of the probability line, while the non-recurrent to the left. This was particularly obvious in the SCC cohort. The left graph shows a very distinct separation of the recurrent and non-recurrent cancers based on their CNV profiles. The distinction is perhaps less striking in the AC cohort, with only one recurrence status obviously misclassified. A single recurrent cancer is deemed non-recurrent according to the logistic regression model (it is sitting to the left of the probability line).

.

5.2.2 CNA Profiles based on discrete segmentation (DNACopy)

When the logistic regression model was fitted on the CNV data normalized by discrete segmentation, the following picture was obtained (Figure 5.3).

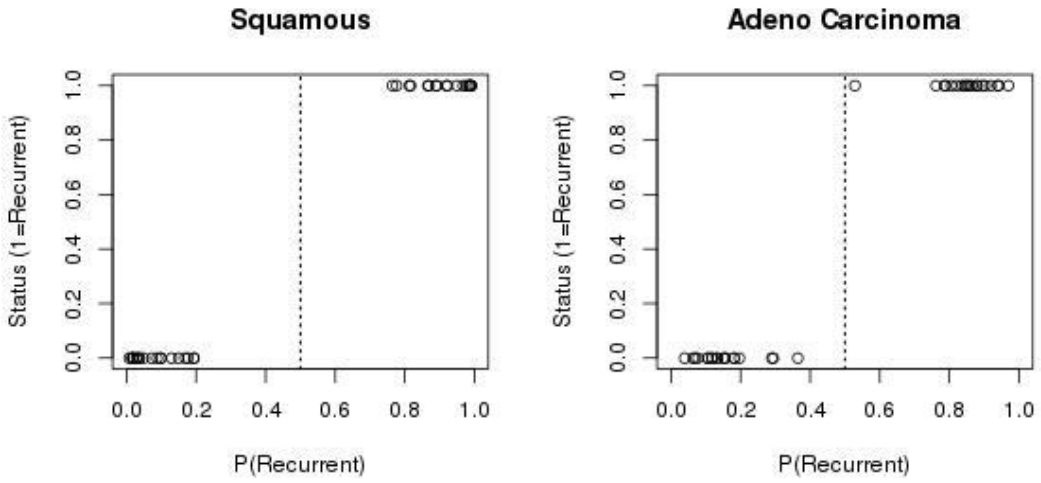


Figure 5.3. Classification for SCC and AC recurrence status, based on the CNV profiles only, obtained using the discrete segmentation. (DNACopy)

The model fit is even more striking when discrete segmentation is used to visualize the CNV in the sequencing data. Both cohorts show no cases of misclassification – all recurrent cancers sit on the right of the probability line, whilst all non-recurrent are grouped to the left.

5.4 Regressing the Recurrence status as a function of the fixed covariates and the copy number profiles

In this analysis, both the covariates and the sequencing CNV data were used in the logistic regression analysis. The classification figures for sequence data using the smooth segmentation in both cohorts (SCC and AC) is given in Figure 5.4, while discrete segmentation/DNAcopy was used im Figure 5.5. Both show no sample misclassification.

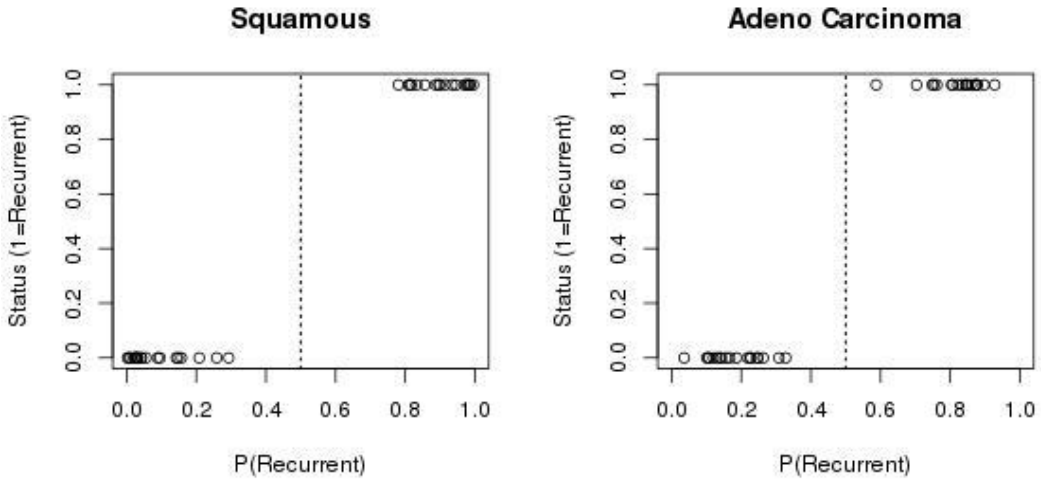


Figure 5.4. Classification for SCC and AC recurrence status based on the variables (Age and Gender) and CNV profiles, using the sequencing data obtained after smooth segmentation.

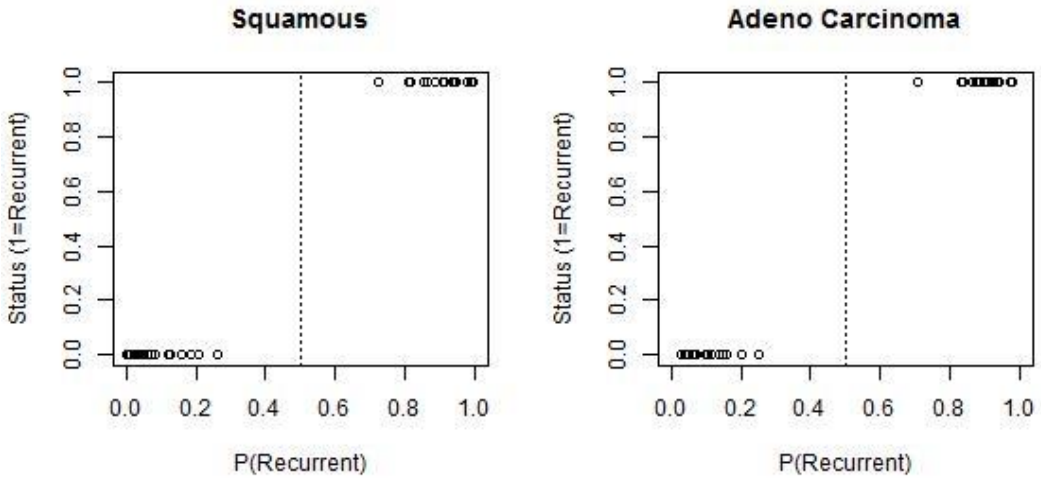


Figure 5.5. Classification for SCC and AC recurrence status based on the variables (Age and Gender) and CNV profiles, using sequencing data after discrete segmentation (DNACopy).

The above analyses show an excellent fit of a logistic regression model, when the CNV data from NG-seq is used. With the clinical data on recurrence already available for the analysed tumour samples, the fitted model was able to differentiate between the investigated clinical characteristic (recurrence) in the two cohorts and classify them correctly, based on their CNV profiles. Only one sample from the AC cohort was misclassified as a non-recurrent, when smooth rather than discrete segmentation was used in the analysis of the sequencing data. At the same time, when age and gender were used as independent variables in the model, the results were significantly more random, with no obvious pattern observed. These results strongly suggest that the genomic profiles based on CNV in recurrent and non-recurrent stage I NCSLCs are significantly different. This was observed in both of the histological subgroups.

In order to test the model, it was applied to a larger group of cancers (approximately 80 cases of AC and SCC) and used to predict their histological subtype. The fitted model once again showed excellent distribution with no samples misclassified (Figure 5.6).

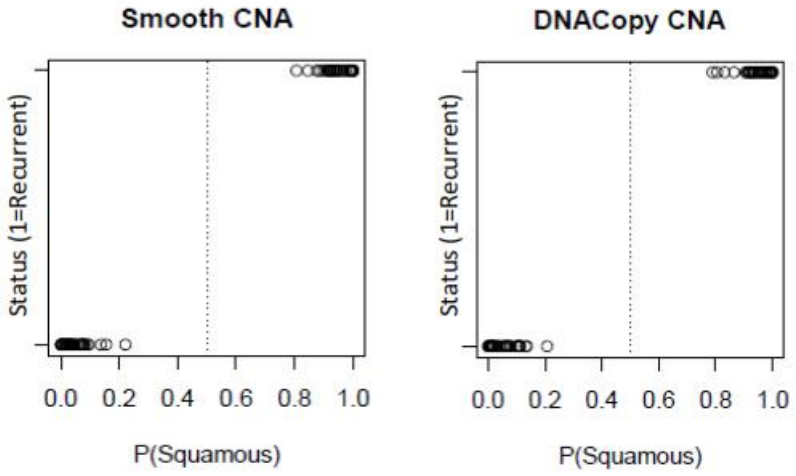


Figure 5.6. Classification for SCC and AC histological subtype based on CNV profiles, using sequencing data after discrete segmentation. The binary response is 0 for AC and 1 for SCC. There is no misclassification.

A further test of the model was performed by randomly assigning a consecutive number (1-38 for the SCC and 1-48 for the AC cohort) to each sample in the study. The logistic regression model was fitted in an attempt to differentiate samples with odd from samples with even numbers based on their CN profiles. In this case the model failed to differentiate odd from even numbered samples in both cohorts.

5.5 Discussion

The main objective of this study was to develop a prediction model for recurrence of early stage NSCLC by analysing multiple samples and stratifying them based on their CNV signatures. The initial attempt to look for single regions of the genome where the recurrent and non-recurrent cohorts showed markedly different CNV did not yield convincing results. This led to a change of mind-set and the focus shifted from looking for specific regions to developing a model that can separate recurrent from non-recurrent cancers by incorporating the CNV patterns that occur along the entire genome (their genomic signatures). By separating the genome into numerous windows (just over 17 000 for each sample in this study), and considering the CNV for each window as an independent variable, we aimed to take into account the genetic events that occur globally along the genome and their significance when trying to differentiate the stage I cancers into clinically significant subtypes. This approach reflects the complex nature of cancer recurrence, which was perhaps simplified in our initial approach of looking for a single or small number of significantly different regions. The underlying biology of lung cancer recurrence is unclear. It is very likely that the alteration of several genes and multiple drivers are required for cancer progression. The mechanisms involved in cancer recurrence most likely require a combination of inactivation (genomic loss) of tumour suppressor genes and amplification (genomic gain) of proto-onco genes. Recurrence, either local or distant, could be determined and/or facilitated by more than one pathway. Depending on the changes in the cancer genome, these pathways could occur simultaneously. The genomic changes (amplifications and deletions) can lead to activation of aberrant pathways of different cellular functions such as neo-angiogenesis, invasion and metastasis. They can occur in different loci along the entire genome, which is probably the reason why the logistic regression model demonstrated such a good fit when differentiating recurrent from non-recurrent cancers.

The logistic regression model, which utilizes the changes occurring in each genomic window as an independent variable is a novel approach to CNV analysis, which might address issues with normalization and reproducibility of such data. One of the most challenging aspects of Ng-seq is the huge amount of raw information generated. Normalisation of this data has proven a significant challenge, with most study groups developing their own technique and thus making results hard to reproduce by other groups. To deal with this problem, we performed a normalisation using the CNAnorm

package (Gusnanto *et al,* 2012). CNAnorm is a Bioconductor (open source software for bioinformatics) package, used to estimate CNV in cancer samples. CNAnorm performs ratio, GC content correction and normalization of data obtained using very low coverage (one read every 100 - 10,000 bp) high throughput sequencing. CNAnorm is readily available for download as freeware. Having a single software package, which can perform multiple steps of data preparation makes future validation of these results in an independent cohort seem encouragingly straightforward.

This logistic regression model could help researchers cope with several established challenges. It aims to make a “prediction” (is the cancer likely to recur) with low classification error, whilst having minimum dependency on the different preparation steps, required in CNV analysis (e.g. optimal window estimation, normalisation due to sample contamination, mapping to a reference genome). By incorporating each genomic window as independent variables other cancer characteristics and their relationship to CNV could be investigated, such as the likelihood of metastasis or even response to adjuvant therapy.

In this study, we have investigated the use of logistic regression to model the likelihood of recurrence of early stage lung cancer in patients, who underwent radical surgery. The model enables the inclusion of clinical characteristics (such as age and gender) as fixed covariates and CNV profiles as random predictors in a single modelling framework. The model exhibits a good fit and, whilst in a cross-validation, shows minimal classification error it is not without its shortcomings. It remains a model fit, which is able to separate the investigated samples in two cohorts, but with limited usefulness in making an accurate predication of whether an independent sample is likely to recur or not. Essentially it demonstrates that recurrent cancers differ significantly from non-recurrent cancers by their copy number signatures. This difference is determined by multiple events of gain and/or loss, which occur along the entire genome. The model essentially determines the “score” of each window by judging how often the CNV are associated with the investigated feature (recurrence vs. non-recurrence) (figure 5.7) and uses cross validation to determine the classification of each sample.

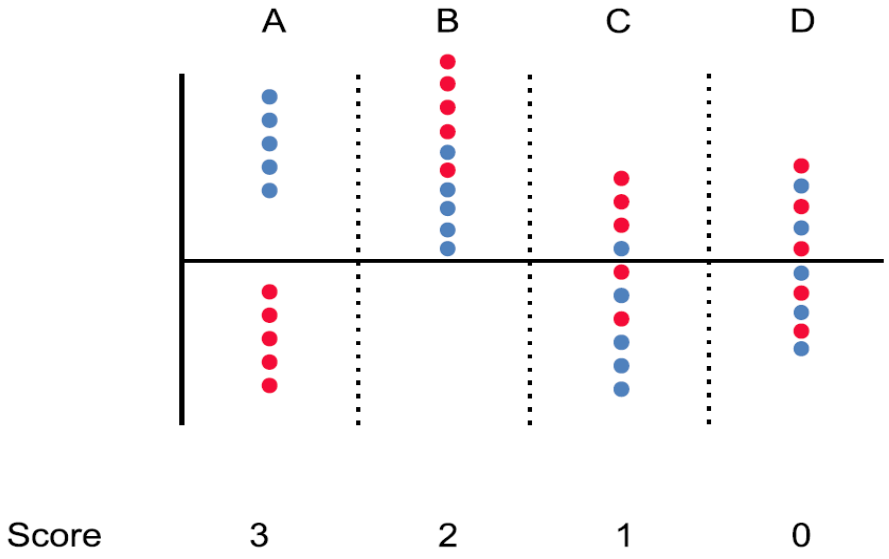


Figure 5.7. Scoring system. The red dots represent the CNV (gain or loss) in non-recurrent samples and the blue in recurrent. The more distinct the grouping is – the higher the "score" for the respective genomic window.

However, its actual predictive power in its current form is limited. It was unable to convincingly predict the recurrence likelihood of any samples, which were not used in the development of the original model. For example, when a random sample was excluded from the cross-validation and creation of the model and then the model was used to predict the likelihood of its recurrence the success rate was approximately 60%. This was in stark contrast with the "model fit", which showed no case misclassification based on the recurrence status. In order to test the model, it was applied to a larger group of cancers (approximately 80 cases of AC and SCC) and used to predict their histological subtype. The fitted model once again showed excellent distribution with no samples misclassified, whilst the actual accuracy of prediction whether an "independent" sample is AC or SCC based on their CNV profile was just over 90%. The reason for this difference in performance of the logistic regression model is unclear, but could very likely be related to relatively low number of cases used when focusing on recurrence. Perhaps, with larger cohorts, derived from different subpopulations, its predictive power will increase. In order to discover the true potential of the logistic regression model and whether it can truly be integrated successfully into clinical practice, its validation in a larger, independent cohort will be necessary. While it is clearly able to differentiate the genomic signatures of recurrent from non-recurrent stage I NSCLC, this will allow a more precise evaluation of its actual predictive power.

Chapter 6
Conclusion

6.1 Relevance of study

NSCLC, which comprises 85% of all lung cancer cases, is the leading cause of cancer mortality worldwide. In the last decade only minor improvements in the clinical outcome have been achieved. This does not adequately reflect the significant research, which has gone into deciphering the genetic abnormalities that drive the process of carcinogenesis. Whilst numerous discoveries of structural abnormalities and functional alterations have been made, using a variety of different methodologies, these have not translated well into everyday clinical medicine. The reasons for this are complex. Histopathological heterogeneity of NSCLC, complicated and expensive methodology and uncertainty in how to identify statistically significant recurrent genetic alterations, when samples vary substantially in their characteristics are just some of the key issues, that have proved difficult to surmount.

Despite this histopathological heterogeneity, all subgroups of NSCLC were treated, until recently, in a similar fashion. The prospect that poor five-year survival and poor response rates to treatment are in part due to a homogenous response to a heterogeneous disease (Borczuk *et al,* 2010) should be strongly considered. Traditionally, NSCLC subtypes (AC, SCC) have been treated as the same biological entity and the treatment strategies have been guided predominantly by their stage based on the TNM system. However, there is growing evidence that histological subtypes in NSCLC respond differently to targeted therapies. Two of the most prominent examples are the superior efficacy of the folate antimetabolite Pemetrexed in patients with non-SCC (presumably due to the higher expression of thymidylate synthase in SCC) (Scagliotti *et al,* 2009) and a higher response rate upon treatment of AC with the EGFR tyrosine kinase inhibitors Gefitinib and Erlotinib, reflecting the higher prevalence of EGFR mutations in this subtype (Mok *et al,* 2009, Langer *et al,* 2010). Furthermore, histological subtyping might play an important role in explaining why previous studies, aiming to identify genetic models predicting recurrence in stage I NSCLC have failed to deliver a clinical impact. In 2009 Broet *et al* used high-resolution microarrays to generate tandem DNA copy number and gene

expression profiles for 85 stage IB lung adenocarcinomas/large cell carcinomas. The group identified specific CNV linked to relapse-free survival and selected genes within these regions exhibiting copy number– driven expression to construct a novel integrated signature (IS) capable of predicting clinical outcome. They noted that failure to incorporate histological subtype might reduce model robustness and predictive accuracy. Using two previously published pure gene expression–based models, the 5- and 16-gene signatures from Chen *et al,* in 2007 and a 50-gene prognostic signature from Beer *et al,* in 2002, they were not able to significantly discriminate between low-risk and high-risk patients in their own cohort (Beer *et al*, 2002, Chen *et al*, 2007).

For patients with early-stage disease, the 5-year survival rates after surgery are as low as 40% to 55% (Mountain CF, 1997, Adebonojo SA, 1999). This makes the issue of accurately identifying subgroups, which might benefit from adjuvant chemotherapy very important (Wakeleea H, 2007). The role of adjuvant chemotherapy for stage IB tumours, however, remains controversial. Preliminary results of the CALGB 9633 trial suggest a potential survival benefit for adjuvant chemotherapy in stage IB disease, but updated results from the same trial show no benefit in overall survival. However, recent clinical trials have shown that adjuvant therapy following resection of lung tumours can lead to improved survival in early-stage NSCLC. In 2004, Kato *et al,* showed that adjuvant chemotherapy with uracil-tegafur improved survival among patients with completely resected pathologic stage I lung ACC (T1N0M0 or T2N0M0) whilst in 2005 Winton *et al* found that early-stage patients who received a combination of Vinorelbine and cisplatin after surgery had an improved overall survival in those patients who did not receive the adjuvant therapy (94 months compared with 73 months) (Winton *et al*, 2005). This suggests that patients with stage I NSCLC represent an excellent opportunity for applying genomic strategies, which will stratify patients into cohorts with low and high risks of recurrence. Currently, there is no established pathway to identify those patients with surgically treated early stage NSCLC that have high risk of cancer recurrence. The ability to identify such high risk cases, particularly in the early post-operative period will allow stratification for additional surveillance or adjuvant therapy. This could lead to an improved survival in these patients. I strongly feel that this work deals with a very specific subgroup of patients with NSCLC, in which cancer recurrence and overall survival can differ significantly and a clear distinction between the two needs to be made. Several of the cases in this study, from both histological cohorts, survived for more than two years after the original

tumour relapse. In fact, one patient was treated for two separate disease recurrences, which occurred more than 24 months from each other. This data made a survival analysis unsuitable to the aims of this particular study and was not performed.

In an editorial for the Annals of Thoracic surgery titled “Molecular biological staging of lung cancer” (published in 2008) Thomas D’Amico suggested that the current staging system (TNM) may have outgrown its usefulness as far as predicting outcomes is concerned, and outlined the necessary characteristics for an optimal cancer staging system (D’Amico, 2008). It must achieve accurate assessment of extent of disease, effective prognostic stratification, and appropriate selection of therapy (D’Amico, 2008). The current TNM system does not differentiate between the subtypes of NSCLC. Patients with identical TNM and histological features show significant differences in the further development of their disease, despite receiving identical treatments. Molecular methods may have a significant role in helping to sub-stratify patients with NSCLC in prognostic groups who might benefit from additional treatment or more aggressive follow up.

To select a subgroup of patients with stage I disease that might benefit from adjuvant therapy, investigators have attempted to identify factors that predict poor prognosis. Recent interest has focused on the identification of biologic markers that predict early recurrence and death in patients with NSCLC. This has been necessitated by the desire for an individualized therapeutic approach, which in light of recent technological advances and improved understanding of tumour genomes seems somewhat generic. Tumour markers may serve to support the current TNM system in improving risk stratification. D’Amico refers to this as biologic cancer staging and suggests its targets - oncogenes, oncogenic protein products, growth factors and/or receptors (D’Amico, 2008).

This study aimed to take these issues into consideration, while targeting an area of lung cancer genomics with potentially very practical applications.

Three key aspects were considered in the design of this study:

1. Identifying a biomarker that would have translatable therapeutic implications. The focus fell on recurrence in cases of stage I NSCLC, that had undergone radical surgical therapy. The high recurrence rates (Hoffman *et al,* 2000) and poor 5-year survival post-surgery (40-50%) (Mountain *et al,* 1997, Adebonojo *et al,* 1999), combined with

the absence of guidelines for adjuvant chemotherapy or specific surveillance were the key features that defined the suitability of the cohort.

2. Utilizing a methodology, which could be reproduced reliably on a large scale and thus integrated into routine clinical practice.
3. Making a clear differentiation between the subtypes of NSCLC (AC and SCC) and treating them as separate biological entities.

6.2 Impact on surveillance and therapy

In the last few years, numerous studies have delved into the lung cancer genome in an attempt to decipher the pathological processes that dictate carcinogenesis. While many significant discoveries have been made, the actual impact of molecular approaches to NSCLC in everyday clinical practice has not been proportional. Numerous obstacles have to be overcome before molecular discoveries are successfully translated into clinical benefit, such as complex data analysis and its integration (Ocak *et al,* 2009), protocols for preserving and handling of fresh tissues and complex validation of results with large multicentre randomized control trials (D'Amico, 2008).

This study attempted to take these difficulties into account and consider possible clinical applications, including the management and surveillance of stage I NSCLC.

6.2.1 Adjuvant chemotherapy in stage I NSCLC

Current treatment protocols do not routinely offer adjuvant chemotherapy to patients with radically resected stage I NSCLC (Pisters al, 2007), despite the high recurrence rate (compared to other leading cancers) and the benefit suggested by some studies. This largely reflects the difficulty in identifying which patients might benefit from additional therapy when weighed against its side effects and complications. By using molecular biomarkers of recurrence this issue may be overcome. Logistic regression modelling for predicting recurrence in stage I NSCLC following radical surgery based on CNV could have greatest impact in the postoperative management. It can be applied to sequencing data of tumour samples obtained after the surgical resection of any stage I NSCLC. There would be no major time constraints and the treatment process would not be hindered or slowed down, as this would be done during the standard convalescence period after lung cancer

surgery. In Leeds Teaching Hospitals, patients are routinely reviewed approximately 8 weeks following the operation. Usually, at that point they have recovered sufficiently to be considered for additional therapy. If at this stage prediction data is available to the MDT and it suggests a high probability of recurrence, then the patient can be offered adjuvant chemotherapy and proceed with the treatment. With the emergence of massive parallel sequencing platforms, which allow simultaneous processing of multiple samples, DNA from all stage I NSCLC operated in a unit over a period of time (e.g. 2 weeks) could be sequenced together, thus streamlining the process further and justifying a regular departmental “slot” for using a sequencer.

6.2.2 Surveillance of stage I NSCLC following radical resection

The issue of surveillance of patients with NSCLC following radical resection is one of the most controversial in the field. No universal guidelines exist and the practices vary greatly from unit to unit and often physician to physician. At the time of writing, there are five practicing consultant thoracic surgeons in the Leeds Teaching Hospitals Trust, who work in collaboration with a Lung Cancer MDT. The surveillance protocols following surgery differ from surgeon to surgeon (depending on individual preferences) in several aspects, such as timing of appointments, obtaining routine CXR and involvement of other relevant specialties (respiratory physicians, oncologist). The routine place of a CT scan in this process is also not clearly established. This snapshot is likely representative of the practices in most large thoracic centres.

In a study, focusing on the postoperative surveillance of a large cohort (346 cases) of stage I NSCLC, which underwent radical surgical resection Pairolero made several important findings (Pairolero *et al,* 2004). They observed that most of the recurrences recorded in the study occurred within the first 2 years after surgery. They also noted that only 53% of patients with recurrent disease were symptomatic and more than half of the patients with symptomatic recurrence presented and were diagnosed after non-scheduled examinations.

A logistic regression prediction model could potentially prove to be extremely useful when addressing the issues of timely surveillance of radically resected stage I NSCLC. By helping to identify at an early stage which patients are more likely to experience a recurrence, it could allow physicians

to tailor a more rigorous surveillance regime, especially in the first 12-24 months. If a patient's genomic profile is found to have a high risk of future recurrence the awareness of medical staff can be raised and a CT surveillance routinely performed at regular intervals (e.g. at 3, 6, 12 and 24 months) post-surgery. This could lead to early detection of both local and distant cancer recurrence and allow for timely therapeutic intervention. This benefit could be achieved without any additional discomfort for the patient, as calculation is based on sequencing data from extracted DNA from a cancer sample obtained post-surgery, thus alleviating the need for any extra clinical appointments or invasive tests.

While this study has shown promising results, it does have its limitations. The small sample size, combined with a "model fit", based on cross validation could be responsible for the clear distinction, which was achieved between recurrent and non-recurrent tumours. These factors could be masking a convenient sample distribution. Although testing the model with histological subgroups and cohorts with randomly assigned numbers suggests otherwise, this remains a possibility. Due to its retrospective character, this study relied heavily on gathering information from different sources, which were not specifically designed for this particular purpose. Although great care was taken to obtain precise data for the tumour samples, small inaccuracies could have subtly influenced the results. Using CNV data has certain shortcomings, which could affect the future of this study. Differences in CN are a naturally occurring phenomenon and are not always associated with abnormal activation and/or inactivation of genes. While using NG-seq of DNA allows the "charting" of these events it does not truly "interrogate" the genome about their precise nature and the molecular pathways they unlock. Finally, although NG-seq is rapidly becoming more affordable and widespread, the cost of whole genome sequencing remains relatively high and could hinder its routine introduction in clinical practice.

6.3 Conclusion

Lung cancer is a spectrum of diseases with numerous alterations in expression patterns resulting from acquired genetic and epigenetic mechanisms (Varella-Garcia, 2010). While numerous genomic changes in individual specimens have been discovered, few of these are recurrent among large numbers of tumours. This has proven to be a major obstacle to forming a precise and universally accepted definition of molecular subtypes

in NSCLC and hinders the formation of algorithms for individualized treatment.

The role of molecular methods in the management of NSCLC is slowly increasing. This pilot study attempted to define a genomic pattern associated with recurrence in radically treated stage I NSCLC using CNV and suggest a feasible application for such an algorithm in clinical practice. The results show that no single area of the genome can be identified as “governing” the process of cancer recurrence, which is likely a result of multiple complex events involving inhibition of tumour suppressor genes, activation of oncogenes, mutations etc. By using a novel prediction model, which takes into consideration abnormal gain or loss of material along the entire genome, this study has shown CNV could be used to differentiate recurrent from non-recurrent stage I NSCLC and guide its further management. Before any practical application of the logistic regression model can be considered, a further validation of the model in a larger cohort of radically treated cases of stage I NSCLC must be performed. The advances in software and processing power have determined the emergence of extensive databases, containing vast amounts of cancer data such as details on surgical procedures, adjuvant and neo-adjuvant therapy, concomitant diseases (benign or malignant), survival and recurrence data. This will greatly facilitate the identification of such a cohort, although collaboration between several institutions, perhaps in the form of a trial, might be necessary in order to achieve truly substantial number of cases.

Appendix A
Laboratory protocols

DNA extraction with macrodissection

Performed using the QIAamp DNA micro kit (Qiagen, Sussex, UK) according to the manufacturer's instructions:

DNA extraction protocol from FFPE tissue blocks (targeted at obtaining at least 70% of cell content):

Preparation: - Each new HE slide was reviewed by a pathologist and the tumour area was marked. The pathologist also commented on the size of the tumour area and the tumour cell content.

- Between seven and ten slides were sectioned (7 microns in thickness)

Procedure

1. Adequate preparation surface of the working area (70% Ethanol used to clean the surface).
2. Five glass baths filled with solvents:

    - xylene
    - 100% ethanol
    - 90% ethanol
    - 70% ethanol
    - diH20

3. Dewaxed sections placed on a rack and submerged for:

    - 5 minutes in the xylene
    - 3 minutes in each ethanol containing bath
    - Left in diH20 until microdissection commences

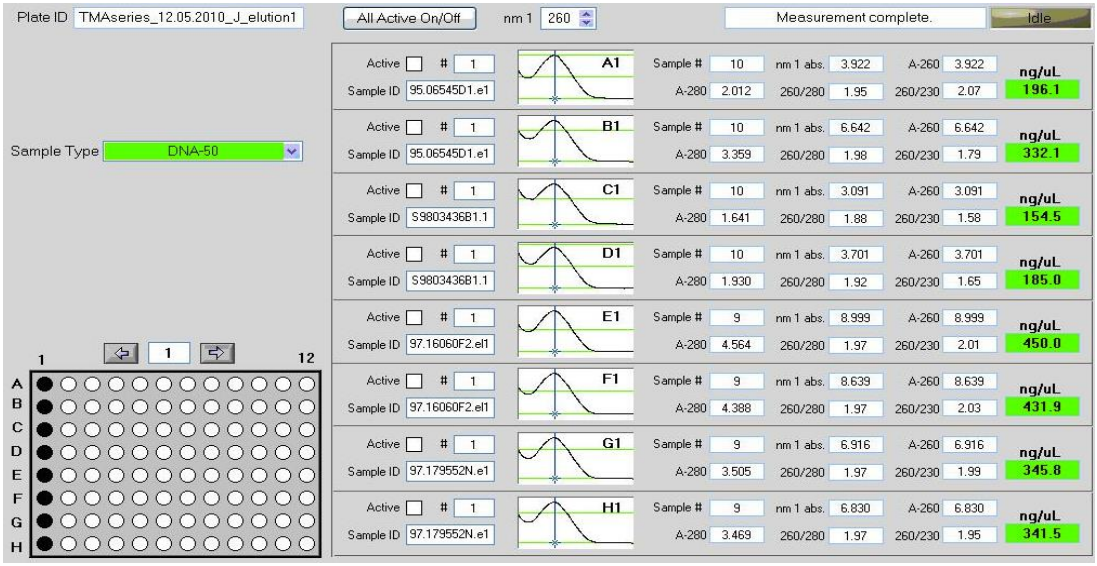
4. The tissue is scraped from the desired area of each tumour slide and placed in a 1.5 ml labelled centrifuge tube.
5. 180 μl of Buffer ATL and 20 μl of proteinase K were added.
6. The samples were mixed by pulse-vortexing for 10 s.
7. The samples were placed in water bath for incubation at 56 degrees for 48 hours.

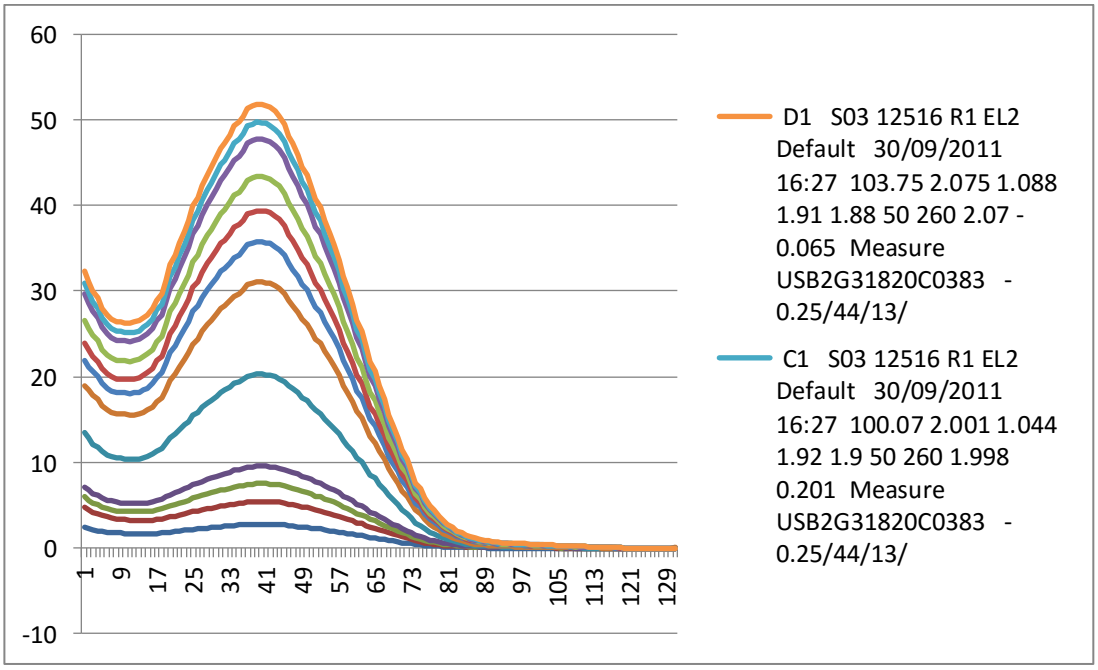
    - Samples can be reviewed after 24 hours and additional 20 μl of proteinase K can be added if necessary

8. The samples were removed from water bath. 200 μl of Buffer AL added and mixed by pulse-vortexing for 15 s.
9. The samples incubated at 70 °C for 10 minutes and briefly centrifuged after cooling down.
10. 200 μl of 100% ethanol added. The samples were pulse-vortexed for 10s and left to incubate at room temperature for 5 minutes.
11. The mixture is transferred to a spin column and centrifuged at 8000 rpm for 1 minute.
12. The spin column was transferred to a clean centrifuge tube and the tube containing the filtrate was discarded.
13. 500 μl of Buffer AW1 added and the samples were centrifuged again at the same speed.
14. The spin column was transferred to a clean centrifuge tube and the tube containing the filtrate was discarded.
15. 500 μl of Buffer AW2 added and the samples were centrifuged at 14 000 rpm for 3 minutes.
16. Spin column placed in a clean labelled 1.5 ml tube. Filtrate discarded.
17. 100 μl of Buffer AE added.
18. Samples incubated for 5 min at room temperature and centrifuged at 8000 rpm for 1 minute.
19. Spin column placed in a clean tube (clearly labelled as Elution 2) and steps 17. and 18. were repeated. The filtrate was clearly labelled as Elution 1.
20. Samples were stored at 4 °C

Nanodrop

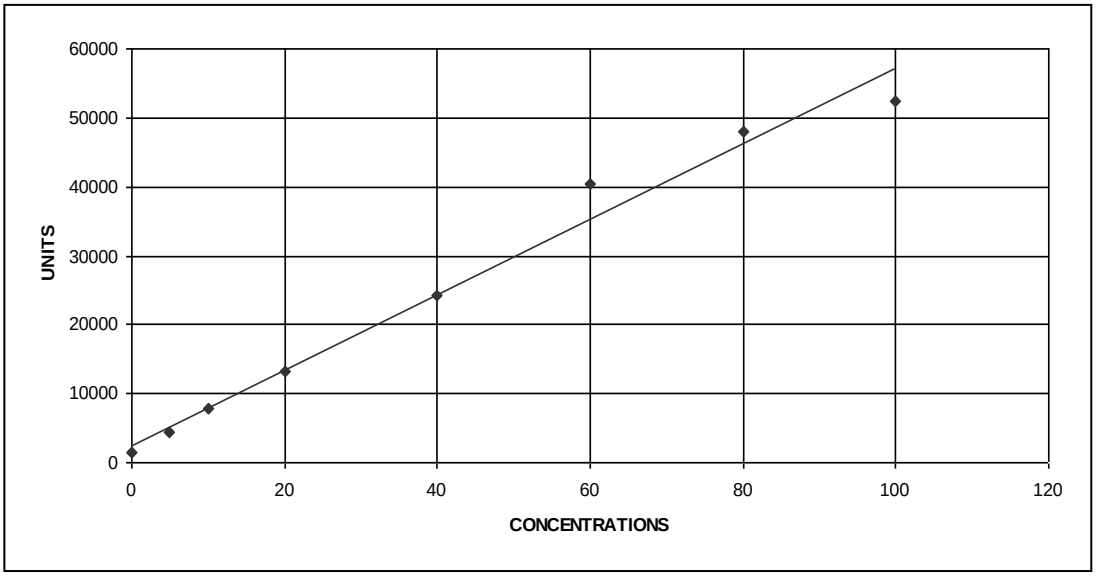
1. An initiation cycle was performed at the start of each measurement with 2uL of dH2O
2. A blanking cycle was performed before each set of measurements with 2 uL of buffer AE
3. DNA samples pulse-vortexed for 10 s. prior to measurement
4. 2 uL of DNA samples used for measurement
5. Two measurements taken for each sample





Pico green quality assay

1. Working solution was prepared by diluting Quant-iT™ dsDNA BR reagent 1:200 in Quant-iT™ dsDNA BR buffer..
2. 200 µL of the working solution were loaded into each microplate well.
3. 10 µL of each DNA standard were added to separate wells and mixed
4. 10 µL of each investigated DNA sample were added to separate wells and mixed
5. The plate was loaded in a reader and the fluorescence was measured
6. A standard curve was used to determine the DNA amounts.



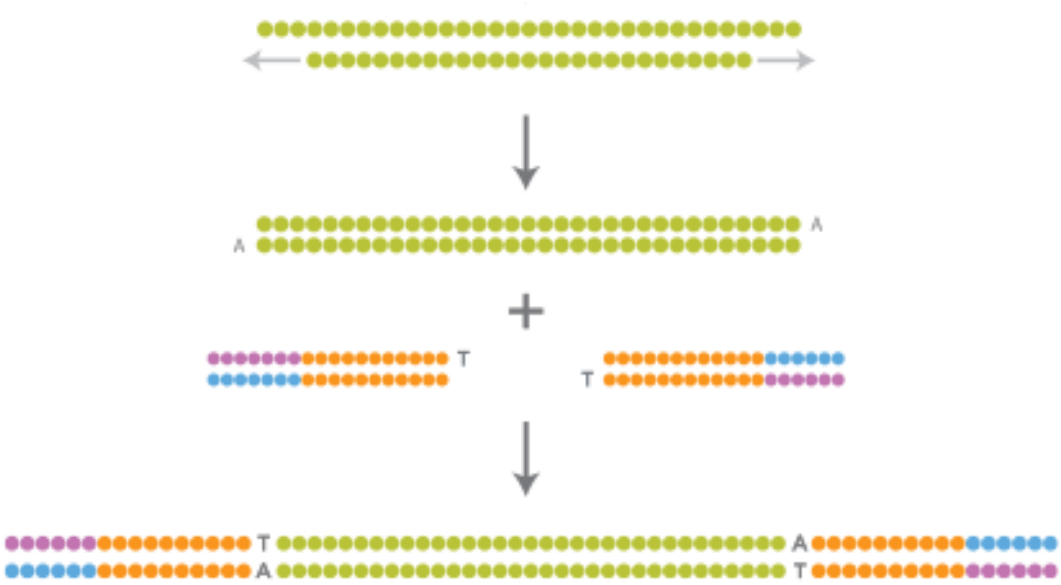
Next generation sequencing

The steps involved in using a next-generation sequencing platform are: (1) DNA library preparation (including shearing the DNA to desired size, end-polishing, adaptor ligation, nick-translation- amplification, and gel purification of libraries); (2) quantification of the product from step one; (3) emulsion PCR; (4) depositing templated beads onto the instrument for sequencing.

End repair was performed by using the End-It DNA End Repair Kit (Epicentre Biotechnologies, Madison, WI, USA)

DNA libraries were prepared for the samples.

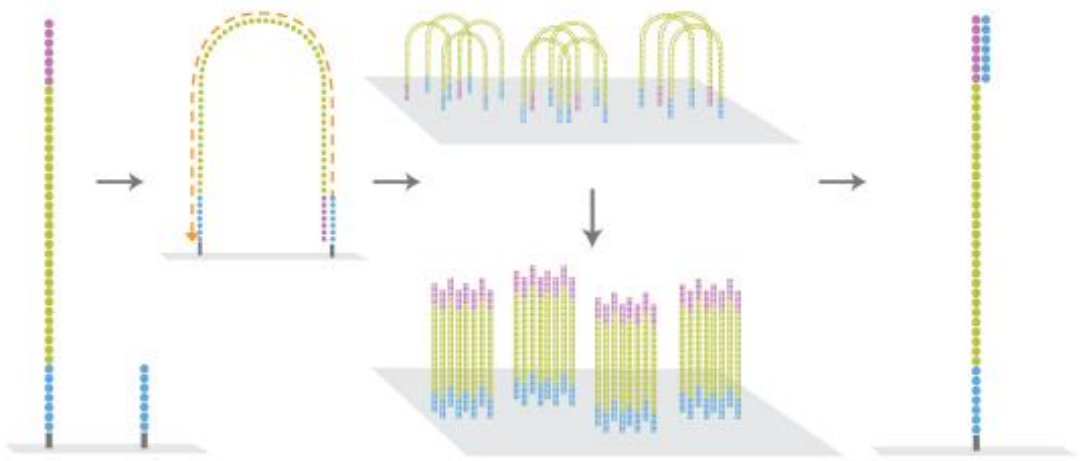
1. DNA was first sheared into a random library of 100-300 base-pair long fragments. This was performed on a Covaris S2 Sample Preparation System (Covaris Inc., Woburn, MA, USA) and checked for appropriate size distribution on an Agilent Bioanalyser DNA 1000 LabChip.
2. After fragmentation the ends of the obtained DNA-fragments are repaired. End repair was performed by using the End-It DNA End Repair Kit (Epicentre Biotechnologies, Madison, WI, USA)
3. A-Addition. An A-overhang is added at the 3'-end of each strand using Klenow DNA polymerase.
4. Adaptors which are necessary for amplification and sequencing are ligated to both ends of the DNA-fragments.
5. These fragments are then size selected and purified using a 2% high purity agarose gel.



Samples were enriched using a 12-cycle enrichment PCR. For low concentration DNA samples, an 18-cycle enrichment PCR was performed before the gel cut stage rather than afterwards. Libraries were then examined using an Agilent Bioanalyser DNA 1000 LabChip and Invitrogen's Quant-iT Picogreen dsDNA BR assay kit to assess for DNA quality and concentration, respectively. This information was used to pool equal amounts of each sample library for cluster amplification and either 51 or 76 cycles of Illumina sequencing by synthesis, resulting in 45/70 bp of genomic DNA sequence and 6 bp of tagged adapter. Sequencing was initially done with 51-bp reads but the move was made to 76-bp reads as machine and analysis package upgrades resulted in better base calling for longer

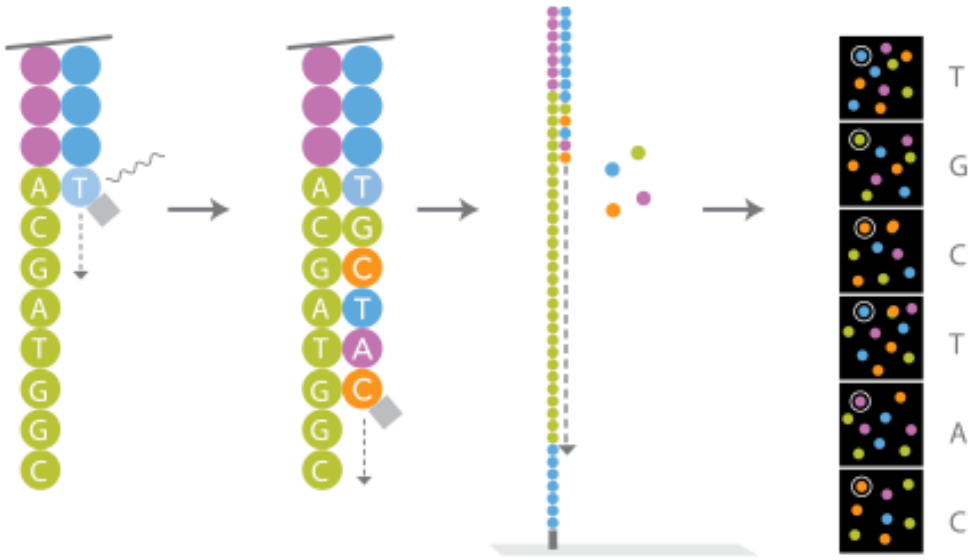
Cluster Generation

The Cluster Generation is performed on the Illumina Cluster Station. Single DNA-fragments are attached to the flow cell by hybridizing to oligos on its surface that are complementary to the ligated adaptors. The DNA-molecules are then amplified by a so called *bridge amplification* which results in a hundred of millions of unique clusters. Finally, the reverse strands are cleaved and washed away and the sequencing primer is hybridized to the DNA-templates.



Illumina’s sequencing by synthesis technology is the most successful and widely-adopted next-generation sequencing platform worldwide. It supports supports massively parallel sequencing using a proprietary reversible terminator-based method that enables detection of single bases as they are incorporated into growing DNA strands. A fluorescently-labeled terminator is imaged as each dNTP is added and then cleaved to allow incorporation of the next base. Since all four reversible terminator-bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias. The end result is true base-by-base sequencing that obtains accurate data for a broad range of applications.

During sequencing the huge amount of generated clusters are sequenced simultaneously. The DNA-templates are copied base by base using the four nucleotides (ACGT) which are fluorescently-labeled and reversibly terminated. After each synthesis step, the clusters are excited by a laser which causes fluorescence of the last incorporated base. After that, the fluorescence label and the blocking group are removed allowing the addition of the next base. The fluorescence signal after each incorporation step is captured by a built-in camera, producing images of the flow cell.


A
T
C
G
A
T
G
G
C
A
T
C
G
G
C
A
T
T
A
G
C
G
C
T
G
C
T
A
C

Pre-Cancer Genomics Group Library Prep Protocol – Copy Number Assay using NEBNext DNA Library Prep Master mix with NEB adaptors and primers.

200ng of DNA is the standard amount for the CNV assay using this protocol. If the DNA has a concentration of more than 200ng/µL, dilute an aliquot to a concentration of 200ng/µL. This will reduce the chance of pipetting errors, which can be a problem for volumes less than 1ul. The concentration must be determined using a fluorescence assay (e.g. Pico Green), as this measures the amount of double stranded DNA.

A) SHEARING

Dilute DNA in TE buffer, making the final volume 250ul. Add sample to a shearing tube, and clip on lid. If using crimped lids – use crimping tool.

Using Covaris S2 system.

Make sure fresh distilled water is used in the Covaris tank each time. Also check level of water bath and top up if required.

Turn on water bath and Covaris before opening SonoLite software. Degassing pump should start automatically – this needs to run for at least 30 minutes before use. Water bath temperature should be set to 20 degrees.

Covaris settings:

The only parameter that needs changing is the 'cycle repeat' number in the 'Batch' Tab. Everything else is pre-programmed.

Settings for DNA shearing in 'Run' Tab (don't change):

	Duty Cycle	Intensity	Cycles/burst
1000bp	19.9%	9.9	1000
500cpb	15%	8	500

Batch: 500cpb

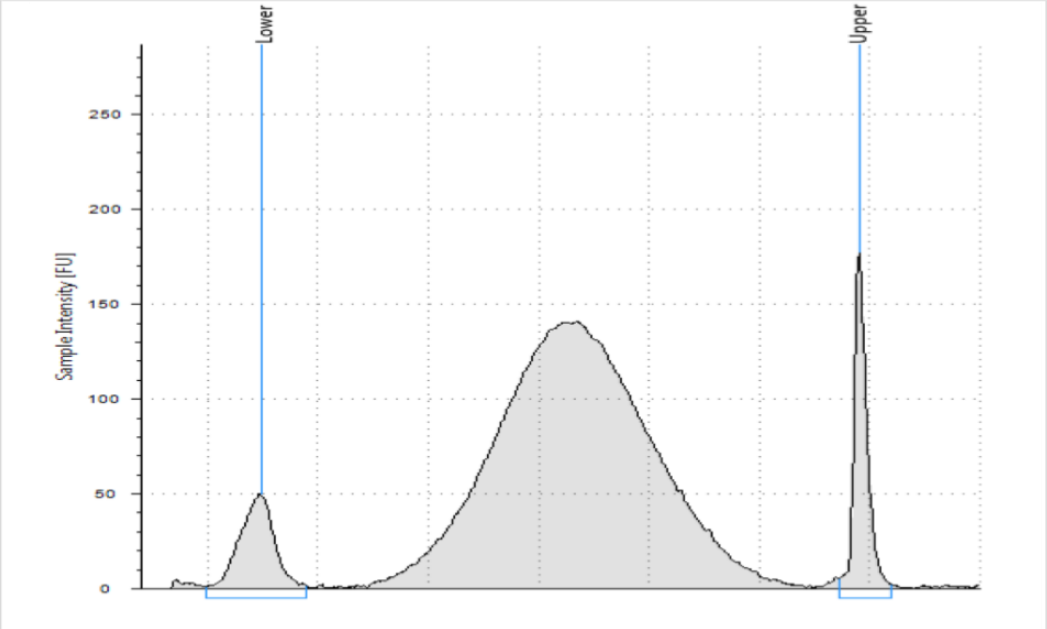
1000bp

Cycles = 25

Once sample is sheared, remove the lid (if a crimped lid was used, use the de-crimping tool). Pipette out the sample using long tips, and process the

solution through a MinElute column according to the Qiagen protocol. Elute in 11µL.

Check 2µL of each sheared sample on the Agilent Tapestation D1K High Sensitivity Screentape (see Agilent protocol). The trace of the shear should look like this.



Switch off Covaris and water bath. Empty Covaris tank.

Resulting DNA can be stored at -20C until required for next step.

B) End-Repair of Fragmented DNA

	Volume x1
NEBNext End repair reaction buffer (x10)	5
NEBNext End repair enzyme mix	2.5
dH20	33.5

Make up master mix using the volumes in the table above. Add 41µL master mix to each sample to the 9µLof DNA from previous step.

Incubate at 20C for 30min.

Clean up using a QiaQuick Column, following Qiagen Bench Protocol.

Elute in 21µL EB Buffer

C) dA-Tailing of End Repaired DNA

	Volume x1
NEBNext dA-Tailing Reaction Buffer	2.5
Klenow (3'>5' exo)	1.5

Make up master mix using the volumes in the table above. Add 4ul master mix to 21µL of DNA from previous step.

Incubate at 37deg for 30min.

Clean up with a MinElute column (Qiagen Bench protocol), eluting in 12.5µL EB buffer.

Resulting DNA can be stored at -20C until required for next step.

D) Adaptor Ligation of dA-Tailed DNA

	Volume X1
Quick Ligation Reaction Buffer (x5)	5
NEBNext Adaptor	2.5
Quick T4 Ligase	2.5
dH2O	2.5

Make up master mix using the volumes in the table above and add 12.5µLof master mix to the 12.5µL of dA-tailed DNA from the previous step. Incubate at 20C for 15min.

Add 3µL of USER enzyme mix by pipetting up and down. Incubate at 37C for 15min.

Clean up using a QiaQuick Column, (Qiagen Bench Protocol).

Elute in 50µL EB Buffer

E) Size Select Adaptor Ligated DNA Using Ampure XP Beads

1. Add 40ul (0.8x concentration) suspended AMPure XP beads to 50µL of DNA solution. Mix well by pipetting up and down 10 times. Incubate for 5 minutes at room temperature
2. Place the tube into the magnetic stand to separate the beads from the supernatant.
3. After the solution is clear (approx. 5 min) carefully transfer the supernatant to a new tube (do not discard) Discard the beads that contain the larger fragments.

4. Add 10µL (0.2x concentration of original volume of 50µL) re-suspended AMPure XP beads the supernatant. Mix well by pipetting up and down 10 times and incubate for 10 min at room temperature.
5. Place the tube into the magnetic stand to separate the beads from the supernatant. After the solution is clear (approx.. 5 min) carefully remove and discard the supernatant. Be careful not to disturb the beads as they contain DNA targets.
6. Add 200 µL of freshly prepared 80% ethanol to the tube. Incubate at room temperature for 30 seconds, and carefully remove and discard the supernatant.
7. Repeat step 6 once.
8. Air dry beads for 10 min. Tube must be free of ethanol before proceeding to next step, as ethanol can inhibit downstream applications.
9. Elute DNA in 22 µL EB buffer, pipetting up and down 10 times.
10. Without disturbing the bead pellet, carefully transfer 20 µL of the supernatant to a clean PCR tube and proceed to enrichment.

F) PCR Enrichment Adaptor Ligated DNA

	Volume
NEB High Fidelity 2x PCR master mix	12.5
Universal PCR Primer (25uM)	1.25

Make up PCR master mix using the volumes in the table above.

Add 13.5 µL master mix to 10 µL of DNA from previous step.

Add 1.25 µL of Indexed primer. Mix thoroughly using pipetting and spin dpown. Make a note of which indexed primer is used for each library.

Use PCR program ‘Enrich12’ If DNA is from a fresh source material (cell line, Fresh frozen etc)

Use PCR program ‘Enrich 15’ If DNA is from FFPE DNA

PCR Program Enrich 12/15

30 seconds at 98°C

12/15 cycles of: 10 seconds at 98°C

30 seconds at 65°C

30 seconds at 72°C

5 minutes at 72°C

Hold at 4°C

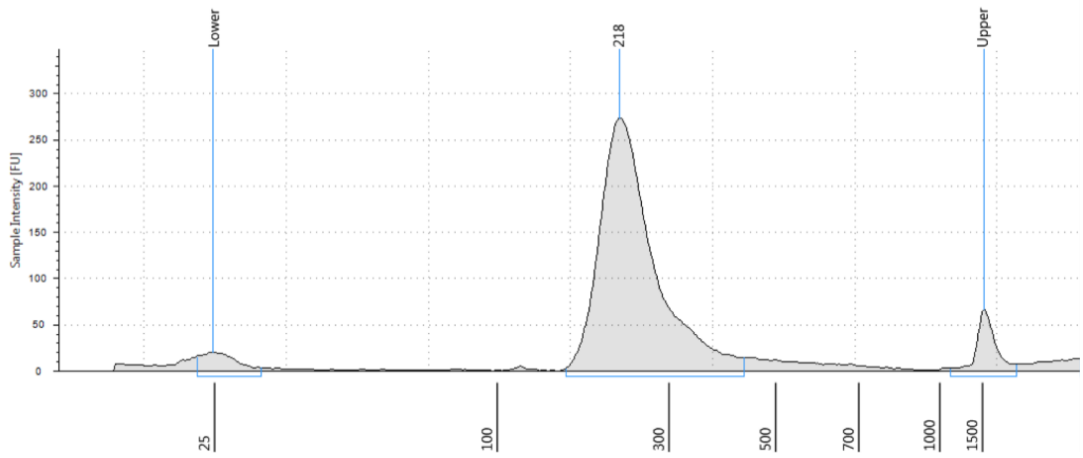
Clean-up with an Ampure bead purification using a 1x concentration (see appendix)

Elute in 40 µL of EB Buffer, as previously described.

LIBRARY QC

Proceed to DNA quantification (using PicoGreen) and Agilent Bioanalyser analysis of each library.

The Tapestation 1DK High Sensitivity trace should look this:



2f4

Please refer to Appendix B if your library is contaminated with adaptor peaks around the 115bp region

APPENDIX

A) Standard Bead Cleaning Protocol

1. Vortex AMPure XP beads to re-suspend
2. Add nX (n being the concentration specified in each step of the protocol) re-suspended AMPure XP beads to reaction mixture. Mix

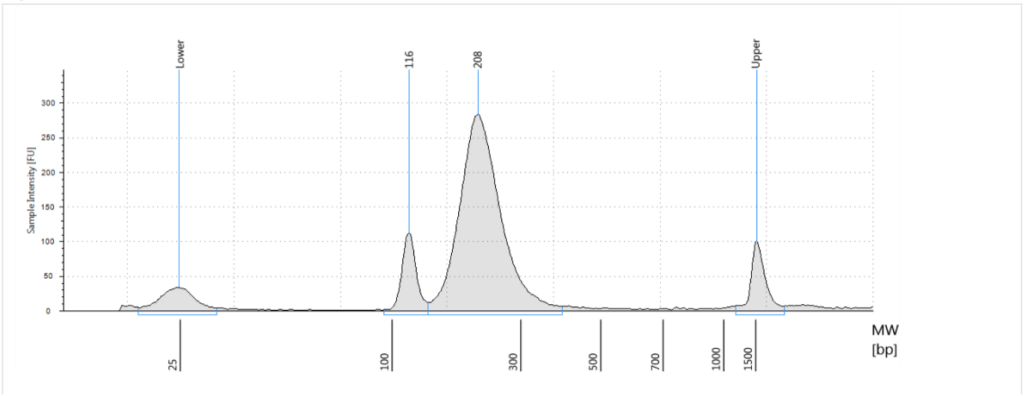
well by pipetting up and down 10 times. Incubate for 5 minutes at room temperature

3. Place the tube into the magnetic stand to separate the beads from the supernatant. After the solution is clear (approx.. 5 min) carefully remove and discard the supernatant. Be careful not to disturb the beads as they contain DNA targets.
4. Add 200 µL of freshly prepared 80% ethanol to the tube. Incubate at room temperature for 30 seconds, and carefully remove and discard the supernatant.
5. Repeat step 6 once more.
6. Air-dry beads for 10 min. Tube must be free of ethanol before proceeding to next step, as ethanol can inhibit downstream applications.
7. Elute DNA in the volume of EB buffer specified for that stage of the protocol, pipetting up and down 10 times.
8. Without disturbing the bead pellet, carefully transfer the supernatant to a clean tube.

B) Adaptor Contamination

If the final library trace shows adaptor contamination around the 115bps it means that there was too much adaptor to bind to the available DNA (see below).

C1



08k

Wavelength	MW [bp]	Conc. [pg/µl]	Molarity [pmol/l]	% of Integrated Area	Peak Comment	Observations
Sample	25	472	28600			Lower Marker
Sample	116	654	8550	12.21		
Sample	208	4700	34200	87.79		
Sample	1,500	580	586			Upper Marker

This contamination can create problems when analysing Illumina sequencing data. If the ratio of the adaptor concentration vs library concentration (values provided by Tapestation output) is < or = to 10 the library is satisfactory and can be submitted for Illumina Sequencing. If this ratio is >10 (the example above has a ratio of 13.9), it is recommended that you perform an additional clean-up step to the remaining 10 µL of adaptor ligated DNA. Use standard bead cleaning protocol (see appendix) using the beads at a concentration of 1.8x and elute in 21 µL EB buffer.

Perform enrichment PCR as described below

	Volume
NEB High Fidelity 2x PCR master mix	25
Universal PCR Primer (25uM)	2.5

Make up PCR master mix using the volumes in the table above.

Add 27.5 µL master mix to 20 µL of DNA from previous step.

Add 2.5 of Indexed primer. Make a note of which indexed primer is used for each library.

Use PCR program ‘Enrich18’

PCR Program Enrich18 30 seconds at 98°C

18 cycles of: 10 seconds at 98°C

30 seconds at 65°C

30 seconds at 72°C

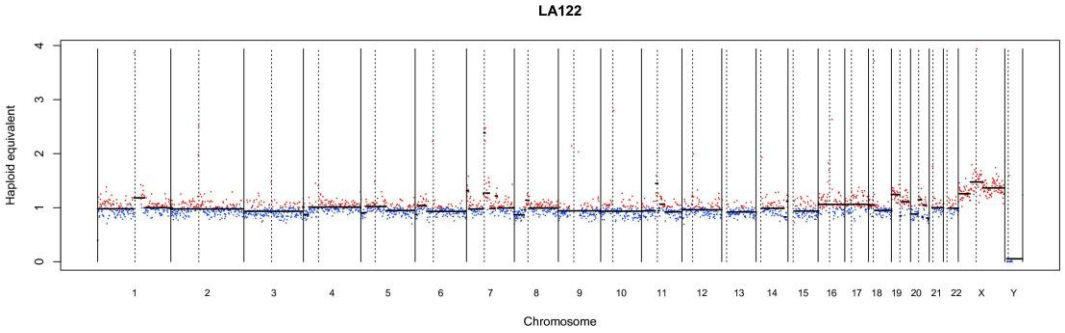
5 minutes at 72°C

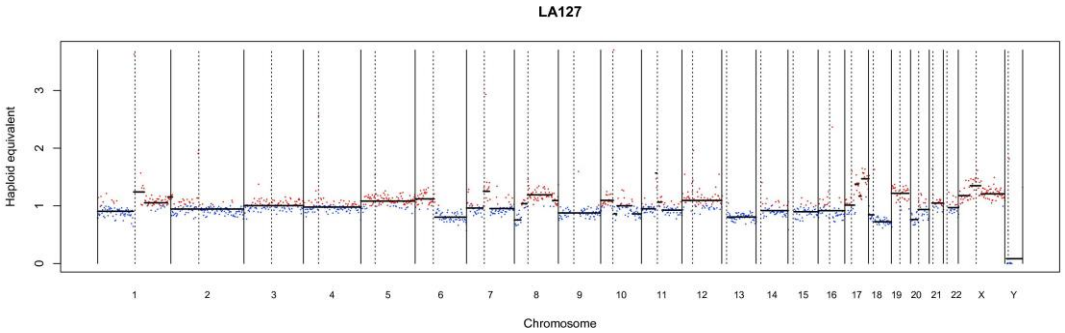
Hold at 4°C

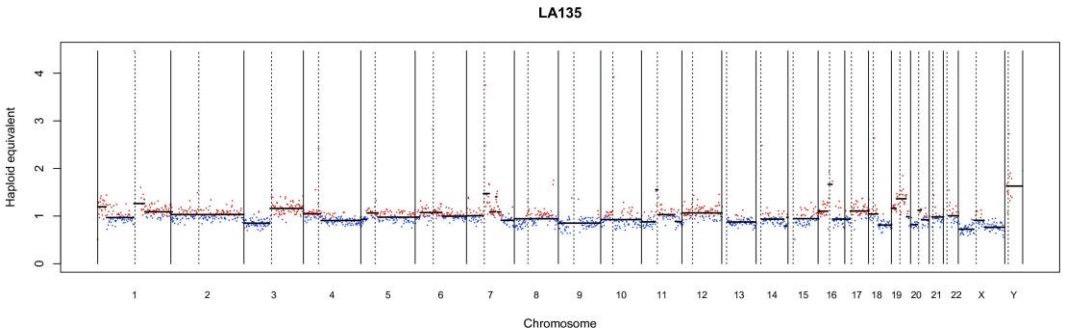
Clean-up with an Ampure bead purification using Standard Bead Cleaning Protocol at a 1x concentration.

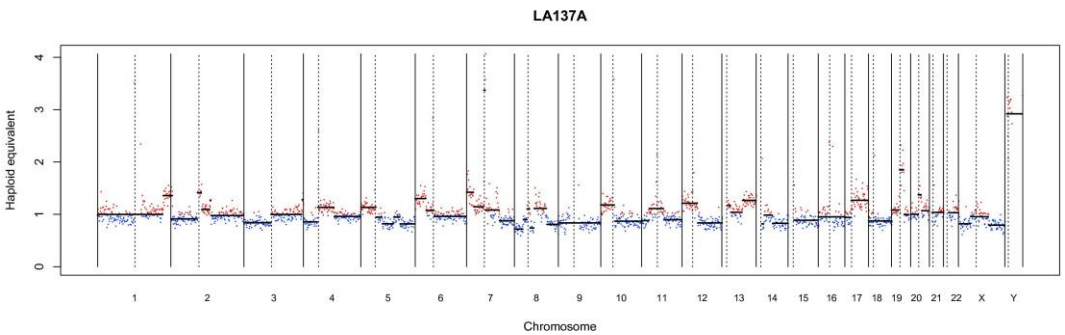
Elute in 40 µL of EB Buffer, as previously described and perform library QC step.

Appendix B
Karyograms

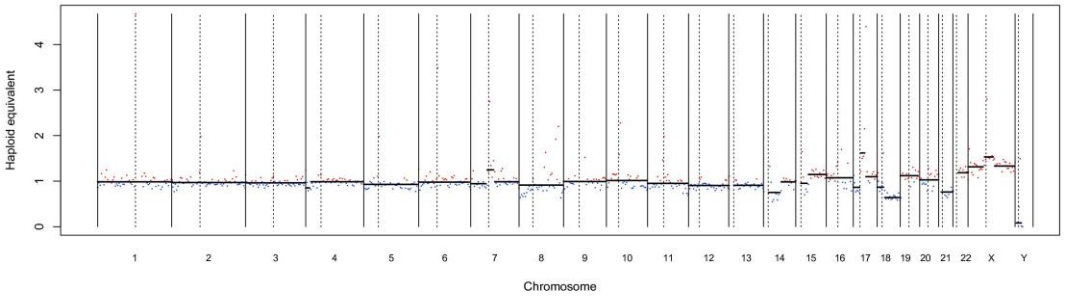




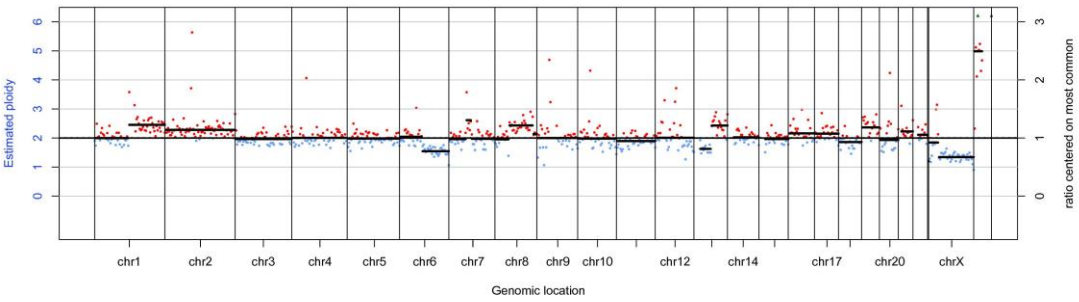




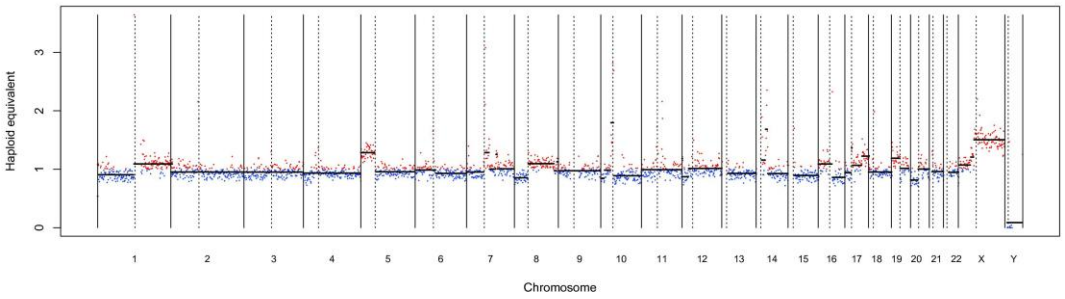
LA146



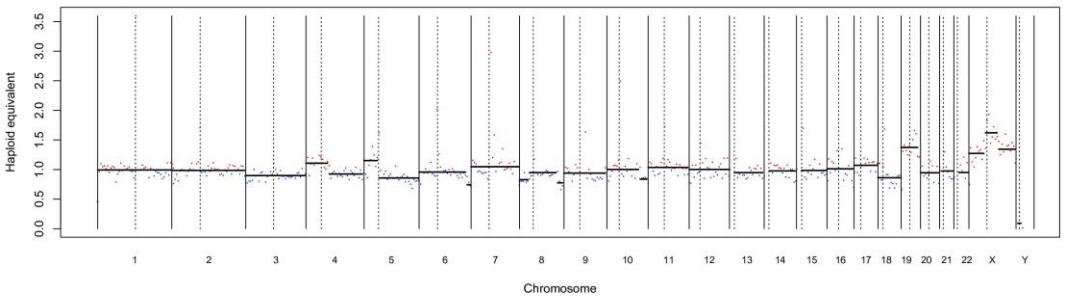
LA149_1X4



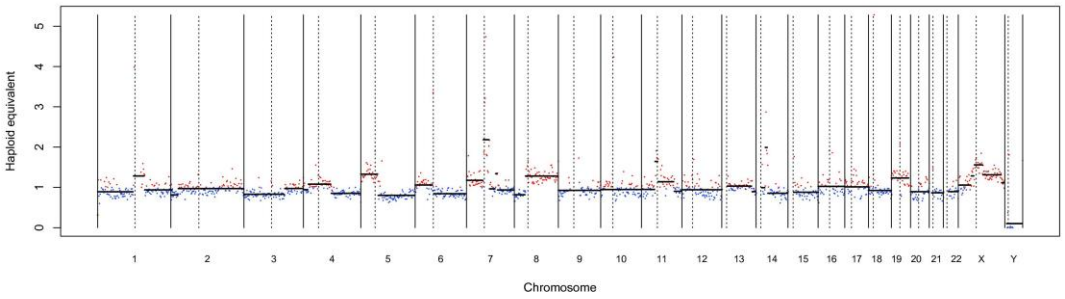
LA152



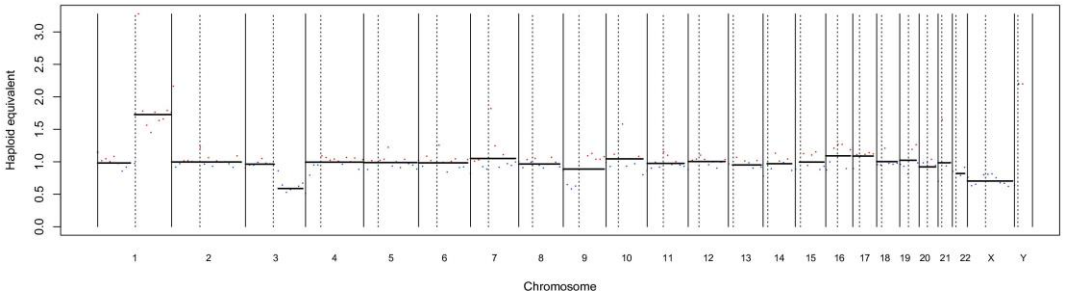
LA153



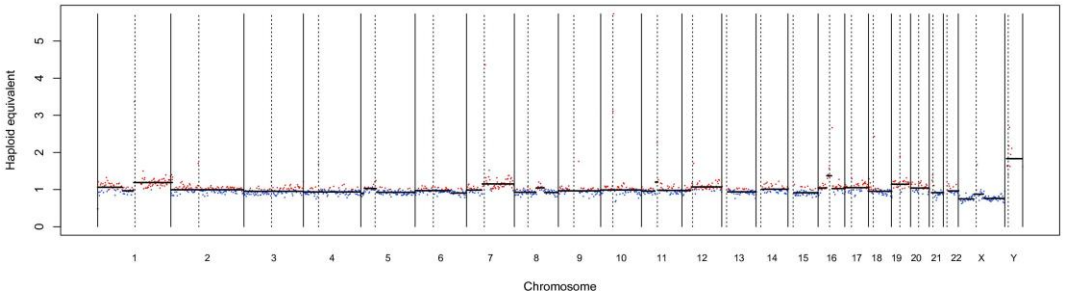
LA158



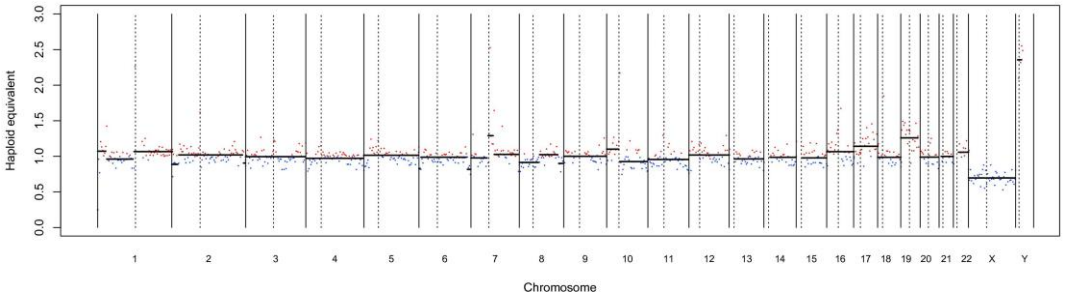
LA160



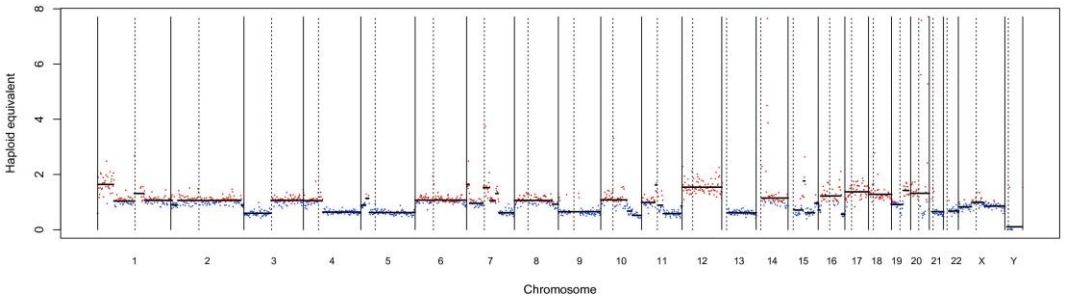
LA169



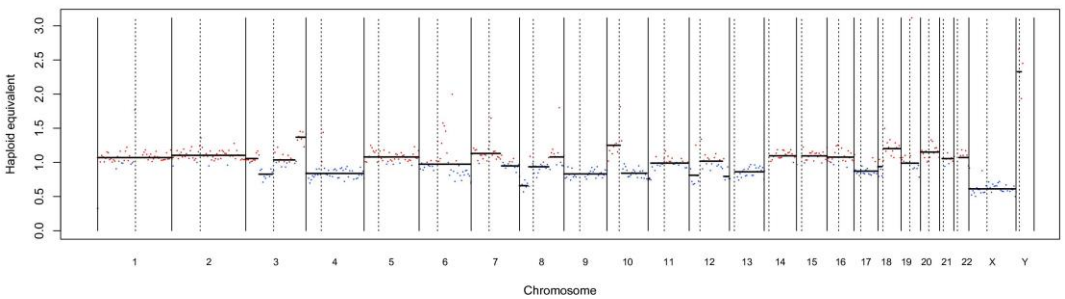
LA170



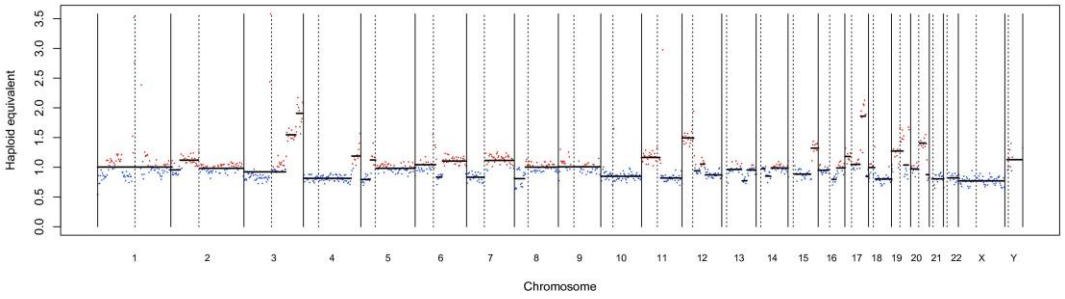
LA172



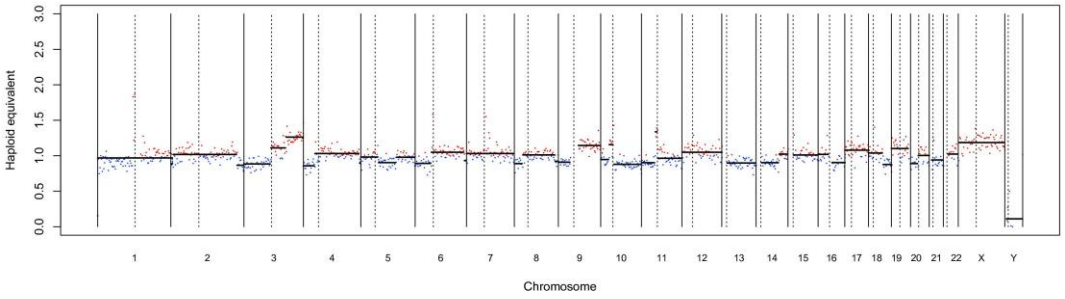
TMA-4



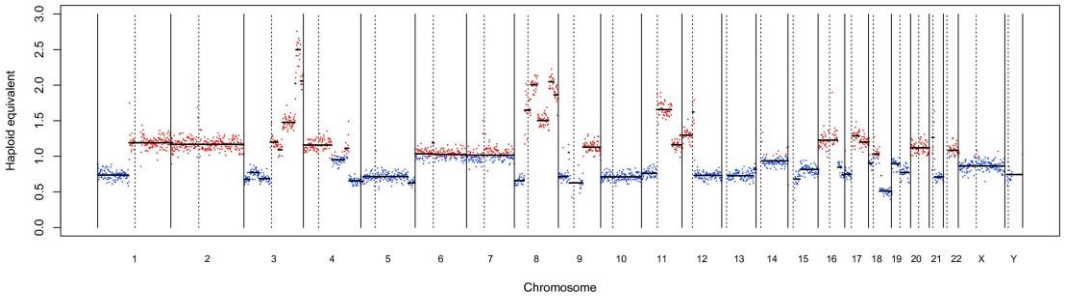
TMA-6



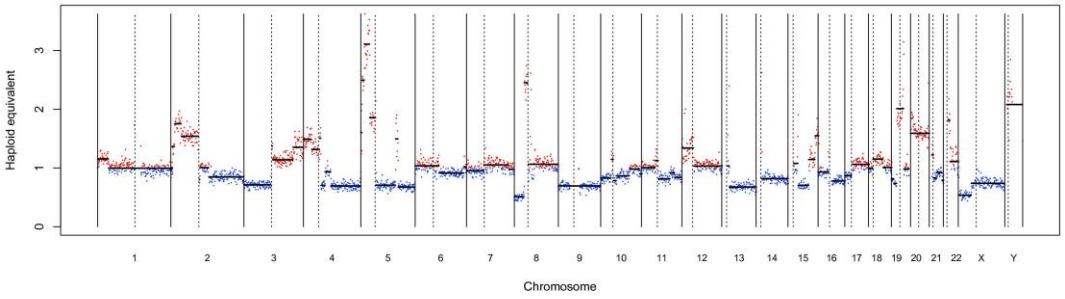
TMA-16



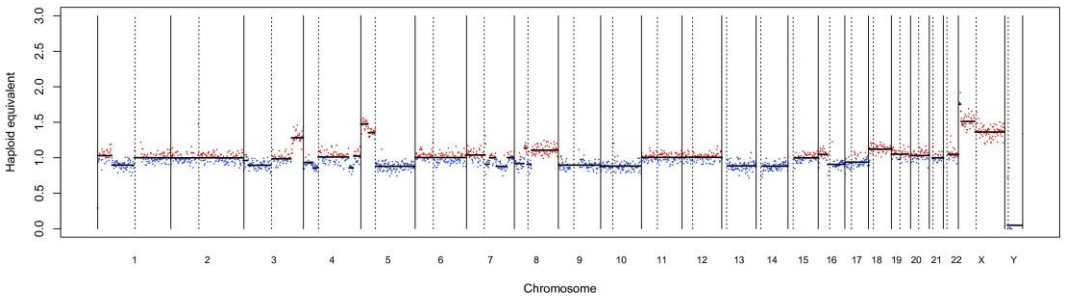
TMA-20



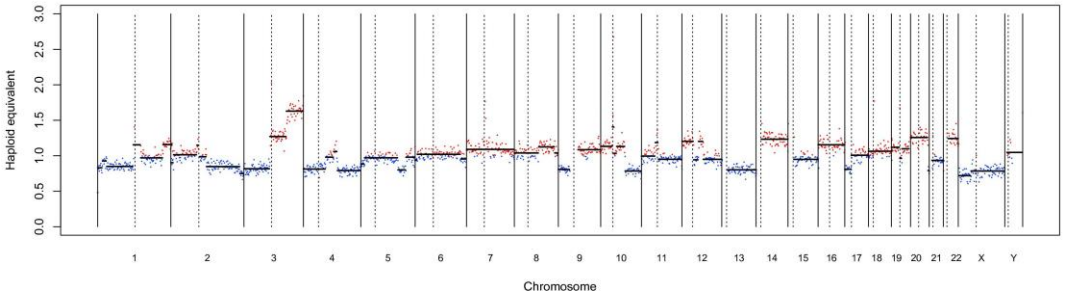
TMA-25



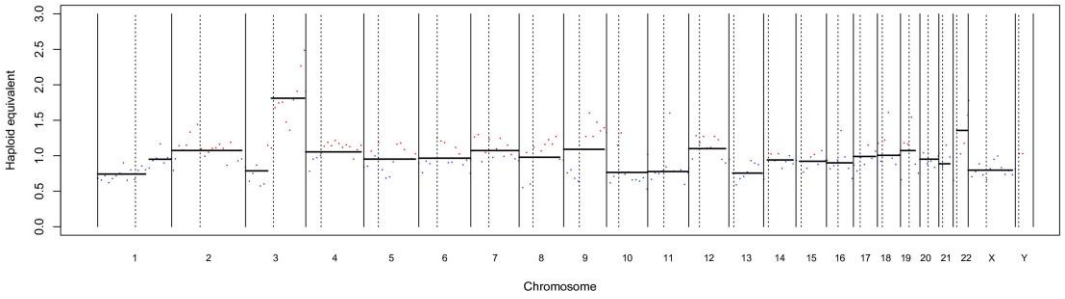
TMA-33



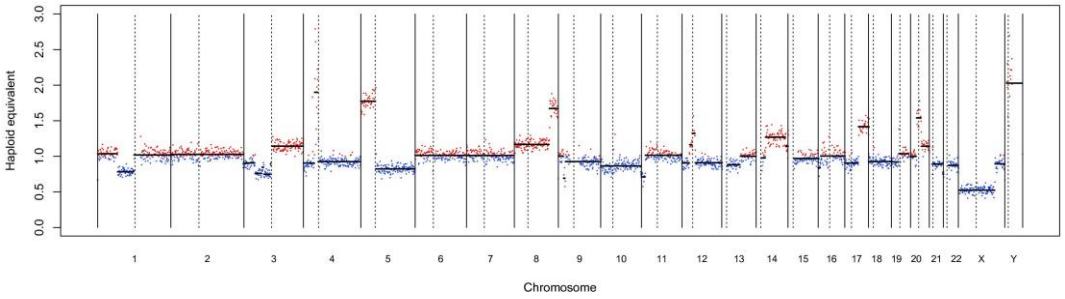
TMA-37



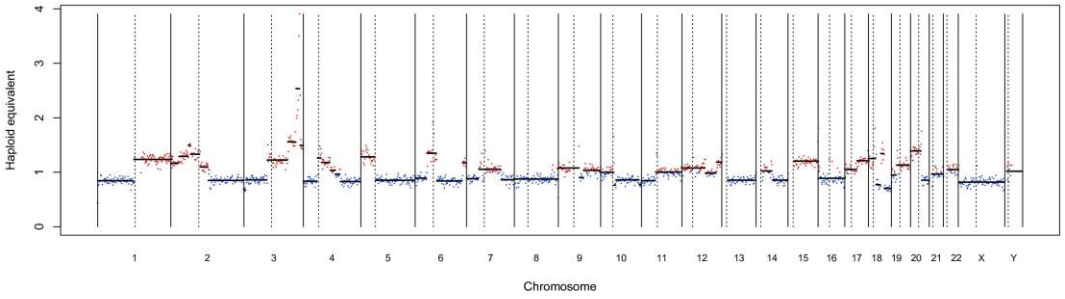
TMA-39



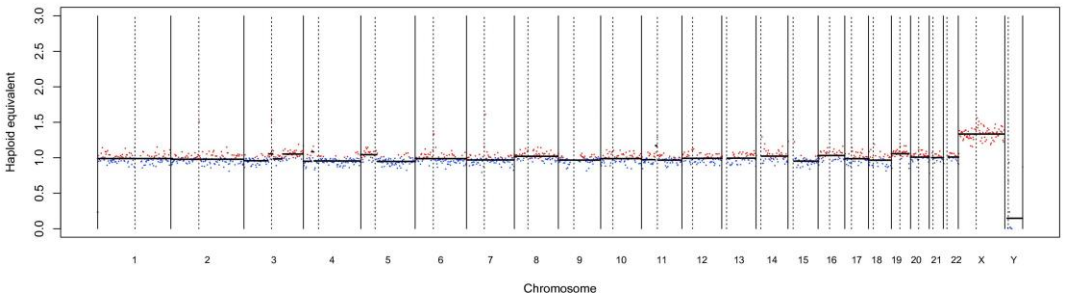
TMA-40



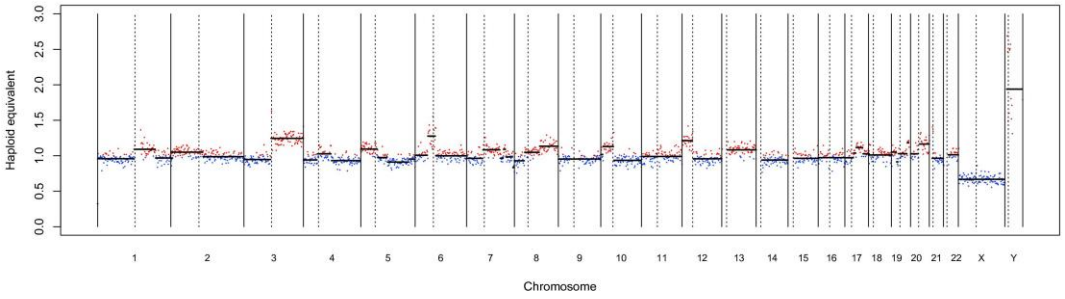
TMA-41



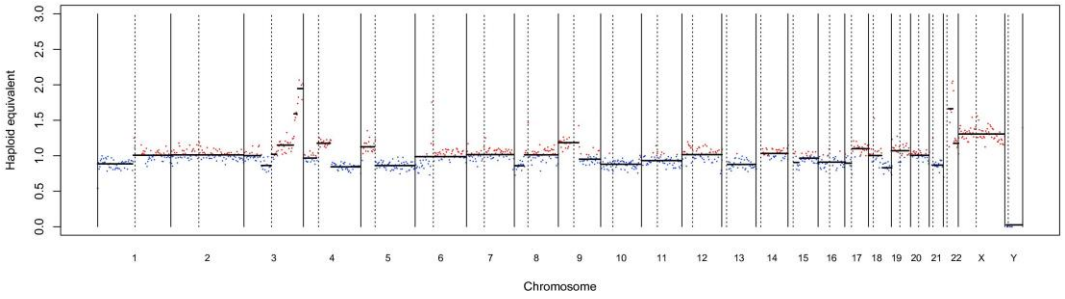
TMA-60



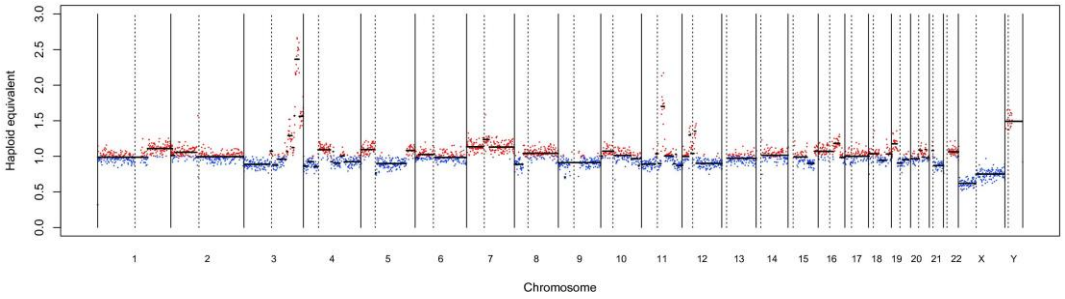
TMA-63



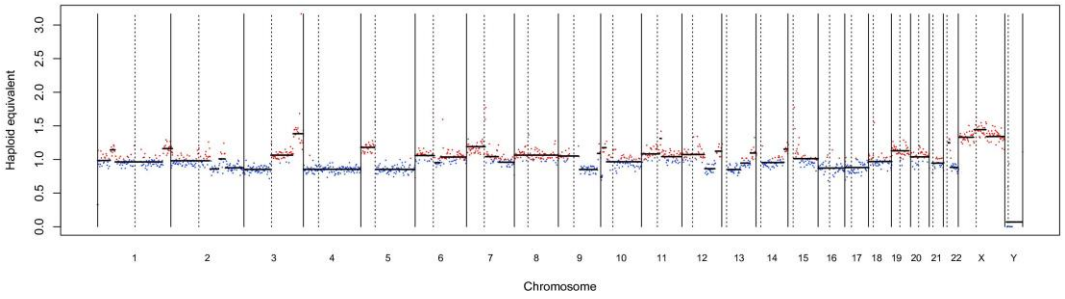
TMA-64



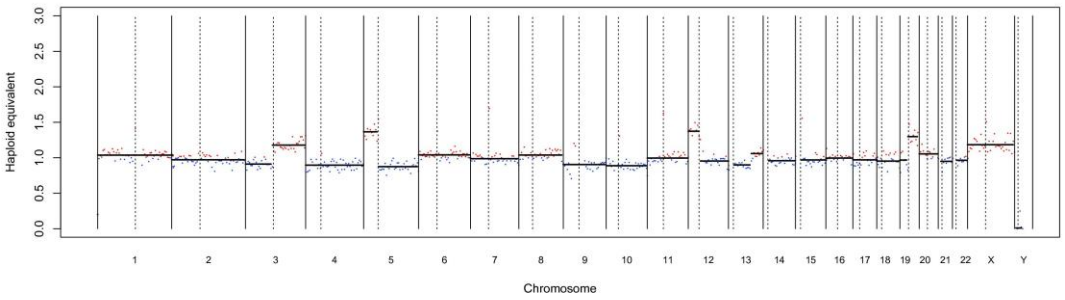
TMA-74



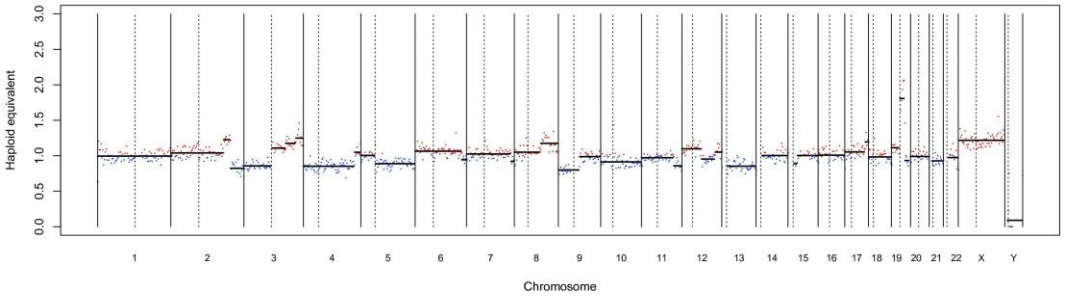
TMA-80



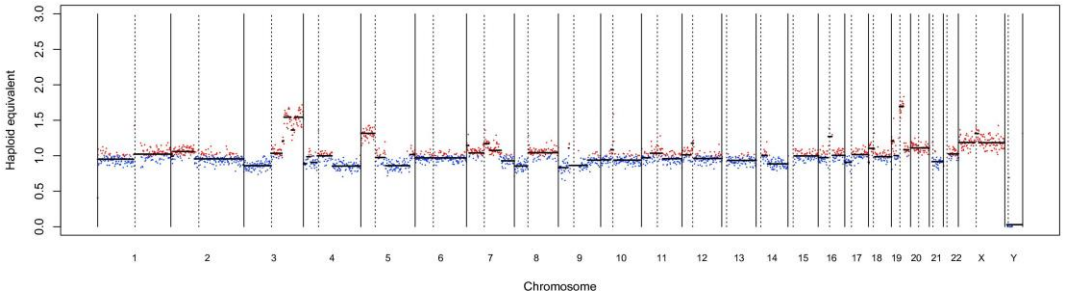
TMA-84



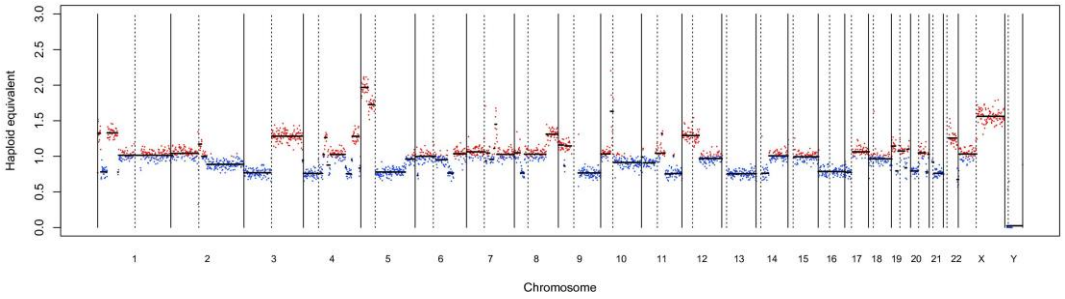
TMA-86



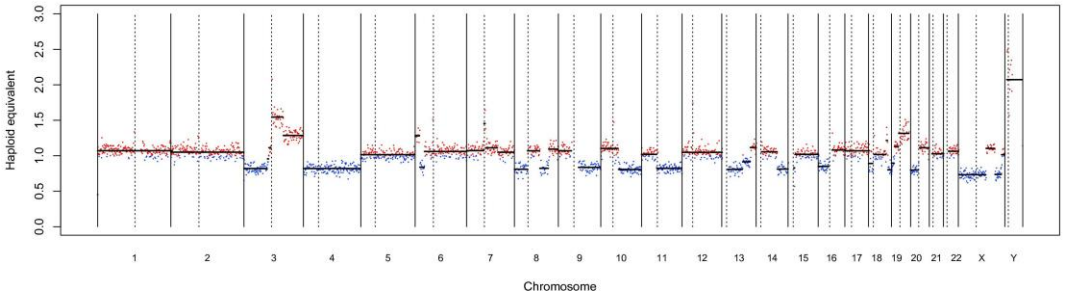
TMA-88



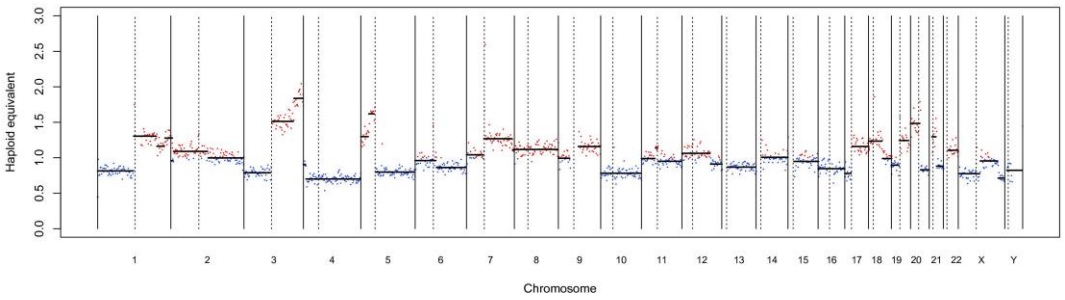
TMA-93



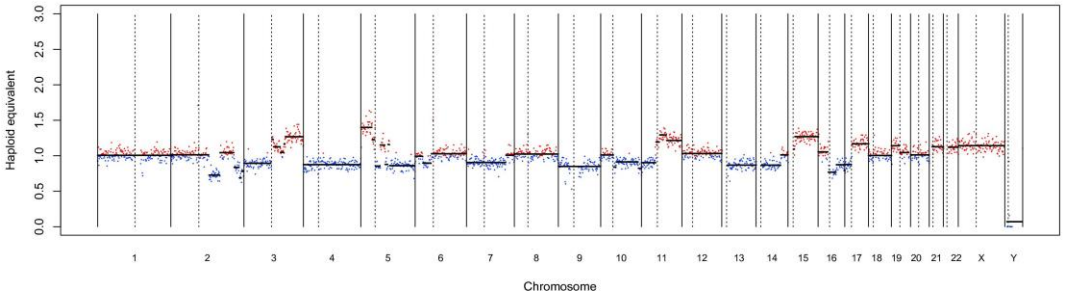
TMA-95



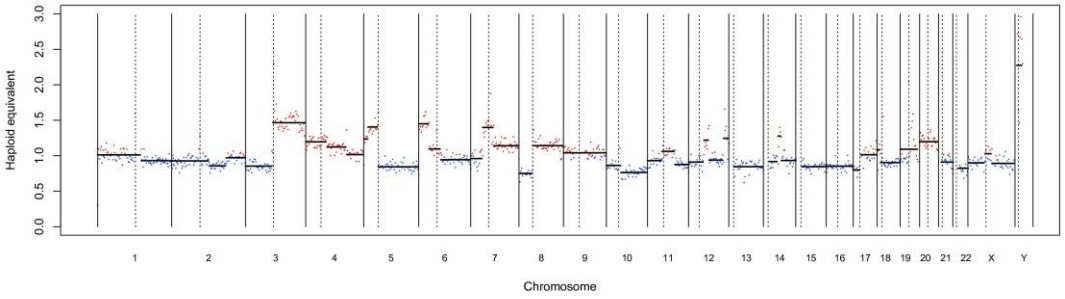
TMA-96



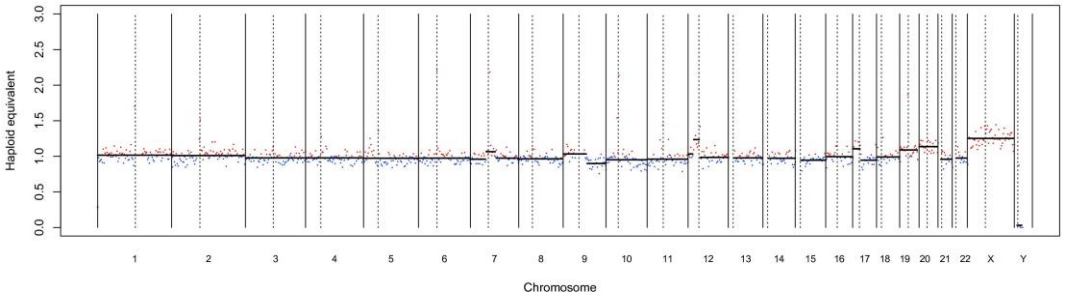
TMA-97



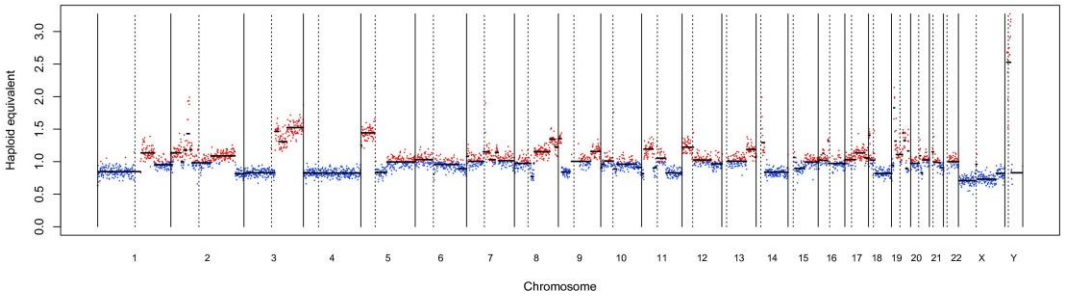
TMA-98



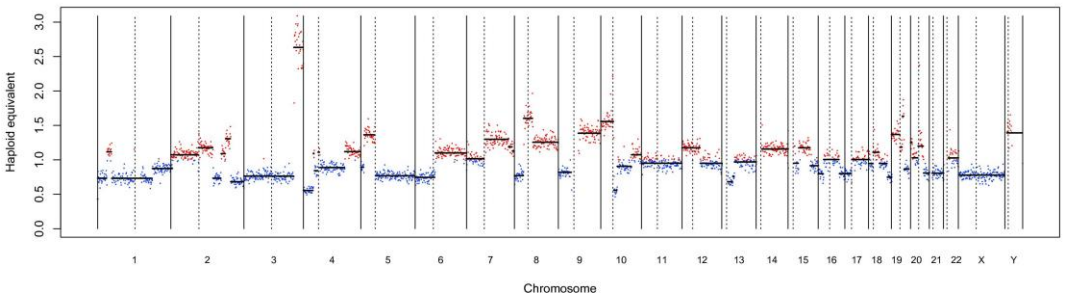
TMA-113



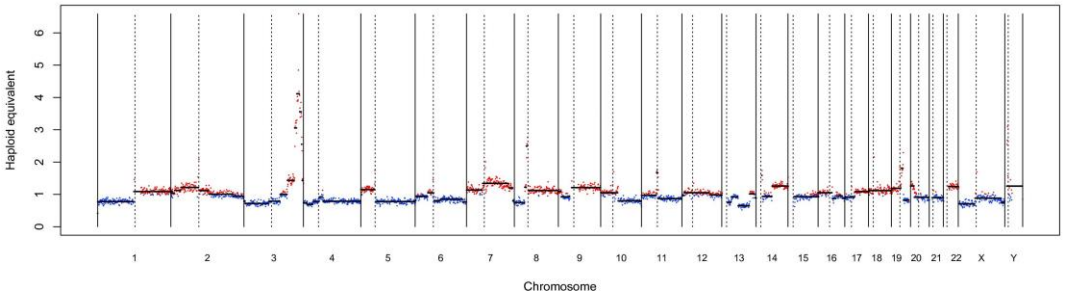
TMA-122



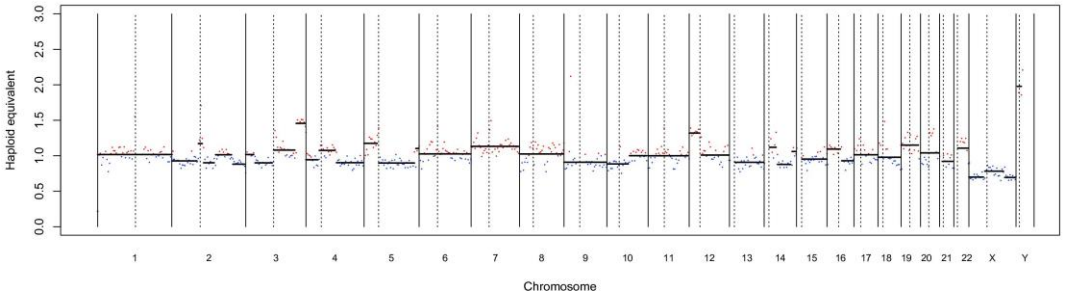
TMA-127



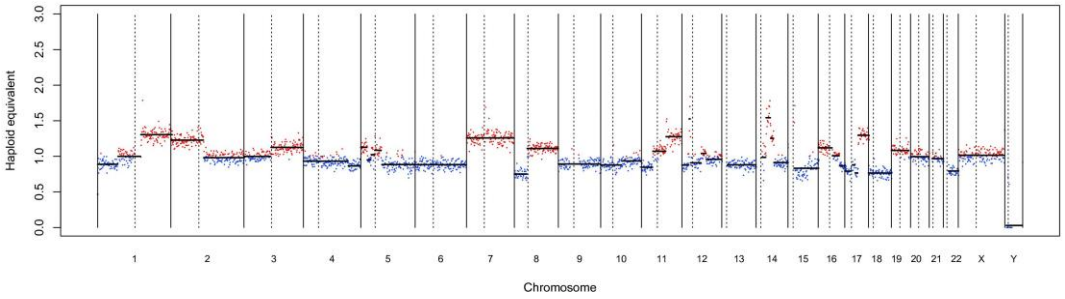
TMA-130



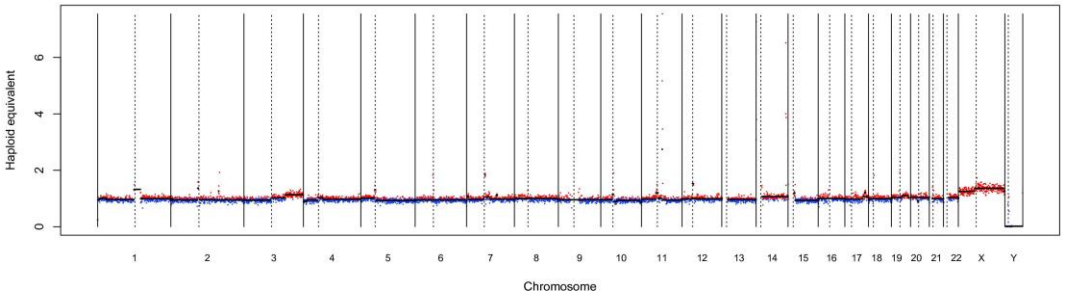
TMA-162



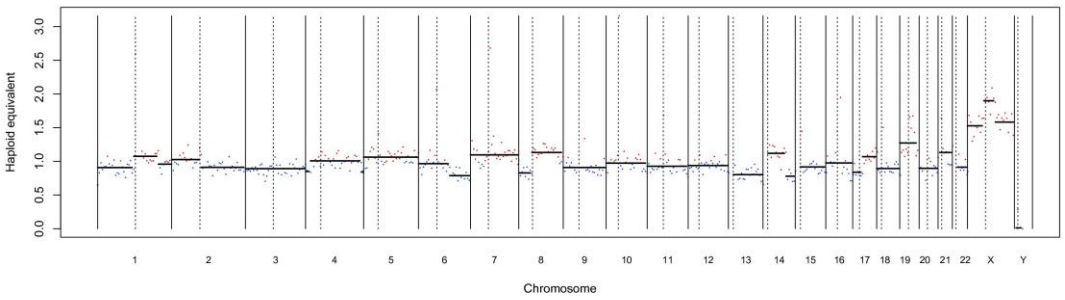
TMA-171

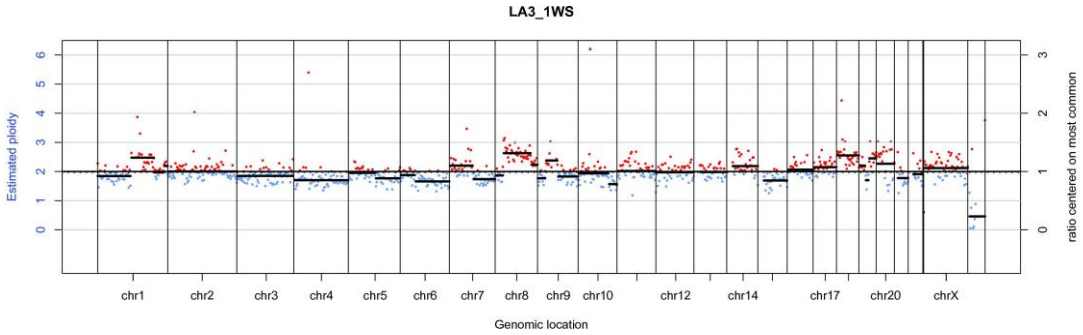


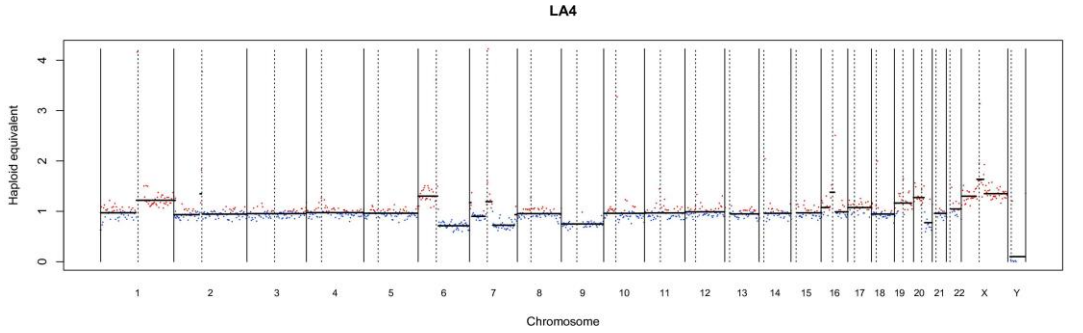
TMA-172

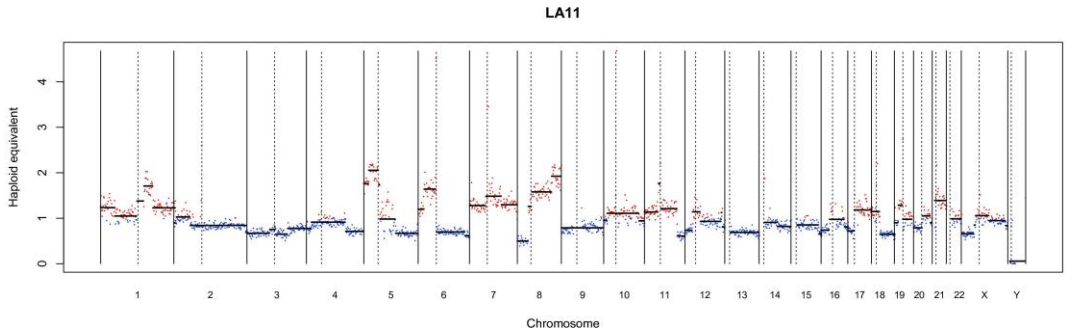


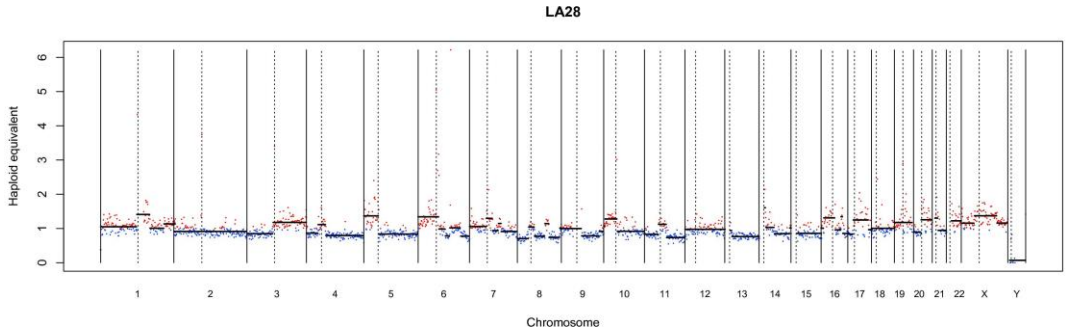
LA1A

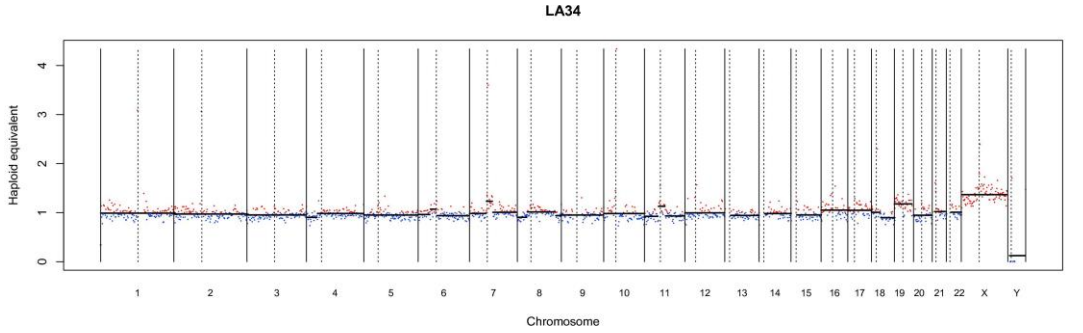



LA3_1WS
Estimated ploidy
ratio centered on most common
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr12
chr14
chr17
chr20
chrX
Genomic location

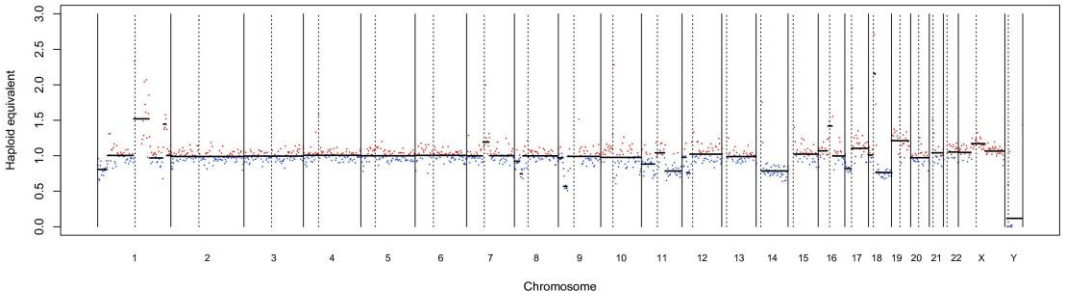

LA4
Haploid equivalent
Chromosome


LA11
Haploid equivalent
Chromosome

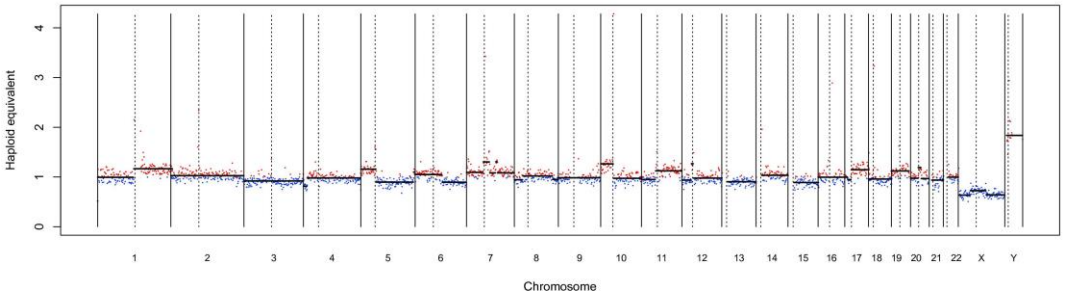

LA28
Haploid equivalent
Chromosome


LA34
Haploid equivalent
Chromosome

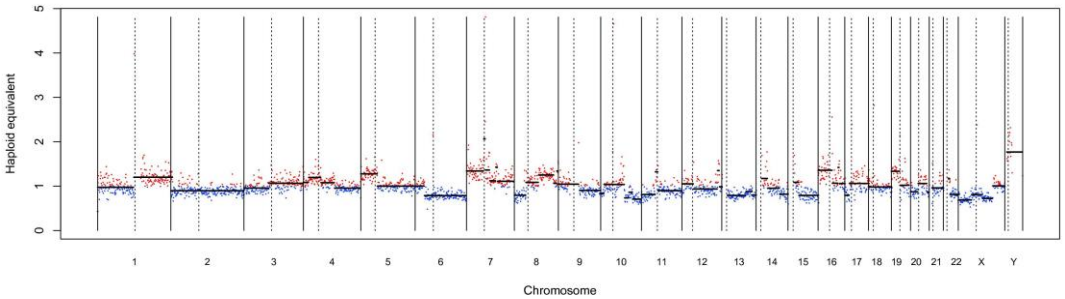
LA37



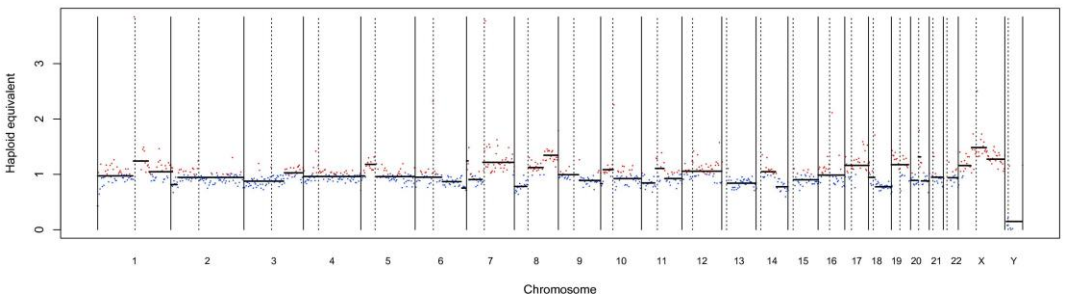
LA56



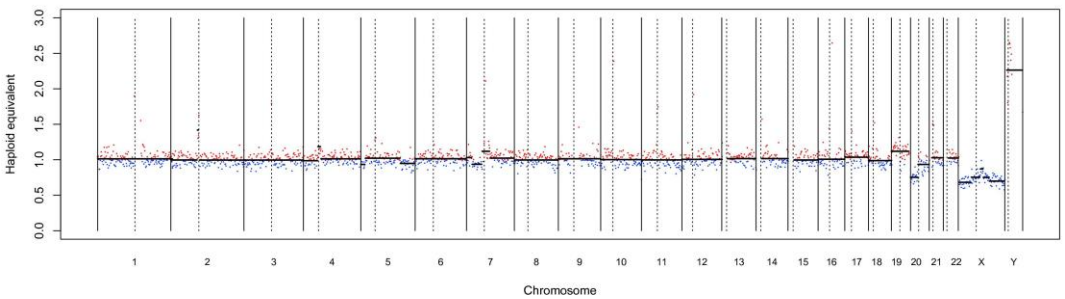
LA57

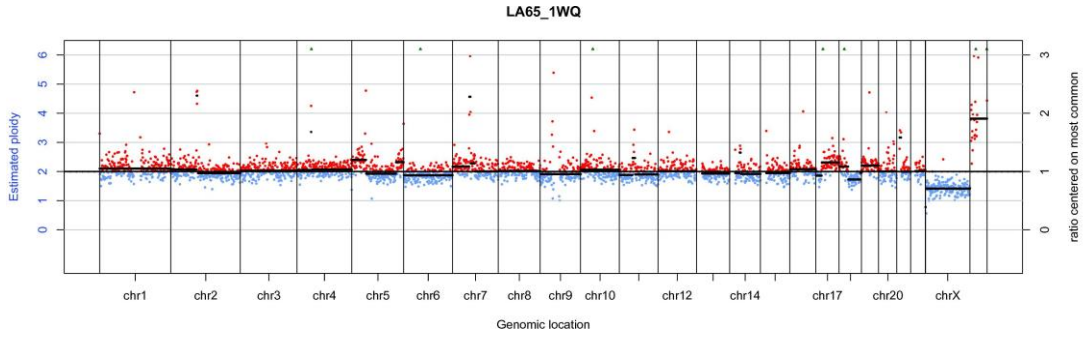


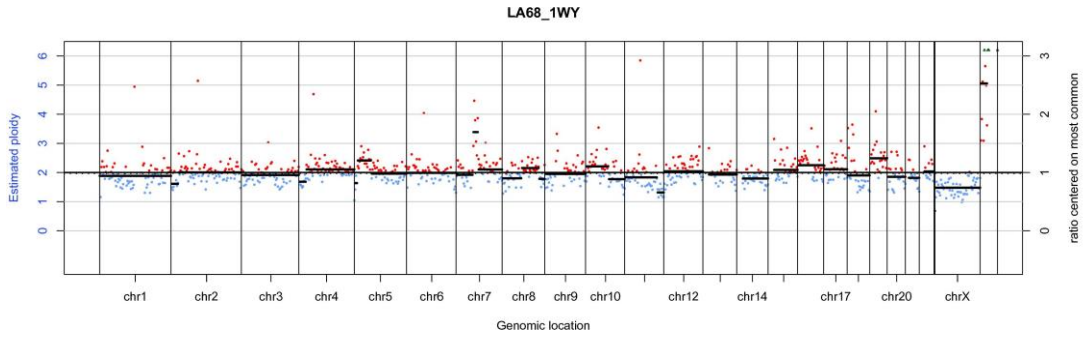
LA59

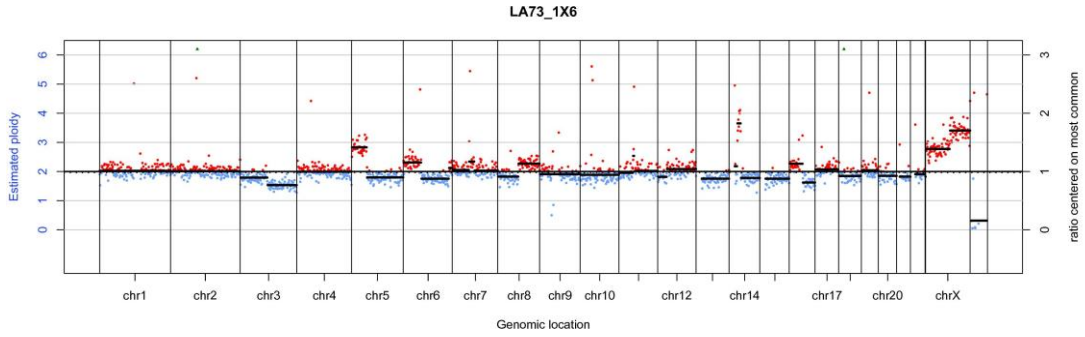


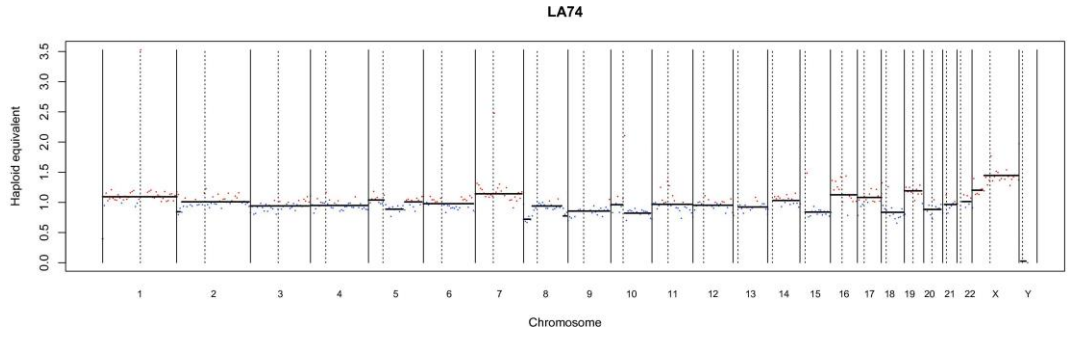
LA61

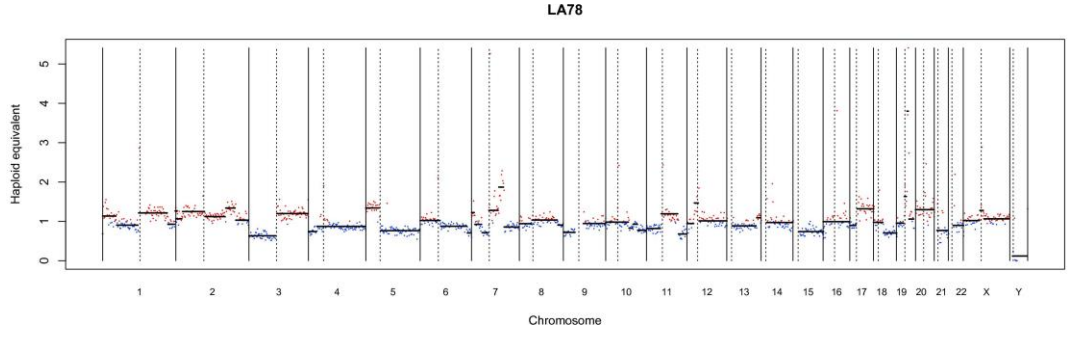



LA65_1WQ
Estimated ploidy
ratio centered on most common
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr12 chr14 chr17 chr20 chrX
Genomic location

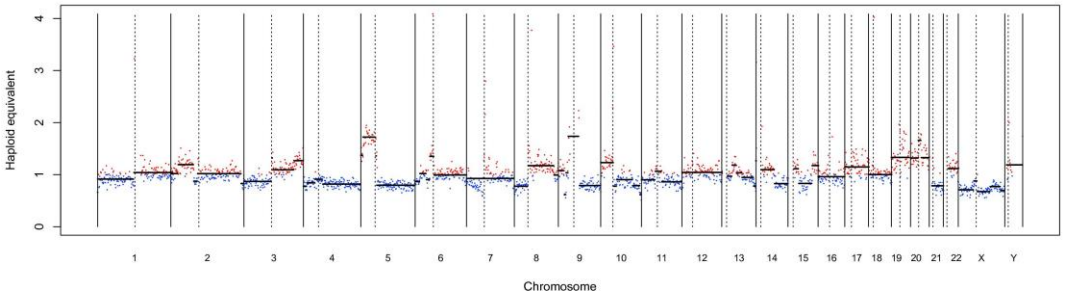

LA68_1WY
Estimated ploidy
ratio centered on most common
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr12 chr14 chr17 chr20 chrX
Genomic location


LA73_1X6
Estimated ploidy
ratio centered on most common
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr12 chr14 chr17 chr20 chrX
Genomic location

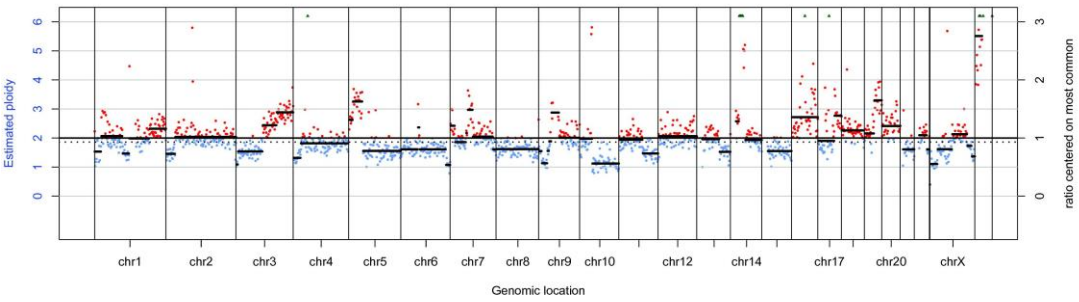

LA74
Haploid equivalent
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome


LA78
Haploid equivalent
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome

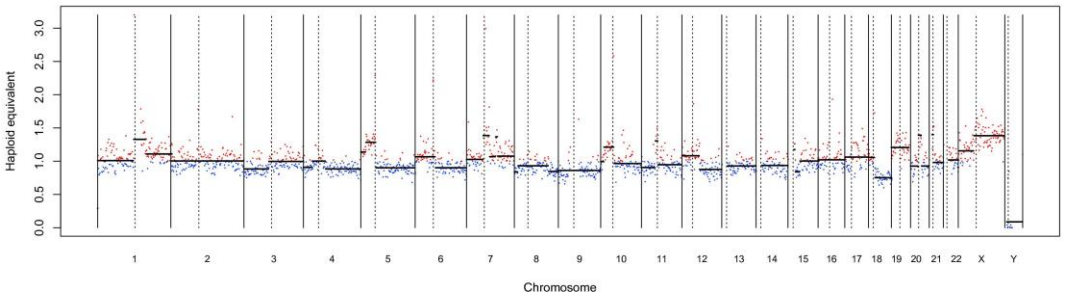
LA83



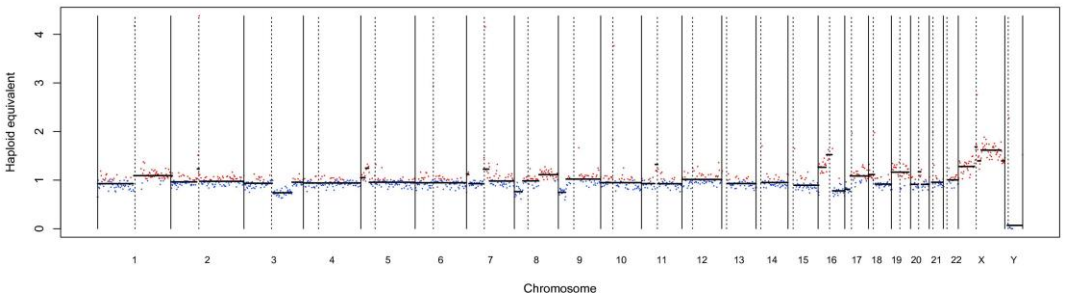
LA84_1WN



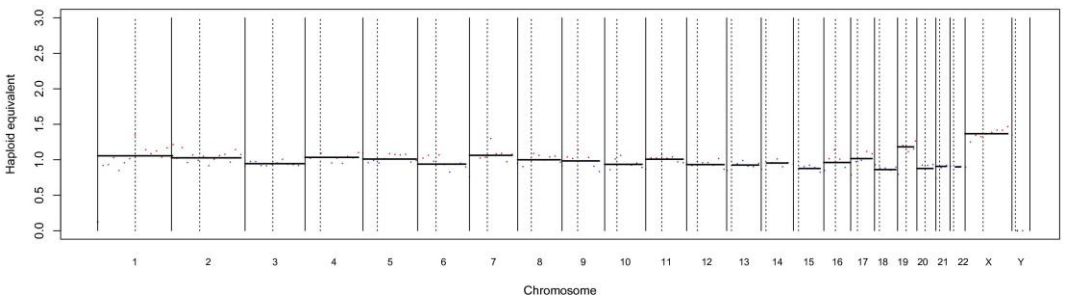
LA87

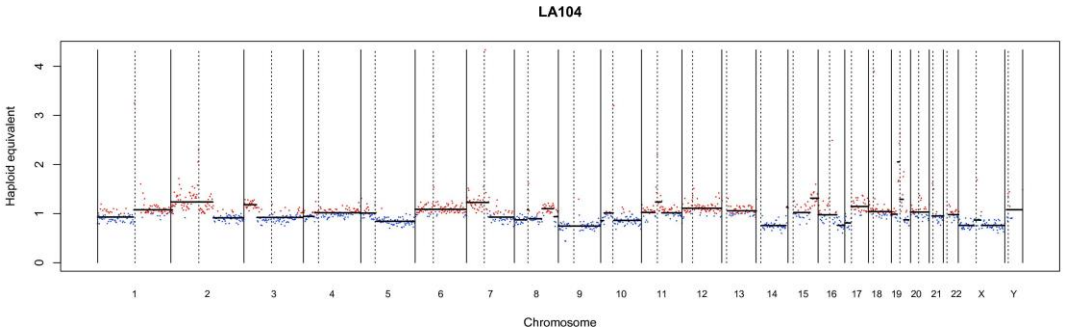


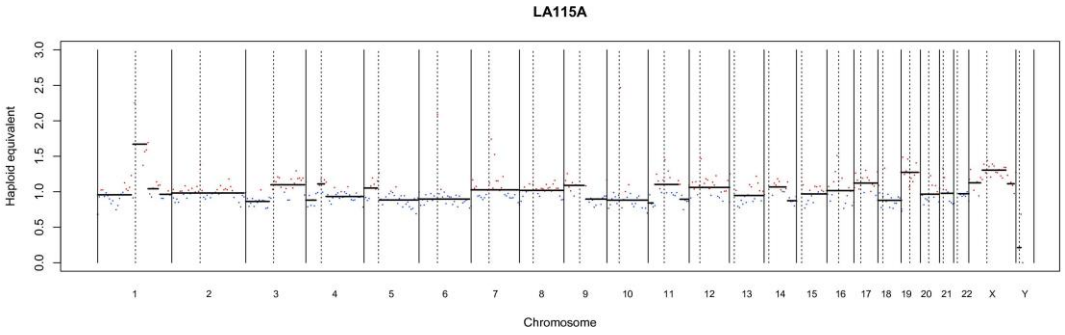
LA95

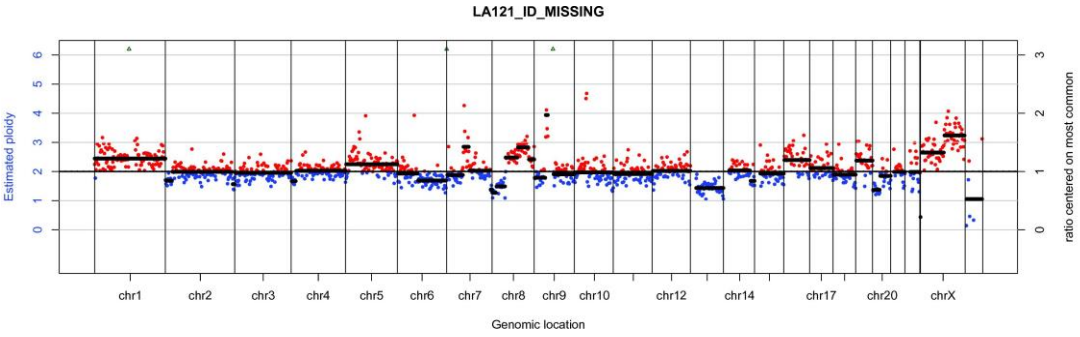


LA99




LA104
Haploid equivalent
0
1
2
3
4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Chromosome


LA115A
Haploid equivalent
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Chromosome


LA121_ID_MISSING
Estimated ploidy
0
1
2
3
4
5
6
ratio centered on most common
0
1
2
3
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr12
chr14
chr17
chr20
chrX
Genomic location

Appendix C

Publications

‘‘Stratifying tumour subtypes based on copy number alteration profiles using next-generation sequence data”. Gusnanto A, Tcherveniakov P, Shuweihdi F, Samman M, Rabbitts P, Wood HM. Bioinformatics. 2015 Apr 5

‘‘A computational index derived from whole-genome copy number analysis is a novel tool for prognosis in early stage lung squamous cell carcinoma. Belvedere O, Berri S, Chalkley R, Conway C, Barbone F, Pisa F, MacLennan K, Daly C, Alsop M, Morgan J, Menis J, Tcherveniakov P, Papagiannopoulos K, Rabbitts P, Wood HM.” Genomics. 2012 Jan;99(1):18-24. doi: 10.1016/j.ygeno.2011.10.006. Epub 2011 Oct 25.

Presentations

“Discovering genomic biomarkers of progression in stage I NSCLC”. Poster presentation. Presented at the Royal College of Surgeons (En). June 2010

“Genomic biomarkers of recurrence in stage I non-small cell lung cancer" presented at the annual ESTS meeting in Marseille, France. June 2011

“Genomic biomarkers and their potential role in management of non-small cell lung cancer”. West Yorkshire Deanery regional teaching of lung cancer. Leeds, July 2011.

“My life in research. A surgical perspective”. Oral presentation. March 2012. Departmental audit meeting of the thoracic unit in St, James’s Hospital, Leeds. November 2011

“A logistic regression model for predicting recurrence in stage I non-small cell lung cancer based on copy number variation”. Oral presentation at the 29th EACTS Annual Meeting. Amsterdam. 6 October 2015.

List of References

Adebonojo SA, Bowser AN, Moritz DM, Corcoran PC. (1999) Impact of revised stage classification of lung cancer on survival: a military experience. Chest, 115, 1507–13.

Adhikary S, Eilers M. (2005) Transcriptional regulation and transformation by Myc proteins. Nat Rev Mol Cell Biol, 6, 635–645.

Arriagada R, Bergman B, Dunant A, Le Chevalier T, Pignon JP, Vansteenkiste J; International Adjuvant Lung Cancer Trial Collaborative Group. (2004) Cisplatin-based adjuvant chemotherapy in patients with completely resected non-small-cell lung cancer. N Engl J Med, 60, 350:351.

Arriagada R, Dunant A, Pignon JP, Bergman B, Chabowski M, Grunenwald D, Kozlowski M, Le Péchoux C, Pirker R, Pinel MI, Tarayre M, Le Chevalier T. (2010) Long-term results of the international adjuvant lung cancer trial evaluating adjuvant Cisplatin-based chemotherapy in resected lung cancer. J Clin Oncol, 28, 35-42.

Arteaga, C.L., Moulder, S.L. and Yakes, F.M. (2002) HER (erbB) tyrosine kinase inhibitors in the treatment of breast cancer. Semin. Oncol, 29, 4-10.

Bass AJ, Watanabe H, Mermel CH, Yu S, Perner S, Verhaak RG, Kim SY, Wardwell L, Tamayo P, Gat-Viks I, Ramos AH, Woo MS, Weir BA, Getz G, Beroukhim R, O'Kelly M, Dutt A, Rozenblatt-Rosen O, Dziunycz P, Komisarof J, Chirieac LR, Lafargue CJ, Scheble V, Wilbertz T, Ma C, Rao S, Nakagawa H, Stairs DB, Lin L, Giordano TJ, Wagner P, Minna JD, Gazdar AF, Zhu CQ, Brose MS, Cecconello I, Jr UR, Marie SK, Dahl O, Shivdasani RA, Tsao MS, Rubin MA, Wong KK, Regev A, Hahn WC, Beer DG, Rustgi AK, Meyerson M. (2009) SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet, 41, 1238-42.

Beer DG, Kardia SL, Huang C, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. (2002) Gene expression profiles predict survival of patients with lung adenocarcinoma. Nat Med, 8, 816–24.

Belvedere O, Berri S, Chalkley R, Conway C, Barbone F, Pisa F, Maclennan K, Daly C, Alsop M, Morgan J, Menis J, Tcherveniakov P, Papagiannopoulos K, Rabbitts P, Wood H. (2012). A computational index derived from whole-genome

copy number analysis is a novel tool for prognosis in early stage lung squamous cell carcinoma. Genomics, 99, 18–24.

Bigbee WL, Gopalakrishnan V, Weissfeld JL, Wilson DO, Dacic S, Lokshin AE, Siegfried JM. (2012) A Multiplexed Serum Biomarker Immunoassay Panel Discriminates Clinical Lung Cancer Patients from High-Risk Individuals Found to be Cancer-Free by CT Screening. J Thorac Oncol, 7, 698-708.

Bloomfield CD, Goldman A, Hossfeld D, de la Chapelle A. (1984) Clinical significance of chromosomal abnormalities in acute nonlymphocytic leukemia. The Fourth International Workshop on Chromosomes in Leukemia. Cancer Genet Cytogenet, 11, 332–350.

Boelens MC, Kok K, van der Vlies P, van der Vries G, Sietsma H, Timens W,Postma DS, Groen HJ, van den Berg A. (2009) Genomic aberrations in squamous cell lung carcinoma related to lymph node or distant metastasis. Lung Cancer, 66, 372-8.

Borczuk A, Toonkel R Powell C. (2010) Genomics of lung cancer. Proc Am Thorac Soc, 6, 152-58.

Broet P, Camilleri-Broet S, Zhang S, Alifano M, Bangarusamy D, Battistella M, Wu Y, Tuefferd M, Régnard JF, Lim E, Tan P, Miller LD. (2009) Prediction of clinical outcome in multiple lung cancer cohorts by integrative genomics: implications for chemotherapy selection. Cancer Res, 69, 1055–1062.

Bunn P. Molecular biology and early diagnosis in lung cancer. (2002) Lung Cancer 38, 5-8.

Chanin TD, Merrick DT Franklin WA Hirsch FR. (2010) Recent developments in biomarkers for the early detection of lung cancer: perspectives based on publications 2003 to present. Curr Opin Pulm Med, 10, 242-47.

Chen DT, Hsu YL, Fulp WJ, Coppola D, Haura EB, Yeatman TJ, Cress WD. (2011) Prognostic and predictive value of a malignancy-risk gene signature in early-stage non-small cell lung cancer. J Natl Cancer Inst, 103, 1859-70.

Chen HY, Yu SL Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF, Chan WK, Li HN, Liu CC, Singh S, Chen WJ, Chen JJ, Yang PC. (2007) A Five-Gene Signature and Clinical Outcome in Non–Small-Cell Lung Cancer. N Engl J Med, 356, 11-20.

Chitale D, Gong Y, Taylor BS, Broderick S, Brennan C, Somwar R, Golas B, Wang L, Motoi N, Szoke J, Reinersman JM, Major J, Sander C, Seshan VE, Zakowski

MF, Rusch V, Pao W, Gerald W, Ladanyi M. (2009) An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumours. Oncogene, 28, 2773–2783.

Chujo M, Noguchi T, Miura T, Arinaga M, Uchida Y, Tagawa Y. (2002) Comparative genomic hybridization analysis detected frequent overrepresentation of chromosome 3q in squamous cell carcinoma of the lung. Lung Cancer, 38, 23-9.

D'Amico T. (2008) Molecular biologic staging of lung cancer. Ann Thorac Surg. 85, 37-42.

D'Amico T, Massey M Herndon J, Moore MB, Harpole DH Jr. (1999) A biologic risk model for stage I lung cancer: immunohistochemical analysis of 408 patients with the use of ten molecular markers. J Thorac Cardiovasc Surg, 117, 736-43.

Dalton, W.S. & Friend, S.H. (2006) Cancer biomarkers—an invitation to the table. Science, 312, 1165-1168.

Dela Cruz C, Tanoue L Matthay R. (2009) "Lung Cancer: Epidemiology and Carcinogenesis in General Thoracic Surgery. " General thoracic surgery. Ed. Locicero J Shields T. 7th revised ed. Philadelphia: Lippincott Williams and Wilkins.

de Ronde J, Klijn C, Velds A, Holstege H, Reinders M, Jonkers J, Wessels L. (2009) KC-SMARTR: An R package for detection of statistically significant aberrations in multi-experiment aCGH data. BMC Res Notes, 3, 298.

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. Nature, 455, 1069-1075.

Douillard JY, Tribodet H, Aubert D, Shepherd FA, Rosell R, Ding K, Veillard AS, Seymour L, Le Chevalier T, Spiro S, Stephens R, Pignon JP; LACE Collaborative

Group. (2010) Adjuvant cisplatin and vinorelbine for completely resected non-small cell lung cancer: subgroup analysis of the Lung Adjuvant Cisplatin Evaluation. J Thorac Oncol, 5, 220-8.

Downward J. (2003) Targeting RAS signalling pathways in cancer therapy. Nat Rev Cancer, 3, 11–22.

Durbin, R. M.; Abecasis, G. R.; Altshuler, R. M.; Auton, G. A. R.; Brooks, D. R.; Durbin, A.; Gibbs, A. G.; Hurles, F. S.; McVean, F. M.; Donnelly, P.; Egholm, M.; Flicek, P.; Gabriel, S. B.; Gibbs, R. A.; Knoppers, B. M.; Lander, E. S.; Lehrach, H.; Mardis, E. R.; McVean, G. A.; Nickerson, D. A.; Peltonen, L.; Schafer, A. J.; Sherry, S. T.; Wang, J.; Wilson, R. K.; Gibbs, R. A.; Deiros, D.; Metzker, M.; Muzny, D.; Reid, J. (2010). A map of human genome variation from population-scale sequencing. Nature, 467, 1061–1073.

Feder M, Siegfried JM, Balshem A, Litwin S, Keller SM, Liu Z, Testa JR. (1998) Clinical relevance of chromosome abnormalities in non-small cell lung cancer. Cancer Genet Cytogenet, 102, 25-31.

Field JK, Oudkerk M, Pedersen JH, Duffy SW. (2013) Prospects for population screening and diagnosis of lung cancer. Lancet, 382, 732-41.

Fong KM, Sekido Y, Gazdar AF, Minna JD. (2003) Molecular biology of lung cancer: clinical implications. Thorax, 58, 892–900.

Gabrilovich DI. (2006) INGN 201 (Advexin): adenoviral p53 gene therapy for cancer. Expert Opin Biol Ther, 6, 823–832.

Go H, Jeon YK, Park HJ, Sung SW, Seo JW, Chung DH. (2010) High MET gene copy number leads to shorter survival in patients with non-small cell lung cancer. J Thorac Oncol, 5, 305–313.

Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S. (2012) Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics. 28, 40-7.

Gusnanto A, Taylor CC, Nafisah I, Wood HM, Rabbitts P, Berri S. (2014) Estimating optimal window size for analysis of low-coverage next-generation sequence data. Bioinformatics, 30, 1823-9.

Herbst R, Heymach J Lippman S. (2008) Molecular Origins of Cancer: Lung Cancer. N Engl J Med, 359, 1367 - 1380.

Hirsch FR, Varella-Garcia M, Cappuzzo F (2009) Predictive value of EGFR and HER2 overexpression in advanced non-small-cell lung cancer. Oncogene, 28, 32–37.

Hicks J, Krasnitz A, Lakshmi B, Navin NE, Riggs M, Leibu E, Esposito D, Alexander J, Troge J, Grubor V, Yoon S, Wigler M, Ye K, Borresen-Dale AL, Naume B, Schlicting E, Norton L, Hagerstrom T, Skoog L, Auer G, Maner S, Lundin P, Zetterberg A. (2006) Novel patterns of genome rearrangement and their association with survival in breast cancer. Genome research, 16, 1465-1479.

Hoffman PC, Mauer AM, Vokes EE. (2000) Lung cancer. Lancet, 355, 479-85.

Huang YT, Heist RS, Chirieac LR, Lin X, Skaug V, Zienolddiny S, Haugen A, Wu MC, Wang Z, Su L, Asomaning K, Christiani DC. (2009) Genome-wide analysis of survival in early-stage non-small-cell lung cancer. J Clin Oncol, 27, 2660-2667.

Johansson M, Karauzum SB, Dietrich C, Mandahl N, Hambraeus G, Johansson L, Clausen PP, Mitelman F, Heim S (1994): Karyotypic abnormalities in adenocarcinomas of the lung. Int J Oncol, 5, 17–26.

Kato H, Ichinose Y, Ohta M et al. (2004) A randomized trial of adjuvant chemotherapy with uracil-tegafur for adenocarcinoma of the lung. N Engl J Med, 350, 1713–21.

Kijima T, Maulik G, Salgia R. Molecular Alterations in Lung Cancer. Methods Mol Med, 75, 29-38.

Kim TM, Yim SH, Lee JS, Kwon MS, Ryu JW, Kang HM, Fiegler H, Carter NP, Chung YJ. (2005) Genome-wide screening of genomic alterations and their clinicopathologic implications in non-small cell lung cancers. Clin Cancer Res, 11, 8235-42.

Klijn C, Holstege H, Schut E, Reinders M, de Ridder J, Jonkers J, Wessels L. (2007) Candidate cancer gene discovery using KC-smart: A novel method for statistical multi-experiment array CGH data analysis. Cellular Oncology, 29, 121-121.

Kubokura H, Tenjin T, Akiyama H, Koizumi K, Nishimura H, Yamamoto M, Tanaka S. (2001) Relations of the c-myc gene and chromosome 8 in non-small cell lung cancer: analysis by fluorescence in situ hybridization. Ann Thorac Cardiovasc Surg, 7, 197–203.

Lamont JP, Kakuda JT, Smith D *et al.* (2002) Systematic postoperative radiologic follow-up in patients with non-small cell lung cancer for detecting second primary lung cancer in stage IA. Arch Surg, 137, 935-938.

Langer CJ, Besse B, Gualberto A, Brambilla E, Soria JC. (2010) The evolving role of histology in the management of advanced non-small-cell lung cancer. J Clin Oncol, 28, 5311–5320.

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25, 1754-60.

Liu H, Kho AT, Kohane IS, Sun Y. (2006) Predicting survival within the lung cancer histopathological hierarchy using a multi-scale genomic model of development. PLoS Med, 3, e232.

Little AG. (2009) No nodes is good nodes. Ann Thorac Surg, 85, 4-5.

Lockwood WW, Wilson IM, Coe BP, Chari R, Pikor LA, Thu KL, Solis LM, Nunez MI, Behrens C, Yee J, English J, Murray N, Tsao MS, Minna JD, Gazdar AF, Wistuba II, Macaulay CE, Lam S, Lam WL. (2012) Divergent Genomic and Epigenomic Landscapes of Lung Cancer Subtypes Underscore the Selection of Different Oncogenic Pathways during Tumour Development. PLoS One, 7, e37775.

Mahabeleshwar GH, Das R, Kundu GC. (2004) Tyrosine kinase, p56lck-induced cell motility, and urokinase-type plasminogen activator secretion involve activation of epidermal growth factor receptor/extracellular signal regulated kinase pathways. J Biol Chem, 279, 9733-42.

Meldrum C, Doyle, M, Tothill R. (2011) Next-Generation Sequencing for Cancer Diagnostics: a Practical Perspective. Clin Biochem Rev, 32, 177–195.

Metzker M. (2010) Sequencing technologies—the next generation. Nat. Rev. Genet, 11, 31-46.

Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT *et al*. (2009) Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med, 361, 947–957.

Molina J, Yang P Cassivi S *et al*. (2008) Non–Small Cell Lung Cancer Epidemiology, Risk Factors,Survivorship. Mayo Clin Proc, 83, 584-94.

Mountain C. Revisions in the International System for Staging Lung Cancer. (1997) Chest, 111, 1710-17.

Mountain C, Dresler C. (1997) Regional lymph node classification for lung cancer staging. Chest, 111, 1718-23.

Muller-Tidow C, Diederichs S, Bulk E, *et al*. (2005) Identification of metastasis-associated receptor tyrosine kinases in non-small cell lung cancer. Cancer Res, 65, 1778-82.

Mulshine, J, Sullivan D. (2005) Clinical practice. Lung cancer screening. N. Engl. J. Med, 352, 2714-2720.

Nahta, R, Esteva1 F. (2003) HER-2-Targeted Therapy – Lessons Learned and Future Directions. Cancer Res, 9, 5078–5048.

Ocak S, Sos M, Thomas R. (2009) High-throughput molecular analysis in lung cancer: insights into biology and potential clinical applications. Eur Respir J, 34, 489-506.

Pairolero P, Williams, D, Bergstralh E *et al.* (1984) Postsurgical stage I bronchogenic carcinoma. Ann Thorac Surg, 38, 331-336.

Panani A, Roussos C. (2006) Cytogenetic and molecular aspects of lung cancer. Cancer Lett, 28, 1-9.

Pao W, Iafrate A, Su Z. (2011) Genetically informed lung cancer medicine. J Pathol, 223, 230-40.

Paris P, Andaya A, Fridlyand J, Jain A, Weinberg V, Kowbel D, Brebner J, Simko J, Watson J, Volik S, Albertson DG, Pinkel D, Alers J, van der Kwast T, Vissers K, Schroder F, Wildhagen M, Febbo P, Chinnaiyan A, Pienta K, Carroll P, Rubin M, Collins C, van Dekken H. (2004). Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumours. Hum Mol Genet. 13, 1303-13.

Pei J, Balsara BR, Li W, Litwin S, Gabrielson E, Feder M, Jen J, Testa JR.Genomic imbalances in human lung adenocarcinomas and squamous cell carcinomas. Genes Chromosomes Cancer. 2001 Jul;31(3):282-7.

Pignon J, Tribodet H, Scagliotti G, Douillard J, Shepherd F, Stephens R, Dunant A, Torri V, Rosell R, Seymour L, Spiro S, Rolland E, Fossati R, Aubert D, Ding K, Waller D, Le Chevalier T. LACE Collaborative Group. (2008) Lung adjuvant cisplatin evaluation: a pooled analysis by the LACE Collaborative Group. J Clin Oncol. 20, 3552-9.

Pirker R, Szczesna A, von Pawel J, *et al*. (2008) FLEX: a randomized, multicenter, phase III study of cetuximab in combination with cisplatin/vinorelbine (CV) versus CV alone in the first-line treatment of patients with advanced non-small cell lung cancer (NSCLC). J Clin Oncol, 26, 54-60.

Pisters K, Evans W, Azzoli C, Kris M, Smith C, Desch C, Somerfield M, Brouwers M, Darling G, Ellis P, Gaspar L, Pass H, Spigel D, Strawn J, Ung Y, Shepherd F. (2007) Cancer Care Ontario; American Society of Clinical Oncology.Cancer Care Ontario and American Society of Clinical Oncology adjuvant chemotherapy and adjuvant radiation therapy for stages I-IIIA resectable non-small-cell lung cancer guideline. J Clin Oncol, 25, 5506-5518.

Pollack J, Sorlie T, Perou C *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumours. Proc Natl Acad Sci, 99, 12963–12968.

Richardson G, Johnson B. (1993) The biology of lung cancer. Semin Oncol, 20, 105–27.

Rosell R, Bivona T, Karachaliou N. (2013) Genetics and biomarkers in personalisation of lung cancer treatment. Lancet, 24, 720-31.

Ross J, Cronin M. (2011) Whole cancer genome sequencing by next-generation methods. Am. J. Clin. Pathol, 136, 527–539.

Rubins J, Unger M, Colice G. (2007) American College of Chest Physicians. Follow-up and surveillance of the lung cancer patient following curative intent therapy: ACCP evidence-based clinical practice guideline (2nd edition). Chest, 132, 355S-67S.

Sandberg A. (1991) The Chromosomes in Human Cancer and Leukemia. Mutat Res, 247, 231-40.

Sanger F, Nicklen S, Coulson A. (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci, 74, 5463-5467.

Shokralla S, Spall J, Gibson J *et al*. (2012) Next-generation sequencing technologies for environmental DNA research. Mol Ecol, 21, 1794-805.

Sato M, Shames D, Gazdar A, Minna J. (2007) A translational view of the molecular pathogenesis of lung cancer. J Thorac Oncol, 2, 327-43.

Scagliotti G, Hanna N, Fossella F *et al*. (2009) The differential efficacy of pemetrexed according to NSCLC histology: a review of two Phase III studies. Oncologist, 14, 253–263.

Senzer N, Nemunaitis J. A review of contusugene ladenovec (Advexin) p53 therapy. (2009) Curr Opin Mol Ther, 11, 54-61.

Shepherd F, Rodrigues Pereira J, Ciuleanu T *et al*. (2005) Erlotinib in previously treated non–small-cell lung cancer. N Engl J Med, 353, 123-32.

Siegel R, Naishadham D, Jemal, A. (2013), Cancer statistics 2013. CA Cancer J Clin, 63, 11-30.

Slebos R, Hruban R, Dalesio O *et al*. Relationship between K-ras oncogene activation and smoking in adenocarcinoma of the human lung. J Natl Cancer Inst, 83, 1024–7.

Spira A, Ettinger D. Multidisciplinary Management of Lung Cancer. N Engl J Med, 350, 379-92.

Staaf J, Isaksson S, Karlsson A *et al*. (2013) Landscape of somatic allelic imbalances and copy number alterations in human lung carcinoma. Int J Cancer, 132, 2020–2031.

Stiles B, Altroki N, Yankelevitz D. (2009) Screening for Lung Cancer: Challenges for Thoracic Surgery. General thoracic surgery. Ed. Locicero J Shields T. 7th revised ed. Philadelphia: Lippincott Williams and Wilkins.

Strauss G, Herndon J, Maddus A *et al*. (2004) Randomized clinical trial of adjuvant chemotherapy with paclitaxel and carboplatin following resection in stage IB non-small-cell lung cancer (NSCLC): report of cancer and leukemia group B (CALGB) protocol 9633. J Clin Oncol. 26, 5043-51.

Strauss G & Skarin A. (1994) Use of tumour markers in lung cancer. Hematol. Oncol. Clin. North. Am, 8, 507-532.

Sung H & Cho J. (2008) Biomarkers for the lung cancer diagnosis and their advances in proteomics. BMB reports. 41, 615-625.

Sy S, Wong N, Lee T, Tse G, Mok T, Fan B, Pang E, Johnson P, Yim A. (2004) Distinct patterns of genetic alterations in adenocarcinoma and squamous cell carcinoma of the lung. Eur J Cancer, 40, 1082-94.

Taguchi A, Politi K, Pitteri S, Lockwood W, Faça V, Kelly-Spratt K, Wong C, Zhang Q, Chin A, Park K, Goodman G, Gazdar A, Sage J, Dinulescu D, Kucherlapati R, Depinho R, Kemp C, Varmus H, Hanash S. (2011) Lung cancer signatures in plasma based on proteome profiling of mouse tumour models. Cancer Cell, 13, 289-99.

Tai A, Yan W, Fang Y, Xie D, Sham J, Guan X. Recurrent chromosomal imbalances in non-small cell lung carcinoma: the association between 1q amplification and tumour recurrence. Cancer, 100, 1918-27.

Tang X, Shigematsu H, Bekele B *et al*. (2005) EGFR tyrosine kinase domain mutations are detected in histologically normal respiratory epithelium in lung cancer patients. Cancer Res, 65, 7568-72.

Testa J, Liu Z, Feder M, Bell D *et al*. (1997) Advances in the analysis of chromosome alterations in human lung carcinomas. Cancer Genet Cytogenet. 95, 20–32.

The International Cancer Genome Consortium. (2010) International network of cancer genome projects. Nature, 46, 993-98.

The Cancer Genome Atlas Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature, 489, 519–525.

Travis W, Colby T, Corrin B, Shimosato Y Brambilla E. (2010) Histological Typing of Lung and Pleural Tumours. World Health Organization International Histological Classification of Tumours. Ed. Springer. 3rd ed. New York, 156.

Ulahannan D, Kovac M, Mulholland P, Cazier J, Tomlinson I. (2013) Technical and implementation issues in using next-generation sequencing of cancers in clinical practice. Br J Cancer, 109, 827–835.

Varella-Garcia M. (2010) Chromosomal and genomic changes in lung cancer. Cell Adh Migr, 4, 100–106.

Venkatraman E, Olshen A. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics, 23, 657-663.

Visbal A, Leighl N, Feld R, Shepherd F. (2005) Adjuvant Chemotherapy for Early-Stage Non-small Cell Lung Cancer. Chest, 128, 2933-43.

Voelkerding K, Dames S, Durtschi J. Next-generation sequencing: from basic research to diagnostics. Clin Chem, 55, 641–658.

Wakeleea H, Dubeyb S, Gandarac D. (2007) Optimal adjuvant therapy for non-small cell lung cancer — how to handle stage I disease. Oncologist, 12, 331–7.

Weihua Z, Tsan R, Huang WC, *et al*. (2008) Survival of cancer cells is maintained by EGFR independent of its kinase activity. Cancer Cell, 13, 385-93.

Weir B, Woo M, Getz G *et al*. (2007) Characterizing the cancer genome in lung adenocarcinoma. Nature, 450, 893-8.

Wheeler D, Srinivasan M, Egholm M *et al*. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature, 452, 872-876.

Winton T, Livingston R, Johnson D *et al*. (2005) Vinorelbine plus cisplatin vs. observation in resected non-small-cell lung cancer. N Engl J Med, 352, 2589–97.

Winton T, Livingston R, Johnson D *et al*. (2004) A prospective randomized trial of adjuvant vinorelbine (VIN) and cisplatin (CIS) in completely resected stage IB and II non-small-cell lung cancer (NSCLC) intergroup JBR 10. Proc Am Soc Clin Oncol, 22, 7018.

Wistuba I, Lam S, Behrens C *et al*. (1997) Molecular damage in the bronchial epithelium of current and former smokers. J. Natl. Cancer Inst, 89, 1366-1373.

Wood H, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, McKinley C, Egan P, Ross L, Hayward B, Morgan J, Davidson L, MacLennan K, Ong TK, Papagiannopoulos K, Cook I, Adams DJ, Taylor GR, Rabbitts P. (2010) Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens, Nucleic Acids Res, 38, e151.

Yakut T, Schulten H, Demir A *et al*. (2006). Assessment of molecular events in squamous and non-squamous cell lung carcinoma. Lung Cancer, 54, 293–301.

Yan W, Song L, Liang Q, Fang Y. (2005) Progression analysis of lung squamous cell carcinomas by comparative genomic hybridization. Tumour Biol, 26, 158-64.

Zamoyska R, Basson A, Filby A, Legname G, Lovatt M, Seddon B. (2003) The influence of the src-family kinases, Lck and Fyn, on T cell differentiation, survival and activation. Immunol Rev, 191, 107-18.