Biomarkers of Lung Cancer Risk and Progression

Fiona Taylor
Academic Unit of Molecular Oncology
Registration Number: 120250033

A thesis submitted to the School of Medicine, University of Sheffield in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Academic Unit of Molecular Oncology/ Department of Oncology
School of Medicine
December 2017

Acknowledgments

Firstly, I would like to thank my PhD supervisor Professor Angie Cox for her time, support, patience and encouragement in helping me complete my PhD thesis. I am also very grateful for the guidance and input that I have received from my two other PhD supervisors, namely Professor Woll and Dr Dawn Teare.

I would like to thank everyone that has helped me with my laboratory work at the Academic Unit of Molecular Oncology to include Dr Ian Brock and Dr Anil Ganesh. I gained experience in library preparation and bioinformatics by working with Dr Antonio Milano, Dr Lucy Crooks, Dr Emilie Boardman at Sheffield Children's hospital as well as Dr Paul Heath and Dr Matthew Wyles at the Sheffield Institute for Translational Neuroscience (SITRaN). Sequencing runs with the Illumina HiSeq2500 were carried out at the Sheffield Children's Hospital. I thank Dr Emilie Boardman for chip loading, commencing the sequencing run and for creating our FASTQ files using data obtained from the HiSeq2500 to enable further analyses. I am very grateful for the time, effort and support of our bioinformatics expert Dr James Bradford who was responsible for bioinformatics analyses to establish copy number ratios from the FASTQ files. Furthermore, I would also like to thank Dr Henry Wood at the University of Leeds for his advice regarding bioinformatics processing and low coverage sequencing.

I was fortunate enough to spend time in two laboratories to learn about standard operating procedures for cfDNA analyses. I thank Prof Jacqui Shaw at the University of Leicester and Prof Caroline Dive at the Cancer Research UK Manchester Institute for allowing me to work with their laboratory teams. I am very appreciative of the time that their post docs, Dr Sumitra Mohan (Manchester) and Dr Barbara Ottolini (Leicester), spent with me. Dr Barbara Ottolini carried out all Ion Torrent experiments, I was fortunate enough to observe several experiments, and I partook in results analyses.

Finally, I would like to thank all the people that participated in the ReSoLuCENT study as well as the data managers Dr Janet Horseman and Lesley Bruce, who facilitated access to clinical data and the laboratory technician Dr Helen Shulver who facilitated access to collected samples.

List of Abbreviations

AAI	average allelic imbalance
AC	adenocarcinoma
ALK	anaplastic lymphoma kinase gene
ATCC	American Type Culture Collection
AUC	area under curve
BCL11A	B-Cell CLL/Lymphoma 11A gene
BEAMing	beads, emulsion, amplification and magnetic
Bp	base pair
BRCA2	breast cancer gene 2
BRAF	proto-oncogene B-Raf, serine/threonine kinase
CAPP-Seq	cancer personalised profiling by deep sequencing
CCND1	cyclin D1 gene
CDKN2A	cyclin-dependent kinase inhibitor 2A gene
CDK4	cyclin-dependent kinase 4 gene
CDX	circulating tumour cell derived xengrafts
CfDNA	cell-free DNA
CGH	comparative genomic hybridisation
Chemo	chemotherapy
ChIP-Seq	chromatin immunoprecipitation sequencing
CI	confidence interval
CNA	copy number aberration
CNA kit	circulating nucleic acid kit
COPD	chronic obstructive pulmonary disease
COSMIC	catalogue of somatic mutations in cancer
CT	computed tomography
Ct	cycle threshold
CTC	circulating tumour cell
CtDNA	circulating tumour DNA
CV	coefficient of variance
ddH_2O	double-distilled water
ddPCR	digital droplet polymerase chain reaction
DDR2	discoidin domain receptor 2 gene
DHPLC	denaturing high performance liquid chromatography

DNA	deoxyribonucleic acid
dNTPs	deoxynucleotides
dsDNA	double stranded DNA
dUTP	2'-deoxyuridine, 5'triphosphate
EBUS-TBNA	endobronchial ultrasound transbronchial needle aspiration
ED	extensive disease
EDTA	ethylenediaminetetraacetic acid
EGFR	epidermal growth factor receptor gene
EGFR TKI	epidermal growth factor receptor tyrosine kinase inhibitor
ECOG	Eastern Cooperative Oncology Group
ELISA	enzyme-linked immunosorbent assay
FDA	Food and Drug Administration
FFPE	formalin fixed paraffin embedded
FGFR	fibroblast growth factor receptor gene
FHIT	fragile histidine triad gene
FISH	fluorescence *in situ* hybridisation
GAME_ON	Genetic Association and Mechanisms in Oncology
GAPDH	glyceraldehyde-3-phosphate dehydrogenase gene
GR	genomic representation
GWAS	genome wide association studies
HER2	human epidermal growth factor receptor 2 gene
HR	hazard ratio
HSCIC	Health and Social Care Information Centre
hTERT	human telomerase reverse transcriptase gene
ICLAC	International Cell Line Authentication Committee
iDES enhanced CAPP-Seq	integrated digital error suppression enhanced CAPP-Seq
IQR	interquartile range
ILCCO	International Lung Cancer Consortium
Indel	insertion/deletion
IRS2	insulin receptor substrate 2 gene
KDR	kinase insert domain receptor gene
KIT	*KIT proto-oncogene receptor tyrosine kinase gene*
LD	limited disease

LDCT	low dose computed tomography
LLP	Liverpool Lung Project
LOH	loss of heterozygosity
KRAS	kirsten rat sarcoma viral oncogene homolog
MAF	mutant allele fraction
MCL1	MCL1, BCL2 family apoptosis regulator gene
MDM2	MDM2 proto-oncogene
MECOM	MDS1 and EVI1 complex locus gene
MET	MET proto-oncogene, receptor tyrosine kinase
MiRNA	micro ribonucleic acid
MPCR	multiplex PCR
MmPCR	massively multiplex PCR
MRNA	messenger ribonucleic acid
MPS	massive parallel sequencing
MSI	microsatellite instability
MYC	MYC proto-oncogene, bHLH transcription factor
MYCN	MYCN proto-oncogene , bHLH transcription factor
ng/ml	nanogram per mililitre
NGS	next generation sequencing
NHS	National Health Service
NIHR	National Institute for Health Research
NKX2-1	NK2 homeobox 1 gene
NLST	National Lung Screening Trial
NOS	not otherwise specified
NPV	negative predictive value
NSCLC	non-small cell lung cancer
OR	odds ratio
P53	p53 protein
PA score	plasma aneuploidy score
Path	pathology
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PDGFRA	platlet derived growth factor receptor alpha gene
PEACE	Posthumous evaluation of advanced cancer environment

PEG	polyethyl glycol
PI3KCA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha gene
PGA score	plasma genomic abnormality score
PGA2 score	adapted plasma genomic abnormality score
PPV	positive predictive value
PS	performance status
PTEN	phosphatase and tensin homolog gene
qPCR	quantitative real time polymerase chain reaction
Ratiosn	standardised values of copy number ratios
RatiosnZscoreSq	standardised copy number ratio Z scores
RB1	retinoblastoma-1 gene
ReSoLuCENT	Resource for the study of lung cancer in North Trent
REL	REL proto-ongogene, NK-kB subunit
RNA	ribonucleic acid
RNA Seq	ribonucleic acid sequencing
ROBO1	roundabout guidance receptor 1 gene
ROC	receiver operating characteristic
RT-qPCR	real-time quantitative polymerase chain reaction
SCLC	small cell lung cancer
Scorpion ARMS	scorpion amplification refractory mutation system
SMRT	single molecule real time sequencing
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SOP	standard operational procedure
SOX2	SRY- box 2 gene
SQ	squamous cell carcinoma
SRS	stereotactic radiosurgery
ssDNA	single stranded DNA
STOMP	Small cell lung cancer trial of Olaparib as maintenance programme
STR	short tandem repeat
TAm-Seq	tagged amplicon deep sequencing
TCGA	The Cancer Genome Atlas
TKI	tyrosine kinase inhibitor

TNM	tumour node metastases
TP53	tumour protein p53 gene
TRACERx	TRAcking Cancer Evolution through therapy (Rx)
UKLS	United Kingdom Lung Cancer Screening
WGA	whole genome amplification
WGS score	whole genome summed Z score
WT	wild type

Abstract

Lung cancer causes high mortality because most people present late with advanced disease that is not amenable to curative treatment. Screening high-risk groups with low dose CT imaging of the thorax has been shown to reduce lung cancer mortality by 20%, but at the cost of a high false positive rate. Population stratification with molecular biomarkers could improve the cost-benefit of lung cancer screening programmes and reduce false positives.

Tumour cells shed DNA into the blood, enabling tumour-derived genetic alterations to be detected non-invasively by analysing circulating cell-free DNA (cfDNA). The aim of this study was to determine the screening and prognostic potential of total cfDNA levels and two genomic instability scores based on the detection of copy number aberrations in cfDNA samples of lung cancer cases and controls collected in the ReSoLuCENT study (A Resource for the Study of Lung Cancer Epidemiology in North Trent). Controls were identified as low or high risk for the development of lung cancer over five years using the Liverpool Lung Project risk model.

CfDNA was extracted from the plasma of 52 untreated lung cancer cases, 32 high risk controls and 10 low risk controls and quantified total cfDNA levels by SYBR green real-time qPCR. Low coverage whole genome sequencing with Illumina HiSeq 2500 was completed for a subset of cases (N=62) and controls (N=40). Two published genomic instability scores were adapted and tested; the plasma genomic abnormality (PGA2) and the copy number aberration (CNA) score. Screening potential was evaluated by performing Receiver Operating Characteristic (ROC) curves to assess the ability of the test to discriminate between lung cancer cases and controls by calculating area under the curve (AUC). Logistic regression was used to further assess the ability of total cfDNA levels and genomic instability scores to predict case or control status. Prognostic value was determined by Kaplan Meir and Cox regression survival analyses.

In this preliminary study, there was no difference in total cfDNA levels between early stage lung cancer cases and high risk controls. The PGA2 score was higher in high risk controls compared to lung cancer cases and was not further evaluated. In comparison, the CNA score had good discriminatory ability for high risk controls compared to all lung cancer cases (stage I-IV) with an AUC of 0.74 but poorer discriminatory ability for early stage cases (I-IIIA) with an AUC of 0.60.

Although total cfDNA levels and CNA scores above the median value were associated with poor survival, both were statistically significant in univariable but not multivariable cox survival regression analyses. Therefore, total cfDNA levels and the CNA score had limited prognostic value when other factors were taken into account. Total cfDNA levels are not recommended as a screening tool because total levels lack specificity for cancer. The screening performance of the CNA score may be improved by targeting recurrent copy number aberrations and by combining the score with alternative tumour-derived genetic alterations in cfDNA such as point mutations or methylation changes.

Publications

F Taylor, J Bradford, PJ Woll, MD Teare, A Cox. Non-invasive detection of lung cancer by identifying copy number aberrations in circulating cell-free DNA with low coverage whole genome sequencing. *Manuscript in preparation*.

J Mckay et al (**F.Taylor** author 72 of 142). Large-scale genetic analysis identifies novel loci and histological variability in susceptibility to lung cancer. Nature Genetics. 2017; 49(7):1126-1132.

F Taylor, J Bradford, PJ Woll, MD Teare, A Cox. Unbiased detection of somatic copy number aberrations in cfDNA of lung cancer cases and high risk controls with low coverage whole genome sequencing. Adv Exp Med Biol. 2016; 924: 29-32.

F Taylor, M Dawn Teare, A Cox, P Woll. Circulating cell-free DNA: a potential biomarker in lung cancer. Lung Cancer Management. 2013; 2 (5): 407-422.

Oral presentations

Non-invasive detection of lung cancer by identifying copy number aberrations in circulating cell-free DNA with next generation sequencing. CRUK Winter Lung Cancer Centre of Excellence Winter Workshop, York, Dec 2016.

Poster Presentations

F Taylor, J Bradford, MD Teare, PJ Woll, A Cox. Unbiased detection of somatic copy number aberrations in cfDNA of lung cancer cases with low coverage whole genome sequencing. . Circulating Nucleic Acids in Plasma and Serum IX, Sept 2015, Berlin.

F Taylor, J Bradford, MD Teare, PJ Woll, A Cox. Low coverage molecular profiling of circulating cell-free DNA in non-small cell and small cell lung cancer. EACR Cancer Genomics, Cambridge, June 2015.

B Ottolini, A Cox, PJ Woll, **F Taylor**, K Page, D Guttery, L Primrose, H Pringle, M Martins, C Pritchard, D Fennell, MD Teare, J Shaw. CfDNA profiling of non-small cell lung cancer using Ion Torrent NGS. National Cancer Research Institute Conference, Liverpool, November 2014.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Introduction to lung cancer

1.1.1 Background

Cancer is a major worldwide health problem. Lung cancer is the second most common cancer in the United Kingdom and causes the highest number of cancer deaths (1, 2). The five-year age-standardised survival rate for lung cancer is 9% (2). This rate is low and treatment advances have had a minimal impact in improving survival outcomes (3). Non-small cell lung cancer (NSCLC) patients diagnosed with stage I or II (early disease) can be treated surgically with curative intent and have a one year survival rate of 71% and 59% respectively (4). However, most patients with NSCLC are diagnosed with stage IV (advanced disease) and the one-year survival rate is significantly lower, 16% (4).

1.1.2 Lung cancer screening

Early detection of lung cancer has been shown to reduce lung cancer mortality. In the National Lung Screening Trial (NLST), high risk individuals aged 55 to 74 years (current smokers or ex-smokers having stopped < 15 year prior to study participation with at least a 30 pack year history of smoking) (N=53,454) were randomised to be screened by low dose CT (LDCT) imaging of the thorax or chest radiography. In this study, lung cancer mortality was reduced by 20% with LDCT imaging compared with chest radiography (5). However, this benefit was at the cost of a high false positive rate and consequently a number of participants had unnecessary invasive interventions and follow-up tests. Nearly 25% of screening LDCT scans were positive, with 96.4% of detected nodules being false positives and 320 people needing to be screened to prevent one cancer death. Furthermore, the probability that a LDCT abnormality represented an indolent tumour was 18%, leading to concerns regarding over-diagnosis (6). Three European LDCT screening randomised studies for lung cancer reported non-significant mortality reductions (DANTE, MILD, Danish Lung Cancer screening trial), but were criticised for small numbers of participants and the short length of follow up in a systematic review (7). In the NELSON study (N=15,822), of the 7155 participants randomised to LDCT, 3% were diagnosed with lung cancer and <1% had interval cancers (8). The negative predictive value (NPV) was high (99.8%) and the positive predictive value (PPV) was 40.4%. In comparison, in the NLST study the NPV was >99% and the PPV after one round

of screening was 2.4% and 5.2% after two rounds (9). However, it is difficult to compare these studies because there were different criteria for a positive test and different study designs. Mortality outcomes are still awaited for the NELSON study. Lung cancer screening would be more cost-effective if the identification of high risk individuals was improved so that both the number of patients requiring LDCT imaging and the false positive rate were reduced.

1.1.2.1 Risk prediction models to improve lung cancer screening

Various risk prediction models have been developed to enrich the screened population and identify people at highest risk of developing lung cancer, based on family history and environmental risk factors (10, 11). The application of risk prediction models in lung cancer screening programmes aims to identify a higher number of lung cancer cases for a set screened population (10). However, there is variability in the effectiveness of risk prediction models. In a recent systematic review of risk prediction models, the ability of models to distinguish between cases and controls measured by the areas under the receiver operating curve was between 0.57 and 0.88 (12).

1.1.2.1.1 The Liverpool Lung Project risk model

The Liverpool Lung Project (LLP) risk model was developed from data collected in a case-control study combined with lung cancer age-incidence data (11). The LLP model estimates the absolute risk of developing lung cancer over a 5-year period for both smokers and non-smokers (11). The model includes the most important lung cancer risk factors of age, sex and smoking (pack years), as well as other risk factors such as occupational exposure to asbestos, pneumonia, prior malignant cancer (other than lung cancer) and family history of lung cancer (11). The LLP risk model has been validated in three large Western independent populations (two large case-control studies and one prospective cohort population study) with good discriminatory ability demonstrated by areas under the receiver operating curves of 0.76, 0.67 and 0.82 respectively (13). The LLP_{v2} has been updated to include chronic obstructive airways disease and tuberculosis (14, 15). A cut off of a ≥5% 5-year risk of developing lung cancer in the prospective cohort study had a sensitivity of 57.4% and specificity of 81.1% compared to a cut off of ≥2.5% that gave a higher sensitivity of 74.3% but lower specificity of 67.4% (13).

Using the LLP risk model, the United Kingdom Lung Cancer Screening (UKLS) programme recruited participants aged 50-75 years and randomised those with a ≥5% risk over 5-years of developing lung cancer to observation or CT screening (16). Of 2028, participants that underwent CT screening, 42 (2.1%) had lung cancer. The false positive rate was 3.6% and 23% of participants required interval CT scans to follow up lung nodules (16). In comparison, the cumulative risk of a false positive biopsy result over 10 years was 7.0% for mammography screening for women aged 40 years (17).

Risk-prediction models such as the LLP risk model are very useful to identify individuals at higher risk, but the use of molecular biomarkers has the potential for greatly improved diagnostic accuracy (15). For clinical utility, such a molecular biomarker test should be based on a sample that is easy to obtain, (e.g. blood), easy to perform in an NHS lab, reproducible, cost-effective, and have high sensitivity and specificity. Currently available tests do not meet these criteria (see Section 1.5).

1.1.3 Histology and staging

There are two main types of lung cancer. NSCLC accounts for approximately 80% of cases, and small cell lung cancer (SCLC) 20% of cases (4, 18). There are several subtypes of NSCLC, and the two most common types are squamous cell carcinoma and adenocarcinoma (19). The histological subtype determines chemotherapy options and the recommended genetic tests to identify patients that may benefit from molecular targeted treatments. Lung cancer is staged according to the tumour, node, metastases (TNM) staging system, and the eighth system superseded the seventh edition in January 2017 (20).

1.1.4 Risk factors

1.1.4.1 Acquired risk factors

Smoking is a major risk factor for the development of lung cancer and causes 85% of cases (21). Other acquired risk factors include exposure to radon, asbestos, heavy metals, diesel exhaust, silica (22) and ionising radiation (23). Furthermore, there is an increased risk of lung cancer with benign pulmonary disease (24) and HIV infection (25).

1.1.4.2 Inherited risk factors

The risk of developing lung cancer is 50% greater if an individual has an affected first degree relative (parent or sibling) compared to having no affected relative (Odds ratio (OR)= 1.51, 95% confidence interval (CI): 1.39-1.63) (26). Inherited genetic syndromes can predispose patients to cancer, including lung cancer. Inherited *RB1* (27) and *TP53* mutations (28) are rare, but associated with a high risk of lung cancer at an early age (29) and an increased risk in subjects who smoke (27, 30).

1.1.4.2.1 Inherited susceptibility loci

Other than rare inherited syndromes, the majority of the inherited predisposition to lung cancer is believed to be caused by alterations in a number of low penetrance susceptibility genes (31).

A variety of genome wide association studies (GWAS) have been carried out. Studies that have analysed more than 100,000 single nucleotide polymorphisms (SNPs) are catalogued online by the National Human Genome Research Institute (32). Susceptibility loci have been identified at chromosome loci 15q25 (33), 5p15 (34) and 6p21 (35). These findings were replicated by The International Lung Cancer Consortium (ILCCO) in a large study of nearly fifty thousand subjects for 15q25 and 5p15 but not 6p21 (36). In this study, the odds ratios for one susceptibility variant were low even for the strongest associations, for example a variant at 15q25 was found to have an odds ratio of 1.26 for the development of lung cancer in white subjects (95% CI 1.21-1.32). However, the odds ratio increased to 2.64 (95% CI 1.86-3.74) when three variants were considered (two variants for 5p15 and one variant for 15q25). Interestingly, the three main loci identified (15q25, 5p15, 6p21) account for less than 10% of familial risk, and further large genome wide studies are required to identify rare low risk variants and structural variations, which are also thought to contribute to lung cancer susceptibility (37). Furthermore, genetic susceptibility loci can differ depending on subject ethnicity and histological subtype of the tumour (38, 39). In a meta-analysis of four GWAS, the presence of a rare variant of *BRCA2* approximately doubled the risk of developing squamous cell lung cancer (odds ratio 2.47), which is the strongest reported genetic association thus far in lung cancer (40).

By exploring regions identified by GWAS, genes that may be important in the aetiology of lung cancer can be identified (39). For example, 15q24-25.1 contains a candidate gene

coding for nicotinic acetylcholine receptor subunits and this region is strongly associated with lung cancer development in smokers (33, 36, 41). A very recent GWAS has identified ten additional loci and highlighted seven genes that may be important in the development of lung cancer (39). For example, three variants associated with the development of lung adenocarcinoma were located near genes related to telomere length (39). As well as aiding understanding of the development of lung cancer, the identification of loci that increase the inherited risk of lung cancer could enable individuals inheriting a combination of risk alleles to be targeted for screening or smoking cessation programmes (41).

1.1.5 Genetic changes in lung cancer

1.1.5.1 Carcinogenesis

The development of lung cancer is a stepwise process from normal bronchial epithelium into regions of metaplasia, dysplasia, carcinoma insitu and invasive carcinoma (42). Progressive histological changes are associated with increased genetic alterations. Supported by the tumour microenvironment, the increased genetic alterations lead to the acquirement of different cancer hallmarks (43). These hallmarks include, the induction of angiogenesis, sustainability of growth signals, evasion of growth suppressors, resistance against cell death, as well as developing capability to invade and metastasise and escape the immune system (43).

In squamous cell carcinoma changes in the chromosome loci 3p are among the earliest changes of carcinogenesis with loss of alleles at different sites (3p21, 3p22-24, 3p25) (42, 44). Eventually changes occur in 9p21 (*CDKN2A/p16*), 8p21-23, 17p13 (*TP53*) and 13q14 (*RB1*) (42)(Figure 1-1). 3p alterations are also early changes seen in adenocarcinoma and small cell lung cancer and therefore may be useful in the early detection of lung cancer (44).

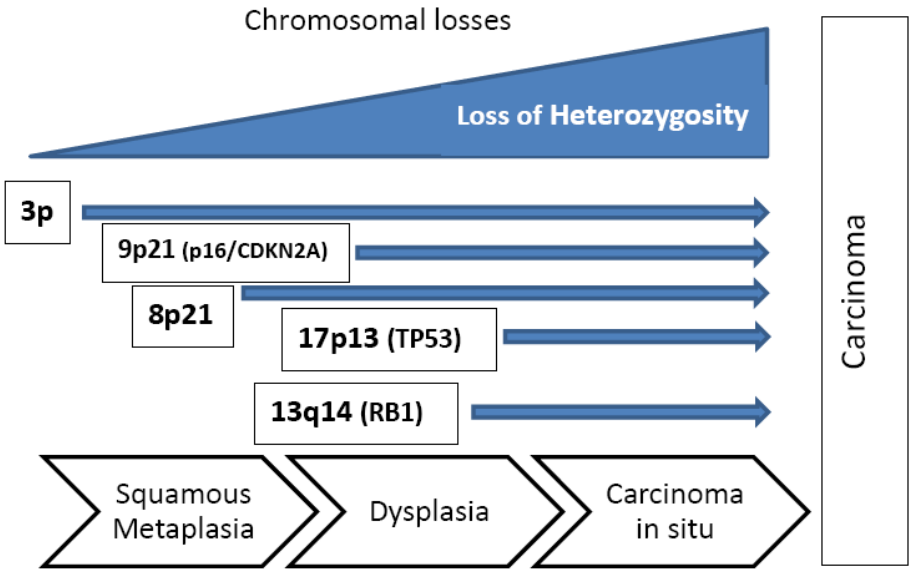


Figure 1-1 The histological development of squamous cell carcinoma and accumulative genetic changes.

Adapted from Wistuba 1999 (42) with permission.

1.1.5.2 Genomic instability

Genomic instability is an 'enabling characteristic' of cancer development that can lead to clonal cell proliferation (43). There are many forms of genetic instability, which increase the mutation rate and therefore the chance of cancer.

Somatic alterations that arise from genomic instability include:

- point mutations: substitution of a nucleotide
- insertions/deletions: addition/removal of a single or multiple nucleotides leading to a shift in the reading sequence
- allelic losses/gains (such as those mentioned above)
- structural chromosomal rearrangements caused by inversion, duplication, translocation (balanced variants), deletion or amplification (unbalanced variants that lead to a change in the number of base pairs in the genome and are also known as copy number aberrations (CNAs))
- epigenetic changes including the modification of DNA by methylation

Mutations can be silent and have no consequence or there can be loss or gain of gene function (45). Silencing of tumour suppressor genes and activation of oncogenes promotes

autonomous stimulation of growth signalling pathways and cell proliferation that can lead to the development of cancer (46-48).

In lung cancer, patterns of somatic mutation differ depending on tumour histology, ethnicity and smoking habits (49-53). Driver oncogenes, although mostly rare, have been identified in lung cancer to be important therapeutic targets, such as *EGFR, ALK, HER2* and *BRAF*, in adenocarcinoma, and *DDR2*, *FGFR* in squamous cell carcinoma (47, 54-58). More recently, 108 of 110 SCLC tumours sequenced harboured mutations in both of the tumour suppressor genes *TP53* and *RB1* (53). This high incidence indicates that mutations in these genes are essential to the pathogenesis of SCLC (51, 53). Other important genes that are deregulated in SCLC are the oncogenes *PIK3CA*, *EGFR* and *MET*, and tumour suppressor genes *CDKN2A* and *PTEN* (59, 60).

Large collaborative efforts to understand the cancer genome, epigenome, transciptome and proteome by processing tumour DNA, RNA and protein have resulted in a greater understanding of the mutational landscape of tumours including lung cancer (61, 62).

1.1.5.2.1 Copy number aberrations in lung cancer

The human genome is diploid and consists of two copies of each chromosome; a change to the total number of chromosomes is called aneuploidy (63). Copy number aberrations (CNAs) are variations in the number of copies of one or more regions of DNA (63). CNAs can vary in size from around 50 bp to several megabases (64), or be as large as a chromosomal arm or whole chromosome (65). Duplication of DNA results in copy number gain, usually termed amplification if more than one copy is gained, and deletion of DNA is termed copy number loss, or deletion if both copies are lost (63).

In lung cancer, CNAs are common, occur early in carcinogenesis (53, 66), are progressive and are present in both SCLC (53) and NSCLC (67). Mostly, focal amplifications and deletions are reported because this level of resolution enables the identification of candidate genes and potentially actionable mutations for therapeutic manipulation (68-70).

1.1.5.2.1.1 Copy number aberrations identified in non-small cell lung cancer

1.1.5.2.1.1.1 Squamous cell lung cancer

In a study of squamous cell lung cancer, there was an average of 323 segmental CNAs identified per tumour, when tumour DNA from 178 surgically resected (76% stage I-II) specimens were profiled for CNAs with Affymetrix 6.0 SNP arrays (>900,000 probes)(68). In the same study, the average number of focal aberrations per tumour was 47 and the average number of broad (defined as ≥50% of a chromosomal arm) aberrations was 23. Novel and previously identified cancer-related genes were noted to occur at peaks of significant amplification and deletion. These data were combined with SNP array data from a further 306 (N=484, 94% stage I-IIIA) tumour-normal pairs. In this combined study, 49 focal deletions and 33 focal amplifications were reported as significant aberrations based on their amplitude and frequency across samples (q value <0.05) (69). The top ten focal amplifications were 3q26.33 (*SOX2*), 8p11.23 (*FGFR1*), 11q13.3 (*CCND1*), 8q24.21 (*MYC*), 7p11.2 (*EGFR*), 4q12 (*PDGFRA*, *KIT,KDR*), 2p16.1 (*REL, BCL11A*), 9p13.3, 19q13.2, 1q21.2 (*MCL1*) and the top ten focal deletions were 9p21.3 (*CDKN2A*), 8p23.2, 2q22.1, 10q23.31 (*PTEN*), 5q11.2, 1p13.1, 3p13, 19p13.3, 3p25.3, 18q23 (69). Low coverage copy number profiling identified significant copy number gains (copy number ratio >0.25) for more than 50% of tumour samples from patients with early stage squamous cell carcinoma for chromosomal regions, 3q, 5p, 7p, 7q, 8q and 19q, and significant copy number loss (copy number ratio <0.25) for region 3p (71).

Gain or amplification of chromosome 3q is one of the most common CNAs described for squamous cell carcinoma (68). More specifically, amplification of the chromosomal region 3q22-29 has been reported to be critical for progression of metaplasia to carcinoma (72, 73). There was a higher number of CNAs identified in squamous metaplastic lesions that progressed to carcinoma (N=6) compared to those that regressed (N=23) (73). The presence of three CNAs (loss of 3p26.3-p11.1 and gain of 3q23-28, and 6p25.3-24.3) identified by array CGH predicted the development of squamous cell lung cancer from metaplastic lesions in an independent group of high risk patients with an accuracy of 92% (74). Another study of 23 cases described gains of 1q25-32, 12q23-24.3 and 17q12-22 as important for progression to carcinoma (72).

1.1.5.2.1.1.2 Adenocarcinoma

Lung adenocarcinoma and lung squamous cell carcinoma tumours have different copy number profiles (69). Affymetrix 6.0 SNP copy number profiles of 660 (89% stage I-IIIA) tumour normal pairs from patients with lung adenocarcinoma had just 25% of focal aberrations in common with profiles obtained from 643 squamous cell carcinoma tumours (69). Fifty-two significant (q value <0.05) focal deletions and 30 focal amplifications were detected. The top ten most significant regions of focal deletion were 9p21.3 (CDKN2A), 9p23, 4q35.1, 22q13.32, 1p13.1, 15q11.2, 16q23.1, 11q25, 9q21.11 and 13q12.11. The top ten most significant regions of focal amplification were 14q13.3 (*NKX2-1*), 8q24.21 (*MYC*), 5p15.33 (*TERT*), 1q21.3 (*MCL1*), 12p12.1 (*KRAS*), 12q14.1 (*CDK4*), 11q13.3 (*CCND1*), 12q15 (*MDM2*), 3q26.2 (*MECOM/TERC*) and 7p11.2 (*EGFR*) (69).

1.1.5.2.1.2 Copy number aberrations identified in small cell lung cancer

In SCLC, CNAs are generally over larger segments such as chromosomal arms, compared to CNAs detected in adenocarcinoma and squamous cell carcinoma that are more focal (66, 75). One-hundred and ten SCLC tumours were processed by Affymetrix 6.0 SNP array to identify CNAs (53). Copy number amplifications greater than 1 Mb in size were detected for chromosomes 3q (*SOX2, PIK3CA*) 5p, 8q (*MYC*) and 1p and amplifications less than 1 Mb in size for 1p (*MYCL1*), 2p (*MYCN*), 13q (*IRS2*), 4q and 8p (*FGFR1*). Copy number deletions of size greater than 1 Mb were identified for chromosomes 3p (*FHIT, ROBO1*), 3q, 13q (*RB1*), 17p (*TP53*) with deletions less than 1 Mb for chromosomes 5q, 9p (*CDKN2A*), 15q and 4q.

1.1.5.3 Identifying genetic mutations

The molecular analysis of lung tumour tissue is important in order to identify potential treatment options for patients with advanced lung cancer. The minority of patients (<20%) with lung adenocarcinoma (76) with tumour samples positive for *EGFR* mutations or *ALK* re-arrangements may respond to targeted therapies with Gefitinib (77), Erlotinib (78), Afatinib (79), Osimertinib (80) or Crizotinib (54). Further targeted therapies are in development in clinical trials, and to be eligible a patient's tumour sample may require genotyping to identify somatic alterations (76, 81). Because most patients with lung cancer present when the tumour has progressed beyond surgical resection, only a small diagnostic biopsy may be taken. This tissue sample can be inadequate for detailed molecular analysis (82) and also limits further research. Biopsies can be repeated but this is an invasive procedure with

potential risks. Alternative approaches are required and circulating cell-free DNA (cfDNA) may provide means to give indirect access to the tumour DNA.

1.2 Circulating cell-free DNA

1.2.1 Background

Extracellular nucleic acids are identified in serum, plasma and lymph as well as non-circulating fluids such as ascites, urine and saliva (83). They comprise cfDNA, microRNA (miRNA) and messenger RNA (mRNA). First described in blood by Mandel and Metais in 1948 the field has significantly developed over the last 20 years as the use of circulating nucleic acids as potential biomarkers across different diseases and in prenatal medicine has been realised (84). Circulating cfDNA describes double stranded DNA (dsDNA) fragments in the blood, either circulating freely, linked to proteins or encapsulated (85).

1.2.2 Circulating cell-free DNA in pregnancy

Cell-free fetal-derived DNA is detected in the maternal circulation from the 7th week of gestation and levels increase as pregnancy progresses (86). Low fractions of fetal-derived cfDNA are present in the maternal circulation, varying from 0.4%-11.9% in early pregnancy to 2.3-11.4% in late pregnancy (86). Fetal genetic abnormalities have been detected in maternal blood non-invasively by analysing cfDNA samples (87). Next-generation sequencing methods have increased the sensitivity for detecting fetal genetic abnormalities in maternal blood when circulating fetal DNA fractions are low (87, 88). Fetal aneuploidy was detected non-invasively by whole genome next generation sequencing of maternal cfDNA (88), even with a fetal DNA fraction ≤10% (87). This has facilitated the translation of non-invasive prenatal testing into clinical practice.

In a large prospective screening study (N=15,841), non-invasive prenatal diagnostic testing of fetal aneuploidies (trisomies 21, 13, 18) in maternal blood by highly parallel next-generation sequencing had higher diagnostic sensitivity and PPV and lower false positive rates compared to the standard screening methods of nuchal translucency ultrasound and serum biochemical analytes (89). Furthermore, the number of invasive diagnostic tests that can potentially harm the developing fetus such as amniocentesis were reduced (90). Nonetheless, testing can fail if the fraction of fetal-derived cfDNA in the blood is too low (91).

Yet, this can be overcome by repeat testing at a later gestational date when circulating fetal fractions are expected to be higher in the blood (92).

In addition to detecting fetal aneuploidies, the complete fetal genome has been sequenced (93) and sub-chromosomal abnormalities have been identified non-invasively (94, 95). Thus, expanding the potential diagnostic role of non-invasive prenatal tests to identify micro-deletions (94), copy number variants (95) and single gene disorders (96). An additional application of prenatal cfDNA testing may be to screen for cancer to aid early detection (97). In the minority of pregnant women with multiple aneuploidies detected by cfDNA whole genome copy number analyses, seven of 39 had an underlying malignancy (97). However, the challenge remains to develop a cost-effective test and the use of non-invasive prenatal testing in the NHS is currently limited (98, 99).

1.2.3 Circulating cell-free DNA in cancer

The plasma concentration of cfDNA (measured in ng/ml) is higher in patients with cancer compared to healthy controls (100, 101). However, raised cfDNA levels are not specific to cancer and have been observed in many other illnesses and after exercise (102-105). Similar genetic and epigenetic changes have been demonstrated in tumour and cfDNA for many tumour types including breast cancer, lung cancer, colorectal cancer and melanoma (84). This has generated much interest in cfDNA as a potential cancer biomarker, particularly in lung cancer where there is limited tumour tissue available.

CfDNA is readily obtained from a simple blood test and there is potential for repeated sampling at different points in the patient pathway. Moreover, cfDNA may be more representative of tumour heterogeneity than a needle biopsy, as it comprises DNA from different clonal populations of tumour cells (106). Recently, clonal and subclonal mutations were detected with variant allele frequencies varying from 0.15% to 23.3% in the plasma of cases with stage I and stage II NSCLC, demonstrating intratumour heterogeneity non-invasively (107).

1.2.3.1 Structure of circulating cell-free DNA

CfDNA can circulate in the blood stream linked to proteins, be encapsulated within exosomes, microparticles or apoptotic bodies or circulate freely as nucleosomes, virtosomes,

or DNA traps (85). CfDNA in the blood is highly fragmented in both cancer cases and controls (108). In healthy controls, cfDNA fragments are characteristically small measuring less than 200 base pairs (bp) with a peak fragment size of approximately 164 bp (109). In cancer patients, cfDNA fragments can range in size from 1kb to just 100 bp but they are typically small measuring from 160 bp to 200 bp (93, 110, 111). CfDNA fragments of 166 bp in length are commonly reported (93, 112-114). This is the length of DNA that is wrapped around a histone protein (142 bp) combined with a linker fragment (24 bp) and is known as a mono nucleosome (109, 115).

Tumour-derived DNA has been noted to be shorter or more fragmented than non-tumour derived DNA (85, 114). A greater abundance of shorter DNA fragments (<145 bp) have been observed in plasma from colorectal cancer cases compared to controls (108). More than 50% of colorectal cancer cases had fragments less than 100 bp compared to 25% of controls (108). Fragments less than 166 bp have been noted to be smaller in size by multiples of 10 bp (109, 112, 114). Ten base pairs is the equivalent length of a turn of the DNA helix around a histone protein that could resist nuclease activity and account for the observed distributions (116) (109). Understanding the structure and origin of cfDNA in cancer cases and controls is important to enhance clinical utility (85).

1.2.3.2 The origin of circulating cell-free DNA

CfDNA in healthy controls originates primarily from haemopoetic cell death (lymphocytes and myelocytes) rather than solid tissue cell death (109). In comparison, cfDNA in cancer patients originates mostly from the death of host tumour micro-environment and tumour cells rather than haemopoetic cells (85), with tumour proportions varying widely from 3% to 93% (110).

Nucleosome positioning varies between different cell types due to epigenetic differences (109). These differences were exploited to identify the origin of the primary tumour by deep sequencing cfDNA from cancer cases and correlating patterns of fragmentation and nucleosome spacing to published datasets of human cell lines and primary tumour tissues (109).

1.2.3.3 *Mechanism of release of circulating cell-free DNA*

The fragment size of cfDNA can infer its structure and mechanism of release from cells (85). Apoptosis is believed to be a major mechanism for the release of cfDNA for both cancer cases and controls yielding a mono-nucleosome size of 166 bp (114). Small fragments in multiples of 166 bp have been observed as a 'DNA ladder' pattern on gel electrophoresis of cfDNA (110). This pattern was similar to the 'DNA ladder' pattern caused by apoptotic cell death (110). Phagocytosis of necrotic cells by macrophages can also lead to the release of small DNA fragments of size 185 bp to 926 bp into the circulation (117). The presence of very small cfDNA fragments in the blood (<145 bp) may result from further cfDNA degradation following phagocytosis, or by nucleases in the blood after release of DNA from cells (108). Larger, less degraded cfDNA fragments of more than 10,000 bp are attributed to the direct release of DNA from necrotic cells (110, 118, 119). In addition, large cfDNA fragments could be actively released from tumour cells. Leukaemic cells incubated in anti-apoptotic conditions released high levels of extracellular DNA despite low caspase activity (118). Figure 1-2 summarises different release mechanisms of cfDNA into the blood.

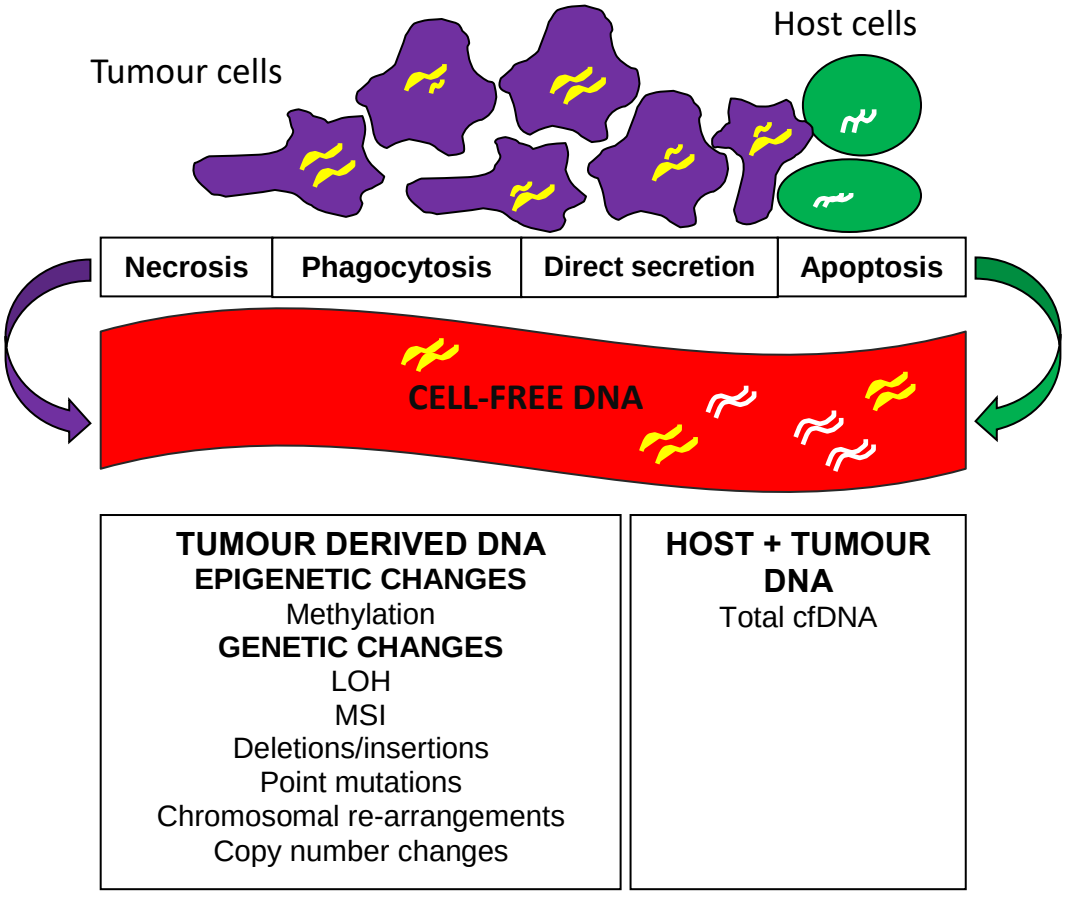


Figure 1-2: The postulated mechanisms for cfDNA release into the blood and identified genetic and epigenetic alterations.

LOH: loss of heterozygosity, MSI: microsatellite instability

Small membrane bound cell fragments called microvesicles that contain tumour DNA are secreted by tumour cells into the circulation, including exosomes, apoptotic bodies and microparticles (85). The numbers of microvesicles are increased in the blood of cancer patients (120, 121). Apoptotic bodies released from cells undergoing apoptosis are engulfed by phagocytes and release their DNA into the blood, thus contributing to levels of circulating cfDNA. Exosomes (30-100nm) and microparticles (200-1000nm) contain protein, DNA and RNA and may function as intracellular messengers (85, 122). Exosomes can be isolated in order to extract DNA and RNA for biomarker analyses (123, 124). It has been suggested that exosomes may directly release their DNA into blood and that they are a 'rich' store of cfDNA (125).

Another postulated mechanism for generation of cfDNA is circulating tumour cell (CTC) lysis. However, a single CTC contains 6 pg of DNA therefore, thousands of CTCs per ml of plasma would be required to obtain the typical cfDNA levels greater than 17 ng/ml seen in advanced cancer (126). In fact, less than ten CTCs are often present in 7.5 mls of plasma in advanced cancer cases and therefore CTC lysis is not likely to make a significant contribution to the levels of cfDNA in the blood (126, 127). Although, the number of CTCs detected in the blood of cancer cases can vary depending on the cancer subtype and stage (127).

1.2.3.4 Elimination of circulating cell-free DNA from the body

The half-life of fetal derived cfDNA in maternal plasma was found to be 16 minutes after delivery of the baby (86). In comparison, the half-life of cfDNA in a colorectal cancer patient after surgery was 114 minutes (128). Plasma nucleases degrade cfDNA but renal and hepatic clearance may also be important in the elimination of cfDNA from the circulation (86).

In cancer patients, the rapid clearance of cfDNA enables real-time monitoring of tumour dynamics in response to chemotherapy (129) and targeted treatments (130). A reduction in the levels of tumour-derived cfDNA or circulating tumour DNA (ctDNA) are early markers of tumour response in different cancer types (129, 131).

1.3 Characterisation of circulating cell-free DNA

The three steps in processing cfDNA are blood sampling and processing, DNA extraction and analysis. Many techniques are in use, some poorly validated, making comparisons between studies difficult, and limiting reproducibility of results (132). All studies require detailed standard operational procedures (SOPs). There are many potential factors that can affect the yield of cfDNA and these are shown in Figure 1-3 along with recommendations for blood collection, processing and storage.

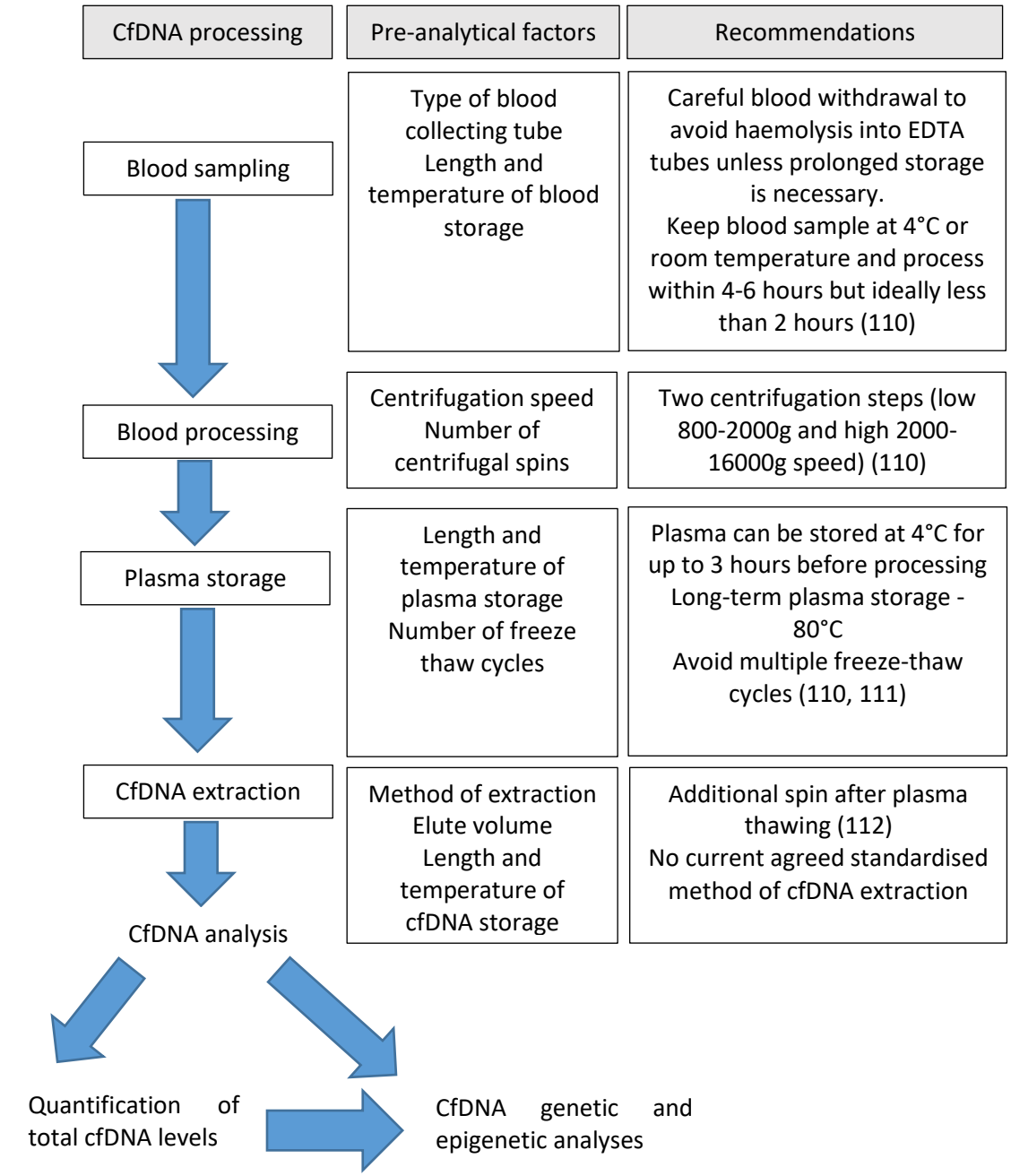


Figure 1-3: Important pre-analytical and analytical factors and recommendations for optimal cfDNA processing (133-135).

1.3.1 Pre-analytical processing

1.3.1.1 Plasma vs Serum

CtDNA is present at very low concentrations (ng/ml) so contamination with lymphocyte genomic DNA must be minimised to reduce false readings. Serum cfDNA levels are higher than those in plasma due to the lysis of white blood cells during the clotting process, which causes the release of genomic DNA into the serum (136, 137). There was a strong correlation between total cfDNA levels and white cell counts in serum but not plasma samples (138). Consequently, most cancer studies now focus on extracting cfDNA from plasma rather than serum.

1.3.1.2 Blood sampling and processing

Blood for cfDNA analysis is collected in tubes with anti-coagulants or preservatives that aim to prevent clotting and reduce cell lysis. The type of anti-coagulant used can effect cfDNA yield when blood processing is delayed (139). Blood collected in tubes containing heparin or citrate had significantly higher cfDNA levels at 24 hours compared to blood collected in tubes containing ethylenediaminetetraacetic acid (EDTA) (139). EDTA blood collecting tubes are relatively cheap, readily available and in routine use in the NHS.

The length of storage of blood prior to processing is a critical pre-analytical factor. Prolonged storage of blood collected in EDTA tubes led to progressive increases in total cfDNA levels at four, seven and 25 hours compared to one and two hours (140). Other studies have shown no difference in total cfDNA levels until after six hours of storage (135). Increased cfDNA levels after prolonged storage is due to the lysis of white blood cells, which results in increased genomic DNA (141).

It is recommended that blood collected in EDTA tubes is processed by double centrifugation (134, 135). A slow spin (800g-2000g) separates plasma from other blood components. Followed by a further faster spin (2000-16000g) to remove all cells and debris whilst avoiding cell lysis (133, 134). Plasma should be frozen and stored at -80°C and repeat freeze/thaw cycles avoided (133).

Although more expensive than EDTA tubes, alternative blood collecting tubes with different cell preservatives have been tested (141-144). CellSave (144) and Cell-Free DNA™ BCT

(Streck) tubes have enhanced the clinical utility of cfDNA analyses by enabling longer storage periods prior to blood processing without affecting cfDNA yields (145) or ctDNA profiling (141). Thus enabling a more practical and flexible approach to blood collection to facilitate greater clinical utility.

CfDNA yields, single nucleotide variant (SNV) genotyping and copy number aberration profiles were unchanged when blood collected in CellSave tubes were compared to blood collected in EDTA tubes that were processed within four hours of blood collection, but were significantly altered when EDTA tubes were processed after four days (144). There were no significant differences between CellSave and Cell-Free DNA™ BCT (Streck) tubes after 96 hours of storage for DNA quality, somatic variant detection and mutant allele frequencies (141). With CellSave tubes, whole blood can be processed four days after blood collection, there is no requirement for centrifugation and samples can be sent in the post to a central laboratory (144). Furthermore, with CellSave tubes CTCs can be analysed from the same blood sample as ctDNA advocating combinatory analyses (144).

1.3.1.3 Circulating cell-free DNA extraction from plasma

The low concentration of cfDNA in the blood necessitates efficient extraction methods. The DNA yield, its purity for PCR analysis (absence of protein, EDTA and ethanol contaminants) and the ability to detect small fragments 50-100 bp are all important factors that affect the accuracy of downstream results (137, 146, 147). CfDNA yield can vary greatly dependent on the extraction method (140, 146-148). Maximising the quantity of extracted DNA is important to increase the sensitivity of downstream mutation analysis.

For a DNA extraction method to be useful, the results must be reproducible between laboratories (132, 149). Many DNA extraction methods have been used and some have been compared (137, 140, 147, 149). Qiagen commercial kits are most frequently used for DNA extraction from plasma (135, 150). Cells are lysed to release their contents and DNA is captured on a silica-gel membrane. Thereafter, a number of washing steps are performed and finally the DNA is eluted from the silica-gel membrane. Advantages of the Qiagen blood kit include that it is simple to use, it produces results quickly, and has the potential for robotic automation (151). Disadvantages are that there is loss of small fragments <150bp, because they are filtered through the membrane instead of adsorbing to it (146). For example the QIAamp® blood kit (Qiagen) yielded significantly less cfDNA from spiked serum (50ng/ml)

compared to the triton/heat/phenolchloroform method, 18.6% (SD±4.34%) and 38.6% (SD±7.15%) respectively (140). It is difficult to compare techniques across studies due to local modifications of protocols and different methods for establishing cfDNA quantity and integrity.

It is important to standardise methods of cfDNA extraction because there are a wide number of methods in use resulting in varied cfDNA yields. As part of the European SPIDA-DNAplas collaboration, plasma was sent to fifty-six different laboratories in Europe, cfDNA was extracted using local methods and returned for quantification (150). CfDNA yield varied between 2.87 pg/µl and 224.02 pg/µl, but there was less variability when extraction methods designed specifically for cfDNA were used (150).

1.3.2 Analytical processing

1.3.2.1.1 Quantification of total cfDNA levels

A variety of methods are in use to quantify total cfDNA levels, which are summarised in Table 1-1. Amplification of a single gene such as *hTERT* (152) or *GAPDH* (140) by quantitative polymerase chain reaction (qPCR) is most commonly used. The choice of target gene may influence absolute DNA yields. Higher levels of cfDNA were reported in the same patient plasma samples with single copy *β-globulin* gene than *GAPDH*, but this may have been due to false positive amplification by the primer and poor primer design (149). CfDNA is highly fragmented and therefore higher molecular weight amplicons give lower estimation of cfDNA yield because fragments smaller than the amplicon are not quantified (133). Accurate quantification is important to ensure adequate levels of cfDNA prior to further genetic analyses. The impact of DNA fragmentation was assessed in one study, which reported that Picogreen and qPCR were less accurate in determining the concentration of smaller DNA fragments but more sensitive than (Nanodrop) spectrometry (153).

Method of total cfDNA quantification	Detection Limit	Advantages	Disadvantages
Fluorescent qPCR			
A reference gene can be amplified using specific primers from just a few copies to thousands/millions. Emitted fluorescence is measured. Examples of reference genes include *GAPDH, β-actin, hTERT*, *Alu* sequences • SYBR green: intercalating dye binds to dsDNA and fluoresces • Taqman probe: binds to predetermined sequence and fluorescent signal is emitted upon probe hydrolysis	0.01 ng/ml (SYBR green) (154)	Robust Reproducible Automated No post-PCR processing required Taqman probe: primer-dimers and nonspecific PCR products are not detected (155)	Risk of introducing contaminants Accuracy can be affected by DNA fragmentation (153) SYBR green is less specific, as also measures primer dimers and contaminants. Samples cannot be multiplexed in SYBR green assays (156) Taqman probe: selective amplification dependent on primer specificity
Picogreen			
Direct fluorescent nucleic acid dye that binds to dsDNA	0.025 ng/ml	Less expensive than qPCR Quick No PCR required	Detects all DNA fragments- less specific Accuracy can be affected by DNA fragmentation (153)
Qubit® fluorometer			
Direct fluorescent nucleic acid dye that binds to dsDNA	0.1 ng/ml	Quick, no PCR required	Low sensitivity
Nanodrop spectrophotometry			
Calculate concentration by UV light absorbance 260nm	1000 ng/ml (153)	Simple and very quick 260/280nm ratio gives measure of DNA purity	Low sensitivity Nonspecific* Affected by contaminants such as protein
Commercial DNA dipstick kits			
1 µl DNA quantified by comparison to DNA standards	100 ng/ml (157)	Simple Fast-minutes	Nonspecific*, poor reproducibility Limited range of detectable concentrations, very low sensitivity

Table 1-1: Methods for determining cfDNA yield.

*Measures single- and double-stranded DNA, RNA, oligonucleotide

1.3.2.1.2 Detection of tumour-derived alterations in circulating cell-free DNA

Tumour-derived cfDNA or ctDNA can be analysed to reveal both genetic and epigenetic changes (158). The presence or absence of genetic alterations can be investigated and ctDNA can be quantified to track tumour evolution and response to treatment in real-time with longitudinal samples (159). The ctDNA mutant allele fraction describes the fraction of mutant alleles compared to the total number of alleles (mutant plus wild-type alleles) (160). CtDNA mutant allele fractions correlate with tumour burden and ctDNA is more likely to be detected in cancer patients with advanced compared to early stage disease (161). To quantify levels of ctDNA in patients with low tumour burden or to detect small changes in allele fractions highly sensitive and specific methods are required.

A variety of methods are in use to determine cfDNA genetic alterations and examples are summarised in Table 1-2. Initial methods to detect cfDNA genetic alterations relied on PCR amplification of a predetermined sequence and real-time quantification to determine the presence of mutant alleles, with some studies validating detected genetic mutations by direct sanger sequencing (162, 163). More recently, reduced costs and technological advances have enabled the precise genetic make-up of a region to be determined by next generation highly parallel sequencing (164-166). This has also improved the sensitivity of mutation detection in cfDNA samples. In contrast to sequencing of targeted regions of known mutations, the complete plasma cfDNA genome or exome has also been intensively scanned to identify copy number changes, point mutations and re-arrangements (106, 131, 160, 167). Highly parallel sequencing, remains expensive but has the advantage of sequencing unknown DNA aberrations and is sensitive and quantitative (165, 168). Sensitivity and specificity can vary depending on a number of factors, including methods of DNA preparation, type of sequencing platform, depth of coverage and bioinformatics analyses. How many times a genomic region is read is known as the depth of coverage and is a strong determinant of analytical sensitivity (160). A higher coverage increases analytical sensitivity as there is more certainty of detecting a true mutation at low allele fraction, but increasing coverage increases cost. Reducing the proportion of the genome sequenced can reduce cost and enable very high coverage (>10,000X). Targeted panels for high coverage next generation sequencing can range from testing a few selected genes to hundreds of genes. Primers can be designed to target amplicons for enrichment (166) or prior to sequencing prepared DNA library fragments with regions of interest can be captured and

enriched, for example by hybrid capture using biotinylated DNA oligonucleotides that target regions of interest (160).

In lung cancer, ctDNA mutant allele fractions have been detected as low as 0.02% by high coverage targeted Illumina sequencing (10,000X) (CAPP-Seq) (160), 0.004% when the same method was used with additional PCR error suppression techniques (iDES-enhanced CAPP-Seq) (169), and 0.01% by ddPCR (170). By confirming the presence of a point mutation on the complementary DNA strand, mutations were detected at an even lower allele fraction of 0.00004% (171).

Method	Advantages	Disadvantages
Real-time quantitative polymerase chain reaction (RT-qPCR) (172)		
Allele specific PCR products amplified in real-time and quantified Examples of methods to enhance analytical sensitivity • Scorpion ARMs (scorpion amplification refractory mutation system) (163)-self- probing (reduces amplicons binding together) and fluorescent detection • Mutation enriched PCR (162)-2 step PCR using intermittent restrictive digestion to selectively eliminate WT genes	Distinguish between mutant and WT allele by difference in only one single nucleotide. Screen 'hot spot' region where mutation known to be so tumour analysis not needed Detection sensitivity 0.5% mutant alleles	Selective amplification dependent on primer specificity Only detects mutations that primers are designed for May miss mutations if only 'hot spot regions' are screened
Bead based digital PCR in emulsion (BEAMing) (173)		
Allele specific PCR products amplified then tag WT and mutant alleles with different fluorescent probes, which can then be counted by flow cytometry	Quantitative-can count fraction of positive alleles Can determine fraction of activating mutations that have switched to resistant type Detect rare mutant alleles Detection sensitivity 0.01%-1% mutant alleles	Personalised assay so may need tumour sample Can only assess known mutations
Denaturing high-performance liquid chromatography (DHPLC) (174)		
Allele specific PCR product is heated and fragment mobility assessed by observation of the chromatogram by UV light (2 peaks=heterozygous, 1 peak=homozygous)	Simple quick and cheap compared to sequencing analysis	Selectivity dependent on primer sensitivity Less sensitive compared to other methods Detection sensitivity 3% mutant alleles (78) Can only assess known mutations
Digital droplet PCR (ddPCR) (170)		
Allele specific Positive and negative fluorescent signals read by flow cytometer and used to calculate mutant allele concentration	Quantitative-can determine concentration of mutant alleles Detection sensitivity 0.001% mutant alleles Quick Cost effective	Selectivity dependent on primer sensitivity Can only assess known mutations Only one or a few mutations can be tested
Direct Sequencing (172, 175)		
Sanger sequencing Chain termination sequencing- Irreversible chain termination after incorporation of a fluorescently labelled nucleotide followed by fragment size separation by electrophoresis. Laser excitation causes fluorescence to be emitted and coloured chromatography peaks enable the identification of nucleotides and their ordering	A) Simultaneously detect tumour specific alterations: chromosomal rearrangements, chromosomal copy number variations, single nucleotide polymorphisms in all known cancer genes. Detects new mutations. No need for tumour sample B) High single base accuracy. Long read lengths can be obtained 700-900bp (aids sequencing of repetitive regions)	Less specific/mutation missed if high background of WT cfDNA ie. rare mutation or small amount of DNA Poor detection sensitivity 10%-30% mutant alleles Long run time for single assay Expensive

Method	Advantages	Disadvantages
Second-generation sequencing- MPS/NGS/Highly parallel (164, 166, 167)		
Whole genome/Whole exome or Targeted sequencing Direct detection of each nucleotide base incorporated into a newly synthesised DNA strand in real time. Precise method depends on the technological platform eg. Illumina, Ion Torrent Examples of methods to enhance analytical sensitivity of targeted sequencing • TAm-Seq- detection sensitivity 2% with sensitivity and specificity >97% (166) • CAPP-Seq-detection sensitivity 0.02% (detection sensitivity 0.1% with 100% sensitivity and 99% specificity (160) • M-PCR- detection sensitivity 0.1% with sensitivity and specificity >99% (159)	As above (A) Simultaneously analyse hundreds or thousands of bases (MPS) High throughput Quantitative- can determine allele fraction Increased sensitivity by targeting regions: exomes, amplicons of oncogenes or tumour suppressor genes	Variable expense depending on proportion of the genome sequenced and coverage, costs are reducing Bioinformatic infrastructure required MPS: risk of over sampling but can be limited by error rate of sequencer

Table 1-2 : Methods to detect tumour-derived cfDNA genetic alterations in lung cancer.

MPS: massively parallel sequencing. NGS: next generation sequencing. TAm-Seq: tagged amplicon deep sequencing (target and amplify long genomic regions (1000s of bases) from just one DNA fragment by PCR prior to library preparation and sequencing). CAPP-Seq: cancer personalised profiling by deep sequencing (targeted hybrid capture of prepared DNA library fragments using biotinylated DNA oligonucleotides to select regions known to be recurrently altered in NSCLC). M-PCR: multiplex PCR (personalised PCR assay based on tumour-detected mutations to target regions of library DNA fragments prior to sequencing).

1.3.2.1.2.1 Tumour-derived copy number aberrations have been identified in circulating cell-free DNA

In 2012, Leary et al. identified CNAs and re-arrangements in cfDNA of colorectal and breast cancer patients by whole genome sequencing with an average genome coverage of 9X (167). CNAs have also been identified in cfDNA of cancer patients by array CGH (113, 176, 177) and low coverage whole genome sequencing (178). CfDNA CNAs have been reported in plasma samples of prostate cancer patients (N=9) (0.1X coverage) (178), patients with hepatocellular carcinoma (N=4) (17X coverage) (168) and patients with metastatic breast cancer (N=58)(0.1X coverage) (179). Using array CGH, resistance mechanisms to androgen receptor targeted agents were explored in cfDNA samples of castrate resistant prostate cancer cases (176). However, whole genome amplification was required to obtain adequate quantities of cfDNA (2.5 µg) for array CGH analyses (176).

1.4 Circulating cell-free DNA and lung cancer

There are many potential clinical applications of cfDNA in cancer and Figure 1-4 summarises the potential benefits for patients with lung cancer (128, 180, 181). Particularly in lung cancer, the small amounts of available tumour tissue and potentially low tumour cellularity from diagnostic biopsy specimens obtained by bronchoscopy and endobronchial ultrasound transbronchial needle aspiration (EBUS-TBNA) warrants an alternative approach to the detection of tumour genetic alterations for genotyping and research (182).

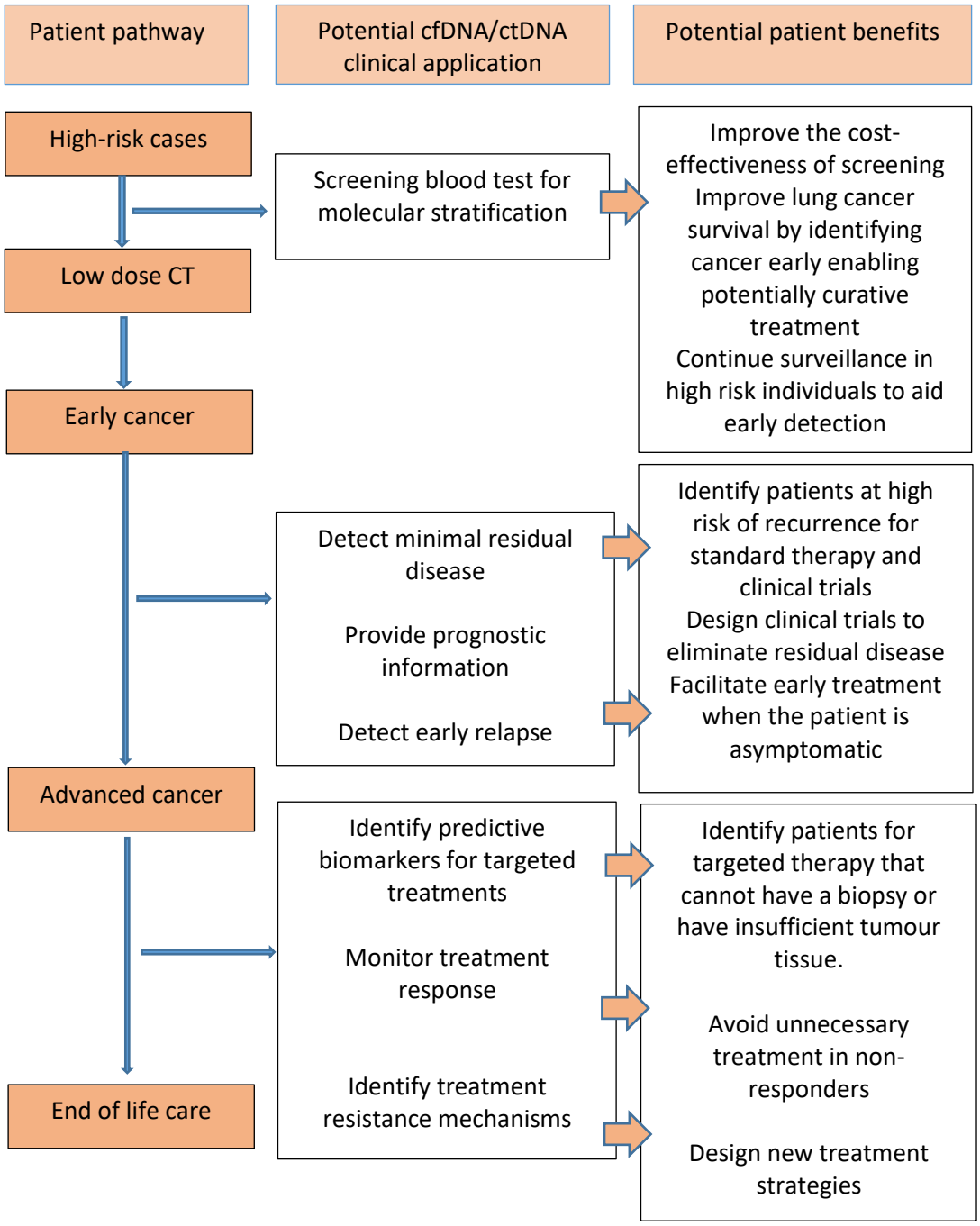


Figure 1-4: The potential applications and benefits of cfDNA technologies in lung cancer.

1.4.1 Detection of biomarkers to predict and monitor treatment response

Genotyping in lung cancer is vital to identify patients that may benefit from targeted therapies to improve disease free survival and survival outcomes (183). *EGFR* mutation status was not identified in nearly one in five lung cancer patients due to insufficient tissue, poor performance status and long turnaround time (184). Genetic testing of cfDNA by means of a simple blood test is non-invasive and facilitates genetic analyses at multiple points in the patients pathway, without the need for an invasive tissue biopsy (158).

Studies of cfDNA in lung cancer patients often focus on the ability to detect oncogenic driver mutations that are potential clinical therapeutic targets in patients with advanced disease (Appendix A). There is varied concordance between mutations identified in tumour tissue and cfDNA from 59% to 100% (185, 186). Discrepancies could occur for several reasons. Poor assay sensitivity can cause false negative results in cfDNA (174, 187, 188). In a large study of 1162 patients with locally advanced or metastatic disease and matched tumour and cfDNA samples, the false negative rate for *EGFR* testing was 10% (188). Concordance can vary between studies due to differences in patient factors as well as technical differences in laboratory methods for pre-analytical and analytical processing of cfDNA (189). In addition, if plasma cfDNA collection and the biopsy were not taken at the same time, the tumour may have evolved in the interval resulting in loss of some mutations and gain of others (190). Furthermore, cfDNA may be more representative of tumour heterogeneity because DNA in the blood has come from multiple tumour cell clones rather than a small sample of cells taken at biopsy (159, 168).

Currently, non-invasive cfDNA *EGFR* testing is recommended only when tumour tissue is not available, because a tumour could harbour a sensitising mutation not detectable in plasma (190, 191). If a cfDNA sample tests negative for a mutation then a repeat tumour biopsy must be re-considered because genotyping plasma is less sensitive then tumour tissue (189).

The FDA have approved the RT-PCR cobas® *EGFR* Mutation Test v2 to detect plasma *EGFR* sensitising mutations (exon 19 deletion and exon 21 L858R substitution mutations) in patients too unwell for biopsy (192). Gefitinib an *EGFR* tyrosine kinase inhibitor is licensed for patients with *EGFR* sensitising mutations detected in plasma when no tumour tissue is available. In our local cancer centre, *T790M* cfDNA testing is soon to be available for patients unable to have a biopsy or with insufficient tumour tissue for processing, to establish

whether treatment with the third generation *EGFR* TKI Osimertinib may be of benefit. However, a tumour biopsy is still recommened in patients with negative *T790M* plasma results due to a 30% false negative rate compared to tumour genotyping (193).

Serial measurements of mutations may be an alternative method to imaging to assess tumour response to treatment (129, 194). The dynamic changes of ctDNA levels are demonstrated by targeted approaches and highly sensitive techniques. Due to the non-specific nature of measuring total cfDNA levels, small changes in ctDNA levels may not be detected as a change to total cfDNA levels and therefore would not be reliable for measuring tumour response to treatment.

1.4.2 Detection of resistance mechanisms

CfDNA biomarkers have been incorporated into early phase clinical trials for predictive and prognostic assessment as well as to identify resistance mechanisms. Resistance mechanisms to first generation (e.g Gefitinib, Erlotinib) and third generation (eg. Osimertinib) *EGFR* TKIs have been studied in cfDNA samples (Apendix Table A). These studies have revealed tumour heterogeneity and a multitude of intra and inter-patient resistance mechanisms to include gene amplifications and point mutations, which may have been missed by tumour biopsy due to sampling bias (195). Understanding response and tumour resistance mechanisms is vital to designing new drugs and treatment strategies to maximise clinical impact (195, 196).

1.4.3 Detection of disease relapse and minimally invasive disease

Following potentially curative treatment, the presence of ctDNA may indicate minimally residual disease and may therefore represent a higher risk of disease relapse (197). In the NSCLC TRACERx (TRAcking Cancer Evolution through therapy) study, cfDNA extracted from pre and post-surgical plasma samples of early lung cancer cases were profiled using personalised multiplex-PCR (mPCR) next generation sequencing assays based on individual tumour mutation profiles (159). The presence of ctDNA was confirmed if at least two SNVs (clonal or subclonal) were detected in plasma. Study participants had clinical assessments, blood profiling and CXRs performed every three to six months. The detection of ctDNA was associated with disease relapse in 13 of 14 cases following surgery for early stage lung cancer. The median time between the detection of ctDNA and CT imaging confirmation of disease relapse was 70 days (range 10-346 days). Furthermore, in three cases the proportion of

ctDNA increased during adjuvant chemotherapy suggesting resistance, and all three cases relapsed within one year of surgery.

In addition to lung cancer, the presence of ctDNA has been associated with disease relapse in early stage breast (198) and colon cancer (197). In breast cancer, serial blood samples were collected following neo-adjuvant chemotherapy and surgery for 55 cases (198). CtDNA was detected by a personalised ddPCR assay (based on clonal somatic mutations identified in matched tumour samples) and the presence of ctDNA was predictive of disease relapse with a hazard ratio of 25.1 (95% CI 4.08-130.5, $p<0.0001$). The median time that ctDNA was detected prior to clinical relapse was eight months. Although case numbers were few, further genetic analysis of ctDNA with highly parallel sequencing identified somatic alterations more in keeping with alterations identified in metastatic deposits rather than the primary tumour.

For 230 stage II colon cancer cases, targeted sequencing of tumour DNA was carried out for 15 genomic regions known to be recurrently mutated (197). Then, the somatic mutation with the highest allele fraction (compared to the mean allele fraction of a group of healthy controls) was used to create a personalised assay for ctDNA detection with high coverage targeted sequencing. The detection of ctDNA after the completion of adjuvant chemotherapy was predictive of disease relapse with a hazard ratio of 11 (95% CI 1.8-68, $p=0.001$). These findings could lead to the development of personalised therapeutic strategies to target and eradicate micro-metastases (159, 198).

1.5 Non-invasive biomarkers to potentially aid early lung cancer detection

1.5.1 Circulating cell-free DNA

The identification of ctDNA in the blood may aid early lung cancer detection (160). Although, it is more difficult to detect cfDNA tumour-derived genetic alterations in early stage compared to advanced stage cancer due to lower tumour burden (161). Despite a high analytical sensitivity of 0.02% for detecting 139 recurrently mutated genes, the sensitivity for detecting ctDNA in stage I lung cancer cases was 50% compared to 100% for stage II-IV, specificity for both sup-groups was 96% (160).

1.5.1.1 Total circulating cell-free DNA levels

Total cfDNA levels have been reported to distinguish early lung cancer cases (I-IIIA) from healthy controls, cases with benign nodules and cases with chronic inflammatory lung disease with a Receiver Operating Characteristic (ROC) area under the curve (AUC) of 0.80 (199), AUC 0.90 and AUC 0.78 (101) respectively. Further studies to demonstrate the potential screening or diagnostic tool in lung cancer are shown in Table 1-3**.** The role of cfDNA levels in screening for lung cancer remains undefined, and poor sample handling as well as methodological differences could contribute to differing results between studies.

Prospective studies are more informative than case-control studies for biomarker assessment because bias and reverse causation are reduced. In a prospective study of 1035 former or current heavy smokers attending for CT screening, there was no significant difference in median total cfDNA levels in participants found to have lung cancer compared to those who did not (4.8 ng/ml (N= 38, IQR 3.4-8.0 ng/ml) and 3.9 ng/ml (N= 947, IQR 2.1-6.1 ng/ml respectively) (200). In this study the median value for lung cancer patients was relatively low in comparison to a case-control study by the same group utilising the same methods (152). It was then noted that over 40% of samples were analysed three years after being frozen and that nearly a third of DNA was lost annually due to degradation in storage (201).

Study	Method of DNA		Subjects (NSCLC unless stated)	Total cfDNA level ng/ml (Range)(± SD)	Cut-off ng/ml	Sensitivity Specificity	AUC-ROC[a] (95% CI)
	Extraction	Quantification					
Szpechcinski 2016(199)	QIAamp DNA blood midi kit	RT-qPCR β actin	Stage I-IIIA N=65 Benign lung noduels N=28 (Healthy controls N=16)	Mean 4.00 ± 1.60 Mean 3.06 ± 1.37 p=0.0009 (Mean 1.01 ± 0.90 p<0.0001)	2.8	86.4% 61.4%	0.80 (0.70-0.84)
Szpechcinski 2015(101)	QIAamp DNA blood midi kit	RT-qPCR β actin	Stage I-IIIA N=50 Healthy controls N=40	Mean 8.0 ± 7.8 (Median 5.9, range 1.12-41.0) Mean 2.3 ± 1.5 (Median 1.87, range 0.72-6.49) p<0.0001	2.8	90% 80.5%	0.90 (0.81-0.95)
			Chronic inflammatory lung disease N=101	Mean 3.36 ± 1.8 p<0.0001	5.25	56% 91%	0.76 (0.68-0.83)
Ulivi 2013(202)	QIAamp DNA blood mini Kit	RT-qPCR GAPDH	Stage I-IV N=100 Healthy controls N=100	Median 47.2 (0.7–251) Median 9.2 (2.2–184) p<0.0001	25	80% 91%	0.90 (0.86-0.93)
Catarino 2012(203)	QIAamp DNA blood mini Kit	RT-qPCR hTERT	Stage I-IV N=104 Healthy controls N=205	Mean 270 (-) Mean 122 (-) p<0.0001	20	79% 83%	0.88 (0.84-0.92)
Van der Drift 2010(204)	MagNA Pure LC Total Nucleic Acid Isolation Kit	RT-qPCR β globin	Stage I-IV N=46 Respiratory clinic attendees with lung cancer excluded (52% COPD) N=20	Median 52 (5-3597) Median 29 (0-175) p=0.03	32	67% 52%	0.66 (0.53-0.80)
Kumar 2010(205)	QIAamp DNA blood mini kit	Picogreen dsDNA kit	Stage III, IV N=100 Benign lung disease N=100	Mean 122.7 ± 47.4 Mean 74 ± 19.8 p<0.001	104.5	52% 95%	0.83 (0.77–0.89)
Szpechcinski 2009(206)	QIAamp DNA blood midi kit	RT-qPCR β actin	Stage I-IIIA N=30 Healthy controls N=16	Mean 12.0 (1.5-64.4) Mean 2.7 (0.9-7.0) p<0.001	7	50% 100%	0.87 (0.74-0.95)

Yoon*(207)	QIAamp DNA blood mini kit	qRT PCR β actin	NSCLC+SCLC (8.8%) Stage I-IV and ED/LD N=102 Controls attending for lung cancer screening mostly smokers N=105	Median 22.6 (3.1-730.5) Median 10.4 (1.6-89.9) p<0.0001	-	-	0.86 (0.81-0.91)
Paci 2009(132)	QIAmp DNA blood mini Kit	qRT PCR hTERT	Stage I-IV (most early) N=151 Healthy controls N=79	Mean 12.8 (-) Mean 2.9 (-) p<0.001	2	85% 47%	0.79 (0.71-0.83)
Ludovini 2008*(208)	Qiamp blood kit**	RT-qPCR hTERT	Stage I-III N=76 Healthy smokers N=66	Mean 60 ± 99.8 Mean 5 ± 8.8 p<0.0001	3.25	80% 60%	0.82 (0.75-0.88)
Herrera 2005 (209)	QIAamp DNA blood mini kit	RT-PCR β actin	Surgical candidates N=25 Healthy controls N=11	Mean 14.6µg/l (3-30µg/l) Mean 10.6µg/l (7-14µg/l) p=0.18	-	-	0.63 (0.44-0.82)
Guan-Shun 2004(210)	QIAamp DNA blood midi kit	PICOgreen dsDNA kit	NSCLC+SCLC (24%) stage I-IV (mostly advanced) N=67 Benign lung disease N=36 Healthy controls N=44	Median 110.7 (10^{th} -90^{th} percentile, 22.9-383.5) Median 45.5 (7.5-121.0) Median 11.6 (2.5-31.8) p<0.001	53.8	70% 80%	cancer vs all controls 0.86 (0.80-0.91)
Sozzi* 2003(152)	QIAamp DNA blood mini kit	RT-qPCR hTERT	NSCLC I-IV (most early stage) N=100 High risk attended for screening N=100	Median 24.3 (-) Median 3.1 (-) No p value	25	46% 99%	0.94 (0.91-0.97)
Sozzi 2001(157)	QIAamp DNA blood mini kit	DNA DipStick TM kit	NSCLC stage I-III N=81 Healthy controls N=32	Mean 318 Mean 18 No p value	26-125	86% 100%	0.84 (0.77-0.90)

Table 1-3: The utilisation of cfDNA levels as a potential screening or diagnostic tool in lung cancer.

[a]AUC-ROC: area under the curve – receiver operator characteristics, a measure of discriminatory power, a value of 1 has excellent discriminatory power where as a value of 0.5 has no discriminatory power. All studies in this table were conducted with plasma samples. *matched cases to controls.**specific type of kit not stated.

Most recently, six genes from The Cancer Genome Atlas known to be commonly methylated in squamous cell and adenocarcinoma lung cancer were investigated in matched sputum and cfDNA samples in a case-control study of early stage lung cancer (stage I and IIA) and controls with histologically confirmed benign lung nodules. With an optimised method to minimise DNA loss, the best combination of three genes had a sensitivity and specificity of 98% and 71% for sputum and 93% and 62% for plasma with AUC 0.89 and AUC 0.77 respectively (211). Prior to this study, the sensitivity and/or specificity of cfDNA methylation analyses were too poor to be useful for screening, or studies focused on cases with advanced lung cancer (212-214), or compared cases to healthy rather than high risk controls (215).

Alternative blood biomarkers have been investigated in lung cancer to aid early detection to include circulating tumour cells (CTCs) (216, 217), microRNAs (218) and proteins (219, 220).

1.5.2 Circulating tumour cells

An advantage of detecting somatic alterations in CTCs are that they are specific to cancer, and are not identified in genomic DNA or caused by other disease processes (221). Similar to cfDNA, CTCs are present in minute quantities in the circulation and have to be isolated from blood constituents (222). CTCs can be enriched by identifying their epithelial cell markers or physical characteristics (size or deformability) or all blood cells can be analysed (223). Once isolated (223), CTCs can be quantified (by the number of cells per volume of blood), single cell components (DNA, RNA, protein) can be extracted and analysed (224) and the morphology of CTCs can be studied (225). Furthermore, CTCs can be grown in vitro or injected subcutaneously to create CTC derived xenograft (CDX) models. The CDX models carry the same genetic signature as the primary tumour, model response to therapeutics and enable the identification of tumour resistance mechanisms (226).

The FDA have approved a cell search system that selects and counts CTCs by an epithelial cell marker for prognostic monitoring, in certain types of cancer but not lung cancer (227). This method in early stage lung cancer patients only identified one or more CTCs in approximately one third of patients studied (N=125) and had poor diagnostic ability (217). With epithelial-mesenchymal transition, CTCs may lose their epithelial cell markers and escape capture (228). In a study of patients with COPD (N=168), CTCs were isolated by size and identified by defined cytopathological features (229). In this study, 3% of COPD patients had between 19

to 67 CTCs detected between one to four years prior to the identification of a lung lesion on annual CT screening and subsequent diagnosis of stage IA NSCLC. No COPD patient without CTCs detected subsequently developed cancer (median follow up 60 months) and there were no CTCs detected or cancer diagnoses for smoking (N=42) or non-smoking healthy controls (N=35) with a mean follow up of five years. However, due to limited data CTCs cannot be recommended to aid early lung cancer detection. Furthermore, the technology is relatively expensive and if robust standardised methods for cfDNA were validated, this would be a simpler, more widely applicable and cheaper methodology. In addition, for CTC genome analyses whole genome amplification is required to obtain adequate DNA quantities for sequencing that can lead to the introduction of artefacts (230)

1.5.3 Circulating RNA

Cell-free miRNAs circulate in the blood and evade degradation by RNAses by attachment to protein complexes, or containment within exosomes/microvesicles, apoptotic bodies or HDL structures (231, 232). As regulators of gene expression, levels of miRNA are deregulated in cancer (231) and certain miRNAs are associated with lung cancer development and aggressiveness (218, 233). Tumour-derived RNA includes miRNA (non-coding small RNAs) and long non-coding RNA. Due to their small size, miRNAs are more stable than long non-coding RNA and they can be detected by relatively simple assays (233). MiRNAs can be detected by RT-qPCR, microarray hybridisation (233) or more recently sequencing (124).

In lung cancer, miRNA have most frequently been studied as diagnostic and prognostic biomarkers (218, 234). A 13 miRNA panel was validated in a large LDCT screening study of high risk individuals (heavy smokers aged ≥50 years, N=1,115), the diagnostic sensitivity and specificity for lung cancer were 78% and 75% with an AUC of 0.85 (234). A 24 miRNA panel was validated prospectively in a large screening study (current or ex heavy smokers aged ≥50 years, N=939) and diagnostic sensitivity and specificity were 87% and 81% (218). In this study, the combination of detecting a lesion by LDCT and a positive miRNA test reduced the false positive rate from 19.4% for LDCT alone to 3.7%. In a different study, miRNA was isolated from tumour-derived exosomes of early stage NSCLC cases and sequenced, the presence of a panel of miRNA had an AUC of 0.94, indictaing excellent discriminative ability (124). However, in lung cancer studies there is great variability between published miRNA panels with different numbers and types of miRNAs reported. These differences are most

likely due to patient factors, a lack of standardised methods for collecting, processing and quantifying/normalising miRNA expression, as well as variability in the material studied eg. whole blood, serum, plasma, and exosomes.

1.5.4 Proteins

In response to tumour-antigens, the immune system produces autoantibodies that circulate in the blood and can be identified in serum by enzyme-linked immunosorbent assays (ELISA), a relatively straightforward assay that can be performed in most clinical laboratories (235). A six-panel autoantibody test was technically (236) and clinically validated in three matched (for age, sex and smoking history) early lung cancer case-control groups with reported AUC of 0.63-0.71 and a diagnostic sensitivity of approximately 40% and specificity of 90% (220). Diagnostic specificity was increased to 93% by the addition of a seventh autoantibody to the panel (p53, NY-ESO-1, CAGE, GBU4-5, HuD, MAGE, SOX2) to create the Early CDT®-Lung test (Oncimmune Ltd., Nottingham, United Kingdom)(237). This test has been audited in clinical practice (N=1,600) (238) and discriminated between people with benign and malignant detected lung nodules (N=296) with similar values of sensitivity and specificity (239). A randomised prospective clinical trial to evaluate the ability of the Early CDT®-Lung test to identify high risk individuals is ongoing with the primary aim to reduce the number of stage III and IV lung cancer cases diagnosed (240). So far, 10,000 ex-smokers or smokers aged 50-75 years have been recruited in Scotland out of a planned 12,000 participants (241).

Blood Levels of pro-surfactant B are increased in NSCLC compared to high risk controls and can be identified by ELISA (219, 242). The ability of Pro-surfactant B levels to distinguish between high risk controls and lung cancer cases was tested in a large prospective screening study (N=2,485) that recruited individuals with a 2% risk of developing lung cancer in a 3-year period by using risk prediction models (219). In this discovery set, there was good discriminative ability with an AUC of 0.69 and 0.74 when adjusted for lung cancer risk factors. In a validation set, with samples from a different study the AUC was 0.68. Levels of pro-surfactant B were found to be higher than matched controls for cases with adenocarcinoma but not squamous cell carcinoma. Reliance on single biomarkers that vary by the histological subtype of lung cancer is not recomended. A screening test needs to differentiate between all histological types of lung cancer and high risk controls to reduce the false negative rate.

The combination of pro-surfactant B levels and N^1,N^{12}-diacetylspermine (a serum metabolite) differentiated between cases that developed lung cancer (N=108) and matched healthy controls (N=216) in a validation set with an AUC of 0.81 for cases with serum samples collected 0-6 months prior to diagnosis and an AUC of 0.73 overall (243). A combinatory biomarker approach is therefore advocated.

1.5.5 Other non-invasive biomarkers in lung cancer

Urine can be collected non-invasively and is an abundant source of biomarkers. Urine metabolites were evaluated by liquid chromatography-mass spectrometry and four metabolites differentiated between lung cancer cases stage I-II (N=213) and matched population controls (N=536) (age, sex, race) with AUC 0.71 (244). For 178 cases and 351 controls (matched on age, sex, race, date of sample collection), the AUC was improved when two of the four urinary metabolites were combined with lung cancer risk factors (from 0.78 to 0.80) (245). However, the occurrence of these metabolites in other cancer subtypes is unknown; metabolism can vary with dietary and drug intake and levels of metabolites can be affected by renal function (244). Alternatively, the detection of volatile organic compounds in exhaled breath show promise for early lung cancer detection but there is a lack of standardised validated methods (246).

1.6 Aims and objectives of the project

Lung cancer is a genetic disease caused by inherited and acquired genetic changes. The hypothesis of my PhD is that the identification of acquired genetic changes of lung cancer in the blood will enable the development of clinically useful biomarkers. There is a significant need to improve the overall survival of people with lung cancer by detecting asymptomatic individuals with early stage disease. Molecular stratification of people at high risk of lung cancer may reduce the number of people needing CT imaging and minimise the false positive rate, thereby improving the cost-effectiveness of a lung cancer-screening programme.

To improve diagnostic accuracy it is important that a test has high diagnostic sensitivity. Current proposed non-invasive biomarker tests report sensitivity of 40%-87% (218, 220). High diagnostic sensitivity was reported for specific miRNA panels but there was poor consensus between different studies because the numbers and types of miRNAs tested differed (218, 220, 234). It was proposed to further investigate the use of cfDNA as a non-invasive blood biomarker to aid early lung cancer detection by using blood samples collected in the ReSoLuCENT study (A Resource for the Study of Lung Cancer Epidemiology in North Trent) study (247) (see Section 2.2.2). Previously, sequencing methods have focused on a single gene or a panel of genes, to reduce cost and increase test sensitivity. Low coverage molecular profiling of cfDNA represents a cost effective, unbiased approach to identify somatic copy number alterations across the whole genome.

1.6.1 Optimising plasma DNA extraction and evaluating total circulating-cell free DNA levels as a potential screening tool (Chapter 3)

The initial aim of my PhD project was to optimise the quantity of extracted plasma cfDNA to enable sensitive downstream genetic analysis of samples collected in ReSoLuCENT. Method standardisation is an important step towards national standardisation and future clinical implementation (84). This facilitates appropriate blood handling and processing, plasma cfDNA extraction and analysis, in order to further biomarker development in lung cancer.

The first aim and objective was:

- To identify the most efficient cfDNA extraction method by comparing the percentage recovery of tumour formalin fixed paraffin embedded (FFPE) DNA from healthy volunteer plasma.

Many studies have demonstrated that total cfDNA levels can discriminate between lung cancer cases and controls (see Section 1.3.2.1.11.5.1.1). Most recently, total cfDNA levels had excellent discriminative ability to distinguish between early lung cancer cases (I-IIIA) and healthy controls with AUC 0.90 (101). However, no study has assessed the discriminatory ability of total cfDNA levels to distinguish early lung cancer cases and high risk controls identified by a risk prediction model. It was hypothesised that total cfDNA levels would be higher in lung cancer cases compared to high risk controls. Quantification of total cfDNA levels is a simple and cheap test that could be carried out in any laboratory able to perform RT-qPCR assays and is therefore a cost-effective approach warranting further evaluation.

The second aim and objective was:

- To evaluate total cfDNA levels quantified by SYBR green RT-qPCR as a potential screening tool for lung cancer by comparing total cfDNA yield for lung cancer cases and high risk controls.

1.6.2 Low coverage sequencing to identify copy number aberrations in circulating cell-free DNA (Chapter 4)

Copy number aberrations (CNAs) occur early in lung carcinogenesis, are progressive and occur in both NSCLC and SCLC. The hypothesis was that due to the release of tumour DNA from cancer cells, lung cancer cases would have more cfDNA CNAs compared to high risk controls and therefore the detection of CNAs may serve as a screening and prognostic tool for lung cancer. A genomic instability score can quantify the magnitude and number of CNAs and it would be expected that lung cancer cases would have higher genomic instability scores compared to high risk controls.

The aims and objectives were

- To evaluate analytical performance and to validate the detection of tumour-derived CNAs by low coverage whole genome sequencing of cfDNA samples
    a. To optimise DNA library preparation for cfDNA samples by comparing library quantities and detection of copy number ratios with different input amounts of cfDNA ng/ml and PCR cycles as well as two different PCR mastermixes.
    b. To determine the lower limit of detection for identifying CNAs by low coverage whole genome sequencing of cfDNA samples by adding tumour FFPE DNA in known quantities to extracted cfDNA from the pooled plasma of healthy volunteers.
    c. To determine test reproducibility across sequencing runs by comparing the detection of copy number ratios in cell line DNA
    d. To demonstrate the identification of tumour-derived CNAs in cfDNA samples of lung cancer cases collected in the ReSoLuCENT study by low coverage whole genome sequencing. The objectives were to compare cfDNA CNAs to those detected in matched tumour FFPE DNA. In addition, to compare cfDNA CNAs to CNAs known to be common to the three main subtypes of lung cancer.
- To evaluate the clinical validity of low coverage whole genome sequencing to identify CNAs in selected lung cancer cases and controls
    a. To explore the sreeening and prognostic value of two genomic instability scores based on the number and magnitude of CNAs identified in cfDNA samples. The objectives were to compare scores between lung cancer cases and high risk controls to assess screening value and to aseess the relationship between score and overall survival to assess potential as a prognostic tool.

2 Materials and Methods

2.1 Materials

2.1.1 General laboratory equipment and consumables

Laboratory equipment	Supplier
AB104-S Balance	Mettler, Toledo
ABI 7900 Genotyping Platform	Applied biosystems
Benchtop Micro Centrifuge Heraeus Pico 17	Thermo Fisher Scientific
Benchtop Rotamixer	HATI
Benchtop Temperature Controlled Centrifuge	MSE Sanyo
Class II Microbiological Safety Cabinet	Envair
CO_2 Incubator MCO175	Sanyo
Covaris® S220 Focused-ultrasonicator	Covaris
Heating Block	Grant Boekel BBA
Ice machine	Scotsman Ice Machine
Magnet stand-96	Thermo Fisher Scientific
Nanodrop Spectrophotometer ND-1000	Labtech International
P2, P10, P20, P100, P200, P1000 Gilson Pipettes	Fisher Scientific
Power pack	Bio-Rad
QBD4 Incubator for Eppendorfs	Grant Boekel BBA
Thermal cycler: GeneAmp PCR system 96 well	Applied Biosystems
Thermal cycler: Light Cycler 480 96 well	Roche
Thermal cycler: PTC-200 96 well	MJ Research
Titramax 1000 Incubator and Shaker	Heidolph
UV Sterilisation Cabinet	Bignet
Vortex Genie 2	Scientific Industries
Water Purification Unit	Lab Technologies
Water bath	Grant Instruments
Western Gel Mini Protean II Cell	Bio-Rad

Laboratory consumables	Supplier
0.2ml Microcentrifuge Tubes	Starlab
0.5ml, 1.5ml, 2ml Microcentrifuge Tubes	Fisher Scientific
1.5 ml DNA Lo-bind Microcentrifuge Tubes	Eppendorf
1.5 ml Cryovials	Scientific laboratory suppliers (SLS)
15 ml Sterile Conicol Tubes	BD Falcon
50 ml Sterile Conicol Tubes	BD Falcon
96 Well PCR Plates	Applied Biosystems
384 Well PCR Plates	Starlab
6 ml EDTA Blood Phlebotomy Tubes	BD
GIBCO Distilled DNase/Rnase Free Water	Life Technologies
Graduated 10µl Microfilter Tips	Starlab
Nitrile Powder Free Gloves	Fisher Scientific

Pipette Tips	Starlab
Plate Seals	Biorad

2.1.2 Laboratory solutions

All laboratory solutions were made up with ddH_2O, were purchased from Sigma-Aldrich Co. unless stated otherwise and were of molecular biology grade.

TAE buffer (10x, pH8.0): 0.4 M Tris-base, 200mM glacial acetic acid, 10 mM EDTA (pH adjusted to 8.0).

Phosphate Buffered Saline (PBS)(Dulbeccos A, 1x): Sodium chloride 0.137M, Potassium Chloride 0.003M, Disodium Hydrogen Phosphate 0.008M, Potassium Dihydrogen Phosphate 0.0015M.

6X Sample loading buffer: Glycerol 60%, Tris-HCL pH 7.6 10mM, EDTA 60mM, Bromophenol blue 0.03%, Xylene Cyanol FF 0.03%.

2.1.3 Buffers and reagents for molecular biology techniques

2.1.3.1 DNA processing

Xylene (Fisher Scientific)

Absolute Ethanol (Fisher Chemical)

Isopropanol 99.5% extra pure (ACROS organics)

Nuclease free distilled water (Gibco by Life Technologies)

Ready to use TE (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA) X1 buffer (Thermo Fisher Scientific)

NaOH 2M (Illumina)

2.1.3.1.1 Phenol choloform method:

Triton X-100 (BDH Prolab)

Phenol chloroform isoamyl alcohol mixture (Sigma Life Science)

Sodium acetate trihydrate (Fisher Scientific)

2.1.3.1.2 Commercial kits purchased for DNA extraction:

QIAamp® DNA FFPE Kit (Qiagen)

Chemagic DNA Buffy Coat Kit (Perkin Elmer)

QIAamp® Blood Mini Kit (Qiagen)

QIAamp® Circulating Nucleic Acid Kit (Qiagen)

FlexiGene DNA kit (Qiagen)

2.1.3.1.3 Commercial kits purchased for Illumina whole genome sequencing:

NEBNext® Ultra DNA library Prep Kit (NEW ENGLAND BioLabs®Inc.)

NEBnext® multiplex oligonucleotides (NEW ENGLAND BioLabs®Inc.)

Dual lane HiSeq Rapid PE Flow Cell (Illumina)

TruSeq Rapid PE Cluster kit (Illumina)

TruSeq Rapid Duo cBot Sample Loading Kit (Illumina)

Truseq rapid SBS 200 cycle kit (Illumina)

2.1.3.2 PCR

PCR water

Nuclease free distilled water (Gibco by Life Technologies)

2.2 Plasma samples

2.2.1 Healthy volunteer study

Healthy volunteers from the University of Sheffield were recruited to the study 'Optimisation of plasma nucleic acids'. Up to 50 mls of blood were withdrawn from each participant, and if required further samples were taken after a minimum four-week interval as per protocol. Peripheral whole blood samples were processed as described in Section 2.3 then plasma and lymphocyte buffy coats were stored in 1.5 ml cryovials. This study had ethical approval from South Yorkshire Research Ethics Committee (08/H13010/40) and was authorised by Sheffield Teaching Hospitals NHS Foundation Trust (STH14669) (Appendix B). Written informed consent was obtained from all participants.

2.2.2 The ReSoLuCENT study

The aim of the ReSoLuCENT study (A Resource for the Study of Lung Cancer Epidemiology in North Trent) (247) was to collect high quality detailed epidemiological and biological samples and data from lung cancer patients with a strong family history or early onset disease, and family based controls. This multi-centre national institute for health research (NIHR) portfolio study was opened in 2006 after ethical approval from West Midlands Research Ethics Committee (05/MRE07/72) and was authorised by Sheffield Teaching Hospitals NHS Foundation Trust (STH13872) (Appendix C). Recruitment to the study completed in August 2016. The Chief Investigator of the study is Professor Penella Woll.

Participants with lung cancer were eligible for the study if they had active disease and were either aged 60 or less at diagnosis or had a strong family history of lung cancer (1st degree relative with lung cancer aged ≤ 60 years or ≥ 2 1st or 2nd degree relatives with lung cancer at any age). Recruited controls were either co-habiting partners or 1st degree relatives of the case. All participants had a blood test and samples were processed using optimised SOPs (as described in Section 2.3) to obtain plasma and lymphocyte buffy coat layer so that contamination and degradation of cfDNA was minimised. In addition, permission was sought from cases to obtain surplus tumour FFPE tissue sections. All participants completed a highly detailed questionnaire to include smoking, occupational, and medical histories.

Each participant in the study was allocated a unique number that identified the recruiting centre, enabled the matching of cases and controls, and linked samples and data. All participants were registered with the Health and Social Care Information Centre (HSCIC) and lung cancer development, disease recurrence and cause of death were recorded.

2.2.2.1 Allocation of a Liverpool lung cancer risk score

Cases and controls aged 40-80 years were allocated a score based on the LLP Model (11)(see Section 1.1.2.1.1). A score of ≥2.5% was chosen to define high risk controls because a study of 4900 lung cancer cases and 1703 healthy controls found that a score of ≥2.5% identified a higher proportion of lung cancer cases compared to a score of ≥5.0%. The identification of lung cancer cases improved from 45.5% with score ≥5.0% compared to 66.7% with score ≥2.5% (248). Although, this was at a cost of increasing the number of controls incorrectly identified as lung cancer from 15.1% to 33.4% (248).

2.3 Blood processing and sampling

Blood processing was optimised to minimise lymphocyte DNA contamination of plasma (140). Peripheral whole blood was withdrawn by venepuncture into EDTA blood collecting tubes to prevent blood clotting. Blood samples were kept on ice and processed within one hour of venepuncture by double centrifugation. The first centrifugation at 800 g, 4 °C for 10 minutes in a benchtop centrifuge formed three separate layers of plasma, lymphocytes and red blood cells. The plasma layer was removed into a 15 ml conical tube leaving at least 2 mm depth of plasma above the lymphocyte (buffy coat) layer in order to avoid contamination of the plasma with genomic DNA. The lymphocyte layer was stored at -80°C in a 1.5ml

cryovial. A second centrifugation of the plasma at 1600g, 4°C for 10 minutes allowed removal of any remaining cells whilst avoiding cell lysis. The resulting pellet consisting of platelets, cells and cellular debris was left in the tube, and the plasma supernatant was aliquoted into 1.5 ml cryovials for storage at -80°C. Small aliquots were utilised to avoid repeat thawing of plasma, as cfDNA levels can be affected (134).

Prior to further processing, thawed plasma was centrifuged at 1000g 4 °C for five minutes, to remove DNA contaminants or any precipitated material (249).

2.4 DNA extraction from tumour tissue

Tumour genomic DNA was isolated from FFPE tissue sections both to spike healthy volunteer plasma for the evaluation of cfDNA extraction methods, and to determine somatic mutations with highly parallel sequencing.

The protocol for the QIAamp® DNA FFPE kit (Qiagen, West Susex UK) was followed with the introduction of an overnight incubation step to ensure complete lysis of plasma proteins and contaminants. This commercial kit extracts DNA from plasma by adsorption of DNA onto a silica membrane.

Using a fresh scalpel for each sample, tumour tissue was scraped into a 2 ml tube. To eliminate the paraffin, 1 ml of xylene was added and the mixture was vortexed and centrifuged (17 000 g for 2 minutes). After removing the supernatant by pipetting, 1 ml of absolute ethanol was added to the pellet to extract any residual xylene followed by a repeat centrifugation. The supernatant was discarded and the remaining pellet was air dried to evaporate residual ethanol. Once dry, to lyse proteins and contaminants the pellet was re-suspended in 180 µl of buffer ATL and 20 µl of proteinase K (20mg/ml). The mixture was incubated at 56 °C in a Titramax 1000 Incubator and Shaker (Heidolph, Germany) and left overnight. An additional 20 µl of proteinase K was then added, and the mixture was incubated at 56 °C for one hour and then at 90 °C for a further hour. The very high temperature enabled the partial reversal of formalin induced nucleic acid crosslinking. Two-hundred microliters of buffer AL was added and 200 µl of absolute ethanol. After thorough mixing, the solution was transferred to a QIAamp MiniELute column. Upon centrifugation of the column (6000 g for 1 minute), DNA became bound to the silica membrane across the

bottom of the column whilst contaminants passed through the membrane into the collection tube to be discarded. Residual contaminants were washed away by two sequential centrifugation steps (600 g for 1 minute) with 500 µl of wash buffer AW1 and AW2. To dry remaining buffer, the column was centrifuged at 17 000 g for 4 minutes. Afterwards, DNA bound to the membrane was eluted into a clean 1.5 ml aliquot by adding 50 µl of elution buffer ATE (10mM Tris-Cl, pH 8.3, 0.1mM EDTA, 0.04% NaN3 (sodium azide)) for a 5 minute incubation followed by centrifugation at 17 000 g for 2 minutes.

2.5 DNA extraction from plasma

2.5.1 Phenol-chloroform method

An organic solvent extraction method based on phenol-chloroform has been reported by our group to give higher DNA yields than the Qiagen® Blood Midi Kit (Qiagen) (140). Therefore, this method was evaluated.

Up to 1 ml of thawed plasma was separated equally into two, 2 ml tubes. A solution of PBS (20% of the sample volume) and the detergent Triton X-100 (2% of the sample volume) was added to each tube to reduce the surface tension of the mixture. The mixture was incubated at 98 ^{o}C for 5 minutes to denature proteins and then cooled on ice. The resulting solid was transferred to a 15 ml falcon tube. An equal volume of phenol-chloroform was added to extract the protein. The mixture was vortexed and separated into two 1.5 ml aliquots. Centrifugation at 17,000 g for 15 minutes caused three layers to form. The bottom yellow liquid layer was phenol-chloroform, the semi-solid white middle layer was protein and the top clear aqueous layer contained the DNA. The aqueous layer was transferred to a clean 2 ml tube. Sodium acetate 3M stored at -20 ^{0}C was warmed to room temperature and added to the DNA supernatant at 10% of the supernatant volume, followed by 1 ml of ice-cold absolute ethanol to precipitate DNA. The samples were left overnight at -20 ^{o}C to allow full precipitation of longer length DNA.

The next day, samples were centrifuged at 17,000 g for 15 minutes. The supernatant was discarded leaving an invisible pellet of DNA. The pellet was washed with 1 ml of 70% ice-cold ethanol and centrifuged at 17,000 g for 10 minutes. The ethanol was removed and the pellet was air dried for over an hour. Once dry, the pellet was re-suspended in DNA/RNAase free

water at 4% of the original plasma volume and left at 4°C for one week to ensure that the DNA had fully dissolved.

2.5.2 Qiagen QIAamp commercial kits

Qiagen QIAamp® kits adsorb DNA onto a silica membrane within a column and are the most widely used commercial kit for the extraction of DNA from plasma. The QIAamp® Circulating Nucleic Acid Kit (Qiagen) was evaluated in comparison to the QIAamp® Blood Mini Kit (Qiagen).

Standard operating procedures (SOPs) for the QIAamp® Blood Mini Kit centrifugation method and QIAamp® Circulating Nucleic Acid Kit vacuum method were followed for the extraction of DNA from 1 ml and 1-3 mls of plasma respectively. These protocols were previously validated in Prof J. Shaw's laboratory at the University of Leicester.

2.5.2.1 QIAamp Blood Mini Kit method

One ml of thawed and centrifuged plasma was added to 100 µl of Qiagen protease (24 mg/ml) in a 15 ml conical tube and vortexed for 15 seconds. If less than 1 ml of plasma was available, then the volume was made up to 1 ml by the addition of PBS. One ml of the lysis buffer AL was added and the solution was vortexed for 15 seconds. The mixture was incubated at 56°C for 10 minutes, to allow enzymatic digestion of protein contaminants, inactivation of DNases and release of nucleic acids from bound proteins, lipids and vesicles. After incubation, 1 ml of absolute ethanol was added, this improves the binding of DNA to the silica membrane. Up to 600 µl of the plasma mixture was added at a time to the column, which was then centrifuged for 1 minute at 6000 g. During centrifugation, DNA in the plasma mixture binds to the silica membrane and contaminants pass through the membrane to be discarded. Residual contaminants were removed in two wash steps by adding 500 µl of buffer AW1 and AW2 followed by centrifugation at 6000 g for 1 minute and 20,000 g for 3 minutes respectively. The purified DNA was separated from the column membrane by incubating with elution buffer AE for 5 minutes followed by centrifugation at 6000 g for 1 minute. Buffer AE (10mM Tris-Cl, 0.5mM EDTA, pH 9.0) was added to the column in two steps of 70 µl and 30 µl, in order to maximise DNA yield whilst maintaining adequate DNA concentration by maintaining a low volume.

2.5.2.2 QIAamp® Circulating Nucleic Acid Kit method

In brief, 1, 2 or 3 mls of thawed plasma were added to a 50ml falcon tube with 100 µl, 200 µl or 300 µl of Qiagen proteinase K and 0.8 ml, 1.6 ml or 2.4 ml of the lysis buffer ACL, depending on the input plasma volume respectively. After vortexing for 30 seconds, the mixture was incubated at 60 °C in a water bath for 30 minutes. Either, 1.8 ml, 3.6 ml or 5.4 ml of the binding buffer ACB (for 1ml, 2ml or 3ml plasma respectively) was added and after vortexing for 30 seconds, the mixture was incubated on ice for 5 minutes.

Instead of using centrifugal force to enable the passage of liquids through the silica membrane (described in Section 2.5.2.1) a vacuum method was used to increase efficiency and reduce labour time. This was particularly important for larger plasma volumes greater than 1 ml. The mixture was poured into a column inserted in a vacuum manifold (QIAvac 24 plus, Qiagen) (the columns had extenders to accommodate higher plasma volumes). The recommended vacuum pressure of -800 mbar was applied and the mixture was slowly pulled through the column. DNA became bound to the silica membrane and salt and pH conditions ensured that proteins and contaminants flowed through the column to be discarded. Once the mixture had passed through the column, residual contaminants were removed by applying a vacuum during sequential washes of 600 µl ACW1, 750 µl of ACW2 and 750 µl of absolute ethanol. The columns were removed from the vacuum manifold and placed in a 2 ml collection tube, which was then spun at 13,100 g for 3 minutes to remove residual liquid. To ensure the removal of all ethanol, the column was dried in a new 2 ml collection tube with the lid open on a heat block at 56 °C for 10 minutes. The column was then placed in a 1.5 ml Lo-Bind tube. The amount of AVE elution buffer added to the column was dependent on the plasma input volume. For 1 ml, 2ml or 3ml of plasma, 50 µl, 100 µl or 150 µl respectively of AVE elution buffer was applied for 3 minutes prior to centrifugation at 13,100 g for 1 minute to elute the nucleic acids.

2.6 DNA extraction from cell lines

The 250 ml FlexiGene DNA kit (Qiagen) was used to extract genomic DNA from cell lines with cell count 1-2 x10^6 cells. Following the manufacturer's instructions, cell pellets were re-suspended in 300 µl of buffer FG1. For each sample, 300 µl of buffer FG2 and 3 µl of Qiagen protease were mixed and 300 µl of this mixture was added to the re-suspended cell pellet. After briefly vortexing, the mixture was incubated in a water bath at 65 °C for 10 minutes to

facilitate cell lysis and DNA release. DNA was precipitated by further mixing with 600 µl of isopropanol followed by centrifugation at 10,000 g for 3 minutes. The tube holding the pellet was inverted to remove excess liquid, and 600 µl of absolute ethanol were added to wash the pellet. After another centrifugation step of 10,000 g for 3 minutes the supernatant was removed and the tube was inverted again for 5 minutes to dry the DNA pellet. Three hundred microlitres of buffer FG3 were added and the mixture was incubated for at least 30 minutes at 65 °C to dissolve the DNA pellet.

2.7 DNA extraction from peripheral blood mononuclear cells

To determine somatic mutations cfDNA was compared to peripheral blood mononuclear cells (PBMC) genomic DNA to allow inherited variants to be excluded. The 2 ml Chemagic DNA Buffy Coat Kit (Perkin Elmer, Baesweiler Germany) and Streptavidin M-PVA magnetic beads (Perkin Elmer) were used with the Chemagic Magnetic Separation Module I robot (Perkin Elmer) to allow high sample throughput. In this automated process, PBS was added to buffy coats prepared as described in Section 2.3, to give a final volume of 2 mls in 50 ml conical tubes. White blood cells were lysed to release DNA with the addition of 20 µl of protease and 5 ml of lysis buffer. After 20 minutes of mixing, 12 ml of binding buffer and 0.8 ml of re-suspended magnetic beads were added. In the presence of the binding buffer, DNA bound to the carboxyl group attached to the magnetic beads. A magnetic rod was inserted and the beads with the attached DNA bound to the rod. To wash off impurities, the magnetic rod was transferred from one wash buffer to another for a total of 3 washes. The DNA was eluted from the beads after a 10 minute incubation in 500 µl of TE buffer pH 8.0 (Thermo Fisher Scientific, Loughborough UK).

2.8 Extracted DNA storage

DNA was aliquoted and stored at 4 °C if it was to be processed within six months of extraction. However, all cfDNA samples were processed within seven days of plasma extraction. For long-term storage, DNA was frozen at -20 °C.

2.9 Quantification of extracted DNA

2.9.1 Quantification of cell line, tumour and genomic DNA

2.9.1.1 DNA quantification with the Qubit® 2.0 fluorometer

Cell line, tumour FFPE and genomic DNA were quantified with the Qubit® 2.0 fluorometer (Thermo Fisher Scientific) using the Qubit® dsDNA BR Assay (Thermo Fisher Scientific). The intercalating dye in this assay emits fluorescence when bound to dsDNA, enabling specific and accurate quantification in the range 100 pg/µl to 1000 ng/µl. The manufacturer's instructions were followed.

For each sample, 1 µl of Qubit® dsDNA BR Reagent was added to 199 µl of Qubit® dsDNA BR Buffer to form a Mastermix. In 0.5ml Qubit® assay tubes, 2 µl of the DNA sample was added to 198 µl of Mastermix, and 10 µl of each Qubit® dsDNA BR standard S1 and S2 were added to 190 µl of Mastermix in separate tubes. After vortexing and an incubation of at least 2 minutes, the assay tubes were inserted into the Qubit® 2.0 fluorometer. The Quant-IT dsDNA BR assay protocol was chosen and the Qubit® 2.0 fluorometer was calibrated with the freshly prepared standards 1 (0 µg/ml) and 2 (100 µg/ml) prior to each quantification. The concentration of DNA in a sample was calculated by the fluorometer with the equation below.

$$Sample\ DNA\ concentration = Qubit\ fluorescence\ value\ X\ (\frac{200}{input\ volume\ in\ \mu l})$$

2.9.2 Quantification of plasma cell-free DNA

2.9.2.1 DNA quantification with conventional PCR

In the polymerase chain reaction (PCR), a segment of DNA is amplified from just a few copies to millions of copies during a sequence of heating and cooling reactions. There are three main steps to a PCR reaction or cycle. The first step heat denatures dsDNA to form single strands. Second is the annealing step, whereby primers bind selectively to the complementary target DNA sequence. Finally, DNA polymerase synthesises a new DNA strand in the extension step. These steps are repeated and the quantity of DNA between the primers is doubled with each PCR thermal cycle.

2.9.2.1.1 SYBR Green RT-qPCR

To accurately quantify low levels of amplifiable cfDNA fragments prior to sequencing, SYBR green RT-qPCR was performed. Quantitative PCR reactions were carried out with the Real-Time PCR system from Applied Biosystems 7900HT and results were analysed with the Sequence Detector Software version 2.4 (Applied Biosystems Thermo fisher Scientific, Loughborough UK).

2.9.2.1.1.1 Absolute quantification with a standard curve

Fluorescence caused by the binding of SYBR green to dsDNA can be measured in 'real time' during PCR cycling, this is demonstrated by an amplification curve (Figure 2-1). The Ct or threshold cycle is the PCR cycle at which the fluorescence from the amplification reaches the threshold line. The Ct value of the test sample can be compared to a standard curve consisting of Ct values from known template DNA concentrations and the concentration of the test sample can thus be calculated. A 10,000 pg/µl standard of mixed individual human genomic DNA (Promega, Madison USA) was serially diluted 1:10 to create a total of five standards down to 1 pg/µl. A non-template negative control of distilled nuclease free water was included to allow exclusion of contamination of water and assay reagents.

To accurately quantify DNA using a standard curve assay quality controls were as follows. The standard curve correlation coefficient R^2 had to be greater than 0.99, and the efficiency of the PCR reaction between 90-105% (slope -3.0- -3.6). The efficiency of the PCR reaction was calculated from the slope of the curve by the equation: efficiency = $(10^{-(1/slope)} -1) \times 100\%$. In addition, a minimum of three replicates of each standard with standard deviation <0.167 was preferable (Figure 2-2).

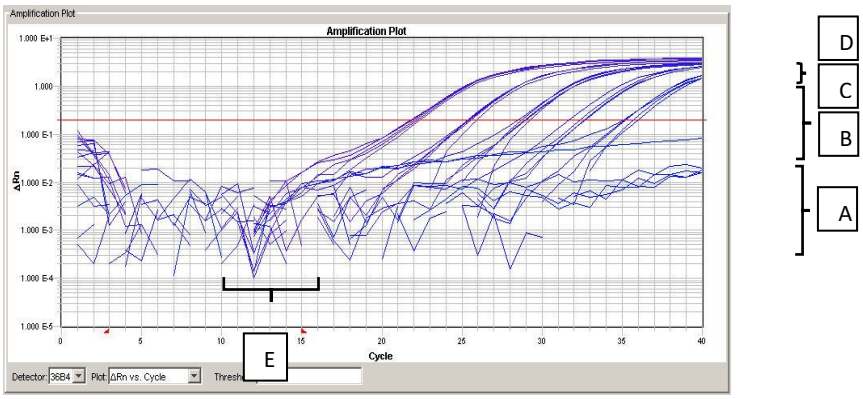


Figure 2-1: An amplification curve for serially diluted DNA standards.

Cycle number (x-axis) is plotted against ΔRn (y-axis) (fluorescence normalised to the passive reference dye minus the baseline). An amplification curve has several phases: A) Background B) Exponential C) Linear D) Plateau E) Baseline. The red line in the exponential phase denotes the threshold at which the Ct value is determined. The threshold was set to be in the middle of the exponential growth phase of the samples.

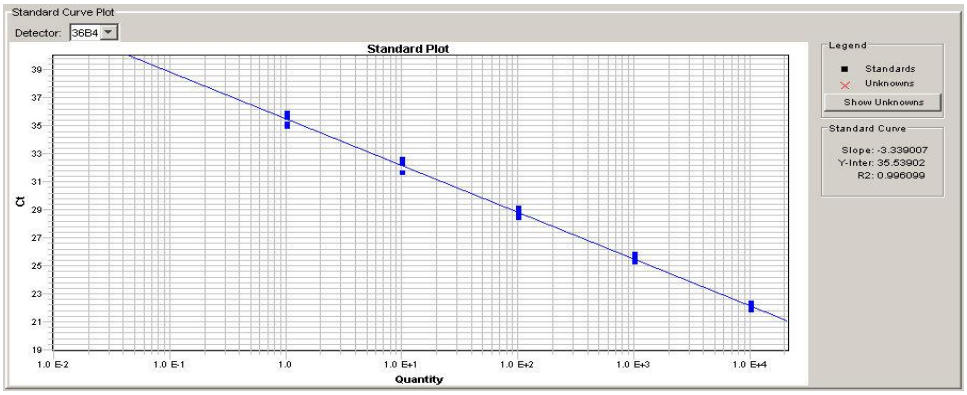


Figure 2-2: An example of an acceptable standard curve.

R^2=0.996, slope = -3.34. Y-axis represents Ct value and the x-axis DNA quantity.

2.9.2.1.1.2 GAPDH primer and SYBR green mastermix

Primers for the 81 bp housekeeping gene *GAPDH* (Glyceraldehyde-3-phosphate Dehydrogenase) were used to generate the quantified PCR product. *GAPDH* primers were purchased from Sigma-Aldrich Co. Aldrich (Ebersberg, Germany) and were re-suspended in sterile nuclease free water as per manufacturer's recommendation to attain a solution of concentration 100 picomoles per 1µl (Forward primer: 5' AACAGCGACACC CATCCTC (SY080702259-039). Reverse primer: 5' CATACCAGGAAATGAGCTTGACA (SY080702268-007)). A short amplified region (amplicon) was important to minimise the risk of non-amplification of small DNA fragments typical of cfDNA. The primer concentration in the mastermix reaction was optimised by a member of our group. The final 7 µl mastermix constituents are displayed in Table 2-1.

Mastermix reaction	X 1 well
2x SYBR Green Mastermix (Applied Biosystems Thermo fisher) *contains SYBR® Green I Dye, AmpliTaq Gold® DNA Polymerase, dNTPs with dUTP, Passive Reference, and optimized buffer components*	5µl
GAPDH 10 pM forward + reverse primer (Sigma-Aldrich)	0.3µl +0.3µl
Nuclease free distilled water	1.4µl

Table 2-1: Components of the mastermix reaction for SYBR green RT-qPCR.

dNTPs: deoxynucleotides. dUTP: 2'-deoxyuridine, 5'triphosphate.

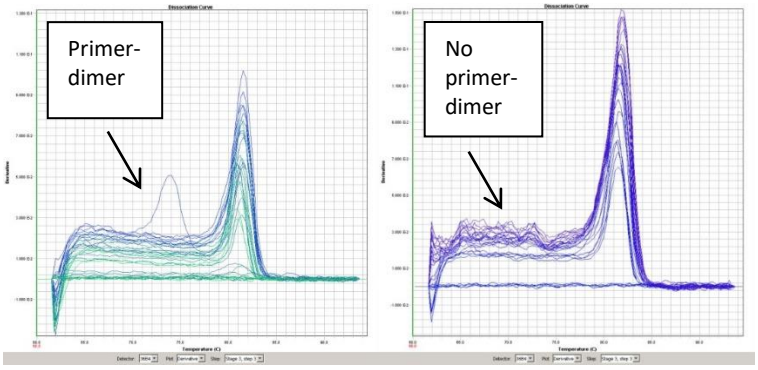
2.9.2.1.1.3 PCR plate preparation

A 384 microwell plate was prepared in the PCR preparation room in a ultra violet (UV) cabinet. To degrade potentially contaminating DNA in the cabinet the UV light was turned on for 5 minutes before and after plate set up (250). Each DNA sample was prepared in triplicate with five replicates of standard. The total volume in each well was 10 µl and consisted of 7 µl of the mastermix reaction solution and 3 µl of the DNA standard or 1:5 diluted test DNA sample. Once the plate preparation was complete, it was sealed and briefly centrifuged prior to RT-qPCR.

2.9.2.1.1.4 Thermocycling conditions and melting curve analysis

The recommended thermocycling conditions were 50 °C for 2 minutes, 95 °C for 10 minutes followed by 40 cycles of 95 °C for 15 seconds and 60 °C for 1 minute. A final step of 95 °C, 60 °C and 95 °C each for 15 seconds was carried out to check for non-specific amplification by forming a melting curve. Melting curves map temperature against change in fluorescence due to SYBR green dye interacting with dsDNA. The melt curve of the amplicon was 81 °C,

which was defined by the sequence of the primer GAPDH. A peak outside of this temperature represented primer-dimer formation or the amplification of non-specific products and the resulting DNA quantity was invalid due to the presence of fluorescence arising from other DNA species (Figure 2-3).



A) Unacceptable melting curve B) Acceptable melting curve

Figure 2-3: Melting curve analysis for the amplicon GAPDH with a melting point of 81 °C.

The y-axis shows change in fluorescence (known as derivative) and the x-axis shows temperature (°C).

2.10 Quality assessment of extracted DNA

2.10.1 Quality assessment of genomic, cell line and tumour DNA

2.10.1.1 Agarose Gel Electrophoresis

Tumour FFPE DNA samples were run on a 1.5 % agarose gel to determine DNA fragment size. To make the gel, 2.25 g of acarose powder (Sigma-Aldrich) were added to 150 mls of 1 x TAE buffer in a 500 ml conical flask. After heating in a microwave for 2 minutes to dissolve the acarose powder, 6 µl of the intercalating dye ethidium bromide 10 µg/µl was added. Once the mixture was cooled it was allowed to set to form a gel in a casting tray with inserted comb to create indents in the gel for sample loading. The gel was loaded with 3 µl of each tumour DNA sample mixed with 2 µl of 5 x loading buffer. For each gel, 5 µl of a Hyperladder (Bioline, London UK) were loaded, to allow the size of DNA electrophoresis bands to be determined. The running buffer was 1 x TAE and the gel was run at constant voltage 100 volts for 1 ½ hours using a BioRad Power Pac.

2.10.1.2 Nanodrop Spectrophotometry

DNA purity and quantity was evaluated by Nanodrop Spectrophotometry with the Nanodrop Spectrophotometer (ND-1000, Labtech International, East Sussex UK) and ND-1000 software for the detection of nucleic acids. The ratio of light absorbed by DNA at wavelength 260nm, and light absorbed at 230nm by carbohydrates, salts and phenol or 280nm by proteins were calculated. A 260nm:280nm ratio and a 260nm:230nm ratio between 1.7 and 2.2 demonstrated satisfactory DNA purity. Nanodrop Spectrophotometry lacks specificity for dsDNA quantification because single stranded DNA (ssDNA), RNA and free nucleotides all absorb UV light at 260nm.

2.10.2 Quality assessment of plasma cell-free DNA

2.10.2.1 Agilent TapeStation 2200

The 2200 TapeStation (Agilent Technologies, Cheshire UK) detects fluorescent stained dsDNA fragments that have been size separated by capillary gel electrophoresis to form bands.

The manufacturer's instructions were followed. In brief, 2 µl of high sensitivity D1000 sample buffer containing the intercalating fluorescent dye (Agilent Technologies) was added to 2 µl of cfDNA or 2 µl of high sensitivity D1000 ladder (Agilent Technologies), in a 96 well plate. The plate was vortexed to ensure the coating of all DNA fragments with the dye, centrifuged to remove droplets from the sides of wells, and inserted into the Agilent Tapestation with a high sensitivity D1000 ScreenTape (Agilent Technologies). The ScreenTape is a gel with 16 independent channels and each sample is automatically loaded into one channel prior to electrophoresis. A high quality camera captures the gel image and analysis software (version A.01.04) determines the position and fluorescence intensity of each band to assess fragment size, and sample molarity by calculating area under the curve. Fragments ranging in size from 35 bp to 1000 bp are detected with a quantitative range from 10 pg/µl to 1000 pg/µl and accuracy of ± 20%.

2.11 DNA spiking of plasma samples

To allow for the comparison of different methods of cfDNA extraction, tumour FFPE DNA quantified by nanodrop spectrometry (see Section 2.10.1.2) was added in known amounts to half of the pooled plasma from healthy volunteers. Tumour DNA is fragmented and

therefore somewhat representative of cfDNA in cancer patients. Control plasma and plasma spiked with tumour DNA were stored at -80°C, prior to DNA extraction.

2.11.1 The percentage of DNA recovery

The amount of DNA extracted from plasma was determined by taking the average quantity of DNA in 1 ml of plasma calculated from the Ct values of three qPCR replicates. The Ct value of each replicate was compared to a standard curve to determine the quantity of DNA in pg/µl. The complete workflow is demonstrated in Figure 2-4.

In order to compare methods of plasma DNA extraction the percentage of spiked DNA recovered from healthy volunteer plasma was calculated using the equation:

$$\%\ DNA\ recovery = \frac{(DNA\ ng/ml\ extracted\ spiked\ plasma - DNA\ ng/ml\ extracted\ matched\ unspiked\ plasma)}{tumour\ DNA\ spike\ ng/ml}\ X\ 100\%$$

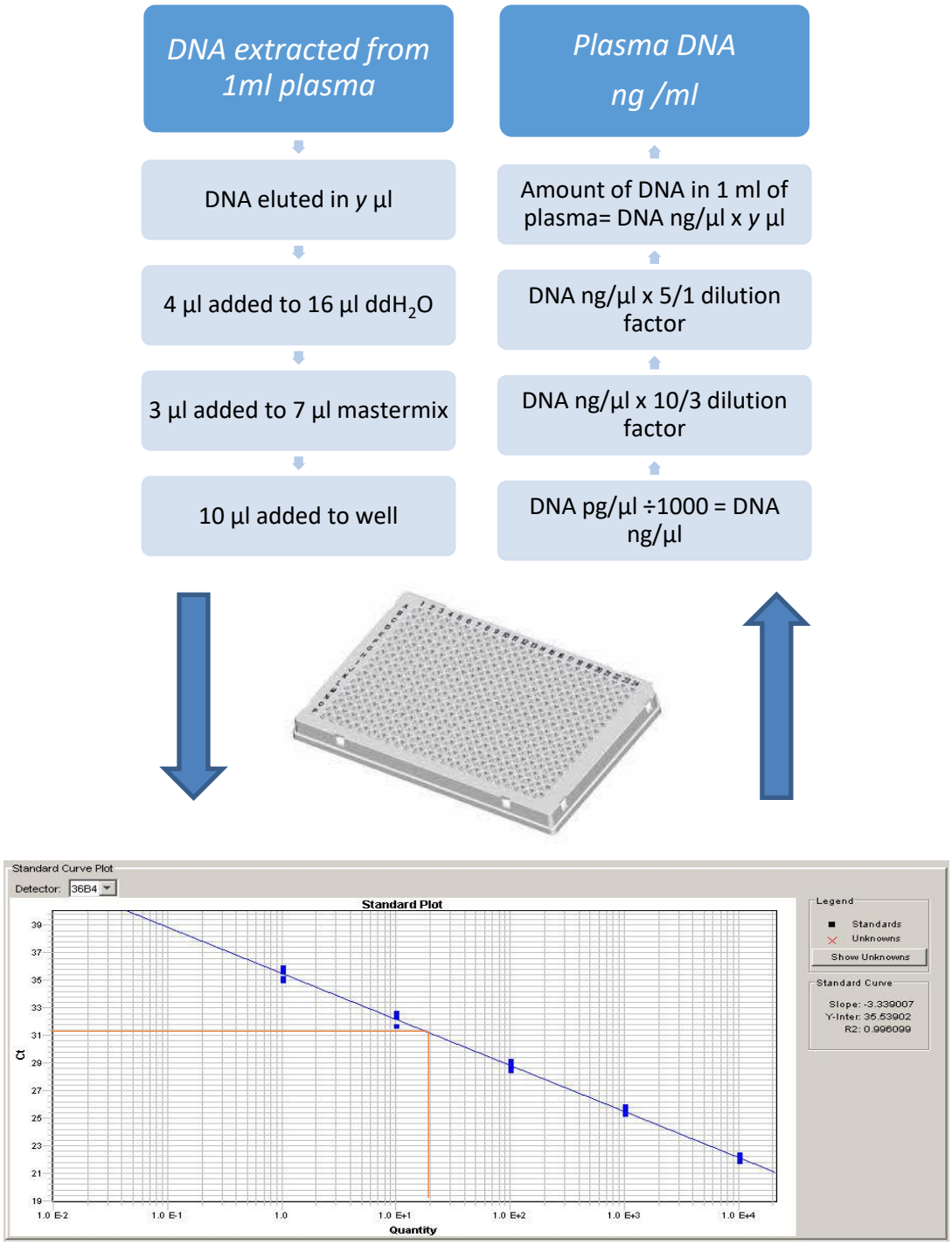


Figure 2-4: Sample workflow for SYBR green RT-qPCR plate set up and method for calculating the amount of DNA extracted from 1 ml of plasma (ng/ml).

A standard curve is displayed with R^2 >0.996 and slope -3.3. The red line demonstrates how extrapolating the Ct value of a sample leads to a quantitative value relative to the standard curve. The x-axis shows DNA quantity and the y-axis the Ct value.

2.12 Cell lines and cell culture

Human cancer cell lines to include a colorectal cancer cell line (HCT 116) and lung cancer cell lines (H69, A549, H460 and SK-MES-1), were obtained from the American Type Culture Collection (ATCC), (Bethesda, USA). Cell lines were cultured in a 37 ^{o}C incubator with 5% CO_2 according to ATCC guidelines in DMEM medium (Lonza, Slough UK) or RPMI 1640 medium (Lonza). Medium was supplemented with 10% fetal calf serum (GIBCO, Thermo fisher). Standard aseptic techniques were employed and procedures were carried out in a grade 2 safety cabinet. DNA extracted from cell lines (see Section 2.6) was used as a positive control for whole genome sequencing runs on the Illumina HiSeq2500.

2.13 Cell line authentication

STR (short tandem repeat) profiling was carried out by the core genomics facility to assess for cell line contamination in accordance with the International Cell Line Authentication Committee (ICLAC). Ten STR loci were amplified by PCR with the GenePrint® 10 System (Promega) on a MJ Research PTC-200 thermocycler (GMI, Minnesota USA). The number of tandem repeats at each allele were detected on the 3730 DNA analyser (Applied Biosystems). STR profiles for cell lines were compared and matched to reference profiles held in the COG Cell Line and Xenograft STR Database (251). The cell line was confirmed as being from the same donor if there was a greater than 80% match. This allowed for genetic drift with increasing passage and variability in testing between laboratories.

2.14 Low coverage whole genome sequencing to identify copy number aberrations

Low coverage whole genome sequencing was carried out to identify CNAs in ReSoLuCENT samples. The HiSeq 2500 system (Illumina, Essex UK) is a next generation highly parallel sequencing platform with the ability to perform whole genome sequencing. There are four main steps: library construction, template preparation, sequencing and data analyses (Figure 2-5).

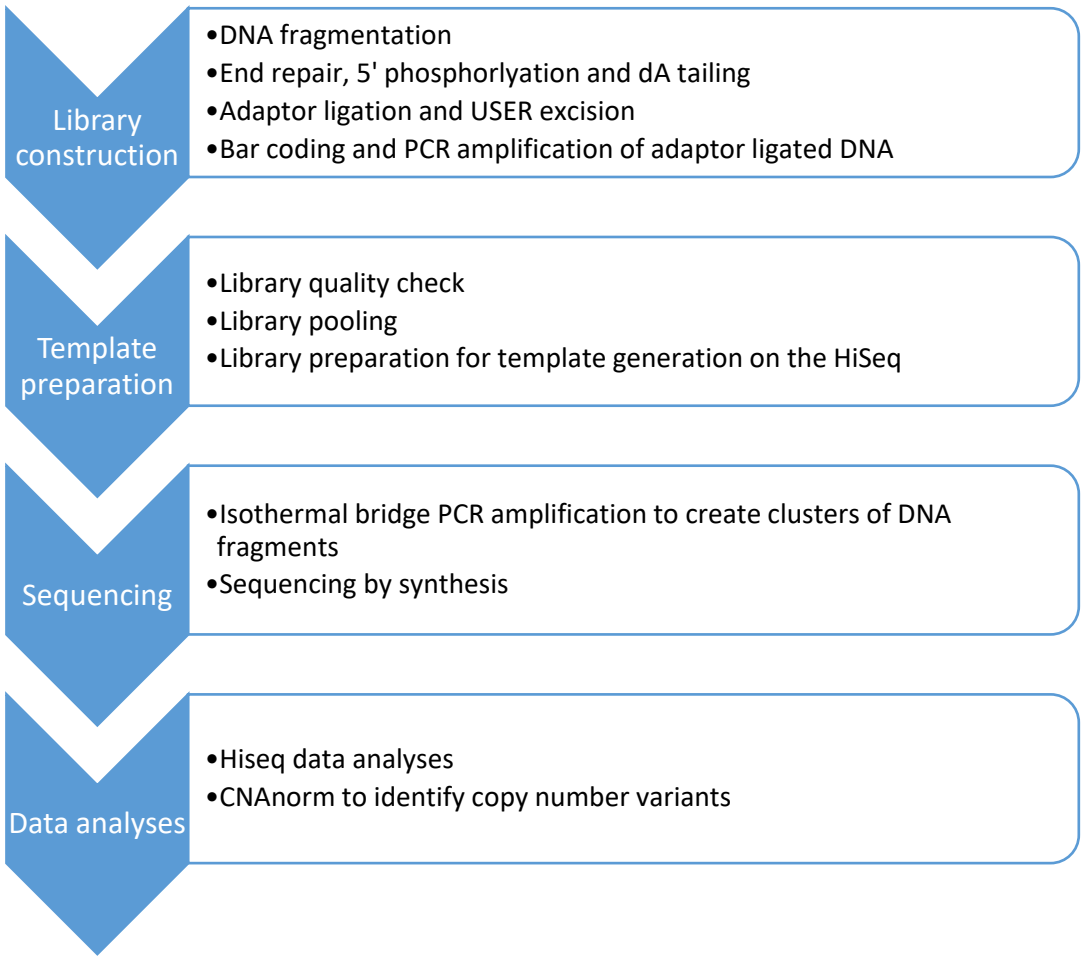


Figure 2-5: The four main steps to complete DNA sequencing with Illumina HiSeq 2500.

2.14.1 Library construction

The NEBNext® Ultra DNA library Prep Kit (NEW ENGLAND BioLabs®Inc., Hitchin UK) was used to prepare libraries for Illumina sequencing. This kit was chosen because it was validated for between 5 ng to 1 µg of tumour FFPE DNA, was cost effective and required only 3 hours of hands on laboratory time. Figure 2-6 summarises the processes involved.

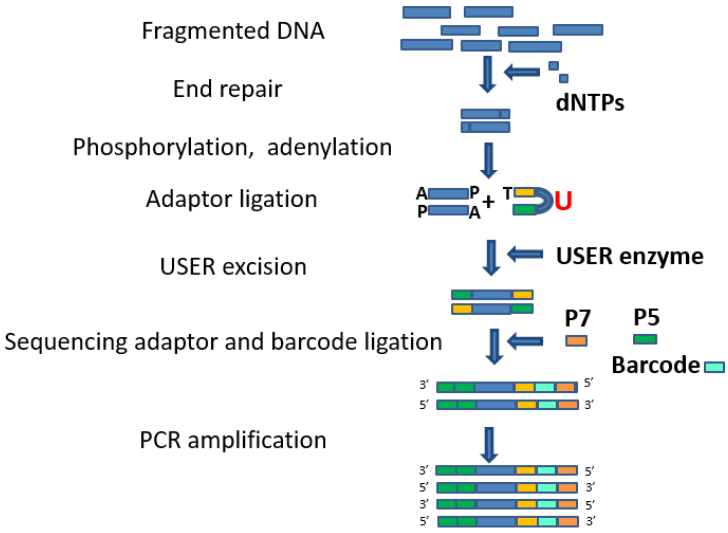


Figure 2-6: An overview of library construction with the NebNEXT® Ultra DNA library Prep kit.

Adapted from www.neb.com (2017) and reprinted with permission from New England Biolabs, Inc.

2.14.1.1 DNA fragmentation

Random DNA fragmentation is an important step in library preparation to ensure unbiased coverage of the whole genome. For each DNA fragment, the number of nucleotides that can be sequenced with Illumina technology is limited to 2x250 bp for forward and reverse reads in bi-directional sequencing, therefore it is important to have short fragments to achieve adequate genome coverage. Furthermore, short fragments are important to avoid overlapping reads where by the same fragment is sequenced twice thus reducing sequencing efficiency.

Tumour FFPE DNA and genomic DNA were diluted in nuclease free sterile water and mechanically sheared by high frequency ultrasonic acoustic waves with the Covaris® E220 Focused-ultrasonicator (Covaris, USA). The standard settings were followed for a target fragment length of 200 bp except for length of treatment time, which was optimised according to DNA concentration (Table 2-2)(252). A peak fragment size of 200 bp was chosen

to give an insert size compatible with sequencing paired end reads of 100 bp. Shearing cfDNA was not indicated because cfDNA fragment size is generally <200 bp (see Section 1.2.3.1). Furthermore, it was important to minimise the loss of DNA given the low quantities present in cfDNA. CfDNA and fragmented samples were run on the Agilent 2100 Tapestation to evaluate the size distribution of fragments prior to library construction (Figure 2-7)(see Section 2.10.2.1).

Covaris® S220 settings for a target peak fragment length of 200bp	100 ng DNA in 130ul	1000 ng DNA in 130ul
Treatment time in seconds	360	430
Peak incident power: power emitted during each ultrasonic acoustic wave burst	175 watts	
Duty factor: percentage of time that the covaris instrument applied power during each burst	10 %	
Cycles per burst: number of sound waves/acoustic oscillations per burst	200	
Temperature	7 °C	
Water level	6	

Table 2-2: The Covaris® E220 Focused-ultrasonicator settings utilised to shear genomic and tumour DNA.

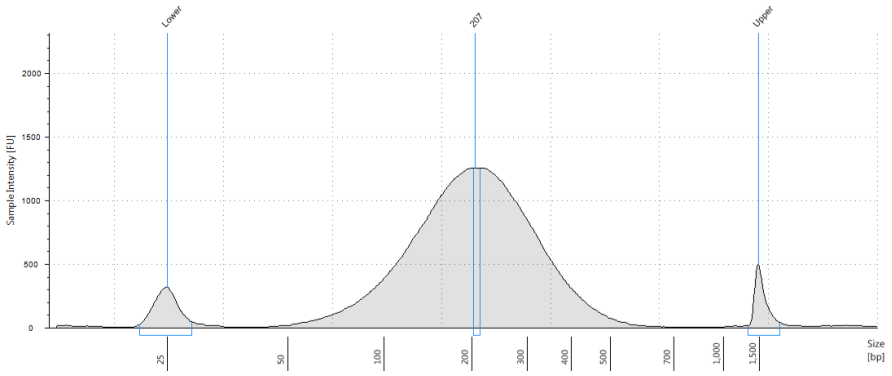


Figure 2-7: An example of the Agilent Tapestation 2000 output after shearing 1000 ng of genomic DNA.

Sample Intesity (FU) is stated on the y-axis and fragment size on the x-axis. The peak fragment size is 207 bp. Upper and lower markers are demonstrated.

2.14.1.2 End repair, 5' phosphorylation and dA tailing

Shearing DNA produces different sized fragments with inconsistent 3' and 5' ends. Therefore, recessed ends were filled in and overhung ends were degraded so that all fragment ends were uniformly blunt. After end repair, 5' ends were phosphorylated and 3' ends were adenylated (dA tailing) to enable adaptor ligation.

Three microlitres of NEBNext® End repair enzyme mix and 6.5 µl of NEBNext® End repair reaction buffer were added to 55.5 µl of DNA. The reaction mixture was placed in a thermocycler at 20 °C for 30 minutes followed by 65 °C for 30 minutes.

2.14.1.3 Adaptor Ligation and USER excision

NEBNext® Adaptors have a single 'T' overhang and bind specifically to the single 'A' overhang of the adenylated 3' DNA fragment end, this minimises detrimental adaptor-adaptor ligation. For PCR amplification of adaptor ligated DNA the stem adaptor loop between the 5' and 3' end is opened by eliminating 'U' in the middle of the loop (Figure 2-8).

Fifteen microlitres of Blunt/TA ligase Master Mix, 2.5 µl of 1:10 diluted NEBNext® Stem Loop Adaptor (1.5 µM) and 1 µl of Ligation enhancer were added to the reaction mixture. To allow adaptor ligation, the mixture was incubated at 20 °C for 15 minutes. Three microlitres of User enzyme were added followed by a 37 °C incubation for 15 minutes. A clean up step was carried out to eliminate PCR contaminants with the same volume of AMPure® XP beads (Beckman Coulter, Inc London UK) as the DNA product as described in Section 2.14.2. Therefore, fragments of size > 100 bp were retained and adaptors were eliminated. Products were eluted in 28 µl (with High Fidelity PCR Master Mix) or 17 µl (with Hotstart Q5 PCR Master Mix) of 0.1X T.E (Tris-Acetate pH 8.0) and stored at -20◦C prior to barcoding and PCR amplification.

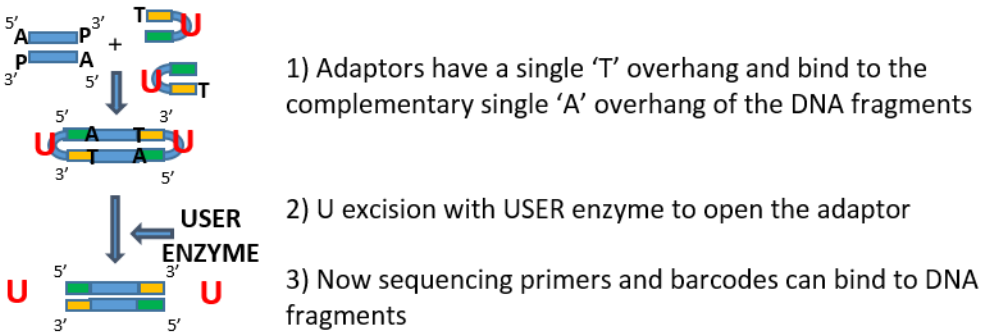


Figure 2-8: Adaptor ligation and U excision to open up the stem loop adaptor with USER enzyme during DNA library preparation.

Adapted from www.neb.com (2017) and reprinted with permission from New England Biolabs, Inc.

2.14.1.4 Barcoding and PCR amplification of adaptor ligated DNA

NEBNext® Master Mix, a unique NEBNext® barcode (NEW ENGLAND BioLabs®Inc.) and the Universal primer mix containing the universal primers P5 and P7 were added sequentially in quantities displayed in Table 2-3 to thawed adaptor ligated DNA. Adaptor ligated DNA fragments were selectively amplified by PCR with thermocycling conditions shown in Table 2.4. During the first few cycles, the barcodes and PCR primers P5 and P7 were ligated to DNA fragments. Universal PCR primers P5 and P7 have complementary sequences to oligonucleotides bound to the solid surface of the Illumina flow cell and are therefore required for hybridisation of library fragments to the flow cell (Figure 2-6).

	High Fidelity 2X PCR MM	Q5 Hot start HiFi PCR MM	
Adaptor ligated DNA fragments	23 µl	15 µl	15 µl
Master Mix (MM) containing DNA polymerase, dNTPs, Mg^{2+} and a propriety buffer	25 µl	25 µl	25 µl
Barcode/Index primer	1 µl (25 µM)	2 µl (10 µM)	10 µl (10 µM)
Universal PCR primer	1 µl (25 µM)	2 µl (10 µM)	
Sterile nuclease free water	0 µl	6 µl	0 µl
Total volume	50 µl		

Table 2-3: Reagent quantities utilised prior to PCR amplification of adaptor ligated DNA dependent on the DNA polymerase and concentration of primers.

	PCR cycling conditions for the High-Fidelity PCR master mix			PCR cycling conditions for the Q5 Hot Start HiFi PCR Master Mix		
Step	Temp ◦C	Time in secs	No. of cycles	Temp ◦C	Time in secs	No. of cycles
Initial denaturation	98	30	1	98	30	1
Denaturation	98	10	8-15*	98	10	7-16*
Annealing	65	30		65	75	
Extension	72	30				
Final extension	72	5 mins	1	65	5 mins	1
Hold	4 ∞					

Table 2-4: PCR cycling conditions for the amplification of adaptor ligated DNA with the NEBNext® Ultra DNA Library Prep kit.

*12-16 cycles were utilised for 5 ng cfDNA, 10-12 cycles for 50 ng cfDNA, 8-9 cycles for 100 ng tumour FFPE DNA and 7-9 cycles for 100 ng genomic DNA (see Section 4.3.1.1).

The NEBNext® High-Fidelity PCR master mix contained Q5 Phusion High-Fidelity DNA polymerase with 3' to 5' directional exonuclease activity. The NEBNext® Q5 Hot Start HiFi PCR Master Mix replaced the NEBNext ®High-Fidelity PCR master mix as an update to the NEBnext® Ultra kit. This resulted in the elimination of DNA polymerase activity at room temperature and therefore increased enzyme efficiency

The number of PCR cycles were determined by input DNA quantity and quality (see Table 4-1). It was important to avoid over amplification since bias would be introduced by the preferential amplification of smaller fragments. After PCR amplification, a second clean up step was carried out with AMPure® XP beads to eliminate PCR contaminants, using the same volume of beads as library product as described in Section 2.14.2. More selective size selection of 320 bp fragments (DNA insert + adaptor + primer) was not performed to minimise DNA loss given the low quantities of cfDNA. The final DNA product or library was eluted in 33 µl of 0.1X T.E., 29 µl of the final product was removed to avoid bead carryover and stored at -20◦C prior to sequencing.

2.14.2 Reaction clean up with AMPure® XP Beads

Reaction clean-ups were manually carried out with re-suspended AMPure® XP Beads (Beckman Coulter). The paramagnetic beads are coated in carboxyl molecules that bind reversibly to DNA fragments in the presence of polyethyl glycol (PEG) and salt.

The size of DNA fragments that bind to the beads or the size of fragments left in solution is dependent on the concentration of PEG, which is determined by the bead: DNA volume ratio. The lower the volume ratio of beads to DNA, the larger the DNA fragments that bind to the beads and the larger the size of the small fragments that remain in the supernatant. For example, a ratio of one was expected to retain fragments >100 bp and exclude primers, which are < 50 bp. In comparison, a ratio of 0.7 facilitates the elimination of fragments < 150 bp that remain in the supernatant, and is used to eliminate adaptor dimers.

In a 96 well plate, the DNA sample was pipette mixed ten times with a pre-specified volume of paramagnetic beads and incubated for 5 minutes. DsDNA bound to the beads, and a magnet was used to separate the beads from the supernatant during a 5-minute incubation. The supernatant containing unincorporated dNTPS and PCR contaminants such as salts, enzymes and excess primers was removed. The remaining DNA-bead pellet was washed

twice for 30 seconds with 200 µl of freshly prepared 80% ethanol. After air drying the pellet for 5 minutes to remove excess ethanol, the pellet was re-suspended in 0.1X T.E elution buffer by pipette mixing 10 times. A magnet was applied after a 2-minute incubation, for 3 minutes, to separate the beads from the eluted DNA and the eluted DNA was removed to a clean tube leaving behind a few microlitres to avoid bead carryover.

2.15 Library Quality Control

2.15.1 Determination of library quality and quantity with the Agilent Tapestation 2200

Adaptor barcode ligated dsDNA Libraries were diluted 1:5 and analysed for quality and quantity with the Agilent Tapestation 2200 (Figure 2-9). Each library electrophenogram was checked for a shift in the size of the peak fragment size. A gain in the peak fragment size of at least 126-128 bp from baseline indicated successful barcode adaptor ligation.

The presence of a peak at approximately 125 bp indicated adaptor-dimer formation and a peak at 60 bp primer-dimer formation. Significant adaptor-dimer or primer-dimer contamination can reduce sequencing efficiency because they form clusters on the flow cell and are sequenced. Adaptor-dimer contamination was avoided by diluting adaptors 1:10 for input DNA <100 ng. The molar concentration (pmol/l) of a library was calculated by the area under the curve.

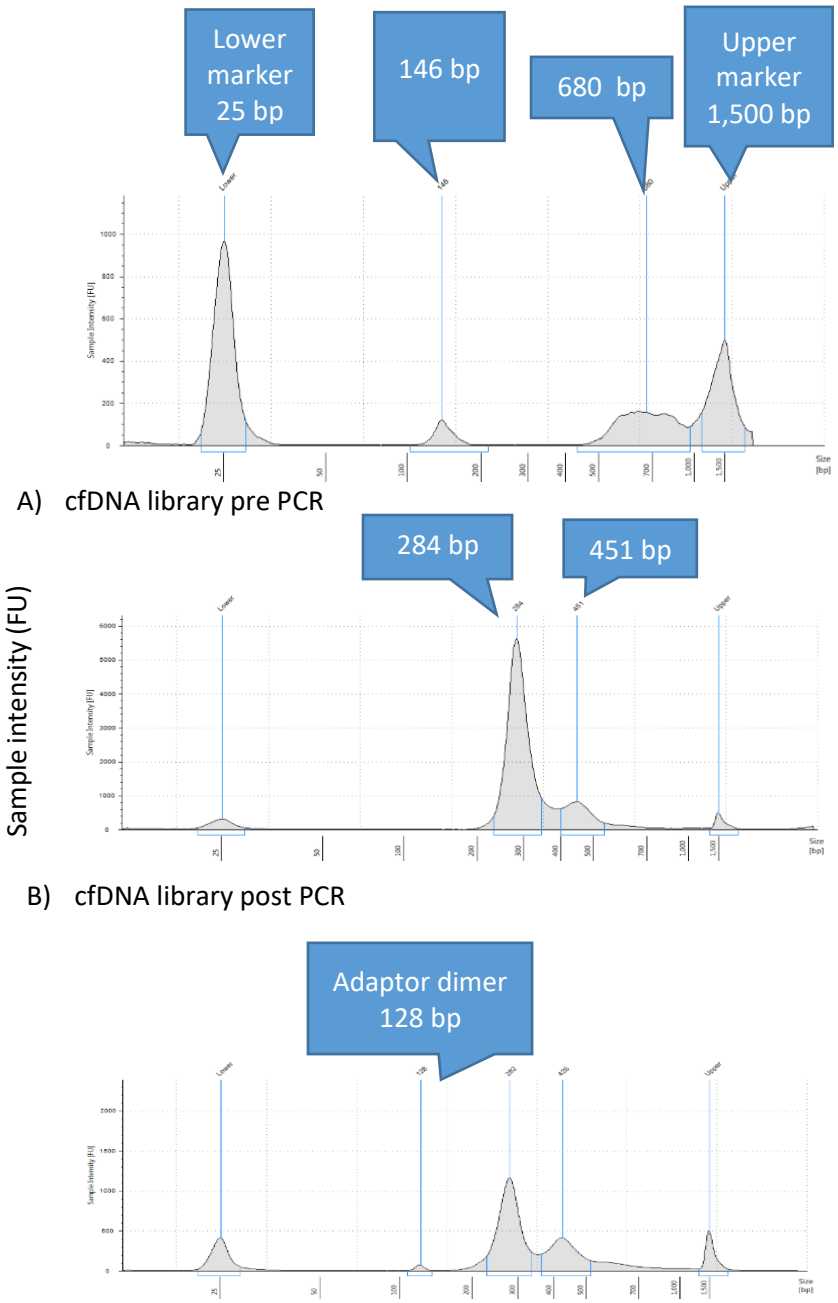


Figure 2-9: Representative Agilent bioanalyser traces of pre and post library cfDNA products. The y-axis represents sample intensity (FU) and the x-axis fragment size in bp. A) cfDNA library pre PCR. B) cfDNA library post PCR. C) cfDNA of a different library post PCR with presence of adaptor-dimer.

2.16 Equimolar pooling of library products

Accurate library quantification and pooling was vital to obtain the optimal density of clusters on the flow cell to ensure high quality base calls and a high number of reads. An in house equation was utilised to calculate the volume of a library for pooling (see below). This equation was based on the molar concentration of multiple pooled genomic DNA libraries that had optimal cluster densities to give a final pooled library of 4nM. The concentration of each sample in the final pooled library was 4nM divided by the total number of samples pooled.

Step 1:

$$Tapestation\ region\ pmol\ per\ l\ \times adjustment\ factor\ 0.0028 = adjusted\ concentration\ nM$$

Step 2:

$$2000\ \div\ (adjusted\ concentration\ nM \times adjustment\ factor\ dependent\ on\ total\ no. samples) = volume\ in\ 500\ \mu l$$

Step 3:

$$Volume\ in\ 500\ \mu l\ \times \text{adjustment factor } 1.75\ \times 3 = \text{adjusted volume in } 1500\ \mu l$$

Step 4:

$$Volume\ of\ buffer\ EB + 0.1\%\ Tween = 1500\ \mu l - \sum adjusted\ library\ volume\ in\ 1500\ \mu l$$

2.17 Whole genome low coverage sequencing

2.17.1 Illumina next generation sequencing

Illumina NGS relies on isothermal bridge PCR amplification of adaptor ligated DNA fragments to create clusters of fragments from a single DNA template to enable highly parallel sequencing (253). Reversible dye terminator chemistry and the identification of fluorescence signals unique to each nucleotide enables the sequence of nucleotides of a DNA template to be identified (253).

The Illumina HiSeq 2500 was prepared for sequencing according to the manufacturer's instructions by Dr Emily Boardman. Rapid sequencing runs were completed using a single flow cell to obtain approximately 300 million reads of length 100 bp, within 27 hours. Low coverage was ensured by multiplexing a maximum of 48 samples so that the expected number of reads per sample was approximately 6 million. The human genome consists of 3 billion nucleotides and therefore 6 million reads of length 100 bp equates to reading 600 million nucleotides and a genome coverage of 0.2X (=600 million/3 billion). Dr Emily Boardman loaded the flow cell and pooled library onto the HiSeq and commenced the sequencing run.

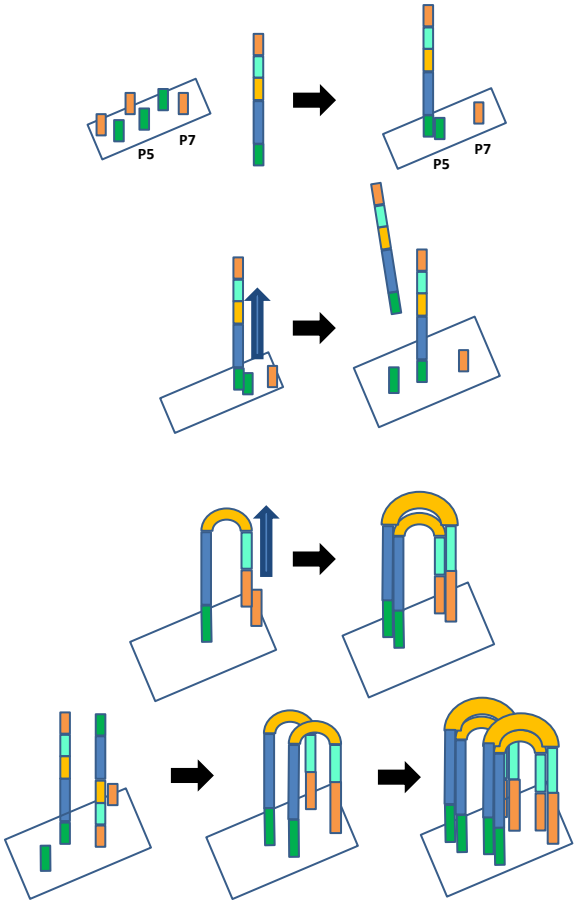
2.17.1.1 Pooled library preparation

Standard protocols were followed to prepare pooled libraries for sequencing on the Illumina HiSeq 2500 (Illumina) with the TruSeq Rapid PE Cluster kit (Illumina). A PhiX control v3 (Illumina) was spiked into the pooled sample library as a positive control and to maintain library complexity (see Section 2.17.1.3.1).

Five microliters of pooled library and separately 5 µl of diluted PhiX control v3 were incubated for 5 minutes with 5 µl of 0.2N NaOH (Illumina) to denature dsDNA. On an ice-block to prevent double strands reforming, 990 µl of pre-chilled Illumina buffer HT1 was added to each eppendorf to attain concentration 20 pM. The mixtures were diluted further with buffer HT1 dependent on the concentration (pM) required to obtain optimal cluster density. 416 µl of the DNA mixture was vortexed with 4 µl of the equivalent concentration of PhiX control v3 (1%) to create a final mixture ready for sequencing.

2.17.1.2 Cluster formation

Figure 2-10 demonstrates how clusters were generated. The upper and lower glass surface of the Illumina flow cell is coated in two types of oligonucleotides that are complementary to the universal PCR primers (P5 and P7) ligated to DNA libraries. This enables hybridisation of the ssDNA fragments to the flow cell. Isothermal PCR bridge amplification generates hundreds of thousands of clusters consisting of clonally amplified DNA fragments. Once cluster formation is completed sequencing begins.



1) Hybridisation of barcode adaptor ligated ssDNA to complementary primers on the flow cell

2) A new strand is synthesised by DNA polymerase and then the template strand is washed away

3) The new strand hybridises to a complementary primer on the flow cell. A new DNA strand is synthesised

4) The dsDNA is denatured leaving two ssDNA fragments. The process is repeated multiple times (from step 2) to generate clusters

Figure 2-10: Isothermal bridge amplification to generate clusters for Illumina sequencing.

dsDNA: double stranded DNA. ssDNA: single stranded DNA. Adapted with permission from Bentley et al 2008 (253).

A 1.5 ml eppendorf holding the final library mixture was placed in the HiSeq 2500 by Dr Emily Boardman. The mixture entered the dual lane HiSeq Rapid PE Flow Cell (Illumina) for on-board automated cluster generation with the TruSeq Rapid PE Cluster kit (Illumina). However, if only one lane of the flow cell was utilised then up to 24 libraries were pooled and cluster generation was commenced on the cBot (Illumina) by Dr Emily Boardman with the TruSeq Rapid Duo cBot Sample Loading Kit (Illumina) and completed on the HiSeq 2500.

2.17.1.3 DNA sequencing with Ilumina HiSeq 2500

Fluorescently labelled nucleotides (A, C, G, T) were washed over the flow cell. The nucleotide that was complementary to the nucleotide of the DNA template competed to be incorporated into the new strand. The other nucleotides were washed away and laser excitation caused fluorescence to be emitted. The flow cell was imaged and clusters emitting

a signal were identified. The wavelength and intensity of the signal determined the called base. Importantly, only one nucleotide at a time was incorporated into the growing DNA strand due to the presence of a blocking 3'-OH group (253), therefore homopolymers were accurately detected. The blocking group and fluorescence group were cleaved after imaging to facilitate binding of the next complementary nucleotide as the next cycle began (253). The number of sequencing cycles determined the read length. The Truseq rapid SBS 200 cycle kit (Illumina) was used for sequencing.

2.17.1.3.1 PhiX control and phasing/pre-phasing errors

The PhiX control v3 was an adaptor ligated DNA library, established from the well characterised small viral genome PhiX. PhiX has almost equal AT and GC content and therefore its addition to a pooled library preserves a balance between AT and GC content. This is important because fluorescence signals are more easily differentiated in complex libraries, which leads to accurate cluster identification and precise correction of phasing and pre-phasing errors. During sequencing, if a newly synthesised strand within a cluster falls behind by one base this is called phasing and if it is ahead by one base this is called pre-phasing.

2.18 Data analysis

The interpretation of sequencing data to obtain somatic CNAs requires many different analyses steps. Both Dr Emily Boardmand and Dr Lucy Crooks performed bioinformatics analyses to construct the FASTQ files once sequencing was completed. Dr James Bradford took the FASTQ files and processed the data according to agreed instructions to establish copy number ratios with the software CNAnorm. Figure 2-11 summarises the data analysis process. The bioinformatics scripts were supplied by Dr James Bradford and are found in Appendix D.

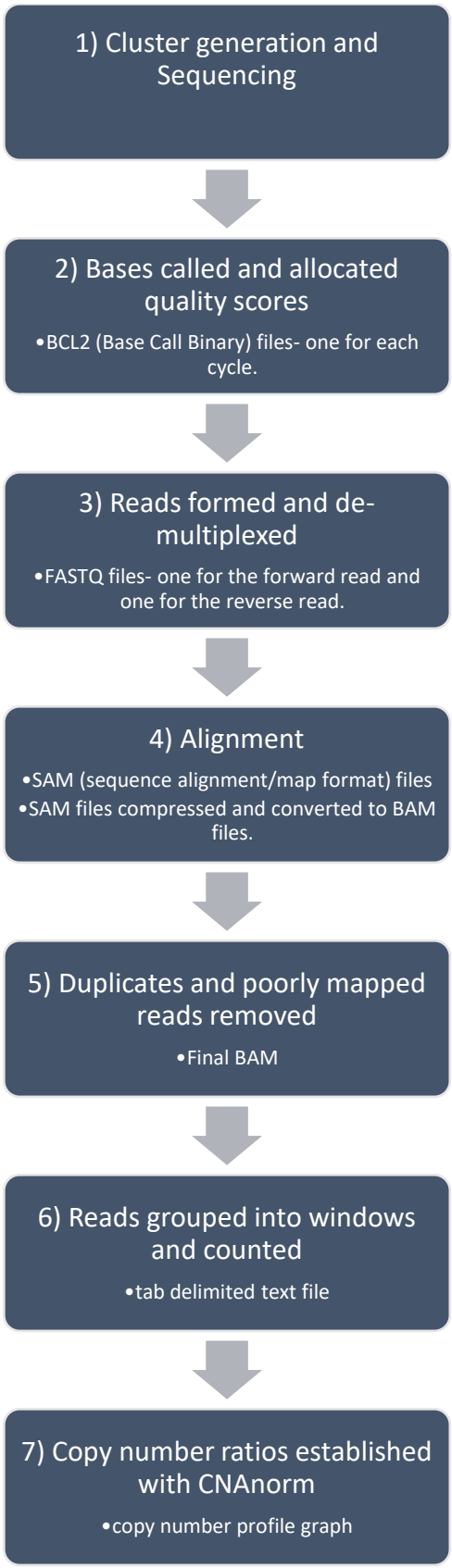


Figure 2-11: A summary of the steps to analyse sequencing data to obtain copy number aberrations.

2.18.1 Cluster identification/template generation and base calling

Data analysis initially took place on the HiSeq 2500 with software version 2.2.68. During the first 4-7 sequencing cycles, clusters on the flow cell were identified and their positions were marked by X and Y co-ordinates. These co-ordinates were used to determine the colour and intensity of emitted fluorescence from each cluster. Up to cycle 25, clusters were removed if the fluorescence intensity was low or if clusters were of poor quality.

Each of the four nucleotide bases emitted fluorescence at a unique wavelength represented by four different colour channels. Base calling occurred in real-time for every cluster in each sequencing cycle. After cycle 12, corrections were applied to the fluorescence intensity if two or more bases emitted fluorescence from a cluster in the same cycle (cross talk correction). After cycle 25, phasing and pre-phasing corrections were applied to correct for DNA polymerase errors and Phred like quality scores were allocated to called bases. The base quality score took into account a number of cluster parameters to include signal to noise ratio and fluorescence intensities. Once the sequencing cycles were completed all clusters had to pass a quality control step to remove unreliable clusters from further data analyses. To pass, the ratio of the brightest base intensity divided by the sum of the brightest and second brightest base intensities had to be greater than 0.6 for all but one base calls in the first 25 cycles. The sequencing run was failed if fewer than 80% of clusters passed this step.

2.18.2 Formation of reads and de-multiplexing

The raw base calls from each cycle were combined to create sequences of bases or reads. Adaptors were trimmed and reads were separated by their indices (de-multiplexed).

2.18.3 Alignment

The BWA: Burrows-Wheeler Alignment tool was used with default parameters to align reads to the reference human genome GRCh38. This tool was optimal for DNA sequencing and short paired-end reads (254). Each read had a mapping quality score that quantified the Phred-scale probability that the alignment was correct.

2.18.4 Removal of duplicates and poorly mapped reads

PICARD software (version 2.1.0) was utilised to mark duplicate reads, which had identical start and end chromosomal co-ordinates (255). Duplicate reads were removed to reduce PCR bias. Furthermore, only uniquely mappable reads with mapping quality score >37 were retained therefore eliminating reads that mapped to a large number of regions (usually within repetitive regions) or those with poor quality base calls.

2.18.5 Determination of sample coverage

To determine the amount of the genome of each sample that was sequenced known as the coverage it was assumed that no reads overlapped. The coverage was calculated by determining the number of bases sequenced (total number of mapped reads x read length) divided by the size of the human genome, 3 billion bases (256). This calculation was carried out after bases in reads, with low mapping quality (<20), marked as duplicates, without a mapped mate pair were removed. Furthermore, after the fore mentioned filtering steps, bases were subtracted in the second observation from an insert with overlapping reads.

2.18.6 Copy number analysis to determine somatic copy number aberrations

2.18.6.1 Creating read profiles for cell-free DNA/tumour DNA and matched genomic DNA

Uniquely mapped reads with high quality scores (>37) from matched genomic DNA (control) and tumour or cfDNA (test) were separated into windows using the bam2windows.pl PERL script (257). Windows were non-overlapping and of fixed size of 1 Mb, to facilitate comparisons across individuals and to reduce signal noise. The number of reads in each window were counted for the test and control, and GC content determined by comparison to the reference human genome GRch38.

2.18.6.2 Removal of ENCODE blacklisted genomic regions

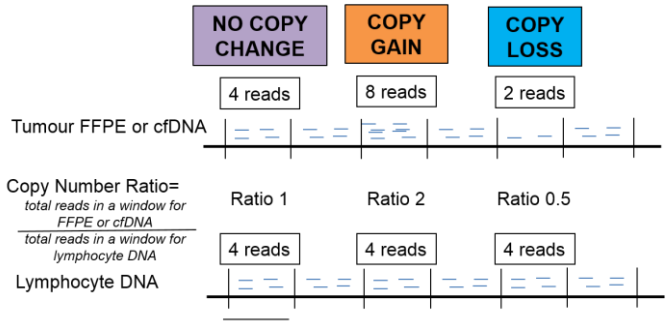
Certain genomic regions can have a misleadingly high number of reads aligned because they consist of repetitive regions (peri-centromeric, telomeric ends, satellite regions) or are a result of sequencing artefact or poor DNA quality. To reduce false positives, any window that overlapped regions blacklisted by ENCODE (The encyclopaedia of DNA elements) were removed prior to input into CNAnorm. Both ‘DAC’ and ‘DUKE’ ENCODE black listed regions

based on hg19 coordinates were downloaded from UCSC and converted to GRCh38 coordinates using the "liftover' tool (258).

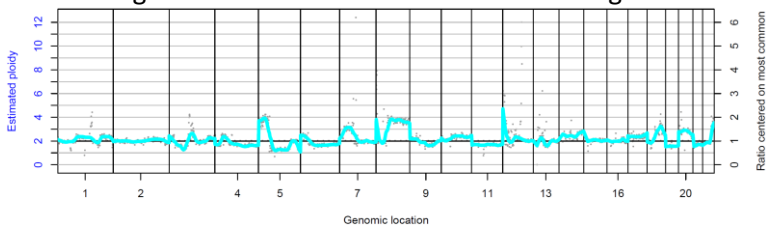
2.18.6.3 Analysis with CNAnorm to identify copy number gains and losses

Read copy number profiles were established for each sample in R (3.2.3) (259) using CNAnorm (version 1.16.0) (260) from Bioconductor (version 3.2) (261). This programme was chosen because CNAs were detected from just 5 ng of FFPE DNA at low coverage and germline aberrations were eliminated by comparison to matched genomic DNA (262). The characteristics of the tested tumour FFPE DNA are similar to those expected for cfDNA, small fragments and low quantities. Furthermore, CNAnorm adjusts for, GC content, contamination caused by the presence of non-tumour cell derived DNA, aneuploidy, and different levels of coverage between matched test and control samples (260). Correction of GC content is important to eliminate bias introduced by PCR in library preparation and sequencing. Figure 2-12 outlines the steps involved to determine copy number ratios with CNAnorm.

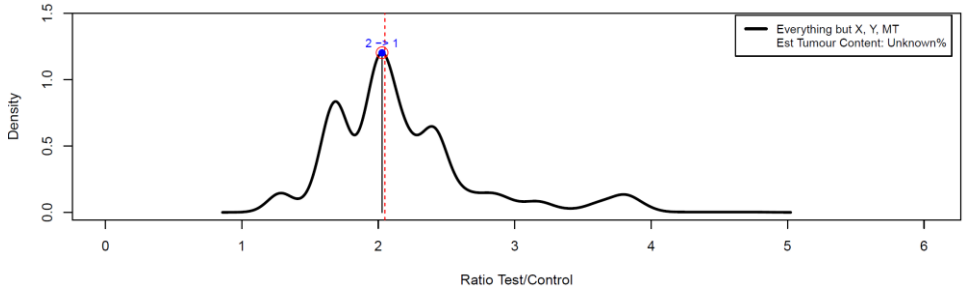
- Copy number ratios were determined for each window by comparing read counts for matched test and control samples



- Copy number ratios were corrected for GC content
- Smooth segmentation was carried out to reduce signal noise and random error



- Adjustments were made for sample contamination, ploidy and varied coverage with 'closest normalisation'. The plot is shown below. The peak of the most frequent copy number ratio closest to the median is identified, normalised to a ratio of 1 and the data is shifted.



- Segmentation of normalised copy number ratios to identify point changes in copy number
- Formation of a copy number profile. Each dot represents a 'window' and black lines or segments represents consecutive windows of a similar ratio. The colours orange show copy number gains and blue shows copy number losses.

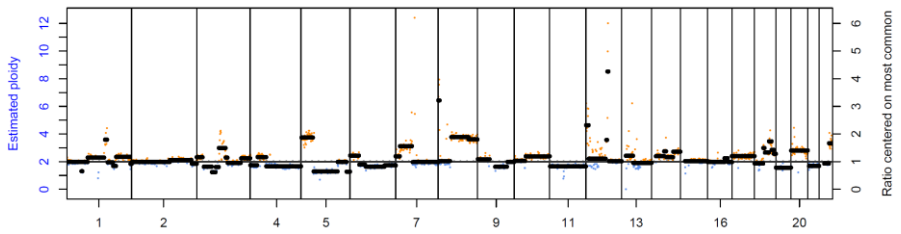


Figure 2-12: The determination of somatic copy number aberrations with CNAnorm.

To establish somatic CNAs the following steps were performed by Dr James Bradford. Firstly, read profiles for cfDNA/tumour DNA (test) were normalised against read profiles for genomic DNA (control) and corrected for GC content to establish copy number ratios. GC content was calculated by dividing the number of G and C nucleotides in a window by the total number of nucleotides in GRCh38 in the same genomic region. Secondly, these ratios were effectively 'smoothed' to reduce signal noise and eliminate significant outliers (260). Thirdly, data was normalised by 'closest normalisation', this corrected for differences in ploidy, genome coverage and contamination between matched test and control samples. Closest normalisation identified the most frequent copy number ratio closest to the median ratio and shifted the data so that this ratio now defined two copies, which is the expected ploidy for a diploid genome. Finally, segmentation analysis was carried out in CNAnorm with the programme DNAcopy (263). This analysis identified point changes in copy number ratios across the genome and grouped genomic regions into segments where copy numbers were constant (263). These data were then plotted against chromosome location to attain a copy number profile for each sample.

A copy number ratio of one was equivalent to two copies of DNA and was expected for a diploid genome. A ratio of two was equivalent to four DNA copies and described a copy number gain whilst a ratio of 0.5 was equivalent to one DNA copy and copy number loss.

2.18.6.4 A 1 Mb window of fixed size was chosen for copy number analyses

The use of different window sizes to group reads across the genome were explored in collaboration with Dr James Bradford. A small window size resulted in increased noise and a greater number of windows with no reads. In comparison, a larger window size resulted in the loss of small CNAs. To balance resolution and signal noise, a window size of 1 Mb was chosen to define copy number aberrations and to determine a genomic instability score.

A window size was defined as a fixed number of bp rather than by a fixed number of sequencing reads. With window size defined by bp, the number of windows remain the same across samples allowing direct comparison of CNAs across samples.

2.19 Genomic Instability Score

The hallmarks that lead to the development of cancer are acquired though successive mutational events caused by genomic instability (43). Therefore, tumour DNA is genetically diverse compared to normal host cell DNA. In cancer, CNAs leading to loss or gain of genome segments account for a high proportion of genomic variation (264) and may therefore provide a highly specific and sensitive biomarker for the detection of genomic instability. Genomic instability scores aim to quantify genetic aberrations by measuring the magnitude and/or the number of CNAs.

Based on the detection of tumour-derived CNAs in cfDNA by whole genome sequencing (see Section 2.18.6), several tumour genomic instability scores have been tested in cancer patients and healthy controls. Furthermore, cfDNA genomic instability scores have been used to differentiate cancer cases and controls (178, 265) and may have prognostic value (266). A score may also be useful to quantify and track serial changes in cfDNA and therefore monitor treatment response (168). Scores measuring the number and magnitude of CNAs across the whole genome, chromosomal arms and smaller chromosomal segments have been explored (167).

Two published genomic instability scores were adapted to evaluate whether scores differentiated lung cancer cases and controls. Genomic instability scores were calculated using copy number ratios of 1 Mb windows after low coverage sequencing of chromosomes 1-22 (see Section 2.18).

2.19.1 The Plasma Genomic Abnormality 2 score

The Plasma Genomic Abnormality score (PGA score) quantifies genomic instability by summing the squared $\log_2$ copy number ratio values of the most significant CNAs from whole genome sequencing data (266). To create this score, copy number ratios of 1 Mb windows were calculated as the ratio of observed read count to the mean number of reads from a healthy control group (N=7) for the corresponding window (265). Then copy number ratios were $\log_2$ transformed and normalised for GC content (265). Log_2 copy number ratios were squared and ranked, and the ratios ranked in the 95^{th} to the 99^{th} percentiles were summed (265).

In the study by Xia et al 2015, higher median Plasma Genomic Abnormality (PGA) scores were found for eight patients with early stage (I-IIA) lung adenocarcinoma compared to eight normal controls (19.5 (range 5.9-64.5) vs 9.3 (range 7.4-11.1) p=0.01) (265) Approximately 20 million reads were obtained for each individual, and the genome coverage was reported as 0.53X.

The PGA score was adapted to create the PGA2 score, by ranking squared copy number ratio Z scores and summing the 95th to 99th percentile scores. Both scores convert gains and losses to positive values to allow both extremes of the distribution to be examined. For our samples, copy number ratios were calculated by comparing the number of reads in a 1 Mb window for cfDNA to the number of reads in the corresponding window of matched genomic lymphocyte DNA (see Section 2.18.6). This differs to the PGA score, where cfDNA copy number ratios were calculated by comparing the number of reads in a window to the mean number of reads in the corresponding window of cfDNA samples from a group of healthy controls. For the PGA2 score, the Z score was calculated by subtracting the mean copy number ratio of all 1 Mb windows for the sample from the copy number ratio of the 1 Mb window and dividing by the standard deviation of all copy number ratios across the genome for that sample.

$$\text{Z score of a 1 Mb window} = \frac{copy\ no.ratio\ of\ a\ 1\ Mb\ window - mean\ (copy\ no.ratio\ of\ all\ 1\ Mb\ windows)}{standard\ deviation\ (copy\ no.ratio\ of\ all\ 1\ Mb\ windows)}$$

Then, copy number ratios were $\log_2$ transformed, squared, ranked and the 95th to 99th percentile scores were summed (266). For both the PGA and PGA2 scores, the top 1% of squared $\log_2$ copy number ratios were discarded (266). This is because it was observed that high magnitude CNAs were caused by low quality sequencing libraries or were located near centromeres or telomeres (266). Regions surrounding centromeres and telomeres contain highly conserved and repetitive DNA sequences (267) and therefore it can be difficult to map accurately short reads to these regions leading to false positive results (266).

2.19.2 Copy Number Aberration score

A less selective approach to measure genomic instability is to sum genetic variation across the whole genome. The whole genome summed Z (WGS) score or 'global Z score', sums the

Z scores of the copy number ratios across all 1 Mb windows of the genome (178). A Z score statistic establishes whether CNAs in a cfDNA sample are present when compared to a reference group of controls. This is achieved by calculating for the test sample the number of standard deviations from the mean of the reference plasma samples for each 1 Mb window. Therefore, a Z score is calculated by the following equation:

$$\text{Z score of a 1 Mb window} = \frac{copy\ no.ratio\ sample - mean\ (copy\ no.ratio\ of\ the\ reference\ control\ group)}{standard\ deviation\ (copy\ no.ratio\ of\ the\ reference\ control\ group)}$$

In a study by Heitzer et al 2013, low coverage sequencing (>0.1X) with Illumina MiSeq was carried out for cfDNA samples from prostate cancer patients and healthy controls (178). WGS scores varied from -1.1 to +2.8 for healthy controls (N=10) and 125 to 1156 for prostate cancer cases (N=9). Prostate cancer patients (N=9) were distinguished from healthy controls (N=10) by the WGS score in hierarchical cluster analyses. A significant CNA was defined as being greater than three SDs away from the mean of the copy number ratio of the corresponding 1 Mb window of the control set. In silico analyses simulated different mixtures of prostate cancer (N=102) and normal control DNA (N=500), and showed that the sensitivity for the detection of CNAs was >80% and the specificity >80% when there was 10% circulating tumour DNA (178). When cell line DNA was mixed with normal control DNA at different proportions, samples with 1% cell line DNA remained separate from the healthy control group in hierarchical cluster analysis (178).

In another study, Chan et al 2013 normalised reads in cfDNA samples to matched genomic DNA and then established Z scores by comparing the $\log_2$ copy number ratios of 1 Mb windows to a reference from 16 healthy controls (168). CNAs identified in pre-surgical cfDNA samples for four patients with hepatocellular cancer almost disappeared in their matched post-surgical samples.

The Copy Number Aberration (CNA) score is a whole genome wide assessment score of genomic instability adapted from the Whole Genome Summed Z (WGS)(113). Z scores were calculated for each 1 Mb window by subtracting the mean copy number ratio of the low risk control group (N=10) from the copy number ratio of each 1 Mb window and dividing by the standard deviation of the copy number ratio from the low risk control group.

$$\text{Z score of a 1 Mb window} = \frac{copy\ no.ratio\ sample - mean\ (copy\ no.ratio\ low\ risk\ controls)}{SD\ (copy\ no.ratio\ low\ risk\ controls)}$$

For each low risk control, the mean and standard deviation for all copy number ratios across all 1 Mb windows were calculated. Then, the average of the mean across the low risk controls (N=10) and the standard deviation were taken and these valuses were used to calculate the Z score. CNA scores were not calculated for low risk controls so that they did not serve as their own controls. Z scores were squared and summed to calculate the CNA score.

In comparison, Heitzer et al. normalised the number of reads in a cfDNA sample to the mean number of reads of a heatlhy control group. Z scores were then calculated within each 1 Mb window by subtracting the mean copy number ratio of the control group from the copy number ratio of the sample and dividing by the standard deviation of the control group (N=19) (113). Then, Z scores were squared and summed to create the final score. Neither the PGA score (by Xia et al.) or the WGS score (by Heitzer et al.) normalised the number of reads in a window for cfDNA to the number of reads in a window of matched genomic DNA, therefore all germline aberrations may not have been eliminated.

2.20 Statistical analysis

Statistical analysis were carried out with Microsoft Excel 2016, Graph Pad Prism version 6.0 and StataMP version 12. A p value <0.05 was deemed to be statistically significant. All statistical tests were two-sided.

2.20.1 Comparison of independent groups

In Chapter 3, non-parametric Mann Whitney U tests were used to evaluate methods of DNA extraction by comparing the percentage of DNA recovery for tumour FFPE DNA added in known quantities to plasma. For a comparison of distributions across more than two groups, a non-parametric Kruskal-Wallis rank test was performed. The characteristics of unselected or selected cases and controls were evaluated by comparing the distribution of continuous variables with the Mann Whitney U test. Categorical variables were compared using the chi-squared test, chi-squared test for trend (where categorical variables were ranked eg. disease stage) or Fisher's exact test (if there were less than five people in the defined category). A

non-parametric test for trend was carried out to compare cfDNA levels or genomic instability scores across disease stages in StataMP version 12 in Chapter 3 and 4 respectively. In Chapter 4, DNA fragment sizes and important sequencing parameters were compared across different groups by performing Mann Whitney U tests. This test was also used to compare genomic instability scores between cases and controls.

2.20.2 Correlations between variables and measures of agreement between tests

In Chapter 4, Spearman's Rank Correlation coefficients were calculated to compare copy number ratios and/or copy number ratio segments for different quantities of input cfDNA and different numbers of amplification cycles during library preparation. The Bland-Altman test was used to compare the degree of agreement between old and new library preparation kits as well as the degree of agreement between sequencing runs as a measure of reproducibility. In addition, the coefficient of variance was calculated for H69 cell line DNA CNA scores between sequencing runs as a measure of reproducibility. Spearman's Rank correlation coefficients were calculated for segmental copy number ratios and copy number ratios to compare 100% tumour FFPE DNA with descending proportions of tumour FFPE DNA spiked into cfDNA extracted from the pooled plasma of healthy volunteers. Copy number ratios and segments were also assessed for correlation to compare tumour FFPE DNA and matched cfDNA samples. Pearson's correlation coefficients were used to measure the correlations between $\log_{10}$ cfDNA levels and $\log_{10}$ genomic instability scores.

2.20.3 Evaluation of circulating cell-free DNA levels and genomic instability scores as potential screening tools

As cfDNA levels (Chapter 3) and CNA scores (Chapter 4) were positively skewed, a $\log_{10}$ transformation was applied in an attempt to make the distribution more symmetrical. The analyses described below were carried out in StataMP (version 12).

2.20.3.1 Logistic regression

Logistic regression analyses were used to compare binary outcome variables. Each variable was tested in univariable analysis and included in multivariable analyses if $p \leq 0.25$ or if it was an important *a priori* factor to adjust the analysis for. Robust standard errors were calculated to better estimate variance.

2.20.3.2 Receiver Operating Characteristic curves

The accuracy of a variable in the ability to distinguish between cases and controls and therefore function as a screening tool was summarised by plotting Receiver Operating Characteristic (ROC) curves. The ROC curve is a standard tool for biomarker evaluation and graphs diagnostic sensitivity (true positive rate) against 1-specificity (false positive rate or proportion of controls with a positive test) (268). ROC curves were generated based on the predicted probability of being a case, derived from the logistic regression analysis for univariable and multivariable models. The area under the curve (AUC) gives a measure of test performance and facilitates the comparison of screening tools. The AUC is the average value of sensitivity for all specificity values (268). A test with an AUC of 0.5 is of no use because the proportion of cases and controls with a positive test are equal; AUC values close to 1.0 indicate a test with good discrimination. For different score cut-offs the diagnostic sensitivity was estimated by dividing the number of true positives by the number of true positives and false positives. The diagnostic specificity was estimated by dividing the number of true negatives by the number of true negatives and false negatives.

2.20.4 Evaluation of circulating cell-free DNA levels and genomic instability scores as potential prognostic tools

2.20.4.1 Survival analyses to determine prognostic factors

Kaplan-Meier survival curves were used to compare the outcomes of individuals with scores above or below the median value after adjusting for time from diagnosis to blood sample collection. The test for proportional hazards was performed to ensure that the relative hazard ratio between the two groups was constant over time.

Cox regression survival analyses was carried out to determine the prognostic value of tested variables for all cases after adjusting for time from diagnosis to blood sample collection. The date of study recruitment was the date that the patient was registered for the study on the electronic database and occasionally lagged behind the date the blood sample was collected. First, all variables were tested in univariable analysis. Variables with p value ≤0.25 were included in the final model or a variable was included if it were an important factor to adjust the analysis for. This value was chosen because it was expected that a variable with p value >0.25 would not be predictive of survival in a model with other predictors (269).

For cox regression there is the assumption that the factors comparing different groups are constant over time. The test of proportional hazard assumption was performed for the final model to test that this assumption was correct by showing that all factors had p value >0.05 and therefore did not significantly vary over time.

3 Optimising plasma DNA extraction and evaluating total circulating cell-free DNA levels as a potential screening tool in lung cancer

3.1 Introduction

Circulating cfDNA is a potential biomarker in lung cancer because total cfDNA levels are raised compared to healthy controls, and the genetic changes in the primary tumour are identified in the blood (84). The challenge is to extract maximal amounts of cfDNA whilst minimising contamination with genomic DNA from blood lymphocytes, to enable sensitive detection of tumour-related genetic alterations. This will facilitate the translation of cfDNA as a biomarker in clinical practice by minimising false positive and false negative results. The three steps in processing cfDNA are blood sampling; DNA extraction and analysis (see Section 1.3.1 and 1.3.2).

Many studies have reported the utilisation of total cfDNA levels as a screening or diagnostic tool for lung cancer (see Section 1.5.1.1). However, no study has considered whether cfDNA levels can differentiate between high risk controls selected by a lung cancer risk model and lung cancer cases. Early detection of lung cancer is vital to improve patient outcomes and current screening strategies use risk models to identify those at highest risk of lung cancer (10). The measurement of total cfDNA levels is a non-invasive, cheap test and therefore wide spread clinical use in the NHS would be feasible.

3.2 Aims and objectives

The first aim and objective was

1) To identify the most efficient cfDNA extraction method by comparing the percentage recovery of tumour FFPE DNA after being added to healthy volunteer plasma.

Two cfDNA extraction methods were chosen after a review of the literature. DNA extracted from FFPE tumour specimens was added to healthy volunteer plasma to model the short degraded fragments characteristic of cfDNA. The amount of cfDNA extracted from the plasma was quantified by SYBR green *GAPDH* RT-qPCR and the percentage of recovered tumour DNA calculated to allow method comparison.

The second aim and objective was

1) To evaluate total cfDNA levels quantified by SYBR green RT-qPCR as a potential screening tool for lung cancer by comparing total cfDNA yield for lung cancer cases and high risk controls.

The hypothesis was that lung cancer cases would have higher cfDNA levels than high risk controls and therefore cfDNA levels could aid early lung cancer detection.

A description of the ReSoLuCENT study, followed by a description of the cases selected for cfDNA analysis is presented. Total cfDNA levels of cases and controls were compared and levels were also compared between subgroups including treated and untreated cases and advanced and early stage cases.

Univariable and multivariable logistic regression were performed to establish whether cfDNA levels predicted case or control status. ROC curve analyses were performed to establish sensitivity and specificity of cfDNA levels in distinguishing cases and controls.

3.3 Results

3.3.1 A comparison of plasma DNA extraction methods

3.3.1.1 Recruitment of healthy volunteers

Fourteen healthy volunteers were recruited to the study 'Optimisation of plasma nucleic acids' from April 2013 to March 2016 to provide negative control plasma and pooled plasma for spiking experiments (see Section 2.2.1). The median age of healthy volunteers was 34 years (range 24-38 years), seven (50%) volunteers were male and seven (50%) were female.

3.3.1.2 Tumour DNA yield and quality

DNA was extracted from colorectal tumour FFPE sections (see Section 2.4) to be used for plasma spiking experiments to model the short degraded fragments of cfDNA. Extracted tumour DNA was added in known quantities to pooled healthy volunteer plasma at a proportion of 65 ng per 1 ml unless otherwise stated. Plasma DNA extraction methods were compared by calculating the percentage of tumour DNA recovered as described in Section 2.11.1.

The quantity and purity of extracted tumour DNA from two colorectal cancer cases were determined by nanodrop spectrophotometry. The yield of DNA from these cases were 404 ng/µl and 181 ng/µl in 50 µl. The 260/280nm ratios were 1.99 and 2.03, whilst the 260/230nm ratios were 2.44 and 2.35 respectively. Extracted tumour DNA yielded a PCR product using BRAF V600E short amplicon (160 bp) and BRAF V600E long amplicon (600 bp) primers (Figure 3-1). These findings were consistent with degraded fragmented DNA expected from FFPE tissue sections, and a representative model of the fragmented DNA circulating in the plasma of cancer patients. This DNA was used to spike healthy control plasma.

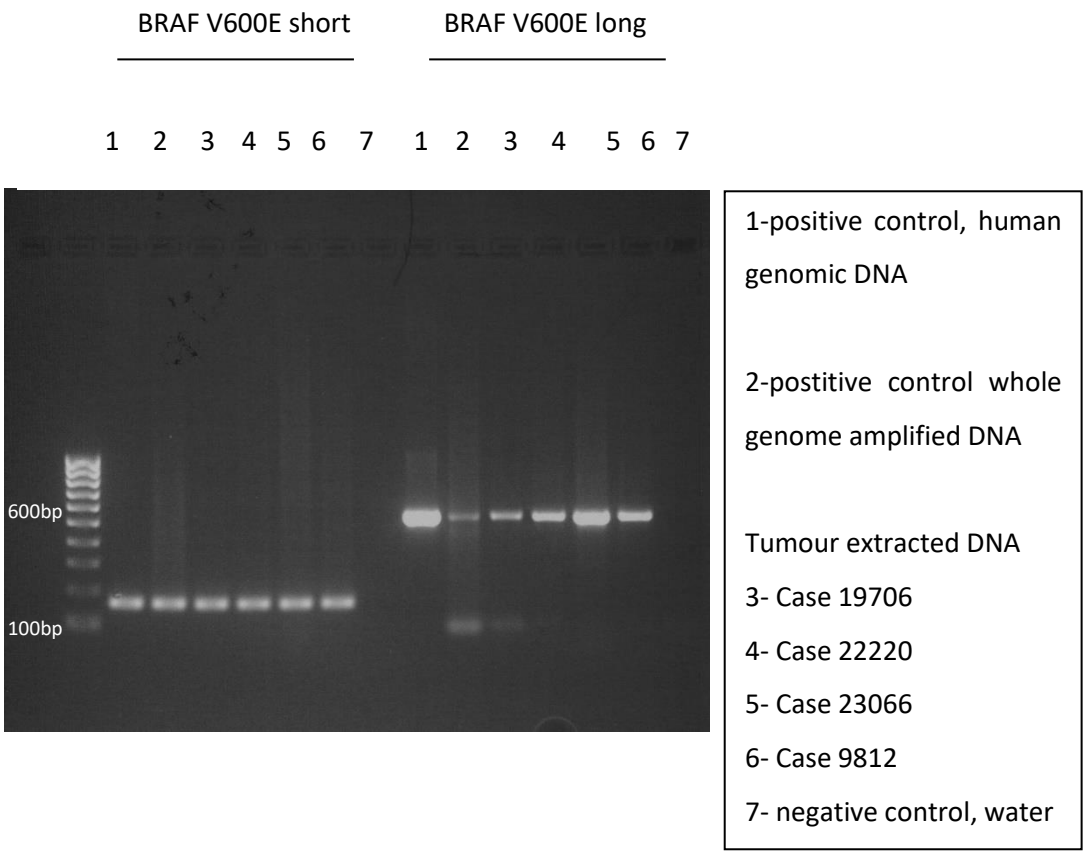


Figure 3-1: Agarose gel electrophoresis of colorectal FFPE tumour DNA extracted from four different cases after PCR for BRAF V600E (short and long exons).

3.3.1.3 QIAamp® Blood Mini Kit compared to the phenol chloroform method

The first aim was to compare two methods of plasma DNA extraction, the QIAamp® blood mini kit (Qiagen) and the phenol chloroform method. Healthy control plasma samples of 1 ml were spiked with 65 ng of tumour FFPE DNA. There was no significant difference in the median percentage of DNA recovered between the QIAamp® blood mini kit (Qiagen) and phenol chloroform method as assessed by three independent experiments (Figure 3-2). DNA yields were very low with both methods. A median of 1.8% (range 1.3- 2.7%) of DNA was recovered from 1 ml of plasma with the QIAamp® blood mini kit (Qiagen) compared to 2.2% (range 2.0- 3.4%) with phenol chloroform (p=0.40 Mann Whitney U test). The phenol chloroform method was time consuming and laborious, therefore the QIAamp® blood mini kit (Qiagen) was further evaluated.

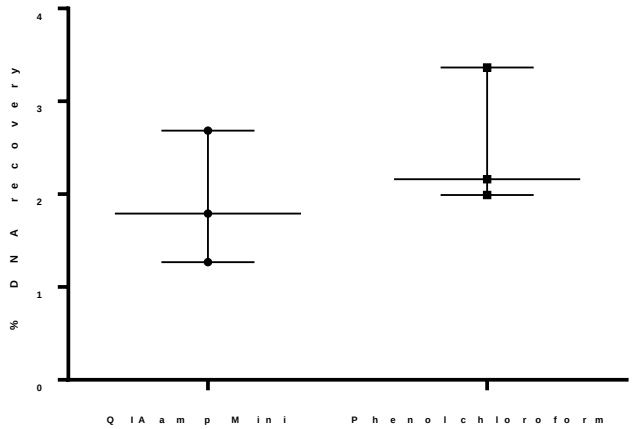


Figure 3-2: The percentage of DNA recovery from 1ml of plasma spiked with 65ng of tumour FFPE DNA for the QIAamp® blood mini kit and phenol chloroform method (N=3)

Median and interquartile range (IQR) shown. Median 1.8% (range 1.3- 2.7%) vs 2.2% (range 2.0- 3.4%) respectively p=0.40 Mann Whitney U test. Each point represents the percentage of DNA recovered from one independent experiment.

3.3.1.4 The QIAamp® circulating nucleic acid kit compared to the QIAamp® blood mini kit

Plasma DNA recovery remained very low with the QIAamp® blood mini kit (Qiagen) therefore the QIAamp® circulating nucleic acid (CNA) kit (Qiagen) designed specifically for the extraction of cell-free nucleic acids was tested. To increase cfDNA yields the quantity of plasma was increased from 1 ml to 3 mls. Plasma samples of 3 mls were spiked with 100 ng of FFPE tumour DNA, and the vacuum rather than spin method was used (see Section 2.5.2.2), to accommodate larger plasma volumes and more efficient sample processing. The median percentage of DNA recovery for the CNA kit (Qiagen) was 20.4% (range 19.1- 20.6%) compared to 7.0% (range 4.6- 9.2%) for the QIAamp® blood mini kit (Qiagen) (p=0.10 Mann Whitney U test) (Figure 3-3).

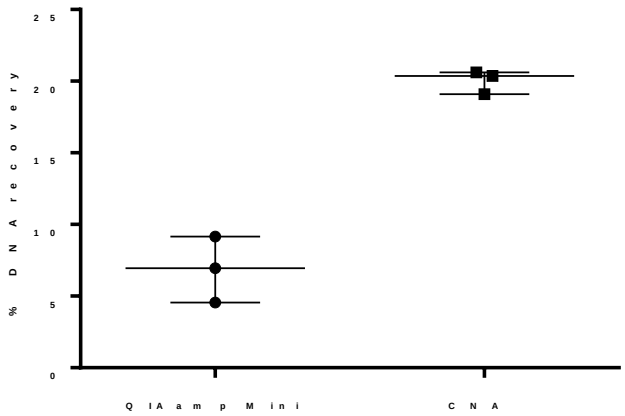


Figure 3-3: The percentage of DNA recovery from 3 mls of plasma spiked with 100 ng of tumour FFPE DNA with the CNA and QIAamp® blood mini kit (N=1).

Median and IQR shown. Each point represents one independent repeat and is the average of three qPCR replicates. Median recovery for CNA kit 20.4% (range 19.1- 20.6%) compared to 7.0% (range 4.6- 9.2%) for the QIAamp® blood mini kit, p=0.10 Mann Whitney U test.

3.3.2 The ReSoLuCENT study

3.3.2.1 Recruitment of participants to the ReSoLuCENT study

There were 887 cases and 538 controls recruited to the multicentre ReSoLuCENT study from 6th April 2006 to 31st August 2016. An intermediate dataset including recruited participants up to the 8th of November 2013 forms the sample set described in this Chapter. Of these 1121 participants, 682 (61%) had a diagnosis of lung cancer and 439 (39%) were related or unrelated controls. Of the cases, two were ineligible, after further pathological review indicated a diagnosis of metastatic thyroid cancer and mesothelioma. Sixty-eight percent of cases and 80% of controls were recruited in South Yorkshire (Sheffield, Doncaster and Rotherham). Figure 3-4 displays recruitment figures for each site.

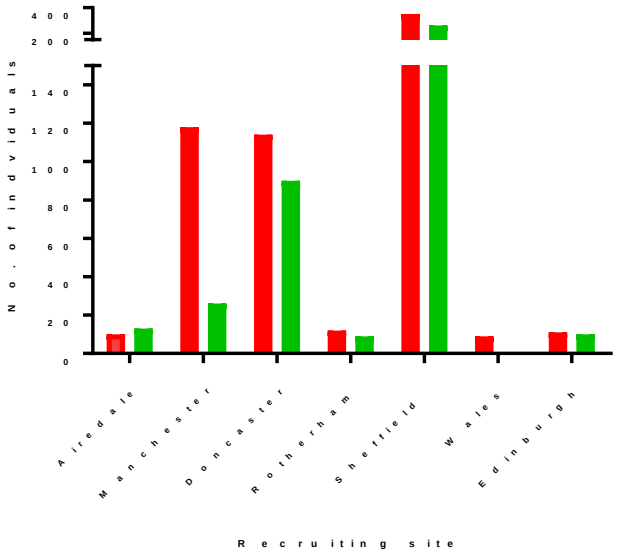


Figure 3-4: The number of cases and controls participating in ReSoLuCENT across recruitment sites.

Red: Cases. Green: Controls.

3.3.2.2 Characteristics of cases and controls recruited

The characteristics of cases (N=680) and controls (N=439) in the ReSoLuCENT study are shown in Table 3-1. The median age of cases at recruitment was significantly older than controls, 56.6 years (range 20.8-83.3 years) compared to 52.1 years (range 19.0-83.2 years), $p<0.0001$ Mann Whitney U test. There were a higher proportion of females in the control group (78% vs 50%) and never smokers (32% vs 7%). Most participants were White British (87%).

Characteristic		Cases (N=680)	Controls (N=439)	P value (statistical test)
Gender	Male	338 (50%)	172 (39%)	p=0.0006 (Chi-squared 2x2)
	Female	342 (50%)	267 (78%)	
Age at recruitment	Median (Range)	56.6 (20.8- 83.3)	52.1 (19.0- 83.2)	p<0.0001 (Mann Whitney U)
Age at diagnosis	Median (Range)	56.2 (20.7-83.1, N=678)	-	-
Ethnicity	White British	585 (86%)	386 (88%)	p=0.24 (White British vs Non White British vs Unknown) (Chi-squared 2x3)
	Black Caribbean	1 (0.1%)	-	
	Black African	1 (0.1%)	-	
	Chinese	3 (0.4%)	-	
	Other	5 (0.7%)	2 (0.5%)	
	Unknown	85 (13%)	51 (12%)	
Smoking Status	Current	187 (28%)	142 (32%)	p<0.0001 (Chi-squared 2x4)
	Ex	353 (52%)	120 (27%)	
	Never	50 (7%)	142 (32%)	
	Unknown	90 (13%)	35 (8%)	
Status as of 31st November 2015	Alive	111 (16%)	422 (96%)	p<0.0001 (Chi-squared 2x2)
	Dead	569 (84%)	17 (4%)	

Table 3-1: The characteristics of cases (N=680) and controls (N=439) participating in ReSoLuCENT.

To be eligible for ReSoLuCENT cases, had to have a confirmed pathological diagnosis of lung cancer. The histological subtype was categorised according to the World Health Organisation 2004 classification of lung tumours (270). Five hundred and three cases (74%) were diagnosed with NSCLC. Of these cases, 204 (41%) had adenocarcinoma (AC), 152 (30%) squamous cell carcinoma (SQ) and for 120 (24%) the histological subtype was not otherwise specified (NOS). In addition, 27 (5%) of NSCLC cases had less common subtypes. These included 10 (2%) cases with neuroendocrine, 5 (1%) with large cell, 2 (0.4%) with bronchcoalveolar, 5 (1%) cases had a combination of NSCLC subtypes diagnosed and there were individual cases of adenocystic (0.2%), spindle variant (0.2%), and pleomorphic/sarcomatoid (0.2%). One hundred and sixty-six (24%) cases were diagnosed with SCLC and 11 (2%) cases had a mixture of SCLC and NSCLC histology.

Lung cancer research is hampered by the poor availability of tumour tissue. In this study, only 29 (4%) of cases had surgically resected tumour tissue available. In contrast, 451 (66%) of cases had a primary bronchial biopsy, 50 (7%) a loco-regional lymph node biopsy, seven (1%) a pleural biopsy, 66 (10%) a biopsy of a secondary metastasis and 72 (11%) had cytology specimens. For five (0.7%) cases the biopsy site was unknown.

At study entry, 295 of 514 (57%) of NSCLC cases had advanced stage IV metastatic disease, 170 (33%) had stage III, 23 (4%) stage II and 6 (1%) stage I disease. In comparison, 118 of 166 (71%) SCLC cases had extensive disease (ED) and 48 (29%) limited disease (LD). Table 3-2 summarises disease stage according to histological subtype.

Stage	NSCLC				MIXED NSCLC AND SCLC	SCLC	Total
	AC	SQ	NOS	OTHER			
Stage I	3	3	-	-	1		7(10%)
Stage II	9	8	5	1	-	-	23 (4%)
Stage III	50	71	38	11	7	-	177 (26%)
Stage IV	137	66	77	15	3	-	298 (44%)
Limited	-	-	-	-	-	48	48 (7%)
Extensive	-	-	-	-	-	118	118 (17%)
Missing	5	4	-	-	-	-	9 (1%)
Total	204 (30%)	152 (22%)	120 (18%)	27 (4%)	11 (2%)	166 (24%)	680

Table 3-2: A summary of the stage and histopathology of lung cancer cases (N=680) recruited to ReSoLuCENT.

AC: adenocarcinoma. NOS: not otherwise specified. NSCLC: non-small cell lung cancer. SCLC: small cell lung cancer. SQ: squamous cell carcinoma.

The ECOG (Eastern Cooperative Oncology Group) performance status (PS) is a measure of an individual's physical activity, and is an important prognostic factor in lung cancer (271). Five hundred and seventy-eight of 680 (85%) of cases had a PS of 0 or 1, indicating the ability to carry out normal activity or light work respectively. On the other hand, 71 of 680 (10%) of cases had a PS of 2, and 19 of 680 (3%) cases had a PS of 3. A PS of 2 indicated that the individual was mobile for more than 50% of the day and a PS of 3 indicated that an individual was mobile for less than 50% of the day, had limited ability to self-care, but was not bed bound.

3.3.2.3 Liverpool Lung Project cancer risk score

As part of another project by Eoin Gray, LLP Risk scores were calculated for both cases and controls, and were available for 781 ReSoLuCENT participants between the ages of 40 to 80 years (Figure 3-5). The LLP model calculates a predicted risk of lung cancer using risk factors such as age; gender and smoking (13)(see Section 1.1.2.1.1). The median risk score for cases (N=521) was significantly higher than the median risk score for controls (N=260), 0.84 (range 0.012-39.36) compared to 0.42 (range 0.005-13.25), $p<0.0001$ Mann Whitney U test. In this subset, there was no difference between cases and controls for age (median 56.6 (range 40-

79 years) vs 56.2 years (range 40-79 years), p=0.86 Mann Whitney U test) or gender (p=0.080 Chi-squared test).

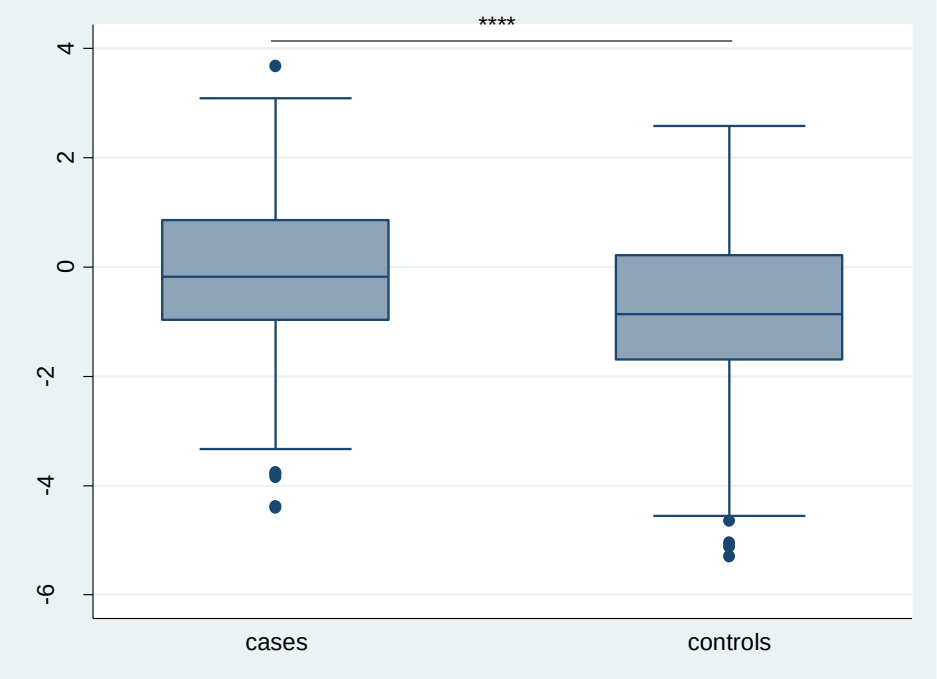


Figure 3-5: A Box plot to compare available Liverpool lung cancer project (LLP) risk scores for lung cancer cases (N=521) and controls (N=260) in ReSoLuCENT.

Median score -0.84 (range -0.012-39.36, N=521) vs -0.42 (range -0.005-13.25, N=260) respectively **** p<0.0001 Mann Whitney U test. Median, IQR, maximm, minimum values and outlier values are shown.

Only 38 controls (15%) had a LLP score ≥2.5% and were thus classed as high risk for the development of lung cancer. Whilst, 110 (21%) of lung cancer cases were classified as high risk. Only 42 cases (8%) and 15 (6%) controls had a LLP score ≥ 5%. Predicted risk is generally low because ReSoLuCENT has recruited young cases and controls and age is a significant risk factor in the LLP risk model.

As of March 31st 2016, 22 controls (5%) were reported by the Health and Social Care Information Centre (HSCIC) (previously The Office of National Statistics) to have developed cancer after study registration (Table 3-3). Only one control was diagnosed with lung cancer. These controls were excluded from molecular analyses and no control analysed subsequently developed cancer to our knowledge. LLP risk scores were available for 18 of 22 controls. The median LLP score was 0.57 (range 0.13-3.79) and only three (17%) controls

had an LLP risk score ≥2.5%. There was no significant difference between the LLP score for controls that were not known to have developed cancer compared to controls that had developed cancer (0.42 (range 0.005-13.25, N=242) vs 0.57 (range 0.13-3.79, N=18), p=0.31 Mann Whitney U test).

Cancer subtype	No. of controls
Basal cell carcinoma	2
Bowel	1
Breast	6
Cervical	4**
Lung	1
Melanoma	3
Prostate	2
Skin undefined	1
Testicular	1
Undefined	1
Total	**22**

Table 3-3: The development of cancer in control subjects in ReSoLuCENT according to HSCIC.

** two controls with cervical cancer subsequently developed breast cancer.

3.3.3 Plasma extracted circulating cell-free DNA total levels of ReSoLuCENT recruits

3.3.3.1 Selection of cases and controls for circulating cell-free DNA analysis

CfDNA was extracted from the plasma of 114 individuals (72 cases and 42 controls) participating in ReSoLuCENT to test the use of cfDNA levels as a screening tool for lung cancer. Levels of cfDNA were quantified by SYBR green *GAPDH* RT-qPCR (see Section 2.9.2.1.1).

Seventy-two of 680 available cases were selected for cfDNA analysis as follows. Plasma from 52 of 54 cases that had not received treatment for cancer prior to blood withdrawal were chosen. One case was eliminated because disease stage was unknown and a further case was not chosen because the histological subtype was rare (adenoid cystic). In addition, plasma from 20 advanced stage treated cases (of 626 treated cases) were processed for which the ctDNA allele fraction had been defined by targeted sequencing with the Ion Torrent Platform (see Appendix E) (N=6) or whereby tumour FFPE tissue was available (N=14).

Plasma from 10 low risk and 32 high risk unrelated controls were chosen from those for whom an LLP risk score was available (N=260). The LLP risk model was used to mimic the selection of high risk controls as carried out in the UKLS lung cancer screening study (see Section 1.1.2.1.1). There were 222 low risk controls with an LLP risk score <2.5 %. There were 38 high risk controls with an LLP risk score ≥2.5%, and 32 were chosen. Low risk and high risk controls were selected to have similar age range and gender as selected early stage cancer cases, although due to small numbers this was more difficult for high risk controls.

Plasma was utilised from blood samples collected from seven different centres participating in the ReSoLuCENT study. Forty-five percent of blood samples were collected from Weston Park Hospital in Sheffield (Figure 3-6). Each site followed ReSoLuCENT SOPs to collect and process blood and to store plasma and genomic DNA. Plasma and genomic DNA samples were couriered, using dry ice to avoid thawing, in batches to Weston Park Hospital for long-term storage.

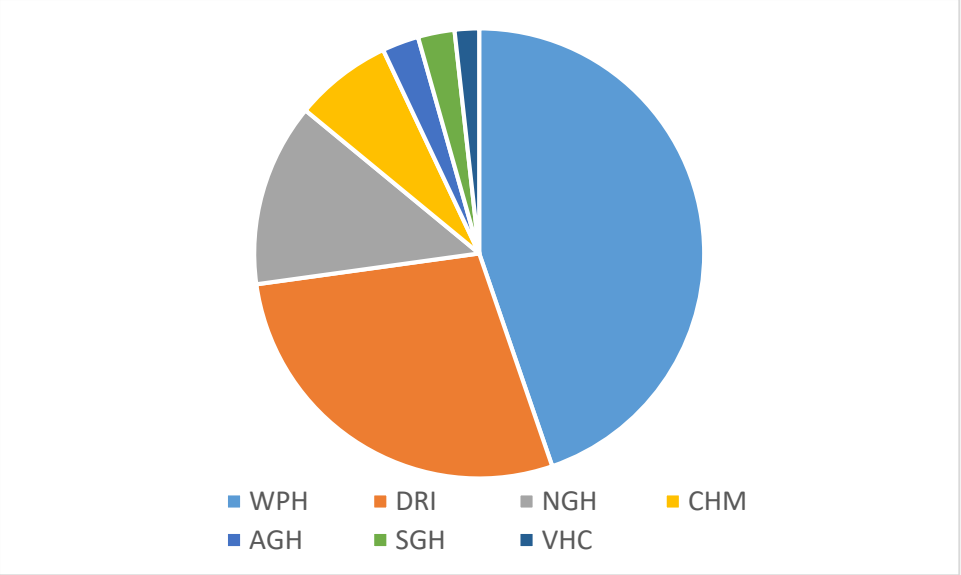


Figure 3-6: A pie chart to demonstrate the proportion of analysed blood samples collected at different centres participating in ReSoLuCENT (N=114)

WPH: Weston Park Hospital, Sheffield (N=51). DRI: Doncaster Royal Infirmary (N=32), Doncaster. NGH: Northern General Hospital, Sheffield (N=15). CHM: Christies Hospital, Manchester (N=8). AGH: Airedale General Hospital (N=3): SGH: Southampton General Hospital (N=3): VHC: Velindre Hospital Cardiff (N=2):

3.3.3.2 Comparison of selected subjects to non-selected subjects

First, it was established that there was no evidence that the cases that were selected for the study were not representative of the whole study. There was no significant difference between the characteristics of the chosen subgroup of cases with cfDNA levels (N=72) compared to the cases whose plasma was not analysed (N= 608) for, age at diagnosis (p=0.11 Mann Whitney U test), gender (p=0.86 Chi-squared test) or smoking status (p=0.49 Chi-squared test for trend). However, due to the presence of selected high risk controls, the control group did differ from those that were not analysed. The median age of the selected control subgroup (N=42) was significantly older than the remaining controls in the ReSoLuCENT study (N=397), 61.1 years (range 34.4-79.2 years) compared to 52.1 years (range 18.9-83.2 years) respectively, p<0.0001 Mann Whitney U test. Furthermore, there was a significantly higher proportion of females and lower proportion of males in the remaining control group (N=397) (62% and 38%) compared to the analysed subset (N=42) (43% and 57%) (p=0.02 Chi-squared test). Smoking proportions were also different between the two groups, with more smokers in the control subgroup (N=42) compared to the remaining controls in the ReSoLuCENT study (N=397) (p=0.003 Chi-squared test for trend).

3.3.3.3 Comparison of all selected cases with controls and analysis of selected subgroups

In the following sections, first the characteristics and cfDNA levels of selected cases (N=72) and controls (N=42) were compared to establish any differences. Then, subgroups of selected cases and controls were compared. Further comparisons of subgroups were carried out with logistic regression and ROC curve analyses to explore the role of cfDNA levels as a potential screening tool.

Subgroups were chosen to enable comparison to the published literature and to determine the role of cfDNA as a screening tool to potentially aid patient stratification in a lung cancer screening programme. The subgroups were, untreated cases (N=52) vs controls (N=40), and untreated early stage cancer cases (N=21) vs high risk controls (N=30). For cases, treatment subgroups were defined by whether anticancer treatment was administered prior to blood withdrawal and disease stage. Cases were subdivided into early stage lung cancer (stage I-IIIA) that is potentially curable with treatment, and advanced stage cancer (stage IIIB-IV), which is not cured by treatment. For controls, LLP risk score and age defined subgroups. Controls aged 50-75 years were included (with the exception of two high risk patients aged >75 years) to mimic the age of participants in the UKLS study (see Section 1.1.2), and were

subdivided by their LLP risk score into low risk (<2.5% risk of developing lung cancer over 5 years) and high risk (≥2.5% risk).

3.3.3.4 Comparison of all selected cases with controls

The characteristics of cases (N=72) and controls (N=42) whose cfDNA levels were determined are summarised in Table 3-4. There was a significant difference between cases (N=72) and controls (N=42) for age (p<0.0001 Mann Whitney U test) and smoking (p<0.0001 Chi-squared test). Not surprisingly, given the selection of high risk controls, controls were significantly older and more were current smokers.

Characteristic		Cases N=72	Controls N=42	P value (statistical test)
Gender	Male	35 (49%)	24 (57%)	p=0.77 (Chi-squared 2x2)
	Female	37 (51%)	18 (43%)	
Age at registration in years	Median (Range)	57.3 (34.6-78.5)	61.2 (34.4-79.2)	p<0.0001 (Mann Whitney U)
Smoking Status	Current	23 (32%)	25 (59%)	p<0.0001 (Chi-squared 2x4)
	Ex	33 (46%)	6 (14%)	
	Never	4 (6%)	11 (26%)	
	Unknown	12 (17%)	0 (0%)	
Length of plasma storage in years		5.7 (1.2-8.9)	6.6 (2.8-9.6)	0.05 (Mann Whitney U)
Status as of August 30th 2016	Alive	10 (14%)	39 (93%)	p<0.0001 (Fisher's exact)
	Dead	62 (86%)	3 (7%)	

Table 3-4: A comparison of the characteristics of cases (N=72) and controls (N=42) that had plasma cfDNA extracted and quantified.

3.3.3.5 Circulating cell-free DNA total levels were higher in cases compared to controls

The median cfDNA level of all cases (treated and untreated) (N=72) was significantly higher than the cfDNA level of controls (N=42) (7.93 ng/ml (range 1.57- 545.10 ng/ml) vs 4.32 ng/ml (range 1.25- 33.99 ng/ml), p<0.0001 Mann Whitney U test) (Figure 3-7). There were four cases with very high cfDNA levels (185 ng/ml, 295 ng/ml, 545 ng/ml and 540 ng/ml). Of these cases, two had a histological diagnosis of SCLC and two had a diagnosis of NSCLC (adenocarcinoma and NOS). All four had stage IV disease with distant spread to at least one metastatic site. One case with cfDNA levels of 540 ng/ml had palliative radiotherapy prior to blood withdrawal, whilst the other three cases had received no treatment. The median time of plasma storage for controls was 6.6 years (range 2.8-9.6 years) compared to 5.7 years (range 1.2-8.9 years) for cases (p=0.054 Mann Whitney U test).

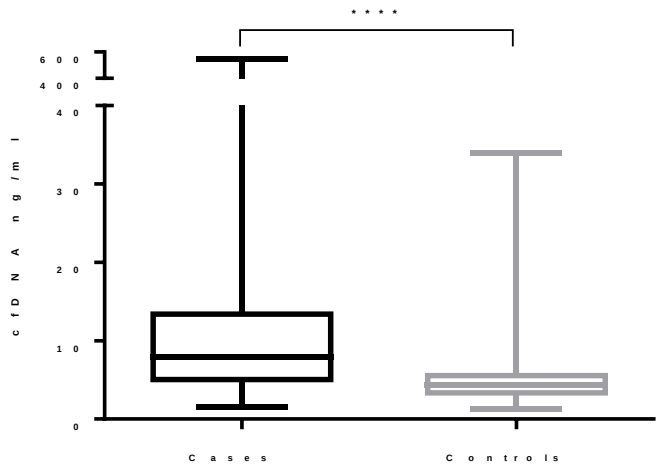


Figure 3-7: A Box plot displaying cfDNA levels in ng/ml for all lung cancer cases (N=72) and controls (N=42).

The median cfDNA level for cases was 7.93 ng/ml (range 1.57- 545.10 ng/ml) vs 4.32 ng/ml (range 1.25- 33.99 ng/ml). **** p<0.0001, Mann Whitney U test. Median, IQR, maximum and minimum values are shown.

3.3.3.6 Treated cases had higher circulating cell-free DNA total levels compared to untreated cases

Anticancer therapy may influence cfDNA levels (129, 272), and therefore the characteristics and cfDNA levels of treated and untreated cases were compared. Twenty lung cancer cases had ongoing or had recently completed cancer treatment prior to their blood withdrawal. Of these cases, eleven had palliative chemotherapy, four palliative radiotherapy, one palliative chemotherapy and radiotherapy, one complete excision of a solitary brain metastasis and no disease elsewhere, one radical surgery with incomplete excision, one radical chemo-radiotherapy and one case was treated with the bisphosphonate zoledronate. A comparison of the characteristics of treated and untreated cases are displayed in Table 3-5. There was no significant difference between the treated and untreated cases for gender (p=0.80 Chi-squared test), age (p=0.85 Mann Whitney U test), stage of disease (p=0.24 Chi-squared test for trend), performance status (p=0.88 Chi-squared test for trend), smoking status (p=0.70 Chi-squared test) and length of time that plasma was stored (p=0.23 Mann Whitney U test). There were higher cfDNA levels in the treated cases compared to the untreated cases, 11.2 ng/ml (range 2.4-540.1) vs 6.9 ng/ml (range 1.6-545.1) (p=0.05 Mann Whitney U test).

Characteristic		Untreated N=52	Treated N=20	P value (statistical test)
Gender	Male	26 (50%)	9 (45%)	p=0.80 (Chi-squared 2x2)
	Female	26 (50%)	11 (55%)	
Age in years	Median (Range)	57.4 (34.6-70.1)	57.05 (37.8-78.5)	p=0.85 (Mann Whitney U)
Stage	I	5 (10%)	0 (0%)	p=0.06 (Chi-squared test for trend, unknown excluded)
	II	9 (17%)	1 (5%) (incomplete excision)	
	III	14 (27%)	6 (30%)	
	IV	24 (46%)	12 (60%)	
	Unknown	0	1 (5%)	
Performance Status (PS)	0	16 (31%)	4 (20%)	p=0.88 (Chi-squared test for trend, unknown excluded)
	1	25 (48%)	13 (65%)	
	2	6 (12%)	3 (15%)	
	3	3 (6%)	0 (0%)	
	Unknown	2 (4%)	0 (0%)	
Smoking Status	Current	17 (33%)	6 (12%)	p=0.70 (Chi-squared 2x4)
	Ex	25 (48%)	8 (15%)	
	Never	3 (6%)	1 (5%)	
	Unknown	7 (13%)	5 (25%)	
Length of plasma storage in years	Median (Range)	5.5 (1.2-8.9)	6.0 (3.5-8.4)	p=0.23 (Mann Whitney U)
CfDNA levels ng/ml	Median (Range)	6.9 (1.6-545.1),	11.2 (2.4-540.1)	p=0.05 (Mann Whitney U)
Status as of August 30th 2016	Alive	12 (23%)	0 (0%)	p= 0.03 (Chi-squared 2x2)
	Dead	40 (77%)	20 (100%)	

Table 3-5: A comparison of the characteristics of treated (N=20) and untreated cases (N=52).

3.3.3.7 Circulating cell-free DNA total levels in untreated lung cancer cases and controls

Given the results above, suggesting that treatment is associated with increased cfDNA levels, treated cases were excluded from further analyses to avoid bias in results due to treatment effect.

3.3.3.7.1 Circulating cell-free DNA total levels increased with advancing disease stage in untreated cases

In our selected subset, it was established whether cfDNA levels increased with advancing disease stage. For the 52 untreated lung cancer cases, cfDNA levels differed between early stage (I-IIIA, N=21) and advanced stage disease (IIIB-IV, N=31) (median 4.74 ng/ml (range 1.57-21.61 ng/ml) vs 10.34 ng/ml (range 2.39-545.1 ng/ml), p=0.0009 Mann Whitney U test). There was a significant trend of increasing cfDNA levels with increasing disease stage

(p=0.001 Non-parametric test for trend) (Figure 3-8). There was a wide distribution of cfDNA levels for cases with stage IV disease from 2.34ng/ml to 545.1ng/ml. Figure 3-9 displays cfDNA levels for untreated stage IV cases (N=24) according to histological subtype. Median cfDNA levels were higher for SCLC cases in comparison to cases with adenocarcinoma (p=0.019), squamous (p=0.024) or other NSCLC subtypes (p=0.026 Mann Whitney U test).

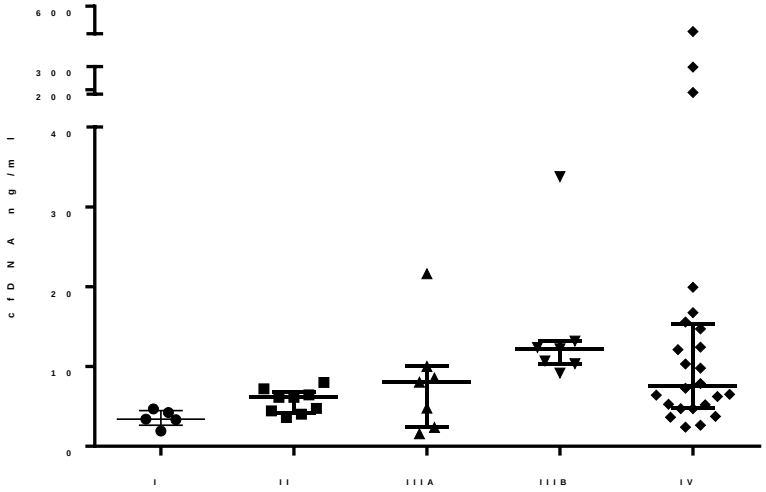


Figure 3-8: CfDNA levels according to the disease stage of untreated cases (N=52).

Median and IQR are shown. Non-parametric test for trend, p=0.001.

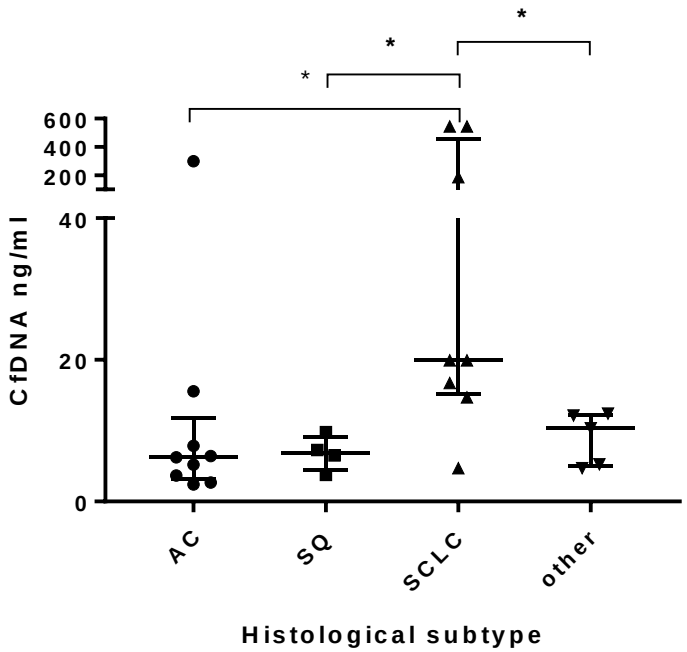


Figure 3-9: CfDNA levels according to the lung cancer histological subtype of stage IV untreated cases (N=24).

Median and IQR are shown. * P<0.05.

3.3.3.7.2 Circulating cell-free DNA total levels were higher for untreated lung cancer cases compared to combined high and low risk controls

Next, comparisons of untreated lung cancer cases (N=52) and controls aged 50 to 75 years old (N=40) (to match the age of participants screened in the UKLS lung cancer screening study), were explored. CfDNA levels of untreated cases (median 6.86 ng/ml (range 1.57-545.15 ng/ml), N=52) were significantly higher than combined high and low risk controls (median 4.43 ng/ml (range 1.25-33.99 ng/ml), N=40) p=0.0021 Mann Whitney U test).

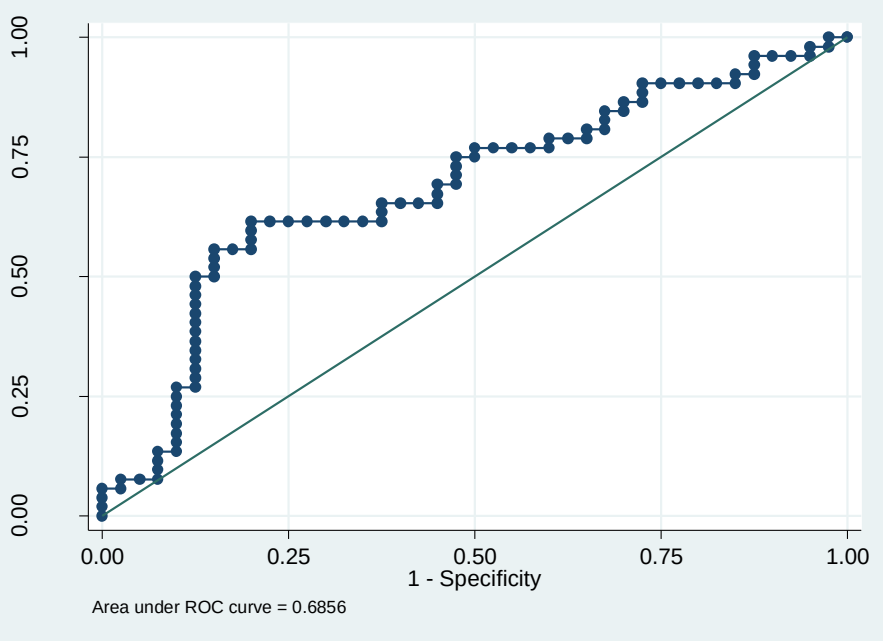
3.3.3.8 Log_{10} cell-free DNA levels were significant predictors of case or control status in univariable and multivariable analysis when comparing untreated cases and controls

A logistic regression model was used to understand whether cfDNA levels predicted case (N=52) or control status (N=40) (see Section 2.20.3.1). As cfDNA levels were positively skewed, a log_{10} transformation was applied. Consistent with our preliminary analysis, for univariable analyses the factors log_{10} cfDNA levels (Odds Ratio (OR) 2.25, p=0.02), years of plasma storage (OR 0.78, p=0.02), age (OR 0.84, p=0.002), and smoking status (never vs ex/current) (OR 4.67, p=0.03) were significant, but gender (OR 1.5, p=0.34) was not. In multivariable analyses, log_{10} cfDNA levels (OR 3.11, p=0.008), smoking status (never vs ex/current) (OR 17.45, p=0.02), age (OR 0.78, p<0.001) and gender (OR 4.14, p=0.02) were significant factors when adjusting for years of plasma storage (Table 3-6).

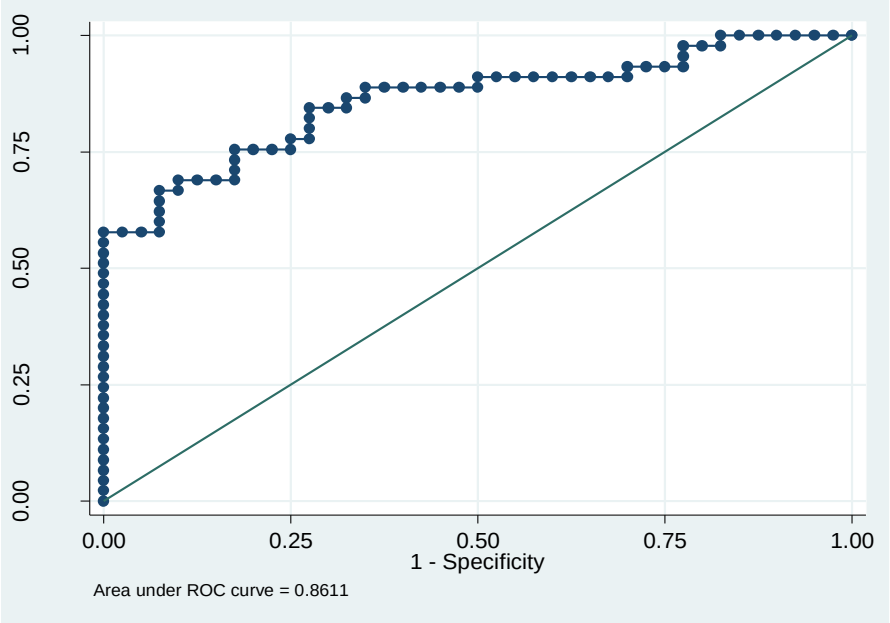
	Odds ratio (95% CI)	**P value**
Univariable analysis		
Log_{10} cfDNA ng/ml	2.25 (1.14-4.44)	0.02
Smoking (never vs ex/current)	4.67 (1.17-18.56)	0.03
Length of plasma storage at -80°C	0.78 (0.64-0.96)	0.02
Age at study registration	0.84 (0.75-0.94)	0.002
Gender (male comparator)	1.5 (0.65-3.47)	0.34
Multivariable analysis		
Log_{10} cfDNA ng/ml	3.11 (1.34-7.21)	0.008
Smoking (never vs ex/current)	17.45 (2.44-124.90)	0.02
Length of plasma storage at -80°C	0.95 (0.69-1.30)	0.76
Age at study registration	0.78 (0.68-0.90)	0.001
Gender (male comparator)	4.14 (1.25-13.67)	0.02

Table 3-6: Univariable and multivariable logistic regression for untreated lung cancer cases (N=52) compared to controls (N=40) to determine significant predictors of case or control status.

ROC curve analysis based on the logistic regression predicted probabilities was carried out to establish potential evidence for the suitability of log_{10} cfDNA levels to discriminate the different groups (Figure 3-10). ROC curves were generated for log_{10} cfDNA levels alone and after adjusting for the predicted probabilities of important variables from multivariable logistic regression (see above). There was fair discriminatory ability for log_{10} cfDNA levels alone (AUC 0.69 (95% CI 0.58-0.78)) and very good discriminative ability after adjustment for smoking, length of plasma storage, age and gender (AUC 0.86 (95% CI 0.78-0.94).



i. Univariable model log_{10} cfDNA levels (AUC 0.69 (95% CI 0.58-0.78))



ii. Multivariable model including age, gender, log_{10} cfdna levels, length of plasma storage and smoking (AUC 0.86 (95% CI 0.78-0.94).

Figure 3-10: ROC analyses for univariable and multivariable models for untreated lung cancer cases (N=52) and controls (N=40) to establish the role of log_{10}cfDNA levels in predicting case or control status

3.3.3.9 *Circulating cell-free DNA total levels for untreated early cancer cases and high risk controls were not significantly different and levels did not predict case or control status*

To be useful as a screening tool, cfDNA levels need to distinguish early stage cancer from high risk controls. However, there was no significant difference in the levels of cfDNA between untreated early stage cancer (stage I-IIIA) (N=21) and high risk controls (N=30) (median 4.74 ng/ml (range 1.57-21.61 ng/ml) vs 4.63 ng/ml (range 1.25-33.99 ng/ml) respectively, p=0.73 Mann Whitney U test) (Figure 3-11). As noted above for all cases the high risk controls (N=30) were older with a higher proportion of current smokers compared to untreated early stage cases (N=21) since they have been selected for high risk LLP risk model scores (Table 3-7).

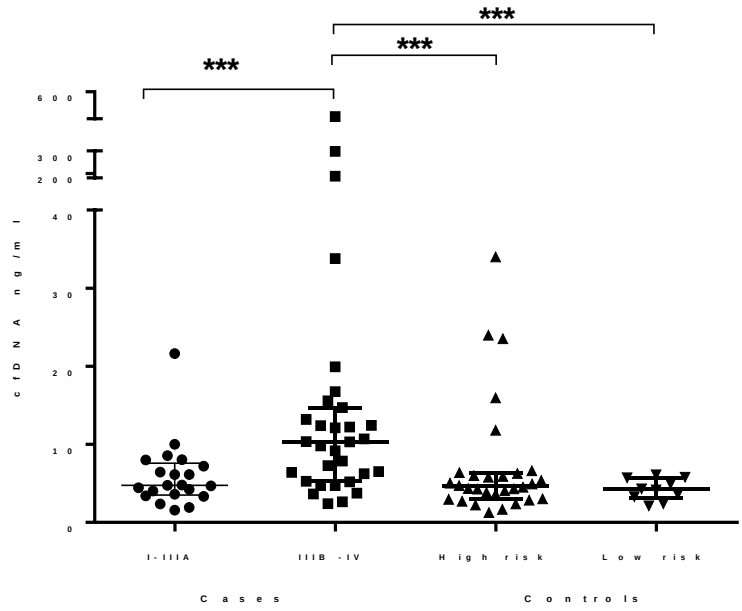


Figure 3-11: A comparison of cfDNA levels ng/ml between early (I-IIIA) (N=21) and late stage (IIIB-IV) (N=31) untreated lung cancer cases and high risk (N=30) and low risk controls (N=10).

*** p<0.001 Mann Whitney U test. Median and IQR shown.

Characteristic		Untreated early stage cancer (I-IIIA) (N=21)	High risk (N= 30)	P value (statistical test)
Gender	Male	9 (43%)	21 (70%)	p=0.08 (Fisher's exact)
	Female	12 (57%)	9 (30%)	
Age in years at registration	Median (Range)	56.3 (40.0-65.2)	61.9 (57.1-72.6)	p<0.0001 (Mann Whitney U)
Age at diagnosis	Median (Range)	56.2 (40.0-65.2)	-	-
Smoking status	Current	6 (29%)	24 (80%)	p=0.002 (Chi-squared 2x2)
	Ex	11 (52%)	6 (20%)	
	Never	0	0	
	Unknown	0	0	
Status as of August 30th 2016	Dead	11 (52%)	1 (3%)	p<0.0001 (Fisher's exact)
	Alive	10 (48%)	29 (97%)	
Length of plasma storage in yrs		6.0 (1.2-8.3)	6.4 (2.8-9.6)	p=0.27 (Mann Whitney U)
CfDNA ng/ml	Median (Range)	4.7 (1.6-21.6)	4.6 (1.2-34.0)	p=0.73 (Mann Whitney U)

Table 3-7: A comparison of the characteristics of early stage lung cancer cases (I-IIIA) (N=21) and high risk controls (N=30).

To determine whether certain factors were predictive of early case vs high risk control status, logistic regression was performed. Smoking was not tested because in this subset all early cancer cases and high risk controls had a smoking history. Log_{10} cfDNA levels were not a statistically significant predictor of case or control status in univariable analysis, neither was length of plasma storage or gender (Table 3-8).

	Odds ratio (95% CI)	**P value**
Univariable analysis		
Log_{10} cfDNA ng/ml	0.95 (0.43-2.09)	0.91
Length of plasma storage at -80°C	0.82 (0.65-1.04)	0.11
Age at study registration	0.60 (0.43-0.84)	0.003
Gender (male comparator)	3.11 (0.96-10.09)	0.06

Table 3-8: Univariable logistic regression for untreated early cancer cases (N=21) and high risk controls (N=30) to evaluate predictive factors to determine case or control status.

3.4 Discussion

3.4.1 Extraction of circulating cell-free DNA from plasma

In this study, three methods of plasma cfDNA extraction were evaluated. There was no significant difference in the mean percentage of DNA recovered from 1 ml of healthy volunteer plasma spiked with tumour DNA for the QIAamp® blood mini kit (Qiagen) compared to the phenol chloroform method. However, the CNA kit (Qiagen) outperformed the QIAamp® blood mini kit (Qiagen) with a higher percentage recovery of DNA from 3 mls of plasma (20.4% vs 7.0%). This finding was consistent with the results of other studies, which are discussed in detail below (135, 273). It was hypothesised that the QIAamp® blood mini kit (Qiagen) was developed for whole blood, and therefore optimised for the extraction of intact higher molecular weight DNA, compared to the CNA kit (Qiagen) developed specifically for short DNA fragments (273).

Tumour DNA from FFPE sections was chosen to spike healthy volunteer plasma as a model for cfDNA in cancer patients. This model was representative because the short degraded tumour DNA fragments were consistent with the characteristics of cfDNA in cancer patients (see Section 1.2.3.1). However, the recovery of tumour FFPE DNA spiked into pooled healthy volunteer plasma was low in this study. Although, DNA recovery improved with larger volumes of plasma. The percentage of DNA recovery with the QIAamp® blood mini kit (Qiagen) from 3 mls of plasma was 7.0% (range 4.6%-9.2%) compared to 1.8% (1.3-2.7%) from 1 ml. Higher plasma volumes result in greater cfDNA yield (273). A linear increase in cfDNA yield was demonstrated up to 3 mls when DNA was extracted with the CNA kit (Qiagen) (273). In order to increase DNA yields further a higher volume of plasma could be used, the CNA kit (Qiagen) is validated for up to 5 mls of plasma.

The recovery of DNA can vary according to the type of DNA added to plasma. Higher levels of DNA have been recovered from plasma spiked with alternative DNA sources (135, 140, 273). In one study, 1 ml of commercial pooled plasma was spiked with λ/*Hind*III DNA in a tenfold serial dilution from 50 ng/ml to 0.05 ng/ml (135) and two fragment sizes were quantified by Taqman PCR. The QIAamp® blood mini kit (Qiagen) and CNA kit (Qiagen) had similar but broad range of percentage recovery for the 23kb fragment (20-70%) but the CNA kit had a higher percentage of recovery of the 564 bp fragment (90-100%). In another study that compared the CNA kit (Qiagen) with the QIAamp® blood mini kit (Qiagen), the CNA kit (Qiagen) consistently demonstrated higher yields and a higher percentage of extracted

lineralised ADH plasmid fragments (>80%) sized 115 bp, 461 bp and 1448 bp (273). The very low recovery achieved may be due to the spiking material tumour FFPE DNA. Tumour DNA may be of poor quality due to damage by formalin fixation and paraffin embedment (274). Commercial reference standards of sheared engineered human cell line DNA are available to add to plasma in different fractions and can also be used to assess the sensitivity/specificity and lower limit of detection of cfDNA genetic profiling (275).

Furthermore, the percentage of DNA recovery reported will be affected by the method of quantification. Amplifiable DNA fragments of plasma-extracted cfDNA were quantified by PCR amplification of *GAPDH* (see Section 2.9.2.1). Tumour DNA was quantified by nanodrop spectrophotometry (see Section 2.10.1.2). Tumour DNA samples had high 230/260nm ratios suggesting the presence of RNA. The nanodrop lacks specificity and is likely to have overestimated the tumour DNA concentration by measuring all nucleic acids. Thus, the percentage recovery are likely to be underestimated. Extracted plasma DNA was quantified by measuring *GAPDH* a housekeeping gene of length 81 bp with SYBR green RT-qPCR. One study used SYBR green RT-qPCR to quantify seven different housekeeping genes and showed that the chosen gene influenced cfDNA yields and suggested averaging the obtained yields to increase accuracy (273). However, this increases cost.

The CNA kit (Qiagen) has surpassed the QIAamp® blood mini kit (Qiagen) as the most common method of plasma DNA extraction (135, 273). Both the QIAamp® blood mini kit (Qiagen) and CNA kit (Qiagen) were easy to use and plasma DNA extraction was completed in less than three hours. In comparison, the phenol chloroform method was labour intensive, taking two days, and involved toxic compounds. Furthermore, the phenol chloroform method had steps that required greater operator skill and therefore there was a higher risk of poor reproducibility. In contrast, the QIAamp® blood mini and CNA kit (Qiagen) have the potential for automation, an important factor when considering standardisation and implementation in a large number of laboratories. In our study, the CNA kit (Qiagen) gave the best yield and recovered 20% of spiked tumour DNA. For these reasons, the CNA kit (Qiagen) is now the preferred method of plasma DNA extraction in our laboratory.

Accurate quantification of cfDNA is important because an insufficient amount of cfDNA can lead to assay failure and inability to identify cfDNA genomic mutations. To overcome this limitation the whole DNA genome can be amplified to increase the amount of DNA prior to

analysis (276). Whole genome amplification (WGA) was required to obtain adequate quantities of cfDNA (2.5µg) for array comparative genomic hybridisation to detect cfDNA CNAs in prostate cancer cases (176). However, suitable controls must be utilised to determine any aberrations that can be introduced by amplification bias or errors (276, 277) and amplification sensitivity and efficacy needs to be evaluated after downstream processing of the WGA product. Negative controls must be included to assess for contamination to avoid false positive results. Contaminating amplified DNA is notoriously difficult to eliminate and can hamper the purity of further amplification reactions. The use of WGA was avoided for these reasons.

Only 4% of recruited cases in the ReSoLuCENT study had surgically resected specimens available. This highlights the importance of developing a non-invasive test or 'liquid biopsy' to greater understand the development and progression of lung cancer in order to improve patient outcomes. CfDNA levels have been examined in this regard in many studies.

3.4.2 Circulating cell-free DNA total levels as a screening tool in lung cancer

Similar to other studies, it was found that cfDNA levels were raised in lung cancer cases compared to controls and increased in late stage lung cancer compared to early stage (202). However, to be useful as a screening tool a marker needs to differentiate between early lung cancer cases and high risk controls. In this study, cfDNA levels were not significantly different between early stage (I-IIIA) lung cancer cases and high risk controls defined by an LLP score ≥2.5%. There was a substantial overlap in the distribution of cfDNA levels leading to poor discrimination.

The poor discriminatory ability of cfDNA levels in this study contrasts with reported AUC values in the literature ranging from 0.63 to 0.94 (see Table 1-3). In addition, the median cfDNA level for lung cancer cases with advanced disease was lower than other reported studies (Table 1-3). The range of values reported could be because of subject differences such as cancer stage and co-morbidities or due to variation in study methods (278). Not all studies match patients to controls for age, sex, co-morbidities and smoking history, which can introduce bias and limits comparisons between studies. Furthermore, the many pre-analytical and analytical factors that can effect cfDNA yield make comparisons between studies difficult (158). The handling and processing of blood can influence cfDNA levels due to the unwanted release of genomic DNA from lysed white blood cells (134, 140). In addition,

other methodological factors such as the method of cfDNA extraction and quantification can also effect cfDNA levels (273)(see Section 1.3.1). For these reasons, it is important that studies report method details to include the type of blood collection tube, time between processing and withdrawal of blood, storage conditions and the speed and number of blood spins to obtain plasma.

ReSoLuCENT was a multi-centre study and samples from seven different centres were utilised in this study. Each centre followed the same strict standard operating procedures for collecting, processing and storing samples; however, it is not possible to eliminate bias introduced by intercentre variability.

It has been suggested that biomarker levels could add further information about the risk of indeterminate lung nodules and therefore reduce over-investigation and over-diagnosis (202, 279). However, raised cfDNA levels are not specific to cancer. Raised cfDNA levels have been found in many other conditions including sepsis (105), inflammatory conditions (103), myocardial infarction (280), obstructive sleep apnoea (104) and even after exercise (102). Alternative tumour specific associated genetic changes may have greater ability to discriminate between early lung cancer and disease free cases (281). Furthermore, cfDNA levels may not be helpful in detecting slow growing tumours which are more likely to be detected by CT screening (200).

Interestingly, higher cfDNA levels were found for treated lung cancer cases compared to untreated cases. Eighteen of 20 (90%) treated cases had disease stage III or IV cancer whilst 38 of 52 (73%) of untreated cases had disease stage III or IV. However, date of the last treatment is not known neither is information regarding disease response to treatment or pre-treatment cfDNA levels.

In this study, higher cfDNA levels were found in cases with advanced stage disease compared to early stage disease. CfDNA levels have previously been shown to correlate with advanced tumour stage, LDH (138) and age (152) but no consistent correlation has been found relating cfDNA levels with stage, histology, age, smoking status, sex (152, 157, 202, 204, 205, 208), number of metastatic sites, performance status (282) or pulmonary inflammatory conditions (204). On the other hand, more specific to tumour cell turnover ctDNA levels do correlate with disease stage in a number of different cancer subtypes to include lung cancer (161).

Most studies now focus on ctDNA levels due to enhanced specificity compared to total cfDNA levels.

3.4.3 Liverpool Lung Project Cancer Risk model

The LLP risk model utilises age, gender, smoking duration, family history of lung cancer, previous history of pneumonia, previous diagnosis of cancer and history of asbestos exposure and has been validated in large National and International studies (13). Interestingly, an LLP score ≥2.5% correctly identified just 21% of our cases and incorrectly identified 15% of controls. In a large validation study, a score of ≥2.5% correctly identified 67% of lung cancer cases and incorrectly identified 33.4% of controls (248). In that study, the mean age of cases was older compared to our study, 66.4 years (SD ±9.1) compared to 55.3 years (SD ±5.5) respectively. In contrast, the mean age of controls in our study was younger, 60.3 years (SD ±9.0) compared to 63.0 years (SD ± 4.3). A blood biomarker may compliment current screening strategies to maximise the likelihood of detecting cancer in younger patients or those with a strong family history of lung cancer.

3.5 Summary and Conclusion

Pre-analytical and analytical methods of plasma DNA extraction can affect the quantity of DNA obtained from blood samples (134, 149). The standardisation and validation of a method of plasma DNA extraction is a vital step towards establishing the use of cfDNA as a potential cancer biomarker. The QIAamp® Circulating Nucleic acid (CNA) kit (Qiagen) was tested with the standard operating procedure provided by Professor Shaw, University of Leicester and this method was validated in our laboratory as the optimal method of plasma DNA extraction. Improvements in cfDNA yield, even if small, will enhance the utilisation of cfDNA as a potential clinical biomarker by increasing the sensitivity of downstream analysis and reducing the chance of assay failure.

Screening tools must be discriminating, reproducible and robust. These data suggest that total cfDNA levels do not discriminate early lung cancer cases from high risk controls. Due to lack of specificity and standardised methods of quantification cfDNA levels are not recommended as a screening tool in lung cancer either alone or as part of a CT screening programme. However, they may be useful in combination with other markers as a screening tool.

4 Low coverage sequencing to identify copy number aberrations in cell-free DNA

4.1 Introduction

The non-invasive detection of somatic genetic alterations in cfDNA has the potential to differentiate between lung cancer cases and controls and therefore aid early lung cancer detection. NSCLC and SCLC have a high number of genetic alterations relative to other cancer subtypes reflecting greater genetic diversity (283) and CNAs are commonly identified in both subtypes (see Section 1.1.5.2.1). Therefore, a genomic instability score based on the number and magnitude of CNAs may aid the molecular stratification of individuals in a lung cancer-screening programme by detecting tumour-derived CNAs in cfDNA samples.

Sequencing to a low read depth across the whole genome is called low coverage sequencing, and reduces cost because a higher number of samples can be multiplexed and sequenced together in one sequencing run (for a definition of coverage see Section 2.18.5). CNAs were identified by the read depth method from just 5 ng of tumour FFPE DNA with low coverage sequencing (X0.1 coverage), and an approximate test cost of £70 per sample (262). Tumour FFPE DNA is degraded and consists of DNA fragments that are short and of similar length to cfDNA fragments, therefore this approach may be suitable for the detection of CNAs in cfDNA.

In this Chapter, two published geomic instability scores, namely the Plasma Genomic Abnormality (PGA) score (265) and the Whole Genome Summed Z (WGS) score (178) were adapted (see Section 2.19) and tested. The PGA was chosen because after sequencing cfDNA at low coverage (0.53X), the score differentiated between early lung cancer cases (N=8) and normal controls (N=8) but the WGS score did not (265). Nevertheless, the WGS score performed well in a study that differentiated prostate cancer cases from healthy controls (178). With the WGS score, genome instability is measured across the whole genome rather than focusing on measuring aberrations with the highest copy number ratio. This is a less selective approach to be able to capture both large amplitude aberrations and large numbers of small aberrations.

4.2 Aims and Hypotheses

This Chapter describes the work carried out to optimise methods for the detection of CNAs by low coverage whole genome sequencing with Illumina HiSeq 2500. A commercial kit was used to prepare DNA for sequencing from DNA samples collected in the ReSoLuCENT study. It was hypothesised that more tumour derived CNAs would be present in cfDNA samples of lung cancer cases compared to high risk controls and therefore quantifying the number and magnitude of CNAs by a genomic instability score may aid lung cancer detection. Higher genomic instability scores would be expected in lung cancer cases compared to high risk controls. In addition, it was hypothesised that a higher genomic instability score would be predictive of a shorter survival time for lung cancer cases.

The aims and objectives were

- To evaluate the analytical performance of low coverage whole genome sequencing in detecting tumour-derived CNAs in cfDNA samples and to validate detection of CNAs.
    a. To optimise DNA library preparation for cfDNA samples by comparing library quantities and detection of copy number ratios with different input amounts of cfDNA ng/ml and PCR cycles as well as two different PCR mastermixes.
    b. To determine the lower limit of detection for identifying CNAs by low coverage whole genome sequencing of cfDNA samples by adding tumour FFPE DNA in known quantities to extracted cfDNA from the pooled plasma of healthy volunteers.
    c. To determine test reproducibility across sequencing runs by comparing the detection of copy number ratios in cell line DNA.
    d. To determine the lower limit of detection for the CNA score by adding tumour FFPE DNA and H69 cell line DNA in known quantities to extracted cfDNA from the pooled plasma of healthy volunteers.
    e. To determine the reproducibility of the CNA score across sequencing runs by comparing the detection of copy number ratios in cell line DNA.
    f. To describe quality control steps that ensured DNA libraries were of good quality and sequencing runs were optimal.
    g. To demonstrate the identification of tumour-derived CNAs in cfDNA samples of lung cancer cases collected in the ReSoLuCENT study by low coverage whole genome sequencing. The objectives were to compare cfDNA CNAs to

those detected in matched tumour FFPE DNA. In addition, to compare cfDNA CNAs to CNAs known to be common to the three main subtypes of lung cancer.

- To evaluate the clinical validity of low coverage whole genome sequencing of cfDNA samples to calculate genomic instability scores based on the identification of CNAs in selected lung cancer cases and controls recruited in ReSoLuCENT.
    a. To explore the sreeening value of two genomic instability scores based on the number and magnitude of CNAs identified in cfDNA samples. The two tested scores were the PGA2 score (see Section 2.19.1) and the CNA score (see Section 2.19.2). The objectives were to compare scores between selected lung cancer cases and high risk controls and to perform logistic regression and ROC curve analyses to establish preliminary evidence for discriminatory ability.

- To assess the relationship between genomic instability score and survival to assess potential as a prognostic tool by calculating Kaplan Meier survival curves to compare the outcomes of individuals with scores above or below the median value and cox regression survival analyses to determine prognostic value of the score alone and in combination with other variables.

4.3 Results

4.3.1 Analytical performance and validation

4.3.1.1 Optimising DNA library preparation for low coverage sequencing

4.3.1.1.1 Optimising library preparation for cell-free DNA samples

CfDNA samples from ReSoLuCENT cases 1518 and 1106 were used for optimisation studies because multiple CNAs were detected and cfDNA yield was high. The input amounts of cfDNA, and numbers of amplification cycles during library preparation were optimised to minimise DNA loss, and to maximise library quantities available for sequencing.

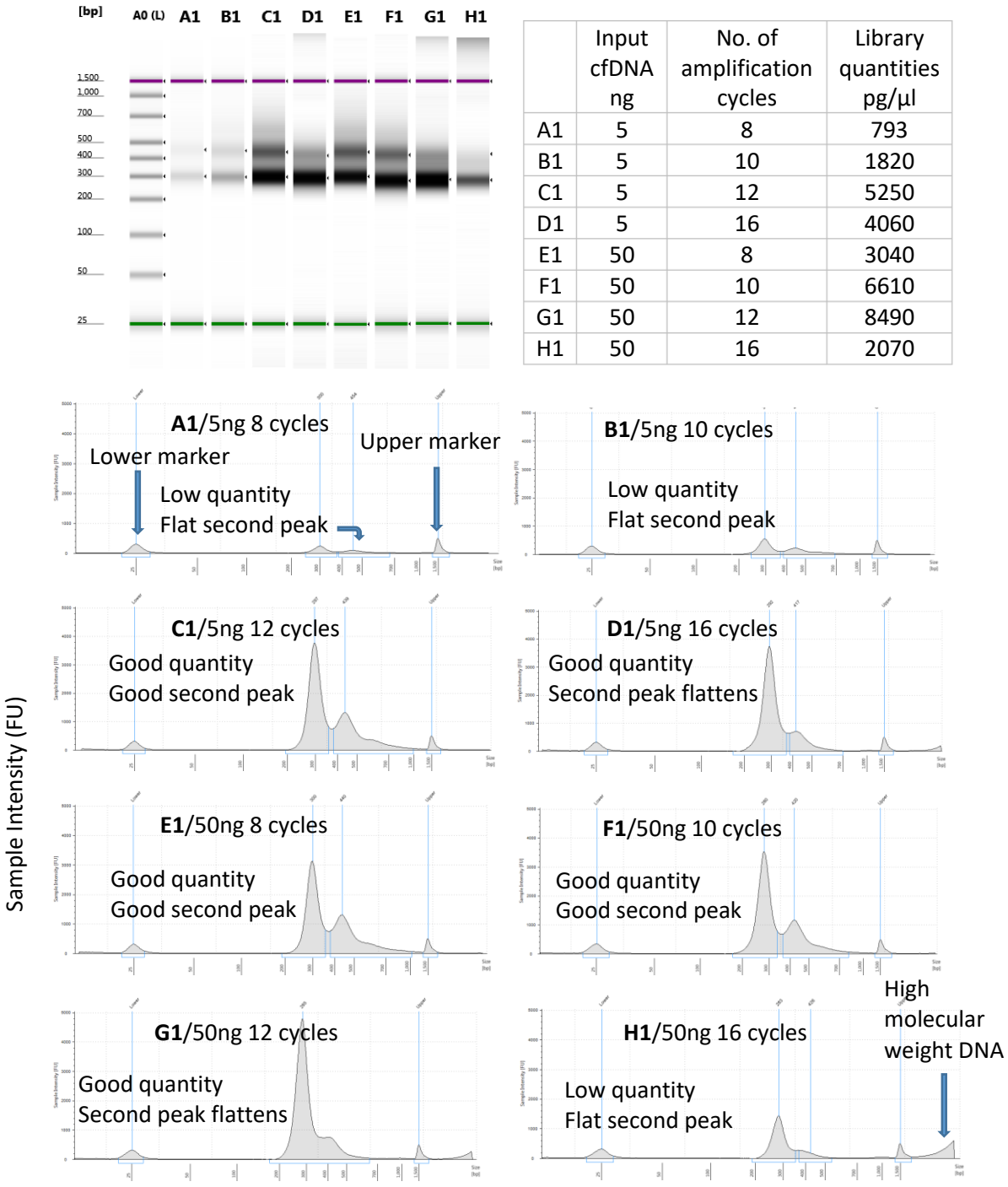
4.3.1.1.1.1 A high number of PCR amplification cycles led to loss of DNA during library preparation

It was important to optimise input cfDNA levels and numbers of amplification cycles to obtain adequate library quantities for sequencing, and avoid over-amplification, which can lead to the loss of DNA (described in the paragraph below). To compare libraries prepared with different amounts of cfDNA and different numbers of amplification cycles, the quality and quantity of prepared libraries were assessed using the Agilent Tapestation 2100 (see Section 2.10.2.1).

For 5 ng of cfDNA, low library quantities were obtained with 8 (793 pg/µl) and 10 (1820 pg/µl) amplification cycles. In comparison, higher library quantities were obtained with 12 (5250 pg/µl) and 16 (4060 pg/µl) cycles. Two peaks were often observed on electrophenogram profiles displaying the size of DNA fragments following amplification (Figure 4-1). The first peak may represent the amplification of mono-nucleosome cfDNA fragments, whilst the second peak may represent the amplification of di-nucleosome cfDNA. The second peak was unexpected because a peak for DNA fragments of approximately 300 bp in baseline cfDNA samples was not observed (see Section 4.3.1.6.1.1) and this is likely to be because the amount of DNA present was too small to be detected by the Agilent Tapestation 2200.

For 50 ng of cfDNA, 10 cycles gave good library quantities (6610 pg/µl) without signs of over amplification. However, when 50 ng were amplified by 16 cycles, the quantity of library fragments decreased (2070 pg/µl). This was due to the exonuclease activity of the DNA polymerase in the NebNEXT® Ultra PCR mastermix leading to DNA loss, as well as the concatenation of fragments demonstrated by the appearance of high molecular weight DNA on the corresponding electropherogram. In addition, for 16 cycles the second fragment peak flattened, indicating over amplification. Representative gel images and electropherogram profiles are displayed in Figure 4-1.

Table 4-1 summarises input DNA levels and the chosen number of amplification cycles to create sequencing libraries with high quantities whilst minimising over amplification. Two nanograms of cfDNA was amplified best by 16 cycles (6840 pg/µl) and 10 ng of cfDNA by 12 cycles (8940 pg/µl) rather than 10 cycles (3510 pg/µl) or 16 cycles (2540 pg/µl).



	Input cfDNA ng	No. of amplification cycles	Library quantities pg/µl
A1	5	8	793
B1	5	10	1820
C1	5	12	5250
D1	5	16	4060
E1	50	8	3040
F1	50	10	6610
G1	50	12	8490
H1	50	16	2070

Figure 4-1: DNA libray quantity and quality when the number of PCR cycles and the cfDNA input amounts from one lung cancer case (1518) were varied, demonstrated by Agilent Tapestation 2200 High Sensitivity Gel images and corresponding electropherograms.

DNA libray quantity and quality was poor when 5ng of cfDNA was amplified by 8 and 10 PCR cycles and library quantity and quality deteriorated for 50ng after 12 cycles.

CfDNA levels	No. of amplification cycles
≥ 2ng < 5ng	16
≥ 5 ng < 25 ng	12
≥ 25 ng ≤ 50ng	10

Table 4-1: CfDNA levels and the optimimal number of amplification cycles chosen to form DNA libraries for sequencing.

4.3.1.1.1.2 The input cell-free DNA quantities and number of PCR cycles did not influence the detection of copy number ratios

To determine whether the detection of CNAs by low coverage whole genome sequencing were affected by different cfDNA quantities and number of PCR cycles to amplify cfDNA during library preparation, copy number ratios were determined for 1 Mb windows and segments (see Section 2.18.6).

Copy number ratios were highly correlated across runs independent of input cfDNA quantities and number of PCR cycles utilised when preparing libraries for sequencing from cfDNA samples (Table 4-2). The median Spearman's rank correlation coefficient was 0.98 (range 0.95-0.99) for case 1518 and 0.98 (range 0.97-0.998) for case 1106.

Case 1518 cfDNA (N=2085)		50 ng				5 ng			
		16 cycles	12 cycles	10 cycles	8 cycles	16 cycles	12 cycles	10 cycles	8 cycles
50ng	16 cycles	1.00							
	12 cycles	0.98	1.00						
	10 cycles	0.95	0.97	1.00					
	8 cycles	0.97	0.98	0.96	1.00				
5ng	16 cycles	0.98	0.99	0.96	0.98	1.00			
	12 cycles	0.96	0.98	0.97	0.98	0.98	1.00		
	10 cycles	0.96	0.98	0.97	0.97	0.98	0.990	1.00	
	8 cycles	0.97	0.98	0.96	0.98	0.98	0.98	0.98	1.00

Case 1106 cfDNA (N=2021)		10 ng			5 ng	2ng
		16 cycles	12 cycles	10 cycles	12 cycles	16 cycles
10ng	16 cycles	1.00				
	12 cycles	0.97	1.00			
	10 cycles	0.98	0.96	1.00		
5 ng	12 cycles	0.996	0.97	0.98	1.00	
2 ng	16 cycles	0.993	0.96	0.98	0.998	1.00

Table 4-2: Spearman's rank correlations for copy number ratios to evaluate the effect of different cfDNA quantities and number of PCR cycles during library preparation for two lung cancer cases (1106 and 1518).

All comparisons were significantly correlated with p value <0.0001. N= number of copy number ratio values per sample.

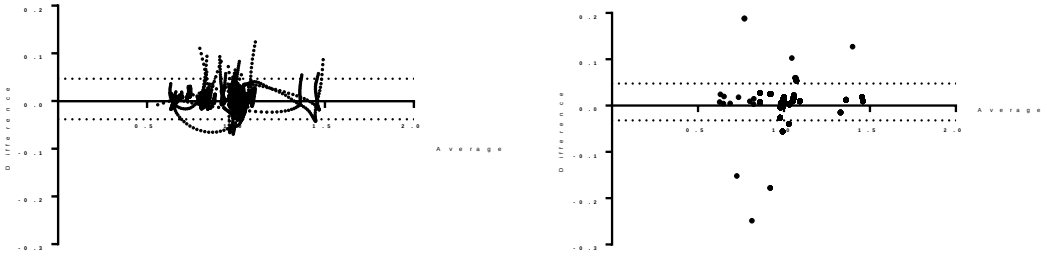
4.3.1.1.2 Optimising library preparation for genomic DNA samples

Agilent Tapestation 2200 electropherogram profiles were similar when libraries were made from 100 ng of genomic DNA and amplified with seven, eight or nine cycles. Generally, higher library quantities were obtained with increasing cycle number (data not shown). Seven amplification cycles were used to amplify genomic DNA samples during library preparation.

4.3.1.1.3 There was no significant difference between the NEBNext® Q5 hot start HiFi PCR master mix and the NEBNext® High-Fidelity PCR master mix

To amplify barcode-adaptor ligated DNA fragments during library preparation the NEBNext® master mix containing DNA polymerase, dNTPs, Mg^{2+} and a propriety buffer was utilised. The NEBNext® Q5 Hot Start HiFi PCR master mix replaced the NEBNext® High-Fidelity PCR master mix, and it contained a more efficient DNA polymerase that was also inactive at room temperature. To compare the two NEBNext® master mixes, libraries for sequencing were prepared using both kits for matched cfDNA and genomic samples from case 1106.

Copy number ratios for individual windows and segments were highly comparable when libraries were prepared from the same cfDNA sample with the old (NEBNext® High-Fidelity PCR) and new (NEBNext® Q5 Hot Start HiFi PCR) master mix (Figure 4-2). The Bland Altman test was used to plot the difference in the copy number ratios for old and new kits for individual 1 Mb windows or segments and compare it to the average of the paired measurements for the old and new kit. The bias or average of the differences for all comparisons was close to zero for both samples indicating that the old and new kit produced similar results. Furthermore, as the average increased the difference in the method did not increase, indicating that results were consistent between methods across a range of copy number ratios. These data were not normally distributed as determined by the D'Agostino-Pearson omnibus K2 normality test and therefore results must be interpreted with some caution because the Bland Altman test is for normally distributed data.



A B

Figure 4-2: A comparison of the NEBNext® Ultra DNA old (NEBNext® High-Fidelity PCR) and new (NEBNext® Q5 Hot Start HiFi PCR) mastermix by using Bland Altman plots to compare copy number ratios (A) and segments (B) identified by low coverage sequencing of 10ng of cfDNA from lung cancer case 1106.

A: comparison of copy number ratios (N=2045) of 1 Mb windows after 12 PCR cycles: Bias -0.004, 95% limits of agreement -0.04- +0.05. Each dot represents the differences between the copy number ratios for the old and new master mixes plotted against the average copy number ratio of the two methods.

B: A comparison of segmental copy number ratios (N=2086) after 12 PCR cycles: Bias 0.0008, 95% limits of agreement -0.03- +0.05. Each dot represents the differences between the segmental copy number ratios for the old and new master mixes plotted against the average segmental copy number ratio of the two methods.

4.3.1.2 The limit of detection for tumour-derived copy number aberrations was between 10-20% by visual inspection

To test the lower limit for the detection of tumour-derived copy number aberrations, tumour FFPE DNA from case 261 was added at known proportions to extracted cfDNA from the pooled plasma of healthy volunteers (see Section 2.2.1). Segmental copy number ratios and copy number ratios for individual 1 Mb windows were compared for each dilution to establish the limit of detection for CNAs.

Figure 4-3 displays scatter diagrams comparing segmental copy number ratios detected for 100% tumour FFPE DNA (10ng) and descending proportions. As the tumour DNA fraction reduced, both the correlation coefficients of segmental copy number ratios and the correlation coefficients of copy number ratio values of individual 1 Mb windows were reduced. Upon performing the experiment for a second time in a different sequencing run, the Spearman's rank correlations were less for both segmental and individual copy number ratio values of 1 MB windows (Table 4-3). In summary, visual inspection shows that tumour

derived CNAs were detected with tumour FFPE DNA fraction of 10%-20% (the lower limit of detection was examined further through a CNA score in Section 4.3.1.4).

Copy number ratios were significantly correlated between cfDNA extracted from pooled healthy plasma with no added tumour FFPE DNA and 100% tumour FFPE DNA. This could be due to sequencing artefact. For example in *Figure* 4-3 a copy number loss of chromosome 19 was observed in all cfDNA samples spiked with tumour FFPE DNA as well as the cfDNA sample without tumour FFPE DNA. Alternatively, cfDNA extracted from pooled healthy control plasma may have become contaminated by tumour FFPE DNA during library preparation.

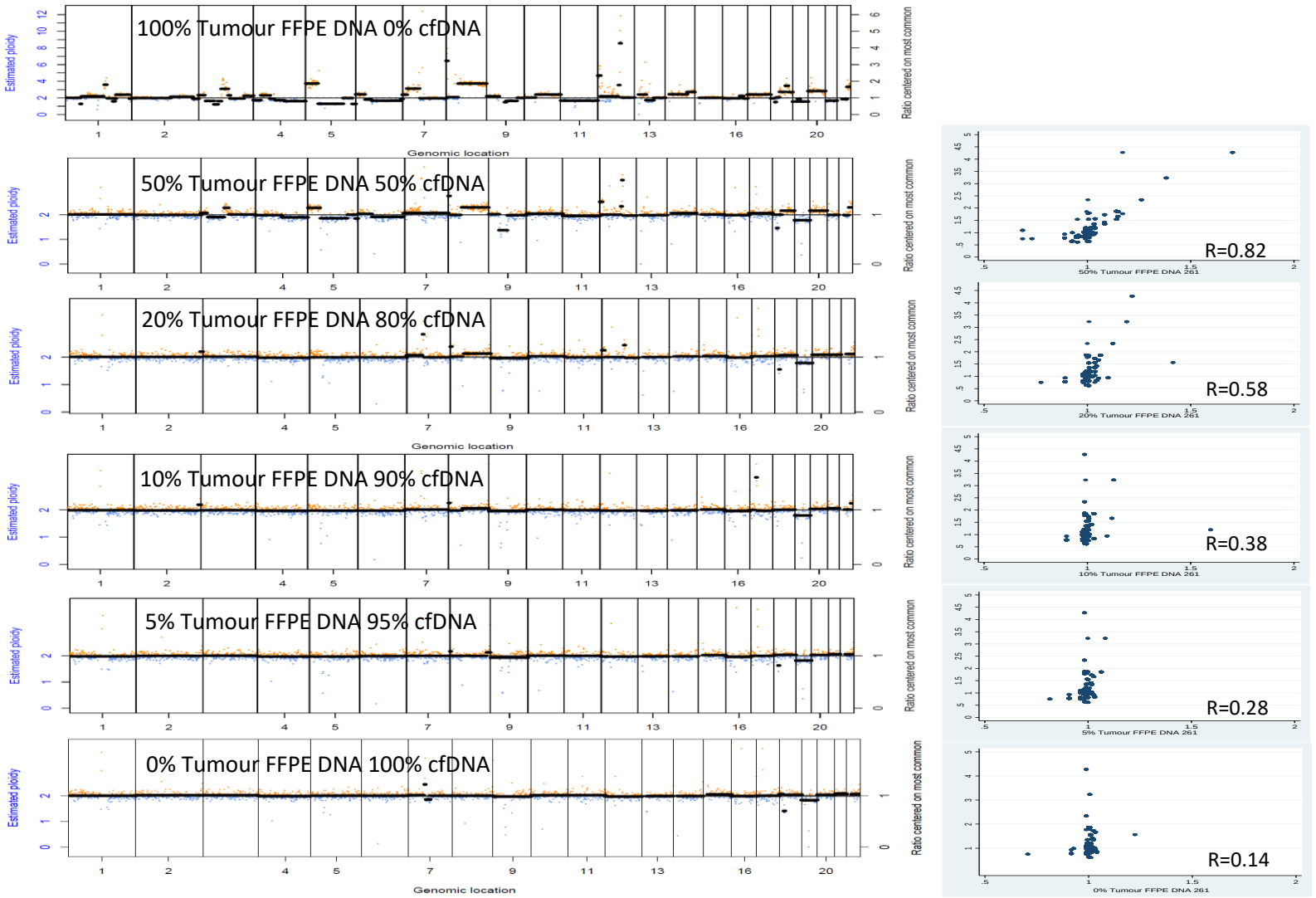


Figure 4-3: Copy number profiles and scatter diagrams to demonstrate the lower limit of detection of copy number ratios and segments when tumour FFPE DNA was spiked into healthy volunteer control cfDNA in descending proportions.
All correlations were significant with p<0.0001 unless otherwise stated (N=2072). The y-axis of the copy number profile graphs shows copy number ratio on the left and estimated ploidy on the right with the x-axis showing chromosome position. The y-axis of the scatter diagram shows the copy number ratios of 100% tumour FFPE DNA and the x-axis shows the copy number ratios for samples spiked with different proportions of tumour FFPE DNA.

Proportion of tumour FFPE DNA compared to cfDNA	Spearman's Rank correlations for Experiment 1 (cfDNA 10ng)		Spearman's Rank correlations for Experiment 2 (cfDNA 10ng)	
	copy number ratios from segments	copy number ratios from 1Mb windows	copy number ratios from segments	copy number ratios from 1Mb windows
50% (5ng)	0.82 (N=2072)	0.89 (N=2044)	0.81 (N=2072)	0.85 (N=2040)
20% (2ng)	0.58 (N=2072)	0.61 (N=2044)	0.40 (N=2072)	0.55 (N=2044)
10% (1ng)	0.38 (N=2072)	0.51 (N=2044)	0.21 (N=2055)	0.28 (N=2044)
5% (0.5ng)	0.28 (N=2072)	0.34 (N=2044)	0.03 (N=2072) p=0.19	0.17 (N=2044)
0% (0 ng)	0.14 (N=2072)	0.10 (N=2044)	-0.03 (N=2072) p=0.07	0.06 (N=2044) p=0.01

Table 4-3: Spearman's Rank correlations to test the identification of tumour derived CNA with different proportions of tumour FFPE DNA from lung cancer case 261.

All correlations were significant with p<0.0001 unless otherwise stated.

4.3.1.3 There was good reproducibility for copy number ratios between sequencing runs

There is potential for bias to be introduced when performing multiple sequencing runs. Reproducibility was evaluated by determining the agreement between sequencing runs of segmental copy number ratios from DNA extracted from the SCLC cell line H69 (see Section 2.20.2). There was strong agreement between sequencing runs when comparing the copy number ratios of segments when consecutive runs were analysed with the Bland Altman statistic (Table 4-4). It must be noted that these data are not normally distributed.

Run		CNV 4 10 ng	CNV 5 10 ng	CNV 6 10 ng	CNV 7 10 ng	CNV 8 10 ng	CNV 9 10 ng
CNV 3 100ng	Bias %	-0.79	-1.14	-0.93	-0.97	-0.94	-0.90
	SD %	3.4	3.4	4.4	6.2	6.0	6.7
	95% CI	-7.5-+5.9	-7.7-+5.4	-9.4-+7.6	-13.1-+11.2	-12.8-+10.9	-14.1-+12.3

Table 4-4: Bland Altman statistic to compare sequencing runs for copy number ratios from segments for cell line DNA H69 (N=2069).

4.3.1.4 The lower limit of detection for the Copy Number Aberration score may be 5%

To gain a preliminary measure of the lower limit of detection for the CNA score, CNA scores were calculated for descending proportions of tumour DNA from case 261 (N=2) and sheared cell-line DNA H69 (N=1) spiked into cfDNA extracted from the pooled plasma of healthy volunteers (see Section 2.2.1)(Figure 4-4)(Table 4-5). The reduction in the CNA score from

100% DNA to 50% DNA was greater for tumour FFPE compared to cell-line DNA. This is consistent with the dilutional effect being more for poor quality DNA due to damage by formalin fixation and paraffin embedment. The CNA score was higher when 5% of DNA was spiked into healthy control plasma compared to 0% for all three experiments, suggesting that the limit of detection may be 5% (although smaller proportions were not tested)(copy number profiles are shown in Appendix F). The CNA scores for cfDNA extracted from pooled healthy volunteer plasma were higher (415 and 325) compared to the median score of high risk controls of 252, although the range of CNA scores in the high risk group was very broad from 149 to 7122.

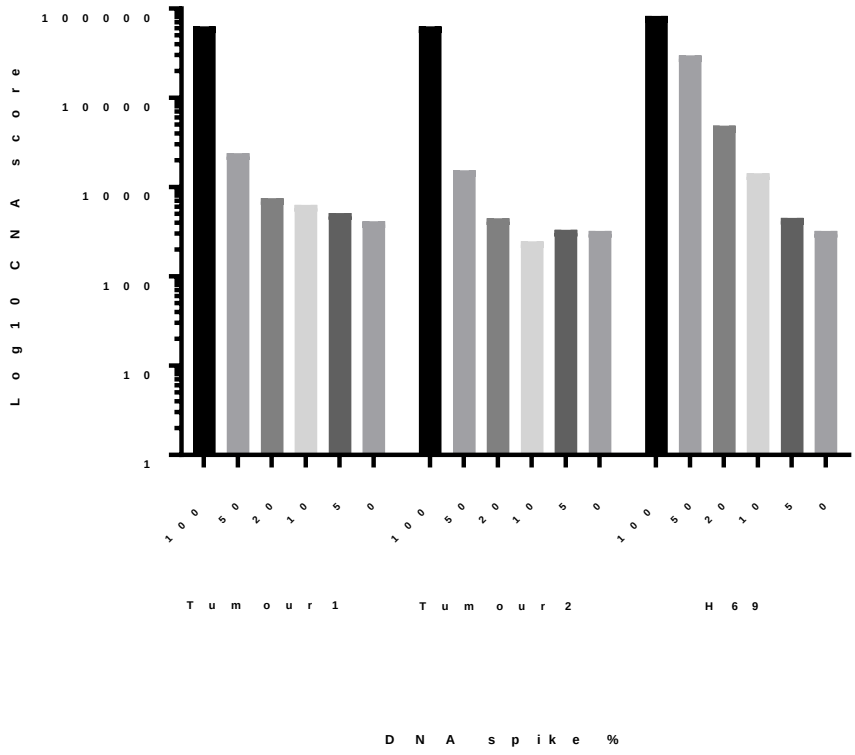


Figure 4-4: CNA scores for different proportions of tumour FFPE (N=2) and H69 cell-line DNA (N=1) spiked into extracted cfDNA from the pooled plasma of healthy volunteers.

Proportion of tumour FFPE or cell-line DNA compared to cfDNA	CNA scores for Experiment 1 tumour FFPE DNA	CNA scores for Experiment 2 tumour FFPE DNA	CNA scores for H69 cell-line DNA
100% (10ng)	63455	63390	83029
50% (5ng)	2412	1553	29962
20% (2ng)	756	448	4883
10% (1ng)	637	248**	1430
5% (0.5ng)	514	334	453
0% (0 ng)	415	325*	325*

Table 4-5: Copy number aberration scores for different proportions of tumour FFPE and cell-line DNA spiked into extracted cfDNA from the pooled plasma of healthy volunteers.

* these samples were sequenced in the same sequencing run and therefore the 0% value was used for both FFPE DNA and H69 DNA experiments.

** the low CNA score for 10% tumour FFPE DNA in experiment two may be explained by a pipetting error.

4.3.1.5 There was good reproducibility for the Copy Number Aberration score between sequencing runs

The median CNA score for cell-line H69 DNA across seven independent sequencing runs was 83073 (range 82675-84661) (Table 4-6). The relative variability across sequencing runs measured by the coefficient of variance (CV) was 0.94% and was calculated by determining the relative difference between the standard deviation of the seven CNA scores divided by the mean.

CNA Run	CNV 3 100 ng	CNV 4 10 ng	CNV 5 10 ng	CNV 6 10 ng	CNV 7 10 ng	CNV 8 10 ng	CNV 9 10 ng
H69 CNA score	84457	83029	84661	83073	82849	83366	82675

Table 4-6: Copy number aberration scores of cell-line H69 DNA across seven sequencing runs.

4.3.1.6 Low coverage whole genome sequencing of lung cancer cases and controls

A pilot study was carried out to determine whether CNAs would be detected in cfDNA of lung cancer cases and controls by low coverage whole genome sequencing. Selected lung cancer cases and controls from the ReSoLuCENT study that had cfDNA levels quantified by SYBR

green RT-qPCR (N=114) (see Section 3.3.3) were chosen for low coverage whole genome sequencing.

Of the 114 individuals, 12 were eliminated from further analyses. Three cases and one control had cfDNA levels less than 2 ng and were not sequenced and seven cases had no cfDNA available after targeted sequencing was performed in a different study. One further case did not have adequate coverage of matched genomic DNA despite re-extraction of DNA from the buffy coat layer, and was not included in analyses. Thus, sequencing data is presented for cfDNA and matched genomic DNA samples for 62 lung cancer cases (51 untreated and 11 treated), 30 high risk and 10 low risk controls (N=102).

The characteristics of the analysed cases and controls (N=102) were consistent with the characteristics of the selected subjects described in Section 3.3.3.3. Similar to the findings in Section 3.3.3.2, the median age at diagnosis of selected cases (N=62) compared to non-selected cases (N=618) was older (median 57.4 years (range 40.0-74.5 years) vs 56.0 years (range 20.7-83.1 years) (p=0.03, Mann Whitney U test). Furthermore, and as expected, the characteristics of controls selected by LLP risk score and analysed for CNAs (N=40) were statistically significantly different from non-selected controls (N=399) for gender (p=0.005, Chi-squared test), age (p<0.0001, Mann Whitney U test) and smoking status (p=0.005, Chi-squared test). The histology and stage of cases (N=62) selected for copy number analyses are shown in Table 4-7.

		NSCLC				SCLC	Total
Stage		Adenocarcinoma	Squamous	Not otherwise specified	other		
	I	4	1	-	-	-	5 (8%)
	II	4*	4	1	1	-	10 (16%)
	III	6*	6**	4	-	2*	18 (29%)
	IV	9*	6**	6**	1	7*	29 (47%)
Total		23 (37%)	17 (27%)	11 (18%)	2 (3%)	9 (15%)	62

Table 4-7: Histology and stage of cases (N=62) selected for copy number aberration analysis.

*one treated case in the sub-group. ** two treated cases in the sub-group.

Consistent with findings in Section 3.3.3.5, median cfDNA levels ng/ml were significantly higher in cases (N=62) compared to controls (N=40), 8.0 ng/ml (range 1.6-545.1 ng/ml) vs 4.4 ng/ml (range 1.2-34.0 ng/ml) (p<0.0001 Mann Whitney U test). In addition, treated cases

(N=11) had significantly higher cfDNA levels compared to untreated cases (N=51), median 17.1 ng/ml (range 4.7-540.1 ng/ml) vs 7.2 ng/ml (range 1.6-545.1 ng/ml) (p=0.03 Mann Whitney U test). The characteristics of untreated early stage cancer cases (N=21) and high-risk controls (N=30) are shown in Table 3-7.

4.3.1.6.1 Quality control

4.3.1.6.1.1 DNA parameters prior to library construction

DNA was extracted from matched genomic (N=102), tumour FFPE (N=10), and plasma samples (N=102) and quantified by the methods described in Section 2.9. Genomic DNA was sheared by ultrasonic acoustic waves to form short fragments of target length 200 bp to form DNA libraries for sequencing (see Section 2.14.1.1). CfDNA was not sheared because cfDNA fragments are typically <200 bp. Prior to DNA library construction, cfDNA and fragmented genomic DNA samples were run on the Agilent 2100 Tapestation. This was important to check for uniformity of fragment size and to ensure the adequate shearing of genomic DNA samples (see Section 2.10).

The median peak fragment size for cfDNA was 147 bp (mean 148 bp, range 118-185 bp, N=84). The median peak fragment size for sheared genomic DNA was 225 bp (mean 227 bp, range 169-338 bp, N=93). Representative examples of gel images from the Agilent Tapestation 2100 for cfDNA and genomic DNA are shown in Figure 4-5. Higher molecular weight DNA of more than 500 bp was present in most samples of cfDNA tested on the Tapestation.

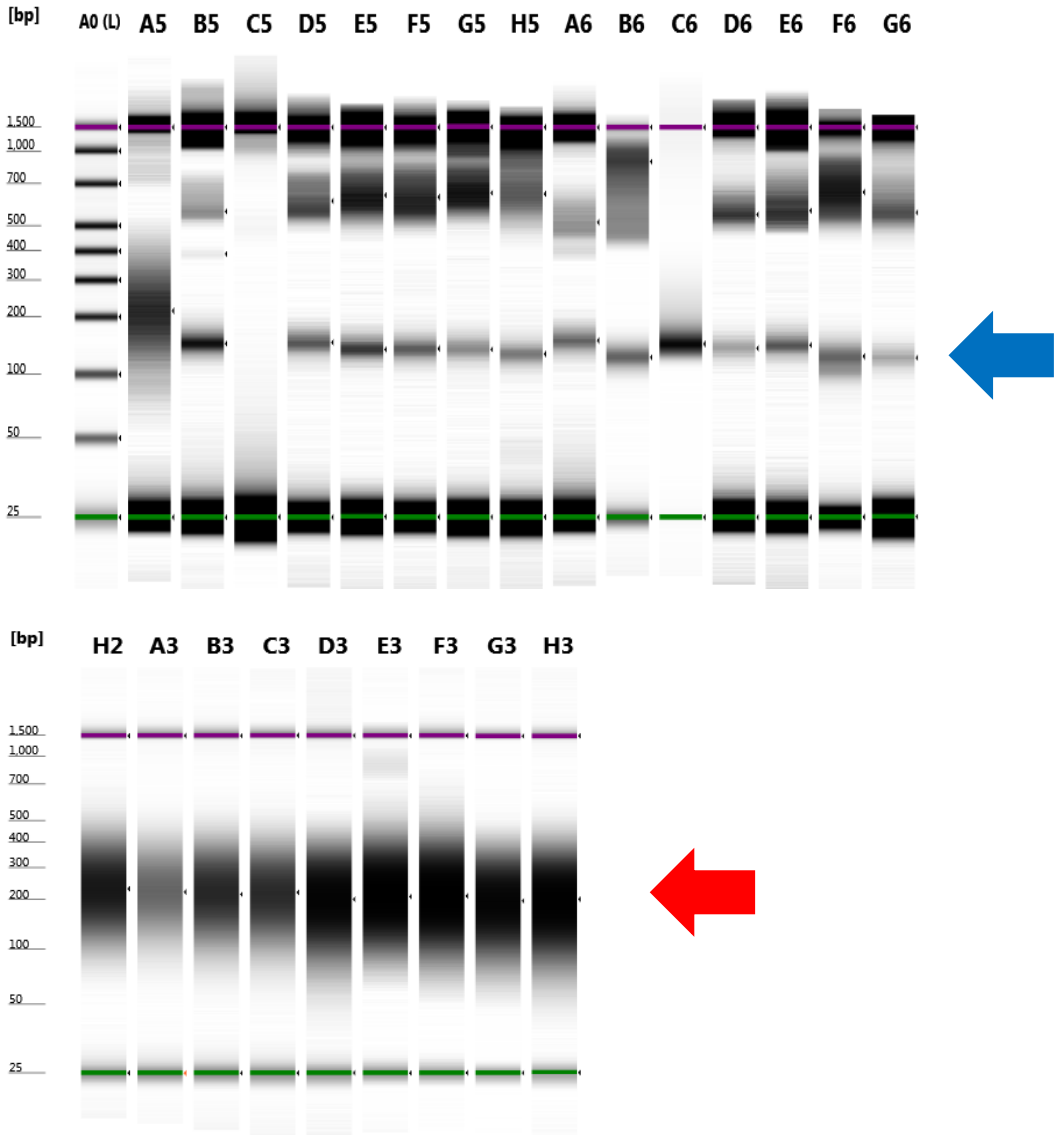


Figure 4-5: Representative gel images demonstrating the fragment sizes of cfDNA and sheared genomic (lymphocyte) DNA from the Agilent Tapestation 2100.

NCI-H69 SCLC cell line DNA.(A5): cfDNA samples (B5-G6) and genomic DNA (lower panel H2-H3). A0- hyperladder. C5-failed run.

The blue arrow shows the expected size of cfDNA fragments at 160 bp. The red arrow shows 200 bp the target length for sheared genomic DNA.

For cfDNA samples, the peak fragment length was available for 84 of 102 participants. The median cfDNA peak fragment length was significantly longer for low risk controls (N=10) compared to high risk controls (N=26) and lung cancer cases (N=48), 158 bp (range 152-170 bp) vs 146 bp (range 128-167 bp) (p=0.0002 Mann Whitney U test) and 146 bp (range 118-185 bp) (p=0.0005 Mann Whitney U test), respectively (Figure 4-6).

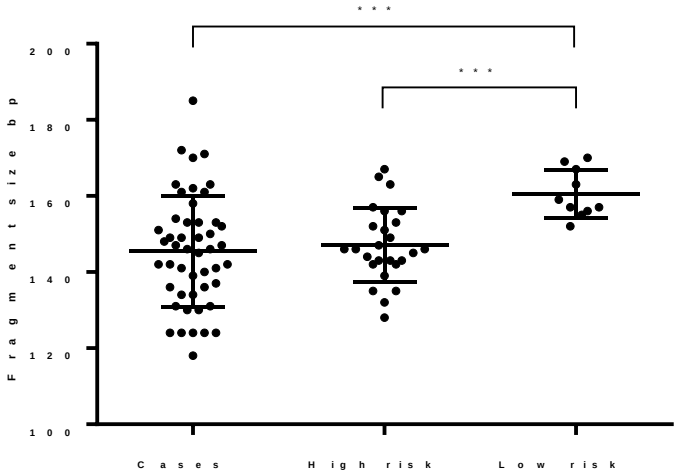


Figure 4-6: The peak fragment size of cfDNA for cases (N=48), high risk controls (N=26) and low risk controls (N=10) prior to DNA library construction.

*** p<0.001. Median and IQR are shown.

In this pilot study, libraries were prepared from cfDNA (N=102), genomic (N=102) and tumour FFPE DNA (N=10) for low coverage whole genome sequencing with Illumina HiSeq 2500 to establish the presence of CNAs. A standard protocol was followed to prepare DNA samples for sequencing using the NebNEXT® Ultra DNA library preparation kit (see Section 2.14.1). DNA libraries were prepared from as much cfDNA as was available, up to a maximum of 50 ng of cfDNA. The median cfDNA quantity to prepare libraries for low coverage sequencing was 20 ng (N=62, range 5.5- 50 ng) for cases and 11.5 ng for controls (N=40, range 3- 50 ng) (p <0.0001, Mann Whitney U test).

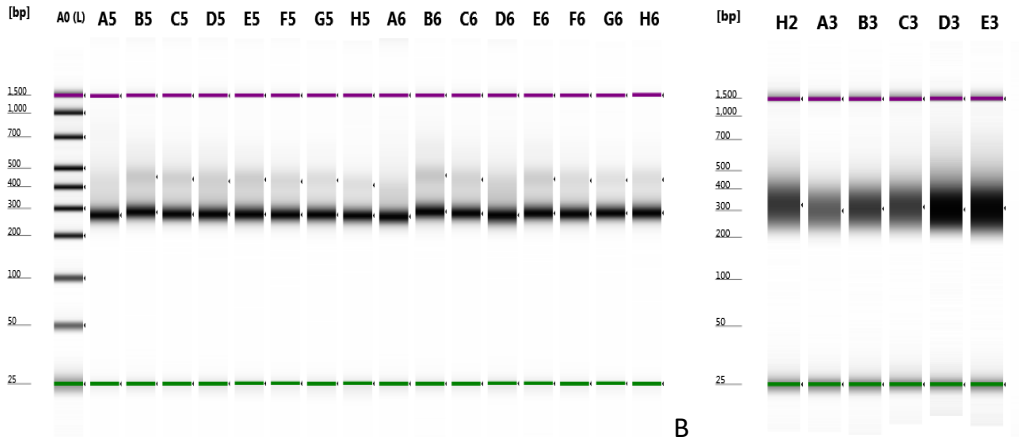
4.3.1.6.1.2 DNA library quality control

An assessment of DNA library quality was important to rule out library contamination by adaptor-dimers or primer-dimers, and to achieve optimal cluster formation during sequencing to maximise sequencing efficiency with the Illumina HiSeq 2500. Library quality was assessed by running libraries from all samples on the Agilent Tapestation 2100 to determine the size and concentration of DNA fragments prior to sequencing (see Section 2.15.1).

4.3.1.6.1.3 No prepared DNA libraries were contaminated by adaptor- dimers or primer-dimers

Electropherograms for all samples from the Agilent Tapestation 2100 were reviewed to assess library fragment size and concentration. No fragments less than 125 bp or less than 60 bp were observed and therefore there was no contamination of libraries by adaptor-dimers or primer dimers respectively (Figure 4-7).

The median DNA fragment size of the main peak for the cfDNA library was 292 bp (range 268-302 bp N=102) and 301 bp for genomic DNA (range 257-337 bp, N=102). For cfDNA, there was often a second peak with a higher molecular weight between 400-500 bp (Figure 4-7). The median library quantity determined by the Agilent Tapestation 2100 was 7.24 ng/µl (mean 6.96 ng/µl, range 1.07-13.70 ng/µl, N=102) for cfDNA and 2.65 ng/µl (mean 3.18 ng/µl, range 0.50-10.40 ng/µl, N=102) for genomic DNA. The size of library DNA fragments were longer than input DNA fragments due to the ligation of barcodes and sequencing adaptors and therefore the libraries passed quality control (see Section 2.15).



A B

Figure 4-7: Representative gel images from the Agilent Tapestation 2100 showing fragment sizes for DNA libraries prepared for sequencing.

cfDNA (A: A5-H6) and genomic DNA libraries (B:H2-E3).

4.3.1.6.1.4 Sequencing quality control

Low coverage whole genome sequencing was carried out with the Illumina HiSeq 2500 (see Section 2.17). Nine sequencing runs were performed in total. All clusters of amplified DNA fragments passed quality control measures to assess the strength and reliability of the emitted fluorescence signal intensity (see Section 2.18.1). A maximum of 24 samples were

sequenced per lane of the flow cell resulting in low genome coverage for each sample. Each run contained a mixture of cfDNA, genomic DNA and tumour DNA libraries. Table 4-8 summarises important quality control parameters for each sequencing run.

Run	No. of samples	Pooled library concentraion pM	Cluster density K/mm^2		Median Q30 base quality scores (range)	Median data output per sample (GB) (range)
			Lane 1	Lane 2		
CNV 1	24	10	830	-	88 (87-90)	732 (237-1,959)
CNV 2	24	10	1167	-	91 (88-92)	1,127 (784-3,880)
CNV 3 (Lane 1)	24	12.5	1090	-	92 (89-93)	1,199 (945-4,855)
CNV 3 (Lane 2)*	2	12.5	-	1085	93 (91-94)	17,311 (17,186-17,436)
CNV 5	48	11	1150	1158	91 (87-92)	1,048 (97-6,602)
CNV 6	44	11	1125	1147	91 (88-92)	1,446 (962-5,106)
CNV 7	48	11	1182	1190	90 (72-95)	1,218 (0-7,632)
CNV 8	46	10	1074	1088	96 (92-96)	1,636 (0-6,528)
CNV 9	43	10	907	910	97 (94-98)	1,436 (0-6,528)

Table 4-8: Important quality control parameters for each sequencing run on the Illumina HiSeq 2500.

*For Run 3 two samples were separated into a different lane to obtain higher coverage and therefore potentially increase the sensitivity for the detection of CNAs.

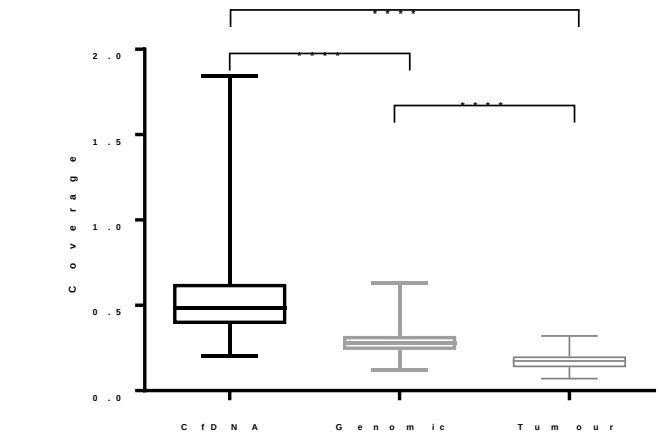
The data for each sequencing run was processed through a bioinformatics pipeline to obtain sequences of bases or reads that were then grouped by their barcodes into their originating samples (de-multiplexed) (see Section 2.18.2). For each sample, reads were aligned or mapped to the human reference genome and poorly mapped or duplicate reads were discarded (see Section 2.18.3 and 2.18.4).

4.3.1.6.1.5 Samples were sequenced at low coverage

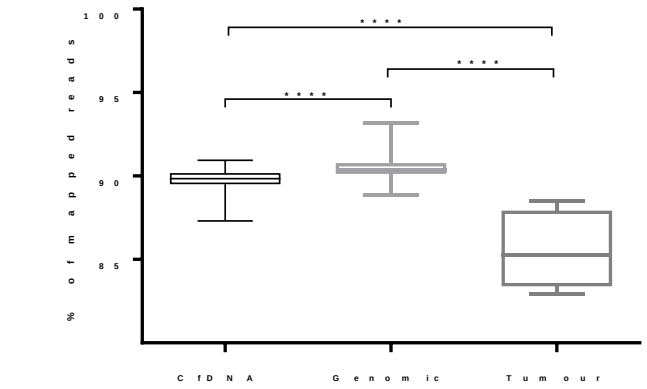
The coverage was calculated by determining the number of bases sequenced (total number of mapped reads x read length) divided by the size of the human genome, 3 billion bases (256) (see Section 2.18.5). The median coverage was highest for cfDNA at 0.49X (range 0.20X-0.63X, N=102), followed by genomic DNA with a coverage of 0.28X (range 0.12X-0.63X, N=102) and tumour FFPE DNA with a median coverage of 0.18X (range 0.07X-0.32X, N=10) (Figure 4-8).

4.3.1.6.1.6 Approximately 90% of reads mapped to the reference genome for each sample

Approximately 10% of reads were unmapped for each sample and these reads were discarded. These included duplicate reads with identical start and stop positions, and reads that mapped to multiple sites of the genome. The median percentage of duplicate reads for all samples was 0.0075% (range 0.0027%-0.014%). There were fewer mapped reads for tumour FFPE DNA (N=10) compared to cfDNA (N=102) ($p<0.0001$) and genomic DNA (N=102) ($p<0.0001$ Mann Whitney U test) (Figure 4-8).



A



B

Figure 4-8: Important sequencing parameters for sheared genomic (N=102), cfDNA (N=102) and sheared tumour FFPE DNA (N=10).

Box plots show the median, IQR and minimum and maximum values. **** $p<0.0001$, Mann Whitney U test. A: Sample coverage. B: Percentage of reads aligning to the reference human genome.

4.3.1.6.1.7 There were some chromosomal regions with no calculated copy number ratios

There were no mapped reads and therefore no copy number ratio values established for nine chromosomal regions varying in size from 1 Mb to 19 Mb. These regions tended to be within or close to centromeres or telomeres (Table 4-9). There were no copy number ratios calculated for the short arm of chromosome 13 or 14. Genomic positions were identified from the UCSC genome browser (284). The total size of the genomic regions with no calculated copy number ratios was approximately 100 Mb, which is equivalent to 3% of the whole genome.

Chromo	Window Start position	Window End position	Size in Mb	Approximate location
1	123000001	142000001	19	1p36.21
9	46000001	59000001	13	within 3MB of a centromere (9q11-12)
13	1	17000001	17	13p
14	1	17000001	17	14p
14	107000001	108000001 (end position of chromosome 107043718)	< 1	Telomere
15	1	16000001	16	Telomere and 15p13-11.2
16	39000001	45000001	6	within 3Mb of a centromere (16q11.2)
21	1	4000001	4	Telomere and 21p13
22	1	9000001	9	Telomere and 22p13-12

Table 4-9: Chromosomal regions with no copy number ratio values obtained for sequenced samples (N=102).

4.3.1.7 Validation of copy number profiles determined by low coverage sequencing

4.3.1.7.1 Tumour FFPE DNA copy number profiles processed in Sheffield were similar to the profiles previously established in the Wood laboratory in Leeds

Tumour FFPE DNA copy number profiles were comparable between laboratories in Sheffield and Leeds, despite independent DNA library preparation and sequencing methods. There were a similar pattern of copy number gains and losses demonstrated (Figure 4-9) (N=3).

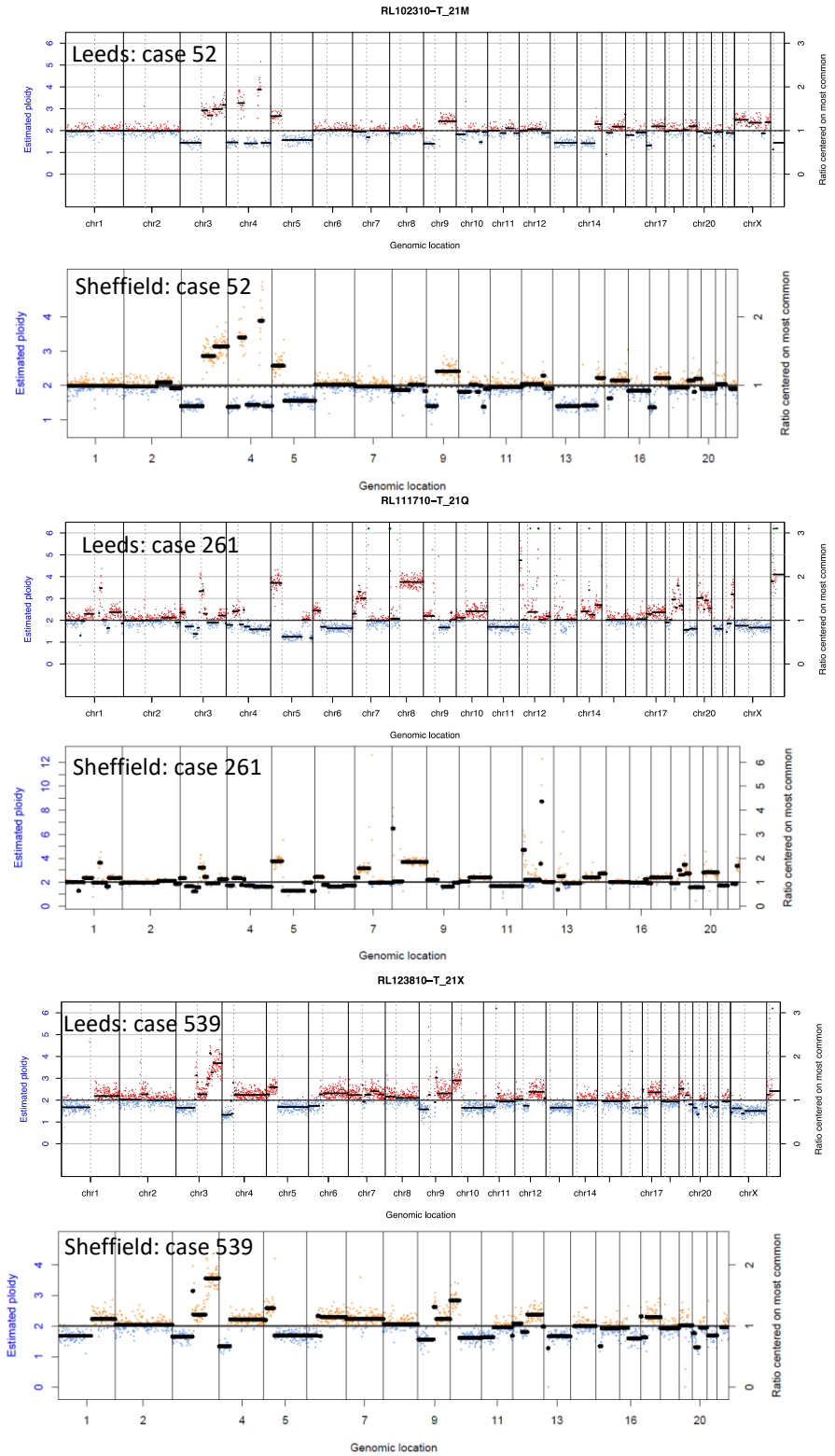


Figure 4-9: A comparison of copy number profiles for tumour FFPE DNA independently processed, sequenced and analysed in Leeds and Sheffield for three different lung cancer cases.

The X and Y profiles have been removed from the Sheffield data. The left axis shows estimated ploidy and the right axis is the copy number ratio. Dots represent copy number ratios of windows and black lines represent segments or windows with similar copy number ratios. Orange or red denotes potential copy number gains and blue denotes copy number losses.

4.3.1.7.2 Tumour FFPE copy number aberrations were detected in matched circulating cell-free DNA samples

Tumour FFPE DNA CNAs were detected in matched cfDNA for some but not all lung cancer cases. Ten matched tumour FFPE DNA and cfDNA pairs were available for comparison. Five cases had tumour tissue available from a primary bronchial biopsy and the other five cases had tumour tissue available following surgical resection of the primary lung tumour. The median time from the tissue sample to blood withdrawal was 30 days (range 0-273 days).

Similar copy number profiles between tumour FFPE and cfDNA were obtained for two lung cancer cases (20%) (52 and 203) (Figure 4-10). For these cases, tumour-derived copy number chromosomal gains and losses were identified in the matched profiles of cfDNA by visual inspection, albeit at lower magnitude, due to dilution of ctDNA by wild type host cfDNA in the blood. For 6 cases (60%) (146, 261, 527, 539, 800 and 805), CNAs were detected for tumour FFPE DNA but not cfDNA (Figure 4-11). Two cases (20%) (240 and 806) had few CNAs detected in both tumour FFPE DNA and cfDNA (Figure 4-12). CfDNA of lung cancer cases 800 and 240 were sequenced at higher coverage (4.26X and 4.18X respectively). However, the detection of CNAs in cfDNA did not improve (data not shown). The clinical characteristics of each case are shown in Table 4-10.

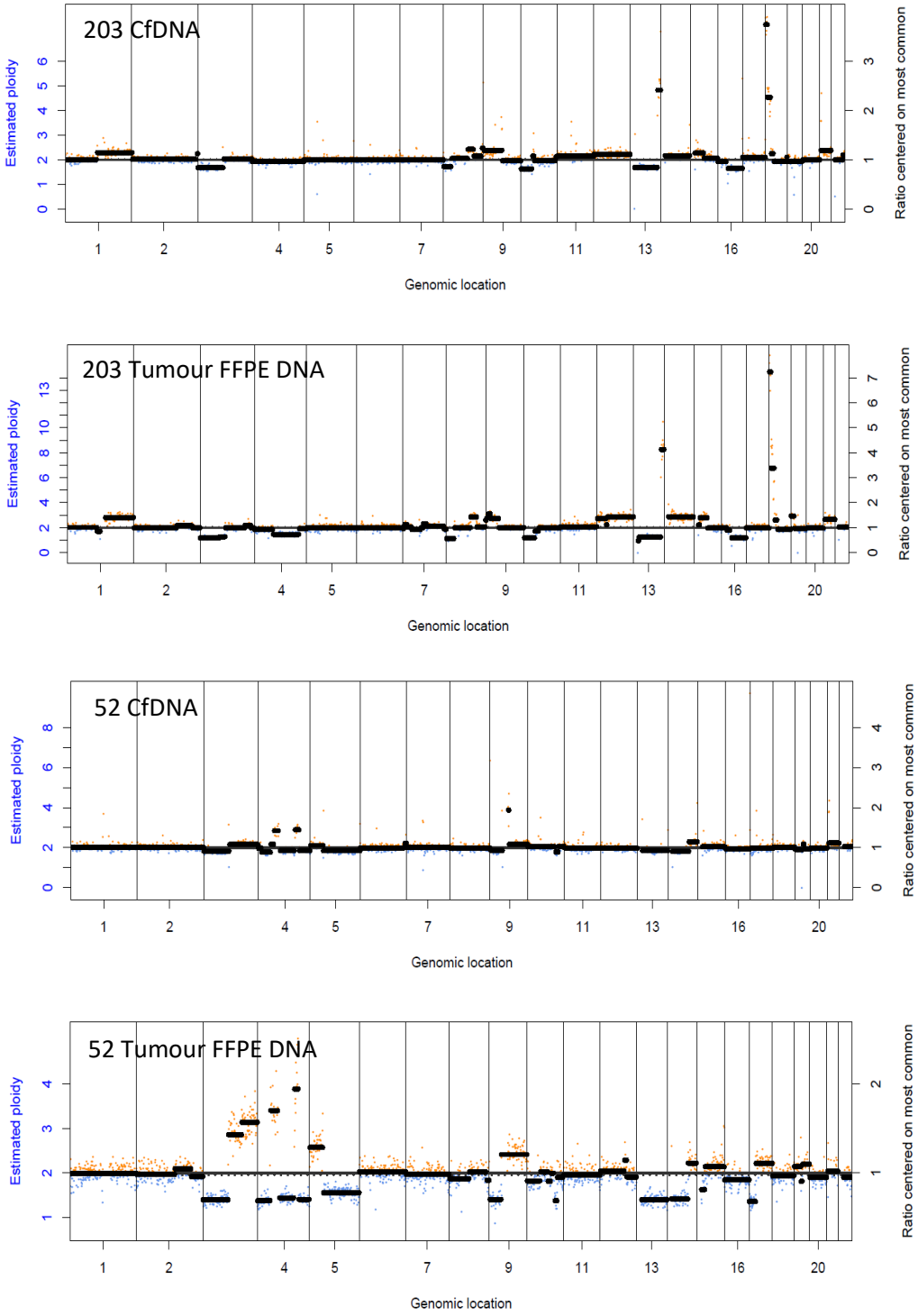


Figure 4-10: Similar copy number profile graphs for tumour FFPE DNA and matched cfDNA for two lung cancer cases.

Dots represent copy number ratios of 1 Mb windows and black lines represent consecutive windows with similar copy number ratios. Orange denotes copy number gain and blue copy number loss. Y-axis shows estimated ploidy and x-axis chromosomal position.

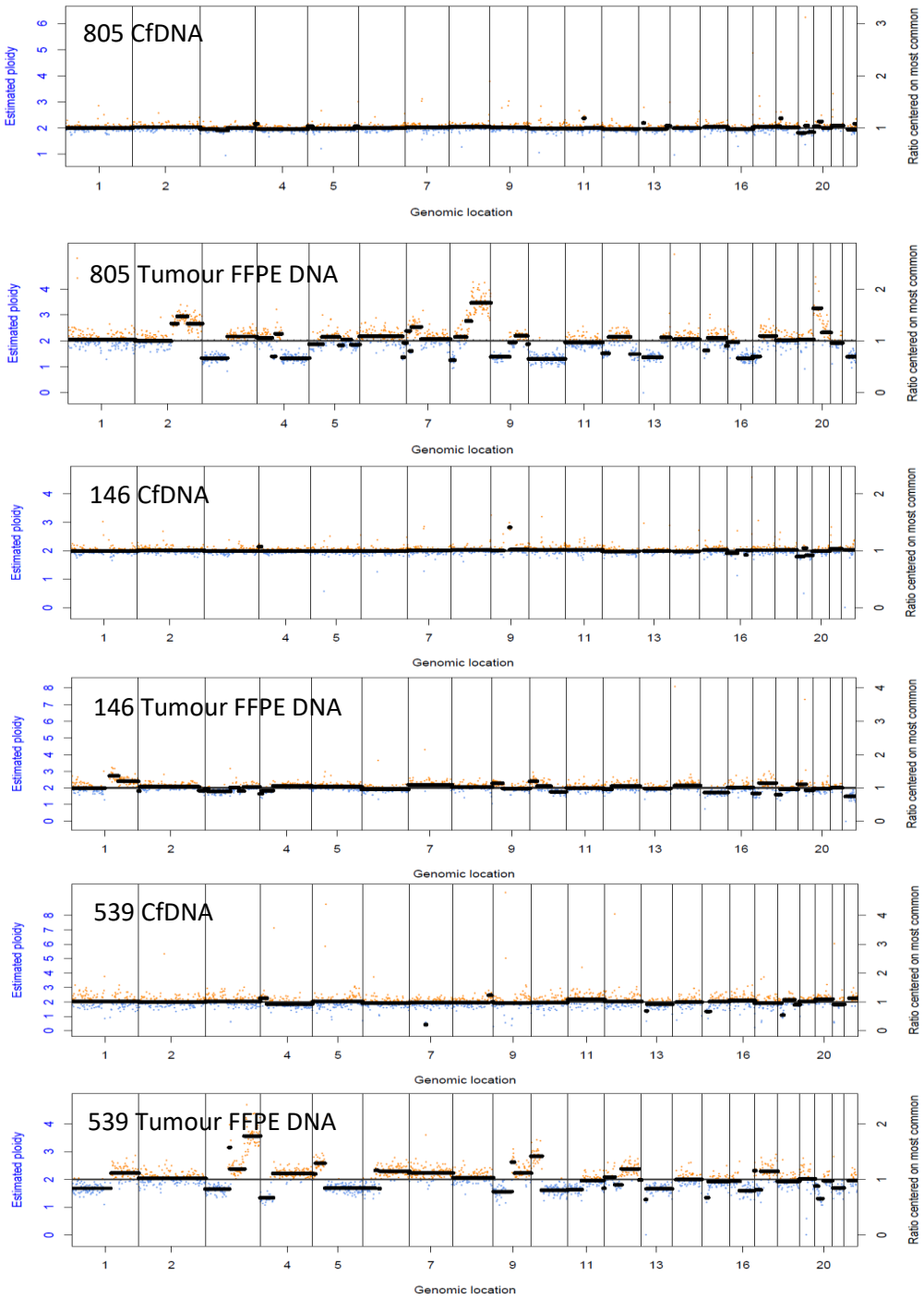


Figure 4-11: Differing copy number profile graphs for three cases with multiple copy number aberrations identified in tumour FFPE DNA but not matched cfDNA.

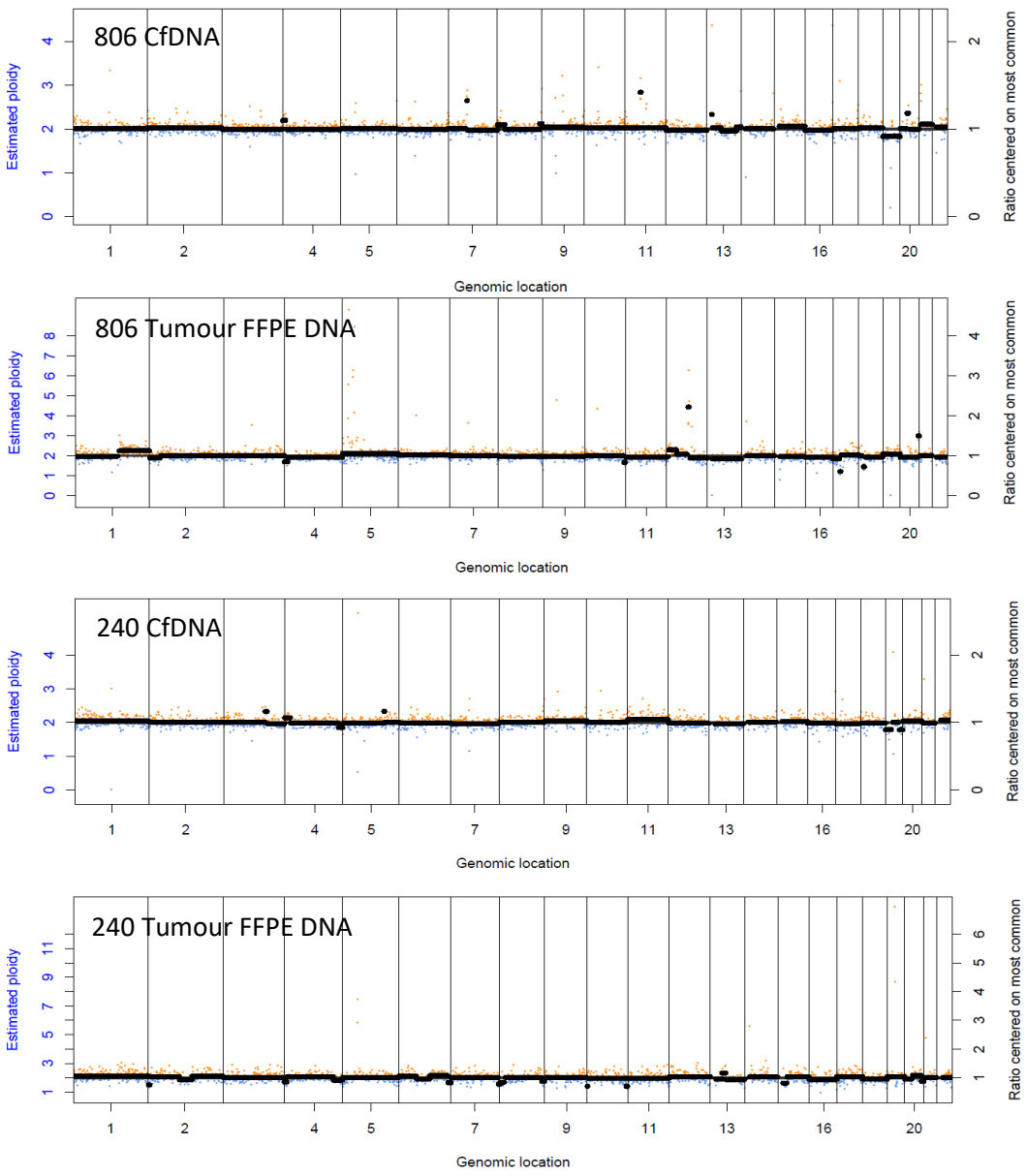


Figure 4-12: Copy number profile graphs for two cases with few copy number aberrations identified in matched tumour FFPE and cfDNA.

Tumour FFPE DNA and matched cfDNA copy number profiles were further compared by comparing the copy number ratios of segments and copy number ratios of 1 Mb windows (Table 4-10). Two cases had good correlation between tumour FFPE DNA and matched cfDNA (203 and 52). This was consistent with the similar pattern of copy number gains and losses observed in copy number profiles. The high number of correlated points >2000 led to high statistical significance for most cases.

Case	CfDNA ng/ml	Length of plasma storage in years	Days from tissue sample to blood collection	Prior treatment to blood collection	Stage	Histology	Tumour FFPE DNA vs cfDNA for copy number ratio segments	Tumour FFPE DNA vs cfDNA for copy number ratio 1 Mb windows
52	**4.7**	**8.3**	**273**	**zometa**	**IV**	**NSCLC-SQ**	**0.68 N=2070**	**0.63 N=2043**
146	15.6	7.5	231	recurrence	IV	NSCLC-AC	-0.20 N=2054	-0.17 N=2044
203	**47.3**	**7.5**	**17**	**chemo**	**IIIB**	**SCLC**	**0.76 N=2070**	**0.78 N=2045**
240	12.1	7.6	12	no	IV	NSCLC-NOS	0.09 N=2084	0.10 N=2044
261	4.7	7.0	0 (same day as surgery)	no	I	NSCLC-AC	0.27 N=2070	0.17 N=2045
527	6.4	5.6	187	recurrence	IV	NSCLC-AC	0.22 N=2072	-0.05 (p=0.04) N=2046
539	2.4	5.5	0 (same day as surgery)	no	IIIA	NSCLC-SQ	-0.15 N=2084	-0.32 N=2040
800	9.8	5.0	22	no	IV	NSCLC-SQ	-0.09 N=2084	-0.23 N=2044
805	**13.9**	**4.9**	**38**	**chemo**	**IV**	**SCLC**	**0.42 N=2086**	**0.33 N=2044**
806	17.1	4.9	153	SRS to brain	IV	NSCLC-mixed	-0.04 (p=0.071) N=2087	-0.29 N=2046

Table 4-10: Spearman's Rank correlations of copy number ratios from 1 Mb windows and segments for the copy number profiles of matched tumour FFPE DNA and cfDNA (N=10).

Unless stated all correlations were significant with $p<0.0001$. Chemo: chemotherapy. SRS: stereotactic radiosurgery.

4.3.1.7.3 Common copy number aberrations found in lung cancer tumours were detected in circulating cell-free DNA samples

Next, it was explored whether common lung cancer CNAs were identified in cfDNA samples by low coverage whole genome sequencing (Table 4-11). The largest studies analysing tumour CNAs for squamous cell carcinoma (N=484) (69), adenocarcinoma (N=660) (69) and small cell carcinoma (N=110) (53) were reviewed to identify the top most common regions of copy number gains and losses. Tumour cell derived cfDNA is further diluted in the blood compared to the tumour due to a background of wild type cfDNA therefore, a copy number ratio greater than 1.10 was chosen to define copy number gain and a copy number ratio less than 0.90 to define copy number loss.

For small cell lung cancer (N=9), 7 out of 9 cases were found to have loss of 3p, which contains the genes *FHIT* and *ROBO1* (53). In addition, gain of 3q (4 of 9) and 5p (7 of 9) and focal loss of 13q (7 of 9) harbouring *RB1* were common CNAs detected. Common CNAs were detected in the cfDNA of squamous cell carcinoma and adenocarcinoma lung cancer cases but at lower proportions compared to small cell lung cancer (Table 4-11).

Interestingly, across all samples there were a high number of chromosome 19 deletions observed to include 25 of 51 untreated cases (49%), 16 of 30 high risk controls (53%) and 2 of 10 (20%) low risk controls (Appendix F). This deletion was not present in H69 cell line DNA but it was present in 2 of 10 (20%) FFPE DNA tumour samples, both of which had multiple aberrations detected (Appendix F)(Figure 10,

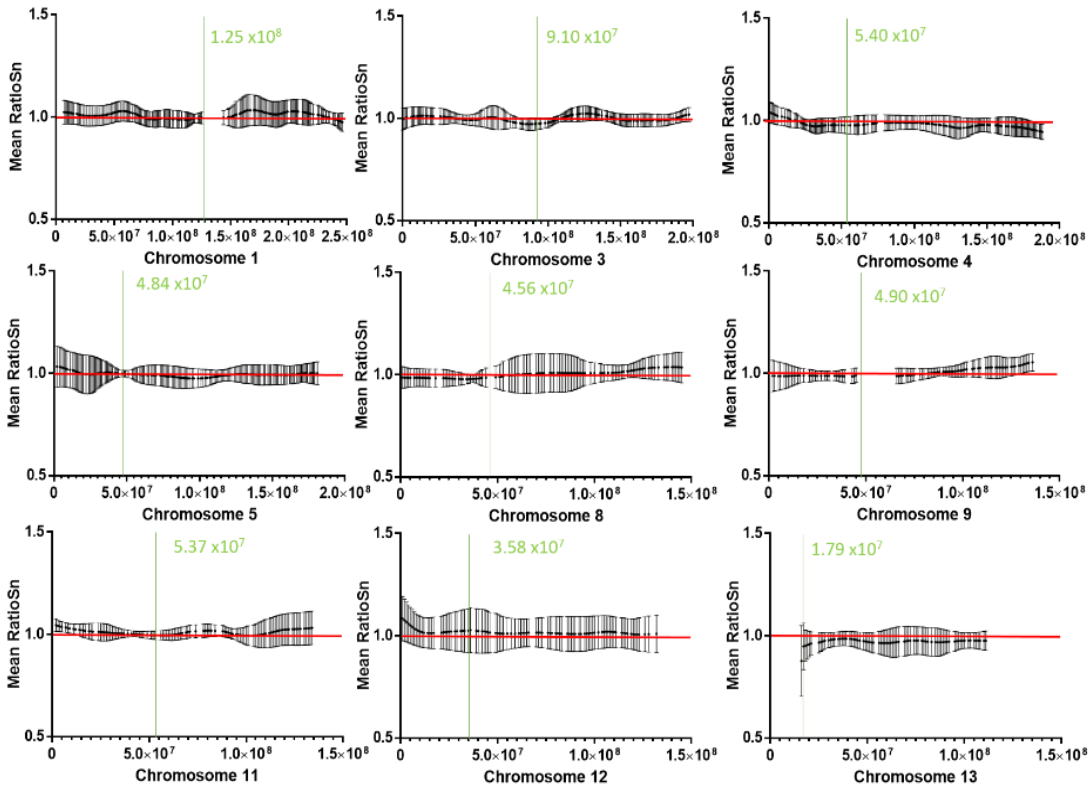
Figure *4-11*, Figure 4-12)

Squamous cell carcinoma (N=17)			Adenocarcinoma (N=23)			Small cell carcinoma (N=9)		
Chromosomal position of CNA		No. cases with CNA detected in cfDNA	Chromosomal position of CNA		No. cases with CNA detected in cfDNA	Chromosomal position of CNA		No. cases with CNA detected in cfDNA
Copy number loss								
9p	0-43000000	1	9p	0-43000000	0	3p	0-90900000	7
8p	0-45200000	2	4q	50000000-190214555	4	3q	90900000-198295559	7
2q	93900000-242193529	0	22q	17400000-50818468	1			
10q	39800000-133797422	1	1p	0-123400000	1			
5q	48800000-181538259	0	15q	19000000-101991189	0			
1p	0-123400000	0	16q	36800000-90338345	0			
3p	0-90900000	2	11q	53400000-135086622	1			
19p	0-26200000	3	13q	17700000-114364328	0			
18q	21500000-80373285	2						
Copy number gain								
3q	90900000-198295559	0	14q	17200000-107043718	0	5p	0-48800000	7
8p	0-45200000	1	8q	45200000-145138636	1	3q	90900000-198295559	4
11q	53400000-135086622	0	5p	0-48800000	1	4q	50000000-190214555	0
8q	45200000-145138636	1	1q	123400000-248956422	2	18q	21500000-80373285	5
7p	0-60100000	0	12p	0-35500000	2	18p	0-80373285	4
4q	50000000-190214555	0	12q	35500000-133275309	2	8q	45200000-145138636	1
2p	0-93900000	0	11q	53400000-135086622	1	8p	0-45200000	1
9p	0-43000000	1	3q	90900000-198295559	0			
19q	26200000-58617616	1	7p	0-60100000	1			
1q	123400000-248956422	0						

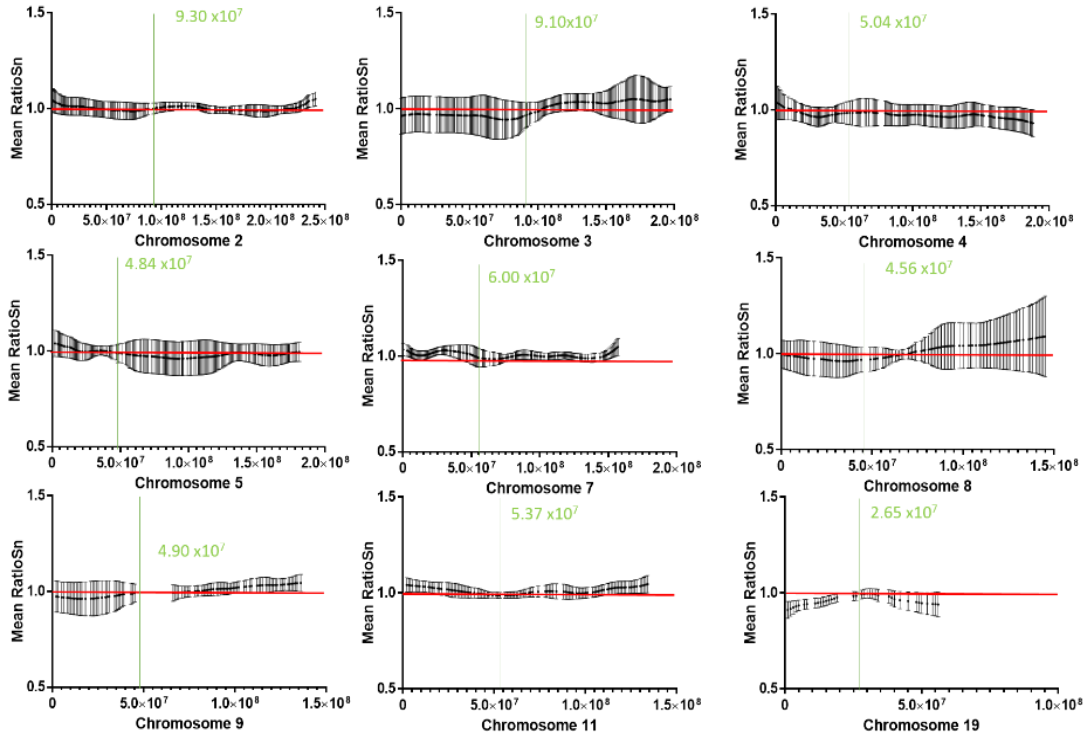
Table 4-11: The most significant CNAs identified from SNP array data in large genomic studies and the number of lung cancer cases with the same CNAs detected in cfDNA samples for the three most common histological subtypes.

CNAs are shown in descending order of significance (determined by q value based on the magnitude and frequency of the aberration across tumour samples).

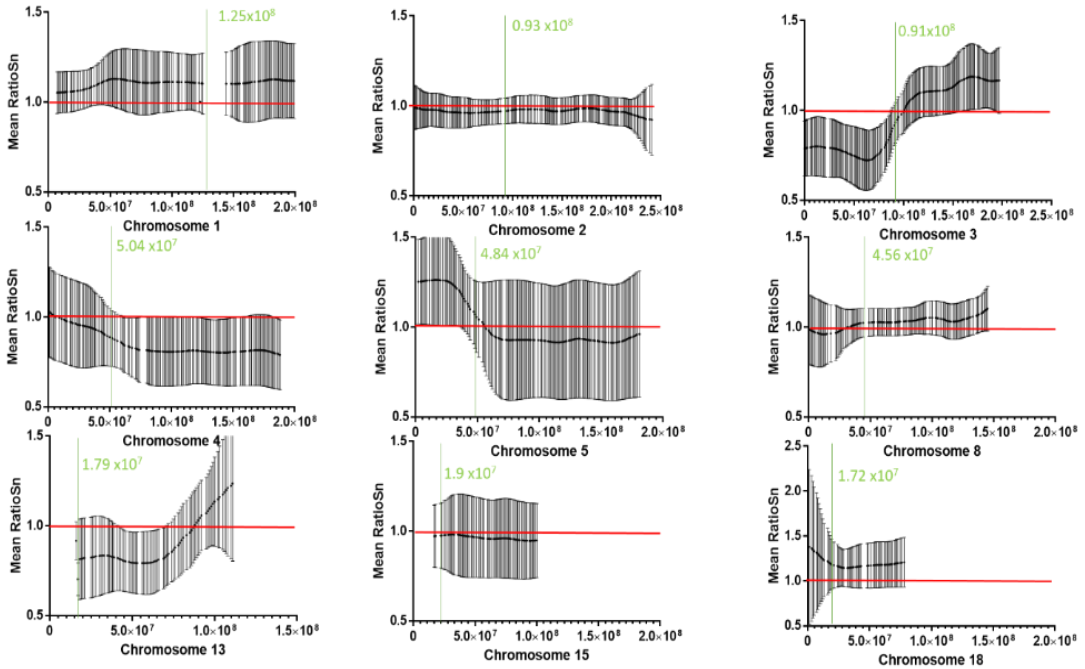
The mean and standard deviation of the copy number ratio of each 1 Mb window for 49 lung cancer cases were plotted against chromosomal position for chromosomes that were known to commonly have CNAs for the three most frequent lung cancer subtypes. Small cell lung cancer cases (N=9) had a greater number and magnitude of CNAs compared to squamous cell carcinoma (N=17) and adenocarcinoma cases (N=23) (Figure 4-13). Although no case with adenocarcinoma had a copy number loss identified for 13q with a copy number ratio less than 0.90, the mean copy number ratio in this region was less than 1.0.



A: Adenocarcinoma (N=23)



B: Squamous cell lung cancer (N=17)



C: Small cell lung cancer (N=9)

Figure 4-13: The identification of common copy number aberrations in cfDNA for the three most frequent histological subtypes of lung cancer.

The mean and standard deviation of the copy number ratio of each 1 Mb window was plotted against chromosomal position for chromosomes known to commonly have CNA. The green line denotes the position of the centromere. The red line highlights the copy number ratio of 1.0 (or ploidy of 2.0).

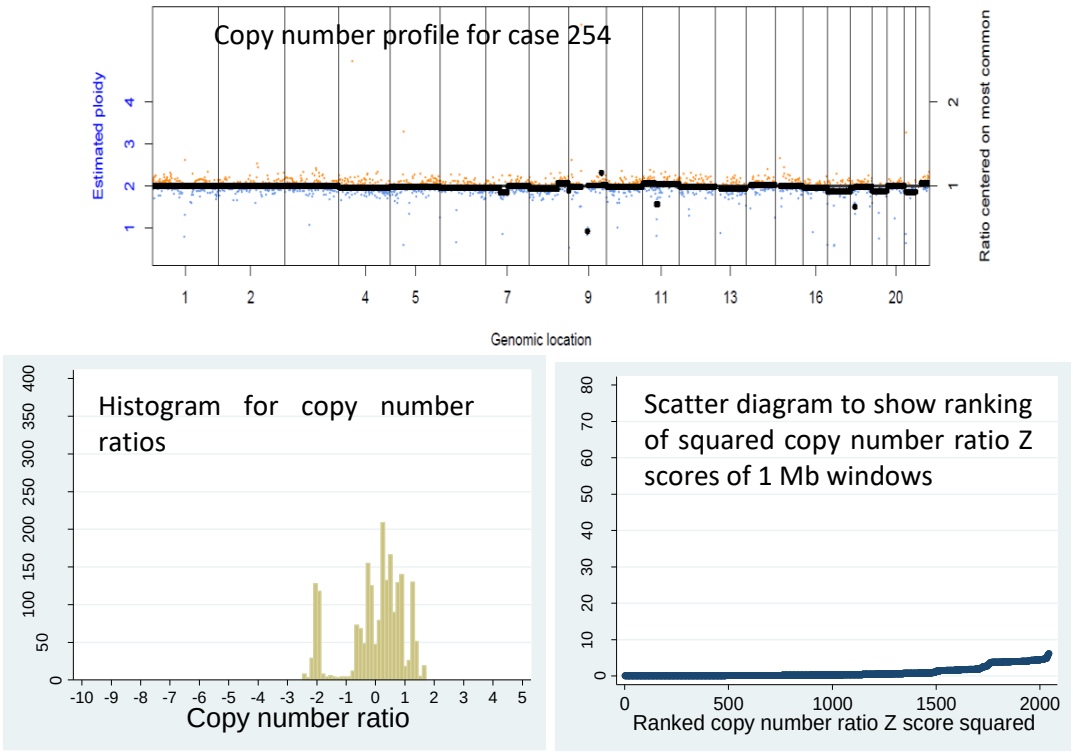
4.3.2 Clinical Validation of circulating cell-free DNA genomic instability scores

Low coverage whole genome sequencing of cfDNA and lymphocyte genomic DNA samples was performed with Illumina HiSeq 2500 (see Section 2.17.1). The sequencing data for 102 selected lung cancer cases and controls was analysed to determine whether two tested genomic instability scores distinguished between lung cancer cases and controls. Selection and characteristics of lung cancer cases and controls are described in Section 3.3.3. Treated cases were removed from analyses because of concerns about introducing bias due to the potential confounding effect of treatment on circulating tumour DNA levels and therefore genomic instability score (see Section 3.3.3.6).

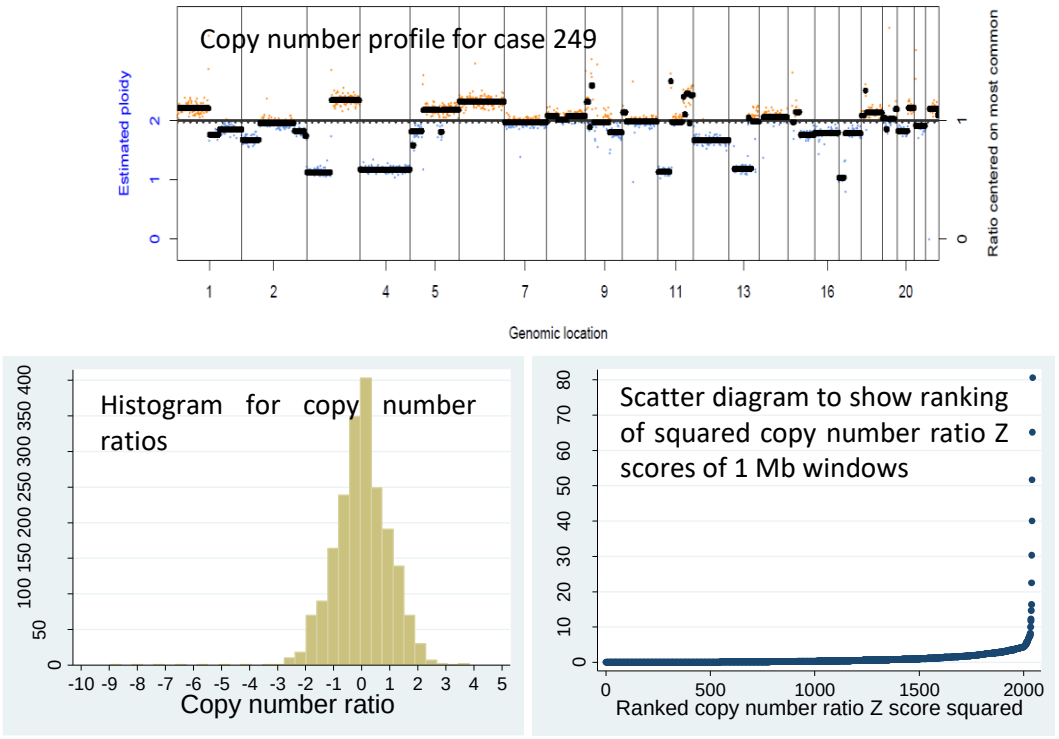
4.3.2.1 The Plasma Genomic Abnormality 2 (PGA2) score

The adapted PGA score (PGA2) was calculated by summing the 95th to 99th centile squared copy number ratio Z scores from 1 Mb sized windows. This adaptation of the Xia et al. method (266) converts gains and losses to positive values to allow both extremes of the distribution to be examined (see Section 2.19.1).

Figure 4-14 displays the copy number ratio profiles, corresponding histogram for copy number ratios and scatter diagram for ranked squared copy number ratios for each 1 Mb window. Low magnitude multiple copy number aberrations were detected for case 249 and the PGA score was 362. In comparison, the copy number profile for case 254 was relatively flat but there were a number of 1 Mb windows with more extreme copy number ratio values and the PGA score was 357.



i) Case 254 stage IIIA NSCLC with few cfDNA CNAs and PGA score 357



ii) Case 249 extensive stage SCLC with multiple cfDNA CNAs and yet the PGA score was 362

Figure 4-14: Two lung cancer cases with similar cfDNA PGA2 scores yet different observed copy number profiles.

4.3.2.1.1 The log_{10} PGA2 score does not correlate with log_{10} circulating cell-free DNA levels

The relationship between the PGA2 score and cfDNA levels was explored on a log_{10} scale because the distribution of the PGA2 score was negatively skewed (see Section 3.3.3.8 for establishment of log_{10} cfDNA levels) for cases (N=51) and controls (N=40) (Figure 4-15). The Pearson's correlation coefficient was -0.06 (p=0.59) indicating that there was no correlation between the log_{10} PGA2 scores and log_{10} cfDNA levels.

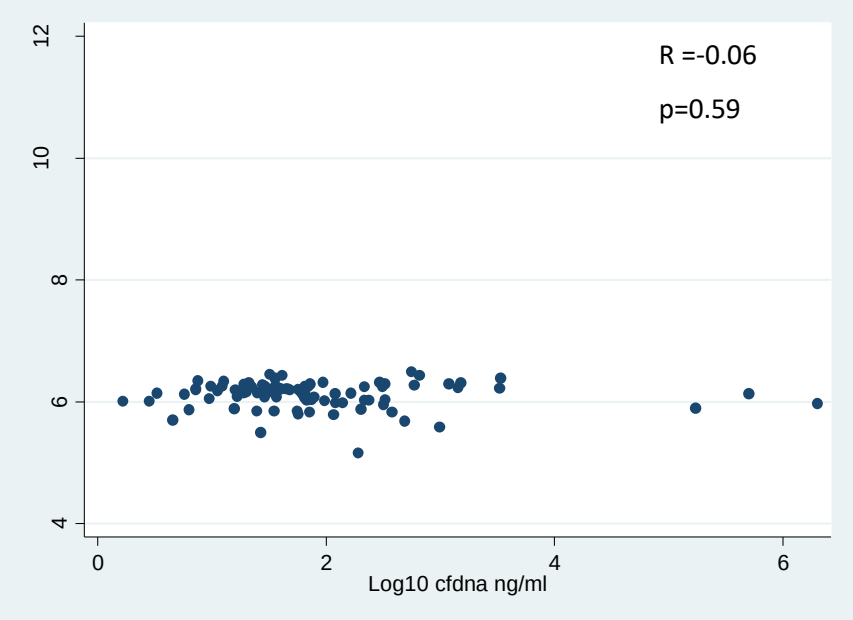


Figure 4-15: A scatter diagram to compare log_{10} PGA2 scores and log_{10} cfDNA levels ng/ml.

Pearson's correlation coefficient R=-0.06 (p=0.59) (N=91).

4.3.2.1.2 There was no difference in PGA2 scores across different disease stages

To assess whether PGA2 scores differed according to disease stage the distribution of scores were compared. There was no significant difference between PGA2 scores across different lung cancer disease stages (p=0.97 Non-parametric test for trend) (Figure 4-16).

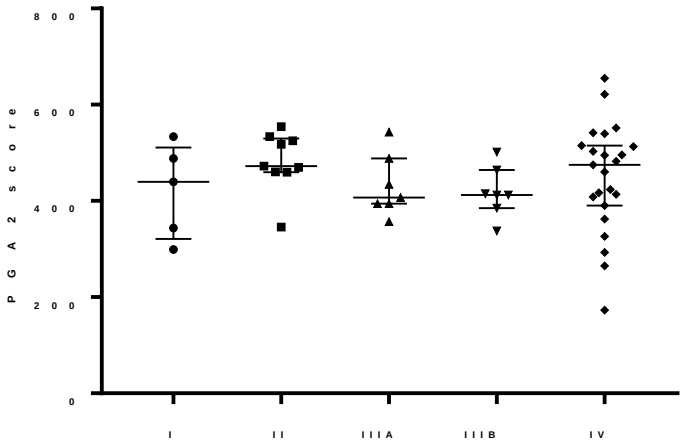


Figure 4-16: PGA2 scores according to the stage of lung cancer.

Non-parametric test for trend, z=0.03 p=0.97 N=51. Median and IQR shown.

4.3.2.1.3 The PGA2 score for high risk controls was higher than the PGA2 score for untreated lung cancer cases and low risk controls

To determine whether the PGA2 score distinguished between untreated lung cancer cases and high risk controls the distribution of scores were compared. The median PGA2 score of the high risk controls (N=30) was significantly higher than the PGA2 score for the low risk controls (N=10), 509 (range 329-628) vs 446 (range 244-570), p= 0.02 Mann Whitney U test. Unexpectedly and against our hypothesis, the median PGA2 score for high risk controls (N=30) was significantly higher than the score for cases with untreated stage I-IIIA (N=21), 460 (range 299-555), p=0.04 Mann Whitney U test, and untreated stage IIIB-IV lung cancer (N=30), 442 (range 173-655), p=0.01 Mann Whitney U test. There was no difference in the median PGA2 score for low risk controls compared to advanced lung cancer cases (p=0.77 Mann Whitney U test) (Figure 4-17).

Summing the 95th to 99th percentile Z score copy number ratios of 1 Mb windows to form a PGA2 score did not perform as expected in our sample set and therefore this score was not further evaluated.

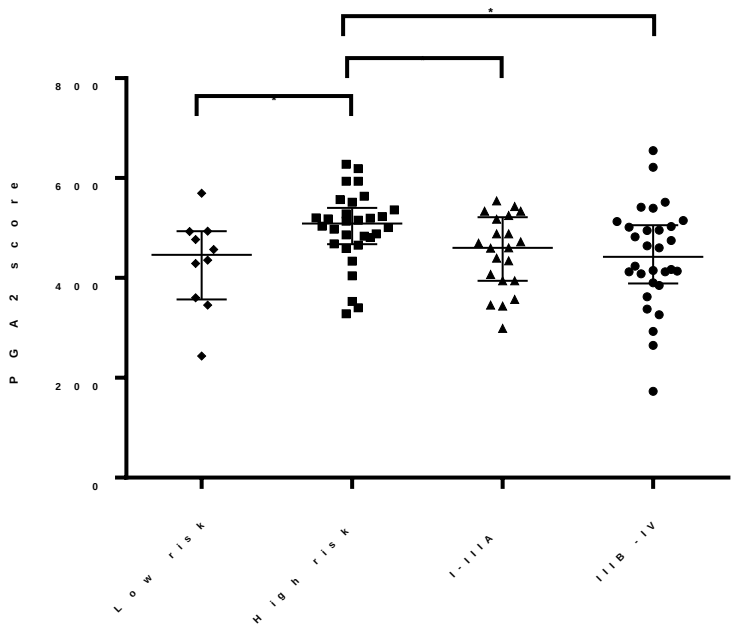


Figure 4-17: PGA2 score for low (N=10) and high risk controls (N=30) and lung cancer cases of stage I-IIIA (N=21) and stage IIIB-IV (N=30).

* $p < 0.05$. Median and IQR shown.

4.3.2.2 The Copy Number Aberration score

An adapted Whole Genome Summed Z score (WGS) (which is now refered to as a Copy Number Aberration (CNA) score) was explored as an unselected measure of genomic instability across the whole genome. Z scores of copy number ratios for each 1 Mb window were created and then the squared Z scores were summed for each 1 Mb window across the genome to create a CNA score (see Section 2.19.2). Since the CNA score was not normally distributed, for most analyses $\log_{10}$ CNA was used (see Section 3.3.3.8 for establishment of $\log_{10}$ cfDNA levels).

4.3.2.2.1 Reference control group

To calculate Z score statistics for each 1 Mb window a reference control group was used. Ten un-related healthy controls at low risk for lung cancer development, with LLP score <2.5%, were used as the reference group. The characteristics of the reference group are displayed in Table 4-12 and their selection from the ReSoLuCENT cohort is described in Section 3.3.3.1.

Characteristic		Low risk controls N=10
Male		3 (30%)
Female		7 (70%)
Age at recruitment	Median (Range)	55.8 years (48-68)
Ethnicity	White British	10 (100%)
Smoking Status	Current	0
	Ex	0
	Never	10 (100%)
	Unknown	0
Status as of August 30th 2016	Alive	9 (90%)
	Dead	1 (10%)

Table 4-12: The characteristics of the low risk healthy control group (N=10).

4.3.2.2.2 Copy Number Aberration scores and circulating tumour DNA allele fractions determined by Ion Torrent targeted sequencing

Six treated cases (222, 291, 338, 765, 1117, 1324) had CNA scores and cfDNA samples that underwent targeted sequencing with the Ion AmpliSeq™ Colon Lungv2 22 gene cancer panel, as part of another project carried out in collaboration with Professor Jacqui Shaw's group at the University of Leicester (see Appendix E for methods(285-288). Five cases had chemotherapy prior to blood withdrawal and one case had palliative radiotherapy to the bone.

Mutations were identified as somatic if they were unique to cfDNA samples and not present in matched genomic DNA after bioinformatics processing. Allele fractions quantify the levels of tumour-derived cfDNA in the circulation (ctDNA) and it was hypothesised that higher CNA scores would be associated with higher ctDNA allele fractions.

4.3.2.2.2.1 Identified somatic mutations in circulating cell-free DNA samples with Copy Number Aberration scores

TP53 missense mutations were the most commonly identified variant unique to cfDNA samples, found in four of six lung cancer cases with allele fractions varying from 1.9% to 77.5%. Two cases had an intronic low frequency SNP identified in *ERBB4* and one case with adenocarcinoma had an additional *EGFR* substitution coding a silent mutation identified (Table 4-13).

Interestingly, one advanced metastatic NSCLC case with a *TP53* allele fraction of 77.5% had very high cfDNA levels (299 ng/ml) and also a very high CNA score of 169,039. Another case with a ctDNA allele fraction of 1.9% (case 338), had a high CNA score of 2696. Although case numbers were very small (N=6), there was no significant correlation between $\log_{10}$ CNA scores and allele fraction (Spearman's Rank correlation coefficient R=0.26, p=0.66)(Figure 4-18).

Case	Stage	Path	CfDNA ng/ml	Unique variants to cfDNA	Gene	COSMIC mutation	CtDNA MAF	Description	CNA score
222	IIIA	SQ	6.4	2	*TP53*	99647 M1441	2.4%	Substitution -missense	983
					ERBB4	-	11.3%	Intronic low frequency SNP	
291*	IV	NOS	299	1	*TP53*	43635 *H179L*	77.5%	Substitution -missense	169,039
338	IV	SQ	10.7	1	*TP53*	9022 *R175H*	1.9%	Substitution -missense	2696
765	IV	AC	10.1	2	*ERBB4*	-	4.6%	Intronic low frequency SNP	4808
					EGFR	-	5.2%	Substitution -coding silent	
1117	IIIB	AC	17.1	1	*TP53*	44142 *Y126S*	2.3%	Substitution -missense	1377
1324	IIIA	SQ	13.8	1	*TP53*	Novel CG>GA R248E	3.5%	In-frame dinucleotide change. Missense	2187

Table 4-13: Clinical characteristics, CNA scores and detected cfDNA mutations using the Ion Torrent Platform for six lung cancer cases.

COSMIC: Catologue of Somatic Mutations in Cancer (60). MAF: mutant allele fraction. * this case was treated with palliative radiotherapy to the bone prior to blood withdrawal.

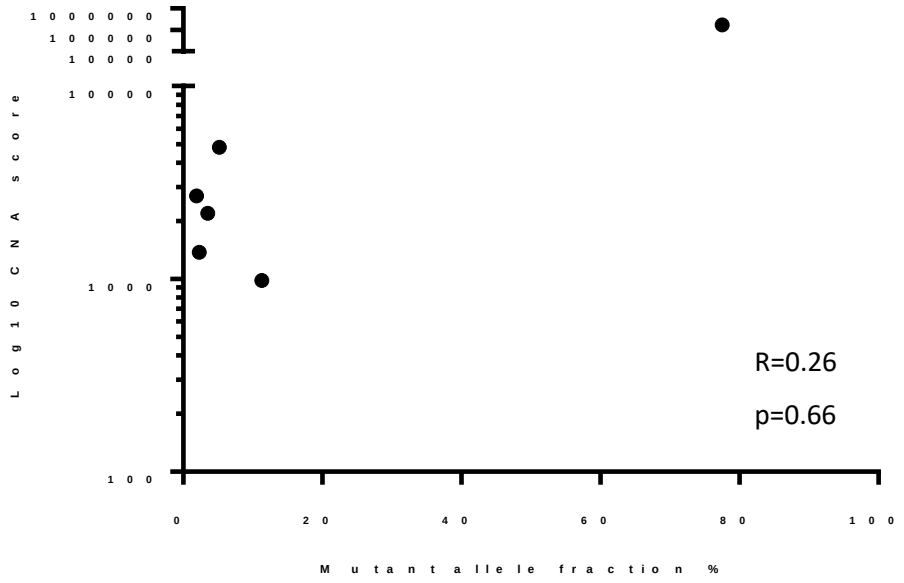


Figure 4-18: Scatter diagram for log_{10} CNA scores and allele fraction determined by Ion Torrent targeted sequencing.

Spearman's Rank correlation coefficient R=0.26 p=0.66

4.3.2.2.3 Log_{10} Copy Number Aberration scores and log_{10} circulating cell-free DNA levels were correlated

To explore the relationship of log_{10} CNA and log_{10} cfDNA levels they were correlated for cases (N=51) and controls (N=30). Log_{10} CNA were positively correlated with log_{10} cfDNA levels ng/ml (Pearson's correlation coefficient R=0.58, p<0.0001)(Figure 4-19).

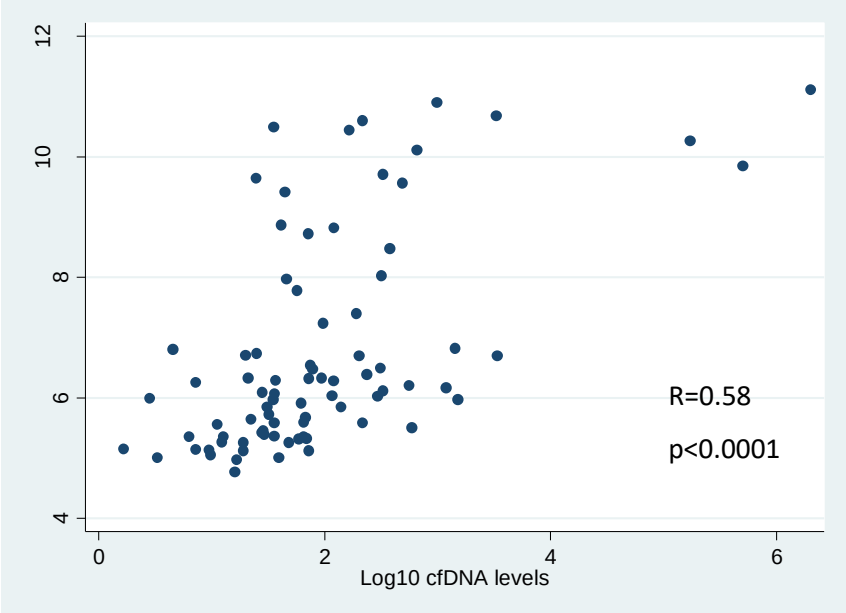


Figure 4-19: A scatter diagram to show the correlation of log_{10} CNA scores with log_{10} cfDNA levels ng/ml for lung cancer cases (N=51) and controls (N=30).

R=0.58, p<0.0001 Pearson's correlation coefficient

4.3.2.2.4 Copy Number Aberration scores were higher for advanced stage cancer compared to early stage cancer cases

To assess whether CNA scores increased with tumour burden the distribution of scores across disease stages were compared. The median CNA score differed significantly across disease stages for untreated lung cancer cases from stage I to stage IV (p<0.001 Non-parametric test for trend)(Figure 4-20). The median CNA score for lung cancer cases with stage I disease was 225 (range 117-904) compared to the median CNA score for cases with stage IV disease 1389 (range 167-66,869).

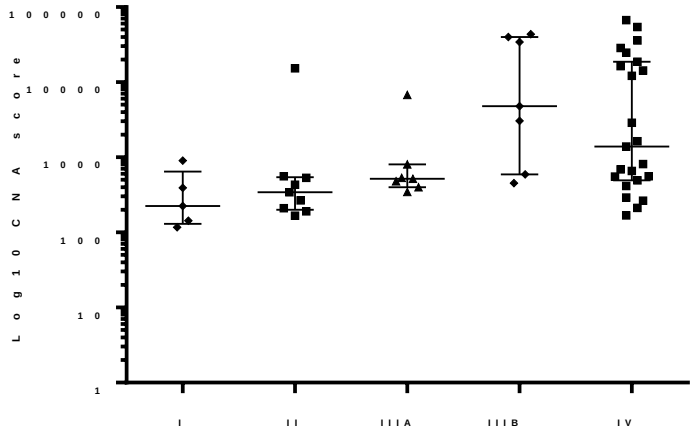


Figure 4-20: CNA scores for lung cancer cases according to disease stage (N=51).

Median and IQR are shown , Non-parametric test for trend z=3.53 p<0.001.

4.3.2.2.5 Small cell lung cancer cases had the highest Copy Number Aberration scores compared to other histological subtypes

The relationship between the CNA score, disease stage and histological subtype were explored. Cases with extensive stage SCLC (N=7) had higher median CNA scores compared to cases with advanced non-squamous NSCLC (N=17) and advanced squamous NSCLC (N=6), median 35,996 (range 14,277-66,869) vs 592 (range 169-43445), p=0.008 and 1512 (560-34,275), p=0.03 Mann Whitney U test (Figure 4-21).

The median CNA score was significantly higher for non-squamous NSCLC in advanced stage cases (N=17) compared to early stage cases (N=14) (median 592 (range 169-43,445) vs 369 (range 117-15,373), p=0.04 Mann Whitney U test). There was a borderline statistically significant difference for the higher median CNA score for advanced squamous NSCLC cases (N=6) compared to the median CNA score for early stage cases (N=7) (median 1512 (range 560-34,275) vs 520 (range 192-6757), p=0.05 Mann Whitney U test).

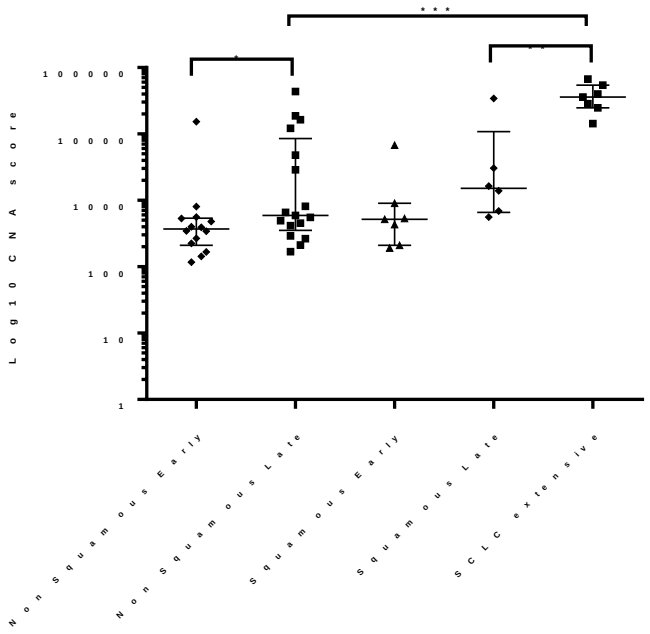


Figure 4-21: CNA score according to histological subtype and disease stage.

Early: stage I-IIIA. Late: stage IIIB-IV. Mann Whitney U test. * p <0.05, ** p<0.005, *** p<0.0005. Median and IQR are shown.

4.3.2.2.6 Copy Number Aberration scores were higher for lung cancer cases compared to high risk controls

To establish whether CNA scores differentiated between lung cancer cases (N=51) and high risk controls (N=30), the distribution of scores were compared. The median CNA score for lung cancer cases (stage I-IV, N=51) was significantly higher than the median CNA score for high risk controls (N=30), 559 (range 117-66,869) compared to 252 (range 149-7122), p value=0.0002 Mann Whitney U test (Figure 4-22).

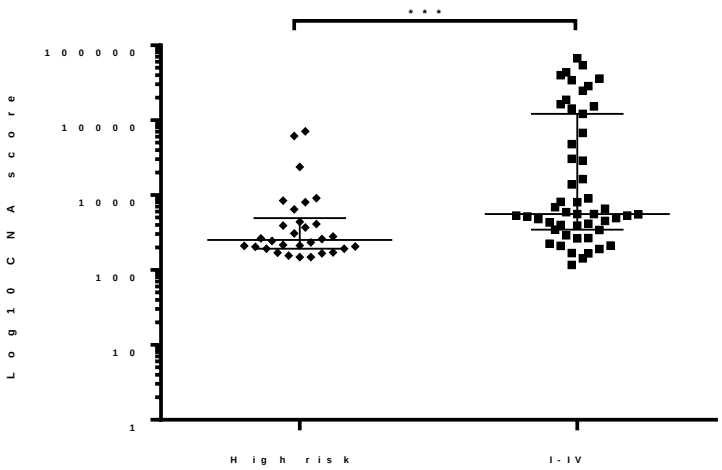


Figure 4-22: CNA score for high risk controls (N=30) and lung cancer cases (N=51).

*** p<0.001. Median and IQR are shown.

4.3.2.2.7 Log_{10} Copy Number Aberration scores predicted status for lung cancer cases and high risk controls in univariable and multivariable analyses

Univariable and multivariable logistic regression were carried out to establish whether the CNA score predicted case-control status for lung cancer cases (N=51) and high risk controls (N=30) (see Section 2.20.3.1). Only 3 out of 51 cases had no history of smoking and all high risk controls were ex or current smokers, therefore smoking was not included as a variable in this analysis.

In univariable analysis, the variables log_{10} CNA (OR 1.89 p=0.004), log_{10} cfDNA levels (OR 2.01 p=0.04) and age at study registration (OR 0.71 p=0.002) were significant. Whilst, length of plasma storage (OR 0.81 p=0.06) and gender (OR 2.42 p=0.07) were of borderline significance. In multivariable analysis, after adjusting for length of plasma storage and log_{10} cfDNA levels, log_{10} CNA (OR 2.16 p=0.03), age at study registration (OR 0.64 <0.0001) and gender (OR 5.41 p=0.02) were significant variables (Table 4-14). Similar to results described in Section 3.3.3.8, younger age and being female were associated with higher lung cancer risk in multivariable analysis. This is a consequence of the eligibility criteria for ReSoLuCENT.

	Odds ratio (95% CI)	P value
Univariable analysis		
Log_{10} CNA score	1.89 (1.23-2.88)	0.004
Log_{10} cfDNA ng/ml	2.01 (1.05-3.85)	0.04
Length of plasma storage at -80°C	0.81 (0.65-1.00	0.06
Age at study registration	0.71 (0.57-0.88)	0.002
Gender (comparator male)	2.42 (0.93-6.34)	0.07
Multivariable analysis		
Log_{10} CNA score	2.16 (1.07-4.34)	0.03
Log_{10} cfDNA ng/ml	2.46 (0.95-7.16)	0.10
Length of plasma storage at -80°C	1.07 (0.71-1.62)	0.74
Age at study registration	0.64 (0.50-0.82)	<0.0001
Gender (comparator male)	5.41 (1.26-23.21)	0.02

Table 4-14: Univariable and multivariable logistic regression analyses for lung cancer cases stage I-IV (N=51) and high risk controls (N=30) to evaluate the relationship of different factors for predicting case-control status.

The relationship between log_{10} CNA score and lung cancer was explored by grouping this variable into quintiles. When univariable logistic regression was carried out, the chance of detecting a lung cancer case broadly increased with quintile except for individuals ranked in the fourth quintile (rank 49-64). However, there remained a significant difference between the fourth quintile compared to the reference group (1-16) with the lowest ranked log_{10} CNA scores (Table 4-15).

	Odds ratio (95% CI)	P value
Univariable analysis		
Log_{10} CNA score five groups (comparator 1-16)		
17-32	2.2 (0.52-9.30)	0.28
33-48	6.6 (1.40-31.05)	0.02
49-64	4.84 (1.09-21.58)	0.04
65-81	16.5 (2.69-101.33)	0.002
Multivariable analysis		
Log_{10} CNA score five groups (comparator 1-16)		
17-32	1.82 (0.31-10.66)	0.51
33-48	3.27 (0.49-21.92)	0.22
49-64	8.31 (1.13-61.34)	0.04
65-81	12.03 (0.45-322.93)	0.14
Log_{10} cfDNA ng/ml	2.32 (0.77-7.07)	0.14
Length of plasma storage at -80°C	1.06 (0.75-1.49)	0.75
Age at study registration	0.64 (0.48-0.87)	0.004
Gender (comparator male)	5.42 (1.31-22.43)	0.02

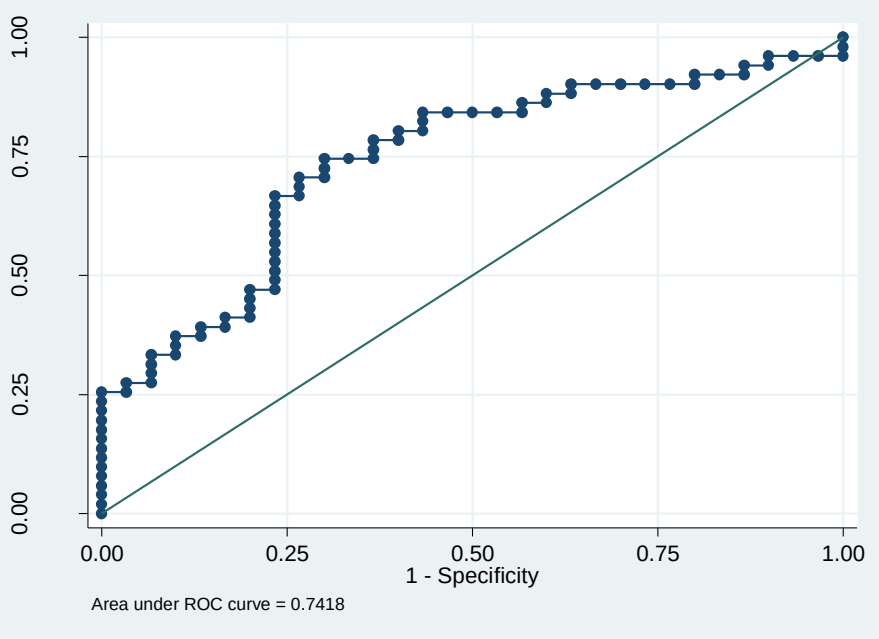
Table 4-15: Univariable and multivariable logistic regression carried out for ranked and grouped log_{10} CNA scores (N=81) to evaluate the relationship of different factors for predicting case-control status.

When multivariable logistic regression was performed for log_{10} CNA subdivided into two ranked groups (based on the significance of the p value in univariable analysis), log_{10} CNA score (OR 5.19 p=0.023), age (OR 0.66 p=0.005) and gender (OR 4.55 p=0.036) were significant predictors. However, log_{10} cfDNA levels (OR 2.31 p=0.13), length of plasma storage (OR 1.06 p=0.72) were not significant predictors of case or control status (Table 4-16).

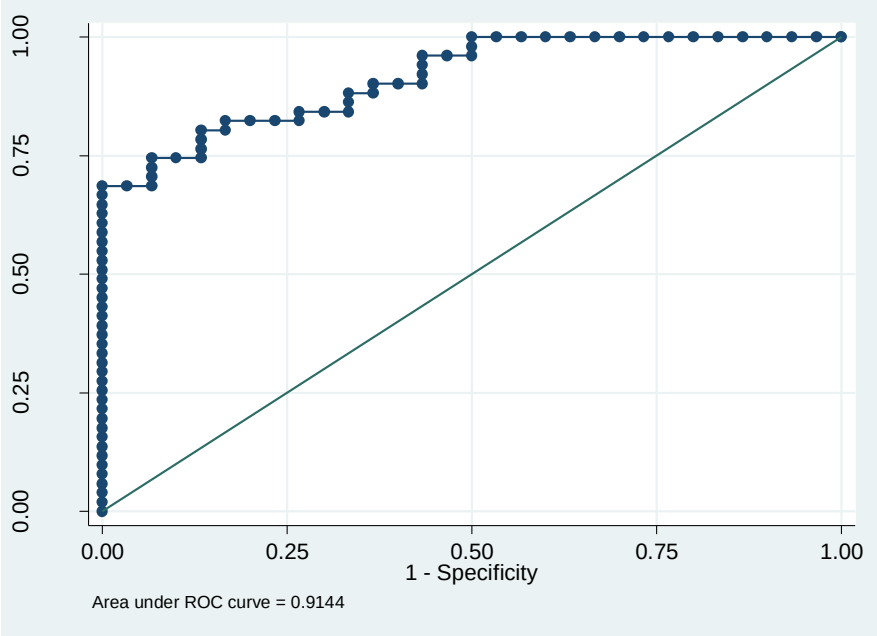
	Odds ratio (95% CI)	**P value**
Univariable analysis		
Log_{10} CNA score two groups 1-32 compared to 33-81	5.85 (2.17-15.77)	<0.0001
Multivariable analysis		
Log_{10} CNA score groups 1-32 compared to 33-81	5.19 (1.26-21.45)	0.023
Log_{10} cfDNA ng/ml	2.31 (0.78-6.78)	0.13
Length of plasma storage at -80°C	1.06 (0.76-1.48)	0.72
Age at study registration	0.66 (0.49-0.88)	0.005
Gender	4.55 (1.11-18.71)	0.036

Table 4-16: Univariable and multivariable logistic regression carried out for ranked and grouped log_{10} CNA scores based on the p value of univariable quintile analyses (N=81) to evaluate the relationship of different factors for predicting case-control status.

ROC curve analyses using the predicted probability of being a case (based on the logistic regression) were performed to compare AUC for the discriminatory ability of the log_{10} CNA score as a continuous variable to distinguish between high risk controls (N=30) and untreated lung cancer cases (N=51) (Figure 4-23). The log_{10} CNA score had good discriminatory ability to differentiate high risk controls (N=30) and untreated lung cancer cases with stage I-IV disease (N=51). The univariable AUC was 0.74 (95% CI: CI 0.63-0.85). The AUC for the log_{10} CNA was greater than the AUC for log_{10} cfDNA levels, which was 0.67 (95% CI: 0.54-0.79), although the 95% confidence intervals overlapped. The AUC was not improved by the combination of log_{10} cfDNA levels and log_{10} CNA scores (AUC 0.74 95% CI 0.63-0.85). The AUC for the multivariable model was very high 0.91 (95% CI 0.86-0.97), demonstrating excellent discriminative ability when multiple variables were combined (Figure 4-23).



i: Univariable model log_{10} CNA (AUC 0.74 95% CI 0.63-0.85)



ii: Multivariable model including log_{10} CNA, age, gender, log_{10} cfdna levels, length of plasma storage (AUC 0.91 95% CI 0.86-0.97)

Figure 4-23: ROC curves for univariable and multivariable models for untreated lung cancer cases (N=51) compared to high risk controls (N=30) to establish the role of log_{10} CNA in predicting case or control status

4.3.2.2.8 Identifying a cut off for the log_{10} Copy Number Aberration score for untreated lung cancer cases and high risk controls

To evaluate the potential clinical usefulness of the CNA score different cut offs were explored to determine the chance that individuals scoring above the cut off were correctly identified as cases (sensitivity) and that individuals scoring below the cut off were correctly identified as controls (specificity). The ROC AUC was 0.74, indicating good discriminatory ability for untreated lung cancer cases (N=51) and high risk controls (N=30). A log_{10} CNA score cut-off of 6.03 gave the best balance of sensitivity 71% and specificity 73% (Table 4-17). A lower cut-off of 5.35 increased sensitivity to 90% and therefore a higher proportion of cases with cancer would be correctly identified, but at the cost of reducing the specificity to 37% and therefore increasing the false positive rate. For our screening test, a high sensitivity is preferred to reduce the false negative rate and therefore minimise the chance of missing a true case.

Log_{10} CNA score cut-off	**Sensitivity %**	**Specificity %**	**Likelihood ratio**
5.25	92	20	1.15
5.35	90	37	1.42
5.73	78	60	2.00
6.03	71	73	2.65
6.33	51	77	2.18
6.70	42	83	2.47
6.82	37	87	2.79
7.24	37	90	3.73
7.96	33	93	5.00

Table 4-17: Examples of different cut-offs for the log_{10} CNA score and corresponding sensitivity, specificity and likelihood ratio values (N=81).

The likelihood ratio defines how much more likely it is that an individual that tests positive has cancer compared to an individual that tests negative.

4.3.2.2.9 There was no difference between the Copy Number Aberration score for early stage cancer cases and high risk controls

To be useful as a potential screening tool, the CNA score needs to differentiate between early lung cancer cases and high risk controls and therefore the distribution of CNA scores were compared between the two groups. There was no significant difference between the median CNA score for high risk controls (N=30) and cases with early stage (I-IIIA, N=21) cancer, 252 (range 149-7122) vs 398 (range 117-15,373), p=0.25 Mann Whitney U test. Advanced stage (IIIB-IV, N=30) cases had a higher median CNA score compared to high risk controls (N=30),

2256 (range 169-66,869) compared to 252 (range 149-7122) respectively p<0.0001 Mann Whitney U test (Figure 4-24).

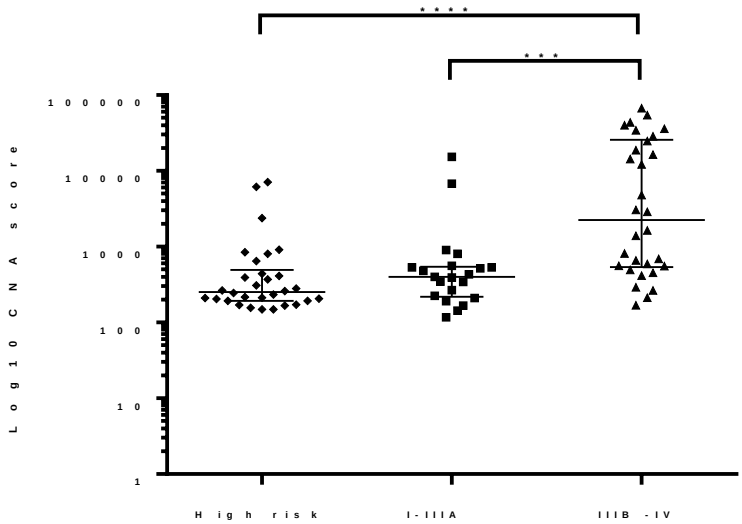


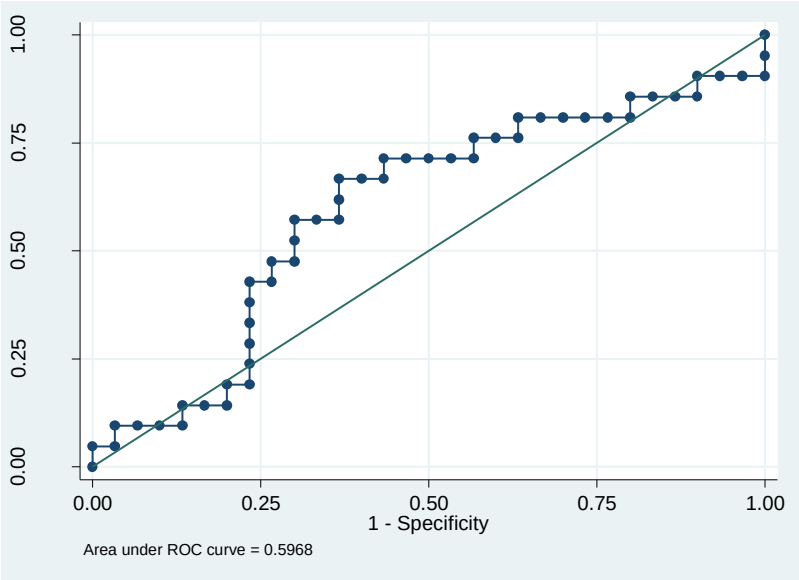
Figure 4-24: CNA score calculated from 1 Mb windows for high risk controls (N=30) and early (stage I-IIIA N=21) and advanced (stage IIIB-IV N=31) lung cancer cases (N=51).

*** p <0.001, **** p<0.0001. Median and IQR are shown.

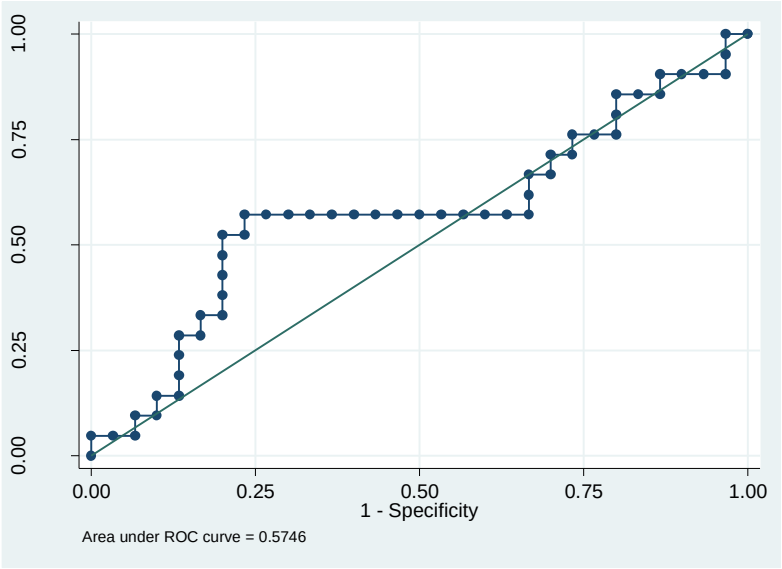
4.3.2.2.10 Log_{10} Copy Number Aberration scores were not predictive of early lung cancer cases compared to high risk controls in univariable analysis

Logistic regression was carried out to evaluate whether log_{10} CNA scores predicted case or control status when early cancer cases (stage I-IIIA, N=21) and high risk controls (N=30) were compared. The log_{10} CNA score was not a significant predictive factor in univariable analysis (OR 1.25 (95% CI 0.74-0.21) p=0.40), and neither were, length of plasma storage (OR 0.82 (95% CI 0.65-1.04) p=0.82) or log_{10} cfDNA levels (OR 0.95 (95% CI 0.42-2.16) p=0.91). Age at diagnosis was a significantly significant predictive factor (OR 0.60 (95% CI 0.43-0.84 p=0.003) and gender was a borderline significant factor (OR 3.11 (95% CI 0.96-10.09), p=0.06).

Consistent with these results, ROC analysis showed that there was poor discriminatory ability when only considering the log_{10} CNA scores of lung cancer cases with early stage (I-IIIA) disease (N=21) compared to the CNA scores of high risk controls (N=30), AUC 0.60 (95% CI 0.43-0.76). Combining log_{10} CNA scores with log_{10} cfDNA levels did not improve discriminatory ability and resulted in an AUC of 0.57 (0.40-0.75)(Figure 4-25)(Table 4-18).



i: Log_{10} CNA (AUC 0.60 95% CI 0.43-0.76)



ii. Log_{10} CNA and log_{10} cfDNA (AUC 0.57 95% CI 0.40-0.75)

Figure 4-25: ROC curves for log_{10} CNA alone and combined with log_{10} cfDNA for untreated early lung cancer (I=IIIA, N=21) compared to high risk controls (N=30) to establish their role in predicting case or control status.

Factor	ROC-AUC for stage I-IV (N=51) (95% CI)	ROC-AUC for early stage I-IIIA (N=21) (95%CI)	ROC-AUC for late stage IIIB-IV (N=30) (95%CI)
Log_{10} CNA	0.74 (0.63-0.85)	0.60 (0.44-0.72)	0.84 (0.73-0.93)
Log_{10} cfDNA ng/ml	0.67 (0.54-0.79)	0.53 (0.38-0.67)	0.76 (0.64-0.87)
Log_{10} CNA and log_{10} cfDNA ng/ml	0.74 (0.63-0.85)	0.57 (0.40-0.75)	0.84 (0.74-0.94)

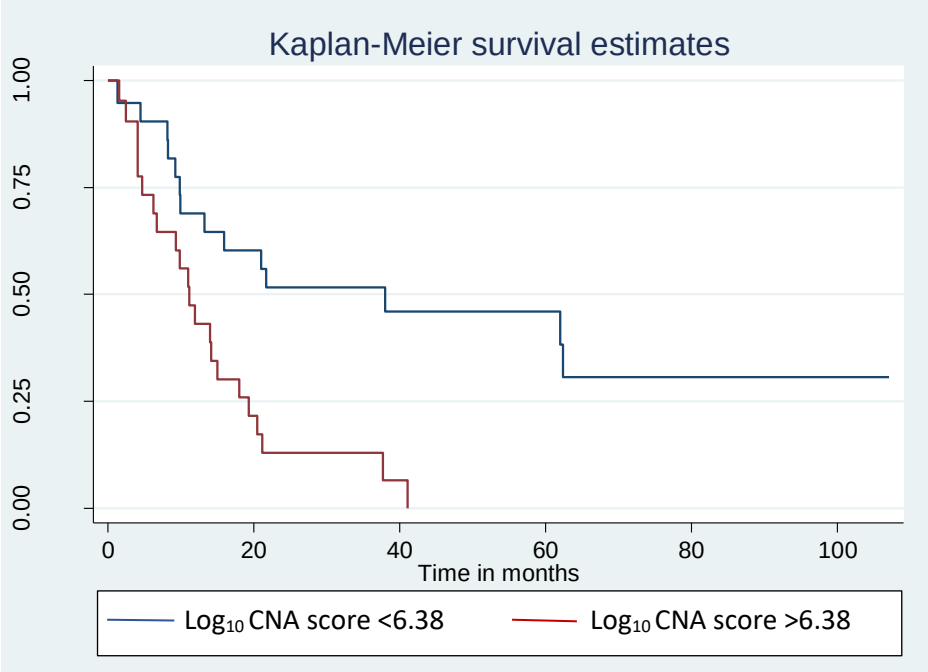
Table 4-18: ROC analyses demonstrating area under the curve for high risk controls (N=30) and untreated lung cancer cases (N=51).

4.3.3 Log_{10} Copy Number Aberration score as a prognostic tool

To evaluate whether the log_{10} CNA score was a prognostic biomarker for lung cancer, HSCIC data with medical record information was used to determine the date that a patient was last known to be alive or date of death, and analyses were adjusted according to time from diagnosis to blood sampling (see Section 2.20.4). All untreated cases were included (N=51), with the exclusion of two cases for whom the date of disease recurrence was unknown. Smoking was not assessed as a prognostic variable because only 3 of 49 cases had never smoked.

4.3.3.1 Lung cancer cases with a log_{10} Copy Number Aberration score higher than the median score had shorter survival

Survival time was shorter for lung cancer cases (N=49) with a log_{10} CNA score greater than the median score of 6.38 compared to cases with log_{10} CNA score less than the median, 38.01 months (95% CI 9.83-not available) vs 11.11 months (95% CI 4.70-15.02) respectively. The hazard ratio for death (logrank) was 3.13 (95% CI 1.54-6.35), p=0.0009. The Kaplan-Meier survival curve is displayed in Figure 4-26.



Time in months	0	20	40	60	80	100
No. at risk log_{10} CNA <6.38	24	14	8	6	3	2
No. at risk log_{10} CNA >6.38	25	5	1	0	0	0

Figure 4-26: Kaplan-Meier survival curve for untreated lung cancer cases (N=49) with log_{10} CNA greater or less than the median CNA value of 6.38.

4.3.3.2 Log_{10} CNA score was a prognostic factor in univariable but not multivariable analyses

Cox regression survival analyses were carried out to establish the prognostic value of important variables (see Section 2.20.4.1). The following variables were significant predictors of survival in univariable analysis; disease stage (early I-IIIA or late IIIB-IV) (HR 5.10 (95% CI 2.28-11.37) p<0.001), log_{10} CNA score (HR 1.22 (95% CI 1.05-1.42) p=0.008), gender (HR 0.48 (95% CI 0.25-0.94) p=0.03) and log_{10} cfDNA levels (HR 1.38 (95% CI 1.02-1.86) p=0.04). Whilst, performance status (0/1 or 2/3) (HR 2.43 (95% CI 0.93-6.36) p=0.07), age at diagnosis (HR 1.05 (95% CI 1.00-1.12) p=0.07), were of borderline significance and, length of plasma storage (HR 1.11 (95% CI 0.95-1.31) p=0.18) and histology (NSCLC non squamous vs NSCLC squamous vs SCLC (HR 1.15 (95% CI 0.52-2.51) p=0.74 and HR 1.82 (95% CI 0.89-3.74) p=0.30) were not significant in univariable analysis. Variables with p value <0.25 were included in the final multivariable model.

Disease stage (early I-IIIA or late IIIB-IV)(HR 9.12 p<0.001) and performance status (0/1 or 2/3)(HR 6.79 p=0.002) were significant predictors for death in multivariable analyses after adjusting for gender, log_{10} CNA score, log_{10} cfDNA levels and age at diagnosis (Table 4-19). There were no significant interactions that required adjustment for in the multivariable models, and the assumption of proportional hazards was met.

Factor	Hazard Ratio (95% CI)	P value
Log_{10} CNA score	0.91 (0.73-1.15)	0.45
Log_{10} cfDNA ng/ml	1.15 (0.76-1.75)	0.52
Length of plasma storage	1.11 (0.92-1.33)	0.27
Age at diagnosis	1.01 (0.92-1.10)	0.15
Gender (comparator male)	0.52 (0.25-1.07)	0.08
Stage (comparator early stage)	9.12 (2.96-27.98)	<0.001
Performance status (comparator PS 0 or 1)	6.79 (2.03-22.71)	0.002

Table 4-19: Hazard Ratios for variables tested in multivariable cox regression survival analyses of untreated lung cancer cases (N=49).

The relationships between log_{10} CNA scores were explored by grouping the variable into quintiles. When cox regression survival analyses were carried out after ranking and grouping cases into five subgroups, the hazard ratios increased as the log_{10} CNA scores increased but the hazard ratio was only statistically significant for the two groups with the highest ranked log_{10} CNA scores, suggesting that only very high scores impact on survival (Table 4-20).

When cases were split into two subgroups according to the statistical significance of the ranked quintile group log_{10} CNA score, for univariable analysis the hazard ratio was 2.14 p=0.02. However, in multivariable analysis when adjusting for gender, age at diagnosis, length of plasma storage and log_{10} cfDNA levels, log_{10} CNA (HR 0.66 p=0.48) remained non-significant and disease stage (HR 8.80 p<0.001) as well as performance status (HR 6.97 p=0.002) remained significant prognostic factors (Table 4-20).

	Hazard Ratio (95% CI)	P value
Univariable analysis		
Ranked and grouped by quintile log_{10} CNA score (comparator 1-10)		
11-20	1.80 (0.51-6.35)	0.36
21-30	2.55 (0.89-7.29)	0.08
31-40	2.81 (1.06-7.47)	0.04
41-49	4.46 (1.54-12.89)	0.006
Multivariable analysis		
Log_{10} CNA score grouped by quintile (comparator 1-10)		
11-20	1.61 (0.42-6.17)	0.49
21-30	1.23 (0.40-3.83)	0.72
31-40	1.08 (0.33-3.48)	0.90
41-49	0.75 (0.17-3.21)	0.70
Log_{10} cfDNA ng/ml	1.14 (0.78-1.68)	0.49
Length of plasma storage	1.13 (0.92-1.38)	0.23
Age at diagnosis	1.00 (0.93-1.07)	0.94
Gender (comparator male)	0.58 (0.26-1.32)	0.20
Stage (comparator early stage)	9.97 (2.17-45.71)	0.003
Performance status (comparator PS 0 or 1)	7.20 (1.75-29.61)	0.006
Univariable analysis		
Ranked and grouped log_{10} CNA score (comparator 1-30) 31-49	2.14 (1.15-3.97)	0.02
Multivariable analysis		
Log_{10} CNA score (comparator 1-30) 31-49	0.66 (0.21-2.05)	0.48
Log_{10} cfDNA ng/ml	1.15 (0.75-1.76)	0.52
Length of plasma storage	1.14 (0.94-1.38)	0.19
Age at diagnosis	1.00 (0.91-1.09)	0.93
Gender (comparator male)	0.53 (0.26-1.08)	0.08
Stage (comparator early stage)	8.80 (2.91-26.42)	<0.001
Performance status (comparator PS 0 or 1)	6.97 (2.05-23.74)	0.002

Table 4-20: Hazard ratios for ranked and grouped log_{10} CNA scores for univariable and multivariable cox regression analyses for untreated lung cancer cases (N=49).

4.4 Discussion

In this pilot study, CNAs were detected in cfDNA of lung cancer cases by low coverage whole genome sequencing with Illumina HiSeq 2500, an unbiased and cost effective method. CfDNA samples were included from the three most common histological types of lung cancer to mimic the general lung cancer population. As expected, a higher number of CNAs with greater magnitude were identified in SCLC cfDNA samples compared to adenocarcinoma and squamous cell carcinoma cfDNA samples.

4.4.1 Analytical performance and validation

4.4.1.1 Tumour FFPE and circulating cell-free DNA samples

For the three main histological subtypes of lung cancer, known common tumour CNAs in cfDNA samples were identified for some but not all cases. Tumour-derived cfDNA is diluted by wild type DNA in the blood (115). Copy number gain and loss were defined by setting an arbitrary copy number ratio threshold (see Section 4.3.1.7.3). CNAs may have been missed if the set threshold was too high but a lower threshold may increase false positives. Alternatively, Z scores of targeted regions could be calculated to more accurately define copy number gains and losses. This approach still relies on setting a threshold but takes into account population differences by comparison to a healthy control group (168, 178). Even though only a small number of copy number aberrations exceeded the set threshold, the mean copy number ratio for a certain genomic region was often less than or greater than 1.0. Furthermore, resolution to identify focal CNAs in our samples was limited and focal aberrations are more common in NSCLC compared to SCLC (66).

CNAs detected in tumour FFPE DNA showed good correlation with CNAs detected in matched cfDNA samples for two out of ten cases. A number of factors may explain the poor correlation of CNAs in matched tumour and cfDNA samples. Anti-cancer treatment prior to blood withdrawal may affect the identification of plasma CNAs by reducing tumour bulk and therefore the shedding of tumour DNA into the circulation (129). Of the ten cases with matched tumour FFPE DNA and cfDNA, four had treatment prior to blood withdrawal. In hierachial cluster analysis, the cfDNA CNA profile of a patient treated with neoadjuvant chemotherapy grouped with the profiles of healthy controls rather than untreated lung cancer cases (265). Tumour-derived cfDNA levels in the circulation are dynamic; levels can increase with disease progression and reduce in response to anticancer therapy (129, 131).

Tumour evolution may be an important factor to explain discrepancies between CNAs detected in tumour DNA and matched cfDNA samples. Tumour evolution results in the development of major clones and subclones of cells that can be spatially separated within tumours and between primary and metastatic sites (289). CNAs identified in tumour DNA from both primary and metastatic sites have been detected in cfDNA samples (113). Furthermore, CNAs from a case with synchronous primary ovarian and breast tumours were detected in cfDNA samples (290). Four of our cases had more than five months pass from tumour sampling to blood withdrawal and new genetic alterations in the tumour may have developed in this time.

Genetic heterogeneity is a characteristic of lung tumours and therefore evaluating only a small sample of tumour tissue taken by a single biopsy can lead to sampling bias (196, 291). CfDNA may be more representative of tumour heterogeneity than a tumour biopsy, because cfDNA comprises of DNA from different clonal populations of tumour cells (106). In a multi-region whole exome/whole genome sequencing study of seven early stage NSCLC primary tumours (adenocarcinoma and squamous cell carcinoma), there were CNAs unique to one or two tumour regions and intra-tumour heterogeneity varied from 4% to 63% (292). However, less than 5% of the tumour was sampled and therefore intra-tumour heterogeneity may have been underestimated (196). In another study of four patients with early stage lung cancer, 43% of ubiquitous (predicted to occur early in tumour evolution and therefore be clonal) and heterogenous (predicted to occur late in tumour evolution and therefore be subclonal) single nucleotide variants (SNVs) identified in tumours were identified in plasma (107). The allele fraction of ubiquitous SNVs ranged from 0.15% to 23.25% whilst the range for heterogenous SNVs was 0.28% to 1.17% (107). TRACERx is a longitudinal study that will provide greater understanding of tumour heterogeneity and evolution in NSCLC by sequencing of tumour, cfDNA and ctDNA (293).

Another possible explanation for the poor correlation of CNAs between tumour FFPE DNA and cfDNA samples could be the introduction of artefact due to repetitive regions, PCR or sequencing errors and poor quality DNA. This can cause false positives and negatives. To reduce false positives from repetitive regions, any window that overlapped regions blacklisted by ENCODE were removed (see Section 2.18.6.2). Furthermore, normalisation of cfDNA and tumour FFPE DNA read profiles against genomic DNA read profiles to create copy

number ratios may help to reduce error from sequencing artefact if present in both samples as well as eliminate inherited aberrations (see Section 2.18.6.1). The coverage of tumour FFPE DNA samples was significantly lower than the coverage of cfDNA samples X0.18 vs X0.49, this may reflect poor quality tumour DNA that has been damaged by formalin fixation and paraffin embedment (274). Deletion of chromosome 19 was identified in cfDNA samples spiked with tumour FFPE DNA, control cfDNA sample (0% FFPE DNA) and control tumour FFPE DNA (100% FFPE DNA). This observation may be an artefact related to the fragmented nature of cfDNA and FFPE DNA and is further discussed in Section 4.4.2.1.

Tumour clones and subclones when diluted by wild type DNA in the circulation may be present at too low an allele fraction to be detected by low coverage whole genome sequencing. The detection of tumour-derived CNAs in cfDNA is dependent on the fraction of ctDNA in the circulation (178, 290) and sample coverage (178). Higher ctDNA fractions in the plasma were associated with a higher proportion of tumour associated CNAs in cfDNA samples (290). In silico analyses showed that when the fractional DNA concentration reduced a higher number of reads were required to detect CNAs (290). In addition, the magnitude and size of CNAs also influences detection (290). Two copy gains were detected with higher sensitivity than one copy gain or loss (290). The detection of ctDNA is also dependent on the number of ctDNA molecules recovered from plasma, highlighting the importance of efficient plasma extraction methods (169). In addition, the short half-life of cfDNA may explain how some patients with advanced cancer have no measurable mutant fragments because cfDNA is rapidly cleared from the circulation (179).

4.4.1.2 The limit of detection of copy number aberrations in circulating cell-free DNA with low coverage sequencing

The limit of detection by visual inspection for the identification of CNAs was shown to be approximately 10-20%, which was determined by adding tumour FFPE DNA to control plasma at different proportions. This may be an underestimate because tumour FFPE DNA can be of poor quality due to the DNA damage that occurs during formalin fixation and paraffin embedment (274). DNA quality is an important factor to ensure that sequencing results are reliable and accurate. The mutant allele fraction may give a more specific measure of ctDNA but is reliant on the tested allele being mutated in the tumour. Furthermore, tumour aneuploidy is not taken into account neither is the increase or decrease in the numbers of circulating mutant copies caused if a mutation is present in a region of copy number gain or

loss (294). Low coverage whole genome sequencing (0.1X) detected tumour-derived CNAs in patients with advanced breast cancer in cfDNA samples with mutant allele fraction greater than 10% (178). In other studies, CNAs were detected in cfDNA samples by whole genome sequencing with mutant allele fractions between 3.7% to 5% (167, 179, 290). Although, the numbers of cases with mutant allele fractions and CNA scores were small, these results compare well to the mutant allele fractions observed for our cases with high CNA scores (see Section 4.3.2.2.2).

With high coverage sequencing, in silico analysis showed that a cfDNA mutant allele fraction of 0.75% had a sensitivity greater than 90% and specificity greater than 99% for the detection of chromosomal arm cfDNA CNAs (167). In silico analyses may not be representative of real values and confirmatory experiments are required. The presence of longer DNA fragments have been associated with a higher detection rate of CNAs, emphasising the need for good quality DNA (179).

Commercial reference standards may more accurately determine analytical sensitivity and specificity and the lower limits of detection. Human cancer cell line DNA has been sheared to 160 bp to produce commercial reference standards with known fractions of mutant alleles and copy number variants (275). Using visual inspection is not a very precise measure for determining the lower limit of detection. Neither is using correlation because the optimal cut-off is unknown. Furthermore, there was correlation of copy number ratios between 100% tumour DNA and cfDNA without any tumour DNA added. This may be due to sequencing artefact for example chromosome 19 loss was demonstrated in both 100% tumour FFPE DNA and cfDNA without spiked tumour FFPE DNA. To get a more precise estimate of the lower limit of detection the circulating mutant allele fraction with Ion Torrent targeted sequencing was explored (see Section 4.3.2.2.2).

4.4.1.3 Fragment size of circulating cell-free DNA samples

The baseline median cfDNA fragment length was significantly longer for low risk controls (158 bp) compared to high risk controls (146 bp) and cases (146 bp). This may represent differences in the mechanism of cfDNA release into the circulation. One-hundred and forty six bp is equivalent to the size of a mono-nucleosome and 158 bp is more consistent with the size of a mono-nucleosome plus linker protein (115). Heitzer et al 2013, reported slightly longer peak fragment lengths for cfDNA in healthy controls compared to patients with

advanced colorectal cancer when cfDNA fragments were size separated by electrophoresis by Agilent Bioanalyser (113). There was a higher proportion of smaller cfDNA fragments (<145 bp and <100 bp) identified by PCR in the blood of colorectal cancer patients compared to healthy controls (108). DNA fragment size differences are important to understand to ensure accurate quantification and mutation analysis by using primers of an appropriate length (295). Furthermore, differences in the size of DNA fragments may aid the differentiation of cases and controls (114).

The baseline cfDNA profiles for both cases and controls in this study contained DNA greater than 500 bp in size. Larger fragments may represent DNA secreted directly from tumour cells into the blood or DNA released by macrophages after the necrotic death of tumour cells rather than the small fragments expected after apoptotic cell death (179). PCR selectively amplifies small DNA fragments and therefore longer DNA fragments in these samples may have been lost during library preparation and sequencing.

The bioanalyser profiles of cfDNA libraries often had a peak fragment size around 290 bp and a second smaller peak around 450 bp. This may represent di-nucleosome fragments that became visible because of amplification during library preparation. Baseline cfDNA bi-phasic profiles of advanced stage colorectal cancer patients (N=32) were associated with higher cfDNA total levels, higher levels of mutant KRAS and higher proportion of copy number aberrations after array CGH (median 10% vs 22%), compared to cases without a bi-phasic profile (113). Similar findings were identified for patients with advanced breast cancer (N=35), whereby a bi-phasic profile was associated with higher number of CNAs after low coverage sequencing (78% vs 7.7%) (179). A saturation of cfDNA degradation mechanisms may occur due to very high cfDNA levels resulting in the release of longer DNA fragments (179). None of our cfDNA baseline profiles were bi-phasic. The biology and dynamics of cfDNA release from tumour cells requires further study across different cancer types, between individuals and within the same individual to assess variation.

4.4.2 Clinical validation

4.4.2.1 Genomic instability scores as a screening tool in lung cancer

In this pilot study, the use of two genomic instability scores based on the detection of tumour-derived CNAs were explored in cfDNA samples of untreated lung cancer cases and

high risk controls. The CNA score had good discriminative ability between lung cancer cases and high risk controls with AUC 0.74. However, when only early stage (I-IIIA) cases were considered there was poor ability to discriminate, with AUC 0.60.

It is more difficult to detect tumour-derived cfDNA in the blood of patients with early stage compared to advanced stage disease due to lower tumour volume and therefore lower number of tumour cells that can potentially release DNA into the blood (160). A test with high analytical sensitivity has a greater chance of detecting tiny amounts of tumour-derived cfDNA (ctDNA) diluted by wild type DNA in the circulation (161). Low coverage whole genome sequencing provides a non-selective approach requiring no prior knowledge of the tumour alterations. In comparison, a personalised approach aims to detect genetic alterations known to be present in matched tumour samples and is tailored to each individual (161). With this approach, highly sensitive technologies such as digital droplet PCR can be used (161) or genomic regions can be targeted at very high coverage to enhance sensitivity of next generation sequencing methods (296) (see Table 1-2). The disadvantages of a personalised approach are that it takes longer, is more costly and that a tumour sample must be available with sufficient quantities of extracted DNA for genetic testing so a repeat biopsy may be required (161). Most importantly, a personalised approach cannot be applied to a screening programme.

Compared to whole genome sequencing to establish CNAs, targeted approaches reduce the proportion of the genome sequenced, therefore the depth of coverage can be increased and analytical sensitivity enhanced. Kirkizlar et al 2015, amplified 3168 SNPs across five genomic regions using massively multiplex PCR (mmPCR) to obtain amplicons of 75 bp (294). Amplicons were sequenced with Illumina HiSeq 2500 and reads were counted to determine CNAs. Clonal and subclonal tumour-derived cfDNA CNAs were detected in 73% of breast cancer cases with stage II disease with an analytical sensitivity of 0.5% (N=11). The average allelic imbalance (AAI) was calculated across specific regions to calculate the proportion of abnormal DNA in a cfDNA sample. CNAs were detected with an AAI between 0.8% and 5.3%. To calculate the allelic imbalance of a region, the number of reads for a heterogenous allele were compared to the total number of reads for both alleles at the same locus. Therefore, this method requires haplotype information for each individual warranting a combinatory biomarker approach (294).

A one-unit increase in the $\log_{10}$ CNA score led to a 2.16 (95% CI 1.07-4.34) increased chance of lung cancer after adjusting for $\log_{10}$ cfDNA levels, age, gender and length of plasma storage. Generally lower $\log_{10}$ CNA scores were associated with lower total cfDNA levels, indicating less ctDNA in the blood stream. Tumour burden generally correlates with ctDNA levels (160). Some studies (110, 118) but not all studies (138, 152, 209) correlate tumour burden with total cfDNA levels. As discussed in Section 3.4.2, the discrepancy is because total cfDNA levels are not specific to cancer and are raised in other medical conditions (102-105, 280).

There are many potential reasons for not detecting CNAs in cancer cases, such as inadequate circulating ctDNA fractions, poor test sensitivity and biological and technical differences (discussed in Section 4.4.1.1). In contrast, some high risk controls had unexpectedly raised CNA scores and visual inspection of copy number profiles demonstrated the presence of segmental copy number gains and losses despite normalisation to matched genomic DNA and elimination of germline aberrations (Appendix F).

Similar to this study, other studies have reported the presence of abnormal DNA in cfDNA samples of controls (114, 160, 171, 297). For example, *TP53* mutations were identified by Ion Torrent targeted deep sequencing (5000X) in approximately 10% of matched controls attending hospital for a non-cancerous medical condition (N=225)(297).

All of our high risk controls were current or ex heavy smokers and genetic alterations may have accumulated in cells secondary to prolonged carcinogen exposure or be a sign of increasing age (171). Somatic CNAs were detected in DNA extracted from bronchial biopsy specimens sampled at different sites along the airways of cases with more than a 20-pack year history of smoking and pre-invasive lung lesions, this is consistent with a 'field characterisation' effect (298). However, the tiny fractions of cfDNA released into the circulation from pre-invasive lesions (<0.1% (296) or benign cells (299) is unlikely to be detected by low coverage whole genome sequencing and therefore may not explain the detection of CNA in our high risk controls.

To our knowledge, none of the high risk controls in our study subsequently developed cancer that could have explained the presence of non-germline CNAs in the blood. The median length of follow up for high risk controls in our study was approximately seven years (range

four to ten years). In a large international prospective study, *TP53* and *KRAS* mutations were identified in 3% and 1% of cfDNA samples of healthy controls that did not subsequently develop cancer (median follow up 75.4 months) and 5.5% and 3.8% of controls that developed cancer at a later date (median time to event 20.8 and 14.3 months respectively) (300). However, in this study *TP53* and *KRAS* mutations in germline DNA were not established. The clinical impact of detecting abnormal DNA in the blood of individuals without cancer needs to be understood to avoid patient anxiety and unnecessary medical interventions (301). Large longitudinal prospective studies with prolonged follow up are required to establish the significance of detecting non-germline cfDNA abnormalities in the blood of non-cancer cases. It would be useful to repeat library preparation and sequencing for high risk controls in our study to determine if our results are reproducible. Mapping, PCR or sequencing artefacts or errors could lead to false positives, or samples may have been cross contaminated.

Deletion of chromosome 19 was a common finding observed across copy number profiles of lung cancer cases and high risk controls. This observation may be related to the fragmented nature of cfDNA and be a sequencing artefact as it was also present in a small proportion of tumour FFPE DNA copy number profiles but not in H69 cell line DNA copy number profiles.

CNA scores were higher when 5% cell line DNA was spiked into pooled healthy control cfDNA compared to samples with no DNA spike. However, the mutant allele fraction would be expected to be a more accurate measure of analytical sensitivity. CfDNA mutant allele fractions were available for six treated lung cancer cases with CNA scores. The cfDNA mutant allele fraction describes the proportion of mutant alleles compared to the total number of alleles. One advanced NSCLC case with a very high CNA score, was found to have a circulating *TP53* mutant allele fraction of 77.5%, a likely clonal mutation due to its high abundance. For another case despite a relatively low circulating *TP53* mutant allele fraction of 1.9%, the CNA score was higher than the median CNA score of advanced cases. It would be useful to know the ctDNA allele fraction and AAI score for all untreated cases and controls, to establish the relationship with CNA scores and to determine whether ctDNA could be detected.

The PGA2 score was calculated by summing squared copy number ratio Z scores ranked in the 95^{th} to 99^{th} percentile. However, this score did not behave as expected in our cohort. The PGA2 score was independent of $\log_{10}$ cfDNA levels and did not increase with advancing

disease stage. Furthermore, high risk controls had significantly higher PGA2 scores compared to lung cancer cases. These results contrast with previous findings by Xia et al. 2015, whereby the PGA score differentiated between early stage adenocarcinoma cases and normal controls (265). The PGA2 score did not capture the multitude of small aberrations present in copy number profiles. Despite eliminating the top 1% of scores, sequencing artefact may still have been present accounting for our results. Normalisation of each 1 Mb copy number ratio calculated to the mean copy number ratio values for the equivalent 1 Mb window of a healthy control group may further aid the elimination of sequencing artefact by subtracting background noise.

Another copy number aberration score that has been evaluated in cfDNA samples of breast and colorectal cancer cases is the Plasma Aneuploidy (PA) score (167). This score sums the Z scores of the top five chromosomal arms with the largest CNAs. To calculate the PA score, first the numbers of reads mapping to a chromosome arm were compared to the total number of reads to obtain a genomic representation (GR) score. A Z score was calculated by subtracting the mean GR score of a group of normal controls from the observed GR score and dividing this by the standard deviation of the GR score of the normal controls. The top five Z scores for the chromosomal arms were converted to P values and the negative sum of the logarithms of the P values were summed (167). A small change in a Z score equates to a large change in a P value and therefore P values can be a better discriminator particularly at the extreme ends of a distribution. Then, the PA score was calculated by taking the observed summed log P values and subtracting the mean of the summed log P values of the normal controls and dividing by the standard deviation of the summed log P values of the normal controls (167). A cut off greater than 5.84 had a specificity of 99% for the detection of CNAs in cfDNA (167). All controls (N=10) had a PA score less than 5.84 and all cancer cases (N=10) had a score higher than 5.84. The two lowest PA scores for cases had the lowest ctDNA fractions (1.4% and 4.7%, overall range 1.4%- 47.9%) (167).

Comparisons of scores between studies is difficult because of differences in the characteristics of study participants and methodology. Furthermore, most studies have included only a small number of cases and controls (167, 178, 179, 265). There are variations in methods for blood collection and processing and extraction of cfDNA from plasma. In addition, there are differences in methods of library preparation, sequencing techniques and machinery and bioinformatics pipelines. In this study, the number of reads in a 1 Mb window

for cfDNA were compared to matched genomic lymphocyte DNA to obtain copy number ratios. Other studies to obtain copy number ratios, compare the number of reads for cfDNA to the mean number of reads of a group of healthy controls. However, little information is given about control groups and they are mostly unmatched to cases for age and gender. Furthermore, details regarding smoking history of controls, lung cancer risk, or previous history of cancer are absent. These clinical parameters may be important confounding factors.

The use of genomic DNA as a comparative to cfDNA has the advantage of eliminating germ-line aberrations. Nevertheless, test cost could be halved if matched genomic DNA was not required to be sequenced. Instead of genomic DNA, the number of reads in cfDNA samples could be compared to the mean number of reads from a control group for each corresponding 1 Mb window. A larger control group would increase the number of reads across the genome, reduce read variability and therefore increase the chance of detecting a true CNA by reducing background noise (169). However, matching to genomic DNA is a more accurate way to reduce false positives and is an advantage of our method compared to other genomic instability scores that do not use genomic DNA (178, 266).

4.4.2.2 Circulating cell-free DNA genomic instability scores as a prognostic tool

Lung cancer cases with a $\log_{10}$ CNA score above the median value lived for significantly less time compared to cases with a score below the median value (11 vs 38 months). However, the $\log_{10}$ CNA score was an independent prognostic factor in univariable but not multivariable analyses. This may be due to small sample size and therefore poor power to observe a significant difference when additional factors are included in the model or could be caused by bias introduced by the interaction of increasing $\log_{10}$ CNA score with increasing disease stage. Consistent with disease stage being a strong prognostic indicator, disease stage was found to have the highest hazard ratio and it was the most significant factor affecting survival in both univariable and multivariable analyses.

There is minimal data published regarding the prognostic value of cfDNA genomic instability scores in cancer cases and larger studies are required. In a study of twenty prostate cancer patients, higher PGA scores were associated with shorter survival in pre-treatment and post-treatment cfDNA samples (266).

4.4.3 Study limitations

It would be useful to validate the detection of CNAs in tested plasma and tumour samples by an alternative method such as array CGH or Affymetrix 6.0 SNP array. However, due to the mostly low quantities of cfDNA extracted whole genome amplification would be required, which may introduce bias. The identification of CNAs in cfDNA samples with array CGH data has been shown to be concordant with low coverage sequencing data (178).

For data analysis,the window size was fixed to 1 Mb and therefore focal CNAs less than 1Mb in size or point mutations were not decteced due to inadequate resolution. In addition, nine chromosomal regions had no mapped reads and therefore no coverage. These regions tended to be close to repetitive regions such as telomeres and centromeres and therefore there can be difficulty with the alignment of short reads. There were no reads aligned to the short arm of chromosome 13 or 14 but these regions do not contain common CNAs important in lung cancer.

This was a retrospective study, although blood samples were collected specifically for cfDNA genetic analyses with optimised methods (140). Different pre-analytical factors and their effect on the CNA score were not assessed. It has been reported that variations in pre-analytical factors can influence the detection of somatic mutations and that pre-analytical factors should be optimised by assessing their impact on the detection of tumour-derived cfDNA (135).

Although our sample set is representative of the general lung cancer population for disease stage and histological subtype, it is a heterogenous group and the statistical analysis of sub-groups is limited due to small case numbers.

To establish the CNA score it was assumed that DNA reads obtained after sequencing were mapped to the genome without bias and were therefore equally distributed across chromosomes. However, nine chromosomal regions had no aligned reads for all samples and this issue is discussed in Section 4.4.3. The magnitude of copy number losses are limited and therefore a copy number loss may not score as high as a copy number gain. Weighting of copy number losses may eliminate this bias for cases with multiple copy number losses compared to gains.

4.5 Summary and Conclusion

This pilot study provides preliminary evidence for the detection of CNAs in cfDNA samples of lung cancer cases by low coverage whole genome sequencing. For some but not all lung cancer cases, good correlation was found between tumour FFPE and cfDNA copy number ratios. Low coverage whole genome sequencing of cfDNA samples was specific and known common tumour CNAs were identified for the three main histological sub-types of lung cancer. The limit of detection for segmental CNAs by visual inspection was 10%-20%. Many factors can influence the detection of cfDNA CNAs. These include anti-cancer treatment prior to blood withdrawal, the fraction of circulating tumour-derived cfDNA, depth of coverage, the size of the aberration and biological and technical differences.

To our knowledge, this is the largest study to explore genomic instability scores based on the detection of cfDNA CNAs in lung cancer cases and high risk controls. The PGA2 score did not perform as expected in our cohort and was not further evaluated. The CNA score had good discriminatory ability to differentiate between high risk controls and lung cancer cases with advanced stage IIIB-IV but not early stage I-IIIA disease. Low coverage whole genome sequencing did not detect the small fractions of tumour-derived cfDNA in the blood when cases had low tumour volumes. A targeted approach that sequences a smaller proportion of the genome combined with a high depth of coverage is predicted to increase analytical sensitivity whilst maintaining cost effectiveness. In a screening programme, tumour DNA is not available and therefore The Cancer Genome Atlas (TCGA) could be used to identify recurrent focal CNAs important in NSCLC and SCLC and these regions could be targeted to maximise the potential of detecting aberrations in patients with early stage disease. Any genomic instability score will require validation in a large independent prospective screening study. Low coverage whole genome sequencing may have greater clinical utility in patients with advanced cancer to identify CNAs that may predict treatment response and identify mechanisms of treatment resistance.

5 Discussion and Future work

Lung cancer causes the highest number of cancer related deaths in the UK and worldwide because most cases present with advanced disease that is not amenable to curative treatment (302). The aim of this study was to develop a non-invasive biomarker to aid early lung cancer detection. For clinical utility, such a molecular biomarker test should be based on a sample that is easy to obtain, (e.g. blood), be easy to perform in an NHS lab, reproducible, cost-effective, and have high diagnostic sensitivity. The hypothesis was that cancer cases would have higher levels of tumour-derived cfDNA (and therefore higher total levels of cfDNA), and greater genomic instability scores compared to high risk controls.

To be useful as a screening tool a biomarker needs to differentiate between early lung cancer cases and high risk controls. In this study, cfDNA levels were not significantly different between early stage (I-IIIA) lung cancer cases (N=21) and high risk controls defined by an LLP score ≥2.5% (N=30). There was a substantial overlap in the distribution of cfDNA levels in these two groups, leading to poor discrimination, with a ROC AUC of 0.53.

More specific to cancer than measuring total cfDNA levels, tumour-derived genetic alterations are detected in cfDNA samples of lung cancer cases (167, 195). To our knowledge, this is the largest study to explore genomic instability scores based on the detection of cfDNA CNAs in lung cancer cases (N=51) and high risk controls (N=30). Our approach was designed to detect CNAs across the whole genome by low coverage sequencing and therefore enable the detection of aberrations across different lung cancer histological subtypes without prior knowledge of any aberrations that may be present in a tumour. This is an essential characteristic for a potential screening tool. Furthermore, by reading only a small proportion of the genome at low depth of coverage, more samples could be multiplexed together in one sequencing run, thus reducing test cost. Low coverage sequencing data was available for 51 untreated cancer cases, 30 high risk and 10 low risk controls.

The PGA2 score was calculated by summing the ranked 95th to 99th percentile squared copy number ratio Z score from 1 Mb windows. There was no association between the PGA2 score and total cfDNA levels and the PGA2 score did not increase with increasing lung cancer stage. The PGA2 score was not explored further because contrary to our hypothesis, PGA2 scores were higher for high risk controls compared to lung cancer cases and low risk controls. By selecting, the 95th to 99th percentiles to capture large amplitude CNAs, smaller amplitude

CNAs were missed and genomic instability was only measured across a small proportion of the genome.

The CNA score assessed genomic instability across the whole genome and was established by summing the squared copy number Z scores for each 1 Mb window, based on the mean and standard deviation of the low risk control group. The log_{10} CNA score was correlated with log_{10} total cfDNA levels. Furthermore, the CNA score increased with increasing disease stage. The distribution of CNA scores was higher in lung cancer cases compared to high risk controls. Log_{10} CNA scores had good discriminative ability in distinguishing between all lung cancer cases and high risk controls, with AUC 0.74. However, when the log_{10} CNA score was adjusted for age, gender, disease stage, log_{10} cfDNA levels and length of plasma storage the AUC improved to 0.91. Yet, when only early stage cases were compared to high risk controls the discriminatory ability was poor with AUC 0.60.

The discriminative ability of the CNA score was affected by an overlap of scores between lung cancer cases and high risk controls and this is discussed in detail in Section 4.4.1.1 and 4.4.2.1. It is not clear why some high risk controls had high CNA scores. Other studies have reported the detection of genetic alterations in cfDNA of controls (114, 171, 297, 300) and this possess a significant challenge to the development of cfDNA screening tests (297, 301). In contrast, some lung cancer cases had unexpectedly low CNA scores. Highly sensitive tests are required to detect ctDNA in the plasma, particularly for early stage cancer cases with low tumour bulk (159, 160). Tumour clones and subclones when diluted by wild type DNA in the circulation may be present at too low an allele fraction to be detected by low coverage whole genome sequencing (107).

The CNA score alone is not recommended for further evaluation as a potential lung cancer screening tool, but it may be useful in combination with other genomic biomarkers, or be useful to monitor quantitative changes longitudinally because it would be expected that with greater cancer burden the CNA score would increase (but remain unchanged in controls without cancer).

5.1 Study limitations

5.1.1 Pre-analytical factors

Retrospective analyses was performed of plasma and genomic DNA samples. It has been reported that cfDNA levels decline by approximately one third for every year in storage (201). Yet, the screening performance of cfDNA levels did not change and cfDNA levels continued to have very good ability to distinguish between lung cancer cases and controls (201). Plasma samples in this study had been stored at -80°C for between one and nine years prior to cfDNA extraction. Work carried out in our laboratory has shown that cfDNA yields at two points seven years apart were well correlated (Pearson's R= 0.78, $p<0.0001$) and there was a median yield drop of 2.8 ng/ml (IQ range 0.59-6.2 ng/ml) (Prof Cox personal comment). To minimise the effect of variable lengths of storage this factor was adjusted for, in logistic regression and cox regression survival analyses. In this study, length of plasma storage was not a significant factor in logistic regression or survival analyses suggesting that this is not a major issue here and vindicates the consistent collection, processing and storage procedures defined in the ReSoLuCENT protocol. However, it may be important to carry out a large prospective study with analysis of cfDNA levels within three months of plasma storage to eliminate bias potentially caused by degradation.

Higher plasma volumes may be necessary to increase the number of ctDNA genome copies present to enable detection. Just one mutant genome may be present in 5 mls of plasma (195). Up to 3 mls of plasma were used in this study, in other studies up to 20 mls of plasma have been collected (167). Automated cfDNA extraction methods may be necessary to deal with high plasma volumes on a large scale.

Much work has been carried out into blood collection, processing and plasma storage (134). The aim is to preserve plasma and avoid white blood cell lysis that increases genomic DNA and further dilutes tumour-derived cfDNA. However, little is known about the optimal physiological condition of the host for cfDNA testing. In this study, the colour of the plasma was highly variable from clear to cloudy and pale yellow to dark yellow/orange. High lipid plasma levels may interfere with cfDNA binding to the column and result in lower cfDNA yields. It is not known whether improved cfDNA yields are seen after overnight fasting.

5.1.2 Case and control selection

The design of ReSoLuCENT to enrich the study sample for cases with inherited changes has introduced bias into this study. Lung cancer cases were selected to be of a younger age (<60 years old) unless there was a strong family history of lung cancer. This led to a lower distribution of ages with a median age of participants of 56 years compared to an unselected lung cancer population when it is expected that half of all lung cancer cases occur in individuals >70 years old (303). Therefore, comparison with the general lung cancer population is potentially limited.

In this study, controls were selected to be high risk if they had more than a 2.5% chance of developing lung cancer over a five-year period according to the LLP risk model (11). The LLP risk model evaluates seven lung cancer risk factors to include older age and smoking history (11). In this study, the median age of selected controls was higher than the median age of selected cases. The age bias may have affected the results; older controls may have more genetic changes related to increasing age. Due to the methods of choosing cases and controls, univariable and multivariable logistic regression analyses showed that younger age was associated with greater odds of having lung cancer, when in fact the natural history of lung cancer is that risk increases with age (303). As a result, the LLP score performs poorly in our selected sample.

The importance of ethnicity in the pathogenesis of cancer is acknowledged (304). Samples used in this study were from mostly white British cases and controls and therefore the findings may not be directly transferable to groups of alternative ethnicity. However, a recent targeted exome sequencing study of 509 non-small cell tumours found there to be no difference between mutation frequencies and copy number alterations between matched black and white individuals (305).

CtDNA levels reduce in response to anti-cancer treatment (129, 131). In this study, the focus was to select cases that had not received anti-cancer treatment prior to blood withdrawal to maximise the chance of detecting tumour-derived cfDNA. However, the numbers of untreated cases were limited in the ReSoLuCENT cohort. Furthermore, the number of cases with early stage cancer were small. Samples from all 21 untreated early stage cases were analysed but to adequately power our study as a guide 30 cases were required to estimate a true sensitivity of 0.92 or more with a 95% confidence interval width of 0.2 (306).

Furthermore, the inclusion of cases with stage IIIA disease, associated with a worse prognosis and potentially greater disease bulk compared to cases with stage I and II disease, may have led to an overestimation of results in the early stage group. This group was included because they can be radically treated.

The small numbers of cases and controls and retrospective design limits this pilot study. On the other hand there is no study that has tried to compare these two groups. Due to the selection methods and small numbers of cases and controls eligible for our study, it was not possible to match chosen cases and controls for age, gender or smoking history to balance the distribution of these potentially confounding factors. To overcome bias in selection of cases and controls, any genetic instability score established in our data set must be validated prospectively in a large independent study and be tested in the population where the score is intended to be used eg. a high risk population attending lung cancer screening. Prolonged follow up is vital to evaluate whether high risk controls with genetic alterations detected in cfDNA subsequently develop cancer and how early these alterations can be detected prior to any diagnosis of lung cancer.

5.2 Future work

5.2.1 Further analyses of data

5.2.1.1 Combining the CNA score with DNA fragment length (determined by sequencing)

Whole genome sequencing enables every DNA insert or fragment length to be defined (114). Jiang et al 2015 studied the DNA insert size of sequenced fragments (median 31 million reads) from selected chromosomal arms known to have aberrations in hepatocellular carcinoma tumours and found that tumour-derived DNA fragments were shorter than non-tumour derived fragments (114). Therefore, short fragments were over-represented with copy number gain and under-represented with copy number loss. This resulted in a difference in the size of fragments between hepatocellular cases and controls, which can be quantified. A score based on the size of sequenced DNA inserts in combination with a CNA score may improve the screening performance of our test.

5.2.2 Further experiments

The lower limit of detection for CNAs of our low coverage approach was approximated to be between 10-20% by visual inspection of copy number profiles and 5% for the CNA score, after spiking tumour FFPE DNA at known proportions into cfDNA samples of healthy controls. This may be an underestimate because tumour FFPE DNA is degraded and damaged by tissue processing methods, which can reduce sequencing efficiency (274). Accurate methods to improve the estimation of the limits of detection or analytical sensitivity of the technique would be to use human cell line DNA with known CNAs sheared to 160 bp and spiked into cfDNA samples of healthy controls at known different proportions or to use commercially available validated standards (275). In addition, a set of healthy controls is required to be a reference to evaluate whether a CNA score from a spiked sample has a significant aberration above a set threshold (for example two standard deviations) to evaluate analytical sensitivity.

ReSoLuCENT recruited controls that were co-habiting with lung cancer cases and detailed smoking histories were collected from participants. Therefore, exploring CNA scores in this cohort may control for environmental damge to DNA and the role of passive smoking may be able to be explored.

5.2.3 Combinatory biomarker approaches

The addition of total cfDNA levels did not improve the discriminatory ability of the CNA score but alternative combinatory approaches are required. More specific combinatory approaches to detect a multitude of tumour derived genetic alterations are recommended to aid early cancer detection. Further supportive evidence for a combinatory approach to enhance specificity is that the oncogenic signatures of tumours can be mutational or copy number rich (62). Tumours with a high number of recurrent mutational events may therefore have a low number of recurrent CNAs and low CNA scores. Genomic data held in The Cancer Genome Atlas (TCGA) for 3,299 tumours was analysed to identify recurrent genetic alteration signatures for 12 tumour subtypes. This study demonstrated that lung squamous cell tumours were copy number rich in comparison to lung adenocarcinoma tumours, which had an equivocal proportion of tumours with copy number rich and mutational signatures. In this study, there was a trend for higher CNA scores for NSCLC cases with squamous cell lung cancer compared to non-squamous cell lung cancer. A combined

targeted approach for the detection of SNVs and CNAs has successfully detected driver genetic alterations in cfDNA samples of early stage breast cancer cases (294).

In 2018, a grant of which I am a co-applicant was successful and led to an award of approximately £50,000 from Weston Park Charity. The aim of this pilot study is to develop a multivariable cfDNA biomarker to aid the detection of lung cancer. The objectives are to establish the sensitivity, specificity and utility of a combined optimised CNA score with tumour mutation burden. The CNA score will be optimised by exploring the addition of DNA fragment size to the score (as described in Section 5.2.1.1) as well the impact of recurrent CNAs, chromosomal re-arrangements and the presene of chromothripsis (reflecting a catastrophic genomic instability event) using established bioinformatics data. The tumour mutation burden will be calculated by identifying common SNVs mutated in NSCLC by high coverage (>X500) targeted sequencing of cfDNA samples.

Alternatively, methylation occurs early in lung cancer tumourogenesis (307-309) and combined approaches have been developed to identify both CNAs and methylation in cfDNA samples (106). By calculating the percentage of 1 Mb bins that were hypomethylated or contained a CNA after whole genome bisulphite cfDNA sequencing (mean 93 million reads), cases with hepatocellular carcinoma and cases with hepatitis B and cirrhosis were distinguished with a sensitivity of 92% and specificity of 88% (106). A four gene cfDNA methylation signature established by evaluating 96 markers by RT-qPCR had excellent discriminatory ability to distinguish NSCLC cases (stage I-IV) from healthy controls with AUC 0.90 (215). Furthermore, a DNA hypermethylation index established from tumour tissue was prognostic in an independent validation cohort of NSCLC stage I cases (310). Methylation cfDNA profiles warrants further investigation in early lung cancer cases and high risk controls.

5.2.4 Future considerations in order to establish a circulating cell-free DNA screening biomarker

In order to establish cfDNA as a cancer biomarker, methods for cfDNA blood collection, processing, plasma DNA extraction, library preparation, sequencing and bioinformatics must be standardised prior to clinical implementation. International efforts are in progress, such as CANCER-ID to evaluate, validate and develop standard operating procedures for ctDNA genetic analyses (311). However, the tiny fractions of ctDNA in the blood of early cancer cases may still limit the clinical utility of cfDNA as a screening tool.

Despite recent advances in technology, ctDNA is not detected in more than 50% of patients with stage I lung cancer (159, 169). High analytical sensitivity is an important factor for detecting ctDNA in early cancer cases because the ctDNA allele fraction can be less than 0.1% (169, 296). A high depth of coverage between 5000X to 10000X is required to detect tiny ctDNA fractions in the blood but the number of samples that can be multiplexed are reduced and therefore test cost increases (160, 297). In a recent study, a variant allele fraction of 0.1% equated to a primary NSCLC tumour volume of 10 cm^3 (159). Yet, much smaller tumour volumes can be detected by low dose CT screening (5) necessitating a test that can detect a variant allele fraction of 0.00014% (159). At such tiny fractions, it is vital to understand the PCR or sequencing error rate or 'background noise' for each genetic alteration in order to reduce false positives (160). Further technological advances are required to increase the analytical sensitivity for the detection of abnormal DNA in the blood to aid early lung cancer detection and avoid false negatives, whilst maintaining cost-effectiveness.

A greater understanding is required to determine the relationship of tumour evolution and cfDNA detected genetic alterations in the blood to maximise clinical impact. The analysis of tumour samples and circulating biomarkers in studies such as TRACERx and PEACE are enhancing our understanding of clonal and subclonal events (107). Most recently, the cfDNA profiles obtained by multiplex-PCR NGS for the first 100 TRACERx participants with early stage NSCLC were published (159). Clonal ctDNA SNVs were in greater abundance in the plasma compared to subclonal SNVs (159), supporting the generation of a cfDNA screening test based on recurrent clonal genomic alterations. CtDNA was detected in nearly all patients with early stage squamous tumours (97%) compared to 19% of cases with adenocarcinoma and 71% of cases with other NSCLC subtypes (159). The release of ctDNA into the circulation in early stage lung cancer was associated with more necrotic tumours (159). In this study, independent predictors for the detection of ctDNA, included non-adenocarcinoma, lympho-vascular invasion and high ki67 proliferation index. These findings highlight the importance of understanding the biology of the release of ctDNA in order to design an appropriate biomarker test and implies that cases with slow growing less necrotic and/or adenocarcinoma tumours are less likely to have ctDNA detected in the blood. This is an important consideration because adenocarcinoma tumours were the most commonly detected tumours when screening high risk groups by low dose CT (8). This implies that even with advances in technologies for ctDNA analyses a ctDNA biomarker may still need to be combined with different biomarkers from alternative body sources such as sputum, urine or

exhaled metabolites to enhance diagnostic sensitivity and specificity to overcome biological differences between tumours. For a screening test to have clinical utility in the NHS it must be cost effective and the more biomarkers that are tested the greater the cost. Our low coverage whole genome sequencing approach had an approximate cost of £120 per sample.

For molecular stratification prior to CT screening, a test with a high diagnostic sensitivity is warranted so that the false negative rate is low. Furthermore, a greater knowledge of abnormalities in healthy control and high risk groups is required to enhance analytical specificity. A screening test must distinguish between 'disease associated cfDNA mutations from exposure associated mutations' (300). A Pre-Cancer Genome Atlas aims to develop understanding of pre-invasive events and the genomic steps required to progress to malignancy and may identify new approaches for early detection and prevention (312).

Most biomarker studies carried out to date are case-control studies with a relatively small number of participants and the biomarker is not always tested in matched high risk groups. A screening prospective study testing a validated biomarker to prove clinical utility is very costly due to the large numbers of individuals involved and the need for longitudinal follow up to determine the outcome of those that test negative and those that test positive with no imaging abnormalities detected. Hence, any tested biomarker in a large prospective screening study must be robust and reproducible with established analytical and clinical validity to determine clinical utility (313). Furthermore, meaningful clinical end-points to assess the efficacy of any biomarker must be established, such as reducing the number of people having imaging investigations and increasing the detection rate of potentially curable lung cancers. A baseline register of biomarker studies may avoid publication selection bias and standard guidelines can aid reporting and comparison of studies (313). Furthermore, the ethical, legal and social implications of any developed biomarker must be considered as per the ACCE framework (189) .

In the future, it may be possible to determine the origin of ctDNA based on the methylation profile (314) or nucleotide footprint (109), which could indicate the next most appropriate imaging modality or investigation. Recently, RNA was isolated from platelets and sequenced in order to test the diagnostic value of RNA within tumour-educated platelets. For six types of cancer, the diagnostic sensitivity was 97% and specificity 94% and the RNA signature correctly identified the cancer subtype with 71% accuracy (315). In this study, the majority

of patients tested had advanced cancer and more data is required to test the diagnostic value in patients with early stage disease. These findings could lead to an early detection test for all cancer subtypes.

In 2016, 100 million dollars was invested in setting up a new company associated with Illumina to identify a blood based pan cancer screening test to aid early detection (316). With technological advancements and reduction in costs a cfDNA blood biomarker test may yet have clinical utility in the NHS to aid early cancer detection. A potential biomarker strategy to aid early detection of lung cancer is shown in Figure 5-1.

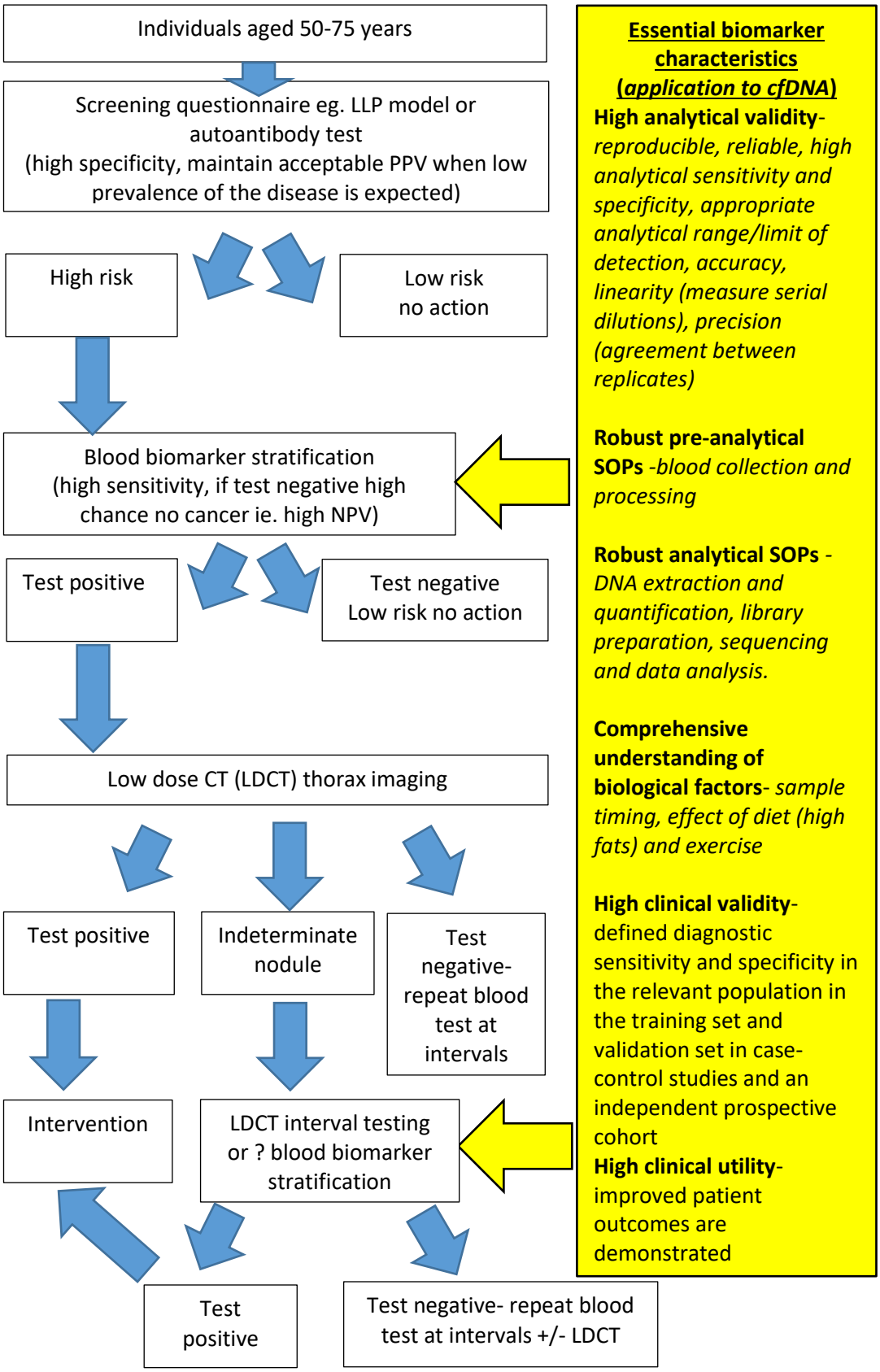


Figure 5-1: A potential biomarker strategy to aid lung cancer screening.

5.3 Summary of PhD work and collaborations

During my PhD, I participated in multiple research projects using blood samples and data collected in the Sheffield ReSoLuCENT study (Figure *5-2*). I have collaborated with Prof J Shaw's laboratory at the University of Leicester and Prof C. Dive's laboratory at the Cancer Research UK Manchester Institute to analyse cfDNA samples collected in the clinical trial for patients with small cell lung cancer called STOMP. In addition, I continue to be an active member of the International Lung Cancer Consortium that has genotyped over 15,000 lung cancer cases and controls including 11190 samples collected in ReSoLuCENT.

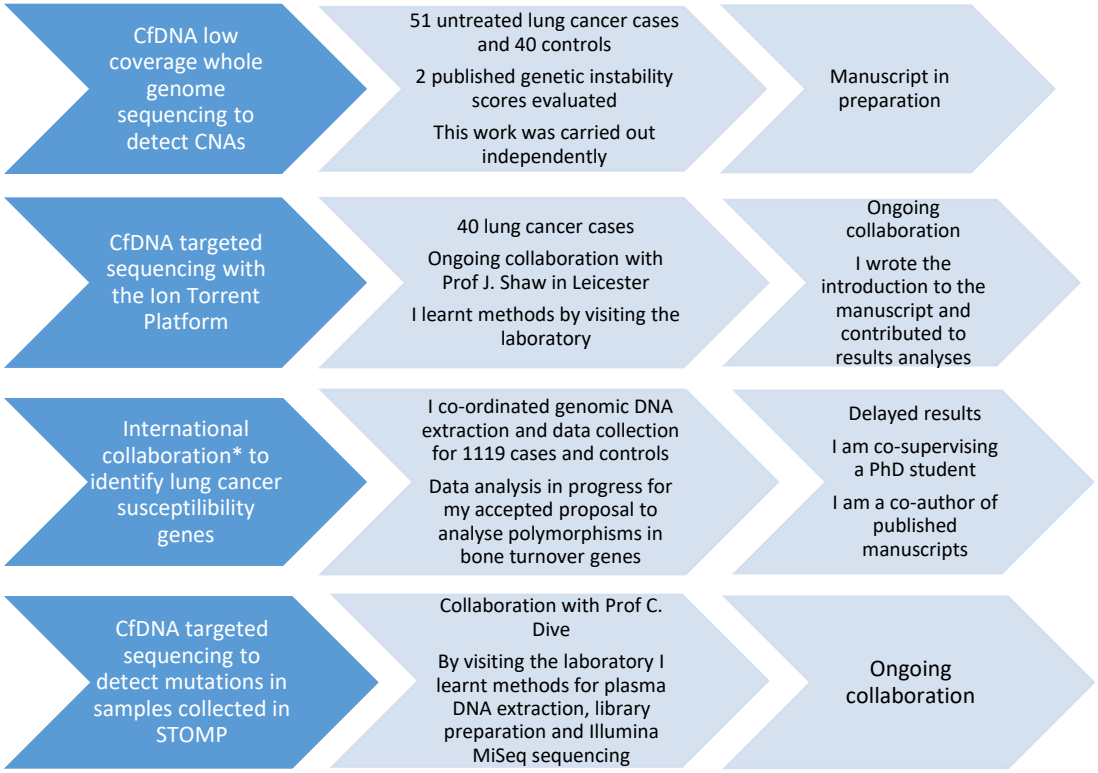


Figure 5-2: Summary of PhD work and collaborations.

STOMP: Small cell lung cancer trial of Olaparib following response to first line chemotherapy. *ILCCO (International Lung Cancer Consortium) GAME_ON study with genotyping results for over 15,000 lung cancer cases and controls.

6 References

1. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet (London, England). 2012;380(9859):2095-128.
2. Office of National Statistics. Cancer Survival in England patients diagnosed 2005-2009 and followed up to 2010 London: Office of National Statistics; 2011 [Available from: http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-239726.
3. Rachet B, Quinn M, Cooper N, Coleman M. Survival from cancer of the lung in England and Wales up to 2001. B J Cancer. 2008;99(Suppl 1):S40-S2.
4. Walters SM, C. Coleman, MP. Peake, MD. Butler, J. Young N. et al. Lung Cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden, and the UK: a population based study, 2004-2007. Thorax. 2013;68(6):551-64.
5. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. The New England journal of medicine. 2011;365(5):395-409.
6. Patz EF, Jr., Pinsky P, Gatsonis C, Sicks JD, Kramer BS, Tammemagi MC, et al. Overdiagnosis in low-dose computed tomography screening for lung cancer. JAMA internal medicine. 2014;174(2):269-74.
7. Humphrey L, Deffebach M, Pappas M, Baumann C, Artis K, Mitchell JP, et al. U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. Screening for Lung Cancer: Systematic Review to Update the US Preventive Services Task Force Recommendation. Rockville (MD): Agency for Healthcare Research and Quality (US); 2013.
8. Horeweg N, Scholten ET, de Jong PA, van der Aalst CM, Weenink C, Lammers JW, et al. Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers. The Lancet Oncology. 2014;15(12):1342-50.
9. Aberle DR, DeMello S, Berg CD, Black WC, Brewer B, Church TR, et al. Results of the two incidence screenings in the National Lung Screening Trial. The New England journal of medicine. 2013;369(10):920-31.
10. Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection criteria for lung-cancer screening. The New England journal of medicine. 2013;368(8):728-36.
11. Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. Br J Cancer. 2008;98(2):270-6.
12. Gray EP, Teare MD, Stevens J, Archer R. Risk Prediction Models for Lung Cancer: A Systematic Review. Clinical lung cancer. 2016;17(2):95-106.
13. Raji OY, Duffy SW, Agbaje OF, Baker SG, Christiani DC, Cassidy A, et al. Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer: a case-control and cohort validation study. Annals of internal medicine. 2012;157(4):242-50.
14. MyLungRisk (MLR) [Available from: https://secure2.utlnet.co.uk/mylungrisk/welcome.aspx
15. Field JK, Oudkerk M, Pedersen JH, Duffy SW. Prospects for population screening and diagnosis of lung cancer. Lancet (London, England). 2013;382(9893):732-41.
16. Field JK, Duffy SW, Baldwin DR, Whynes DK, Devaraj A, Brain KE, et al. UK Lung Cancer RCT Pilot Screening Trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening. Thorax. 2016;71(2):161-70.

17. Myers ER, Moorman P, Gierisch JM, Havrilesky LJ, Grimm LJ, Ghate S, et al. Benefits and Harms of Breast Cancer Screening: A Systematic Review. Jama. 2015;314(15):1615-34.
18. Youlden DR, Cramb SM, Baade PD. The International Epidemiology of Lung Cancer: geographical distribution and secular trends. J Thorac Oncol. 2008;3(8):819-31.
19. Travis WD, Brambilla E, Riely GJ. New pathologic classification of lung cancer: relevance for clinical practice and clinical trials. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013;31(8):992-1001.
20. Detterbeck FC, Boffa DJ, Kim AW, Tanoue LT. The Eighth Edition Lung Cancer Stage Classification. Chest. 2017;151(1):193-203.
21. Parkin D. Tobacco-attributed cancer burden in the UK in 2010. Br J Cancer. 2011;105(Suppl 2):S6-S13.
22. Brown T, Darnton A, Fortunato L, Rushton L, Group BOCBS. Occupational cancer in Britain: Respiratory cancer sites: larynx, lung and mesothelioma. Br J Cancer. 2012;107:S56-S70.
23. Maddams J, Parkin D, Darby S. The cancer burden in the United Kingdom in 2007 due to radiotherapy. Int J Cancer. 2011;129(12):2885-93.
24. Brenner D, McLaughlin JR, Hung R. Previous Lung Diseases and Lung Cancer Risk: A Systematic Review and Meta-Analysis. PLoS ONE. 2011;6(3):e17479.
25. Winstone TA, Man SFP, Hull M, Montaner JS, Sin DD. Epidemic of lung cancer in patients with HIV infection. Chest. 2013;143(2):305-14.
26. Cote ML, Liu M, Bonassi S, Neri M, Schwartz AG, Christiani DC, et al. Increased risk of lung cancer in individuals with a family history of the disease: a pooled analysis from the International Lung Cancer Consortium. European journal of cancer (Oxford, England : 1990). 2012;48(13):1957-68.
27. Kleinerman R, Tarone R, Abramson D, Seddon J, Li F, Tucker M. Hereditary Retinoblastoma and Risk of Lung Cancer. J Nat Can Inst. 2000;92(24):2037-9.
28. Foulkes W. Inherited Susceptibility to Common Cancers. New Engl J Med. 2008;359(20):2143-53.
29. Olivier M, Goldgar DE, Sodha N, Ohgaki H, Kleihues P, Hainaut P, et al. Li-Fraumeni and related syndromes: correlation between tumor type, family structure, and TP53 genotype. Cancer research. 2003;63(20):6643-50.
30. Hwang S-J, Cheng LS-C, Lozano G, Amos C, Gu X, Strong L. Lung cancer risk in germline p53 mutation carriers: association between an inherited cancer predisposition, cigarette smoking, and cancer risk. Human Genet 2003;113(3):223-43.
31. Brennan P, Hainaut P, Boffetta P. Genetics of lung-cancer susceptibility. The Lancet Oncology. 2011;12(4):399-408.
32. Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, et al. A Catalog of Published Genome-Wide Association Studies. 2009 [Available from: www.genome.gov/gwastudies.
33. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet. 2008;40(5):616-22.
34. McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, et al. Lung Cancer susceptibility locus at 5p15.33. Nature Genetics. 2008;40(12):1404-6.
35. Amos CI, Pinney S, Li Y, Kupert E, Lee JS, De Andrade M, et al. A susceptibility locus on chromosome 6q greatly increases lung cancer risk among light and never smokers. Cancer research. 2010;70(6):2359-67.
36. Truong T, Hung RJ, Amos CI, Wu X, Bickeboller H, Rosenberger A, et al. Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. Journal of the National Cancer Institute. 2010;102(13):959-71.

37. Brennan P, Hainaut P, Boffetta P. Genetics of lung-cancer susceptibility. Lancet Oncology. 2011;12(4):399-408.
38. Timofeeva MN, Hung RJ, Rafnar T, Christiani DC, Field J, Bickeboller H, et al. Influence of common genetic variation on lung cacner risk: meta-analysis of 14 900 cases and 29 485 controls. Human Molecular Genetics. 2012;21(22):4980-95.
39. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet. 2017;49(7):1126-32.
40. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet. 2014;46(7):736-41.
41. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature. 2008;452(7187):633-7.
42. Wistuba I, Behrens C, Milchgrub S, Bryant D, Hung J, Minna J, et al. Sequential molecular abnormalities are involved in the multistage development of squamous cell lung carcinoma. Oncogene. 1999;18(3):643-50.
43. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646-74.
44. Zabarovsky E, Lerman M, Minna J. Tumor suppressor genes on chromosome 3p involved in the pathogenesis of lung and other cancers. Oncogene. 2002;21:6915-35.
45. Greenman C, Stephens P, Smith R, Dalgliesh G, Hunter C, Bignell G, et al. Patterns of somatic mutations in human cancer genome. Nature. 2007;446(7132):153-8.
46. Agathanggelou A, Honorio S, Macartney DP, Martinez A, Dallol A, Rader J, et al. Methylation associated inactivation of RASSF1A from region 3p21.3 in lung, breast and ovarian tumours. Oncogene. 2001;20(12):1509-18.
47. Sharma S, Bell DW, Settleman J, Haber DA. Epidermal Growth Factor Receptor Mutations in Lung Cancer. Nat Rev. 2007;7(3):169-81.
48. Sard L, Accornero P, Tornielli S, Delia D, Bunone G, Campiglio M, et al. The tumor-supprsssor gene FHIT is involved in the regulation of apoptosis and in cell cycle control. Proc Natl Acad Sci U S A. 1999;96(15):8489-92.
49. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455(7216):1069-75.
50. Liu P, Morrison C, Wang L, Xiong D, Vedell P, Cui P, et al. Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. Carcinogenesis. 2012;33(7):1270-6.
51. Peifer M, Fernandez-Cuesta L, Sos ML, George J, Seidel D, Kasper LH, et al. Integrative genome analysis identify key somatic driver mutations of small-cell lung cancer. Nat Genet. 2012;44(10):1104-10.
52. Dearden S, Stevens J, Wu YL, Blowers D. Mutation incidence and coincidence in non small-cell lung cancer: meta-analyses by ethnicity and histology (mutMap). Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2013;24(9):2371-6.
53. George J, Lim JS, Jang SJ, Cun Y, Ozretic L, Kong G, et al. Comprehensive genomic profiles of small cell lung cancer. Nature. 2015;524(7563):47-53.
54. Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, Maki RG, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. The New England journal of medicine. 2010;363(18):1693-703.
55. Mazieres J, Peters S, Lepage B, Cortot AB, Barlesi F, Beau-Faller M, et al. Lung cancer that harbors an HER2 mutation: epidemiologic characteristics and therapeutic perspectives. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013;31(16):1997-2003.

56. Lee S, Kim M, Jin G, Yoo S, Park J, Choi J, et al. Somatic Mutations in Epidermal Growth Factor Receptor Signaling Pathway Genes in Non-small Cell Lung Cancers. J Thorac Oncol. 2010;5:1734-40.
57. Weiss J, Sos ML, Seidel D, Peifer M, Zander T, Heuckmann JM, et al. Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. Sci Transl Med. 2010;2(62):62ra93.
58. Oxnard GR, Binder A, Janne PA. New targetable oncogenes in non-small-cell lung cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013;31(8):1097-104.
59. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet. 2012;44(10):1111-6.
60. COSMIC. Catalogue of Somatic Mutations in Cancer Database 2014 [Available from: http://www.sanger.ac.uk.
61. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20.
62. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. 2013;45(10):1127-33.
63. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458(7239):719-24.
64. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12(5):363-76.
65. Bowcock AM. Invited review DNA copy number changes as diagnostic tools for lung cancer. Thorax. 2014;69(5):495-6.
66. Iwakawa R, Takenaka M, Kohno T, Shimada Y, Totoki Y, Shibata T, et al. Genome-wide identification of genes with amplification and/or fusion in small cell lung cancer. Genes Chromosomes Cancer. 2013;52(9):802-16.
67. Huang YT, Lin X, Liu Y, Chirieac LR, McGovern R, Wain J, et al. Cigarette smoking increases copy number alterations in nonsmall-cell lung cancer. Proc Natl Acad Sci U S A. 2011;108(39):16345-50.
68. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519-25.
69. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet. 2016;48(6):607-16.
70. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511(7511):543-50.
71. Belvedere O, Berri S, Chalkley R, Conway C, Barbone F, Pisa F, et al. A computational index derived from whole-genome copy number analysis is a novel tool for prognosis in early stage lung squamous cell carcinoma. Genomics. 2012;99(1):18-24.
72. Ma J, Gao M, Lu Y, Feng X, Zhang J, Lin D, et al. Gain of 1q25-32, 12q23-24.3, and 17q12-22 facilitates tumorigenesis and progression of human squamous cell lung cancer. The Journal of pathology. 2006;210(2):205-13.
73. van Boerdonk RA, Sutedja TG, Snijders PJ, Reinen E, Wilting SM, van de Wiel MA, et al. DNA copy number alterations in endobronchial squamous metaplastic lesions predict lung cancer. American journal of respiratory and critical care medicine. 2011;184(8):948-56.
74. van Boerdonk RA, Daniels JM, Snijders PJ, Grunberg K, Thunnissen E, van de Wiel MA, et al. DNA copy number aberrations in endobronchial lesions: a validated predictor for cancer. Thorax. 2014;69(5):451-7.
75. Clinical Lung Cancer Genome Project; Network Genomic Medicine. A genomics-based classification of human lung tumors. Sci Transl Med. 2013;5(209):209ra153.

76. Sequist LV, Heist RS, Shaw AT, Fidias P, Rosovsky R, Temel J, et al. Implementing multiplexed genotyping of non-small cell lung cancers into routine clinical practice. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2011;22(12):2616-24.
77. Fukuoka M, Wu YL, Thongprasert S, Sunpaweravong P, Leong SS, Sriuranpong V, et al. Biomarker analyses and final overall survival results from a phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non-small-cell lung cancer in Asia (IPASS). Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2011;29(21):2866-74.
78. Pao W, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, et al. EGF receptor gene mutations are common in lung cancers from 'never smokers' and are associated with sensitivity of tumours to gefitinib and erlotinib. Proc Natl Acad Sci U S A. 2004;101(36):13306-11.
79. Miller VA, Hirsh V, Cadranel J, Chen YM, Park K, Kim SW, et al. Afatinib versus placebo for patients with advanced, metastatic non-small-cell lung cancer after failure of erlotinib, gefitinib, or both, and one or two lines of chemotherapy (LUX-Lung 1): a phase 2b/3 randomised trial. The Lancet Oncology. 2012;13(5):528-.
80. Goss G, Tsai CM, Shepherd FA, Bazhenova L, Lee JS, Chang GC, et al. Osimertinib for pretreated EGFR Thr790Met-positive advanced non-small-cell lung cancer (AURA2): a multicentre, open-label, single-arm, phase 2 study. The Lancet Oncology. 2016;17(12):1643-52.
81. Middleton G, Crack LR, Popat S, Swanton C, Hollingsworth SJ, Buller R, et al. The National Lung Matrix Trial: translating the biology of stratification in advanced non-small-cell lung cancer. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2015;26(12):2464-9.
82. Boch C, Kollmeier J, Roth A, Stephan-Falkenau S, Misch D, Gruning W, et al. The frequency of EGFR and KRAS mutations in non-small cell lung cancer (NSCLC): a routine screening data for central Europe from a cohort study. BMJ Open. 2013;3(e002560).
83. Chan AK, Chiu RW, Lo YM. Cell-free nucleic acids in plasma, serum and urine: a new tool in molecular diagnosis. Annals of clinical biochemistry. 2003;40(Pt 2):122-30.
84. Schwarzenbach H, Hoon DSB, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nature reviews Cancer. 2011;13(5):528-38.
85. Thierry AR, El Messaoudi S, Gahan PB, Anker P, Stroun M. Origins, structures, and functions of circulating DNA in oncology. Cancer metastasis reviews. 2016;35(3):347-76.
86. Lo YMD, Zhang J, Leung T, Lau T, Chang M, Hjelm N. Rapid Clearance of Fetal DNA from Maternal Plasma. The American Journal of Human Genetics. 1999;64(1):218-24.
87. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A. 2008;105(42):16266-71.
88. Chiu RW, Chan KC, Gao Y, Lau VY, Zheng W, Leung TY, et al. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci U S A. 2008;105(51):20458-63.
89. Norton ME, Jacobsson B, Swamy GK, Laurent LC, Ranzini AC, Brar H, et al. Cell-free DNA analysis for noninvasive examination of trisomy. The New England journal of medicine. 2015;372(17):1589-97.
90. Larion S, Warsof SL, Romary L, Mlynarczyk M, Peleg D, Abuhamad AZ. Association of combined first-trimester screen and noninvasive prenatal testing on diagnostic procedures. Obstetrics and gynecology. 2014;123(6):1303-10.
91. Bianchi DW, Parker RL, Wentworth J, Madankumar R, Saffer C, Das AF, et al. DNA sequencing versus standard prenatal aneuploidy screening. The New England journal of medicine. 2014;370(9):799-808.

92. Gil MM, Quezada MS, Bregant B, Ferraro M, Nicolaides KH. Implementation of maternal blood cell-free DNA testing in early screening for aneuploidies. Ultrasound Obstet Gynecol. 2013;42(1):34-40.
93. Lo YM, Chan KC, Sun H, Chen EZ, Jiang P, Lun FM, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med. 2010;2(61):61ra91.
94. Peters D, Chu T, Yatsenko SA, Hendrix N, Hogge WA, Surti U, et al. Noninvasive prenatal diagnosis of a fetal microdeletion syndrome. The New England journal of medicine. 2011;365(19):1847-8.
95. Zhao C, Tynan J, Ehrich M, Hannum G, McCullough R, Saldivar JS, et al. Detection of fetal subchromosomal abnormalities by sequencing circulating cell-free DNA from maternal plasma. Clinical chemistry. 2015;61(4):608-16.
96. Chan KC, Jiang P, Sun K, Cheng YK, Tong YK, Cheng SH, et al. Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. Proc Natl Acad Sci U S A. 2016;113(50):E8159-e68.
97. Bianchi DW, Chudova D, Sehnert AJ, Bhatt S, Murray K, Prosen TL, et al. Noninvasive Prenatal Testing and Incidental Detection of Occult Maternal Malignancies. Jama. 2015;314(2):162-9.
98. Morris S, Karlsen S, Chung N, Hill M, Chitty LS. Model-based analysis of costs and outcomes of non-invasive prenatal testing for Down's syndrome using cell free fetal DNA in the UK National Health Service. PLoS One. 2014;9(4):e93559.
99. Chitty LS, Wright D, Hill M, Verhoef TI, Daley R, Lewis C, et al. Uptake, outcomes, and costs of implementing non-invasive prenatal testing for Down's syndrome into NHS maternity care: prospective cohort study in eight diverse maternity units. BMJ (Clinical research ed). 2016;354:i3426.
100. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ. Free DNA in the serum of cancer patients and the effect of therapy. Cancer research. 1977;37(3):646-50.
101. Szpechcinski A, Chorostowska-Wynimko J, Struniawski R, Kupis W, Rudzinski P, Langfort R, et al. Cell-free DNA levels in plasma of patients with non-small-cell lung cancer and inflammatory lung disease. Br J Cancer. 2015;113(3):476-83.
102. Fatouros IG, Destouni A, Margonis K, Jamurtas AZ, Vrettou C, Kouretas D, et al. Cell-free plasma DNA as a novel marker of aseptic inflammation severity related to exercise overtraining. Clinical chemistry. 2006;52(9):1820-4.
103. Zhong XY, Von Muhlenen I, Li Y, Kang A, Gupta AK, Tyndall A, et al. Increased concentrations of antibody-bound circulatory cell-free DNA in rheumatoid arthritis. Clinical chemistry. 2007;53(9):1609-14.
104. Ye L, Ma GH, Chen L, Li M, Liu JL, Yang K, et al. Quantification of circulating cell-free DNA in the serum of patients with obstructive sleep apnea-hypopnea syndrome. Lung. 2010;188(6):469-74.
105. Huttunen R, Kuparinen T, Jylhava J, Aittoniemi J, Vuento R, Huhtala H, et al. Fatal Outcome in Bacteremia is Characterised by High Plasma Cell Free DNA Concentration and Apoptotic DNA Fragmentation: A Prospective Cohort Study. PLoS ONE. 2011;6(7):e21700.
106. Chan KC, Jiang P, Chan CW, Sun K, Wong J, Hui EP, et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proc Natl Acad Sci U S A. 2013;110(47):18761-8.
107. Jamal-Hanjani M, Wilson GA, Horswell S, Mitter R, Sakarya O, Constantin T, et al. Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2016;27(5):862-7.

108. Mouliere F, El Messaoudi S, Pang D, Dritschilo A, Thierry AR. Multi-marker analysis of circulating cell-free DNA toward personalized medicine for colorectal cancer. Molecular oncology. 2014;8(5):927-41.
109. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. 2016;164(1-2):57-68.
110. Jahr S, Hentze H, Englisch S, Hardt D, Fackelmayer FO, Hesch RD, et al. DNA Fragments in the Blood Plasma of Cancer Patients: Quantitations and Evidence for Their Origin from Apoptotic and Necrotic Cells. Cancer research. 2001;61(4):1659-65.
111. Wu T, Zhang D, Chia J, Tsao K, Sun C, Wu J. Cell-free DNA: measurement in various carcinomas and establishment of normal reference range. Clinica chimica acta; international journal of clinical chemistry. 2002;321:77-87.
112. Ivanov M, Baranova A, Butler T, Spellman P, Mileyko V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. BMC Genomics. 2015;16 Suppl 13:S1.
113. Heitzer E, Auer M, Hoffmann EM, Pichler M, Gasch C, Ulz P, et al. Establishment of tumor-specific copy number alterations from plasma DNA of patients with cancer. Int J Cancer. 2013;133(2):346-56.
114. Jiang P, Chan CW, Chan KC, Cheng SH, Wong J, Wong VW, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc Natl Acad Sci U S A. 2015;112(11):E1317-25.
115. Diehl F, Meng L, Dressman D, He Y, Shen D, Szabo S, et al. Detection and quantification of mutations in the plasma of patients with colorectal tumours. Proc Natl Acad Sci USA. 2005;102(45):16368-73.
116. Mouliere F, Rosenfeld N. Circulating tumor-derived DNA is shorter than somatic DNA in plasma. Proc Natl Acad Sci U S A. 2015;112(11):3178-9.
117. Benesova L, Belsanova B, Suchanek S, Kopeckova M, Minarikova P, Lipska L, et al. Mutation-based detection and monitoring of cell-free tumor DNA in peripheral blood of cancer patients. Anal Biochem. 2013;433(2):227-34.
118. Stroun M, Lyautey J, Lederrey C, Olson-Sand A, Anker P. About the possible origin and mechanism of circulating DNA Apoptosis and active DNA release. Clinica chimica acta; international journal of clinical chemistry. 2001;313(1-2):139-42.
119. Wang B, Huang H-Y, Chen Y-C, Bristow R, Kassauei K, Cheng C-C, et al. Increased Plasma DNA Integrity in Cancer Patients. Cancer research. 2003;63:3966-8.
120. D'Souza-Schorey C, Clancy JW. Tumor-derived microvesicles: shedding light on novel microenvironment modulators and prospective cancer biomarkers. Genes & development. 2012;26(12):1287-99.
121. Katsuda T, Kosaka N, Ochiya T. The roles of extracellular vesicles in cancer biology: toward the development of novel cancer biomarkers. Proteomics. 2014;14(4-5):412-25.
122. Valadi H, Ekstrom K, Bossios A, Sjostrand M, Lee JJ, Lotvall JO. Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nature cell biology. 2007;9(6):654-9.
123. Thakur BK, Zhang H, Becker A, Matei I, Huang Y, Costa-Silva B, et al. Double-stranded DNA in exosomes: a novel biomarker in cancer detection. Cell research. 2014;24(6):766-9.
124. Jin X, Chen Y, Chen H, Fei S, Chen D, Cai X, et al. Evaluation of tumor-derived exosomal miRNA as potential diagnostic biomarkers for early stage non-small-cell lung cancer using next-generation sequencing. Clinical cancer research : an official journal of the American Association for Cancer Research. 2017.
125. Fernando MR, Jiang C, Krzyzanowski GD, Ryan WL. New evidence that a large proportion of human blood plasma cell-free DNA is localized in exosomes. PLoS One. 2017;12(8):e0183915.

126. Crowley E, Di Nicolantonio F, Loupakis F, Bardelli A. Liquid biopsy: monitoring cancer-genetics in the blood. Nature reviews Clinical oncology. 2013;10(8):472-84.
127. Allard WJ, Matera J, Miller MC, Repollet M, Connelly MC, Rao C, et al. Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. Clinical cancer research : an official journal of the American Association for Cancer Research. 2004;10(20):6897-904.
128. Diehl F, Schmidt K, Choti MA, Romans K, Goodman S, Li M, et al. Circulating mutant DNA to assess tumour dynamics. Nature medicine. 2008;14(9):985-90.
129. Tie J, Kinde I, Wang Y, Wong HL, Roebert J, Christie M, et al. Circulating tumor DNA as an early marker of therapeutic response in patients with metastatic colorectal cancer. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2015.
130. Wang Z, Chen R, Wang S, Zhong J, Wu M, Zhao J, et al. Quantification and dynamic monitoring of EGFR T790M in plasma cell-free DNA by digital PCR for prognosis of EGFR-TKI treatment in advanced NSCLC. PLoS One. 2014;9(11):e110780.
131. Murtaza M, Dawson SJ, Tsui D, Gale D, Forshew T, Piskorz AM, et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature. 2013;497:108-12.
132. Paci M, Maramotti S, Bellesia E, Formisano D, Albertazzi L, Ricchetti T, et al. Circulating plasma DNA as a diagnostic biomarker in non-small cell lung cancer. Lung cancer (Amsterdam, Netherlands). 2009;64(1):92-7.
133. ECMC and NCRI Biomarkers & Imaging Clinical Studies Group. Cell-free DNA concensus meeting and report 2014 [Available from: http://www.ecmcnetwork.org.uk/cfdna-consensus-meeting-and-report.
134. El Messaoudi S, Rolet F, Mouliere F, Thierry AR. Circulating Cell-free DNA: Preanalytical considerations. Clinica chimica acta; international journal of clinical chemistry. 2013;424:222-30.
135. Page K, Guttery DS, Zahra N, Primrose L, Elshaw SR, Pringle JH, et al. Influence of Plasma Processing on Recovery and Analysis of Circulating Nucleic Acids. PLoS ONE. 2013;8(10).
136. Lee T, Montalvo L, Chrebtow V, Busch M. Quantitation of genomic DNA in plasma and serum samples: higher concentrations of genomic DNA found in serum than in plasma. Immunohematology. 2001;41:276-82.
137. Board RE, Williams VS, Knight L, Shaw J, Greystoke A, Ranson M, et al. Isolation and Extraction of Circulating Tumour DNA from patients with Small Cell Lung Cancer. Ann NY Acad Sci. 2008;1137:98-107.
138. Gautschi O, Bigosch C, Huegli B, Jermann M, Marx A, Chasse E, et al. Circulating Deoxyribonucleic Acid As a Prognostic Marker in Non-Small-Cell Lung Cancer Patients Undergoing Chemotherapy. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2004;22(20):4157-64.
139. Lam N, Rainer TH, Chiu R, Lo YMD. EDTA Is a Better Anticoagulant than Heparin or Citrate for Delayed Blood Processing for Plasma DNA Analysis. Clinical chemistry. 2004;50(1):256-7.
140. Xue X, Teare MD, Holen I, Zhu YM, Woll P. Optimizing the yield and utility of circulating cell-free DNA from plasma and serum. Clinica chimica acta; international journal of clinical chemistry. 2009;404(2):100-4.
141. van Dessel LF, Beije N, Helmijr JC, Vitale SR, Kraan J, Look MP, et al. Application of circulating tumor DNA in prospective clinical oncology trials - standardization of preanalytical conditions. Molecular oncology. 2017;11(3):295-304.
142. Parpart-Li S, Bartlett B, Popoli M, Adleff V, Tucker L, Steinberg R, et al. The Effect of Preservative and Temperature on the Analysis of Circulating Tumor DNA. Clinical cancer

research : an official journal of the American Association for Cancer Research. 2017;23(10):2471-7.
143. Kang Q, Henry NL, Paoletti C, Jiang H, Vats P, Chinnaiyan AM, et al. Comparative analysis of circulating tumor DNA stability In K3EDTA, Streck, and CellSave blood collection tubes. Clinical biochemistry. 2016;49(18):1354-60.
144. Rothwell DG, Smith N, Morris D, Leong HS, Li Y, Hollebecque A, et al. Genetic profiling of tumours using both circulating free DNA and circulating tumour cells isolated from the same preserved whole blood sample. Molecular oncology. 2016;10(4):566-74.
145. Sherwood JL, Corcoran C, Brown H, Sharpe AD, Musilova M, Kohlmann A. Optimised Pre-Analytical Methods Improve KRAS Mutation Detection in Circulating Tumour DNA (ctDNA) from Patients with Non-Small Cell Lung Cancer (NSCLC). PLoS One. 2016;11(2):e0150197.
146. Schmidt B, Weickmann S, Witt C, Fleischhacker M. Improved method for isolating cell-free DNA. Clinical chemistry. 2005;51(8):1561-3.
147. Kirsch C, Weickmann S, Schmidt B, Fleischhacker M. An improved method for the isolation of free-circulating plasma DNA and cell-free DNA from other body fluids. Ann NY Acad Sci. 2008;1137:135-9.
148. Stemmer C, Beau-Faller M, Pencreac'h E, Guerin E, Schneider A, Jaqmin D, et al. Use of magnetic beads for plasma cell-free DNA extraction: toward automation of plasma DNA analysis for molecular diagnostics. Clinical chemistry. 2003;49(11):1953-5.
149. Fleischhacker M, Schmidt B, Weickmann S, Fersching D, Leszinski G, Siegele B, et al. Methods for isolation of cell-free plasma DNA strongly affect DNA yield. Clinica chimica acta; international journal of clinical chemistry. 2011;412:2085-8.
150. Malentacchi F, Pizzamiglio S, Verderio P, Pazzagli M, Orlando C, Ciniselli CM, et al. Influence of storage conditions and extraction methods on the quantity and quality of circulating cell-free DNA (ccfDNA): the SPIDIA-DNAplas External Quality Assessment experience. Clinical chemistry and laboratory medicine. 2015;53(12):1935-42.
151. Nygaard A, Garm Spindler K-L, Pallisgaard N, Andersen R, Jakobsen A. The prognostic value of KRAS mutated plasma DNA in advanced non-small cell lung cancer. Lung cancer (Amsterdam, Netherlands). 2013;79(3):312-7.
152. Sozzi G, Conte D, Leon M, Cirincione R, Roz L, Ratcliffe C, et al. Quantification of Free Circulating DNA As a Diagnostic Marker in Lung Cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2003;21(21):3902-8.
153. Sedlackova T, Repiska G, Celec P, Szemes T, Minarik G. Fragmentation of DNA affects the accuracy of the DNA quantitation by the commonly used methods. Biological Procedures Online. 2013;15(1):5.
154. Walker JA, Kilroya GE, Xing J, Shewale J, Sinha SK, Batzer MA. Human DNA quantitation using Alu element-based polymerase chain reaction. Anal Biochem. 2003;315(1):122-8.
155. Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. Genome Res. 1996;6(10):986-94.
156. VanGuilder HD, Vrana KE, Freeman WM. Twenty-five years of quantitative PCR for gene expression analysis. Biotechniques. 2008;44(5):619-26.
157. Sozzi G, Conte D, Mariani L, Lo Vullo S, Roz L, Lomardo C, et al. Analysis of Circulating Tumour DNA in Plasma at Diagnosis and during Follow-up of Lung Cancer Patients. Cancer research. 2001;61(12):4675-8.
158. Taylor F, Teare MD, Cox A, Woll P. Circulating cell-free DNA: a potential biomarker in lung cancer. Lung Cancer Management. 2013;2(5):407-22.
159. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature. 2017;545(7655):446-51.

160. Newman AM, Bratman SV, To J, Wynne JF, Eclov NCW, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumour DNA with broad patient coverage. Nature medicine. 2014;13(5):528-38.
161. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014;6(224):224ra24.
162. He C, Liu M, Zhou C, Zhang J, Ouyang M, Zhong N, et al. Detection of epidermal growth factor receptor mutations in plasma by mutant-enriched PCR assay for prediction of the repsonse to gefitinib in patients with non-small-cell lung cancer. Int J Cancer. 2009;125(10):2393-9.
163. Kimura H, Kasahara K, Kawaishi M, Kunitoh H, Tamura T, Holloway B, et al. Detection of epidermal growth factor receptor mutations in serum as a predictor of the response to gefitinib in patients with non-small cell lung cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2006;12(13):3915-21.
164. Narayan A, Carriero N, Gettinger SN, Kluytenaar J, Kozak KR, Yock TI, et al. Ultrasensitive Measurement of Hotspot Mutations in Tumor DNA in Blood Using Error-Suppressed Multiplexed Deep Sequencing. Cancer research. 2012;72(14):3492-8.
165. Leary RJ, Kinde I, Diehl F, Schmidt K, Clouser C, Duncan C, et al. Development of personalized tumor biomarkers using massively parallel sequencing. Sci Transl Med. 2010;2(20):20ra14.
166. Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DWY, Kaper F, et al. Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA. Sci Transl Med. 2012;4(136):136ra68.
167. Leary RJ, Sausen M, Kinde I, Papadopoulos N, Carpten JD, Craig D, et al. Detection of Chromosomal Alterations in the Circulstion of Cancer Patients with Whole-Genome Sequencing. Sci Transl Med. 2012;4(162):162ra54.
168. Chan KC, Jiang P, Zheng Y, Liao G, Sun H, Wong J, et al. Cancer Genome Scanning in Plasma: Detection of Tumor-Associated Copy Number Aberrations, Single-Nucleotide Variants, and Tumoural Heterogeneity by Massive Parallel Sequencing. Clinical chemistry. 2013;59(1):211-24.
169. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol. 2016;34(5):547-55.
170. Oxnard GC, PP. Kuang, Y. Mach, SL. O'Connell, A. Messineo, MM. Luke, JJ. Butany, M. Kirschmeier, P. Jackman, DM. Janne, PA. Noninvasive detection of response and resistance in EGFR-mutant lung cancer using quantitative next-generation genotyping of cell-free plasma DNA. Clinical cancer research : an official journal of the American Association for Cancer Research. 2014;20(6):1698-705.
171. Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ, et al. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. Proc Natl Acad Sci U S A. 2016;113(21):6005-10.
172. Kimura H, Suminoe M, Kasahara K, Sone T, Araya T, Tamori S, et al. Evaluation of epidermal growth factor receptor mutation status in serum DNA as a predictor of response to gefitinib (IRESSA). Br J Cancer. 2007;97(6):778-84.
173. Taniguchi K, Uchida J, Nishino K, Kumagai T, Okuyama T, Okami J, et al. Quantitative detection of EGFR mutations in circulating tumor DNA derived from lung adenocarcinomas. Clinical cancer research : an official journal of the American Association for Cancer Research. 2011;17(24):7808-15.
174. Bai H, Mao L, Wang HS, Zhao J, Yang L, An T, et al. Epidermal growth factor receptor mutations in plasma DNA samples predict tumor response in Chinese patients with stages

IIIB to IV non-small-cell lung cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2009;27(16):2653-9.
175. Mack P, Holland W, Burich R, Sangha R, Solis L, Li Y, et al. EGFR Mutations Detected in Plasma Are Associated with Patient Outcomes in Erlotinib Plus Docetaxel-Treated Non-Small Cell Lung Cancer. J Thorac Oncol. 2009;4(12):1466-72.
176. Azad AA, Volik SV, Wyatt AW, Haegert A, Le Bihan S, Bell RH, et al. Androgen Receptor Gene Aberrations in Circulating Cell-Free DNA: Biomarkers of Therapeutic Resistance in Castration-Resistant Prostate Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2015;21(10):2315-24.
177. Wyatt AW, Azad AA, Volik SV, Annala M, Beja K, McConeghy B, et al. Genomic Alterations in Cell-Free DNA and Enzalutamide Resistance in Castration-Resistant Prostate Cancer. JAMA oncology. 2016.
178. Heitzer E, Ulz P, Belic J, Gutschi S, Quehenberger F, Fischereder K, et al. Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. Genome medicine. 2013;5(4):30.
179. Heidary M, Auer M, Ulz P, Heitzer E, Petru E, Gasch C, et al. The dynamic range of circulating tumor DNA in metastatic breast cancer. Breast Cancer Res. 2014;16(4):1-10.
180. Shaw J, Page K, Blighe K, Hava N, Guttery D, Wad B, et al. Genomic analysis of circulating cell free DNA infers breast cancer dormancy. Genome Res. 2011;22(2):220-31.
181. Dawson SJ, Tsui D, Murtaza M, Biggs H, Rueda O, Chin S, et al. Analysis of Circulating Tumour DNA to Monitor Metastatic Breast Cancer. The New England journal of medicine. 2013;368(13):1199-209.
182. Hiley CT, Le Quesne J, Santis G, Sharpe R, de Castro DG, Middleton G, et al. Challenges in molecular testing in non-small-cell lung cancer patients with advanced disease. Lancet (London, England). 2016;388(10048):1002-11.
183. Barlesi F, Mazieres J, Merlio JP, Debieuvre D, Mosser J, Lena H, et al. Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT). Lancet (London, England). 2016;387(10026):1415-26.
184. Spicer J, Tisher B, Peters M. EGFR mutation testing and oncologist treatment choice in advanced NSCLC: global trends and differences. . Ann Oncol (suppl 1). 2015;26:i57-i61.
185. Punnoose E, Atwal S, Liu W, Raja R, Fine BM, Hughes BGM, et al. Evaluation of Circulating Tumour Cells and Circulating Tumour DNA in Non-Small Cell Lung Cancer: Association with Clinical Endpoints in a Phase II Clinical Trial of Pertuzumab and Erlotinib. Clinical cancer research : an official journal of the American Association for Cancer Research. 2012;18(8):2391-401.
186. Rosell R, Moran T, Queralt C, Porta R, Cardenal F, Camps C, et al. Screening for Epidermal Growth Factor Receptor Mutations in Lung Cancer. The New England journal of medicine. 2009;361(10):958-67.
187. Brevet M, Johnson M, Azzoli C, Ladanyi M. Detection of EGFR mutations in plasma DNA from lung cancer patients by mass spectrometry genotyping is predictive of tumour EGFR status and response to EGFR inhibitors. Lung cancer (Amsterdam, Netherlands). 2011;73(1):96-102.
188. Reck M, Hagiwara K, Han B, Tjulandin S, Grohe C, Yokoi T, et al. ctDNA Determination of EGFR Mutation Status in European and Japanese Patients with Advanced NSCLC: The ASSESS Study. J Thorac Oncol. 2016;11(10):1682-9.
189. Bernabe R, Hickson N, Wallace A, Blackhall FH. What do we need to make circulating tumour DNA (ctDNA) a routine diagnostic test in lung cancer? European journal of cancer (Oxford, England : 1990). 2017;81:66-73.
190. Wang S, An T, Wang J, Zhao J, Wang Z, Zhuo M, et al. Potential clinical significance of a plasma-based KRAS mutation analysis in patients with advanced non-small cell lung cancer.

Clinical cancer research : an official journal of the American Association for Cancer Research. 2010;16(4):1324-30.
191. Douillard JY, Ostoros G, Cobo M, Ciuleanu T, Cole R, McWalter G, et al. Gefitinib treatment in EGFR mutated caucasian NSCLC: circulating-free tumor DNA as a surrogate for determination of EGFR status. J Thorac Oncol. 2014;9(9):1345-53.
192. Malapelle U, Sirera R, Jantus-Lewintre E, Reclusa P, Calabuig-Farinas S, Blasco A, et al. Profile of the Roche cobas(R) EGFR mutation test v2 for non-small cell lung cancer. Expert review of molecular diagnostics. 2017;17(3):209-15.
193. Oxnard GR, Thress KS, Alden RS, Lawrance R, Paweletz CP, Cantarini M, et al. Association Between Plasma Genotyping and Outcomes of Treatment With Osimertinib (AZD9291) in Advanced Non-Small-Cell Lung Cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2016;34(28):3375-82.
194. Mok T, Wu YL, Lee JS, Yu CJ, Sriuranpong V, Sandoval-Tan J, et al. Detection and Dynamic Changes of EGFR Mutations from Circulating Tumor DNA as a Predictor of Survival Outcomes in NSCLC Patients Treated with First-line Intercalated Erlotinib and Chemotherapy. Clinical cancer research : an official journal of the American Association for Cancer Research. 2015;21(14):3196-203.
195. Chabon JJ, Simmons AD, Lovejoy AF, Esfahani MS, Newman AM, Haringsma HJ, et al. Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. Nature communications. 2016;7:11815.
196. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014;346(6206):251-6.
197. Tie J, Wang Y, Tomasetti C, Li L, Springer S, Kinde I, et al. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. Sci Transl Med. 2016;8(346):346ra92.
198. Garcia-Murillas I, Schiavon G, Weigelt B, Ng C, Hrebien S, Cutts RJ, et al. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. Sci Transl Med. 2015;7(302):302ra133.
199. Szpechcinski A, Rudzinski P, Kupis W, Langfort R, Orlowski T, Chorostowska-Wynimko J. Plasma cell-free DNA levels and integrity in patients with chest radiological findings: NSCLC versus benign lung nodules. Cancer Lett. 2016;374(2):202-7.
200. Sozzi G, Roz L, Conte D, Mariani L, Andriani F, Lo Vullo S, et al. Plasma DNA Quantification in Lung Cancer Computed Tomography Screening. Am J Res Crit Care. 2009;179(1):69-74.
201. Sozzi G, Roz L, Conte D, Mariani L, Andriani F, Verderio P, et al. Effects of Prolonged Storage of Whole Plasma or Isolated Plasma DNA on the Results of Circulating DNA Quantification Assays. Journal of the National Cancer Institute. 2005;97(24):1848-50.
202. Ulivi P, Mercatali L, Casoni G-L, Scarpi E, Bucchi L, Silvestrini R, et al. Multiple marker detection in peripheral blood for NSCLC diagnosis. PLoS ONE. 2013;8(2):e57401.
203. Catarino R, Coelho A, Araujo A, Gomes M, Nogueira A, Lopes C, et al. Circulating DNA: Diagnostic Tool and Predictive Marker for Overall Survival of NSCLC Patients. PLoS One. 2012;7(6):e38559.
204. Van der Drift MA, Hol BEA, Klaassen CHW, Prinsenc CFM, van Aarssend YAWG, Donderse R, et al. Circulating DNA is a non-invasive prognostic factor for survival in non-small cell lung cancer. Lung cancer (Amsterdam, Netherlands). 2010;68:283-7.
205. Kumar S, Guleria R, Singh V, Bharti AC, Mohan A, Das BC. Efficacy of circulating plasma DNA as a diagnostic tool for advanced non-small cell lung cancer and its predictive utility for survival and response to chemotherapy. Lung cancer (Amsterdam, Netherlands). 2010;70(2):211-7.

206. Szpechcinski A, Dancewicz M, Kopinski P, Kowalewski J, Chorostowska-Wynimko J. Real-time PCR quantification of plasma DNA in non-small cell lung cancer patients and healthy controls. Eur J Med Res. 2009;7(14):237-40.
207. Yoon SO, Park S, Lee SH, Kim JH, S. LJ. Comparison of circulating plasma DNA levels between lung cancer patients and healthy controls. The Journal of molecular diagnostics : JMD. 2011;11(3):182-5.
208. Ludovini V, Pistola L, Gregorc V, Floriani I, Rulli E, Piattoni S, et al. Plasma DNA, microsatellite alterations, and p53 tumor mutations are associated with disease-free survival in radically resected non-small cell lung cancer patients: a study of the perugia multidisciplinary team for thoracic oncology. J Thorac Oncol. 2008;3(4):365-73.
209. Herrera LJ, Raja S, Gooding WE, El-Hefnawy T, Kelly L, Luketich JD, et al. Quantitative analysis of circulating plasma DNA as a tumor marker in thoracic malignancies. Clinical chemistry. 2005;51(1):113-8.
210. Guang-shun X, Ai-rong H, Long-yun L, Yan-ning G, Shu-jun C. Quantification of plasma DNA as a screening tool for lung cancer. Chinese Medical J. 2004;117(10):1485-8.
211. Hulbert A, Jusue-Torres I, Stark A, Chen C, Rodgers K, Lee B, et al. Early Detection of Lung Cancer Using DNA Promoter Hypermethylation in Plasma and Sputum. Clinical cancer research : an official journal of the American Association for Cancer Research. 2017;23(8):1998-2005.
212. Vinayanuwattikun C, Sriuranpong V, Tanasanvimon S, Chantranuwat P, Mutirangura A. Epithelial-Specific Methylation Marker: A Potential Plasma Biomarker in Advanced Non-small Cell Lung Cancer. J Thorac Oncol. 2011;6(11):1818-25.
213. Zhang R, Shao F, Wu X, Ying K. Value of quantitative analysis of circulating cell free DNA as a screening tool for lung cancer: A meta-analysis. Lung cancer (Amsterdam, Netherlands). 2010;69(2):225-31.
214. Ostrow KL, Hoque MO, Loyo M, Brait M, Greenberg A, Siegfried J, et al. Molecular Analysis of Plasma DNA for the Early Detection of Lung Cancer by Quantitative Methylation Specific PCR. Clinical cancer research : an official journal of the American Association for Cancer Research. 2010;16(13):3463-72.
215. Wielscher M, Vierlinger K, Kegler U, Ziesche R, Gsur A, Weinhausel A. Diagnostic Performance of Plasma DNA Methylation Profiles in Lung Cancer, Pulmonary Fibrosis and COPD. EBioMedicine. 2015;2(8):929-36.
216. Maheswaran S, Sequist LV, Nagrath S, Ulkus L, Brannigan B, Collura CV, et al. Detection of mutations in EGFR in circulating lung cancer cells. The New England journal of medicine. 2008;359(4):366-77.
217. Tanaka F, Yoneda K, Kondo N, Hashimoto M, Takuwa T, Matsumoto S, et al. Circulating tumor cell as a diagnostic marker in primary lung cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2009;15(22):6980-6.
218. Sozzi G, Boeri M, Rossi M, Verri C, Suatoni P, Bravi F, et al. Clinical utility of a plasma-based miRNA signature classifier within computed tomography lung cancer screening: a correlative MILD trial study. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2014;32(8):768-73.
219. Sin DD, Tammemagi CM, Lam S, Barnett MJ, Duan X, Tam A, et al. Pro-surfactant protein B as a biomarker for lung cancer prediction. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013;31(36):4536-43.
220. Boyle P, Chapman CJ, Holdenrieder S, Murray A, Robertson C, Wood WC, et al. Clinical validation of an autoantibody test for lung cancer. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2011;22(2):383-9.
221. Diaz LA, Bardelli A. Liquid biopsies: genotyping circulating tumor DNA. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2014;32(6):579-86.

222. O' Flaherty JD, Gray S, Richard D, Fennell D, O'Leary JJ, Blackhall FH, et al. Circulating tumour cells, their role in metastasis and their clinical utility in lung cancer. Lung cancer (Amsterdam, Netherlands). 2012;76(1):19-25.
223. Dive C, Brady G. SnapShot: Circulating Tumor Cells. Cell. 2017;168(4):742-.e1.
224. Alex-Panabieres C, Schwarzenbach H, Pantel K. Circulating Tumour Cells and Circulating Tumour DNA. Ann Rev Med. 2012;63:199-215.
225. Luke JJ, Oxnard GR, Paweletz CP, Camidge DR, Heymach JV, Solit DB, et al. Realizing the potential of plasma genotyping in an age of genotype-directed therapies. Journal of the National Cancer Institute. 2014;106(8).
226. Hodgkinson CL, Morrow CJ, Li Y, Metcalf RL, Rothwell DG, Trapani F, et al. Tumorigenicity and genetic profiling of circulating tumor cells in small-cell lung cancer. Nature medicine. 2014;20(8):897-903.
227. Cristofanilli M, Budd GT, Ellis MJ, Stopeck A, Metera J, Miller M, et al. Circulating tumour cells, disease progression, and survival in metastatic breast cancer. The New England journal of medicine. 2004;351(8):781-91.
228. Grover PK, Cummins AG, Price TJ, Roberts-Thomson IC, Hardingham JE. Circulating tumour cells: the evolving concept and the inadequacy of their enrichment by EpCAM-based methodology for basic and clinical cancer research. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2014;25(8):1506-16.
229. Ilie M, Hofman V, Long-Mira E, Selva E, Vignaud JM, Padovani B, et al. "Sentinel" circulating tumor cells allow early diagnosis of lung cancer in patients with chronic obstructive pulmonary disease. PLoS One. 2014;9(10):e111597.
230. Babayan A, Alawi M, Gormley M, Muller V, Wikman H, McMullin RP, et al. Comparative study of whole genome amplification and next generation sequencing performance of single cancer cells. Oncotarget. 2016.
231. Esquela-Kerscher A, Slack FJ. Oncomirs - microRNAs with a role in cancer. Nature reviews Cancer. 2006;6(4):259-69.
232. Montani F, Bianchi F. Circulating Cancer Biomarkers: The Macro-revolution of the Micro-RNA. EBioMedicine. 2016;5:4-6.
233. Boeri M, Verri C, Conte D, Roz L, Modena P, Facchinetti F, et al. MicroRNA signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. Proc Natl Acad Sci U S A. 2011;108(9):3713-8.
234. Montani F, Marzi MJ, Dezi F, Dama E, Carletti RM, Bonizzi G, et al. miR-Test: a blood test for lung cancer early detection. Journal of the National Cancer Institute. 2015;107(6):djv063.
235. Chapman CJ, Murray A, McElveen JE, Sahin U, Luxemburger U, Tureci O, et al. Autoantibodies in lung cancer: possibilities for early detection and subsequent cure. Thorax. 2008;63(3):228-33.
236. Murray A, Chapman CJ, Healey G, Peek LJ, Parsons G, Baldwin D, et al. Technical validation of an autoantibody test for lung cancer. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2010;21(8):1687-93.
237. Chapman CJ, Healey GF, Murray A, Boyle P, Robertson C, Peek LJ, et al. EarlyCDT(R)-Lung test: improved clinical utility through additional autoantibody assays. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine. 2012;33(5):1319-26.
238. Jett JR, Peek LJ, Fredericks L, Jewell W, Pingleton WW, Robertson JF. Audit of the autoantibody test, EarlyCDT(R)-lung, in 1600 patients: an evaluation of its performance in routine clinical practice. Lung cancer (Amsterdam, Netherlands). 2014;83(1):51-5.
239. Massion PP, Healey GF, Peek LJ, Fredericks L, Sewell HF, Murray A, et al. Autoantibody Signature Enhances the Positive Predictive Power of Computed Tomography

and Nodule-Based Risk Models for Detection of Lung Cancer. J Thorac Oncol. 2017;12(3):578-84.
240. Sullivan FM, Farmer E, Mair FS, Treweek S, Kendrick D, Jackson C, et al. Detection in blood of autoantibodies to tumour antigens as a case-finding method in lung cancer using the EarlyCDT(R)-Lung Test (ECLS): study protocol for a randomized controlled trial. BMC cancer. 2017;17(1):187.
241. ECLS study. Early Cancer detection test- Lung cancer Scotland 2017 [Available from: www.eclsstudy.org.
242. Taguchi A, Politi K, Pitteri SJ, Lockwood WW, Faca VM, Kelly-Spratt K, et al. Lung cancer signatures in plasma based on proteome profiling of mouse tumor models. Cancer cell. 2011;20(3):289-99.
243. Wikoff WR, Hanash S, DeFelice B, Miyamoto S, Barnett M, Zhao Y, et al. Diacetylspermine Is a Novel Prediagnostic Serum Biomarker for Non-Small-Cell Lung Cancer and Has Additive Performance With Pro-Surfactant Protein B. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2015;33(33):3880-6.
244. Mathe EA, Patterson AD, Haznadar M, Manna SK, Krausz KW, Bowman ED, et al. Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. Cancer research. 2014;74(12):3259-70.
245. Haznadar M, Cai Q, Krausz KW, Bowman ED, Margono E, Noro R, et al. Urinary Metabolite Risk Biomarkers of Lung Cancer: A Prospective Cohort Study. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2016;25(6):978-86.
246. Krilaviciute A, Heiss JA, Leja M, Kupcinskas J, Haick H, Brenner H. Detection of cancer through exhaled breath: a systematic review. Oncotarget. 2015;6(36):38643-57.
247. The ReSoLuCENT study 2017 [Available from: http://resolucent.group.shef.ac.uk/.
248. D'Amelio AM, Jr., Cassidy A, Asomaning K, Raji OY, Duffy SW, Field JK, et al. Comparison of discriminatory power and accuracy of three lung cancer risk models. Br J Cancer. 2010;103(3):423-9.
249. Page K, Powles T, Slade MJ, Tamburo De Bella M, Walker RA, Coombes R, et al. The Importance of Careful Blood Processing in Isolation of Cell-Free DNA. Ann NY Acad Sci. 2006;1075:313-7.
250. Chan KCA, Lo YMD. Circulating tumour-derived nucleic acids in cancer patients: potential applications as tumour markers. Br J Cancer. 2007;96(5):681-5.
251. Children's Oncology Group Cell Culture and Xenograft Repository. COG Cell Line and Xenograft STR Database 2016 [Available from: http://strdb.cogcell.org/.
252. Covaris. DNA shearing with E220 Focused-ultrasonicator 2013 [Available from: http://covarisinc.com/wp-content/uploads/pn_010308.pdf.
253. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.
254. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
255. Picard. picard-tools-2.1.0 2016 [Available from: https://github.com/broadinstitute/picard/releases/tag/2.1.0.
256. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45.
257. PERL script 2012 [Available from: precancer.leeds.ac.uk/software-and-datasets/cnanorm.

258. Genome Bioinformatics Group of UC Santa Cruz. Mappability or Uniqueness of Reference Genome from ENCODE 2016 [Available from: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability.
259. Pinheiro J, Bates D, DebRoy S, Sarkar D, Team RC. Linear and Nonlinear Mixed Effects Models. R package version 3.1-124 2016 [Available from: http://CRAN.R-project.org/package=nlme.
260. Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics. 2012;28(1):40-7.
261. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12(2):115-21.
262. Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, et al. 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. 2010;38(14):e151.
263. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557-72.
264. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463(7283):899-905.
265. Xia S, Huang CC, Le M, Dittmar R, Du M, Yuan T, et al. Genomic variations in plasma cell free DNA differentiate early stage lung cancers from normal controls. Lung cancer (Amsterdam, Netherlands). 2015;90(1):78-84.
266. Xia S, Kohli M, Du M, Dittmar RL, Lee A, Nandy D, et al. Plasma genetic and genomic abnormalities predict treatment response and clinical outcome in advanced prostate cancer. Oncotarget. 2015;6(18):16411-21.
267. Grady DL, Ratliff RL, Robinson DL, McCanlies EC, Meyne J, Moyzis RK. Highly conserved repetitive DNA sequences are present at human centromeres. Proc Natl Acad Sci U S A. 1992;89(5):1695-9.
268. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.
269. UCLA: Statsitical Consulting Group 2016 [Available from: http://www.ats.ucla.edu/stat/stata/seminars/stata_survival/.
270. Beasley MB, Brambilla E, Travis WD. The 2004 World Health Organization classification of lung tumors. Seminars in roentgenology. 2005;40(2):90-7.
271. Simmons CP, Koinis F, Fallon MT, Fearon KC, Bowden J, Solheim TS, et al. Prognosis in advanced lung cancer--A prospective study examining key clinicopathological factors. Lung cancer (Amsterdam, Netherlands). 2015;88(3):304-9.
272. Li BT, Drilon A, Johnson ML, Hsu M, Sima CS, McGinn C, et al. A prospective study of total plasma cell-free DNA as a predictive biomarker for response to systemic therapy in patients with advanced non-small-cell lung cancers. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2016;27(1):154-9.
273. Devonshire AS, Whale AS, Gutteridge A, Jones G, Cowen S, Foy CA, et al. Towards standardisation of cell-free DNA measurement in plasma: controls for extraction efficiency, fragment size bias and quantification. Anal Bioanal Chem. 2014;406(26):6499-512.
274. Wong SQ, Li J, Tan AY, Vedururu R, Pang JM, Do H, et al. Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. BMC medical genomics. 2014;7:23.
275. Horizon. Cell free DNA [Available from: https://www.horizondiscovery.com/reference-standards/our-formats/cell-free-dna.

276. Li J, Harris L, Mamon H, Kulke MH, Wei-Hua L, Zhu P, et al. Whole Genome Amplification of Plasma-Circulating DNA Enables Expanded Screening for Allelic Imbalance in Plasma. The Journal of molecular diagnostics : JMD. 2006;8(1):22-30.
277. Croft DT, Jordan RM, Patney HL, Shriver CD, Vernalis MN, Orchard TJ, et al. Performance of Whole-Genome Amplified DNA Isolated from Serum and Plasma on High-Density Single Nucleotide Polymorphism Arrays. The Journal of molecular diagnostics : JMD. 2008;10(3):249-57.
278. Van der Vaart M, Pretorius PJ. Is the role of circulating DNA as a biomarker of cancer being prematurely overrated? Clinical biochemistry. 2010;43(1-2):26-36.
279. Aberle DR, Abtin F, Brown K. Computed Tomography Screening for Lung Cancer: Has it Finally Arrived? Implications of the National Lung Screening Trial. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013;31(8):1002-8.
280. Jing RR, Wang HM, Cui M, Fang MK, Qiu XJ, Wu XH, et al. A sensitive method to quantify human cell-free circulating DNA in blood: relevance to myocardial infarction screening. Clinical biochemistry. 2011;44(13):1074-9.
281. Carozzi F, Bisanzi S, Falini P, Sani C, Vebturini G, Lopes Pegna A, et al. Molecular profile in body fluids in subjects enrolled in a randomised trial for lung cancer screening: Perspectives of integrated strategies for early diagnosis. Lung cancer (Amsterdam, Netherlands). 2010;68(2):216-21.
282. Siera R, Bremnes RM, Cabrera A, Jantus-Lewintre E, Sanmartin E, Blasco A, et al. Circulating DNA is a Useful Prognostic Factor in Patients with Advanced Non-small Cell Lung Cancer. J Thorac Oncol. 2011;6(2):286-90.
283. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546-58.
284. Genome Browser Gateway [Available from: http://genome.ucsc.edu/cgi-bin/hgGateway.
285. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475(7356):348-52.
286. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998;8(3):175-85.
287. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, et al. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. 2008;18(5):763-70.
288. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004;91(2):355-8.
289. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017;168(4):613-28.
290. Chan KC, Jiang P, Zheng YW, Liao GJ, Sun H, Wong J, et al. Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. Clinical chemistry. 2013;59(1):211-24.
291. Jamal-Hanjani M, Quezada SA, Larkin J, Swanton C. Translational implications of tumor heterogeneity. Clinical cancer research : an official journal of the American Association for Cancer Research. 2015;21(6):1258-66.
292. de Bruin EC, McGranahan N, Swanton C. Analysis of intratumor heterogeneity unravels lung cancer evolution. Molecular & cellular oncology. 2015;2(3):e985549.
293. Jamal-Hanjani M, Hackshaw A, Ngai Y, Shaw J, Dive C, Quezada S, et al. Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. PLoS biology. 2014;12(7):e1001906.

294. Kirkizlar E, Zimmermann B, Constantin T, Swenerton R, Hoang B, Wayham N, et al. Detection of Clonal and Subclonal Copy-Number Variants in Cell-Free DNA from Patients with Breast Cancer Using a Massively Multiplexed PCR Methodology. Translational oncology. 2015;8(5):407-16.
295. Mouliere F, Robert B, Arnau Peyrotte E, Del Rio M, Ychou M, Molina F, et al. High fragmentation characterizes tumour-derived circulating DNA. PLoS One. 2011;6(9):e23418.
296. Izumchenko E, Chang X, Brait M, Fertig E, Kagohara LT, Bedi A, et al. Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA. Nature communications. 2015;6:8258.
297. Fernandez-Cuesta L, Perdomo S, Avogbe PH, Leblay N, Delhomme TM, Gaborieau V, et al. Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer. EBioMedicine. 2016;10:117-23.
298. Nakachi I, Rice JL, Coldren CD, Edwards MG, Stearman RS, Glidewell SC, et al. Application of SNP microarrays to the genome-wide analysis of chromosomal instability in premalignant airway lesions. Cancer prevention research (Philadelphia, Pa). 2014;7(2):255-65.
299. Kato S, Lippman SM, Flaherty KT, Kurzrock R. The Conundrum of Genetic "Drivers" in Benign Conditions. Journal of the National Cancer Institute. 2016;108(8).
300. Gormally E, Vineis P, Matullo G, Veglia F, Caboux E, Le Roux E, et al. TP53 and KRAS2 mutations in plasma DNA of healthy subjects and subsequent cancer occurrence: a prospective study. Cancer research. 2006;66(13):6871-6.
301. Pantel K. Blood-Based Analysis of Circulating Cell-Free DNA and Tumor Cells for Early Cancer Detection. PLoS medicine. 2016;13(12):e1002205.
302. National Lung Cancer Audit Report 2013 [Available from: http://www.hscic.gov.uk/catalogue/PUB12719/clin-audi-supp-prog-lung-nlca-2013-rep.pdf.
303. Cancer Research UK. Lung Cancer Incidence by Age [Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/incidence#collapseOne.
304. Spratt DE, Chan T, Waldron L, Speers C, Feng FY, Ogunwobi OO, et al. Racial/Ethnic Disparities in Genomic Sequencing. JAMA oncology. 2016;2(8):1070-4.
305. Campbell JD, Lathan C, Sholl L, Ducar M, Vega M, Sunkavalli A, et al. Comparison of Prevalence and Types of Mutations in Lung Cancers Among Black and White Populations. JAMA oncology. 2017.
306. Machin D, MJ C, SB T, SH T. Sample Size Tables for Clinical Studies, Third Edition: Wiley-Blackwell; 2009.
307. Zhang YA, Ma X, Sathe A, Fujimoto J, Wistuba I, Lam S, et al. Validation of SCT Methylation as a Hallmark Biomarker for Lung Cancers. J Thorac Oncol. 2016;11(3):346-60.
308. Lokk K, Vooder T, Kolde R, Valk K, Vosa U, Roosipuu R, et al. Methylation markers of early-stage non-small cell lung cancer. PLoS One. 2012;7(6):e39813.
309. Selamat SA, Galler JS, Joshi AD, Fyfe MN, Campan M, Siegmund KD, et al. DNA methylation changes in atypical adenomatous hyperplasia, adenocarcinoma in situ, and lung adenocarcinoma. PLoS One. 2011;6(6):e21443.
310. Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, et al. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013;31(32):4140-7.
311. [Available from: http://www.cancer-id.eu.
312. Campbell JD, Mazzilli SA, Reid ME, Dhillon SS, Platero S, Beane J, et al. The Case for a Pre-Cancer Genome Atlas (PCGA). Cancer prevention research (Philadelphia, Pa). 2016;9(2):119-24.

313. McShane LM, Hayes DF. Publication of tumor marker research results: the necessity for complete and transparent reporting. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2012;30(34):4223-32.
314. Sun K, Jiang P, Chan KC, Wong J, Cheng YK, Liang RH, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci U S A. 2015;112(40):E5503-12.
315. Best MG, Sol N, Kooi I, Tannous J, Westerman BA, Rustenburg F, et al. RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. Cancer cell. 2015;28(5):666-76.
316. Illumina. Illumina forms new company to enable early cancer detection via blood based screening 2016 [Available from: https://www.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=2127903.

7 Appendix

Appendix A: Potential clinical applications of cfDNA in lung cancer

Potential clinical applications of cfDNA in lung cancer	Application in lung cancer	Examples of studies in lung cancer	
Predict treatment response to molecular targeted therapies in advanced cases	EGFR-TKI and EGFR mutations	Bai 2009 He 2009 Yung 2009 Kimura 2007 Mack 2009 Mao 2010 Taniguchi 2011 Brevet 2011 Yam 2012	Nakamura 2012 Wang 2014 Weber 2014 Yanagita 2016 Oxnard 2016 Reck 2016 Reckcamp 2016 Kasahara 2017 Remon 2017 Zhang 2017 Muller 2017
	Crizotinib and ALK re-arrangements	Wang 2016 Cui 2017	
Monitor treatment response by tracking mutation profiles in advanced cases	EGFR TKI and EGFR mutations	Oxnard 2014 Wang 2014 Thress 2016 Reckcamp 2016 Piotrowska 2016 Pecuchet 2016 Zhu 2017	
	First line chemotherapy	Wang 2010	
	KRAS mutations	Yanagita 2016	
	BRAF mutations	Janku 2016	

Identify treatment resistance mechanisms and new therapeutic targets in advanced cases	EGFR TKI and EGFR mutant lung cancer	Punnoose 2012 Kuang 2009 Murtaza 2013 Oxnard 2014 Douillard 2014 Chabon 2016	Yanagita 2016 Thompson 2016 Thress 2016 Sundaresan 2016 Kashahara 2017 Chabon 2017
Monitor for minimal residual disease and disease relapse	Allelic imbalances, cfDNA levels Targeted ctDNA profiling	Sozzi 2001 Abbosh 2017	
Early detection	CfDNA levels	Sozzi 2001 2003 Zhang 2010 Szpechcinski 2015	

Appendix B : STH approval letter for Optimisation of plasma DNA studies

Ref: STH14669/LB

Sheffield Teaching Hospitals NHS
NHS Foundation Trust

07 Aug 09

Professor P Woll
Cancer Research Centre
Weston Park Hospital
Whitham Road
Sheffield
S10 2SJ

Dear Professor Woll

Authorisation of project

STH ref: **STH14669**

Study title: **Optimisation of plasma DNA studies**

Chief Investigator: **Professor P Woll, University of Sheffield**

Principal Investigator: **Dr M Teare, University of Sheffield**

Sponsor: **STH NHS FT**
Funder: **Own Account**

The Research Department has received the required documentation for the study as listed below:

1.	**Sponsorship IMP studies (non-commercial)** **Sponsorship responsibilities between institutions** **Responsibilities of investigators** **Monitoring Arrangements**	Not Applicable Not Applicable Not Applicable Not Applicable
2.	**STH registration document: completed and signed**	NHS REC Application Form – V 5.6 - P Woll – 08 May 08 NHS REC Application Form – V 5.6 - D Patel – 09 May 08 STH Finance Form – P Woll – 30 May 08
3.	**Evidence of favourable scientific review**	STH R&D ISR - 31 Jan 08
4.	**Protocol – final version**	Version 1.2 - 13 Feb 08
5.	**Participant Information sheet – final version**	PIS version 1.1 - 13 Dec 07
6.	**Consent form – final version**	Version 1.3 - 16 Jun 08

Appendix C : STH approval letter for the ReSoLuCENT study

Sheffield Teaching Hospitals NHS
NHS Foundation Trust

Date 29 Apr 14

To whom it may concern

Confirmation of Sponsorship of Clinical Research Project

STH ref:	STH13872
NIHR CSP ref:	N/A
REC ref:	West Midlands, 05/MRE07/72
Study title:	ReSoLuCENT: Genetic epidemiology of lung cancer
Chief Investigator:	Professor Penella Woll
Principal Investigator:	Professor Penella Woll
Sponsor:	Sheffield Teaching Hospitals NHS Foundation Trust

Sheffield Teaching Hospitals NHS Foundation Trust has agreed to act as Sponsor for the above mentioned Research Project (non-CTIMP).

This project received R&D Authorisation at Sheffield Teaching Hospitals NHS Foundation Trust on 27 Mar 2006.

Yours sincerely

Professor S Heller
Director of R&D, Sheffield Teaching Hospitals NHS Foundation Trust
Telephone +44 (0) 114 2265934
Fax +44 (0) 114 2265937

Appendix D: Bioinformatics scripts for data processing

- **Shell script for creating read profiles**

```
# Map to human genome GRCh38 with bwa
./bwa mem -M -t 6 [Human_Genome] R1.fq R2.fq > output.sam
# convert to bam
samtools-0.1.19/samtools view -Sb output.sam > output.bam
# sort bam
samtools-0.1.19/samtools sort output.bam output_sorted.bam
# mark duplicates with Picard
java -Xmx2g -jar MarkDuplicates.jar I=output_sorted.bam
O=output_sorted_dups.bam M=output_sorted_dups.metrics
# remove duplicates
samtools rmdup output_sorted_dups.bam output_sorted.nodups.bam
# only retain uniquely mappable reads with qual>37
samtools view -b -q 37 output_sorted.nodups.bam > output_sorted.final.bam

# Bin reads (-w window size, or -r mean number of reads in window)
perl bam2windows.pl --samtools-path=[path to samtools] -gc
gc1000Base_38.txt [-r 1000] [-w 1000000]test sorted.final.bam
ref sorted.final.bam > output.tab
```

- **CNAnorm R script for closest normalisation**

```
#obtain read copy number count files for analysis
get_results = function(w,x,y,z)
{
a=read.delim(file="output.tab", stringsAsFactors=FALSE,
check.names=FALSE) set.seed(31)
CN <- dataFrame2object(a)
#eliminate Y and X chromosome and mitochondrial DNA
toSkip <- c("X","Y", "MT")
#normalisation (GC content, closest normalisation for ploidy and heterogeneity) and eliminate
outliers
CN <- gcNorm(CN, exclude=toSkip)
CN <- addSmooth(CN, lambda=7)
CN <- peakPloidy (CN, exclude=toSkip, method='closest')
pdf(w, height=4.27, width=11.69)
plotPeaks(CN)
dev.off()
CN <- validation (CN,ploidy = (sugg.ploidy(CN) - 1))
#segmentation of normalised copy number ratios
CN <- addDNACopy(CN)
CN <- discreteNorm(CN)
pdf(x, height=4.27, width=11.69)
#establish copy number profile graphs
data(gPar)
gPar$genome$colors$gain.dot <- 'darkorange'
gPar$genome$colors$grid <- NULL
gPar$genome$cex$gain.dot <- .2
gPar$genome$cex$loss.dot <- .2
plotGenome(CN, superimpose='DNACopy', show.centromeres=FALSE,gPar=gPar,
colorful=TRUE)
dev.off()
pdf(y, height=4.27, width=11.69)
plotGenome(CN, superimpose='smooth', show.centromeres=FALSE)
dev.off()
exportTable(CN, file=z, show='center')
```

Appendix E: Methods for determining somatic mutations by targeted highly parallel genome sequencing with the Ion Torrent Platform

This work was carried out in collaboration with Professor Jaqui Shaw at the University of Leicester. We supplied and shipped plasma and genomic DNA collected in the ReSoLuCENT study to Leicester for Ion Torrent analysis. I spent time in the laboratory in Leicester to learn plasma DNA extraction, library preparation and sequencing methods. I assisted in result analysis and checked variants on the Integrated Genomic Viewer.

The Ion Torrent platform is validated for the detection of genome variants from 10 ng of tumour FFPE DNA and was therefore an attractive approach for targeted genome sequencing of cfDNA. There were four main steps required to process samples and identify variants. These were, library construction, template enrichment, sequencing and data analysis (Figure Appendix E1). The manufacturer's instructions for the maintenance and preparation of all instruments were carried out. Standard recommended protocols were followed to prepare and sequence samples.

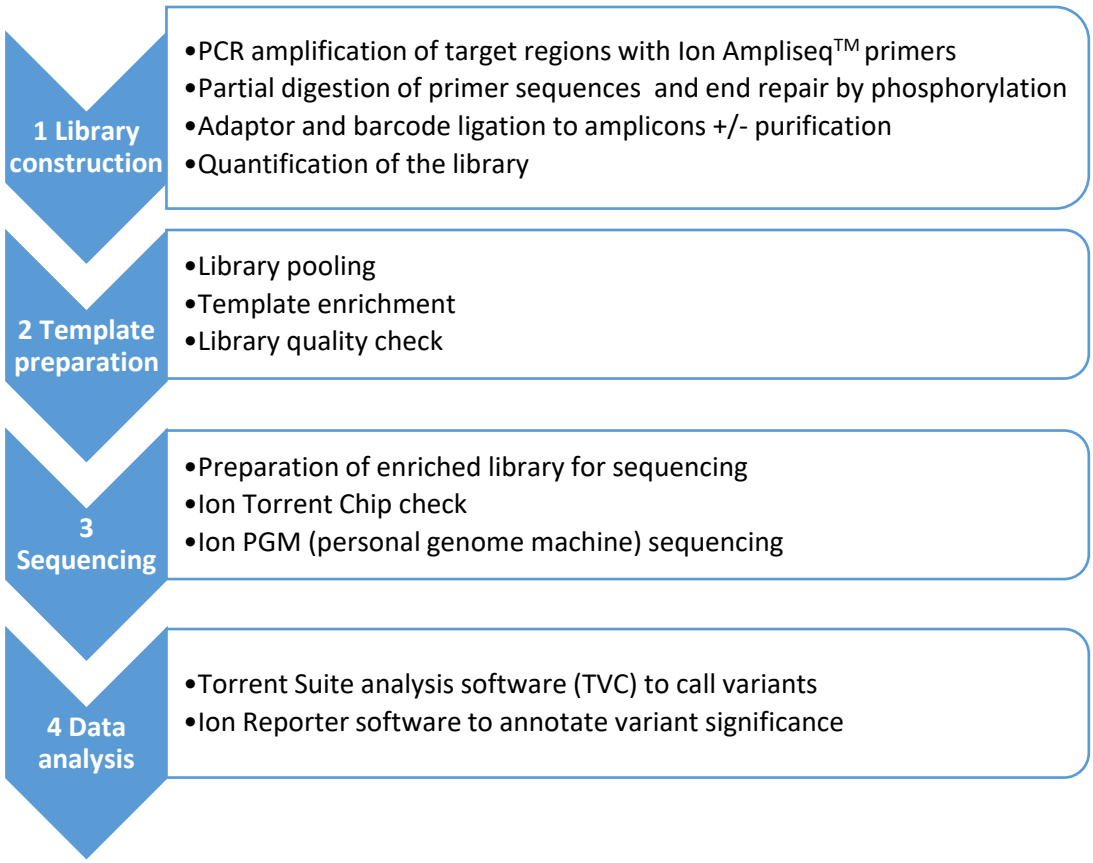


Figure Appendix E1: The four main steps for targeted sequencing with the Ion Torrent Platform

1 Library construction

The Ion Ampliseq™ Library Kit 2.0, Ion Xpress™ barcode adaptors and Ion Ampliseq™ Colon/Lung cancer panel v2 were used to create amplicon libraries (all Thermo Fisher Scientific). Figure Appendix E2 summarises the library construction process.

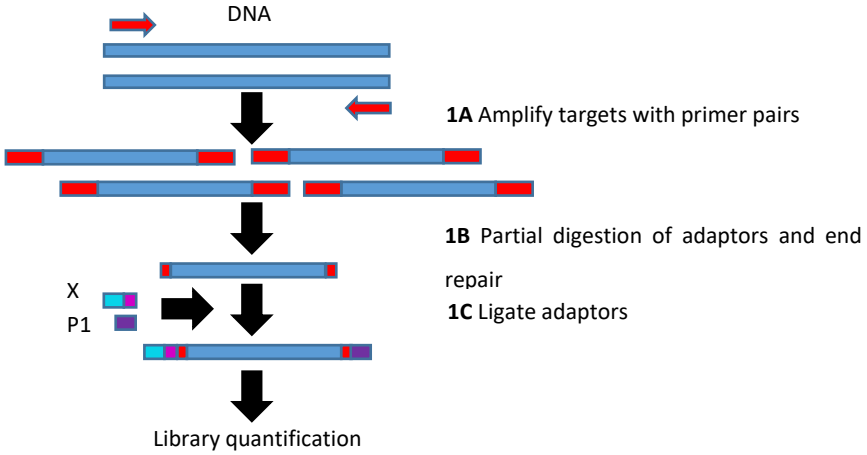


Figure Appendix E2: Ion AmpliSeq™ Library preparation and quantification. P1 is a universal sequencing adaptor. X is an adaptor with a unique oligonucleotide barcode for each sample.

1A) PCR amplification of target regions

Target regions were amplified by primer pairs in the Colon/Lung cancer panel v2. In a 200 µl tube or 96 well plate, 10 µl of 2X Ion AmpliSeq™ primer pool were added to 4 µl of 5X Ion AmpliSeq™ HiFi Master Mix. Up to 10 ng of cfDNA or genomic DNA were added in a maximum volume of 6 µl. Targeted regions were amplified by PCR with thermocycling conditions displayed in Table Appendix E1.

Step	Temp	Time	
Enzyme activation	99 °C	2 mins	
Denature DNA	99 °C	15 secs	*22-28 cycles for cfDNA dependent on the number of primer pairs, 17-22 cycles for genomic DNA.
Anneal/ Extension	60 °C	4 mins	
Hold	10 °C	Hold	

Table Appendix E1: Thermocycling conditions for the PCR amplification of target regions.

1B) Partial digestion of primer sequences and end repair by phosphorylation

After PCR amplification, primer sequences were partially digested and amplicon ends were repaired by phosphorylation to facilitate barcode adaptor ligation, by addition of 2 µl of FuPa Reagent to each sample. The mixture was incubated in a thermocycler with the following conditions, 50 °C for 10 minutes, 55 °C for 10 minutes, 60 °C for 20 minutes and final 10 °C hold for up to 1 hour.

1C) Barcode adapter ligation

A diluted unique barcode adaptor mix was made for each sample by adding 2 µl of Ion P1 adapter to 2 µl of a unique Ion Xpress Barcode X and 4 µl of sterile nuclease free water. 2 µl of this was added to 4 µl of Switch Solution and the partially digested repaired amplicons, followed by 2 µl of DNA ligase. The mixture was incubated at 22 °C for 30 minutes, 72 °C for 10 minutes and finally held at 10 °C for up to 1 hour.

1D) Quantification of barcode adaptor libraries and further library enrichment

Amplicon libraries were purified, enriched, quantified and finally normalised to 100 pM and stored at -20°C. These processed are described in detail in the following sections.

i. First purification

The barcode adaptor libraries were purified using AMPure®XP bead (Beckman Coulter, Inc) to remove enzymes, FuPa and PCR inhibitors. 45 µl of AMPure®XP beads were added to 30 µl of the barcode adaptor ligated amplicons to give a bead to sample ratio of 1.5X. The following amendments were made to the purification method, 150 µl of 70 % ethanol was used to wash the pellet, and DNA was eluted in 50 µl of Platinum® PCR SuperMix High Fidelity (Thermo Fisher Scientific). Two microlitres of Equilizer™ primers (Thermo Fisher Scientific) were added to the eluate and after mixing 50 µl were transferred into a 200 µl tube for an additional amplification step.

ii. PCR amplification, size selection and second purification

Equilizer™ primers bind only to adaptor ligated DNA amplicons. PCR amplification aims to achieve a stronger signal from adaptor-ligated amplicons and to reduce the effect of amplicon loss during the purification steps. Thermocycling conditions were 98 °C for 2 minutes followed by 7 cycles of 98 °C for 15 seconds and 60 °C for 1 minute, and a final hold at 10 °C.

After PCR amplification, AMPure™ beads were used in two steps to size select the amplicon products and eliminate impurities. First, 25 µl of beads were added to 50 µl of sample for a bead: product ratio of 0.5X. This ratio resulted in the separation of large DNA fragments that

bound to the beads from the small DNA amplicons that remained in the supernatant. The supernatant was transferred to a clean 200 µl tube and a second clean-up was performed. Sixty microlitres of beads were added to 50 µl of sample to give a bead to product ratio of 1.2X. This resulted in the barcode adaptor DNA amplicons binding to the beads, whilst the primers and other impurities remained in the supernatant, which was discarded. The remaining pellet was washed twice in 150 µl of 70% ethanol, dried for 5 minutes and the DNA was eluted in 50 µl of Low T.E (Thermo Fisher Scientific). Libraries were quantified with the Qubit® fluorimeter and library quality was assessed with the Agilent Tapestation 2200 and the high sensitivity kit.

2 Template Preparation

To prepare constructed libraries for sequencing, libraries were pooled and templates were formed and quality checked (Figure Appendix E3). These processes are explained in greater detail in the following sections.

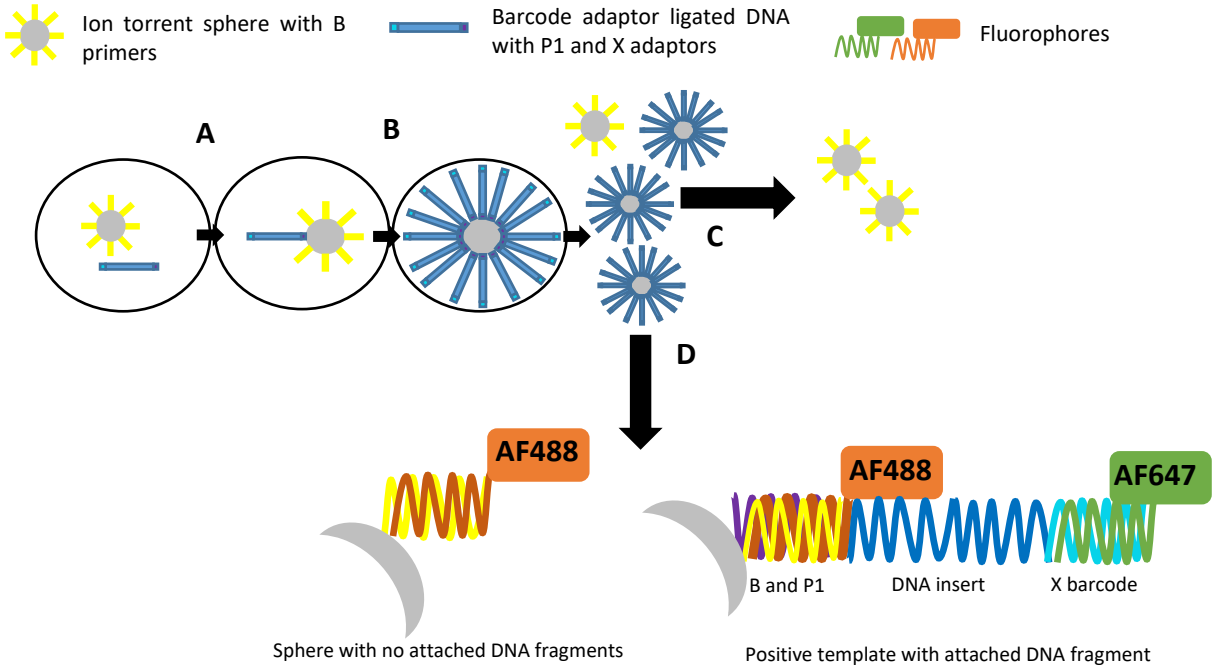


Figure Appendix E3: Template preparation for Ion Torrent sequencing. A) One barcode adaptor ligated DNA fragment binds to a complementary oligonucleotide on the Ion Sphere particle within a water droplet in oil. B) Emulsion PCR forms a positive template covered in millions of clonally amplified DNA fragments. C) Positive templates are selected. D) The proportion of positive templates are calculated by establishing the AF647:AF488 ratio.

***2A+2B)* Library pooling and Template Enrichment by emulsion PCR**

Ten microlitres of each library normalised to 100pM were pooled together prior to enrichment. In the enrichment step, barcode adaptor ligated DNA fragments are amplified by emulsion PCR. Ion Torrent Spheres provide the solid platform for this process. Adaptor ligated DNA fragments bind to the complementary primers attached to the spheres. To create monoclonal templates, it is essential that only one DNA fragment binds to each sphere, this is achieved by having an excess of spheres compared to fragments. Fragment: bead complexes are mixed with emulsion oil to create droplets in a reaction filter. The solution containing the droplets passes through the multiple channels of a Ion OneTouch™ 2 amplification plate that is set at two different temperatures for thermocycling. The solution moves due to peristalsis created by a fluidic pump. After amplification is completed,

each sphere is covered in approximately one million clonally amplified DNA fragments to create a positive template. The emulsion flows into a centrifuging tube along with a 'recovery solution' that contains a detergent. Upon centrifugation, the emulsion is broken down, the spheres are washed and recovered into a collecting tube forming a pellet.

Library templates were prepared with the Ion PGM™ Template OT2 200 kit (Thermo Fisher Scientific). In brief, an amplification solution was made by mixing 500 µl of Ion PGM Template OT2 200 Reagent Mix, 300 µl of PCR Reagent B and 50 µl of Enzyme Mix. Two microlitres of the pooled library were added to the amplification solution, which was then vortexed and briefly centrifuged. Then, 100 µl of re-suspended Ion Sphere Particles (ISPs) were added. After vortexing, the final solution of 1000 µl was immediately injected into the Ion PGM™ OneTouch Plus Reaction Filter (Thermo Fisher Scientific) followed by 1500 µl of oil. The filter was inverted carefully and inserted onto the automated Ion OneTouch™ 2 instrument (Thermo Fisher Scientific). The programme PGM: Ion PGM™ Template OT2 200 kit was run.

2C) Selection of Ion sphere positive templates

Complementary DNA fragments attached to the positive template spheres have biotin incorporated into the P1 adaptor. The biotinylated positive sphere templates bind to the MyOne™ Streptavidin C1 beads and are separated from non-template spheres by magnet transfer, this maximises sequencing yield. The Ion sphere positive templates are then washed. The complementary strand attached to a DNA fragment on the sphere are separated by the addition of NaOH but stay bound to the beads. The beads and therefore complementary fragments are removed using a magnet. Remaining, are single strands attached to the Ion positive spheres that are now ready to be used as a sequencing template. The collecting tube holding the enriched library product was removed from the Ion OneTouch™ 2 instrument. All but 50 µl of liquid were discarded and the pellet of spheres were re-suspended. Two microlitres of the enriched sphere product were retained to check library quality. The remaining solution was transferred to the automated Ion OneTouch™ enrichment system (Thermo Fisher Scientific) for purification and selection of Ion sphere enriched templates with Dynabead® MyOne™ Streptavidin C1 beads (Thermo Fisher Scientific). An 8-well strip was inserted into the Ion OneTouch™ 2 enrichment system. This contained in the following order, the enriched library, 130 µl of Dynabeads® MyOne™ Streptavidin C1 beads re-suspended in Beads Wash Solution, 3 consecutive wells filled with 300 µl of Ion OneTouch™ Wash Solution, an empty well, 300 µl of freshly prepared Melt-Off Solution and an empty well to collect the final selected product.

2D) Library quality check

The quality of the enriched library was determined by establishing the ratio of positively enriched Ion spheres to the total number of spheres. This was achieved by using two types of fluorophore probes. The Alexa Fluor® 488 is attached to an oligonucleotide chain that is complementary to a sequence present in the primers on all spheres and the DNA template. The Alexa Fluor® 647 is attached to an oligonucleotide chain complementary to a sequence only present in the DNA template. Therefore, positive templates are distinguished from spheres that have no DNA templates and the percentage of positive template can be established (Figure Appendix E3).

In brief, 2 µl of the enriched library were added to 19 µl of annealing buffer and 1 µl of Ion fluorophore probes, both from the Ion Sphere™ quality control kit (Thermo Fisher Scientific). The solution was incubated at 95 °C for 2 minutes and 37 °C for 2 minutes to enable the probes to anneal to their targets. AF647 and AF488 fluorescence was measured with the Ion assay on the Qubit® 2.0 fluorometer and the ratio of positively enriched Ion spheres to total Ion spheres was calculated using an Ion Torrent excel file. The optimal result was a library with 10% to 30% of enriched positive template spheres because these are most likely to be monoclonal due to a significant excess of spheres compared to DNA fragments.

3) Ion Torrent PGM sequencing

The Ion Torrent Sequencing 200 Kit v2 (Thermo Fisher Scientific) was used to obtain bi-directional sequencing reads up to 200 bp in length with the Ion Torrent Personal Genome Machine™ (PGM).

3A) Preparation of enriched spheres for sequencing

Enriched positive template spheres were combined with Ion control spheres that have DNA fragments attached whereby the sequence of bases are known. The Ion control spheres are positive controls for the sequencing process and they aid calibration of the Ion Torrent PGM™ to ensure accurate base calling.

Five microlitres of control Ion spheres were combined with the enriched positive template Ion spheres, pipette mixed and centrifuged at 15,000 g for 5 minutes. Next, 15 µl of the supernatant surrounding the resultant pellet of spheres were removed and 12 µl of Sequencing Primer were added prior to re-suspending the pellet. To anneal the sequencing primers, the mixture was incubated at 95 °C for 2 minutes followed by 37 °C for 2 minutes. Three microlitres of Ion PGM™ Sequencing 200v2 Polymerase were added and the mixture was incubated at room temperature for 5 minutes.

3B) Chip loading

The Ion 316 v2 semiconductor chip (Thermo Fisher Scientific) with an expected output of 1.9-2.5 million reads was chosen to facilitate sample multiplexing whilst maintaining high amplicon coverage at a reasonable cost. The prepared Ion sphere library (30 µl) was loaded onto the calibrated Ion 316 v2 chip for sequencing at a rate of 1 µl per second. Subsequent steps of centrifugation and mixing of the sample aimed to lodge one sphere into each well. It was important to avoid the introduction of air into the chip because spheres become dislodged. Any residual liquid was removed from the chip prior to sequencing.

3C) Targeted Sequencing

The Ion PGM™ was used for the sequencing of enriched libraries. During sequencing, each of the four nucleotide bases (A, T, C, G) flowed sequentially over the Ion Torrent semiconductor chip, which consisted of millions of wells. Each well contained one bead that was covered in approximately one million copies of a single DNA fragment. In a well, if the flowing nucleotide was complementary to the nucleotide of the DNA template attached to the sphere it was incorporated by DNA polymerase into a newly synthesised DNA strand, resulting in the release of a hydrogen ion. The resulting change in pH of the solution was measured directly by an Ion Sensor, converted to a voltage and the base was called. A different nucleotide then flowed over the chip and the process was repeated. If no base was incorporated there was no change in pH, and no base was called. If two consecutive bases were identical on the DNA template two nucleotides were incorporated, two hydrogen ions were released, the voltage was doubled and two of the same bases were called. Bi-directional sequencing was carried out because two templates were generated for each fragment during initial library construction therefore both ends of an amplicon were sequenced with a single read run (147).

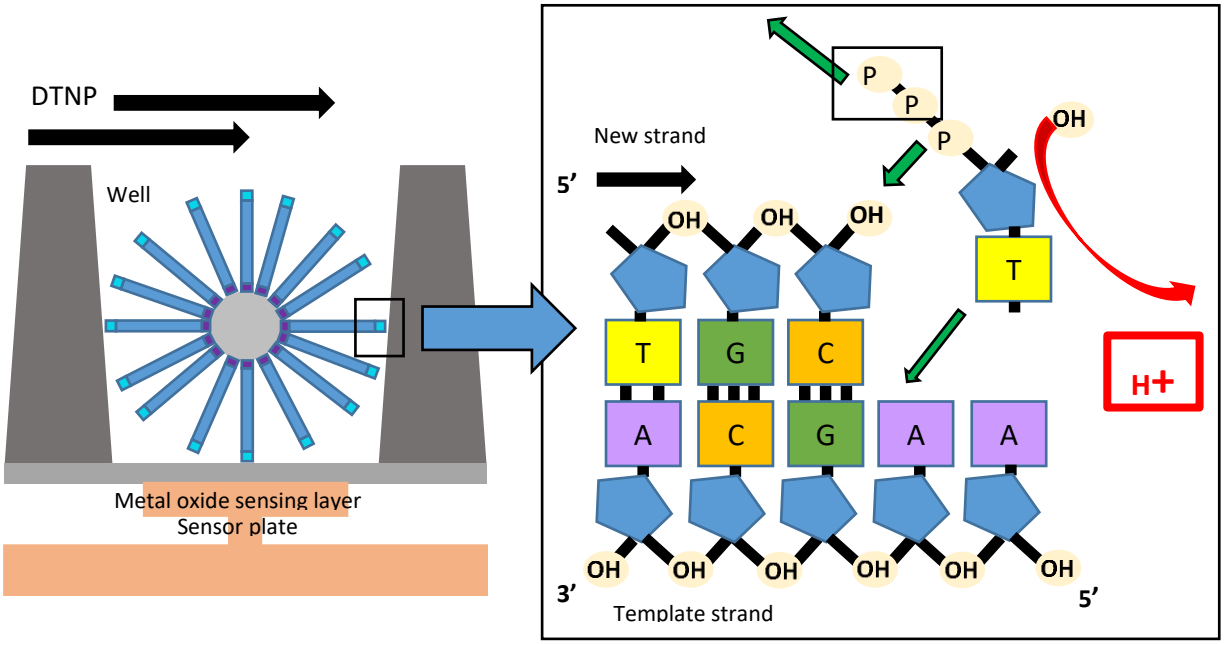


Figure Appendix A) Ion Semiconductor Sequencing (add labels to sensor and ISP). Addapted from Rothberg et al. 2011 (147) (permission not required by the publisher).

4) Ion Torrent Data Analysis

Automated data analysis optimised for the processing of Ion Torrent sequencing data was completed with the Ion Torrent suite software (version 4.0.2) on the Ion Torrent Server. Nucleotide bases were called, assigned Phred like quality scores and ordered to form one read for every DNA fragment. Reads were trimmed, filtered and aligned to the human reference genome 19 (hg19) to enable the identification of variants. The data analysis steps are outlined in Figure Appendix E4.

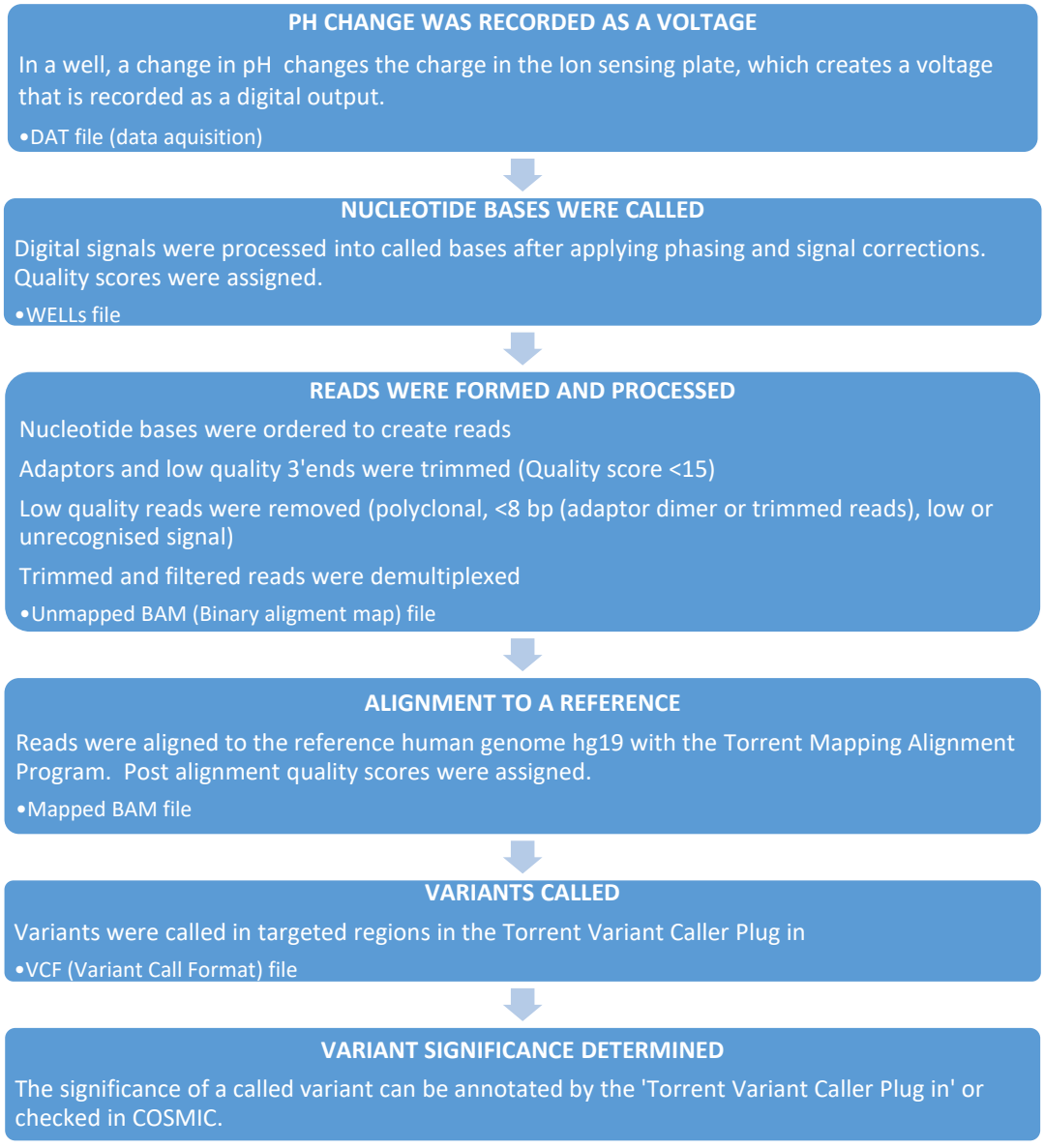


Figure Appendix E4: Data analyses steps to identify variants and their significance

Ion Torrent Phred like quality scores

A quality or Phred like score was assigned to each nucleotide to predict the likelihood of the called base being correct (325). This score enabled sequencing quality to be compared between experiments and established the accuracy of each individual base call (326).

Torrent Variant Caller

Aligned reads were evaluated for base discrepancies from the reference sequence using the Torrent Variant Caller Plugin (TVC version). Somatic low stringency parameters were applied to optimise low frequency variant detection but minimise false negatives. SNVs, MNPs (multi nucleotide polymorphisms) and Indels were called in the targeted regions defined by the specific Ion Torrent Panelv2 BED files.

Manual checking of called variants

The parameters of all called variants were manually checked. Variants with a quality Phred score less than 20 or original amplicon coverage of less than 50 were discarded. All reads of called variants were visualised in the Integrated Genomic Viewer (IGV version 2.3) to determine false positives due to PCR artefact, sequencing errors and strand bias. Variants were excluded if the variant allele was called within 10 nucleotide bases of the end of a read due to the higher error rate of DNA polymerase. Figures Appendix E5 and Appendix E6 display accepted and rejected variant calls.

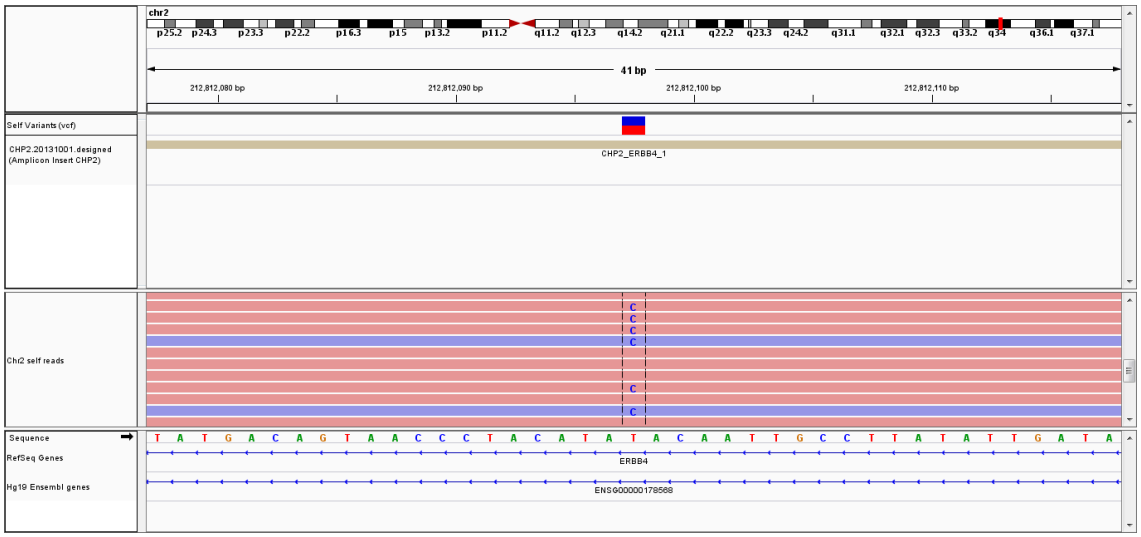


Figure Appendix E5: Example of a read profile of an accepted variant viewed in IGV. Reads in red are on the positive/forward strand and reads in blue are on the negative/reverse strand. The reference allele is T and the variant allele is C.

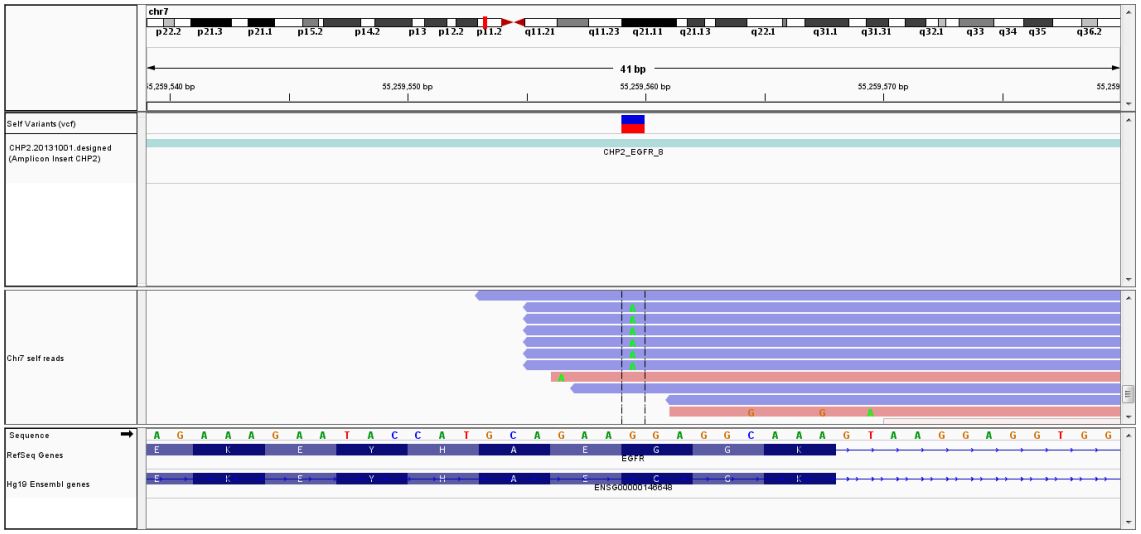


Figure Appendix E6: Example of a read profile of a rejected variant viewed in IGV. The reference allele is G and the variant allele is A. The variant allele is called within five bases of the end of the reverse/negative strand 3' end and is most likely an artefact introduced by mis-priming during PCR amplification.

Determination of the significance of a called variant

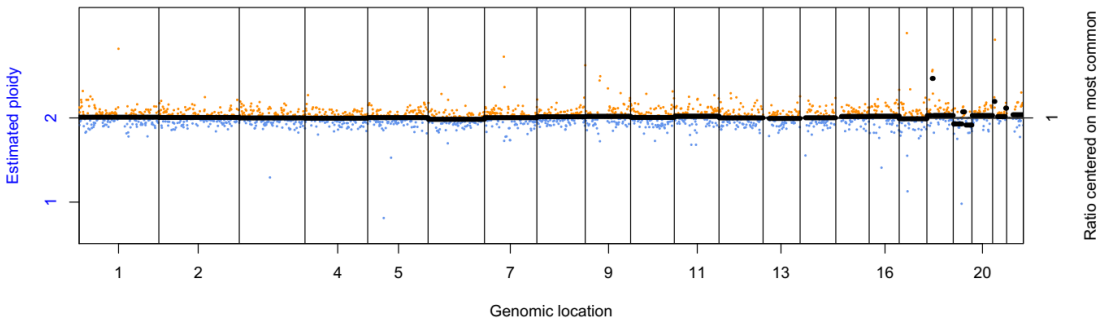
Targeted sequencing can identify discrepancies between the reference and DNA sequence within the length of an amplicon. It is important to determine whether a variant is somatic or germline. We sequenced lymphocyte genomic DNA and if a variant was identified to be present in both germline and cfDNA samples it was eliminated. All somatic variants were checked for their presence in the public database of somatic mutations COSMIC (327).

Appendix F: Copy number profiles for cfDNA samples

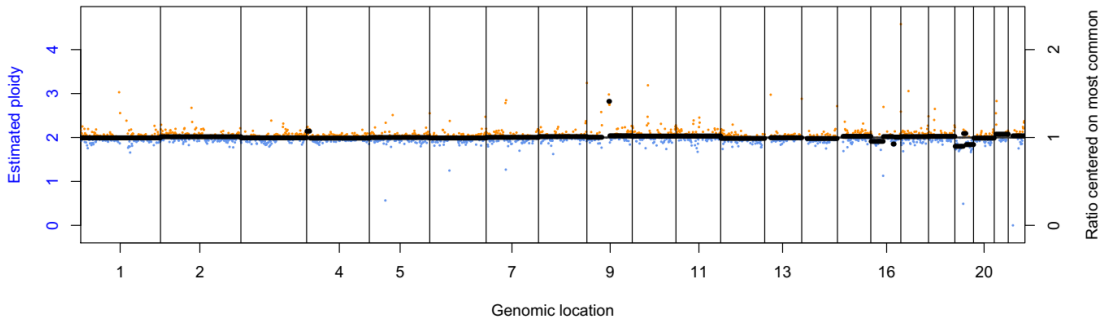
Untreated lung cancer cases (age, gender, centre, smoking status, stage, pathology)(N=51)

(WPH: Weston park hospital Sheffield, NGH: Northern General Hospital, Sheffield, DRI: Doncaster Royal Infirmary, AGH: Airedale, VHC: , CHM: Christies Hospital Manchester, SGH:

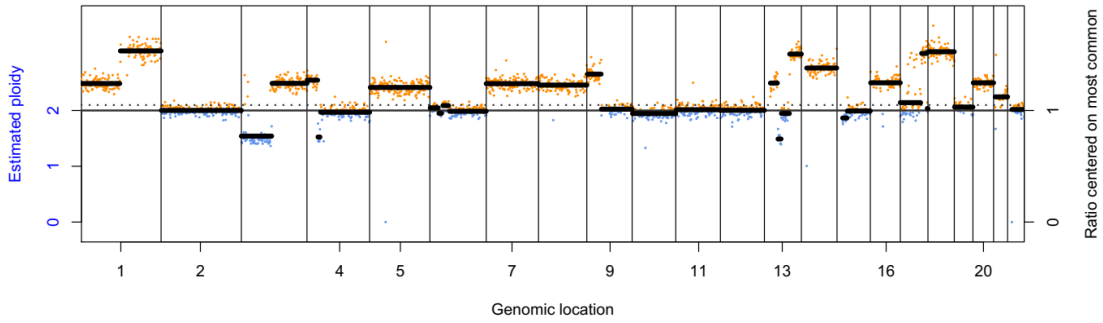
P130 (57 years, male, WPH, ex-smoker, stage IV adenocarcinoma) 290 CNA score



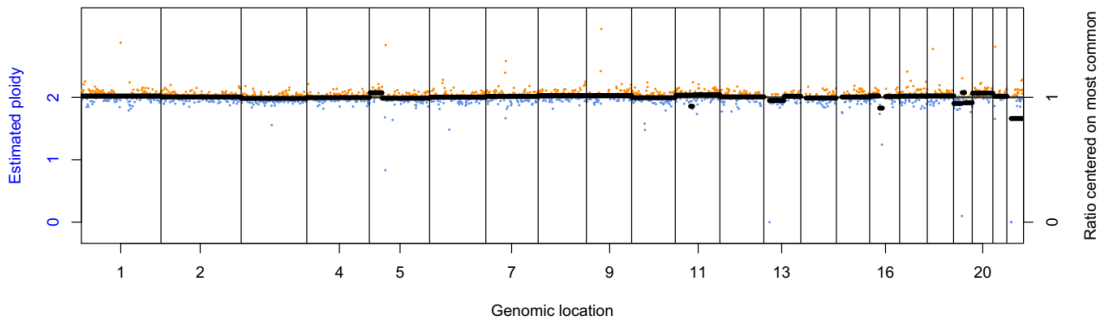
P146 (66 years, female, WPH, ex-smoker, stage IV adenocarcinoma) 494 CNA score



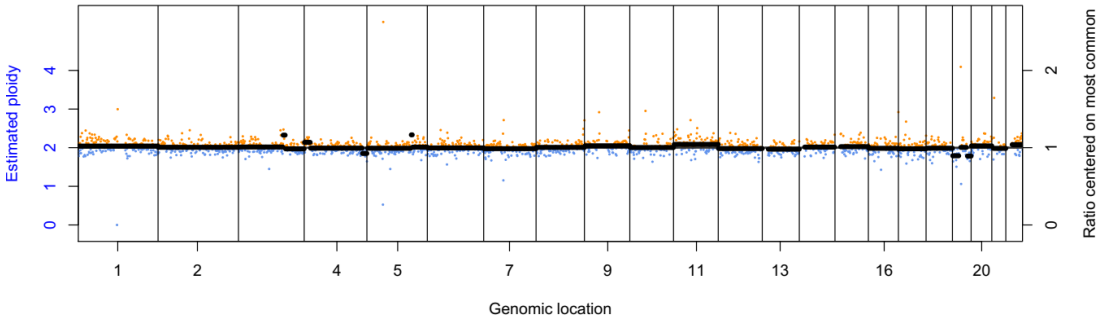
P157 (60 years, female, WPH, current, stage III SCLC) 40062 CNA score



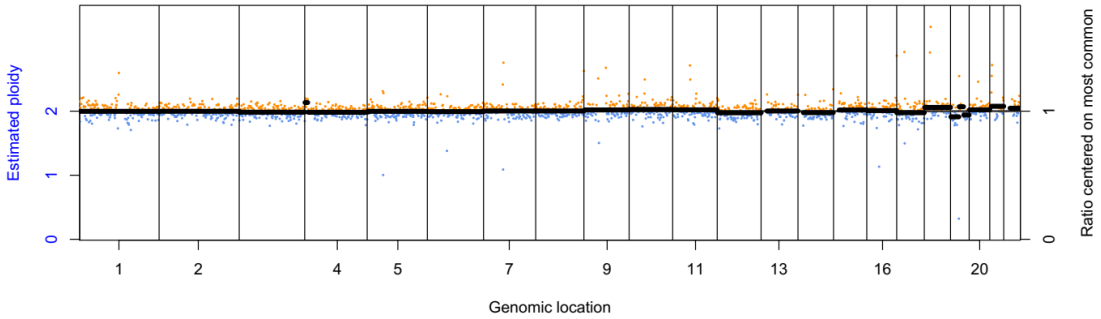
P165 (54 years, female, DRI, current, stage IV NSCLC NOS) 266 CNA score



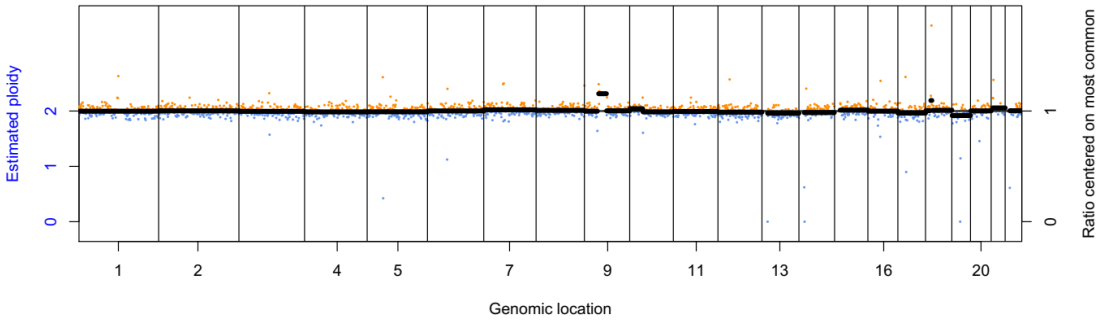
P240 (58 years, male, DRI, current, stage IV NSCLC NOS) 660 CNA score



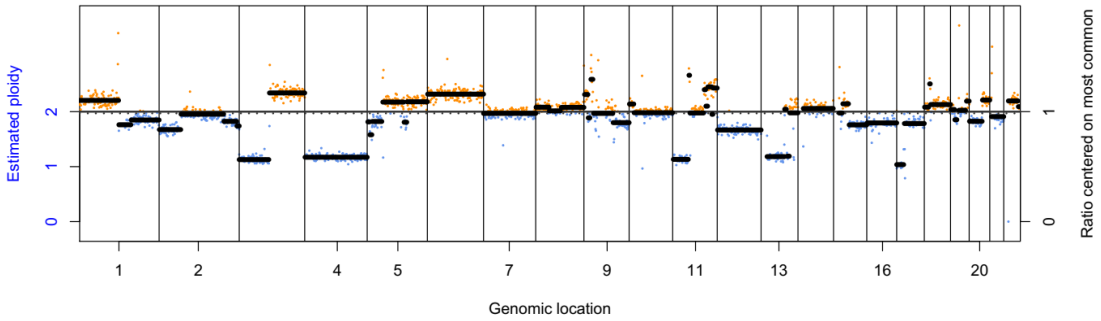
P244 (65 years, female, NGH, current, stage IIA squamous) 211 CNA score



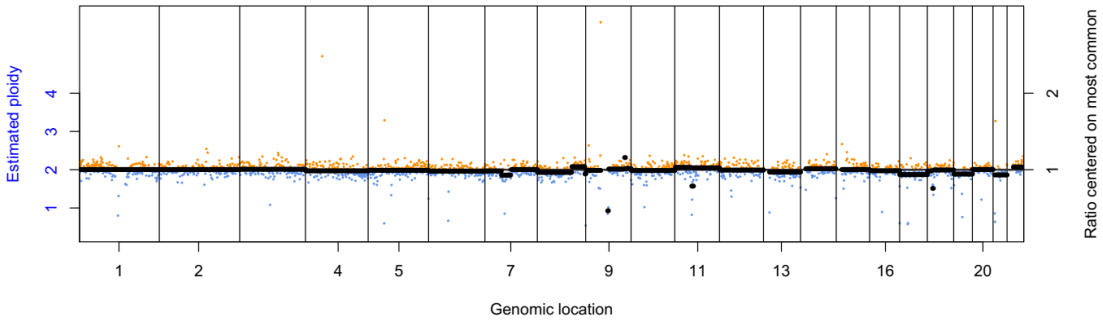
P246 (58 years, female, WPH, never smoked, stage IV adenocarcinoma) 169 CNA score



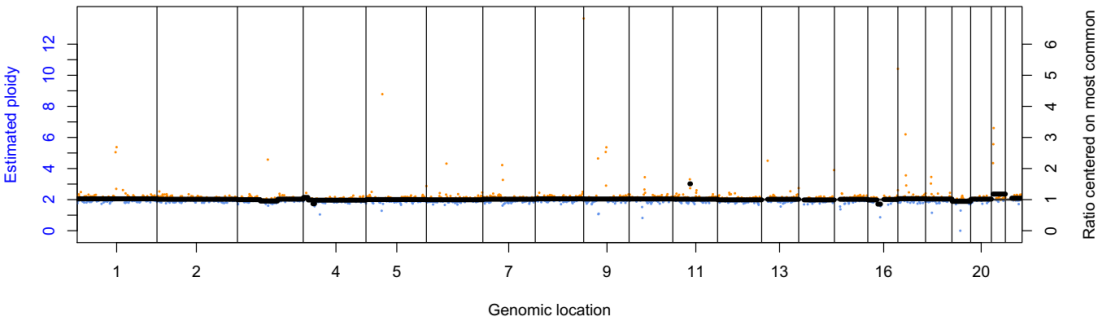
P249 (60 years, male, WPH, ex-smoker, stage IV SCLC) 28704 CNA score



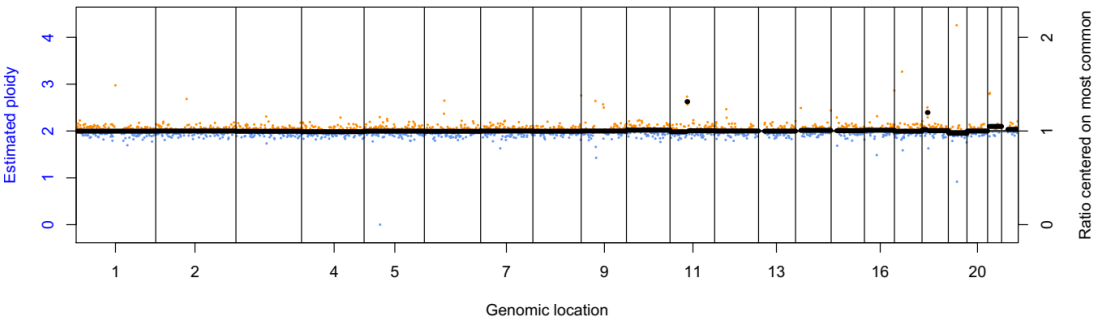
P254 (61 years, female, WPH, ex-smoker, stage IIIA NSCLC NOS) 804 CNA score



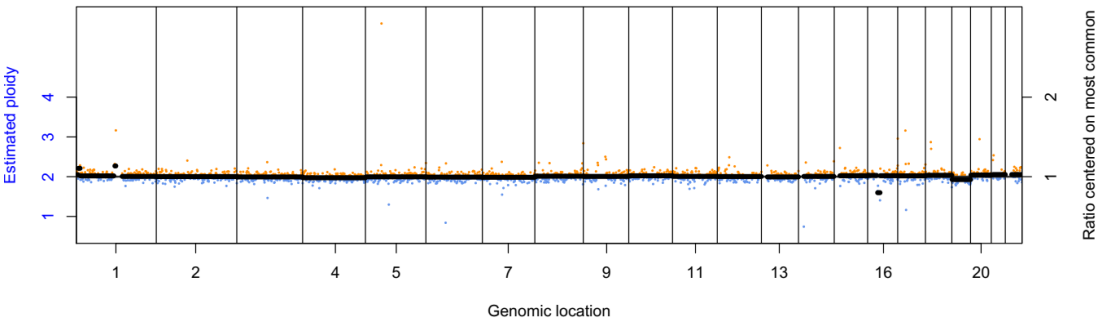
P255 (59 years, female, DRI, current, stage IV squamous) 692 CNA score



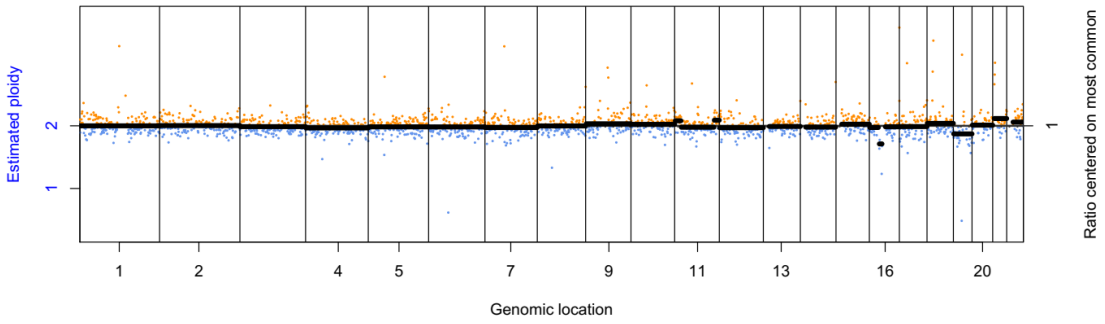
P261 (59 years, male, NGH, current, stage IA adenocarcinoma) 393 CNA score



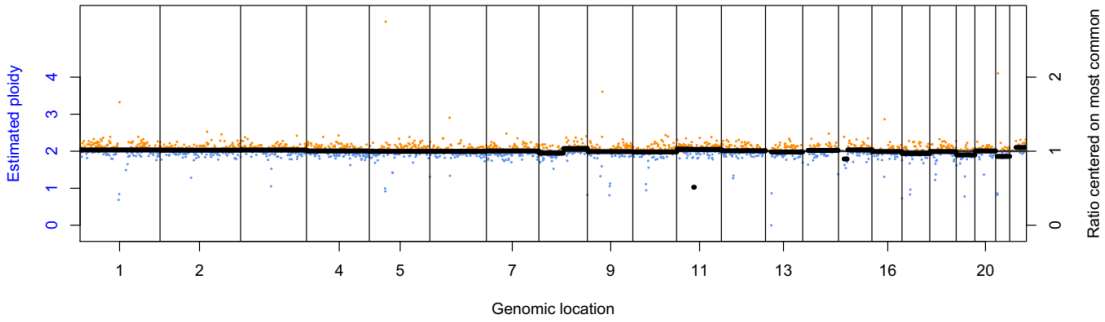
P276 (54 years, female, NGH, ex-smoker, stage IIA adenocarcinoma) 268 CNA score



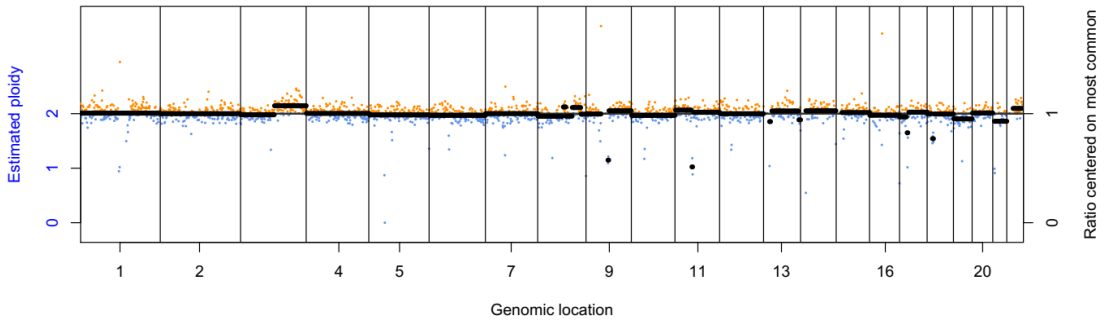
P281 (55 years, male, WPH, ex-smoker, stage IIIA adenocarcinoma) 478 CNA score



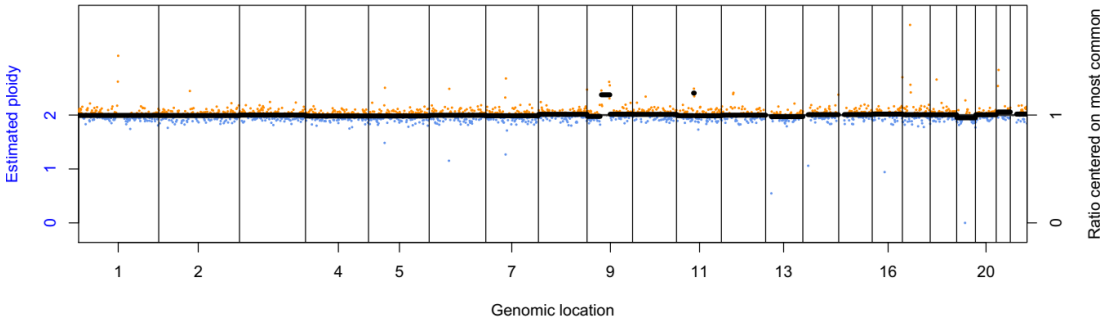
P288 (60 years, male, NGH, current, stage IA squamous) 904 CNA score



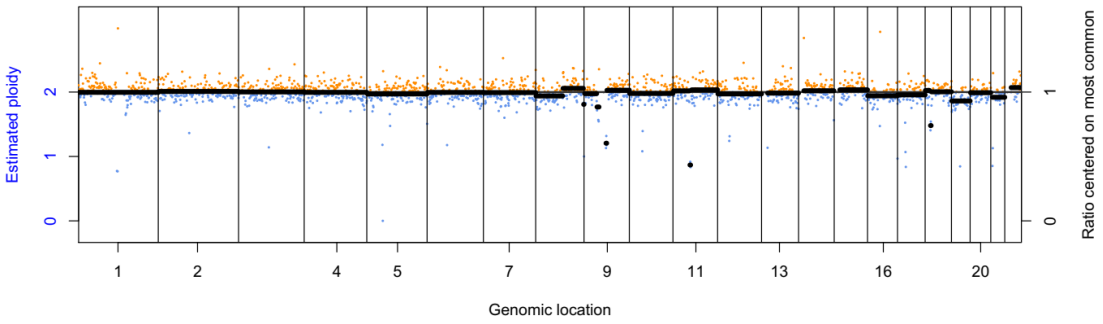
P332 (59 years, female, DRI, ex-smoker, stage IV squamous) 560 CNA score



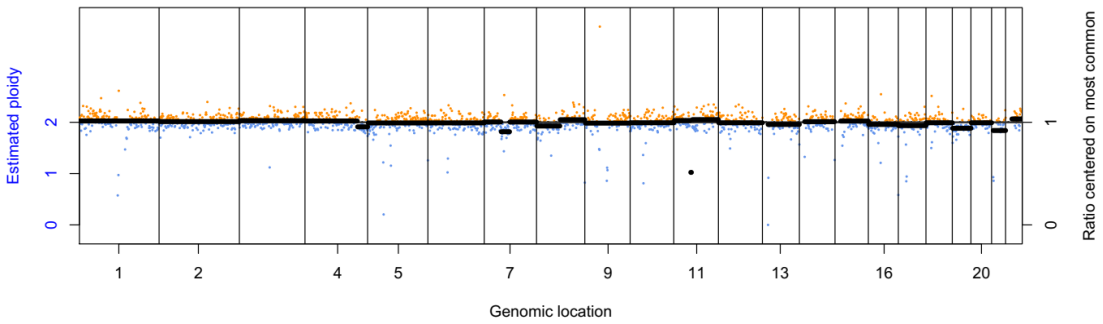
P434 (57 years, male, NGH, ex-smoker, stage IIB squamous) 192 CNA score



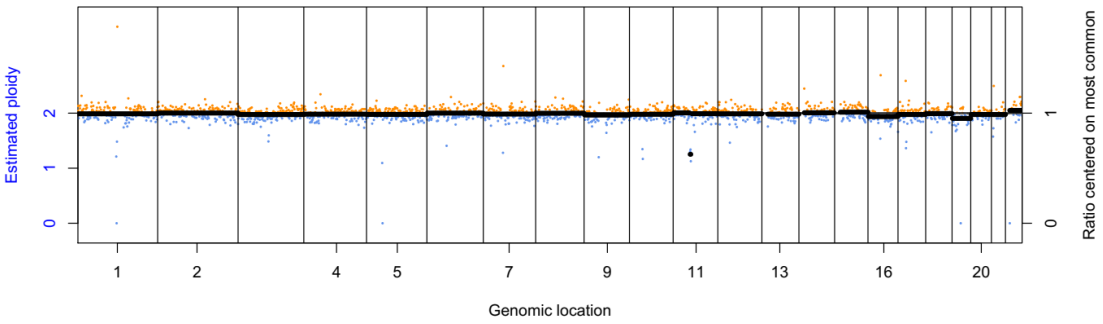
P458 (40 years, female, NGH, ex-smoker, stage IIIA adenocarcinoma) 535 CNA score



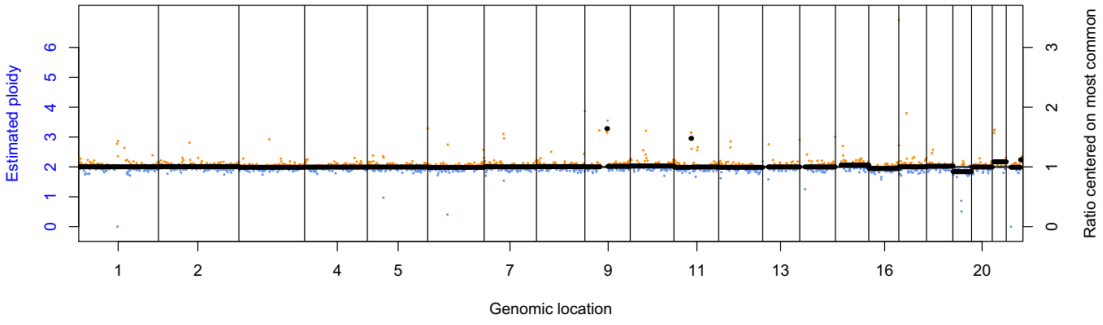
P483 (57 years old, male, NGH, ex-smoker, stage IIA squamous) 531 CNA score



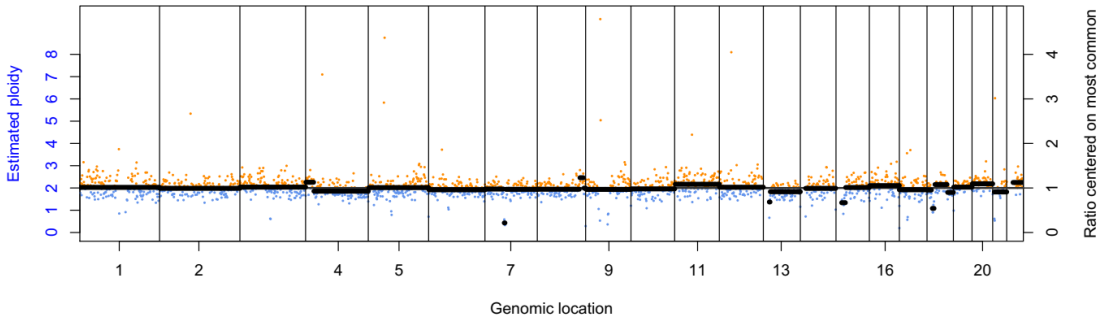
P493 (49 years, male, NGH, ex-smoker, stage II NSCLC NOS) 344 CNA score



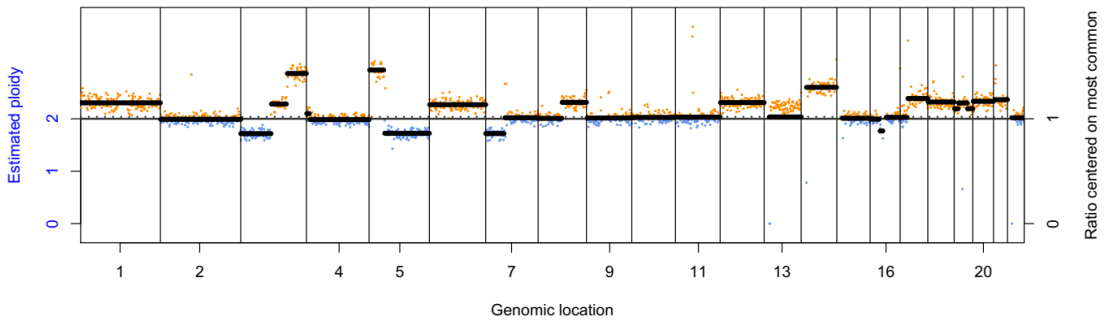
P527 (50 years, male, WPH, unknown, stage IV adenocarcinoma) 556 CNA score



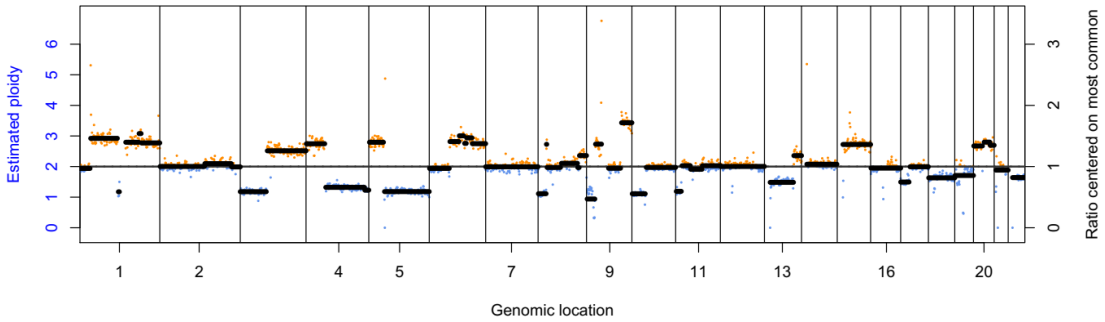
P539 (53 years, male, NGH, unknown, stage IIIA squamous) 520 CNA score



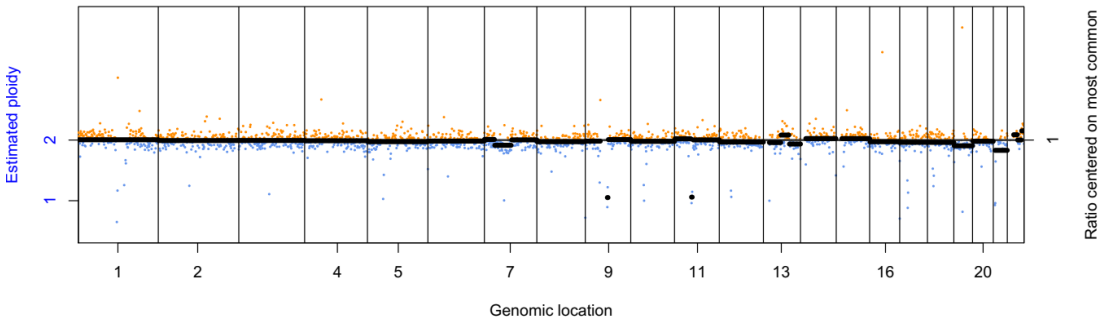
P640 (70 years, male, WPH, ex-smoker, stage IV NSCLC NOS) 16458 CNA score



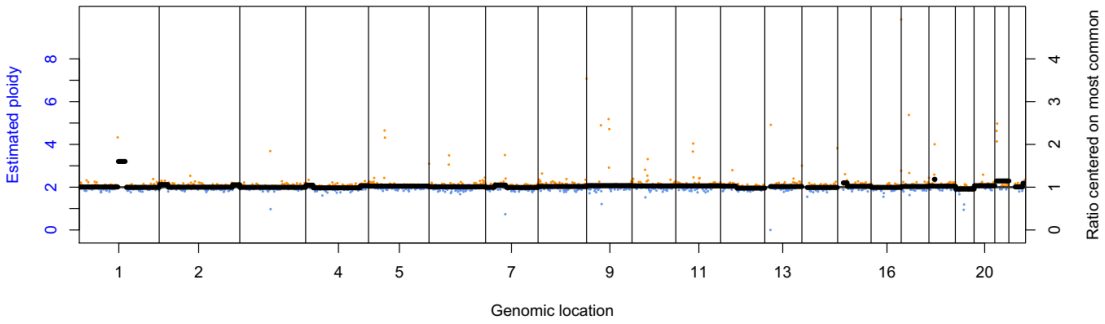
P710 (55 years, male, DRI, ex-smoker, stage IV SCLC) 5454 CNA score



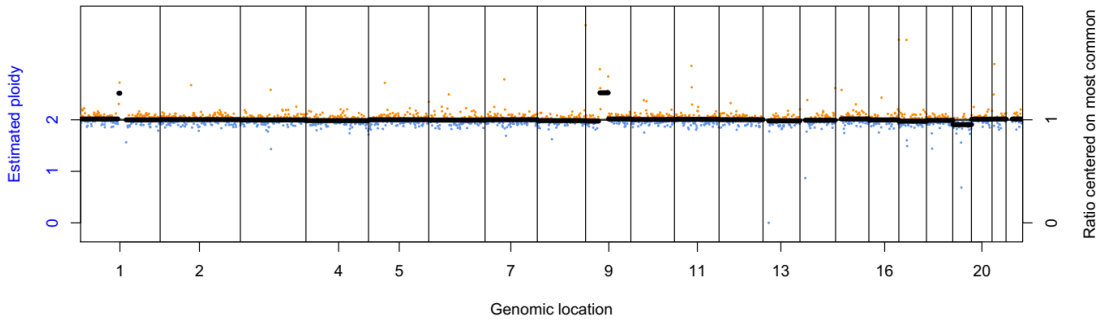
P790 (59 years, female, DRI, current, stage IIIA adenocarcinoma) 398 CNA score



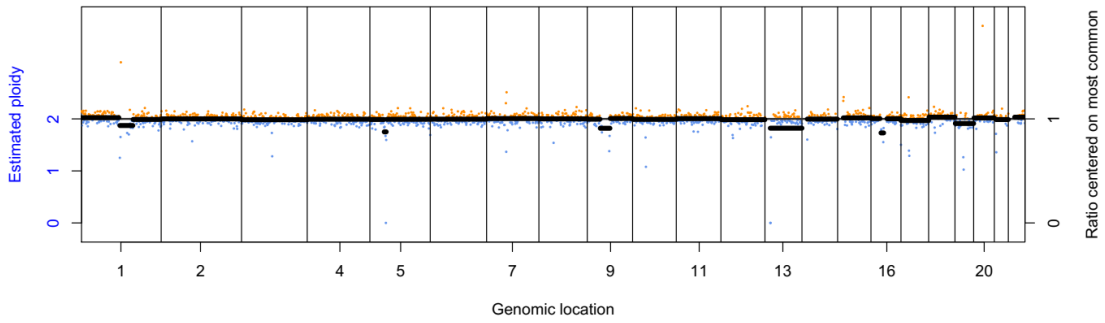
P800 (56 years, female, DRI, never smoked, stage IV squamous) 1635 CNA score



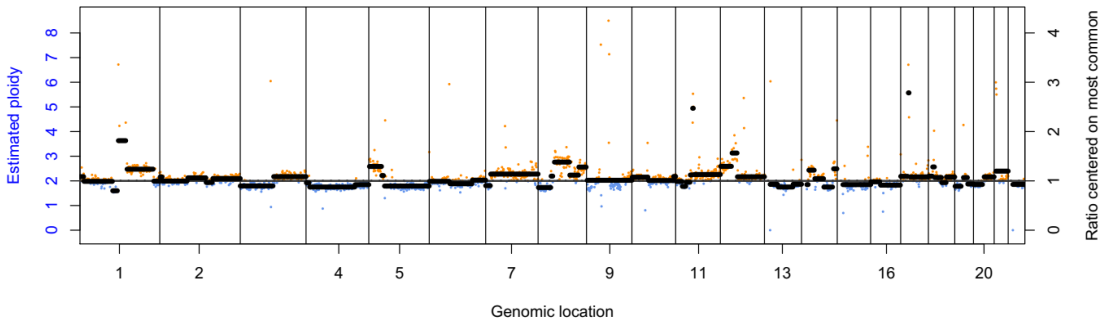
P801 (55 years, male, WPH, ex-smoker, stage IV NSCLC NOS) 213 CNA score



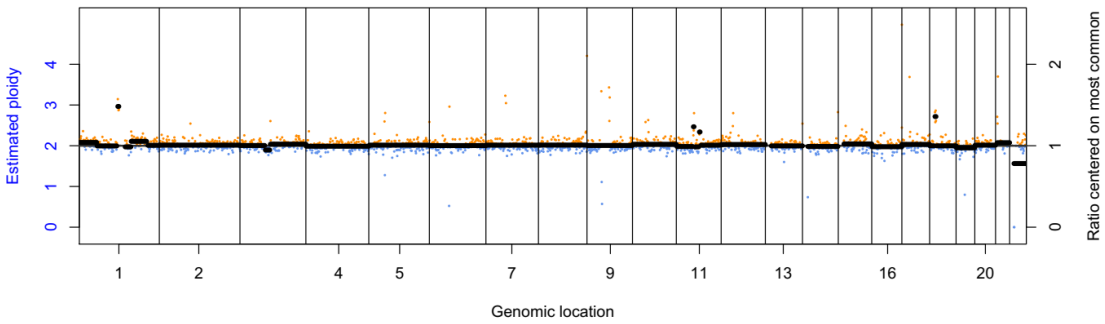
P816 (59 years, female, DRI, ex-smoker, stage IIA adenocarcinoma) 167 CNA score



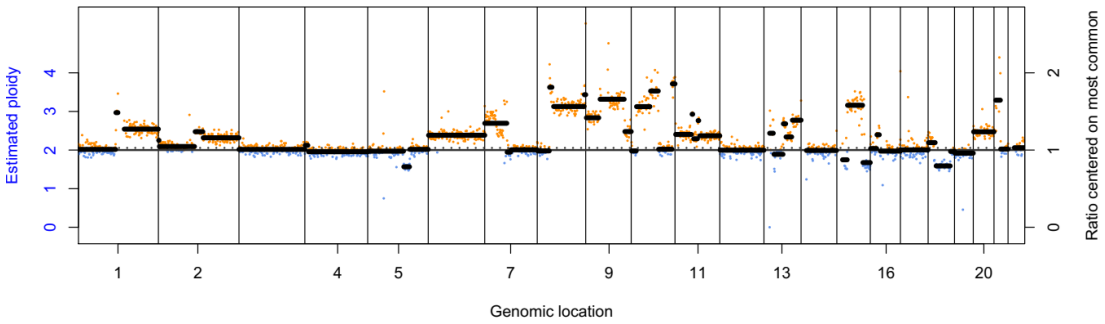
P823 (53 years, female, AGH, current, stage IV adenocarcinoma) 12229 CNA score



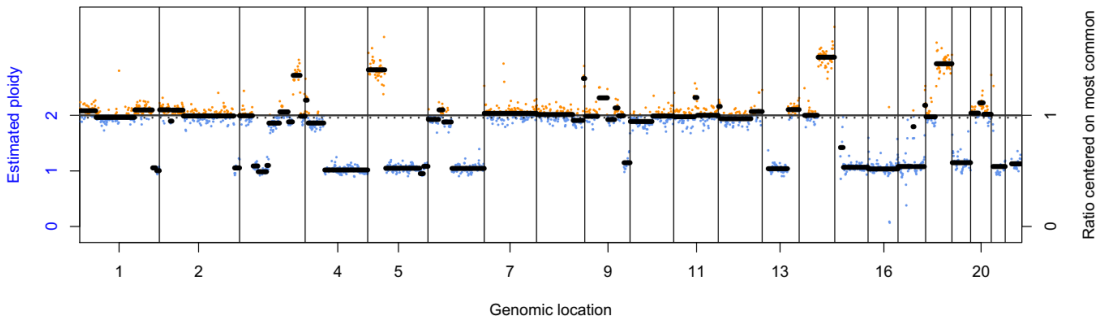
P855 (60 years, male, VHC, current, stage IIIA NSCLC NOS) 346 CNA score



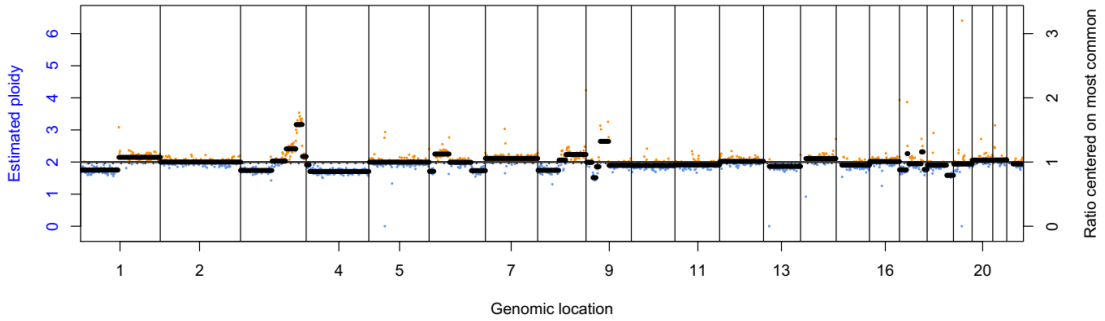
P858 (41 years, male, VHC, ex-smoker, stage IIIB, NSCLC NOS) 43445 CNA score



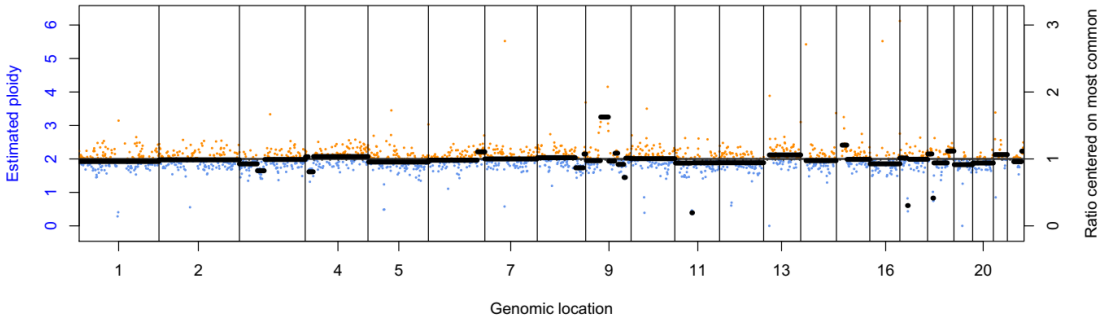
P878 (57 years, female, DRI, current, stage IV, SCLC) 66869 CNA score



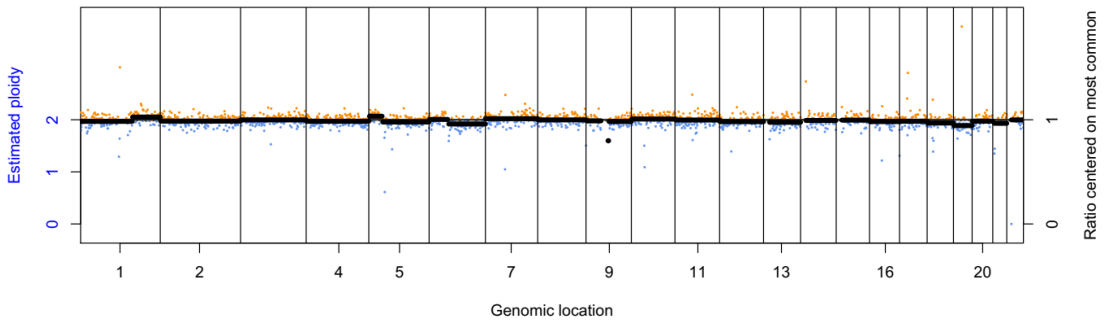
P1024 (56 years, male, CHN, Unknown, stage IIIA squamous) 6757 CNA score



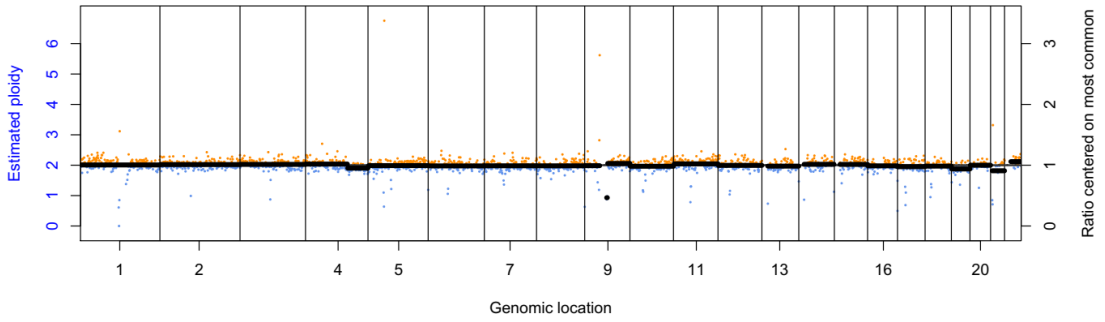
P1027 (61 years, female, CHN, never smoked, stage IV NSCLC NOS) 2877 CNA score



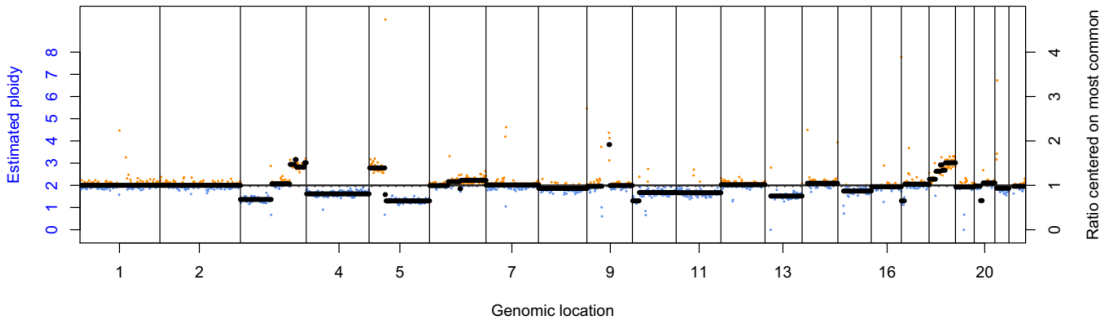
P1052 (60 years, female, DRI, ex-smoker, stage IIIB NSCLC NOS) 453 CNA score



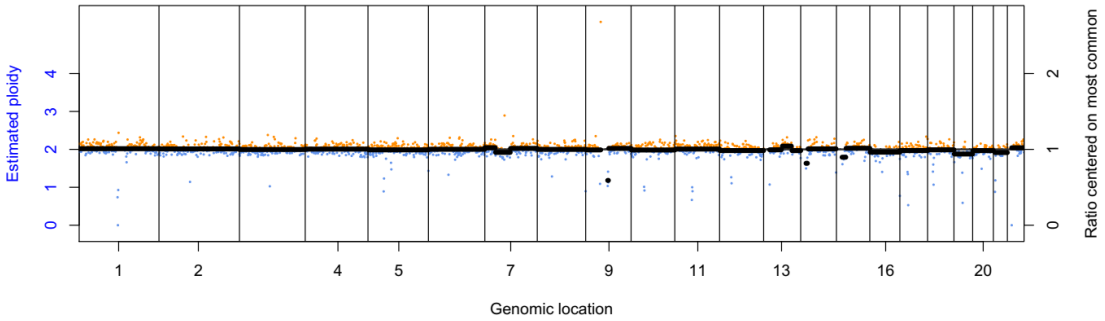
P1103 (61 years, male, DRI, current, stage IV adenocarcinoma) 813 CNA score



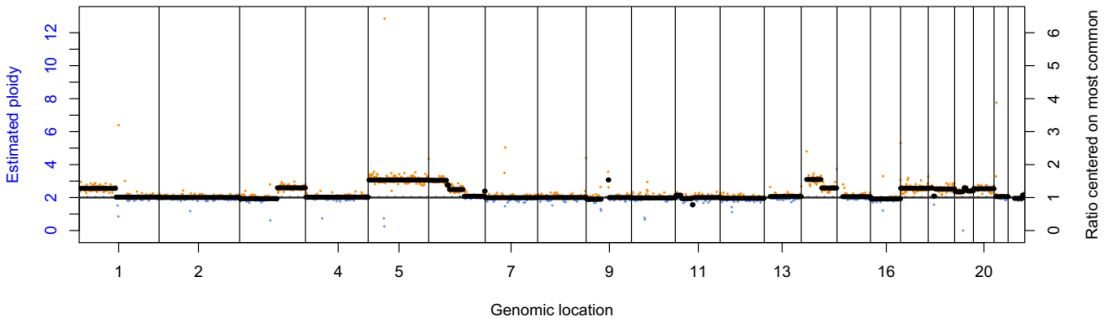
P1106 (51 years, male, DRI, current, stage IV SCLC) 24652 CNA score



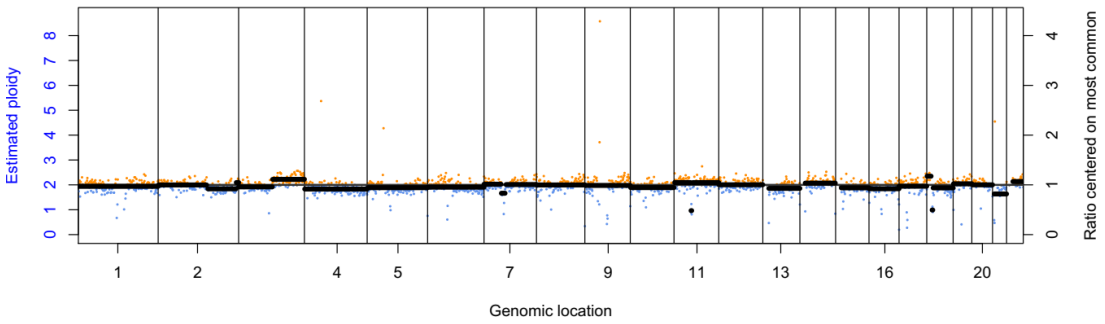
P1111 (60 years, female, WPH, ex-smoker, stage IIIB adenocarcinoma) 592 CNA score



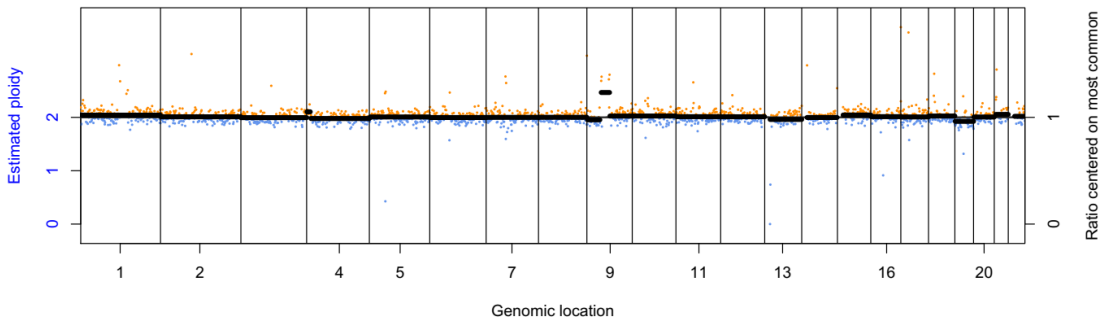
P1134 (52 years, male, CHM, ex-smoker, stage IV, SCLC) 35996 CNA score



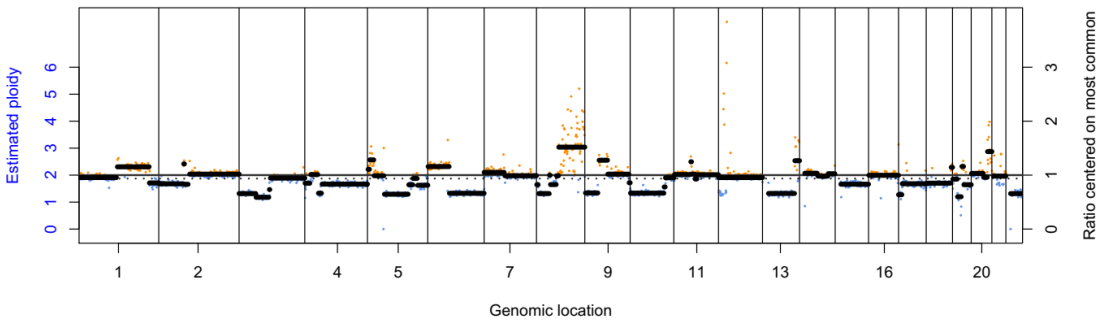
P1151 (61 years, male, WPH, ex-smoker, stage IIIB squamous) 3064 CNA score



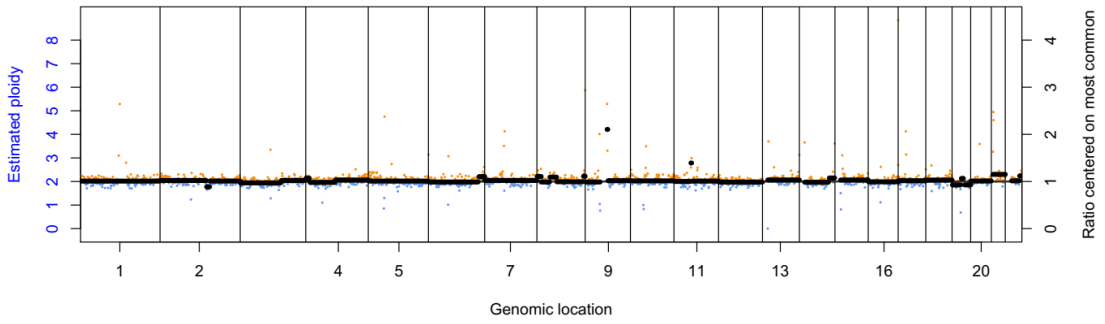
P1155 (60 years, male, WPH, unknown, stage IV adenocarcinoma) 415 CNA score



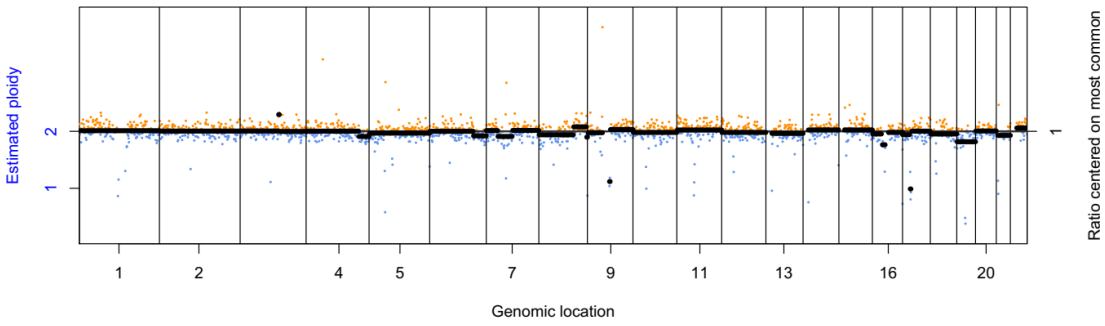
P1156 (59 years, female, DRI, current, stage IIIB, squamous) 34275 CNA score



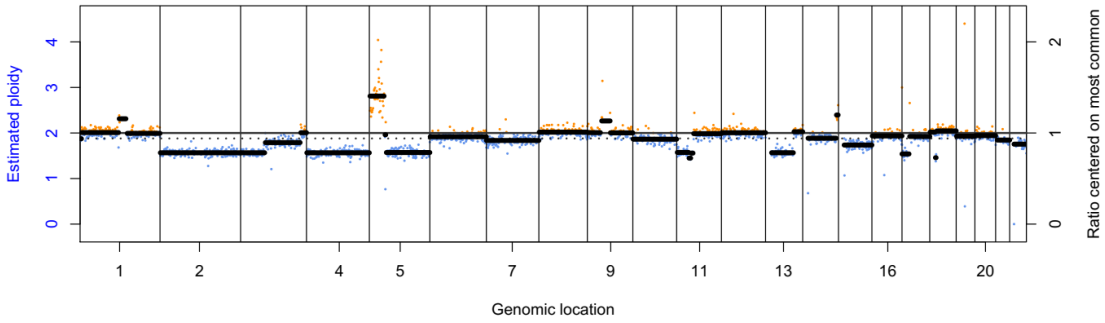
P1165 (58 years, male, WPH, unknown, stage IV squamous) 1389 CNA score



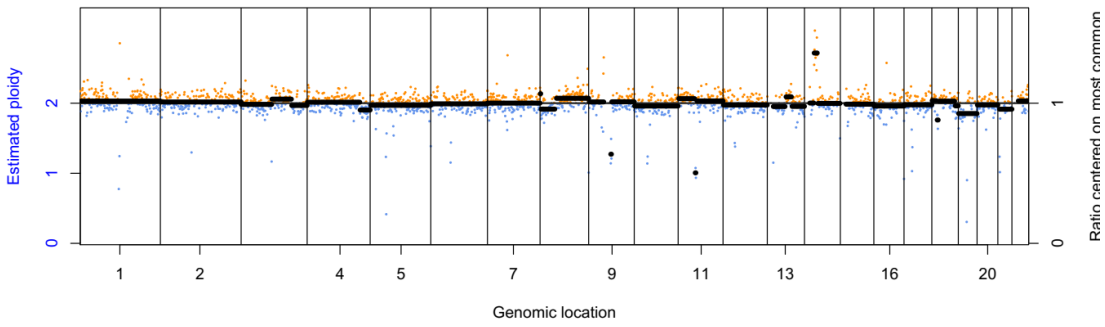
P1182 (51 years, female, NGH, ex-smoker, stage IIB squamous) 430 CNA score



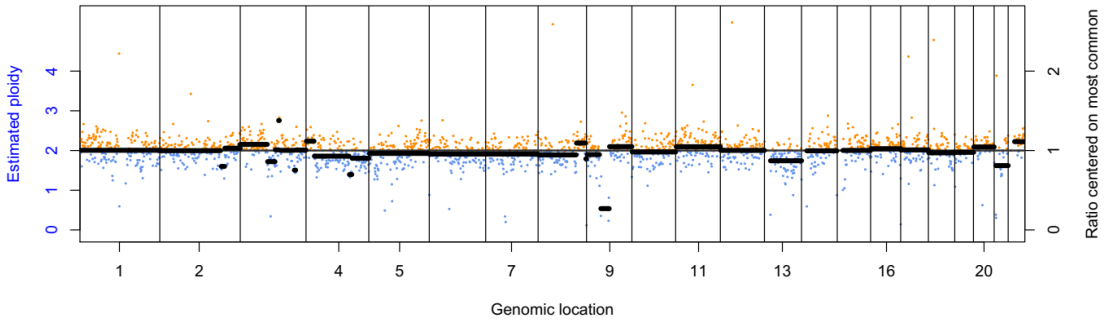
P1271 (60 years, male, WPH, ex-smoker, stage IV SCLC) 14277 CNA score



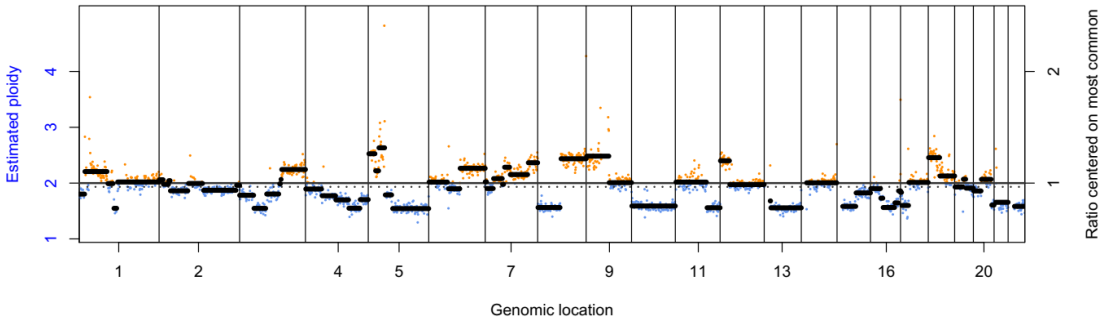
P1301 (59 years, female, NGH, current, stage IIB adenocarcinoma) 558 CNA score



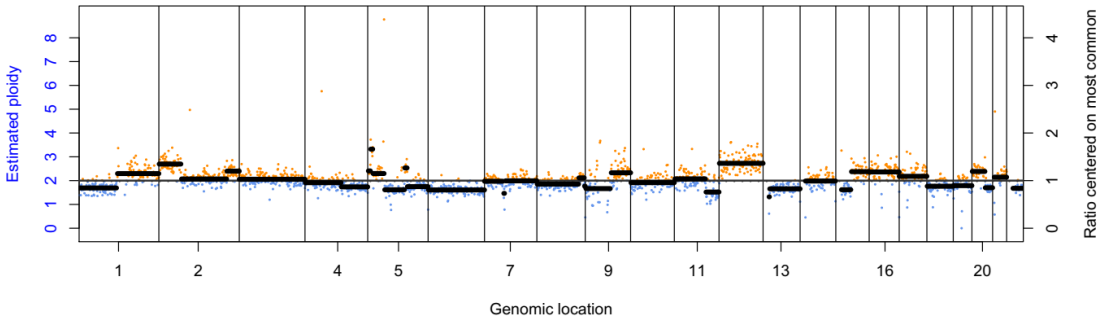
P1353 (60 years, female, SGH, ex-smoker, stage IIIB adenocarcinoma) 4791 CNA score



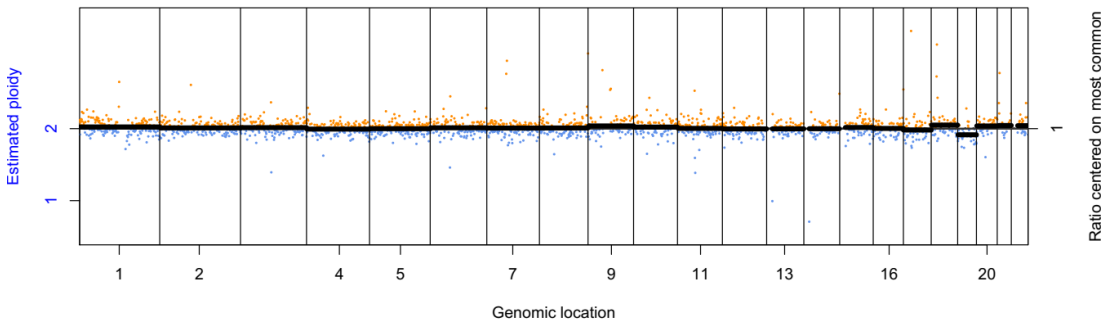
P1473 (52 years, female, NGH, unknown, stage IIA NSCLC other) 15373 CNA score



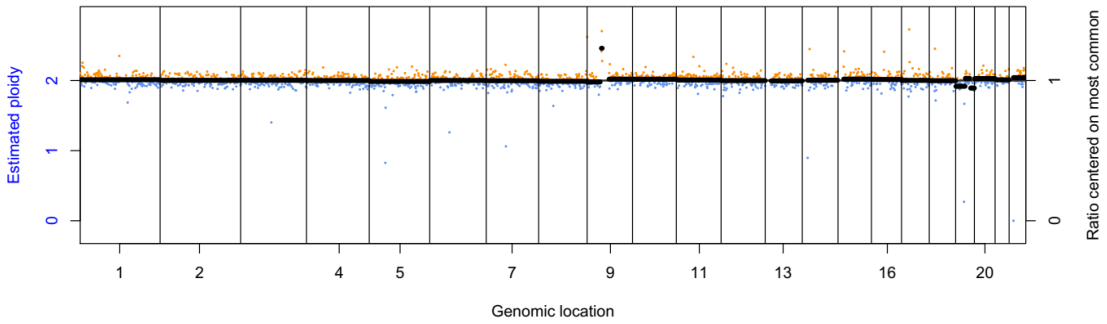
P1518 (52 years, male, CHM, current, stage IV adenocarcinoma) 18823 CNA score



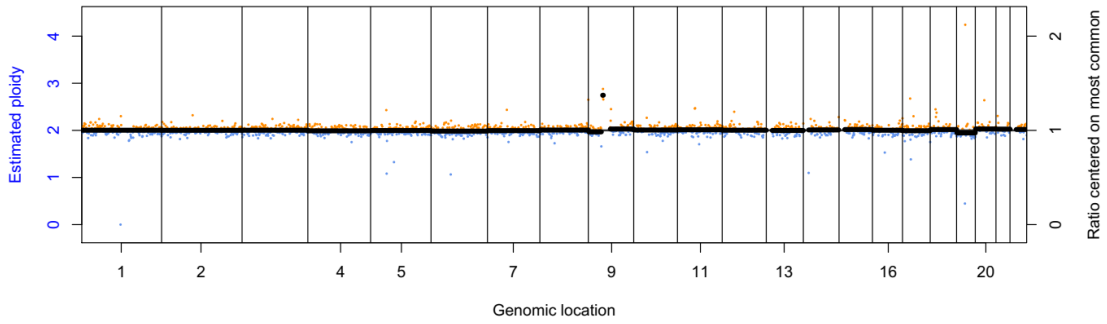
P1634 (51 years, female, NGH, unknown, stage IA adenocarcinoma) 225 CNA score



P1646 (48 years, female, NGH, ex-smoker, stage IA adenocarcinoma) 117 CNA score

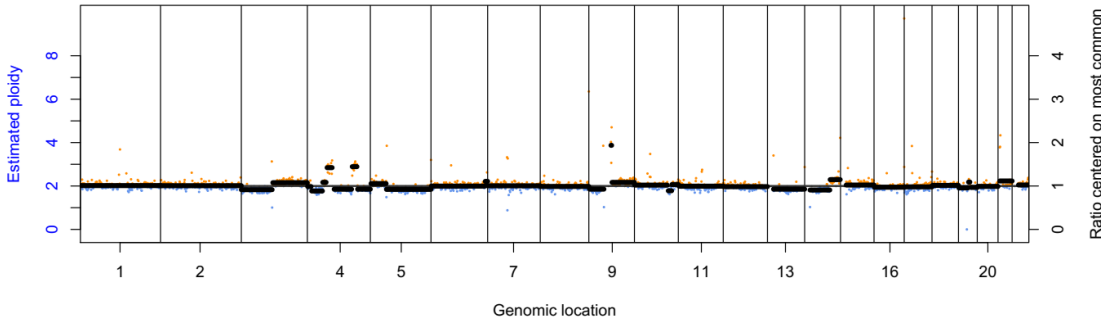


P1767 (53 years, female, NGH, ex-smoker, stage IA adenocarcinoma) 144 CNA score

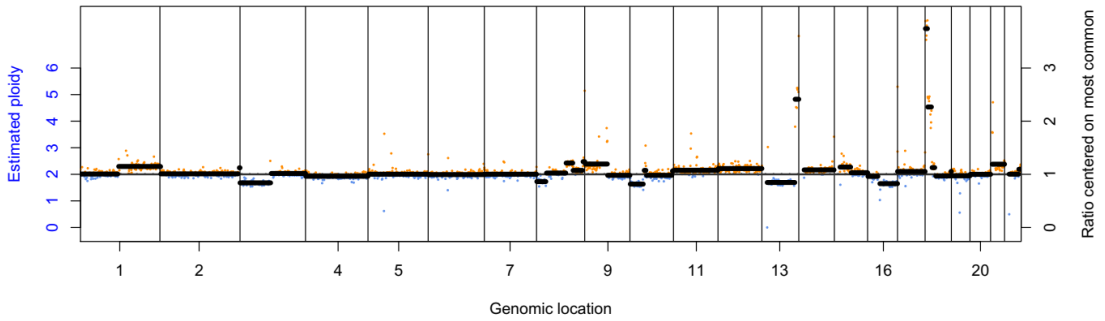


<u>Treated lung cancer cases (age, gender, centre, smoking status,stage, pathology)(N=11)</u>

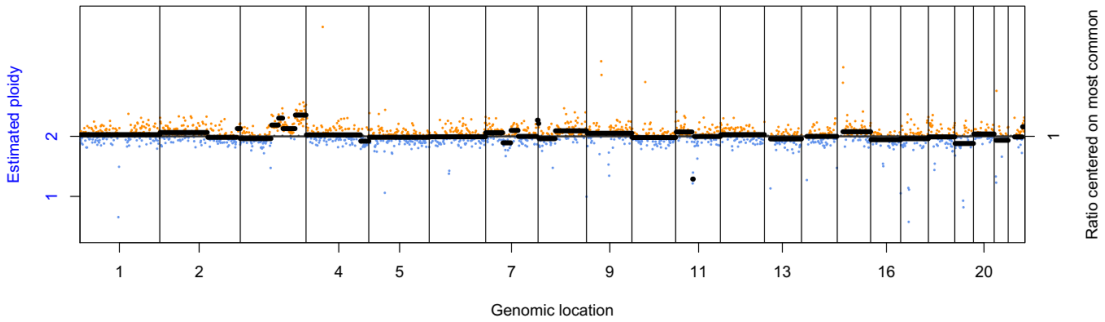
P52 (53 years, female, WPH, ex-smoker, stage IV Squamous) 2269 CNA score



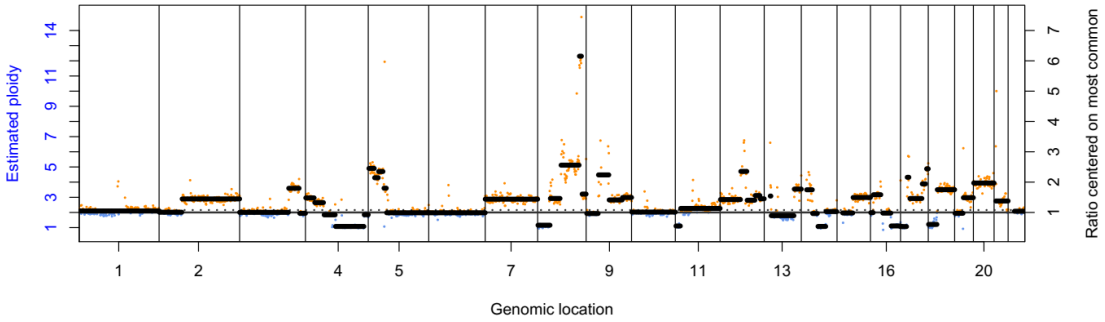
P203 (60 years, female, WPH, ex-smoker, stage IIIB SCLC) 25760 CNA score



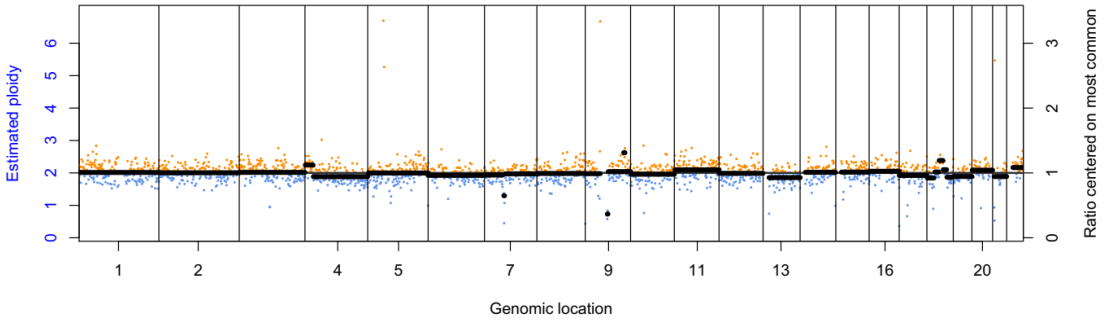
P222 (57 years, male, DRI, ex-smoker, stage IIIA Squamous) 983 CNA score



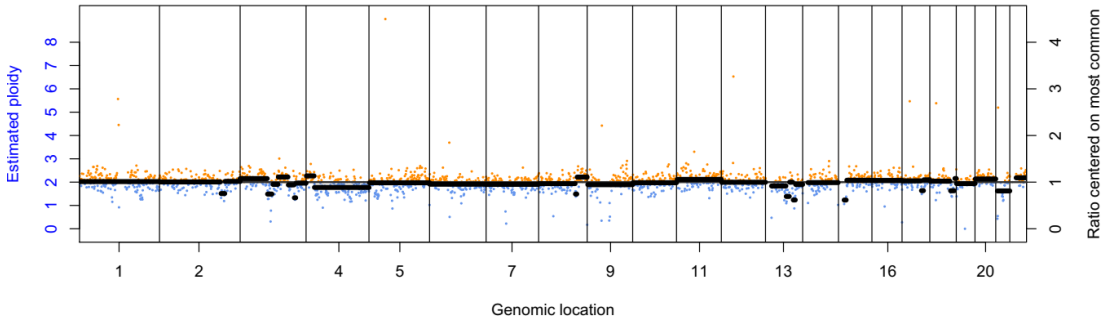
P291 (50 years, female, WPH, current, stage IV NOS) 169039 CNA score



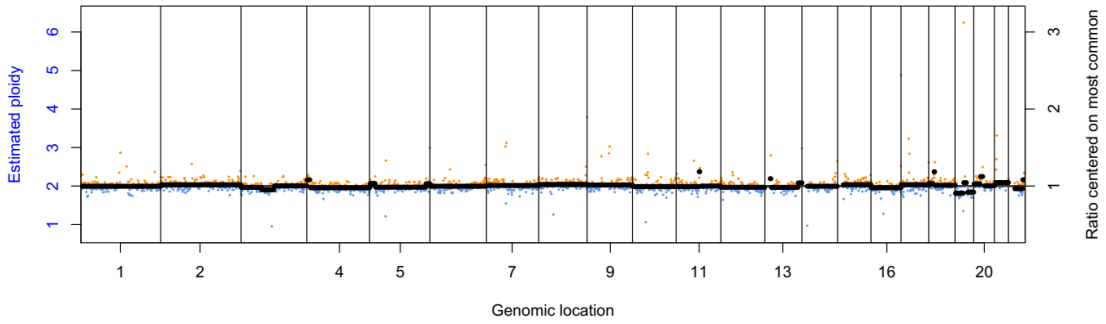
P338 (75 years, male, WPH, ex-smoker, stage IV Squamous) 2696 CNA score



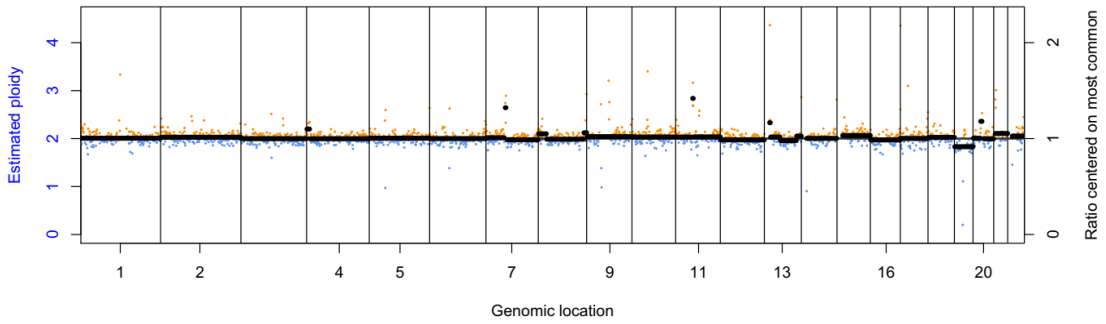
P765 (55 years, female, WPH, unknown, stage IV adenocarcinoma) 4808 CNA score



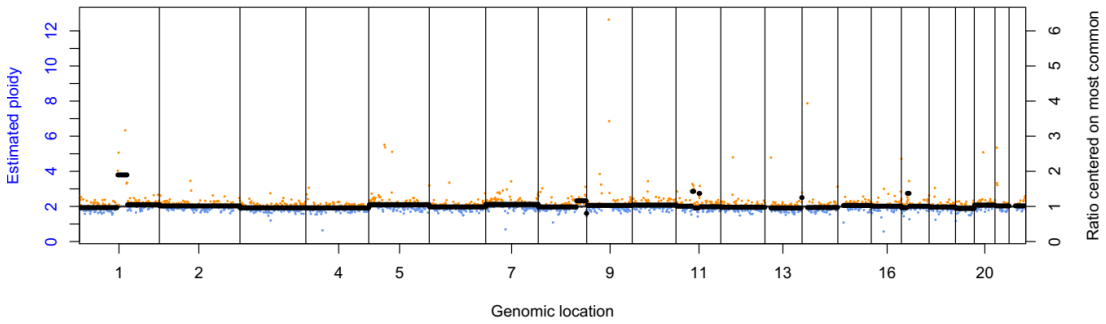
P805 (57 years, female, WPH, current, stage IV SCLC) 551 CNA score



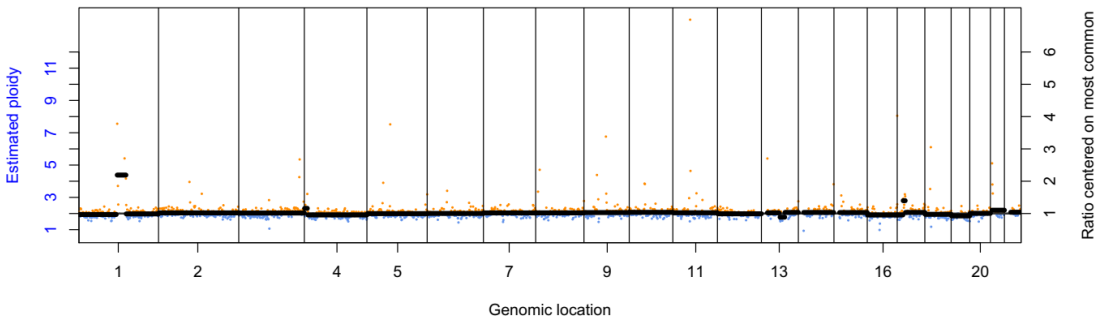
P806 (58 years, male, DRI, ex-smoker, stage IV mixed adeno/squamous) 531 CNA score



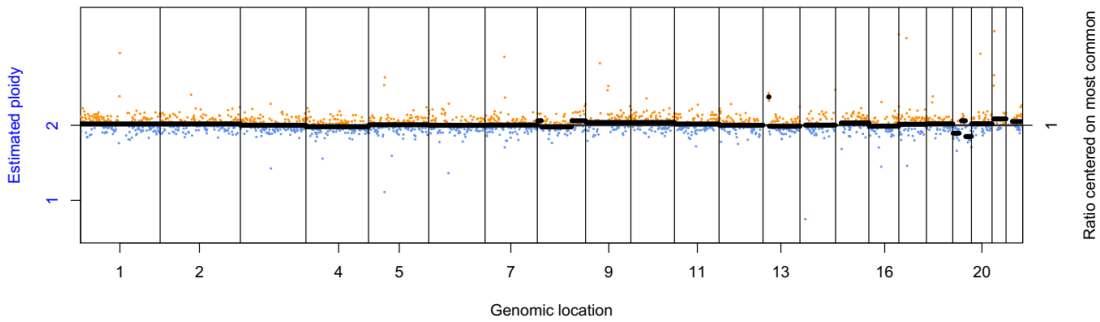
P1117 (52 years, female, WPH, current, stage IIIB adenocarcinoma) 1377 CNA score



P1324 (60 years, male, DRI, never smoked, stage IIIA squamous) 2187 CNA score

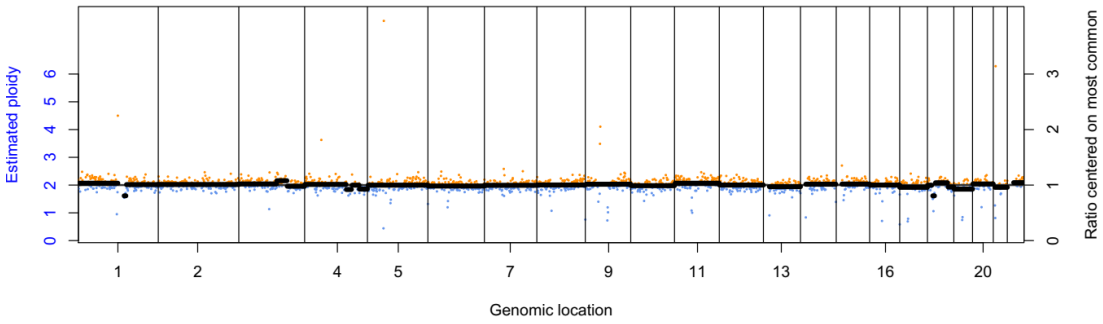


P1367 (69 years, female, DRI, unknown, stage IIA adenocarcinoma) 280 CNA score

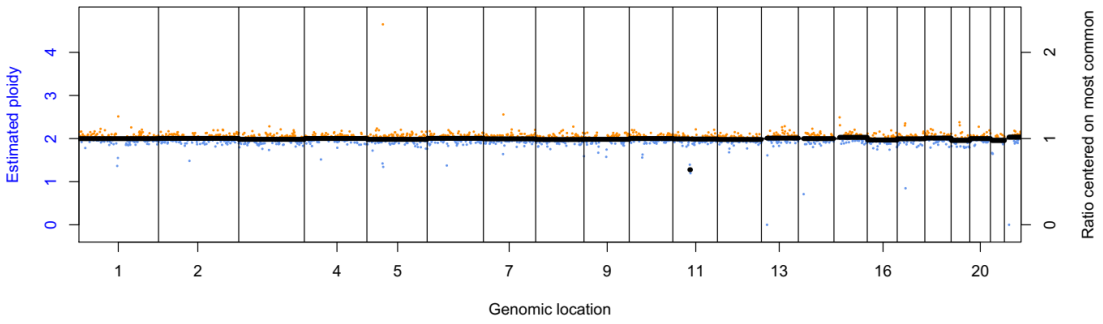


High risk controls (age, gender, centre, smoking history)(N=30)

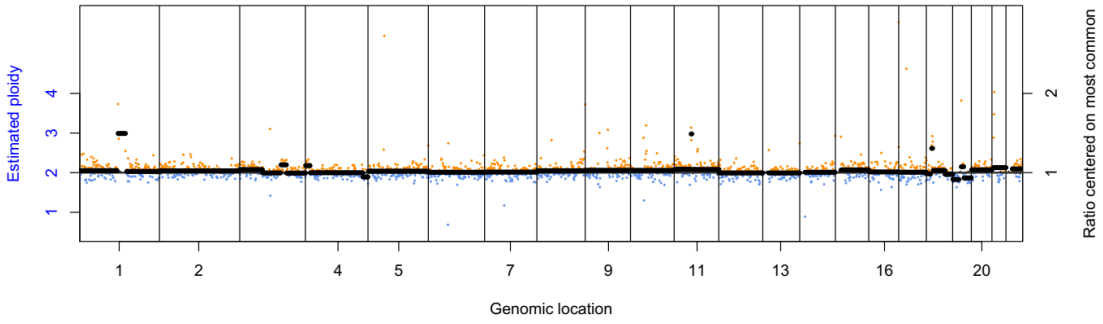
P27 (72 years, female, WPH, current) 842 CNA score



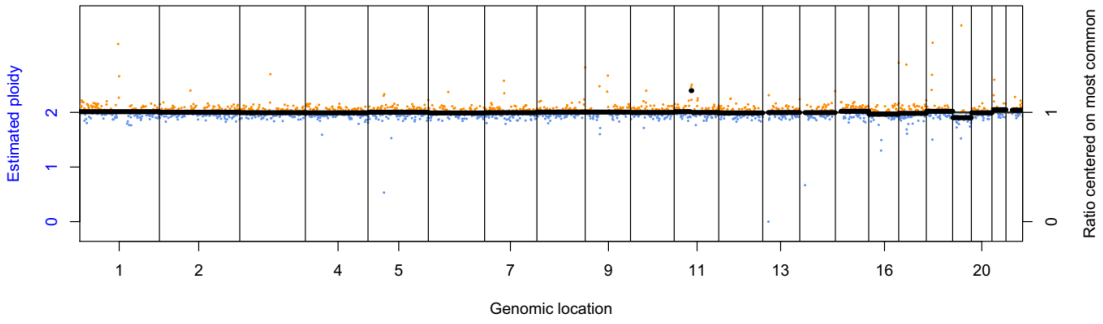
P38 (64 years, male, WPH, current) 210 CNA score



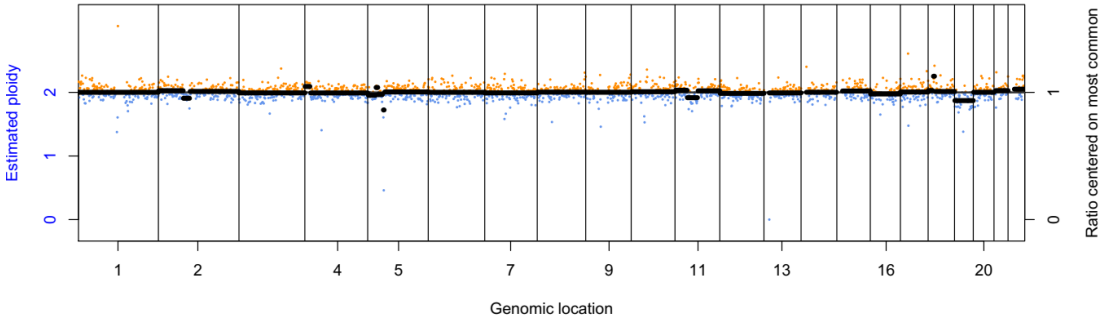
P39 (72 years, male, WPH, current) 916 CNA score



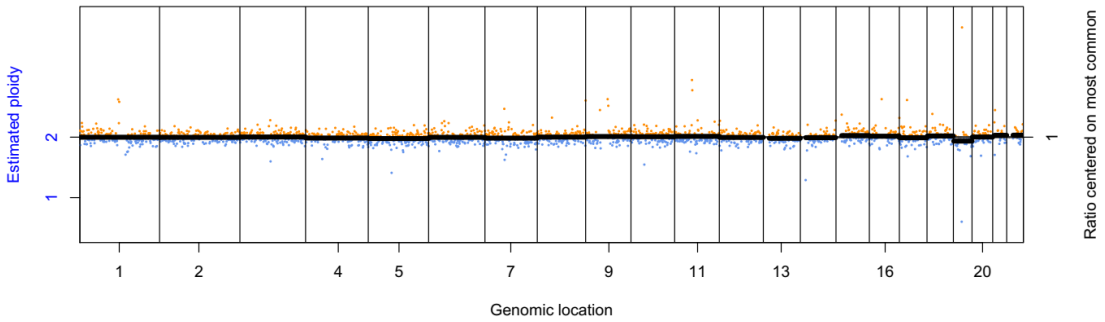
P46 (61 years, female, WPH, current) 194 CNA score



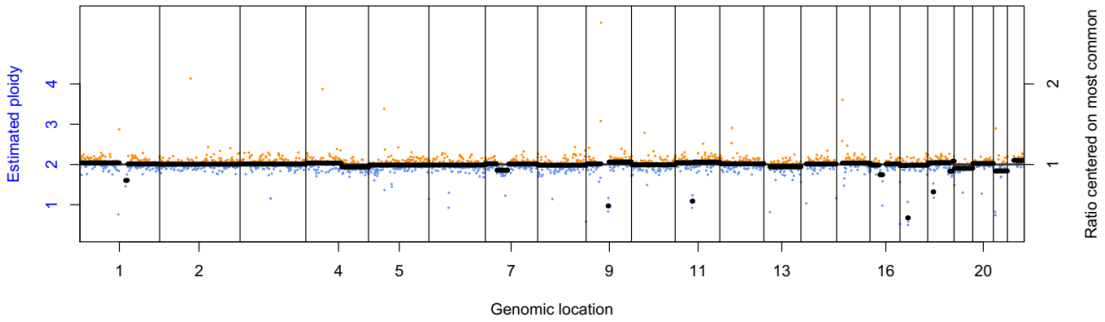
P98 (62 years, female, WPH, current) 266 CNA score



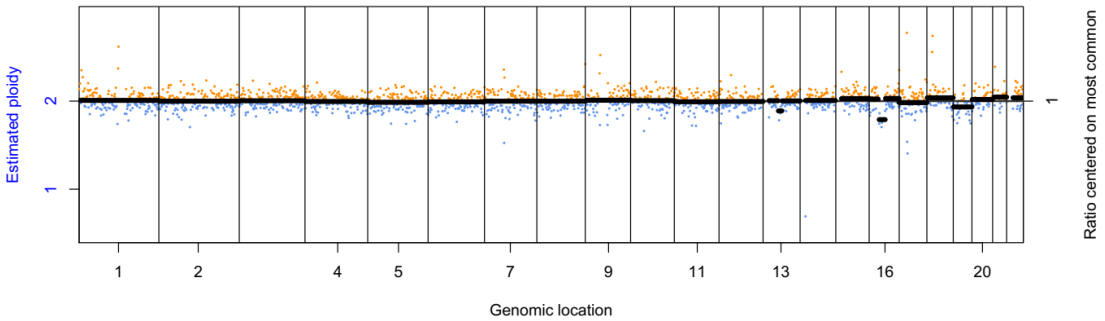
P107 (63 years, female, WPH, current) 172 CNA score



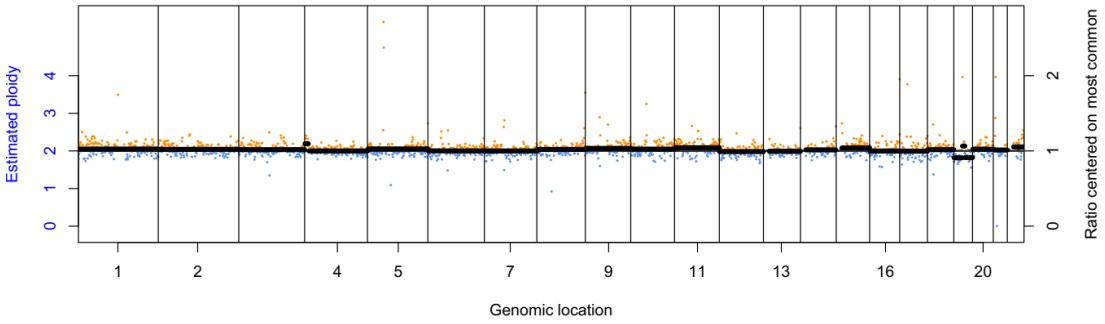
P154 (64 years, male, WPH, ex-smoker) 649 CNA score



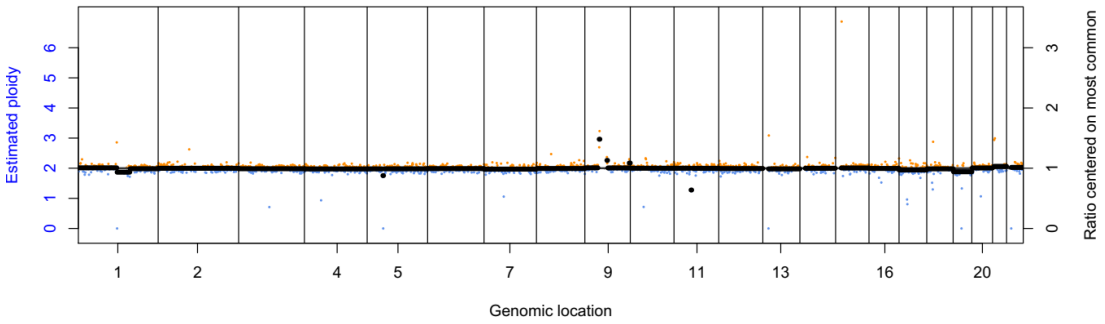
P241 (60 years, male, DRI, current) 170 CNA score



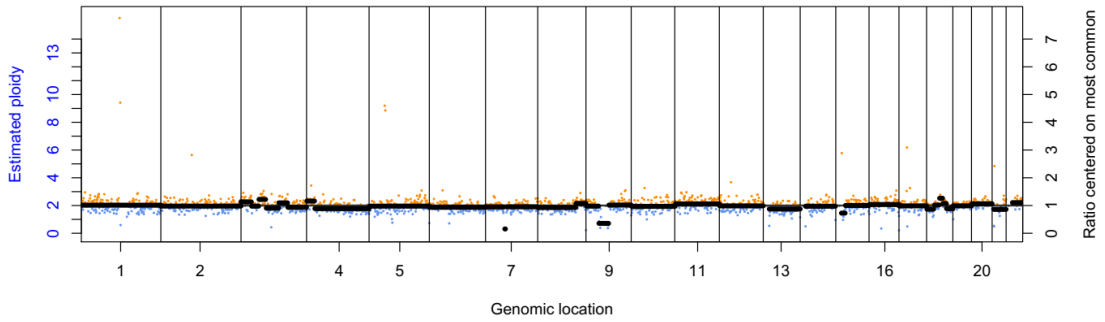
P247 (63 years, male, DRI, current) 803 CNA score



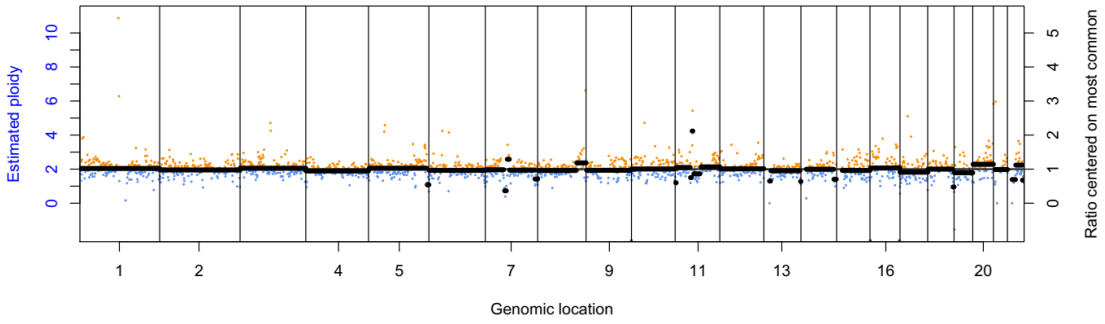
P293 (69 years, male, WPH, current) 280 CNA score



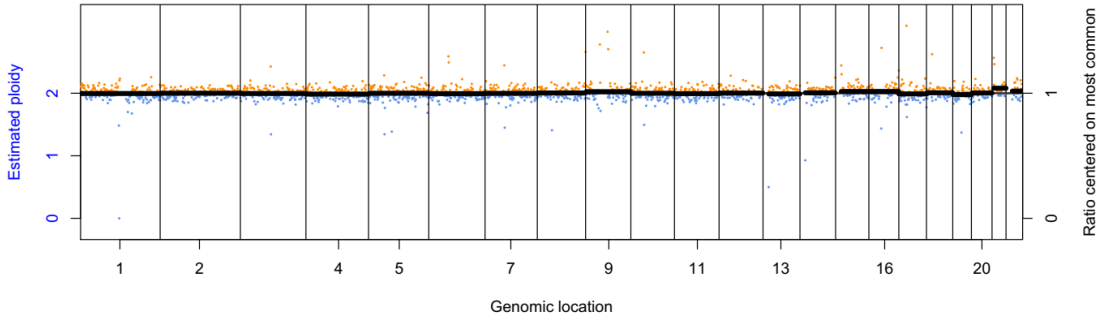
P355 (62 years, male, SGH, current) 6132 CNA score



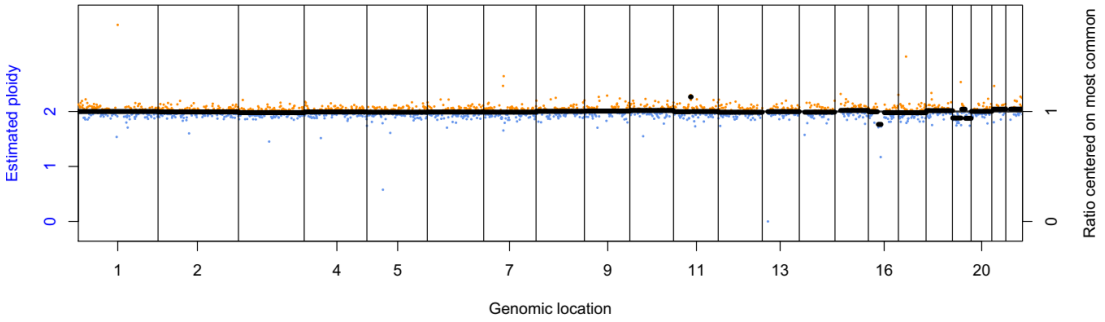
P368 (60 years, male, WPH, ex-smoker) 7122 CNA score



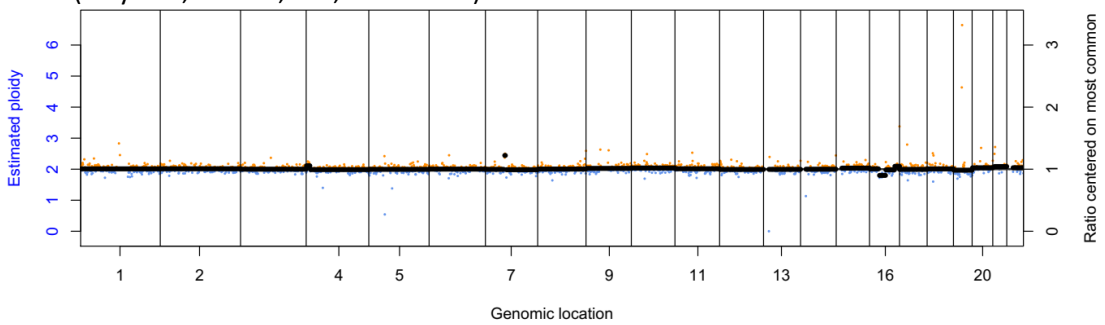
P439 (58 years, male, WPH, current) 149 CNA score



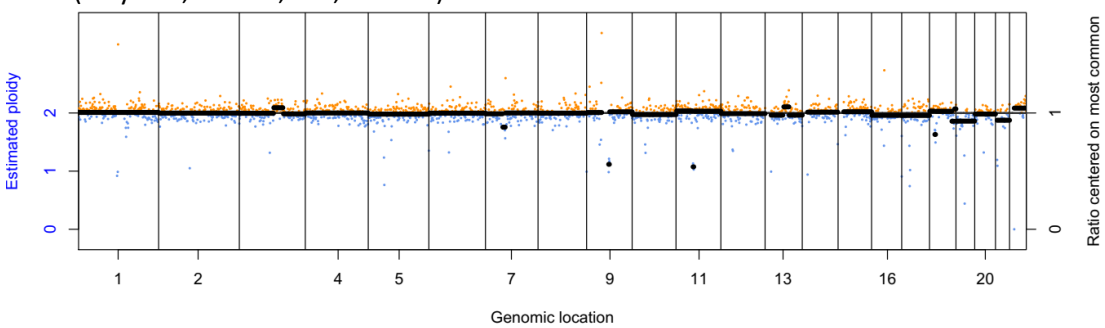
P561 (59 years, male, WPH, current) 192 CNA score



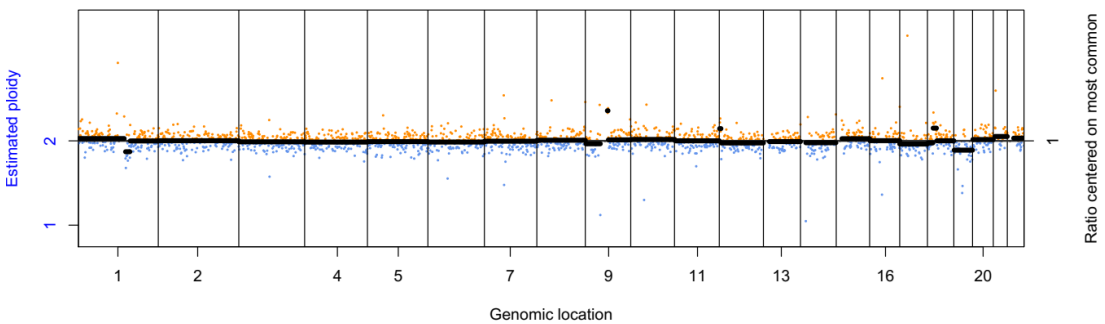
P584 (64 years, female, DRI, ex-smoker) 259 CNA score



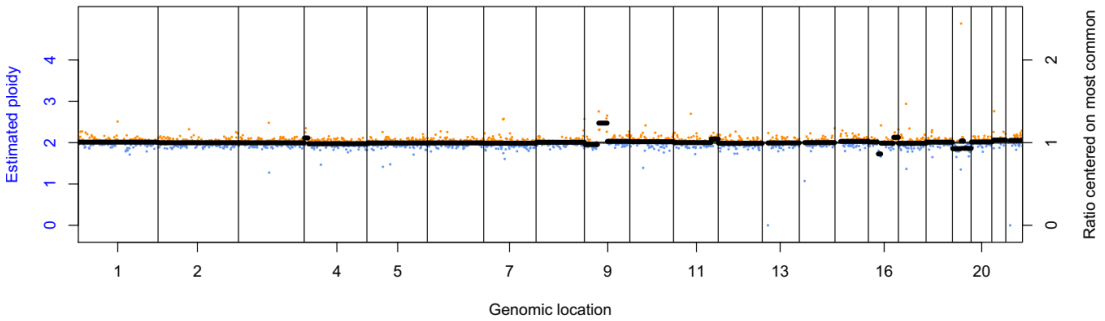
P589 (60 years, female, DRI, current) 442 CNA score



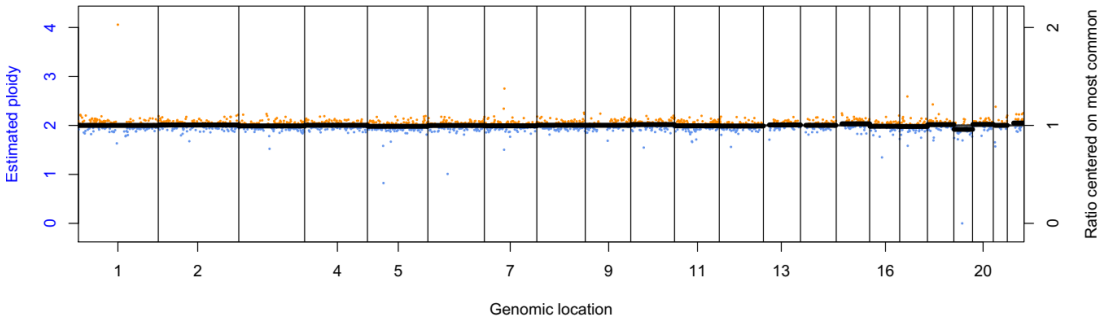
P769 (59 years, male, WPH, current) 233 CNA score



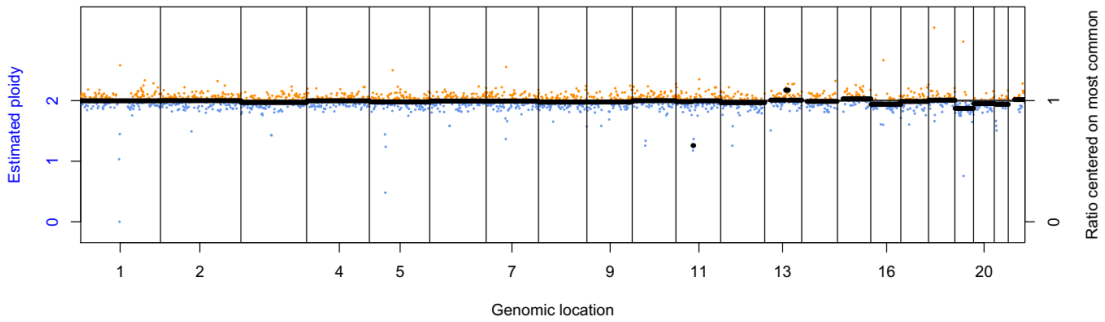
P778 (61 years, female, AGH, current) 307 CNA score



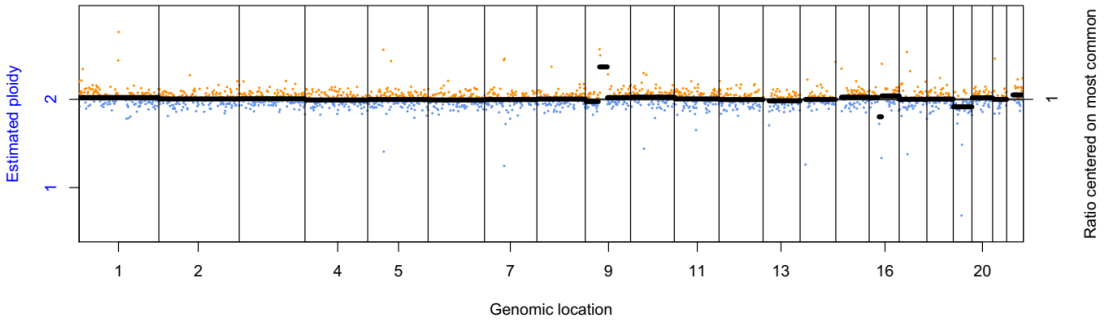
P798 (73 years, male, WPH, current) 204 CNA score



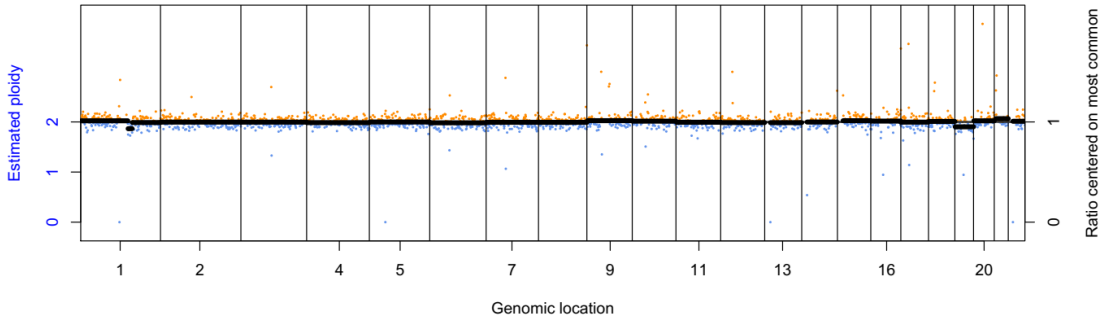
P886 (60 years, male, CHM, ex-smoker) 369 CNA score



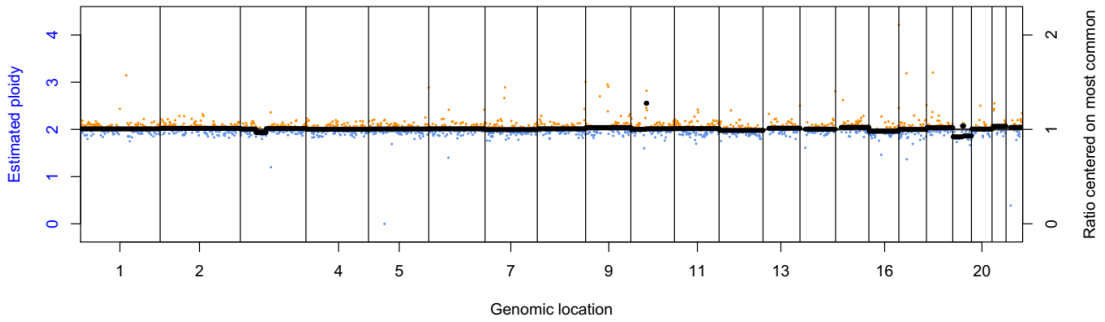
P901 (62 years, male, DRI, current) 169 CNA score



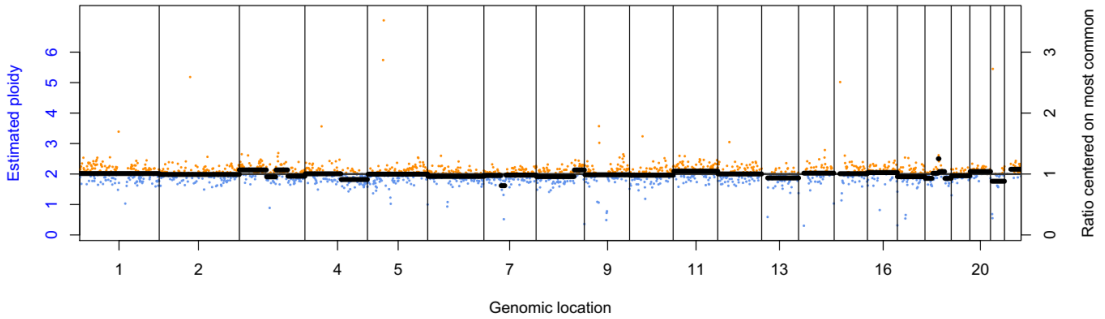
P987 (63 years, male, AGH, current) 217 CNA score



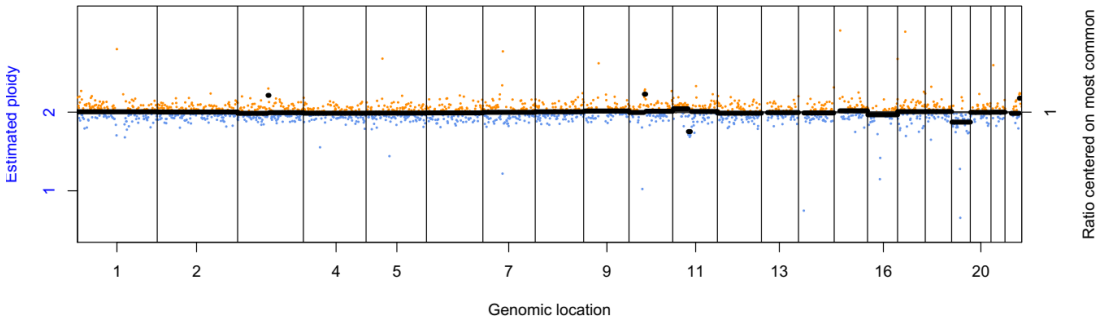
P1034 (61 years, male, CHM, current) 392 CNA score



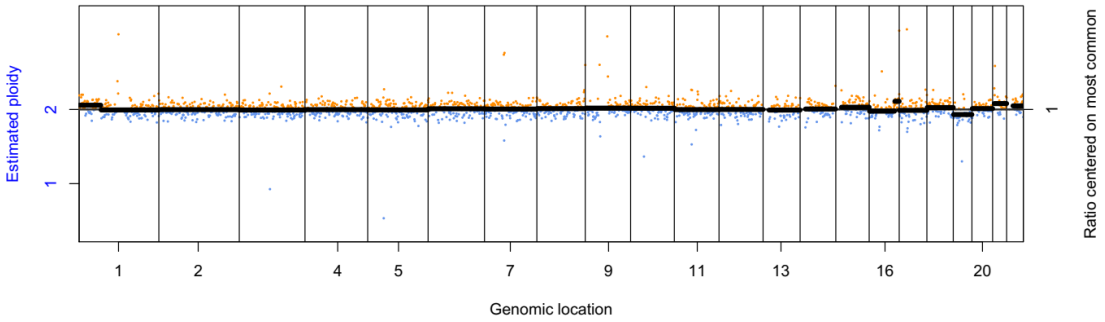
P1053 (62 years, female, DRI, current) 2390 CNA score



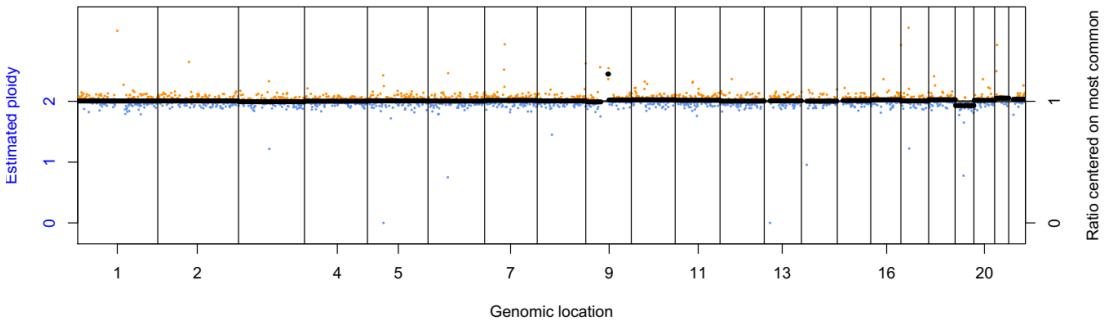
P1086 (70 years, female, CHM, current) 245 CNA score



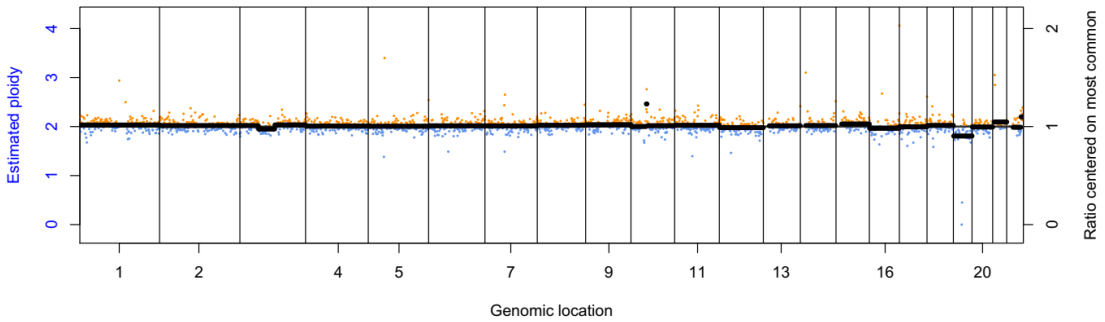
P1160 (57 years male, WPH, current) 211 CNA score



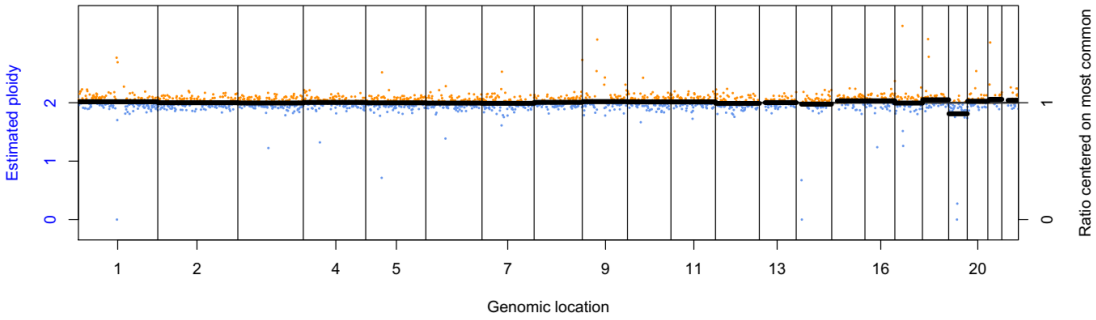
P1296 (65 years, male, WPH, ex-smoker) 149 CNA score



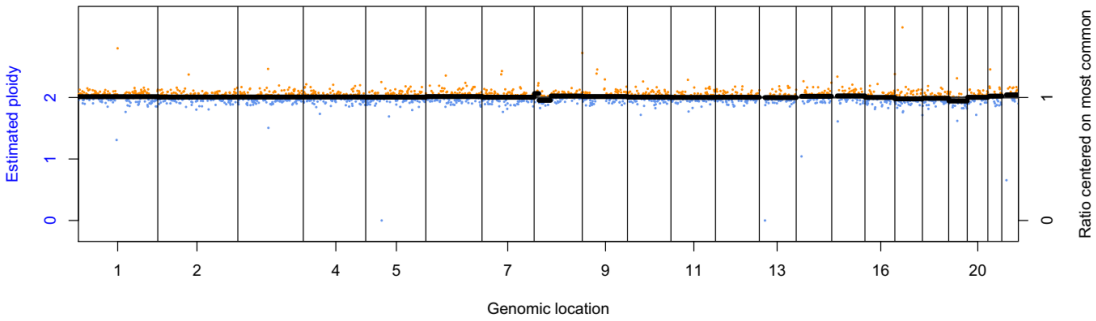
P1326 (58 years, male, WPH, current) 411 CNA score



P1332 (68 years, male, WPH, ex-smoker) 205 CNA score

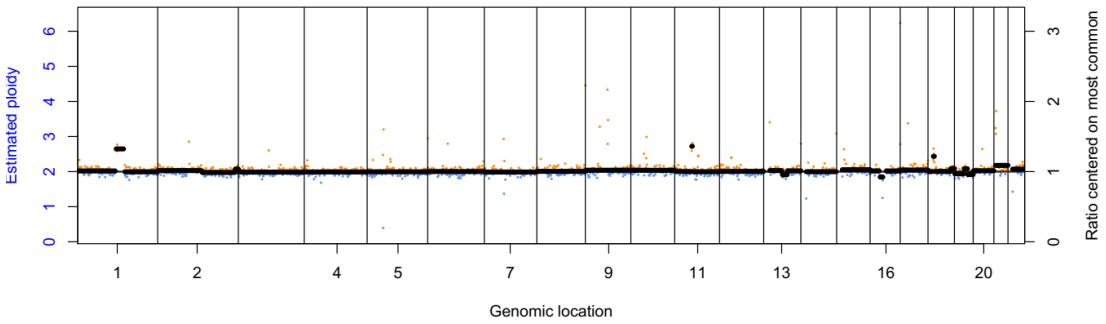


P1339 (60 years, male, WPH, current) 156 CNA score

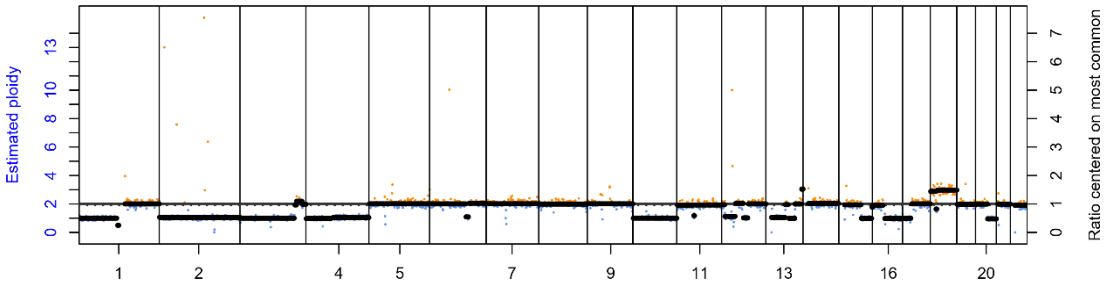


Low risk controls (N=10) (age, gender, centre, smoking history)

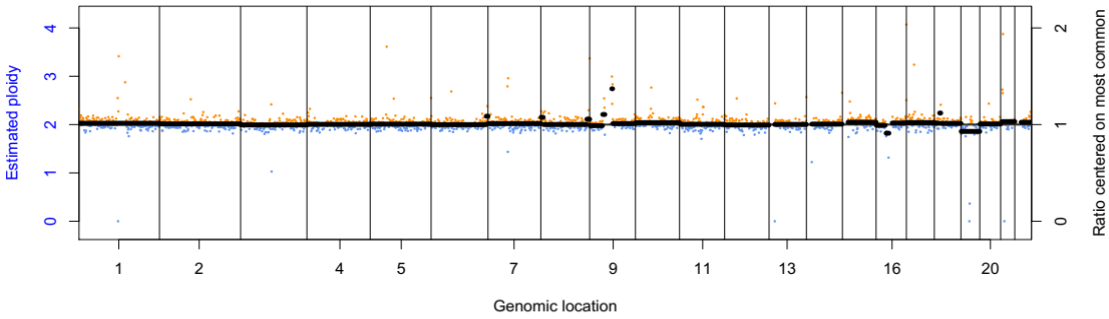
P116 (59 years, female, WPH, never smoked)



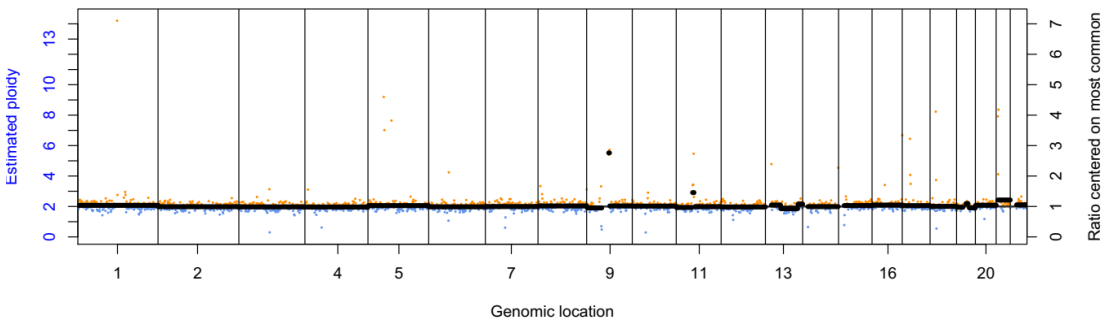
P185 (48 years, female, DRI, never smoked)



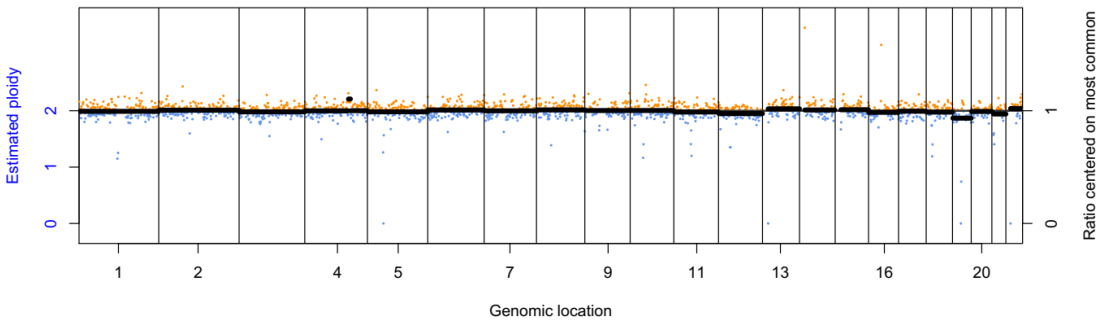
P369 (49 years, male, WPH, never smoked)



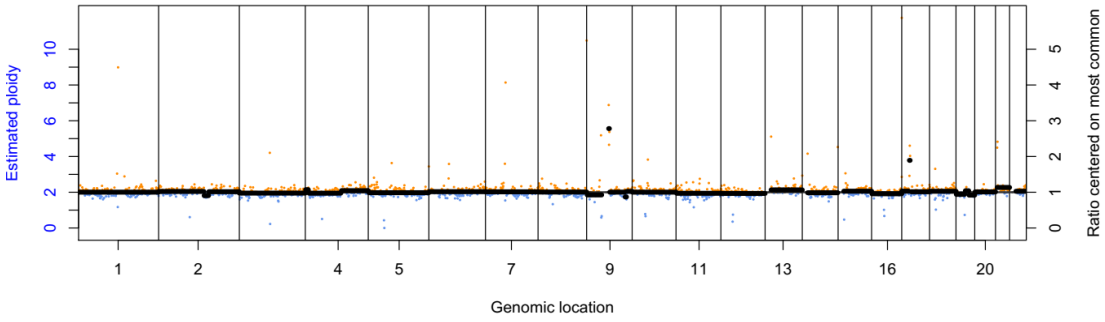
P388 (59 years, male,WPH, never smoked)



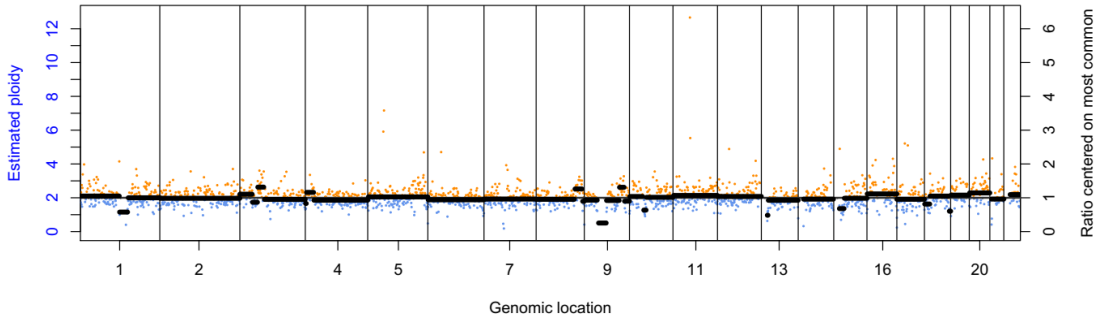
P397 (53 years, female, DRI, never smoked)



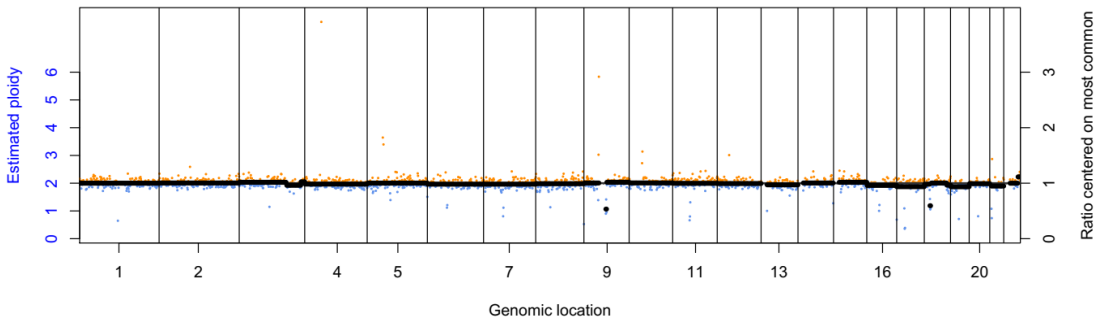
P413 (68 years, female, WPH, never smoked)



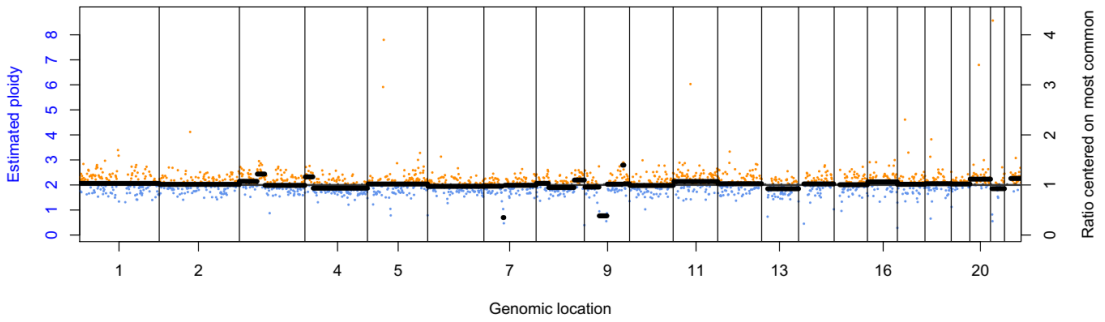
P441 (51 years, female, DRI, never smoked)



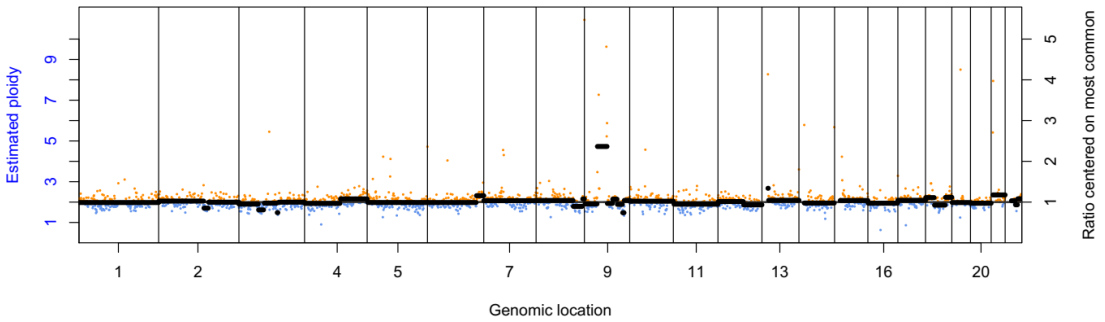
P541 (52 years, female, DRI, never smoked)



P721 (61 years, male, WPH, never smoked)

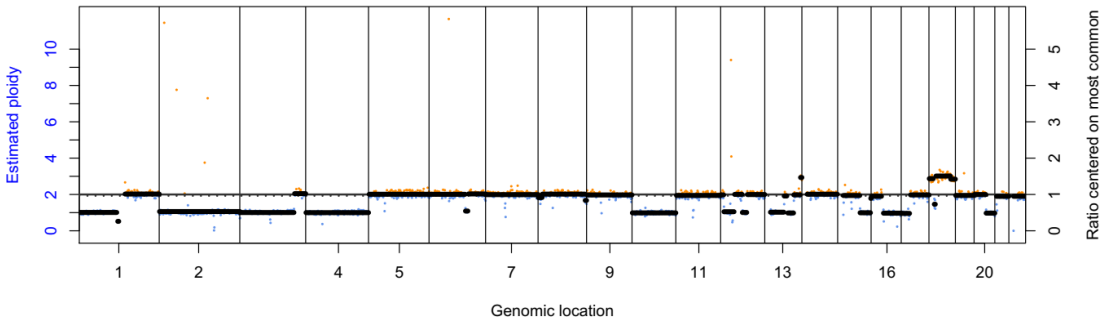


P1099 (65 years, female, WPH, never smoked)

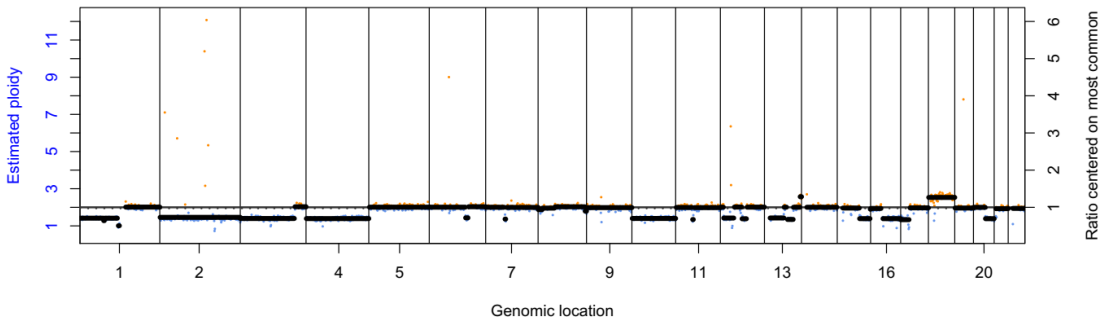


Copy number profiles of H69 cell line DNA spiked into pooled healthy volunteer cfDNA at varying proportions

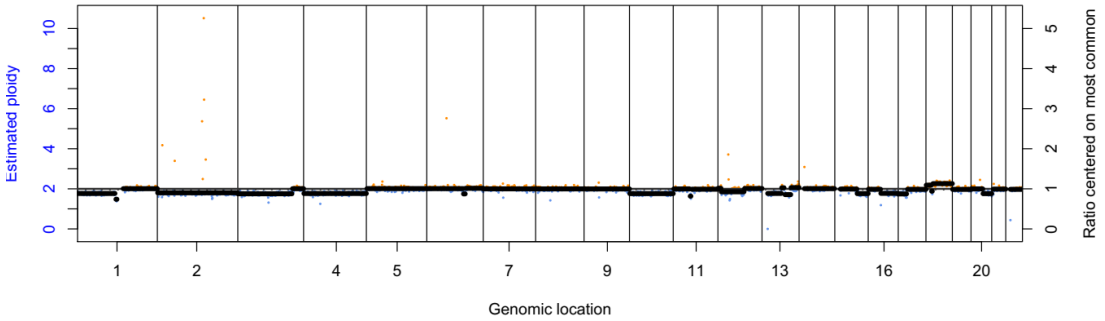
100% H69 DNA



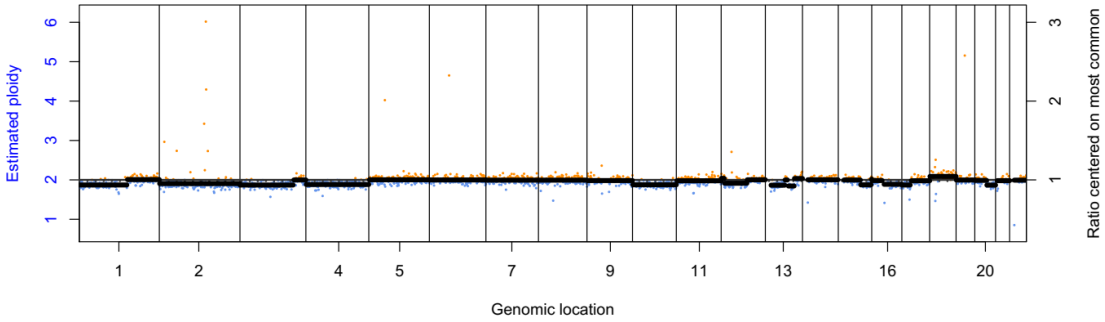
50% H69 DNA



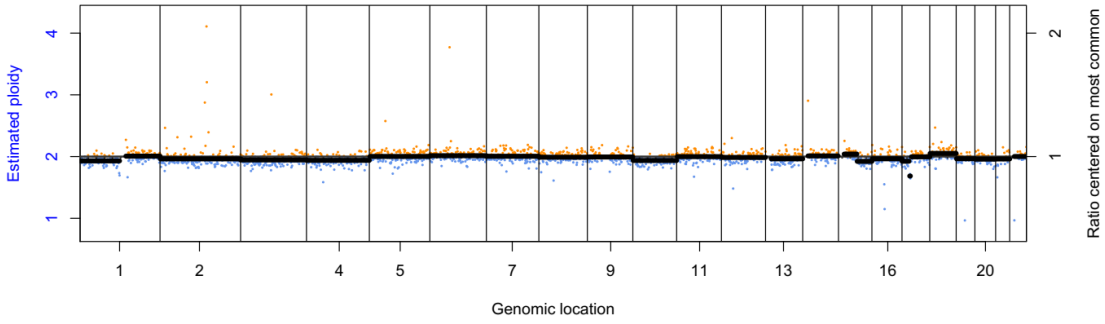
20% H69 DNA



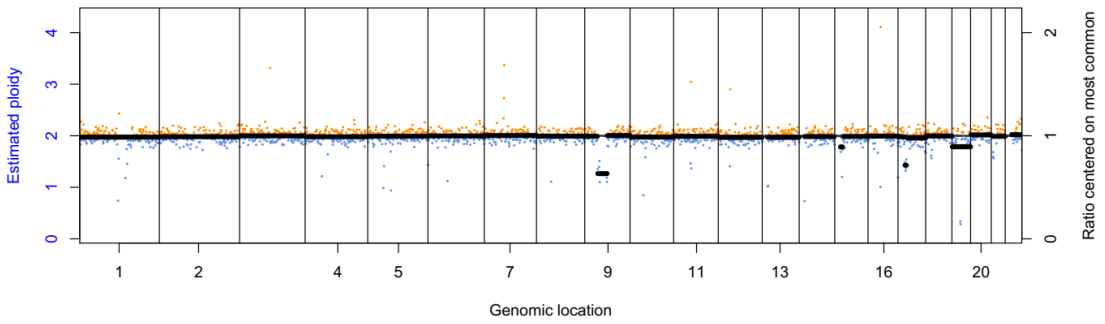
10% H69 DNA



5% H69 DNA

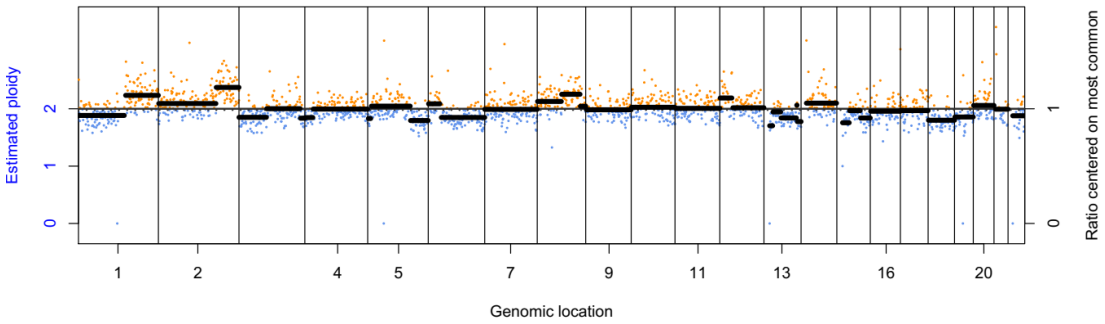


0% H69 DNA (100% pooled healthy control cfDNA)



Tumour FFPE DNA copy number profile

Tumour FFPE DNA for case 527



Tumour FFPE DNA for case 1106