The University Of Sheffield.

Identifying functionally important aberrations of Neurofibromin-1 in non-small cell lung cancer

By:

Greg Wells

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Department of Oncology & Metabolism

Submission Date

October 2018

Acknowledgements

I would like to give my upmost respect and admiration to all the patients who took part in this study. Without whom, this project would have not been possible.

I would like to thank all the support workers, nurses, and clinicians involved in lung cancer clinics at Weston Park Hospital who advised me with the clinical aspects involved in patient recruitment. To all the staff within the Cancer Clinical Trial Centre who helped me access the ReSoLuCENT samples and taught me how to manage the clinical side of the project.

All the staff at Sheffield Diagnostic Genetic Services and Histology departments, at the Sheffield Children's Hospital Trust, for enabling me to carry out the research for this project in their diagnostic laboratories. Prof Simon Cross, Dr Paul Heath, and Dr Dave Hammond from the University of Sheffield, who all passed on valuable knowledge and experience to keep the project moving forward.

My supervisors: Dr Gill Wilson, Prof Sarah Danson, and Prof Ann Dalton, for giving me this opportunity and supporting me through every part of the PhD. I got really lucky with you three.

I would like to thank the BBSRC and AstraZeneca for funding this research and giving me the opportunity to complete part of the project in industry.

Finally, I would like to thank my whole family, for never giving up on me. My wife Allie, for recognising the spark that got me here and mi Nan and Gramps, who missed all this.

Abstract

Lung cancer is the most commonly diagnosed cancer worldwide. 87% of cases are classified as non-small cell lung cancer (NSCLC). Neurofibromin-1 (*NF1)* is a tumour suppressor gene which limits RAS mediated signal transduction within the MAPK pathway. Up-regulation of this pathway leads to uncontrolled cellular proliferation. Somatic mutations in *NF1* are reported in 8-12% of NSCLC patients. It was hypothesised that *NF1* variants have a functional consequence on the activation of the MAPK pathway.

86 NSCLC patients were recruited and consent obtained to access their archived formalin fixed paraffin embedded tumour tissue. Next generation sequencing (NGS) was used to screen the patients for *NF1* variants. Copy number changes of *NF1* were investigated using digital droplet PCR. Finally, the NanoString nCounter was utilised to measure mRNA gene expression signatures which relate directly to MAPK and PI3K/AKT/mTOR activation in order to determine whether the *NF1* variants observed were functionally relevant.

This study confirmed the prevalence of *NF1* variants in NSCLC previously reported. Of the 25 *NF1* variants identified 15 were novel with no previous reports in the literature or in databases. Analysis via the NanoString nCounter determined *NF1* mRNA transcript 2 to be the predominantly expressed transcript in NSCLC. In accordance with previous reports it was shown that the MEK gene expression signatures was able to predict *KRAS* variants and therefore MAPK activation. Expanding on this using data from Pan-Lung TGCA the MEK signatures were shown to also predict *EGFR* and BRAF driver variants. Finally, we demonstrated that *NF1* variants with loss of function of the GAP domain causes upregulation of the PI3K/AKT/mTOR pathway, while cases with co–occurring *NF1* and *RASA1* variants result in up regulation of the MAPK pathway. We also identified a third of NF1 loss of function cases which had an increased MEK signature score, which suggests unknown co-drivers are required in addition to *NF1* for significant MAPK activation.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADC	Lung Adenocarcinoma
AF	Allelic fraction
ALK	Anaplastic lymphoma kinase
ATP	Adenosine triphosphate
AUC	Area under the curve
bp	Base pair
BRAF	v-Raf murine sarcoma viral oncogene homolog B
BWA	Burrows-Wheeler Alignment tool
cAMP	Cyclic adenosine mono phosphate
CDKN	Cyclin-dependent kinase inhibitor
CELF	CUG-BP and ETR-3 like factors
CI	Confidence interval
CNV	Copy number variation
CV	Coefficient of variation
dbSNP	NCBI's Single Nucleotide Polymorphism database
ddPCR	Droplet digital PCR
DDR2	Discoidin domain-containing receptor 2
DIN	DNA integrity number
dsDNA	Double stranded DNA
DUSP	Dual specificity phosphatases
EGFR	Epidermal growth factor receptor
EML4	Echinoderm microtubule-associated protein like 4
ERK	Extracellular-signal-regulated kinases
ETV	*ETS* variant gene
EZ1	Qiagen EZ1 DNA Tissue Kit
FFPET	Formalin fixed paraffin embedded tissue
FGFR1	Fibroblast growth factor receptor 1
FISH	Florescence in-situ hybridisation
FTH1	Ferritin heavy polypeptide 1
GAP	Guanosine-triphosphatase activating proteins
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
gDNA	Genomic DNA
GDP	Guanine diphosphate
GIAB	Genome in a bottle (NA12878)
gnomAD	Genome Aggregation Database
GRCh37/hg19	Human reference genome 37
GTP	Guanine triphosphate
GTPase	Guanosine-triphosphatase
H+E	Haematoxylin and eosin
HRAS	Harvey RAS
IHC	Immunohistochemistry
indels	Insertions and deletions
KRAS	Kirsten rat sarcoma viral oncogene homolog
LIMP2	LIM domain kinase 2
LoH	Loss of heterozygosity

MAPK	Mitogen-activated protein kinase
MEK	Mitogen-activated protein kinase kinase
NCBI	National Center for Biotechnology Information
NF1	Neurofibromin-1
NF1P	*NF1* pseudogenes
NGS	Next generation sequencing
NICE	National Institute of Health and Care Excellence
NOS	Not otherwise specified
NRAS	Neuroblastoma RAS
NSCLC	Non-small cell lung cancer
NTC	No template controls
PHLDA1	Pleckstrin homology like domain family A member 1
PI3K	Phosphoinositide 3-kinase
QFI	Quantitative functional index
QIAamp	Qiagen's QIAamp DNA FFPE Tissue Kit
ReSoLuCENT	Resource for the Study of Lung Cancer Epidemiology in North Trent
RIN	RNA integrity number
ROC	Receiver operating characteristic
ROCK	Rho-associated protein kinase
RPP30	Ribonuclease P/MRP subunit P30
S100A6	S100 calcium binding protein A6
SCH	Sheffield Children's Hospital
SD	Standard deviation
SEC14-PH	Sec14 and pleckstrin homology
SERPINB1	Serpin family B member 1
SNP	Single nucleotide polymorphisms
SNV	Single nucleotide variations
SOX2	Sry-related HGM-box gene 2
SPRY2	Sprout like homolog 2
SQCC	Squamous cell carcinoma
STH	Sheffield Teaching Hospitals
SUZ12	Polycomb repressive complex 2 subunit
TCGA	The Cancer Genome Atlas
TK	Tyrosine kinase
TKI	Tyrosine kinase inhibitors
TP53	Tumour protein 53
truXTRAC	Covaris truXTRAC FFPE DNA kit
TSG	Tumour suppressor genes
VCF	Variant call format
WES	Whole exome sequencing
WPH	Weston Park Hospital
YHREC	Yorkshire and Humber Regional Ethics Committee
ZFP106	Zinc finger protein 106

Chapter 1
Introduction

1.1 Non-Small Cell Lung Cancer

1.1.1 Incidence

In 2018 there were an estimated 1.8 million new cases of lung cancer, making it the most commonly diagnosed cancer worldwide (Bray et al., 2018). In the UK alone 46,388 new cases were reported in 2015, of which 87% were classified as non-small cell lung cancer (NSCLC) (CRUK, 2018). Lung cancer has a poor prognosis, with a 5 year survival rate of less than 10%. This is largely due to late stage diagnosis, with 68% of cases being diagnosed at stage III and IV (CRUK, 2018). At this late stage, curative measures such as surgical intervention and radical radiotherapy are not possible. Patients have a median age of 70 years. In most cases the lung cancer is associated with cigarette smoking and these patients may have co-morbidities related to smoking, such as cardiovascular and cerebrovascular disease (Rojewski et al., 2016). These factors may make delivery of any treatments more difficult. Whilst the number of treatments has increased over the last decade with the emergence of targeted therapeutics, such as tyrosine kinase inhibitors and immunotherapies, the prognosis statistics still remain poor. Such facts illustrate that there is an unmet need to improve early diagnostic techniques and further identify the molecular mechanisms that underlie NSCLC, in order to identify better therapeutic and clinical management options.

1.1.2 Genetic Traits

Genomic anomalies are the defining trait of many cancers. Genomic mutations in key loci give a cell a new modus operandi, redefining the cell's purpose and function. Such genomic aberrations are often the hallmarks of cancer (Hanahan and Weinberg, 2011). These aberrations can vary from a single nucleotide variant to gains and losses of whole chromosomes. The genetic aberrations we observe, and the genes affected in carcinogenesis can be broadly characterised into terms such as driver or passenger mutations and oncogenes or tumour suppressor genes (TSG). Driver mutations give the cell a distinct growth advantage over surrounding cells and tissue and are positively selected during clonal expansion. Passenger mutations do not have any functional relevance, but can be passed down the tumour's evolutionary lineage (Michael et al., 2009). Oncogenes are affected by mutations that result in an increase or constitutive

activation of intracellular signalling pathways, enabling unregulated proliferation. Conversely TSG generally protect against growth signals and provide oncogenic protection, such that inactivation can lead to loss of regulation of critical pathways.

1.1.3 Risk Factors

The main risk factor in the acquisition of somatic mutations leading to the development of lung cancer is cigarette smoking. This is reported to account for 85% of cases in the UK (Parkin, 2011). Differential mutational distribution is observed between current/former smokers and never smokers. Current/former smokers have a higher chance of Kirsten rat sarcoma viral oncogene homolog (*KRAS)* mutations relative to never smokers (41% vs. 5%). Conversely, never smokers have a higher chance of epidermal growth factor receptor (*EGFR*) mutations relative to current/former smokers (38% vs. 14%) (Paik et al., 2012).

Occupational carcinogenic risk factors include; overexposure to arsenic, asbestos, cadmium and related compounds, and diesel engine exhaust (Brown et al., 2012). Inherited germline mutations are less common in NSCLC, but aberrations in the TSG retinoblastoma and tumour protein 53 (*TP53*) have been linked to an increase in tumour susceptibility (Hwang et al., 2003, Kleinerman et al., 2000). Epidemiology is also an influencing factor, with *EGFR* activating mutations being more prevalent in the female gender and those of East Asian ethnicity (Fukuoka et al., 2003, Miller et al., 2004).

1.1.4 Histological Subtypes of NSCLC

Histologically NSCLC has been traditionally divided into three subtypes; large cell lung cancer, squamous cell carcinoma (SQCC), and adenocarcinoma (ADC), with the latter two being the most prevalent (Pao and Chmielecki, 2010). Over the last two decades, molecular-based analysis has started to identify key genetic changes observed across and within these subtypes. This has led to cancer being further characterised based on common genetic variants. Such variants can help inform prognosis and predict sensitivity to targeted therapeutics.

1.1.5 Genetic Landscape

Genetic instability is a key hallmark of cancer. Next generation sequencing (NGS) has enabled the whole exome sequencing (WES) of ADC and SQCC. This has given us an *in situ* snapshot of the many genomic characteristics that define these histological subtypes, shown in Table 1.1. Large cell lung cancer still remains to be characterised in this way, but published WES data exists for over 660 ADC cases and 484 SQCC cases (Campbell et al., 2016). This data has exposed the differences and similarities in the aberrant genomic landscapes. ADC and SQCC contain high mutation burdens, 8.87 per Mb and 8.1 per Mb respectively. Even with knowledge of the genomic aberrations it is still challenging to subdivide these into driver or passenger events, with many cases as yet having no recognised driver variants.

1.1.6 NSCLC Candidate Driver Genes and Tumour Suppressor Genes

ADC has been shown to be driven by somatic mutations in the mitogen-activated protein kinase (MAPK) signalling pathway. Collisson *et al*. (2014) highlighted known driver mutations in 62% of tumours analysed. This included mutations in the following; *KRAS* (32%), *EGFR* (11%), and v-Raf murine sarcoma viral oncogene homolog B (*BRAF)* (7%), consistent with previous ADC WES studies (Imielinski et al., 2012, Ding et al., 2008). Mutual exclusivity between *EGFR* and *KRAS* drivers was observed in all ADC WES studies. Conversely in SQCC, WES data highlighted fewer abnormalities; *KRAS* (<1%), *EGFR* (<2%), and *BRAF* (<2%) in these driver mutations (Hammerman et al., 2012).

One recurrent characteristic seen in SQCC was a copy number increase of a region of chromosome 3 (3q25-26). Copy numbers increases are commonly <5, but have been reported up to 21 copies (Yamamoto et al., 2008). This is reflected in the amplification levels of sry-related HGM-box gene 2 (*SOX2*) and phosphoinositide 3-kinase (*PI3K*) catalytic subunit *PIK3CA,* both found within this locus. *PIK3CA* amplification was observed in 38% of SQCC cases, but rarely in ADC (Yamamoto et al., 2008). Non-synonymous *PIK3CA* driver substitutions p.Glu454Lys and p.His1047Arg are observed in both ADC and SQCC, but make up less than 2% and 4% of cases respectively (Yamamoto et al., 2008, Scheffler et al., 2015, Collisson et al., 2014, Ding et al., 2008, Imielinski et al., 2012, Hammerman et al., 2012). Fibroblast growth factor receptor 1 (*FGFR1*) amplification is another commonly observed

event in SQCC. *FGFR1* is a candidate driver of SQCC and observed in 17-22% of cases (Hammerman et al., 2012, Moch et al., 2010). Discoidin domain-containing receptor 2 (*DDR2*) is also a rare but a reported candidate driver of SQCC, with non-synonymous substitutions observed in 1-4% of cases (Hammerman et al., 2011, Hammerman et al., 2012).

Despite their differences in candidate oncogenic drivers, ADC and SQCC do have parallels. Both are subject to homozygous deletion of the cyclin-dependent kinase inhibitor 2A (*CDKN2A*) locus. However, SQCC is also prone to epigenetic silencing of *CDKN2A* via methylation (Hammerman et al., 2012, Collisson et al., 2014). Both subtypes demonstrate a high frequency of mutations in the TSG *TP53*. This is a key characteristic in many cancers and has been extensively studied. TP53 functions include; cell cycle inhibition, inducing senescence, inducing apoptosis, and activation of DNA repair mechanisms in response to DNA damage and hyper-proliferative signalling from RAS activation (Kathryn et al., 2014). Another frequently mutated TSG in both subtypes, which has been less well examined, but again is observed in many different types of cancers is neurofibromin-1 (*NF1*) (Hammerman et al., 2012, Collisson et al., 2014, Imielinski et al., 2012).

Table 1.1. Frequency of somatic variants and copy number variation in lung adenocarcinoma and squamous cell carcinoma.

Gene	Chromosome Location	Lung Adenocarcinoma		Squamous Cell Carcinoma	
		Somatic variants (%)	Copy number variation (%)	Somatic variants (%)	Copy number variation (%)
KRAS	12p12.1	33	6	1	2
EGFR	7p11.2	14	7	4	7
BRAF	7q34	10	3	4	1 1
RAB25	1q22	--	13	--	2
NF1	17q11.2	12	1	11	1
ALK:EML4	t(2p23.1)(2p21)	1	--	--	--
PIK3CA	3q26.32	7	2	16	38
PTEN	10q23.31	1	2	8	3
MTOR	1p36.22	5	1 1	7	--
RICTOR	5p13.1	5	10	2	8
TP53	17p13.1	47	1	94	--
FGFR1	8p11.23	1	3 2	2	17
DDR2	1q23.3	3	11	1	2
SOX2	3q26.33	<1	4	<1	43
CDKN2A	9p21.3	4	20	17	27
STK11	19p13.3	18	1	2	--

Genetic based division of the two most common histological subtypes of NSCLC. Blue (loss of copy number) Red (Increase in copy number). Data set taken from (Hammerman et al., 2012, Collisson et al., 2014).

1.2 Neurofibromin-1

1.2.1 *NF1* Prevalence in cancer

The prevalence of *NF1* variants in multiple cancers is now becoming widely acknowledged. This is due to the increase of genomic profiling studies utilising NGS, WES, and large scale collaborations such as the Cancer Genome Atlas (TCGA). *NF1* is described as one of the many genes which functions in RASopathy and thereby regulates the RAS family of proteins and subsequently MAPK signalling (Ratner and Miller, 2015). Anderson *et al*. (1993) first described a *NF1* homozygote deletion in a melanoma cell line, and proposed its role as a tumour suppressor gene which functions in progression and development of malignant melanoma (Andersen et al., 1993b). Melanoma has a high overall mutation burden of 14.4 per Mb and *NF1* variants are observed in 13% of all cases (Hodis et al., 2012). Reports in epithelial ovarian and breast cancers have shown *NF1* variants to be observed in 14.5% and 5% of all cases respectively (Perou, 2012, Ross et al., 2013). This is not an extensive list and most cancers which have been subjected to genomic profiling have been shown to carry *NF1* variants at various degrees of prevalence (Gao et al., 2013, Cerami et al., 2012).

Several studies have also investigated mRNA editing of *NF1* in cancer. Marrero and colleagues demonstrated *NF1* mRNA transcript 1 to be predominant in breast cancer, with surrounding normal tissue predominantly expressing *NF1* transcript 2 (Marrero et al., 2012). Marrero proposed the shift from *NF1* transcript 2 to transcript 1 could be a significant event in the development of sporadic breast cancers. Similar results have also been observed in ovarian cancer, with cell lines expressing an 11 fold increase in *NF1* transcript 1 compared to transcript 2 (Iyengar et al., 1999). Iyengar and colleagues suggested the expression of transcript 2 was related to a decreased cellular growth rate of ovarian cancer cell lines.

1.2.2 *NF1* Prevalence in NSCLC

NSCLC genomic profiling studies show 8-12% of patients display somatic *NF1* mutations. Over half of these mutations result in a reading frame shift, a non-synonymous non-sense variant, or a splice site mutation which results in protein inactivation. The remainder have non-synonymous mis-sense mutations. As with *TP53* there is a clear lack of *NF1* mutation hotspots observed in NSCLC cases, shown in Figure 1.1 (Campbell et al., 2016). It is also important to note that *NF1* and *KRAS* mutations in ADC display a significant tendency of mutual exclusivity (p = 0.006), whilst *NF1* and *EGFR* display a tendency of mutual exclusivity, it is not deemed statically significant (p = 0.084) (Imielinski et al., 2012).

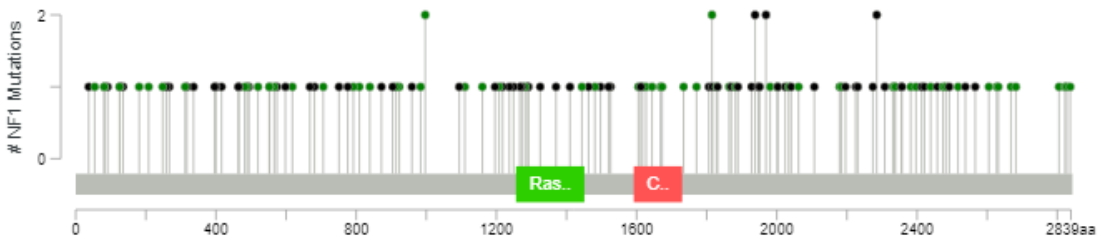


Figure 1.1. NF1 somatic mutations observed in lung adenocarcinoma and squamous cell cancer patients. Green pins (mis-sense mutation), black pins (truncating mutations). *NF1* variants reported in Pan-Lung TCGA (Campbell et al., 2016). Figure generated via Cbioportal.org (Cerami et al., 2012, Gao et al., 2013).

1.2.3 Evolutionary Occurrence and Risk Factors of *NF1* Mutations in NSCLC

Data from Collisson *et al.* (2014) and Imielinski *et al*. (2012) demonstrated that somatic *NF1* mutations occur early in the evolution of ADC. Both studies showed a bias towards patients with a stage I or II diagnosis, 76% and 69% respectively. In addition 85% of the reported *NF1* mutations from these studies were seen in these early stages of the disease. Data from these two studies also revealed 89% of *NF1* mutations were observed in patients with a smoking history, the remainder (11%) had unknown smoking status, but no non-smokers were reported. Another striking correlation in ADC is that *NF1* mutations are found in 25% of oncogene-negative tumours, compared to just 2% of oncogene-positive (Collisson et al., 2014).

1.2.4 Neurofibromatosis

NF1 is most commonly known for its role in the autosomal dominant disease Neurofibromatosis type 1 (MIM: 162200), otherwise known as von Recklinghausen's disease. The disease is reported to affect an average of 1 in 3000 individuals (Friedman, 1999). Neurofibromatosis is commonly diagnosed in early childhood based on clinical manifestations. These include; café-au-lait spots (flat light brown spots on the skin), lisch nodules (small lumps on the iris of the eye), neurofibromas (soft bumps under the skin), plexiform neurofibroma, axillary or inguinal freckling, characteristic skeletal abnormalities (pseudarthrosis, hypoplasia of sphenoid wing, severe kyphoscoliosis), and optic glioma (tumour of the optic nerve) (Tonsgard, 2006).

The disease is the result of mutations in the *NF1* gene leading to functional inactivation of the protein. These mutations leave the individual with increased susceptibility to the development of benign and malignant tumours of the peripheral nerve sheaf. *NF1* mutations in these cases, as with NSCLC lack any significant hotspots and are distributed throughout the coding region (Mattocks et al., 2004, Brinckmann et al., 2007). Whilst germline mutations are passed onto offspring, they have no direct correlation with the clinical phenotype. Family members with identical mutations can demonstrate very different clinical characteristics (Easton et al., 1993). This suggests the phenotypic characteristics are owing to further pathogenic mechanisms, which remain to be fully elucidated. Loss of function of one *NF1* allele displays haplosufficient characteristics, but the disease still has 100% penetrance (Riccardi and Lewis, 1988). Loss of heterozygosity (LoH) is therefore required for disease penetrance, following Knudsen's two-hit hypothesis paradigm (Rasmussen et al., 2000). LoH is suspected to originate in Schwann cells, which then increase paracrine growth signals, resulting in haploinsufficiency of the surrounding cell types, in particular mast cells (Yang et al., 2008). However, genetic aberrations in addition to bi-allelic inactivation of *NF1* are thought to be involved in the transition of tumours from benign to malignant, including; polycomb repressive complex 2 subunit (*SUZ12)*, *TP53*, and *CDKN2A* (Sohier et al., 2017).

1.2.5 *NF1* Structure

The gene encoding NF1 and its cDNA sequence was first identified in 1990 (Cawthon et al., 1990). *NF1* (17q11.2) spans approximately 350 kb and is comprised of 58 exons which yield a 12 kb mRNA transcript (Cawthon et al., 1990, Apolline et al., 2015). The transcript is subject to mRNA editing resulting in multiple protein isoforms, the three most common being; isoforms I-III (Cawthon et al., 1991, Suzuki et al., 1992, Andersen et al., 1993a). *NF1* isoform I (transcript variant 2, NM_000267.3) is the result of the exclusion of exon 31 (commonly named exon 23a using legacy nomenclature) and results in a 2818 amino acid protein. Isoform II (transcript variant 1, NM_001042492.2) is the result of the insertion of exon 31, thus resulting in a 2839 amino acid protein. The splicing of *NF1* pre-mRNA is orchestrated by three different families of RNA binding proteins. This includes CUG-BP and ETR-3 like factors (CELF) and Hu proteins, both of which promote the exclusion of exon 31 (Barron et al., 2010). TIA-1/TIAR proteins promote the inclusion of exon 31 by competitively binding to the same downstream sequence as the Hu proteins (Barron and Lou, 2012). Finally, isoform III (transcript variant 3, NM_001128147.2) is truncated and terminated at exon 15, resulting in a 593 amino acid protein (Mukhopadhyay et al., 2002). Intriguingly, *NF1's* largest intron 36 spans >60kb and has three genes encoded in the antisense strand. These genes are oligodendrocyte-myelin glycoprotein and two ecotropic viral integration sites which are transcribed in the opposite direction to *NF1* (Cawthon et al., 1991, Viskochil et al., 1991).

The resulting NF1 is a large multiple domain protein which is expressed ubiquitously (Philpott *et al* 2017). Although NF1 is expressed ubiquitously the different isoforms are largely observed in different tissues. Isoform I is more commonly expressed in neurons of the central nervous system, whilst isoform II is found in most other tissues (Hinman et al., 2014). The exclusion of exon 31, which is located in the guanosine-triphosphatase (GTPase) activating proteins (GAP) domain, has been shown to have 10 fold increase in RAS regulatory activity compared to isoform I (Hinman et al., 2014, Suzuki et al., 1992). Isoform III is largely observed in cases of neurofibromatosis and lacks any GAP related domain (Mukhopadhyay et al., 2002).

The encoded NF1 protein has an approximate size of 250kDa and contains two well characterised domains. The most extensively studied GAP domain shares highly conserved sequences with GTPase activating proteins such as p129 GAP (Welti et al., 2011). The GAP domain makes up 8% of NF1 and spans amino acid 1235-1452, which translates to exons 27-33 based on *NF1* mRNA transcript 1, shown in Figure 1.2. The second is the Sec14 and pleckstrin homology (SEC14-PH) domain. The SEC14-PH domain spans amino acids 1580 - 1837 which translates to exons 36 to 38 based on *NF1* mRNA transcript 1 (www.uniprot.org/uniprot/P21359). The RAS-GAP and the SEC14-PH are the only regions of NF1 to have their crystal structures resolved via xray diffraction (D'Angelo et al., 2006, Ahmadian et al., 2003).

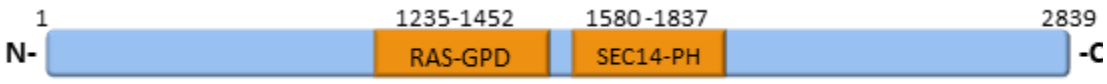


Figure 1.2. Illustration of NF1 transcript variant 1 (NM_001042492.2).

1.3 *NF1* Functions as a Tumour Suppressor Gene

1.3.1 RAS Independent Mechanisms

NF1 functions as a TSG in RAS-dependent and RAS-independent manners. RAS-independent mechanisms include NF1 stimulation of adenylyl cyclase activity, leading to an increase of cyclic adenosine mono phosphate (cAMP) (Brown et al., 2010, Anastasaki and Gutmann, 2014). Brown and colleagues demonstrated that retinal ganglion neurons with *NF1* heterozygous loss of function suffered from decreased growth and increased apoptosis. This was restored by increasing cAMP levels, but not by inhibiting the downstream signalling pathways activated by RAS, including PI3K/AKT/mTOR and the MAPK pathway (Brown et al., 2010).

Further RAS-independent functions of NF1 have been related to the currently undefined region spanning amino acids 1-1163. Starinsky-Elbaz and colleagues demonstrated this by transfecting the giloblastoma cell line U87, which does not express NF1, and mouse embryonic fibroblasts which express very low levels of NF1 with the region of NF1 spanning the amino acids 1-1163 (Starinsky-Elbaz et al., 2009). It was shown that NF1 1-1163 inhibited cell migration of both cell lines. The authors went on to confirmed this by measuring an increase in cell adhesion genes (Starinsky-Elbaz et al., 2009). Starinsky-Elbaz and colleagues proposed that the *NF1* 1-1163 domain acts as a negative regulator of the Rac1-ROCK1/PAK1-LIMK1-ADF/cofilin pathway, via an unknown mechanism (Starinsky-Elbaz et al., 2009).

NF1 has also been identified as a negative regulator of the Rho/ROCK/LIMK2/cofilin signalling pathway which functions in cytoskeleton dynamics. Ozawa and colleagues demonstrated NF1 depletion increased activation of the Rho-ROCK-LIMK2 pathway. This resulted in an invasive phenotype in both HeLa and HT1080 cells (Ozawa et al., 2005). More recent data has shed further light on this process; Vallée and colleagues observed protein-protein interactions using a yeast two-hybrid system between the SEC14-PH domain of NF1 and LIM domain kinase 2 (LIMK2). The authors used a HEK-293 transfected cell line and corroborated these findings demonstrating that the SEC14-PH domain interacts with the KIN and PDZ/SP domains of LIMK2 (Vallée et al., 2012). The authors proposed the SEC14-PH domain inhibits the Rho/ROCK/LIMK2/cofilin pathway through inhibition of

LIMK2 phosphorylation by Rho-associated protein kinase (ROCK). This demonstrates NF1 can influence cytoskeleton remodelling in a RAS independent manner.

1.3.2 RAS Dependent Mechanism

The RAS-dependent pathway is the most documented NF1 mechanism and is enabled through the *NF1* GAP domain (Ahmadian et al., 2003). The RAS family of signalling proteins includes three cell membrane bound GTPases homologs; KRAS, Harvey RAS (HRAS), and neuroblastoma RAS (NRAS) (Castellano and Downward, 2011). The *NF1* GAP domain exerts its tumour suppressor ability over RAS via the hydrolysis of RAS-guanine triphosphate (GTP) to its inactive RAS-guanine diphosphate (GDP) state (Xu et al., 1990). This inactivation is the result of a conformational change of the switch I and II regions of RAS, facilitated by the hydrolysis process of removing the phosphate group (Ahmadian et al., 2003). Driver mutations seen in the RAS family diminish its GTPase activity, no longer allowing GAP proteins such as *NF1* to catalyse the hydrolysis of GTP to GDP (Ahmadian et al., 1999). This locks the protein in its active conformational state, leaving it free to activate downstream effectors.

1.4 MAPK Signalling Pathway

1.4.1 Ligand dependent activation of the MAPK pathway

The RAS family have numerous upstream activators; one of the most comprehensively studied in relation to NSCLC is EGFR. EGFR (also known as HER1 and ErbB) is a transmembrane tyrosine kinase (TK) signalling protein belonging to the HER family (Cadranel et al., 2013). Other members of this family include; HER2, HER3, and HER4. The extracellular domain of EGFR, HER3, and HER4 share an affinity for a number of extracellular ligands (Hartman et al., 2013). During ligand dependent activation, EGFR extracellular domains undergo a conformational change that allows a dimerisation with EGFR, HER2, HER3, or HER4 (Dawson et al., 2005). Dimer formation results in the autophosphorylation of the intracellular TK domain facilitated by adenosine triphosphate (ATP) (Hartman et al., 2013). In its phosphorylated state, the EGFR dimer forms a complex with downstream signalling proteins including, growth factor receptor-bound protein 2 (GRB2) and the guanine exchange factor son of sevenless homolog 1 (SOS1) (Zarich et al., 2006). SOS1 then catalyses RAS-GDP to its RAS-GTP active state, shown in Figure 1.3.

Well studied effectors of the RAS family include RAF and PI3K (Gupta et al., 2007, Vojtek et al., 1993). Activation of these signalling proteins results in a downstream signalling cascade of the MAPK and PI3K/AKT/mTOR signalling pathways. The activation of these pathways if not negatively regulated downstream can ultimately lead to cell proliferation, cell migration, and cell survival (Siegelin and Borczuk, 2014). *In vitro* models demonstrate loss of *NF1* predominantly drive tumourigenisis through activation of the MAPK pathway (de Bruin et al., 2014a).

1.4.2 The MAPK Cascade

The MAPK cascade is a three tier protein signalling network which transfers messages from the membrane bound cytoplasmic RAS family across the cytoplasm to within the nucleus. The RAF family is the first tier of the cascade and has three isoforms Raf-1/c-Raf, B-Raf, and A-Raf, all encoded by different genes. All share three conserved regions (CR1-3). CR1 is located in the N-terminal and includes the RAS binding domain (Matallanas et al., 2011) CR3 is found at the C-terminal and is home to the mitogen-activated protein kinase Kinase 1/2 (MEK1/2) serine/threonine kinase activation sites. All RAF proteins have a narrow range of substrates including MEK1/2, the second tier of the MAPK cascade. MEK1/2 are dual specificity tyrosine/threonine protein kinases, as with the RAF family MEK1/2 has a narrow range of substrates which includes extracellular-signal-regulated kinases (ERK)1/2, the third tier of the MAPK pathway (Roskoski, 2012). Unlike BRAF and MEK1/2, ERK has a wide range of substrates and is able to translocate from the cytoplasm to the nucleus. The subsequent phosphorylation of the threonine and tyrosine residues of ERK1/2 ultimately leads to activation of pro-survival transcription factors such as ELK-1 and CREB and deactivation of apoptotic transcription factor FOXO3 (Mebratu and Tesfaigzi, 2009). ERK1/2 activation also initiates the transcription of target genes involved in negative feedback of the MAPK pathway including the dual specificity phosphatases (DUSP) 4 and 6 (Dry et al., 2010b).

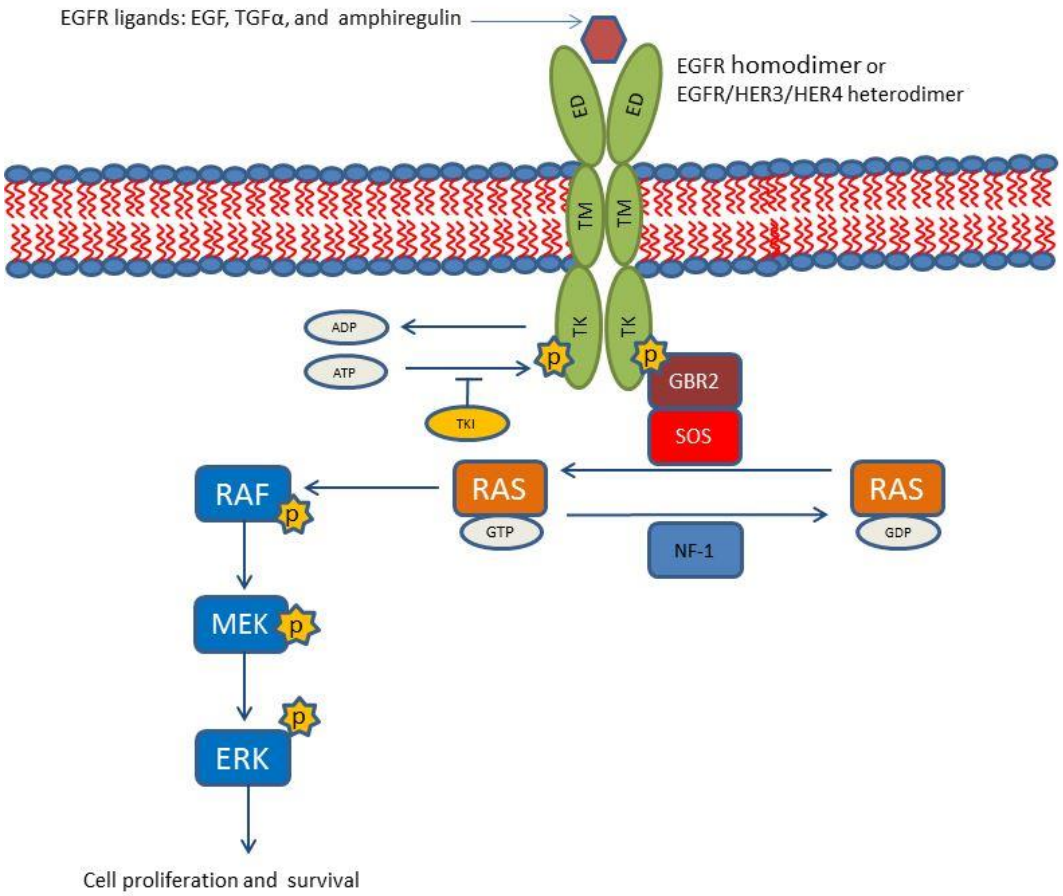


Figure 1.3. Ligand dependent activation of the MAPK pathway. Ligand binding of EGFR extracellular domains (ED) causes dimerisation with EGFR, HER2, HER3, or HER4. EGFR ligands include epidermal growth factor (EGF), transforming growth factor α (TGF α), and amphiregulin. Autophosphorylation is a result of ATP binding to the tyrosine kinase domain (TK) and donating a phosphate group. This is the same domain that TKI bind to, thereby inhibiting phosphorylation. The EGFR dimer illustrated forming a complex with GBR2 and SOS1, which converts RAS into its active GTP bound state. RAS-GTP then starts a signalling cascade through the MAPK pathway (p represents the signalling protein in its active phosphorylated state). *NF1* is a negative regulator of RAS and catalyses the reaction of RAS-GTP to its inactive RAS-GDP bound form. Loss of *NF1* will result in increased activation of the MAPK pathway. Transmembrane domain (TM).

1.5 *NF1* Pseudogenes

Pseudogenes are abundant throughout the genome and originate from a gene or partial gene duplication that has taken place throughout evolution (Tutar, 2012). It is calculated there are 14112 pseudogenes scattered about the human genome (Pei et al., 2012). Pseudogenes, as their name suggests demonstrate a high level of homology with the gene from which they originated. Luijten and colleagues reported 7 *NF1* pseudogenes (*NF1P*) present on multiple chromosomes including; 2, 12, 14, 15, 18, 21, and 22 (Luijten et al., 2000, Luijten et al., 2001). *NF1P1* is shown in Figure 1.4. Their investigation suggested the *NF1P* on chromosome 2 was the first to arise and all *NF1P* to be located in the pericentromeric regions of the chromosomes.

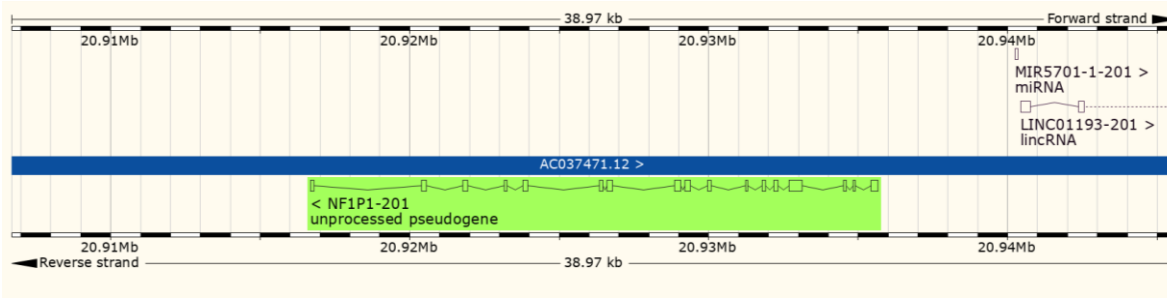


Figure 1.4. NF1 pseudogene 1. Ensembl genome browser illustration of NF1P1 located on chromosome 15.

1.6 Selective Therapeutics for NSCLC

1.6.1 ADC Targeted Therapeutics

Knowledge of the genetic characteristics of NSCLC has led to a paradigm shift in the treatment of patients displaying key genetic drivers. This has enabled the use of selective therapeutics that target key activating driver mutations. However, there are a greater number of identified driver mutations such as *KRAS* that do not currently offer any therapeutic advantage, and are linked to a poor prognosis.

In ADC three driver mutations are currently used as predictive biomarkers for the use of selective therapeutics. These include; *EGFR* in frame exon 19 deletions (19 del) and the Leu858Arg substitution in exon 21, both located in the TK domain (Lynch et al., 2004, Pao et al., 2004). The third is the anaplastic lymphoma kinase (ALK) and echinoderm microtubule-associated protein like 4 (EML4) fusion protein (Cadranel et al., 2013, Katayama et al., 2011).

Tyrosine kinase inhibitors (TKI) that target these oncogenic mutations have been developed and approved by the National Institute of Health and Care Excellence (NICE) (NICE, 2014). The TKIs Gefitinib, Erlotinib, Afatinib and Crizotinib have demonstrated significant increase in progression free survival compared to systemic chemotherapy for *EGFR* and *ALK-EML4* positive cases (Haaland et al., 2014, Maemondo et al., 2010, Rosell et al., 2012).

Clinical trials using *KRAS* as a predictive biomarker for the potential MEK inhibitor (selumetinib) and docetaxel in combination have been investigated in the SELECT-1 trial (NCT00890825, 2018). SELECT-1 initially yielded significant progression free survival compared to docetaxel mono therapy in phase two trials. Patients with metastatic or locally advanced NSCLC demonstrated a medium progression-free survival of 5.3 months (95% confidence interval (CI), 4.6-6.4) in the combination group, compared to and 2.1 months (95% CI 1.4-3.7) in the docetaxel group The difference between groups was 3.2 months with a progression hazard ratio (HR) of 0.58 (80% CI, 0.42-0.79), p = 0.014 (Janne et al., 2013). However, the phase three trial demonstrated no significant difference between the docetaxel vs. docetaxel and selumetinib arms (NCT01933932, 2018). The Medium progression free survival observed for the selumetinib in combination with

docetaxel arm was 3.9 months (interquartile range (IQR), 1.5-5.9) compared to the 2.8 months (IQR, 1.4-5.5) in the docetaxel arm. The difference observed was, 1.1 months with a HR of 0.93 (95% CI, 0.77-1.12), p = 0.44 (Janne et al., 2017).

1.6.2 SQCC Targeted Therapeutics

Unfortunately, SQCC currently has no approved targeted therapies. There have been preclinical and early phase trials investigating the potential of targeted therapeutics in SQCC. A phase two trial investigating the multi target kinase inhibitor Dasatinib in patients with *DDR2* and *BRAF* variants was terminated due to low accrual and lack of efficacy (NCT01514864, 2015). The Lung-MAP trial is currently investigating the use of the FGFR tyrosine kinase inhibitor AZD4547 as a second line therapy in patients with stage 4 SQCC with *FGFR1* amplification (NCT02965378, 2017).

1.6.3 Future NSCLC Targeted Therapeutics

Future advancements for potential targeted treatments for both ADC and SQCC are being investigated in the National Lung Matrix Trial (NCT02664935, 2018). The Matrix trial is a collaboration between Cancer Research UK, Experimental Cancer Medicine Centre Network, AstraZeneca, Pfizer, and Mirati Therapeutics Inc. The trial is currently recruiting for phase II and is investigating 10 potential therapies by stratifying NSCLC patients based on multiple pre-specified actionable biomarkers.

1.7 Emergence of *NF1* Loss as a Resistance Mechanisms to Targeted Inhibitors

Loss of *NF1* and the presence of *NF1* variants have also been linked to resistance in targeted therapies. Several studies have now demonstrated that loss of *NF1* is enough to cause resistance to targeted inhibitors in NSCLC and melanoma (Whittaker et al., 2013, Beauchamp et al., 2014, Maertens et al., 2013b, de Bruin et al., 2014a).

de Bruin and colleagues demonstrated that *NF1* knockdown caused resistance to Erlotinib in a sensitive *in vitro* ADC model by increasing active RAS-GTP; this potentially led to MAPK and PI3K/AKT/mTOR activation (de Bruin et al., 2014a). The authors then demonstrated that transfection of the *NF1*-GAP domain in the model re-established Erlotinib sensitivity by diminishing ERK phosphorylation but not AKT phosphorylation. Additionally, constitutively active AKT and MEK were transfected into ADC PC9 cells. The MEK transfected cells demonstrated an increase in resistance to Erlotinib, whereas AKT only moderately increased cell survival compared to controls (de Bruin et al., 2014a). To further confirm MAPK as the dominant driving pathway, a combination of Erlotinib and the MEK inhibitor Selumetinib resulted in complete loss of ERK phosphorylation in the *NF1* knockdown model (de Bruin et al., 2014a).

Similarly, in melanoma Whittaker and colleagues used PLX4720, a selective $BRAF^{V600F}$ inhibitor in a resistant cell line with *NF1* knockdown and demonstrated that this in combination with Selumetinib re-sensitised the cells. The authors also noted a direct correlation between ERK phosphorylation and cell proliferation. These findings are also supported by Maertens and colleagues using a melanoma mouse *BRAF/NF1* model (Maertens et al., 2013b). The authors demonstrated that PLX4720 alone did not reduce ERK phosphorylation. However, the MEK inhibitor PD0325901 abolished ERK phosphorylation and caused tumour regression *in vivo* (Maertens et al., 2013b). These models suggest that *NF1* exerts its tumour suppressor abilities in a RAS dependent manner, functioning primarily through inhibition of the MAPK pathway.

de Bruin and colleagues established that *NF1* loss was enough to confer EGFR TKI resistance. In order to address whether *NF1* loss occurs in response to exposure to EGFR TKI, the authors used an *in vivo* EGFR mouse model, subjected to prolonged exposure to

Erlotinib (de Bruin et al., 2014a). The authors reported reduced translation of *NF1* mRNA in 10 out of 18 Erlotinib resistant tumours. Notably the remaining 8 all showed other resistance mechanisms, including a *EGFR* Thr790Met gatekeeper mutations and hepatocyte Growth Factor Receptor (MET) amplification (de Bruin et al., 2014a). *NF1* mRNA expression was investigated in a cohort of 34 patients with unknown *EGFR* status, high *NF1* mRNA expression was associated with a statistically significant increased medium survival time, shown in Figure 1.5 (de Bruin et al., 2014a).

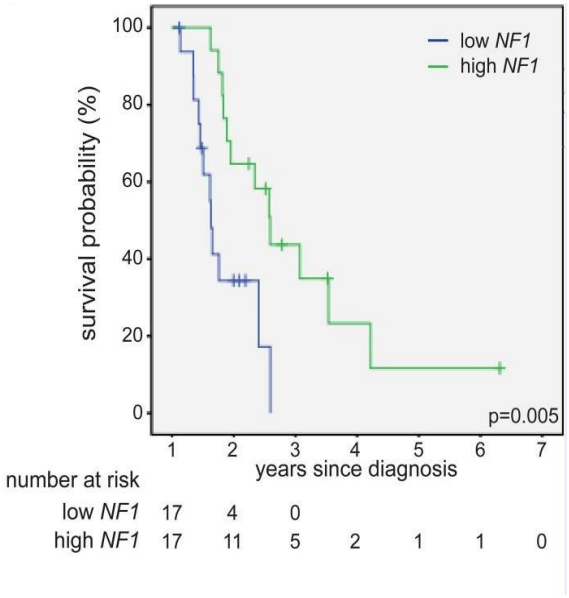


Figure 1.5. Kaplan–Meier plot post EGFR TKI therapy. 34 Patients divided into two groups based on *NF1* mRNA expression (*NF1* high and *NF1* low), median *NF1* mRNA expression cut off point was used to define the groups. *EGFR* status of the tumours is unknown. Figure adapted from de Bruin *et al* (2014a).

NF1 has been identified as a resistance mechanism to the DDR2 inhibitor Dasatinib in SQCC (Beauchamp et al., 2014). The authors used a Dasatinib sensitive SQCC cell line and subjected it to prolonged exposure to Dasatinib until resistance developed. The cell line exome was then sequenced, identifying an *NF1* splice site mutation 1392 G>A. This mutation was directly linked to reduced protein expression. *NF1* was knocked down in the parental cell line which confirmed protein loss does confer resistance to Dasatinib (Beauchamp et al., 2014). A comparison of phosphorylation of AKT, MEK, and ERK between resistant and parental cell lines revealed ERK phosphorylation was maintained in the

resistant cell line but not AKT and MEK (Beauchamp et al., 2014). Beauchamp and colleagues also investigated a Dasatinib sensitive alpha type platelet-derived growth factor receptor amplification dependent cell line. Again *NF1* knockdown resulted in increased Dasatinib resistance (Beauchamp et al., 2014).

Genomic profiling studies have revealed *NF1* somatic mutations are present in both ADC and SQCC in 8-12% of cases (Ding et al., 2008, Collisson et al., 2014, Hammerman et al., 2012, Campbell et al., 2016). In ADC *NF1* mutations are more prevalent in oncogene negative tumours and are mutually exclusive to *KRAS* drivers (Campbell et al., 2016). In vitro models of NSCLC and melanoma have both shown that loss of *NF1* drives the increase of growth signalling through the MAPK pathway (de Bruin et al., 2014a, Maertens et al., 2013a, Whittaker et al., 2013), whilst Beauchamp and colleagues demonstrated that a somatic mutation of *NF1* is enough to confer resistance to Dasatinib, through haploinsufficient mechanisms, in SQCC. However, the full consequences of the spectrum of mutations observed in NSCLC need to be addressed to determine their functional relevance in the clinical environment.

1.8 Hypothesis and Study Aims

It was hypothesised that NF1 variants have a functional consequence on the activation of the MAPK pathway. *NF1* variants are commonly observed in the NSCLC population and in-vitro models have demonstrated that NF1 loss is enough to drive the MAPK pathway. The current study sets out to determine the functional consequences that these *NF1* variants have on clinical cases.

To address this hypothesis we set out the following study aims.

1. Recruit 100 NSCLC patients via Weston Park Hospital (WPH) Cancer Clinical Trials Centre.
2. Screen the patients archive tumour tissue for genetic *NF1* variants and *NF1* copy number changes using a combination of next generation sequencing and digital droplet PCR.
3. Measure MAPK activation in the archive tumour tissue using mRNA gene expression signatures and relate this back to the *NF1* variants, known *MAPK* drivers, and cases with no known *MAPK* drivers.

Chapter 2
Materials and Methods

2.1 Materials

All materials were stored according to manufacturers' instructions unless otherwise stated

2.1.1 Nucleic Acid Extraction

2.1.1.1 Qiagen QIAamp FFPE Tissue Kit and Deparaffinization Solution

- Qiagen QIAamp FFPE Tissue Kit, Catalogue number: 56404
  - QIAamp MinElute® Columns and Collection Tubes
  - Tissue lysis buffer (Buffer ATL)
  - Lysis buffer (Buffer AL)
  - Wash buffers (Buffers AW1 and AW2 concentrate)
  - Elution Buffer (Buffer ATE)
  - Proteinase K
- Deparaffinization Solution Catalogue number: 19093

2.1.1.2 Covaris truXTRAC FFPE DNA kit

- Covaris truXTRAC FFPE DNA kit, Catalogue number: 520161
  - MicroTUBE Screw-Cap FFPE
  - Purification columns and collection tubes
  - Tissue lysis buffer (SDS Buffer)
  - Binding buffer (B1 Buffer)
  - Wash buffers (BW and B5 Buffer)
  - Elution Buffer (PE Buffer)
  - Proteinase K

2.1.1.3 Qiagen EZ1 DNA Tissue Kit and Protocol Card

- EZ1 DNA Tissue Card, Catalogue number: 9015588
- Qiagen EZ1 DNA Tissue Kit, Catalogue number: 953034
  - Reagent Cartridge
  - Disposable Tip Holders
  - Disposable Filter-Tips
  - Sample Tubes (2 ml)
  - Elution Tubes (1.5 ml)
  - Proteinase K
  - Buffer AVE

2.1.1.4 Chemagen (Chemagic) 360

- Chemagic DNA Blood kit, 2ml, Catalogue number: CMG-1097
  - Magnetic Beads
  - Lysis Buffer 1
  - Binding Buffer 2
  - Wash Buffer 3

- Wash Buffer 4
- Wash Buffer 5
- Wash Buffer 6
- Elution Buffer 7
- Protease

2.1.1.5 Qiagen Rneasy FFPE Tissue Kit

- RNeasy FFPE Kit, Catalogue number: 73504
  - RNeasy Mini spin column
  - Lysis buffer (Buffer RBC)
  - Digestion buffer (Buffer PKD)
  - Proteinase K
  - RNase-Free DNase I (lyophilized)
  - RNase-Free Water
  - Dnase Booster Buffer
  - Wash Buffer (Buffer RPE)
  - Rnase Free Water

2.1.2 DNA Quantification & Quality Control

2.1.2.1 ThermoFisher Scientific: Qubit assay kits

- Qubit dsDNA BR Assay Kit, Catalogue number: Q32850
  - Qubit® dsDNA BR Reagent
  - Qubit® dsDNA BR Buffer
  - Qubit® dsDNA BR Standard #1 (100 pg/µL)
  - Qubit® dsDNA BR Standard #2 (1000 ng/µL)

- Qubit dsDNA HS Assay Kit, Catalogue number: Q32851
  - Qubit® dsDNA HS Reagent
  - Qubit® dsDNA HS Buffer
  - Qubit® dsDNA HS Standard #1 (10 pg/µL)
  - Qubit® dsDNA HS Standard #2 (100 ng/µL)

- Qubit RNA HS Assay Kit, Catalogue number: Q33210
  - Qubit® RNA HS Reagent
  - Qubit® RNA HS Buffer
  - Qubit® RNA HS Standard #1 (250 pg/µL)
  - Qubit® RNA HS Standard #2 (100 ng/µL)

2.1.2.2 Quantitative PCR

- Qiagen, QuantiTect Probe PCR Kit, Catalogue number: 204343
  - QuantiTect Probe PCR Master Mix
  - RNase-Free Water
  - ThermoFisher, Hs01694011_s1 TaqMan® Gene Expression Assay, Catalogue number: 4331182, Ferratin heavy polypeptide chain (FTH1) probe and primers

2.1.2.3 Agilent TapeStation 2200

- All ScreenTape assays kits
  - D1000 ScreenTape Assay, Catalogue number: G2991-90030
  - High Sensitivity D1000 ScreenTape Assay, Catalogue number: G2991-90130
  - Genomic DNA ScreenTape Assay, Catalogue number: G5067-5365
  - RNA ScreenTape Assay, Catalogue number: G2964-90022
  - High Sensitivity RNA ScreenTape, Catalogue number: G2964-90121

2.1.3 Sanger Sequencing

- OneTaq® Quick-Load® 2X Master Mix with Standard Buffer, Catalogue number: M0486S, New England Biolabs
- Oligo primers, Catalogue number: VC00021, Sigma Aldridge
- BigDye™ Terminator v1.1 Cycle Sequencing Kit, Catalogue number: 4337449, ThermoFisher
- 96% Ethanol, Catalogue number: 20 823.327, VWR
- Better Buffer, Catalogue number: 3BB1, Microzone Ltd
- Illustra ExoProStar 1-Step, Catalogue number: 11941411
- Agencourt CleanSEQ, Catalogue number: APN000136, Beckman Coulter

2.1.4 NGS Library Preparation

General Reagents

- Agencourt AMPure XP Kit, Catalogue number: A63882, Beckman Coulter
- Dynabeads MyOne Streptavidin T1, Catalogue number: 65602, Thermo Fisher Scientific
- 96% Ethanol, Catalogue number: 20 823.327, VWR
- Nuclease-free Water (not DEPC-treated) Catalogue number: AM9930, Ambion

Agilent's SureSelect XT Low Input Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library

- SureSelect XT Low input Reagent Kit, index 1-96, Catalogue number: G9707A
  - SureSelect XT HS Index Primers A01 through H12

- SureSelect XT Low input Reagent Kit, index 97-192, Catalogue number: G9708A
  - SureSelect XT HS Index Primers A01 through H12

- Hereditary Cancer v3.2 Design ID 3014041
  - Custom biotinylated 120 nt RNA probe set

- Agilent XT Low input Library Preparation Kit for ILM (pre-PCR)
  - Ligation Buffer
  - T4 DNA Ligase
  - End Repair-A Tailing Buffer
  - End Repair-A Tailing Enzyme Mix
  - Herculase II Fusion DNA Polymerase
  - 5× Herculase II Reaction Buffer
  - 100 mM dNTP Mix
  - Forward Primer
  - Adaptor Oligo Mix

- Agilent XT Low input Target Enrichment Kit ILM Hyb Module box 1 (post-PCR)
  - SureSelect Binding Buffer
  - SureSelect Wash Buffer 1
  - SureSelect Wash Buffer 2
  - SureSelect XT HS and XT Low Input Blocker Mix

- Agilent XT Low input Target Enrichment Kit ILM Hyb Module box 2 (post-PCR)
  - SureSelect RNase Block
  - SureSelect Fast Hybridization Buffer

- Herculase II Fusion DNA Polymerase (red cap)
- 5× Herculase II Reaction Buffer
- 100 mM dNTP Mix
- SureSelect Post-Capture Primer MixNGS Illumina Paired End Sequencing

- HiSeq Rapid SBS Kit v2 (200 cycles), Catalogue number: FC-402-4021
- HiSeq PE Rapid Cluster Kit v2, Catalogue number: PE-402-4002
- PhiX Control v3, Catalogue number: FC-110-3001
- Qiagen Buffer EB, Catalogue number: 19086
- 2N NaOH
- TWEEN® 20 Catalogue number: P2287

2.1.5 Bio-Rad Digital Droplet PCR

- ddPCR *NF1* copy number assay, Catalogue number: 10031240
- ddPCR RPP30 copy number assay, Catalogue number: 10031243
- ddPCR Supermix for probes (No DUTP), Catalogue number: 1863023
- DG8 Cartridges, Catalogue number: 1864008
- DG8 Gaskets, Catalogue number: 1863009
- Droplet Generation Oil for Probes, Catalogue number: 1863005
- Oligo primers, Catalogue number: VC00021, Sigma Aldridge

2.1.6 NanoString nCounter

- NanoString custom design XT formulation CodeSet, design ID: Sel-Yale_2
  - Reporter CodeSet
  - Capture CodeSet
- nCounter Standard Master Kit, Catalogue number: NAA-AKIT-192
  - nCounter Cartridge
  - nCounter Prep Plates
  - nCounter Prep Pack, Catalogue number:
    - Racked tips and foil piercers
    - 12-tube strips
    - Tip Sheaths
    - 12-strip tube caps
    - Cartridge well seals
    - Hybridization buffer
- Agilent QPCR Human Reference Total RNA, Catalogue number: 750500

2.1.7 Patient Samples

86 patients were recruited directly in this study (NF1 in NSCLC) based on the eligibility criteria described in section 2.2.1.2. 40 samples were also accessed from the ReSoLuCENT trial which had the same acceptance criteria as this study, as described in section 2.2.1.2. 56/86 samples from this study (NF1 in NSCLC) passed initial pre-analytical acceptance criteria as described in section 4.2, all of which were surgical specimens. 34/40 samples from the ReSoLuCENT trial passed initial pre-analytical acceptance criteria, however a further 4 failed at initial library preparation stage. The remaining 30 ReSoLuCENT samples included in analysis were all biopsies, 17 from metastatic sites and 13 from primary sites. Full description of samples analysed is shown in Table 4.5.

2.2 Methods

2.2.1 Recruitment of Patients to the Neurofibromin-1 in Non-Small Cell Lung Cancer Study

2.2.1.1 Ethics Statement

The initial aim for the academic study 'Identifying functionally important aberrations in neurofibromin-1 in non-small cell lung cancer' was to recruit 100 patients with NSCLC from WPH for genetic analysis.

Local approval and peer review for the project was granted by Weston Park Hospital Clinical Trials Executive in February 2015. The project was approved by Sheffield Teaching Hospitals (STH) Research and Development department (STH18741) after reviewing the project protocol (Appendix 1), patient information sheet (Appendix 2), patient consent form (Appendix 3) and the Integrated Research Application System form. Final ethical approval came from Yorkshire and Humber Regional Ethics Committee (YHREC) (15/YH/0102) (Appendix 4) and was granted in June 2015. The project summary was published on the Health Research Authority website (http://www.hra.nhs.uk/news/research-summaries/nf1-in-nsclc).

2.2.1.2 Patient Recruitment

NSCLC patients were pre-screened for potential eligibility to the study prior to being approached at WPH lung cancer clinics.

Principal patient eligibility criteria included:

- Pathological (histological) confirmation of non-small cell lung cancers with ADC or SQCC subtypes.
- Archived tissue available at STH
- Willingness to have a blood test
- Able to give written informed consent

Once potential patients were identified, details were passed on to the clinical staff directly involved with the patient's care. The patient's Consultant or Registrar made the final decision regarding whether the patient would be approached or not. Patients who were

approached were given the patient information sheet and the STH 'Donation of body tissue samples for research' leaflet and allowed to ask questions regarding the study. Patients who decided to participate gave written consent for a one off blood sample to be taken, access to their archived tumour tissue, and for members of the research team to access their non-identifiable patient information.

2.2.1.3 ReSoLuCENT Patient Samples

To mitigate for low patient accruement and patient samples with insufficient tumour tissue we applied to access 'Resource for the Study of Lung Cancer Epidemiology in North Trent' (ReSoLuCENT) patient samples and non-identifiable patient information (STH13872), YHREC (05/MRE07/72). This application was successful and any shortfall was made up using these samples. A substantial amendment was submitted to YHREC updating the project protocol for inclusion of these samples.

2.2.1.4 Processing and Storage of NSCLC Blood Samples

Blood samples were collected from the consented patients by a trained phlebotomist, either on the day of consent or on their subsequent visit to WPH lung cancer clinic. Up to 50 ml of whole blood was collected for each patient in 9 ml EDTA vacutainers. After the blood draw samples were stored on ice or at 2-8°C for a maximum of 2 hours before processing. One 9 ml EDTA vacutainer was decanted into 2 ml aliquots and stored at -20°C. The remaining 9ml EDTA vacutainers were subjected to a 2 cycle centrifuging step to isolate the plasma for potential circulating free DNA analysis. The first centrifugation step was at 800g for 10 minutes at 4°C. The separated plasma was transferred to a new tube and further centrifuged at 1600g for a further 10 minutes at 4°C. The resulting plasma was transferred into 1.5 ml aliquots and stored at -80°C.

2.2.1.5 FFPET Samples

All patient surgical resected samples or bronchoscopies were undertaken at STH between 2008-2017. These were processed into formalin fixed paraffin embedded tissue (FFPET) within the hospitals histology department. ReSoLuCENT samples were processed in hospitals throughout the UK from 2004–2013. All FFPET samples had slides created; these were stained by routine haematoxylin and eosin (H+E). The slides were examined by a trained pathologist for total cellularity of the section, the proportion of nuclei that are

tumour as opposed to normal, stroma, and inflammatory cells, and to determine if necrosis was present. Professor Simon Cross outlined the regions of tumour on the slides of study samples for guidance when macro dissecting.

2.2.2 Nucleic acid extraction from FFPET samples

2.2.2.1 Qiagen QIAamp FFPE Tissue Kit

Qiagen's QIAamp DNA FFPE Tissue Kit (QIAamp) and Qiagen's Deparaffinization Solution is a DNA extraction method recommended by Illumina and Agilent for downstream NGS of FFPET.

The original QIAamp protocol has had several steps added to optimise the process for downstream NGS. Qiagen's deparaffinization solution was included to remove paraffin. An overnight incubation at 56°C was included to improve proteinase K digestion of proteins. The 90°C incubation, which is used to remove protein / DNA covalent crosslinks, was reduced to 80°C to reduce the amount of double stranded DNA (dsDNA) from denaturation. All reagents were prepared and stored according to the manufacturers' instructions and all centrifuge steps were completed at room temperature.

4-6 x 10 µm sections were added to a 1.5 ml microfuge tube. 320 µl of deparaffinization solution was added and vortexed vigorously for 10 seconds, then centrifuged at 1000 g for 2 seconds. This was incubated at 56°C for 3 minutes then allowed to cool to room temperature (15–25°C). 180 µl buffer ATL was added and mixed by vortexing. This was centrifuged for 1 min at 11,000 g before adding 40 µl of proteinase K to the lower clear phase and mixed gently by pipetting up and down. It was then incubated at 56°C overnight, then transferred to an 80°C incubator for 1 hour and centrifuged at 1000 g for 2 seconds. The lysate (clear lower phase) was transferred to a new 1.5 ml microfuge tube and 200 µl of AL buffer was added followed by 200 µl of 96-100% ethanol which was then vortexed vigorously. The lysate was then transferred to a QIAamp MinElute column and centrifuged at 6000 g for 1 minute. The flow through was discarded and 500 µl of AW1 Buffer added before being centrifuged at 6000 g for 1 minute. The flow through was discarded and 500 µl of AW2 Buffer added before being centrifuged at 6000 g for 1 minute. The flow through

was discarded and the column was centrifuged at 20,000 g for 3 minutes. The QIAamp MinElute column was transferred to a new 1.5 ml microfuge tube, 50 µl of ATE buffer was added to the column and incubated at room temperature for 5 minutes. Finally this was centrifuged at 20,000 g for 1 minute to elute DNA.

2.2.2.2 Covaris truXTRAC FFPE DNA kit

The Covaris truXTRAC FFPE DNA kit (truXTRAC) is a new extraction method released in 2014. The actual column based extraction is similar to the QIAamp method, however, the initial deparaffinization step is replaced by a focused ultrasonication to remove the paraffin. All reagents were prepared and stored according to the manufacturers' instructions and all centrifuge steps were completed at room temperature.

4 x 10 µm sections were put into a Covaris microTUBE Screw-Cap and 100 µl of tissue SDS Buffer added. The Covaris microTUBE was put onto the mounting plate for the Covaris E220 and run using the parameters shown in Table 2.1

Table 2.1. Covaris E220 setting for removing the paraffin from the fixed tissue.

Setting	FFPE DNA
Duty Factor	10%
Peak Incident Power (PIP)	175
Cycles per Burst	200
Treatment Time	300 seconds
Bath Temperature	20°C

Post focused ultrasonication, 20 µl of Proteinase K solution was added to each microTUBE. The Covaris E220 was then rerun using the above setting, with the exception that the time was reduced to 10 seconds. The microTUBE Screw-Caps were incubated overnight at 56°C, then transferred to an 80°C incubator for 1 hour and centrifuged at 1000 g for 2 seconds. Following this the lysate was transferred to a 1.5 ml tube and 140 µl of buffer B1 was added and vortexed thoroughly. 160 µl of 96-100% ethanol was then added and again vortexed thoroughly and centrifuged at 10,000 g for 2 minutes. The liquid layer was transferred to a purification column, avoiding the paraffin base white layer at the top. The purification

column was then centrifuged at 11,000 g for 1 minute. The flow through was discarded, 500 µl of buffer BW was added to the column and centrifuged at 11,000 g for 1 minute. The flow through was again discarded, 500 µl of buffer B5 added and centrifuged at 11,000 g for 1 minute. The flow through was then discarded and the tube centrifuge at 11,000 g for 1 minute. The purification column was transferred to a new 1.5 ml microfuge tube, and 50 µl of buffer BE, pre-warmed to 70°C, added and incubated at room temperature for 3 minutes. This was then centrifuged at 11,000 g for 1 minute to elute DNA. The first eluate was reloaded to the top of the column, incubated for 3 min at room temperature and re-centrifuged at 11,000 g for 1 minute.

2.2.2.3 Qiagen EZ1 DNA Tissue Kit

Qiagen EZ1 DNA Tissue Kit (EZ1) was used for FFPET DNA extraction at Sheffield Diagnostics Genetic Services (SDGS) based at Sheffield Children's Hospital (SCH). It is a largely automated protocol run on the Qiagen's BioRobot EZ1. Unlike the QIAamp and the truXTRAC, which use silica anionic column based technology, the EZ1 uses silica-coated magnetic particles which bind to the DNA and are washed in subsequent steps.

4 x 10 µm sections were suspended in 180 µl of Qiagen's ATL in a 1.5 ml tube. 20 µl of proteinase K was added to the tube before being vortexed vigorously. This was then incubated overnight at 56°C, on a shaking incubator at 300 rpm. It was then transferred to a 90°C incubator for 1 hour. The BioRobot EZ1 rack place was set up with elution tubes in row 1, tips and tip-holders in row 2, and the EZ1 reagent cartridges. After the 90°C incubation the lysate was transferred to sample tubes in row 4 of the eEZ1 rack. The EZ1 DNA Tissue Kit protocol card was inserted into the BioRobot EZ1 along with the EZ1 rack. Elution volume was set to 50 µl and the protocol started.

2.2.2.4 DNA Extraction from Whole Blood Samples

Extraction of DNA from 1-2 ml of whole blood was performed using the automated Chemagic 360 (PerkinElmer) in the medium volume configuration using the Chemagic DNA Blood Kit. The extraction was carried out in two stages. The first stage involved mixing and lysing of the cells in the blood to release the DNA using lysis buffer 1. In the second stage, the DNA was bound to the magnetic beads using binding buffer 2. The beads were then

washed several times in wash buffers 3-6. Finally the DNA was eluted from the beads in 300 µL of TE buffer.

2.2.2.5 Qiagen RNeasy FFPE Tissue Kit

Whilst NGS requires high quality DNA to generate sequence, the Nanostring nCounter used for the gene expression analysis is very tolerant to FFPET derived RNA. All FFPET samples were sectioned at 10 µM and mounted on uncharged slides. Based on tumour content where possible, RNA was extracted from a minimum surface area of least 50 mm^2. For smaller samples this required dissecting tissue from up to 8 slides.

All reagents were prepared and stored according to the manufacturers' instructions and all centrifuge steps were completed at room temperature.

160 µl of deparaffinization solution was added to a 1.5ml tube. Tumour material (previously outlined by a Pathologist) was macro- dissected from the slide with a scalpel and added to the tube. This was vortexed vigorously for 10 s, then centrifuged at 1000 g for 2 seconds, before being incubated at 56°C for 3 min and allowed to cool to room temperature. 150 µl of Buffer PKD was added and mixed by vortexing. This was then centrifuged for 1 min at 11,000 g. 10 µl of proteinase K was added to the lower clear phase and mixed gently by pipetting up and down. This was incubated at 56°C overnight, then transferred to a 80°C incubator for 15 minutes. The lysate (clear lower phase) was transferred to a new 1.5 ml microfuge tube and incubated on ice for 3 minutes. This was then centrifuged for 15 min at 20,000g, before transferring the supernatant without disturbing the pellet to a new tube. 16 µl of DNase Booster Buffer was added followed by 10 µl of DNase I. This was mixed by inverting twice and centrifuged very briefly, to collect residual liquid, and then incubated at room temperature for 15 minutes. 320 µl of Buffer RBC was added and mixed thoroughly by pipetting. 720 µl of ethanol (100%) was added to the sample, and mixed well by pipetting. 700 µl of the sample was transferred to a RNeasy MinElute spin column and centrifuged for 15 seconds, discarding the flow-through. The remainder of the sample (~515 µl) was added to the RNeasy MinElute spin column and centrifuged for 15 seconds. The flow-through was discarded and 500 µl of Buffer RPE added before being centrifuged at 8000 g for 15 seconds. The flow-through was discarded and 500 µl of Buffer RPE added before been centrifuged at 8000g for 15 seconds. The flow-

through was discarded and the column was centrifuged at full speed with the lid open for 5 minutes. The RNeasy MinElute spin column was transferred to a new 1.5 ml tube. 15 µl of RNase-free water was directly added to the spin column membrane and centrifuged for 1 min at full speed to elute the RNA.

2.2.3 DNA Quantification

2.2.3.1 UV spectrophotometry DNA Quantification

Initial DNA quantification and contamination was assessed using the NanoDrop 8000 (Thermofisher Scientific). The NanoDrop uses UV spectrophotometry to measure absorbance of light at specific wavelengths with in the UV range (<400 nm) when analysing nucleic acids. Nucleic acid absorbs light at 260 nm and aromatic amino acids from proteins absorbs light at 280 nm.

The ratio of absorbance at these wavelengths can be used to assess DNA purity. Absorbance (A) at 260 nm / 280 nm (A_{260} / A_{280}) is used for protein contamination and generally a ratio of 1.8 – 2.0 suggests acceptable levels of protein contamination.

UV spectrophotometry was used to calculate the quantity of nucleic acid using the Beer-Lambert equation:

$$A = \epsilon C \ell$$

A is absorbance at the specific wavelength, ϵ = molar extinction coefficient (50 ng-cm / µl for dsDNA), C is concentration (ng/µl), and ℓ is the light pathway (cm). The NanoDrop 8000 has a detection range for dsDNA of 2.5 – 3750 ng/µl.

The NanoDrop pedestal was wiped with lint-free tissue prior to being blanked with 2 µl of nuclease free water. A known DNA concentration standard was then used to check the NanoDrop calibration. 2 µl of eluted DNA from the extractions was then added to the instrument for analysis.

2.2.3.2 Qubit DNA Quantification

DNA concentration was further quantified by the Qubit 2.0 Fluorometer. This is recommended by Agilent Technologies for DNA quantification for downstream analysis

using NGS. The Qubit 2.0 uses target-specific fluorescence and only emits light when specifically bound to dsDNA. The resulting concentration is more accurate than the NanoDrop as only dsDNA is detected. The Qubit 2.0 has a detection range of 2–1000 ng using the Qubit dsDNA BR Assay Kit and 0.2–100 ng using the Qubit dsDNA HS Assay Kit.

As per Qubit dsDNA assay kit instructions 1 µl of Qubit dsDNA reagent was added to 199 of Qubit dsDNA buffer creating a 1:200 Qubit working solution / per sample. For the two standards used in the assay, 10 µl of standard is added to 190 µl of Qubit working solution in Qubit assay tubes. 1-10 µl of sample is added to 190-199 µl of Qubit working solution in Qubit assay tubes resulting in 200 µl. These were vortexed for 3 seconds then incubated at room temperature for 2 minutes. The samples were then measured against the known standards providing the concentration.

2.2.3.3 qPCR DNA Quantification

To optimise for NGS downstream analysis we required a method of measuring the amount of DNA which is unaffected by chemical modifications and degradation introduced by the FFPET process. Sah *at al*., (2013) and Choudhary *at al*., (2014) previously described an absolute quantitative qPCR method to determine the volume of functional DNA (DNA that is amplifiable) (Sah et al., 2013, Choudhary et al., 2014). This method of quantification was used as a pre sequencing quality control check. The genomic copy number is directly related to the sequencing read depth. When a sequence depth of >300 X is required, at least 300 genomic copies starting material, not considering loss during library preparation, is needed.

Ferritin heavy polypeptide 1 (*FTH1*) gene was used as the target for the qPCR assay. A Thermofisher Scientific TaqMan assay with both primers located in a single exon of *FTH1* was used to amplify a 180 base pair (bp) region of the gene. *FTH1* was used as only 0.3% of SQCC and ADC cases have reported amplification or deletion of this gene and 0.4 % of cases have single nucleotide aberration mutations (Campbell et al., 2016).

High quality genomic DNA (gDNA) was extracted from 2 ml of whole blood via ChemGen Mid. This was then quantified using the Qubit 2.0. A four-fold dilution series was created ranging from 50 ng to 0.2ng (15150 to 59 haploid copies) and used as standards.

The standards were analysed on Applied Biosystems real-time PCR 7500 instrument. Cycling parameters were 50°C x 2 mins, 95°C x 15 mins, (95°C x 15 sec, 60°C x 1 min) x 40 cycles. Mastermix for the qPCR reaction shown in Table 2.2.

Table 2.2**. FTH1 qPCR reagent volumes**

Reagent	Volume (µl)
QuantiTect Master mix (2X)	12.5
FTH1 probe and primers (20X)	1.25
Molecular H_20	6.25
DNA	5
Total reaction volume	25

DNA standards (50 ng to 0.2 ng) were run in technical replicates of 3 shown in Figure 2.1. For the standards to be considered reproducible and accurate the 3 technical replicates had to fall within 0.5 Ct of each other (Standard deviation (SD) < 0.289). To demonstrate the amplification observed in the standards was efficient, thereby ensuring the DNA standards were of high quality, a standard curve was generated using the Log_{10} of standard concentration vs. cycle threshold (Ct). Efficiency should be between 95 – 105%. The gradient of the curve was used to calculate efficiency using the following calculation:

$$\text{Efficiency (\%)} = (10^{(-1/\text{gradient})}) -1 * 100$$

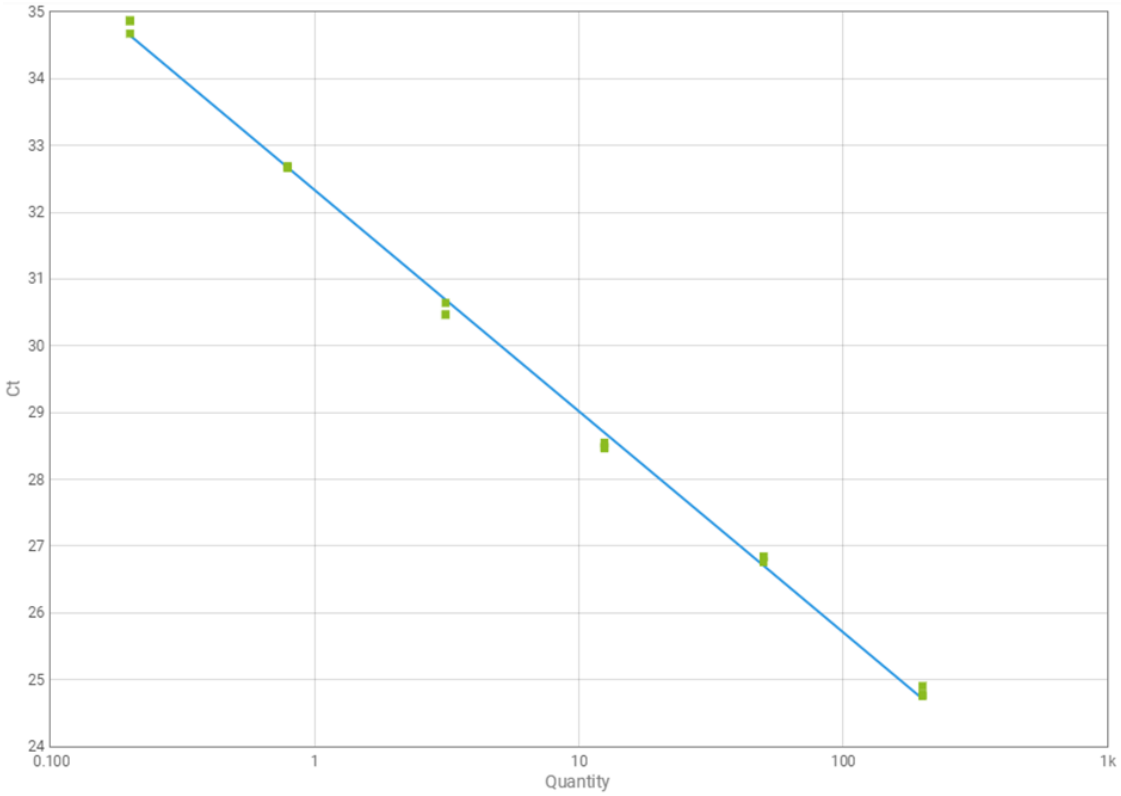


Figure 2.1. Absolute qPCR Standard curve plot. Standards produced from high molecular weight DNA ranging from 200 ng to 0.2 ng (60600 – 59 haploid copies) shown in green. All the standards were run in triplicate and have a < 0.5 Ct between replicates. R^2 = 0.0998, Slope= -3.312, Efficiency = 100.4%.

FFPET extracted DNA was initially quantified using the Qubit 2.0 BR kit. The DNA was then re-quantified via qPCR running three technical replicates against the standard curve. In cases of very poor DNA yield only one replicate was run. The amount of amplifiable FFPET derived DNA was calculated using the following linear regression equation:

$$y = a + bx$$

Where *y* is the dependent variable, *a* is the y axis intercept, *b* is the gradient, and *x* is the independent variable.

A quantitative functional index (QFI) score was calculated as a percentage of amplifiable DNA as follows:

$$\frac{\text{qPCR quantification (ng/µl)}}{\text{Qubit quantification (ng/µl)}} \times 100 = \text{QFI (\%)}$$

2.2.3.4 Agilent TapeStation 2200

The Agilent TapeStation 2200 bioanalyser enables fast high throughput analysis of DNA and RNA using a ScreenTape based electrophoresis system. Data generated can be viewed in a standard pseudo-gel format or electropherogram. Agilent ScreenTape comes in various formats dependent on requirements (Table 2.3).

Table 2.3**. Agilent TapeStation ScreenTape**

ScreenTape	Fragment Size (bp)	Concentration Range
D1000	30 - 1000	0.1 – 50 ng/µl
High sensitivity D1000	30 - 1000	75 – 1000 pg/µl
Genomic DNA	200 to > 60,000	10-100 ng/µl
High Sensitivity RNA	100 - 6000	500–10,000 pg/µL

Manufacturers' instructions were followed for loading the instrument. For DNA the relevant loading buffer was loaded into a 96 well fully skirted plate followed by DNA, RNA, or the ladder (Table 2.4). The 96 well plate was vortexed at 2000 RPM for 1 minute followed by a brief centrifuge. Barcoded ScreenTape, tips, and the 96 well plate were inserted into the TapeStation 2200 for analysis. For RNA analysis the same steps were used with an additional 72°C, 3 minutes incubation, followed by 2 minutes on ice, and a final pulse centrifuge step before loading.

Table 2.4. Agilent TapeStation loading volumes

ScreenTape	Loading Buffer (µl)	DNA / RNA / Ladder (µl)
D1000	3	1
High sensitivity D1000	2	2
Genomic DNA	10	1
High Sensitivity RNA	1	2

2.2.4 Next Generation Sequencing

2.2.4.1 Designing Custom Next Generation Sequencing Probes

In order to enrich the regions of interest during NGS, a custom Agilent SureSelect hybridisation probe set was designed. Hybridisation probes are biotinylated RNA nucleotide sequences complementary to the regions of interest. The probes are incubated with fragmented gDNA, which are then used to pull down and enrich the areas of interest. Hybridisation probes were designed for *NF1*, *EGFR* exons 19 - 21, *BRAF* exon 11, and *KRAS* exon 2-3. However, Agilent recommends custom probe sets should be a minimum of 200 kbp to limit off target sequencing. Therefore a 47 gene panel was designed as described in the following paragraph, consisting of 11537 probes (275 kbp) which were to be used at SCH for their new diagnostic hereditary cancer testing panel (Table 2.5). The panel included the probes complementary to *NF1*, *EGFR*, *BRAF*, and *KRAS*.

Table 2.5**. Agilent SureSelect custom design cancer panel**

Sheffield Children's Hospital, Genetics Department, Hereditary Cancer Panel				
ATM	*PTEN*	*MSH6*	*MITF*	*SDHAF2*
BAP1	*RAD51C (FANCO)*	*MUTYH*	*SDHB*	*SDHC*
BRCA1	*RAD51D*	*PMS2*	*TSC1*	*SDHD*
BRCA2	*STK11*	*SMAD4*	*TSC2*	*TMEM127*
BRIP1	*TP53*	*POLD1*	*VHL*	*KRAS*
CDH1	*APC*	*POLE*	*WT1*	*BRAF*
CHEK2	*BMPR1A*	*FH*	*MAX*	*EGFR*
NBN	*EPCAM*	*FLCN (BHD)*	*NF1*	
PALB2 (FANCN)	*MLH1*	*HNF1A*	*PRKAR1A*	
PPM1D	*MSH2*	*MET*	*RET*	

Genes included in Agilent's SureSelect custom targeted sequencing for a hereditary cancer panel. Agilent Design ID: 3014041

All relevant mRNA transcripts were acquired from the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov) using the reference sequence accession numbers. NCBI gene accession numbers were added to Agilent's SureDesign earray (https://earray.chem.agilent.com/suredesign/) as targets using coding exons and no flanking bases as the region of interest parameters. A BED file was downloaded at the first review step. The BED file contains genomic coordinates for the start and end of each exon for a given target gene. For genes with multiple transcripts, the BED files were merged into one to ensure all possible exons of different transcripts were included. This was done using Galaxy.org BEDTools, MergeBED algorithm (https://usegalaxy.org/). The resulting BED files, including all exons of all transcripts were transferred to a custom excel file, which added ±25 bp of intronic region to the coordinates. These coordinates were then uploaded to Agilent's SureDesign earray. The selection parameters used to design the probes included:

- Density = 5X (The amount of overlapping probes are 5 X deep in a staggered tile manner)
- Masking = Moderately Stringent (Masking hides repetitive regions so probes are not designed in these regions)
- Boosting = Max performance (Probes with a higher GC content are replicated by a higher factor)

Once the probe set was designed the resulting BED file was downloaded. This annotates the target regions, missed regions, and covered regions. Any regions that were missed due to the level of masking in the algorithm were re-submitted with less stringent masking parameters. Figure 2.2 illustrates the 5 X 120 nucleotide tiled probes in relation to exons of interest.

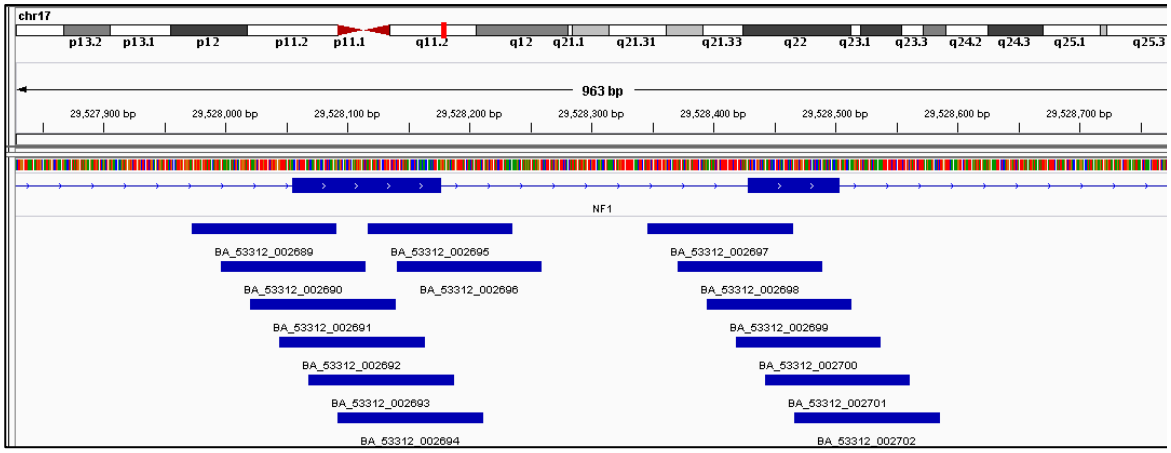


Figure 2.2. Broad Institute's Integrated Genomic Viewer. Agilent's probes covering exons 10 and 11 ± 25 bp of intronic region of *NF1*. Probes are designed to 5X in a tiled manner offset by 20 bp.

The sequences from the resulting probes from the resubmitted missed regions were then requested from Agilent. Each individual probe sequence was then cross referenced with human reference genome (GRCh37/hg19) using the University of California Santa Cruz (UCSC) gnome browser alignment tool BLAT (https://genome.ucsc.edu/). Agilent's hybridisation probes, which had a QSIZE score of >40 and mapped to other regions of the genome were discarded. QSIZE indicates the number of matching base pairs between the sequences uploaded and reference genome.

2.2.4.2 Agilent's XT low input kit Library Preparation

FFPET derived gDNA was used with Agilent's XT low input kit to create NGS libraries following the manufacturers' protocol. Agilent's XT low input kit enables library preparation using between 10–200 ng of starting gDNA. The protocol was optimised using Agilent's DNA integrity number (DIN) and the QFI score summarised in 2.1.2.

Briefly, the gDNA is fragmented to ~200 bp in size and adaptors ligated to the end of the fragments. During the pre-hyb PCR indexes are added during the amplification process alongside P7 and P5 adaptors (complementary to the flow cell). The hybridisation step enriches for the regions of interest followed by a post-hyb PCR resulting in the sequencing libraries (Figure 2.3)

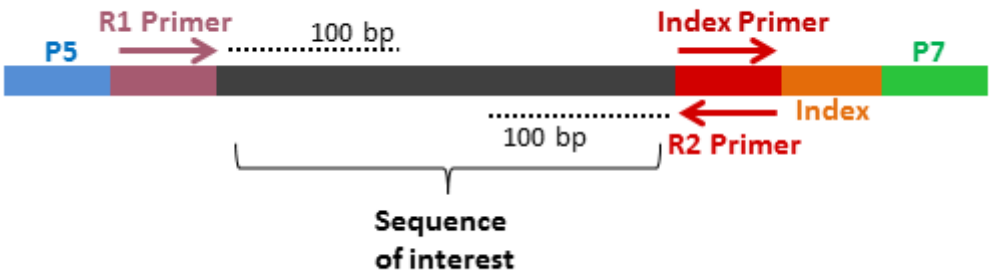


Figure 2.3. Illumina paired end library preparation

Quantity and quality of genomic DNA samples

FFPET derived DNA for sample analysis was quantified using the NanoDrop and Qubit and analysed for integrity using the TapeStation's DIN (2.1.2). The samples were also quantified using qPCR.

Starting amounts of DNA were adapted from Agilent's recommendations considering both the DIN score and qPCR quantification shown in Table 2.6. Samples were diluted in 50 µl of nuclease-free H_2O prior to shearing.

Table 2.6**. SureSelect XT Low Input DNA input**

Protocol Parameter	FFPE Samples	
	DIN > 8	DIN < 8
DNA input for Library Preparation	10 ng to 200 ng DNA, quantified by Qubit Assay	10 ng to 200 ng of amplifiable DNA, based on qPCR quantification

Low Input DNA input modifications based on DNA Integrity Number (DIN) score and qPCR quantification

Shear gDNA to 200 bp

50 µl of H_2O containing up to 200 ng of gDNA was transferred to a 130 µl Covaris microTUBE and briefly centrifuged. These were sheared on a Covaris E220 instrument using the settings shown in Table 2.7.

Table 2.7**. Covaris E220 setting for shearing of gDNA**

Setting	FFPE DNA
Duty Factor	10%
Peak Incident Power (PIP)	175
Cycles per Burst	200
Treatment Time	240 seconds
Bath Temperature	2° to 8°C
Expected fragments sizes in the region of 200 bp	

Repair and dA-Tail the DNA ends

A master mix was prepared for End Repair/dA-Tailing of the sheared gDNA samples. The master mix was prepared according to Table 2.8. It was mixed by pipetting and briefly centrifuged. This was stored on ice till required.

Table 2.8. Preparation of repair/dA-Tail master mix

Reagent	Volume for 1 reaction	Volume for 8 reactions
End Repair-A Tailing Buffer	16 µl	144 µl
End Repair-A Tailing Enzyme Mix	4 µl	36 µl
Total	20 µl	180 µl

20 µl of the End Repair/dA-Tailing master mix was added to each sample well containing 50 µl sheared DNA and mixed via pipetting in a 96 well PCR plate. The plate was sealed and briefly centrifuged. The samples were then incubated on a thermocycler using the following cycling parameters: 20°C for 15 minutes, 72°C for 15 minutes, and 4°C hold. The 96 well plate was then transferred to ice.

Ligate the adaptors

A ligation mastermix was prepared according to Table 2.9, mixed by pipetting, and briefly centrifuged. The ligation mastermix was incubated at room temperature for 30-45 minutes before use.

Table 2.9. Preparation of ligation master mix

Reagent	Volume for 1 reaction	Volume for 8 reactions
Ligation Buffer	23 µl	207 µl
T4 DNA Ligase	2 µl	18 µl
Total	25 µl	225 µl

25 µl of the ligation mastermix was added to the 70 ul of sample in the PCR plate and mixed by pipetting, followed by a brief centrifuge. 5 µl of Adaptor Oligo Mix was then added to each sample and was mixed by pipetting. The plate was sealed, followed by a brief

centrifuge. The samples were then incubated on a thermocycler using the following cycling parameters: 20°C for 30 minutes, and 4°C hold step.

Purify the sample using AMPure XP beads

The samples were then purified to remove all enzymes and substrates from d/A-tailing and adaptor ligation using AMPure XP beads. 80 µl of homogeneous AMPure XP beads (kept at room temperature for at least 30 minutes prior to use) was added to each DNA sample and mixed by pipetting. This was then incubated for 5 minutes at room temperature. The 96 well plate was then added to a magnetic separation device until the solution was clear (approximately 5 to 10 minutes). Whilst still on the magnet the solution was aspirated without disturbing the pelleted beads. Keeping the plate on the magnet, 200 µl of freshly-prepared 70% ethanol was added to each sample well and incubated for 1 minute before aspirating the ethanol. Another 200 µl of 70% ethanol was added and incubated for 1 minute before being aspirated. The PCR plate was then left on the magnet for 5-10 minutes to allow any residual ethanol to dry off. The exact time was determined by the first crack appearing in any of the pellets. 35 µl of nuclease-free water was added to each sample well, the plate was sealed, vortexed, followed by a very brief centrifuge. The PCR plate was incubated at room temperature for 2 minutes. The 96 well plate was then added to the magnetic stand and left for approximately 5 minutes until the solution was clear. The supernatant containing the fragmented adapter ligated DNA was carefully transferred to a new 96 well plate ensuring not to disturb the beads. This was kept on ice until the next step.

Amplify the adaptor-ligated library

The samples were then amplified using the adaptor sequences ligated to the DNA fragments in the previous steps. This was accomplished using one universal primer and one indexed primer which is unique to each sample. Indexes shown in Appendix 5.

A pre-capture master mix was prepared as described in Table 2.10, this was vortexed followed by a brief centrifuge and kept on ice.

Table 2.10. Preparation of pre-capture PCR master mix

Reagent	Volume for 1 reaction	Volume for 8 reactions
5× Herculase II Reaction Buffer	10 µl	90 µl
100 mM dNTP Mix	0.5 µl	4.5 µl
Forward Primer	2 µl	18 µl
Herculase II Fusion DNA Polymerase	1 µl	9 µl
Total	13.5 µl	121.5 µl

13.5 µl of the PCR master mix prepared in Table 2.10 was added to each of the purified DNA library sample. 2 µl of the appropriate SureSelect XT Low Input Index Primer to was then added to each reaction. The plate was sealed, vortexed at high speed for 5 seconds followed by a brief centrifuge. The plate was then added to a thermocycler using the parameters shown in Table 2.11. The number of PCR cycles used is dependent on the starting amount of DNA shown in Table 2.12.

Table 2.11. Pre-Capture PCR Thermocycler parameters

Segment	Number of cycles	Temperature	Time
1	1	98°C	2 minutes
		98°C	30 seconds
2	11 - 14	60°C	30 seconds
		72°C	1 minute
3	1	72°C	5 minute
4	1	4°C	Hold

Table 2.12. Pre-capture PCR cycle number recommendations

Quantity of Input FFPET DNA	Cycles
100 to 200 ng*	11 cycles
50 ng*	12 cycles
10 ng*	14 cycles

* qPCR-determined quantity of amplifiable DNA

Purify the amplified captured libraries using AMPure XP beads

Following the pre-capture PCR the plate was briefly centrifuged before a second AMPure XP bead clean-up was undertaken. This followed the same protocol as previously described on page 47, with the exceptions:

- 50 µl of AMPure XP bead was added to the post-PCR samples in the 96 well plate
- 15 µl nuclease- free water was added to each sample well after the beads were dried
- 15 µl of indexed library was transferred to a new 96 well plate

Assess quality and quantity pre-hybridisation

To assess the quantity and quality for the resulting indexed DNA libraries, the individual samples were analysed on Agilent's TapeStation 2200 using a DNA D1000 ScreenTape using the protocol described in section 2.2.3.4. The individual sample should demonstrate a peak fragment size of approximately 200–400 bp in size on the electropherogram Figure 2.4.

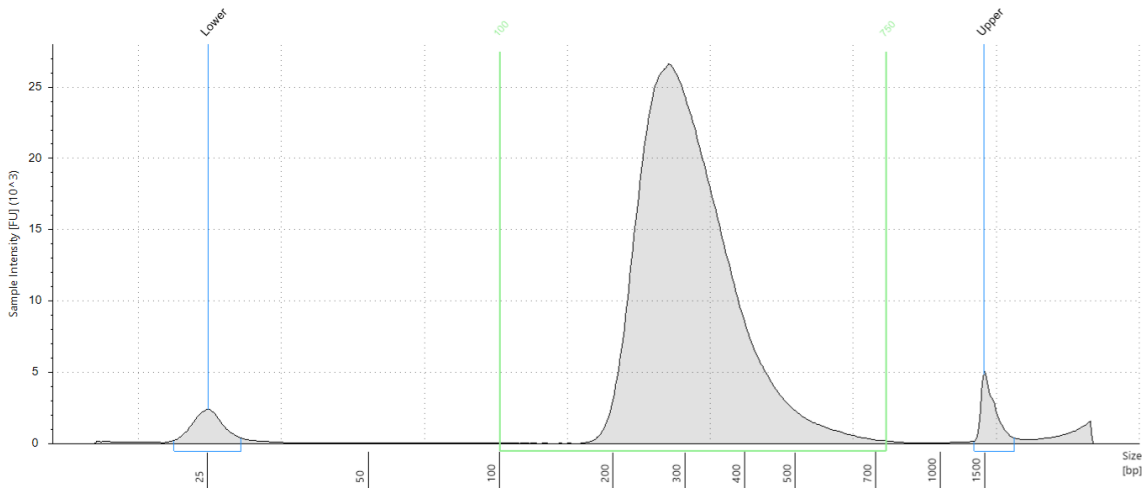


Figure 2.4. Tapestation D1000 electropherogram of pre-hybridisation library. Average peak size 316 bp. Area under the peak was used to determine the concentration.

Hybridization and Capture

The prepared indexed libraries were then hybridized to a target-specific probe set designed in section 2.2.4.1. Up to 1000 ng of the DNA libraries were added to a new 96 well plate in a volume of 12 µl. 5 µl of low input blocker mix was added, the plate sealed and vortexed at high speed for 5 seconds, followed by a quick centrifuge. The plate was then added to a thermocycler. Run parameters are shown in Table 2.13. The thermocycler was paused at the 3rd segment to enable additional reagents to be added.

Table 2.13. Thermocycler program for hybridization with required pause

Segment	Number of cycles	Temperature	Time
1	1	95°C	5 minutes
2	1	65°C	10 minutes
3	1	65°C	1 minute (pause cycler here)
4	60	98°C	1 minute 3 seconds
5	1	65°C	Hold

A 25% solution of RNase Block was prepared according to Table 2.14, this was vortexed, briefly centrifuged and kept on ice until required.

Table 2.14. Preparation of RNase block solution

Reagent	Volume for 1 reaction	Volume for 8 reactions
SureSelect RNase Block	0.5 µl	4.5 µl
Nuclease-free water	1.5 µl	13.5 µl
Total	2 µl	18 µl

A capture library hybridization mix was then prepared according to Table 2.15, vortexed at 1600 rpm for 5 second and centrifuged at 1000 g for 2 seconds.

Table 2.15. Preparation of capture library hybridization mix

Reagent	Volume for 1 reaction	Volume for 8 reactions
25% RNase Block solution	2 µl	18 µl
Capture Library probe set	2 µl	18 µl
SureSelect Fast Hybridization Buffer	6 µl	54 µl
Nuclease-free water	3 µl	27 µl
Total	13 µl	18 µl

The thermocycler containing the DNA libraries and low input blocker mix was paused at segment 3 as described in Table 2.13. 13 µl of the capture library hybridization mix prepared in Table 2.15 was added to the samples. This was mixed by pipetting and re-sealed. The 96 well plate was removed from the thermocycler and centrifuged at 1000 g for 2 seconds before returning it and resuming the programme.

Prepare streptavidin-coated magnetic beads

The Dynabeads MyOne Streptavidin T1 magnetic beads were vigorously re-suspended using a vortex mixer. 50 µl per sample of the re-suspended beads was added to a new 96 well plate.

Wash the streptavidin beads:

200 µl of SureSelect binding buffer was added to 50 µl of the re-suspended beads, this was mixed by pipetting. The 96 well plate was added to a magnetic separator device. After 5 minutes or when the solution was clear the supernatant was aspirated and discarded. The plate was removed from the magnet and the pellet re-suspended in 200 µl of SureSelect binding buffer.

Washing the streptavidin beads was repeated another 2 times, resulting in three total washes. After the final wash the streptavidin beads were left re-suspended in 200 µl of SureSelect binding buffer.

Capture the hybridized DNA using streptavidin-coated beads

When the thermocycler had completed the hybridization protocol the samples were transferred immediately to the 96 well PCR plate containing the washed streptavidin beads. These were mixed by pipetting then resealed. The plate was then mixed at 1600 rpm for 30 minutes at room temperature.

In a fresh 96 well plate 200 µl of wash buffer 2 was added to 6 wells per sample. The 96 well plate was sealed and incubated at 70°C on a thermocycler.

After the DNA libraries and streptavidin beads had finished the 30 minute 1600 rpm room temperature incubation, the plate was placed on a magnetic separator. When the solution was clear, the supernatant was aspirated and discarded. The plate was then removed from the magnet and the pellets re-suspended in wash buffer 1 and mixed by pipetting. The plate was then placed on the magnetic separator. When the solution was clear, the supernatant was aspirated and discarded.

Wash buffer 2:

The 96 well plate was removed from the magnetic separator and the pellets re-suspended in wash buffer 2, pre-warmed to 70°C. These were mixed by pipetting, the plate resealed, and incubated at 70°C for five minutes. The plate was then placed on a magnetic separator for 1 minute, or until the solution was clear. The supernatant was then aspirated and discarded.

The wash buffer 2 washes were repeated 5 more times, resulting in six total washes. After the final wash the beads were re-suspended in 25µl of nuclease-free water and mixed by pipetting. The plate containing bead-bound target-enriched DNA was kept on ice until required.

Amplify the captured libraries

A post capture PCR master mix was prepared as described in Table 2.16.

Table 2.16. Preparation of post-capture PCR reaction mix

Reagent	Volume for 1 reaction	Volume for 8 reactions
Nuclease-free water	12.5 µl	112.5 µl
5× Herculase II Reaction Buffer	10 µl	90 µl
Herculase II Fusion DNA Polymerase	1 µl	9 µl
100 mM dNTP Mix	0.5 µl	4.5 µl
SureSelect Post-Capture Primer Mix	1 µl	9 µl
Total	25 µl	225 µl

25 µl of the post capture PCR master mix prepared in Table 2.16 was added to the 25 µl of bead-bound target-enriched DNA. This was mixed carefully by pipetting. The 96 well plate was placed on a thermocycler using the run parameters shown in Table 2.17.

Table 2.17. Post-capture PCR thermocycler program

Segment	Number of cycles	Temperature	Time
1	1	98°C	2 minutes
		98°C	30 seconds
2	12	60°C	30 seconds
		72°C	1 minute
3	1	72°C	5 minute
4	1	4°C	Hold

After the post-capture PCR program was complete the 96 well plate was briefly centrifuged and placed on a magnetic separation device for 2 minutes, or until the solution was clear. The supernatant was then transferred to a fresh 96 well plate.

Purify the amplified captured libraries using AMPure XP beads

A third AMPure XP bead clean-up was undertaken on the post-capture libraries. This followed the same protocol as described on page 47 with the following exceptions:

- 50 µl of AMPure XP bead was added to the post-PCR samples in the 96 well plate
- 25 µl nuclease- free water was added to each sample well after the beads were dried
- 25 µl of the post-capture libraries was transferred to a new 96 well plate

Assess DNA quantity and quality post-hybridisation

To assess the quantity and quality for the resulting post-capture DNA libraries, the individual samples were analysed on Agilent's TapeStation 2200 using a DNA High sensitivity D1000 ScreenTape. The individual sample should demonstrate a peak fragment size of approximately 200–400 bp in size on the electropherogram shown in Figure 2.5.

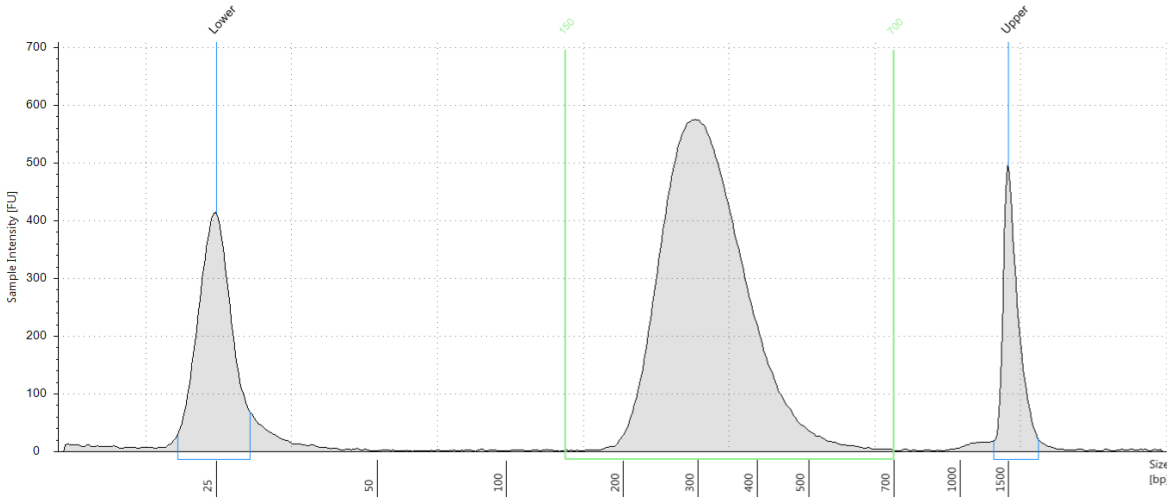


Figure 2.5. Tapestation High Sensitivity D1000 electropherogram of pre-hybridisation library. Average peak size is 321 bp. Area under the peak was used to determine the concentration.

2.2.4.3 Illumina Hi-Seq 2500

Briefly, the DNA libraries are passed through an Illumina 2 lane flow cell. The flow cell has millions of DNA oligos adhering to its surface; these are complementary to adaptors (P5/P7) shown in Figure 2.3. The libraries are denatured before loading and binding to the P5/P7 oligos in a complementary manner. These are then amplified *in-situ* to generate amplicon clusters, via massively parallel bridge amplification. Following cluster generation, the dsDNA is denatured leaving only the amplicon strands, which are now extensions to the oligos adhered to the flow cell via a phosphodiester bond. These strands are then

sequenced by synthesis during the extension of the new complementary strand. In every cycle of sequencing (216 in total) complementary terminated florescent nucleotides were added, the florescence of each cluster read, and the termination molecule and fluorescent cleaved before moving on to the next cycle. Therefore, each cluster generates one sequence read, to be analysed by the bioinformatics pipeline.

2.2.4.4 Pooling of Samples

The initial loading concentration is critical to the density of library binding to the flow cell. The libraries used were quite small respectively (275 kbp), so it was possible to multiplex up to 96 samples in one sequencing run. Samples were pooled to give a final molarity of 4nM. The calculation shown below was performed on each sample used to ensure the individual libraries were pooled equally. Molarity of each sample was calculated using the Agilent's TapeStation.

$$\text{Volume of library (µl)} = \frac{\text{Volume of final pool (µl) x Concentration of final pool (nM)}}{\text{Number of pooled samples x Concentration of library (nM)}}$$

The pooled DNA libraries were made up to the final pool volume using Qiagen's buffer EB with 0.1% tween-20 added and stored at 2-8°C overnight or at -20°C for longer periods.

2.2.4.5 Loading the Hi-Seq 2500

5 µl of the 4nM pooled DNA libraries was added to a 1.5 ml tube labelled 'DNA 20'. In a second tube 2 µl of PhiX and 3 µl of EB buffer with 0.1% tween-20 was added 'PhiX 20'. PhiX is a pre-manufactured library and is used as an internal control. To each tube 5 µl of 0.2N NAOH was added, the tubes were then vortexed and centrifuged briefly. The tubes were incubated at room temperature for 5 minutes. The next HT1 dilution steps were completed in a cold block at 0°C. 990 µl of pre-chilled HT1 buffer was then added to each tube, briefly vortexed and centrifuged. 270 µl of HT1 buffer was then added to 2 fresh 1.5 ml tubes labelled DNA 11 and PhiX 11. 330 µl of 'DNA 20' was transferred to the 'DNA 11' tube and 330 µl of PhiX 20 was transferred to the PhiX 11 tube, briefly vortexed and centrifuged. In a final tube 495 µl of 'DNA 11' and 5µl 'PhiX 11' was added. The final tube

was then incubated at 95°C for 2 minutes then stored on ice prior to loading on the Hi-Seq 2500. At this point the respective molarity based on the dilutions was 11 pM.

The Hi-Seq was set up according to the manufacturers' instructions using Illumina's HiSeq Rapid SBS Kit v2 (200 cycles). The following sequencing parameters were used during set up for analysis of SureSelect libraries. A two lane flow cell was used with the Hi-Seq in rapid run mode.

- Select Paired End
- Select Custom
- Enter 108 cycles for Read 1
- Enter 8 cycles for Index 1
- Leave Index 2 as 0
- Enter 108 cycles for Read 2

2.2.4.6 Low Allelic Fraction controls for NGS

No commercial low allelic frequency standards are available for *NF1*. Therefore, to test the lower limit of detection in relation to samples with low tumour contents or intra-tumour heterogeneity, we created a 3.13% allelic fraction dilution. The dilution was made from two FFPE derived tumour gDNA samples that had previously been sequenced using the methods described in section 2.2.4. Both samples had a similar QFI (59% and 60%) and DIN (6.1 and 6.2) scores demonstrating similar amounts of amplifiable DNA. Multiple heterozygote single nucleotide polymorphisms (SNP) were identified in sample 00136 that were homozygote wildtype at corresponding coordinates in sample 00107. To reduce the bias introduced by the already heterogeneous sample (00136) only heterozygote SNPs with a percentage of reads from 45-55% were considered, all others were excluded from analysis. One of the SNPs was also Sanger sequenced as described in section 2.2.4.7 to confirm via a different method. It is important to note it was necessary to go beyond the regions dictated by the BED file and panel design described in section 2.2.4.1 to identify multiple heterozygote or homozygote in the corresponding samples. Both samples were diluted to 4 ng/µl prior to being serial diluted to create the 3.13 % allelic fraction as shown in Figure 2.6.

Figure 2.6.Illustration of the dilution series used to create the 3.13% allelic fraction control.

2.2.4.7 Sanger sequencing

To confirm one of the *NF1* heterozygote and homozygote SNPs, in samples to be used in the NGS low allelic fraction dilutions, Sanger Sequencing was used as an alternate method of confirmation. Primers were designed flanking the SNP rs55747230 shown in Table 2.18. M13-tails were also added to the primer sequences so the sequencing could fit into a SCH diagnostic workflow. The regions of interest were amplified using a PCR master mix prepared according to Table 2.19. PCR cycling conditions: 94°C x 1 min, (94°C x 30 sec, 57°C x 1 min, 68°C x 1 min) x 33 cycles, 15°C hold.

Table 2.18. Primers flanking rs55747230 in *NF1*

		Sequence (5'->3')
rs55747230	Forward primer	CACGACGTTGTAAAACGACTGGAATTGTGCAGAGTGTGG
	Reverse primer	CAGGAAACAGCTATGACCACATTGGACATACAGTTGAGAGA

M13-tails highlighted in red. Primer shown in black.

Table 2.19. PCR reagent volumes

Reagent	Volume for 1 reaction (µl)
OneTaq	10
Primer F 10 µM	0.5
Primer R 10 µM	0.5
Molecular H_20	4
Genomic DNA (7ng/ µl)	5
Total	20

Following the amplification, amplicons and no template controls (NTC) were analysed on the TapeStation using a D1000 tape to determine product size, non-specific product, and contamination issues. The amplicon products were then diluted in 50 µl H_2O. 2 µl of a 1 in 5 dilution of ExoProStar 1-Step was added to 5 µl of the diluted amplicon. This was incubated at 37°C x 15 mins then at 80°C x 15 mins. Following this forward and reverse BigDye master mixes were prepared according to Table 2.20 using M13 forward and reverse primers. These forward and reverse reactions were then run on a thermocycler using the following parameters 95°C x 1 min, (95°C x 15 sec, 55°C x 10 sec, 60°C x 1 min) x 25 cycles, 15°C hold.

Table 2.20. BigDye reagent volumes

Reagent	Volume for 1 forward reaction (µl)	Volume for 1 reverse reaction (µl)
BigDye Ready React v1.1 Terminator Mix	1	1
Better Buffer	7	7
H_2O	3.8	3.8
Sequencing primer @ 1pmol/µl	3.2	3.2
Exostarred and diluted PCR product	5	5
Total	20	20

The forward and reverse amplicons were cleaned-up using Agencourt CleanSEQ in a 96 well plate. 10 µl of Agencourt CleanSEQ was added to each reaction, then 62 µl of 85% ethanol before been vortexed at 1200 rpm for 30 seconds. The 96 well plate was then added to a magnetic separation device. When the solution was clear (approximately 3 minutes), whilst still on the magnet, the solution was aspirated without disturbing the pelleted beads. Keeping the plate on the magnet, 100 µl of freshly-prepared 85% ethanol was added to each sample well and incubated for 30 seconds before aspirating the ethanol. The 96 well plate was left on the magnet for approximately 10 minutes to allow any residual ethanol to dry off. The 96 well plate was removed from the magnet and 40 µl of nuclease-free water was added to each sample well to elute the sequencing product from the beads. The plate was placed back on the magnet prior to the supernatant containing sequencing product being transferred to a new 96 well plate. This was then sequenced on an ABI 3730 platform. Sequence trace shown in Figure 2.7.

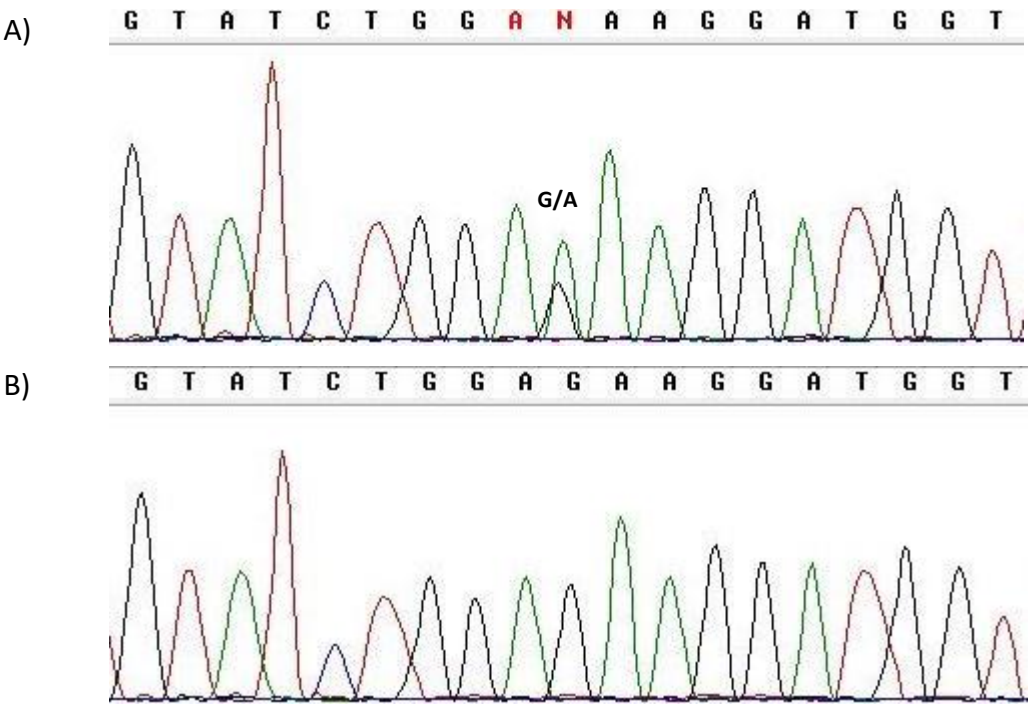


Figure 2.7. rs55747230 sequence trace. A) Reverse trace of sample 00136 the heterozygote spike showing a G and A base. B) Reverse trace of sample 00107 shown to be homozygote G. Forward traces not shown but identical results were observed.

2.2.5 Bioinformatics

2.2.5.1 Generation of aligned BAM files

The primary sequencing output generated by the HiSeq 2500 is a base call file (BCL) format. This file contains reads from each of the clusters and quality scores. The HiSeq 2500 automatically de-multiplexes the samples based on the individual indexes. The BCL files were converted into a FASTQ format using bcltofastq conversion software v1.8.4. Following this the reads were aligned to GRCh37/hg19 using Burrows-Wheeler Alignment tool (BWA) v0.7.15. Once aligned the SAM files produced by BWA were converted to a BAM format using samtools v1.4. The files could then be visualised using software such as the integrated genomic viewer or Alamut. The Genome analysis tool kit software Picard was used to remove optical duplicates from analysis. Samtools was used to generate QC metrics relating to the sequencing quality. The BAM files then underwent a local realignment and a base quality score recalibration resulting in the Phred score for each individual base sequenced.

2.2.5.2 Somatic variant detection

VarScan2 v2.4.0 was used to screen the aligned BAM files for potential variants including nucleotide substitutions, insertions and deletions (indels). VarScan2 is a somatic variant calling software which uses a heuristic/statistic approach to call variants. This is accomplished by defining requirements that need to be met in order for VarScan2 to identify any suspected variant including; desired thresholds for read depth, base quality, variant allele frequency, and statistical significance. VarScan2 outputs a human readable variant call format (VCF) file.

Default VarScan2 parameters:

- Minimum supporting reads at a position to call variants [2]
- Minimum base quality at a position to count a read [15]
- Minimum variant allele frequency threshold [0.01]

VarScan2 uses a tumour sample and a matched normal sample to determine somatic variants. However, to initially screen the tumour samples we used the commercially available genome in a bottle (GIAB) (NA12878) as the normal matched sample. GIAB is a highly characterised genome in which variants have been determined, in parallel, using

five sequencing platforms, seven different aligners and variants determined using three different variant callers. GIAB is used as a validation control for platforms and software. Using this as the normal matched control reduced the need to sequence normal matched samples when variants were not detected. Samples with variants detected then had their matched paired sample sequenced to determine if the variant was somatic or germline in origin.

2.2.5.3 Variant Calling and Annotation

VCF files produced by VarScan2 includes annotation on the variants identified from sequencing. The software program VCFminer was used to filter the files. Criteria for variant exclusion included:

- All GIAB only variants
- Variants outside *NF1* exons ±25 bp and -800 bp of promoter sequence, *EGFR* exons 19-21, *BRAF* exon 11, and *KRAS* exons 2 – 3
- Variants with a population frequency >1% based on the Genome Aggregation Database (gnomAD r2.0.2)
- Variants with a sample AF <3%
- Variants with <15 supporting reads
- Variants that correspond to *NF1* pseudogenes were flagged and removed if no other variant were identified within the specific read

The remaining variants were cross-referenced with the NCBI's Single Nucleotide Polymorphism database (dbSNP) and ClinVar. Variants including indels and nucleotide substitutions were then analysed for possible pathogenicity and loss of function using combination of online resources and *in-silico* tools.

Interactive Biosoftware's Alamut Visual v2.9 was used to interpret sequence variants including missense (synonymous and non-synonymous), nonsense, and indels. This enabled annotation and predictions on frameshifts and downstream premature stop codons. Alamut Visual also has five fully integrated *in-silico* splicing prediction algorithms. These algorithms are: SpliceSiteFinder-like (Zhang, 1998), MaxEntScan, (Yeo and Burge, 2004), NNSPLICE (Reese et al., 1997), GeneSplice, (Pertea et al., 2001), and HSF (Desmet et al., 2009).

Non-synonymous missense variants were also subjected to further *in-silico* analysis using a number of algorithms. SNPs&GO is a machine-learning algorithm which predicts the effect of amino acid changes on protein function (Calabrese et al., 2009). SNPs&GO also uses and outputs from 2 other algorithms; Predictor of human Deleterious Single Nucleotide Polymorphism (PhD-SNP) (Capriotti et al., 2006) and Protein Analysis Through Evolutionary Relationships (PANTHER) (Thomas et al., 2003). PANTHER classifies proteins according to family and subfamily, molecular function, and biological process by aligning the sequence in question to protein families and subfamilies in its own library. All three programmes output a substitution classification; disease or neutral based on a probability score (neutral≤ 0.5) and a reliability score of 1-10 to inform confidence level.

Pmut is another pathogenic mutation predictor which uses neural networks and sequence based information to predict either pathological or neutral effect on protein function (Ferrer-Costa et al., 2004). Sequence based information includes; assessing structural information, residue sequence, and evolutionary properties. The programmes output is a substitution classification; disease or neutral based on a pathogenicity index ranging 0-1 (neutral ≤ 0.5) and a reliability prediction (%).

PROtein Variation Effect Analyser (PROVEAN) is another mutation predictor, which clusters multi-sequence alignment and calculates a score across clusters. If the score is >-2.5 the variant is considered deleterious whereas <-2.5 is considered neutral in its effect on possible pathogenicity (Choi et al., 2012).

Finally Poly-phen-2 is a supervised learning programme which uses evolutionary conservation to access the impact of the variants on the function of the region (Adzhubei et al., 2010). The output is a score 0–1 with 1 being probably damaging, the programme also calculates sensitivity and specificity. For our dataset we used the HumVar output, this is trained on the differences between human disease-causing mutations and common human non-synonymous mutations with no association to disease and a minor AF of >1%.

2.2.6 Mutual Exclusivity

To determine if the variants observed between *NF1* and *BRAF*, *KRAS*, or *EGFR* had any tendency towards mutual exclusivity or co-occurrence we calculated the odds ratio using the calculation below as described by (Gao et al., 2013).

$$\text{Odds ratio} = \frac{(A \times D)}{(B \times C)}$$

Where A = the number of samples where both genes have variants, B = the number of samples with variants in gene 1, C = the number of samples with variants in gene 2, D = the number of samples with no variants in both genes. Possible outcomes shown in Table 2.21.

Table 2.21. Mutual exclusivity based on odds ratio

Odds Ratio	Outcome
0.0 < Odds ratio < 0.1	Strong tendency toward mutual exclusivity
0.1 < Odds ratio < 0.5	Tendency toward mutual exclusivity
0.5 < Odds ratio < 2.0	No association
2.0 < Odds ratio < 10	Tendency towards co-occurrence
Odds ratio > 10	Strong tendency towards co-occurrence

2.2.7 *NF1* Pseudogenes

Luijten *at al.* (2000) and Luijten *at al.* (2001) reported 7 possible *NF1* pseudogenes in their studies. More recent searches on databases including; NCBI and Ensembl now report 12 *NF1* pseudogenes. FASTA sequences for all exonic regions of the functional *NF1* gene and *NF1* pseudogenes ±100 bp of intronic sequence and full intron exon transcripts were downloaded from Ensembl using genome build GRCh38/hg38.

NCBI's alignment tool BLAST was used to calculate homology of pseudogenes compared to the functional *NF1* gene. To identify differences in coding regions a local alignment of all exonic regions ±100 bp for all 12 known pseudogenes was performed against the functional *NF1* gene. This was done utilising European Bioinformatics Institute's (https://www.ebi.ac.uk/) EMBOSS Water algorithm.

2.2.8 Copy Number Analysis

2.2.8.1 ddPCR for copy number Analysis

To investigate copy number variation (CNV) of the *NF1* locus within the tumour samples we employed droplet digital PCR (ddPCR). Briefly, ddPCR partitions a single PCR reaction into 20,000 droplets, in theory; each droplet contains one copy of the target region or endogenous reference and functions as an individual PCR reaction. Samples can be multiplexed containing two individual targets using fluorescent probes. Ribonuclease P/MRP Subunit P30 (*RPP30*) was used as the endogenous reference alongside the target of interest *NF1*. Post PCR each droplet is read individually for florescence and scored either positive for florescence or negative. This yields four possible outcomes when multiplexing 2 targets Figure 2.8.

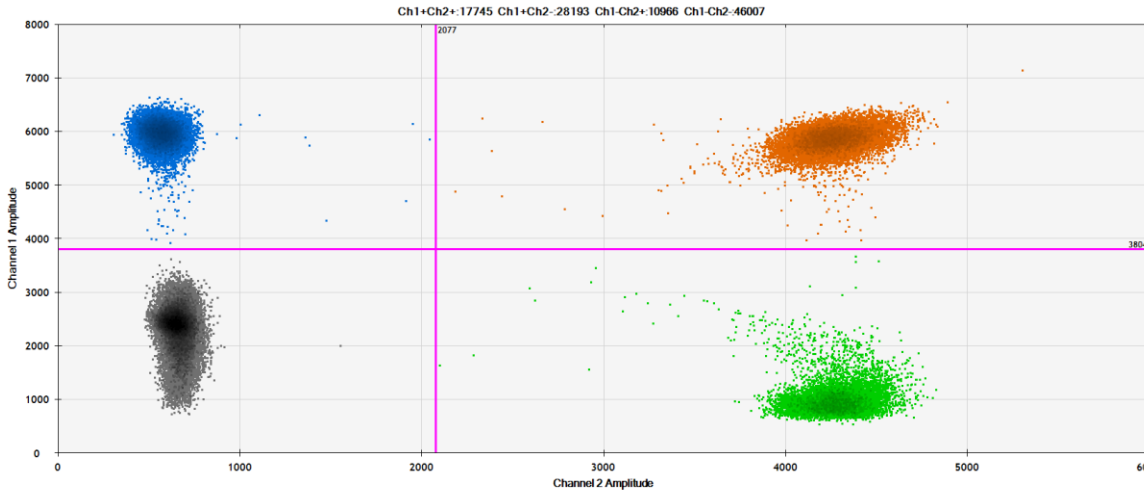


Figure 2.8. 2D plot of droplet florescence amplitude. Black represents negative droplets with no florescence. Blue *NF1* positive droplets. Green *RPP30* positive droplets. Orange both *NF1* and *RPP30* positive droplets.

Between 0.2 and 136 ng of FFPET derived gDNA was added to each reaction, which was analysed in 2 X technical replicates. Bio-Rad *NF1* copy number variation probes were used to target *NF1* labelled with a FAM fluorophore (dHsaCP2507122) and *RPP30* labelled with a HEX fluorophore (dHsaCP2500313). A mastermix was prepared according to Table 2.22.

Table 2.22. ddPCR reagent volumes

Reagent	Volume for 1 reaction (µl)	Volume for 8 reactions plus dead volume (µl)
ddPCR Supermix (no dUTP) (2X)	10	88
NF1 probe and primers (20X)	1	8.8
RPP30 probe and primers (20X)	1	8.8
Molecular H_20	7	61.6
Total	19	167

20.9 µl of mastermix was added to each well used on a 96 well plate. 1.1 µl of gDNA was added, the plate was vortexed at 1600 rpm for 1 minute followed by a brief centrifugation at 500 g for 2 seconds. 20 µl was then transferred to a Bio-Rad DG8 cartridge, 70 µl of droplet generation oil was also added to the DG8 cartridge. A gasket was attached to the cartridge before loading it into the QX20 droplet generator. Once complete 40 µl of emulsified sample was transferred to a fresh PCR plate which was sealed with a foil heat seal and run on a thermocycler using the following parameters, 95°C x 10 mins, (94°C x 30 sec, 55°C x 1 min) x 40 cycles, 98°C x 10 min, 4°C x hold.

After thermocycling the sealed 96 well plate was placed in the QX200 Droplet Reader. QuantaSoft Software was set up for copy number variation, ddPCR Supermix for Probes (No dUTP) was used as the supermix type, Ch1 and Ch2 were labelled with the relevant fluorophores (FAM or HEX) and the target names. The QX200 Droplet Reader was then run to analyse the wells. Once data acquisition was complete the data was analysed using QuantaSoft 1.7.4. Thresholds were set using the 2D amplitude plot shown in Figure 2.8.

2.2.8.2 ddPCR *NF1* 3 copy number control

Commercial standards to use for validation purposes are not available for *NF1* CNV. Briefly, to manufacture a DNA control we amplified the region where the ddPCR primers target, calculated the copy number, and spiked the amplicon into genomic DNA known to be diploid for the *NF1* locus to create a 3 copy number control.

Bio-Rad would not disclose their primer sequence for the *NF1* copy number assay, but did provide a context sequence containing the ddPCR target region. To create the amplicon

spike, primers were designed to encompass as much of this sequence as possible shown in Table 2.23. A PCR master mix was prepared according to Table 2.24 . PCR cycling conditions: 94°C x 1 min, (94°C x 30 sec, 53°C x 1 min, 68°C x 1 min) x 33 cycles.

Table 2.23. Bio-Rads *NF1* context sequence

A	GCCTAGTAGATGAGAACCAGTTCACCTTAACCATTGCAAACCAGGGCACGCCGCTCACC TTCATGCACCAGGAGTGTGAAGCCATTGTCCAGTCTATCATTCATATCCGGACCCGCTGG GAAC

B		Sequence (5'->3')
	Forward primer	ACCAGTTCACCTTAACCATTGC
	Reverse primer	CCAGCGGGTCCGGATATGAA

A) Bio-Rad context sequence which encompasses their assay target region resulting in a 70 bp amplicon. B) Primers shown in yellow used to amplify a 104 bp amplicon used for the amplicon spike.

Table 2.24. PCR reagent volumes

Reagent	Volume for 1 reaction (µl)
OneTaq	12.5
Primer F 10 µM	0.5
Primer R 10 µM	0.5
Molecular H_20	6.5
Genomic DNA	5
Total	25

AMPure XP bead clean-up was undertaken on the amplicon. This followed the same protocol described on page 47 with the exceptions:

- 50µl of AMPure XP bead was added to the post-PCR sample
- 40 µl nuclease- free water to each sample well after the beads were dried

The resulting amplicon was analysed on Agilent's TapeStation 2200 using DNA D1000 ScreenTape. One clear band was observed at 104 bp in size, no signs of non-specific binding and no primer dimers were observed. The amplicon copy number was calculated using the concentration and size of the amplicon and the formula below. The same calculation was used to determine the copies in the genomic DNA been used as the 2 copy diluent.

Number of copies = (Amount (ng) x $6.022x10^{23}$) / (length (bp) x 10^9 x 650)

8690 copies were calculated to be present in 30 ng 2 copy diluent. A serial dilution of the amplicon was used to create a 4345 amplicon standard which was spiked into the 2 copy number diluent. This was used as the 3 *NF1* copy number standard. A set of standards representative of a 3, 2.8, 2.6, 2.4, 2.2 copy numbers were created to represent a 3 copy number change in an 80, 60, 40, and 20 % tumour content sample.

2.2.9 NanoString Gene Expression Analysis

To investigate the gene expression signatures related to the MAPK pathway we decided to utilise the Nanostring nCounter platform. This platform was chosen because it has been shown to be tolerant of RNA analysis from degraded FFPET (Veldman-Jones et al., 2015). Furthermore, it has been validated using both cell lines and clinical tissue for the gene expression signatures we were interested in (Brant et al., 2017, Ahn et al., 2017).

Briefly, it uses in solution hybridisation to bind mRNA targets using a reporter probe and a capture probe. The capture probe was used to pull down target regions, whilst the reporter probe contained a target specific florescent barcode. The unbound RNA was then washed off and the barcodes read using the Nanostring digital analyser.

2.2.9.1 NanoString CodeSets

The NanoString CodeSet contains previously reported MEK and RAS gene expression signatures (Brant et al., 2017, Dry et al., 2010a, Loboda et al., 2010a). The MEK signature proposed by Dry and colleagues consists of 18 up-regulated genes in relation to MEK activation (Dry et al., 2010a). The RAS signature consists of 147 genes in total, 105 of which are up regulated and 47 down regulated in relation to RAS activation (Loboda et al., 2010a). Both signature CodeSets are shown in appendix 6.

In addition to this, three *NF1* probes were designed. These probes were not only used to measure *NF1* expression but could also differentiate between the two most common *NF1* mRNA transcripts (NM_001042492.2, NM_000267.3). These probes are shown in Table 2.25.

Table 2.25. *NF1* Nanostring CodeSet Probes

Exon Boundaries		Sequence (5'->3')
7-8	R	TACCAGATCCCACAGACTGATATGGCTGAATGTGCAGAAAAGCTATTTGA
	C	CTTGGTGGATGGTTTTGCTGAAAGCACCAAACGTAAAGCAGCAGTTTGGC
22-23	R	TGAAGGCAGCTCTGAACATCTAGGGCAAGCTAGCATTGAAACAATGATGT
	C	TAAATCTGGTCAGGTATGTTCGTGTGCTTGGGAATATGGTCCATGCAATT
30-32	R	TCGAAGTGTGTGCCACTGTTTATACCAGGTGGTTAGCCAGCGTTTCCCTC
	C	AGAACAGCATCGGTGCAGTAGGAAGTGCCATGTTCCTCAGATTTATCAAT

Exon numbers related to the longest encoding *NF1* transcript 2 (NM_001042492.2). Reporter probe (R), Capture Probe (C)

2.2.9.2 Nanostring mRNA Hybridisation Preparation

Tumour tissue was enriched via macro-dissection based on pathologist assessment and mark up. Approximately 50-100 mm^2 of tumour tissue was used for each RNA extraction taken from 1-8 10µM slides. RNA extraction described in section 2.2.2.5. The RNA was then quantified using the Qubit RNA high sensitivity kit and diluted to 20 ng/µl where possible using nuclease free water.

The following was made for each 12 reactions. The reporter CodeSet and capture ProbeSet reagents were thawed at room temperature; these were both inverted 5 times to mix once thawed, followed by a brief centrifugation at 500 g for 2 seconds. A mastermix was created by adding 70 µl of hybridization buffer to the tube containing the reporter ProbeSet. This was inverted 5 times to mix, followed by a brief centrifugation at 500 g for 2 seconds. 8 µl of the described mastermix was added to strip tubes. 5 µl of 20 ng/µl of RNA was then added to each of the tubes followed by 2 µl of the capture ProbeSet. The tubes were capped and flicked to mix, and centrifuged for 2 seconds in a microfuge. These were then incubated at 65°C for 21 hours.

Post hybridisation the strip tubes were loaded into the NanoString nCounter prep station. The prep station was then set up using the manufacturers' on screen instructions adding the following: nCounter cartridge, nCounter prep plates, racked tips and foil piercers. The prep station then washed the hybridised RNA and loaded it to the cartridge. Once complete the cartridge was sealed and transferred to the NanoString nCounter digital analyser. The CodeSet RLF file was uploaded to the digital analyser, this contains the

barcodes and related gene names. The cartridge was read within an hour of the prep station finishing at the highest optical setting: 550 fields of view. Once complete the gene count is downloaded as a RCC file for analysis using nSolver v4.0.

2.2.9.3 nCounterData analysis and Gene Expression Calculations

All Nanostring nCounter data were normalised using nSolver Analysis Software version 4.0. In the first instance batch data was QC checked using nSlover.

- Imaging QC: Flag lanes when field of view registration is less than 90%
- Binding Density QC: Flag lanes when binding density is outside the range 0.1 – 2.25
- Positive Control Linearity QC: Flag lanes when positive control R^2 value is less than 0.95
- Positive Control Limit of Detection QC: Flag lanes when 0.5fM positive control is less than or equal to 2 standard deviations above the means of the negative controls
- Manual visualisation of the RAW count data to confirm all RNA controls have expression and all water blanks are below 50 counts.

Raw Nanostring data were then normalised in a three step process.

1. The background thresholding was adjusted using the mean of the internal negative controls plus 2 standard deviations
2. The geometric mean of the internal positive controls was used for technical normalisation
3. Finally the geometric mean of the 21 housekeeping genes was used to normalise for initial differences in input of starting material

The RAS signature score was calculated as described by Loboda *at al*., (2010). The gene counts were expressed as the Log_{10} ratio. The mean of the 47 gene down-regulated arm were then subtracted for the mean of the 105 gene up-regulated arm resulting in the final score per sample.

The MEK signature score was calculated as described by Dry *at al*., (2010). The gene expression counts were expressed as the Log_2 ratio. The mean for the 18 genes was then calculated. The optimised NSCLC MEK signature score as described by Brant *at al*., (2017) was calculated as the 18 gene signature but only using 6 genes.

2.2.9.4 Housekeeping genes

Housekeeping genes were determined suitable or unsuitable using AstraZeneca's in house algorithm. The algorithm analysed the housekeeping genes in two ways

- Correlations
    - The correlation matrix between each gene was calculated and then, for each gene, the average correlation for each gene was calculated. The average correlations and overall average correlation was calculated and any gene which has an average correlation which is less than overall average -2SD is considered potentially unsuitable.

- Standard deviation of average differences
    - The average difference and standard deviation of that average difference between each gene pair was calculated over all samples, and then for each gene the average standard deviation was calculated. From the gene average standard deviations, an overall average standard deviation was calculated and any gene which has an average standard deviation which is greater than overall average SD +2SD may be considered potentially unsuitable.

2.2.10 The Cancer Genome atlas Pan-Lung data set

To make up for shot falls in our cohort we utilised The Cancer Genome atlas (TGCA) using the Pan-Lung data set. This data set was generated in Campbell *et al*. (2016) publication and includes 660ADC and 484 SQCC cases. This was accessed via the University of California Santa Cruz's xenabrowser.net. Gene expression RNAseq was downloaded as gene-level transcription estimates, as in log2(x+1) transformed RSEM normalized count. The CMAP PI3K/AKT/mTOR 190 gene expression signature heatmap was generated using Broad Institutes Morpheus software. Hierarchical clustering was performed via a Euclidean distance metric, using an average linking method and grouped by genotype. Genes used in the CMAP gene expression signature are shown in Appendix 7.

2.2.11 Statistical Analysis and Visualisation

All statistical analysis was performed using SSPS v.22. Unless otherwise stated all assumptions for statistical test used were not violated. GraphPad Prism was used for visual representation. Cbioportal.org was used to generate gene specific oncoprints and lollipop plots (Gao et al., 2013, Cerami et al., 2012). NanoString nSolver 4.0 was used for gene expression normalisation and to generate the correlation plots.

Chapter 3
Optimisation and Validation of Next Generation Sequencing and Digital Droplet PCR

3.1 Introduction

Next Generation Sequencing

NGS has become an immensely powerful tool for analysis of genetic sequence; and is becoming the gold standard in genomic research and diagnostic laboratories. NGS targeted chemistry and workflows are commonly used for detection of germline and somatic mutations including; single nucleotide variations (SNV) and indels. Illumina offers high throughput sequencing using aftermarket commercial custom designed panels.

To generate reproducible next generation sequencing data, pre-analytical requirements include high molecular weight DNA (Arreaza et al., 2016). This is generally obtained from fresh frozen tissue, cell cultures, or from leukocytes. However, diagnosis of lung cancer requires both imaging data and histological review. The histological review is based on analysis of FFPET, as it helps retain the tissues morphological features for the pathologist to examine. Unfortunately, as fresh frozen tissue (the optimal source material for downstream genetic analysis) is rarely available, this makes FFPET the primary source for detection of genetic mutations in solid tumours. FFPET is not ideal for the storage of nucleic acids as extracted DNA is fragmented, of low quality, and is often low in yield. This is the result of the pre fixation time, in which the tissue becomes anoxic, as well as the actual fixation process itself (Srinivasan et al., 2002). Other types of DNA damage induced in this process include: DNA nicks, gaps, protein-DNA crosslinks, C > T / G > A nucleotide transitions, and oxidised bases (Sikorsky et al., 2007).

DNA Extraction

Sikorsky and colleagues have shown DNA fragmentation, nicks, gaps, DNA – protein crosslinks and oxidised bases can all inhibit PCR, a key step of NGS library preparation. It is therefore important to get the maximum DNA yield from the samples and determine if different extraction methods are detrimental to DNA quality. The quantity and quality of the DNA within our patient samples is an unknown variable so in order to maximise sequencing data quantity and quality from the FFPET samples we initially investigated the pre-analytical step of DNA extraction. This investigation included comparing extraction methods and various DNA quantification tools to ensure we obtained as much information as possible from the patient samples.

To address the quality of the DNA extracted, we employed a recent update to the Agilent 2200 software (version 2.1.38.8716). This update gives the ability to calculate the DIN score. The DIN score is based on the level of fragmentation which is seen as a smear on the gel. As score of 1 indicates highly degraded DNA and 10 represents high molecular weight DNA. There is little information in the literature on DIN scores and NGS, but more recent NGS protocols such as Agilent's XT HS kit utilise the score and concentration to determine changes in the library preparation protocol at key steps. It is therefore logical to compare this score across the extractions to determine any effect on DNA quality.

NGS Low Allelic Fraction

Biopsies and surgical specimens of tumours are rarely 100% tumour, but even if they are then intra-tumour heterogeneity must be considered. We aim to validate our NGS sequencing protocol to be able to identify somatic variants in samples with as little as 6% tumour content.

Droplet Digital PCR

ddPCR is an emerging method of measuring copy number changes in clinical samples. It has been shown to tolerate degraded FFPET samples and has demonstrated high correlation to more established methods; such as florescence in-situ hybridisation (FISH), immunohistochemistry (IHC), and SNP arrays (Heredia et al., 2013, Zhang et al., 2016a). ddPCR also removes the subjective judgment that pathologists make with FISH and IHC and is far more cost effective and adaptable than SNP arrays. As with NGS, we investigated ddPCRs capability of detecting CNV of *NF1* in samples with low tumour content.

Pseudogenes

NGS library preparation relies on fragmentation of the genome, and hybridisation enrichment is based on homology. Both of these steps could lead to *NF1P* being enriched and mapped back to *NF1* during alignment to the reference genome, which could result in false positive variants. NGS data analysis alignment does provide some level of security through mapping quality metrics of independent reads. Cunha and colleagues navigated this problem by only using reads with a Q score of >20 (Cunha et al., 2016). Different approaches used include; discarding all reads from analysis that map back to two genomic positions (Pasmant et al., 2015). However, this could result in true positive variants been lost. To address this, we mapped back all *NF1P* to the functional gene to calculate homology and determine variants/differences. These differences will be exploited to flag any possible pseudogene reads mapping back to *NF1* to provide further security in the results.

3.2 Comparison of DNA extractions from FFPET

To address the optimal method for gDNA extraction from FFPE tissue, three extraction methods were compared; Qiagen's EZ1 automated method (EZ1), Qiagen's QIAamp FFPE tissue extraction kit (QIAamp FFPE), and Covaris's truXTRAC FFPE DNA kit (truXTRAC). Anonymised lung FFPE tissue was provided by the SCH histology department for validation purposes. Excess paraffin was removed prior to extraction leaving a surface area of 100 mm^2, of which 48 mm^2 was tissue. 10 µM scrolls were sectioned and 6 scrolls used in each replicate. Each experiment was run with three replicates per extraction method. Three methods of quantification were used; NanoDrop, Qubit, and TapeStation. The fragment size of the genomic DNA was calculated using the TapeStation. The experiment was repeated on three separate occasions to assess inter-extraction variability.

All A_{260} / A_{280} ratios were >1.8 for all three extraction methods as determined by the NanoDrop, suggesting protein contamination is not an issue with any of the methodologies.

3.2.1 Variability of DNA quantification methods

Variable measurement of concentration of the same DNA sample across multiple platforms is a common problem and could provide downstream analytical challenges (Mathot et al., 2013). To investigate this, three methods of quantifying DNA were tested across the three different extraction methods used. All three extraction methods use the same elution volume.

The concentration of extracted DNA via the three different extraction methods was measured by the Qubit, TapeStation, and NanoDrop. All were normally distributed as assessed by Shapiro-Wilk's test ($p > 0.05$). A Levene's test showed that homogeneity of variances was not assumed for the QIAamp FFPE and truXTRAC extractions ($p < 0.05$), but was assumed for the EZ1 method ($p = 0.304$). A Welch ANOVA demonstrated a difference between methods of DNA quantification ($p < 0.001$). A *post hoc* Games-Howell test determined a significant difference between the NanoDrop vs. Qubit ($p < 0.001$) and NanoDrop vs. TapeStation ($p < 0.001$). No significant difference was observed between the Qubit vs. TapeStation ($p > 0.775$) (Table 3.1). To confirm the results from the EZ1 method, in which homogeneity of variances was assumed an ANOVA followed by Tukey *post hoc*

analysis, was also performed in parallel. This resulted in identical grouping, data not shown.

Table 3.1. Games-Howell *Post-hoc* analysis showing the variance between DNA quantification methods using three different extraction methods.

QIAamp FFPE		**Sig**
	Qubit vs. TapeStation	p = 0.888
	Qubit vs. NanoDrop	p < 0.001
	TapeStation vs. NanoDrop	p < 0.001
truXTRAC		
	Qubit vs. TapeStation	p = 0.838
	Qubit vs. NanoDrop	p < 0.001
	TapeStation vs. NanoDrop	p < 0.001
EZ1		
	Qubit vs. TapeStation	p = 0.775
	Qubit vs. NanoDrop	p < 0.001
	TapeStation vs. NanoDrop	p < 0.001

The NanoDrop estimated an increase of >2.5 fold when compared to the TapeStation and Qubit from all extraction methods shown in Figure 3.1. As the Qubit and TapeStation quantification method showed no statistical differences, we proceeded with Qubit quantification which was the most reproducible.

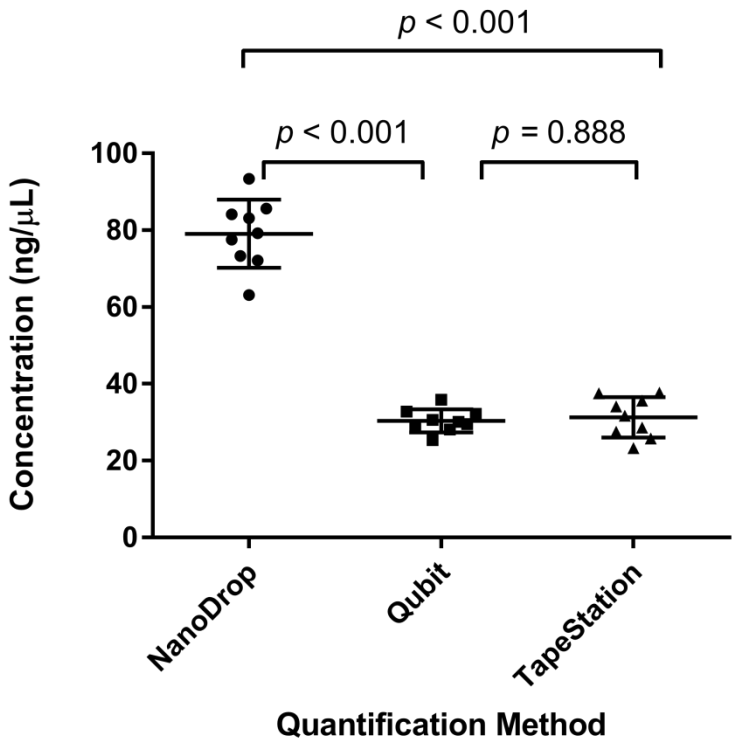


Figure 3.1. Analysis of Variance of DNA Quantification Methods. 9 biological replicates analysed from the QIAamp FFPE extraction method. SD (small whiskers), mean (large whisker) shown on the graph. Statistically significant differences between quantification methods determined by Welch ANOVA (P < 0.001). Significance between individual extractions determined by Games-Howell post-hoc test. The Qubit and TapeStation demonstrated similar means, Qubit (30.4 ng/μL) and TapeStation (31.3 ng/μL) with the NanoDrop displaying a >2.5 fold increase (79.1 ng/μL).

3.2.2 Variability of DNA yield across three extraction methods

To determine the extraction method which delivers the greatest DNA yield in a reproducible manner, the Qubit concentrations across the three extraction methods were compared. This was done using three intra-batch replicates repeated on three independent occasions to assess intra/inter-batch reproducibility.

The concentrations from the QIAamp FFPE, truXTRAC, and EZ1 were normally distributed as assessed by a Shapiro-Wilk's test ($p > 0$ .05). A Levene's test showed that homogeneity of variances was not assumed for any of the three methods ($p < 0.05$). A Welch ANOVA demonstrated a statistically significant difference in DNA yield between extractions ($p < 0.001$). A *post hoc* Games-Howell test determined a significant difference ($p < 0.001$) in DNA yield between the QIAamp FFPE vs. EZ1 and truXTRAC vs. EZ1. No statistical difference was observed between the QIAamp FFPE vs. truXTRAC ($p > 0.556$). However, the QIAamp FFPE extractions demonstrated greater reproducibility (coefficient of variation (CV) = 9.97), compared to truXTRAC (CV = 24.63) (Figure 3.2).

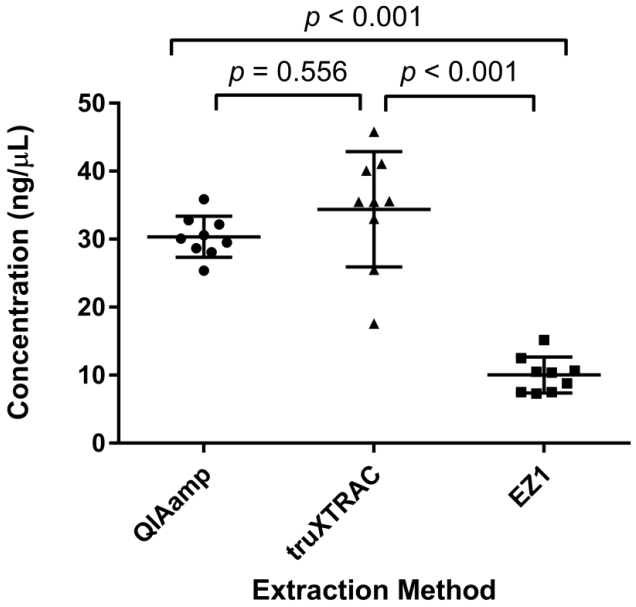


Figure 3.2 Analysis of Variance of Concentration Yield Between Extraction Methods. Qubit measurement of DNA extracted via three different methods. 3 intra batch biological replicates were analysed on three separate occasions giving 9 replicates in total. Significant difference between groups as determined by Welch ANOVA ($P < 0.001$). Difference between individual extractions determined by Games-Howell post-hoc test shown on the above graph. The truXTRAC (34.41 ng/μL), and QIAamp FFPE (30.37 ng/μL) demonstrating a >3 fold increase in yield when compared to the EZ1 (10.04 ng/μL).

3.2.3 DNA Quality

Agilent's DNA Integrity Score

The DIN score for the 9 biological replicates for each of the three extraction methods were normally distributed as assessed by a Shapiro-Wilk's test ($p > 0.05$). A Levene's test showed that homogeneity of variances was assumed for the QIAamp FFPE, truXTRAC, and EZ1 methods ($p > 0.05$). A one-way ANOVA demonstrated no statistical difference between extraction methods and the DIN score ($p = 0.373$) with all extractions having a rounded mean of 2.3. As expected Agilent software failed to calculate a DIN score for the negative control and fresh leukocyte extracted DNA had a high score of 9.3. This indicates the anonymised FFPET sample used is highly degraded. While all three extractions had similar means the QIAamp FFPE had the lowest (Standard deviation (SD) = 0.06) compared to truXTRAC (SD = 0.10) and EZ1 (SD = 0.09). The Pseudo-gel image for 3 replicates for each extraction method is shown in Figure 3.3.

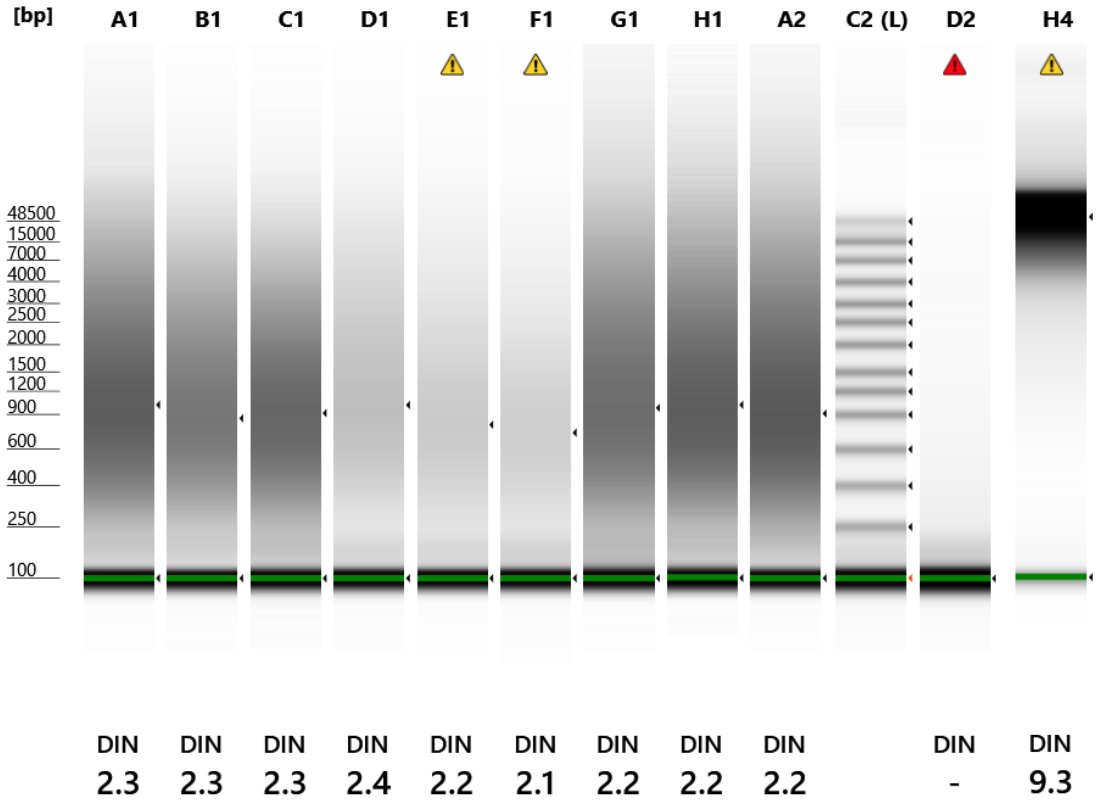


Figure 3.3. Agilent's TapeStation genomic tape pseudo-gel image. QIAamp FFPE lanes A1-C1. EZ1 lanes D1-F1. truXTRAC lanes G1-H2. 48.5 kilobase ladder C2(L). Blank lane D2. Anonymised Leukocyte extracted DNA lane H4 demonstrating high molecular weight DNA. Yellow warning (E1, F1, H4) sample concentration outside recommended range for quantification. Red warning (D2) sample outside functional range for DIN.

3.3 *NF1* Pseudogenes

NCBI's BLAST was used to compare the whole pseudogene sequence (exon and introns) back to its pseudogene genomic location to confirm the sequence downloaded from Ensembl an alternate source. All pseudogene FASTA sequences from Ensembl matched NCBI sequences 100% with E values of 0.0. The whole *NF1P* sequences were then aligned with the functional *NF1* gene. This demonstrated that all *NF1P* intronic and exonic regions have ≥90% homology with the functional *NF1* gene. Homology and pseudogene size shown in Table 3.2.

Table 3.2. *NF1* full pseudogene homology

Pseudogene	Ensemb Transcript ID	Chromosome	Size (bp)	Query covered (%)	Homology (%)
NF1P1	ENST00000604105.1	15	18973	97	90
NF1P2	ENST00000556855.1	15	18985	97	91
NF1P3	ENST00000457709.1	21	4353	79	91
NF1P4	ENST00000549579.2	14	9460	98	91
NF1P5	ENST00000588287.1	18	4342	79	91
NF1P6	ENST00000426025.1	22	9451	98	91
NF1P7	ENST00000612076.1	14	9732	96	92
NF1P8	ENST00000415351.1	2	7660	97	91
NF1P9	ENST00000629326.1	15	19041	97	91
NF1P10	ENST00000613227.1	14	3166	99	91
NF1P11	ENST00000621292.1	14	7300	98	91
NF1P12	ENST00000549593.1	12	962	100	93

In order to compare the exons of *NF1Ps* to *NF1* to identify specific variants/differences we aligned all *NF1P* exons ±100 bp to the corresponding functional *NF1* exons. 2818 differences across the 12 pseudogenes including SNV and indels were identified. These variants were then cross-referenced with NCBI's dbSNP. Surprisingly, over 11% of the variants from our alignments were reported as functional *NF1* SNPS in dbSNPs. Homology of *NF1* exons compared to *NF1P* are shown in Table 3.3.

Table 3.3. *NF1* Exonic pseudogene homology

Pseudogene	Size (bp)	Exons	Corresponding *NF1* exons	Homology (%)
NF1P1	5908	17	18-29, 32-36	91
NF1P2	5568	16	18-29, 32, 34-36	91
NF1P3	1250	4	9-11, 15	94
NF1P4	3881	11	13, 16-24, 36	93
NF1P5	1248	4	9-11, 15	94
NF1P6	3849	11	13, 16-24, 36	92
NF1P7	3730	11	13, 16-24, 36	93
NF1P8	3261	9	16-24	92
NF1P9	5523	16	18-29, 33-36	91
NF1P10	1742	5	21-24, 36	92
NF1P11	2882	8	17-18, 20-24, 36	92
NF1P12	983	2	12-22	93

All *NF1P* exons ±100 bp of intronic region and their respective homology to corresponding *NF1* exons

3.4 NGS Low Allelic Fractions

To test the limits of the NGS protocol and its abilities to detect low level variants we used a 3% allelic fraction standard (6% tumour content) as described in section 2.2.4.6. The 3% mixed standard was sequenced and analysed using the method described in section 2.2.4-2.2.5 using GIAB as the normal matched tissue. Varscan2 was run with default parameters. All variants identified in the sample 00107 (homozygote diluent) and 00136 (heterozygote spike) that did not meet initial criteria described in 2.2.5.3 were excluded from analysis. Remaining variants included the 3% 00136 and false positives. False positives are described here as variants not observed in any of the original samples used to create the 3% allelic fraction (AF) control.

NTC were run on each NGS library preparation batch up to the pre-hybridisation quality check described in section 2.2.4.2. All 8 known SNPs in *NF1* were identified within the 3% dilution with allelic frequencies of 1.85–5.22% and variant read depths of 2-46. 46 false positive were also identified with allelic frequencies of 1–3.85% and read depths of 2- 12 shown in Figure 3.4. Receiver operating characteristic (ROC) analysis for AF the area under the curve (AUC) was 0.962 (95% CI, 0.902 to 1.000) using a cut off of 3.1% AF demonstrating a specificity of 87.5% and a sensitivity of 95.7%. ROC analysis of the variant

read depth revealed similar results, AUC 0.899 (95%CI, 0.740 to 1.000), using a cut off of 10 reads resulted in a specificity of 75.0% and a sensitivity of 95.7%.

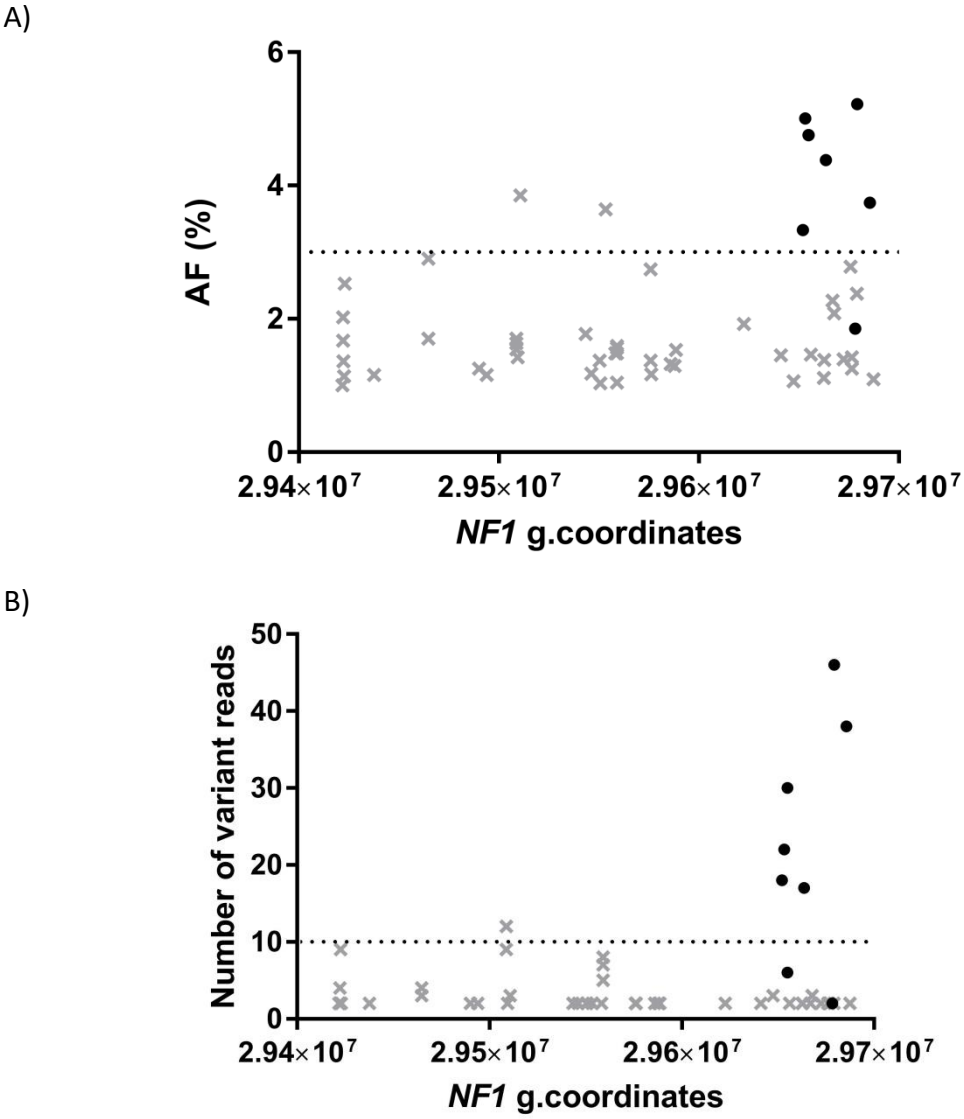
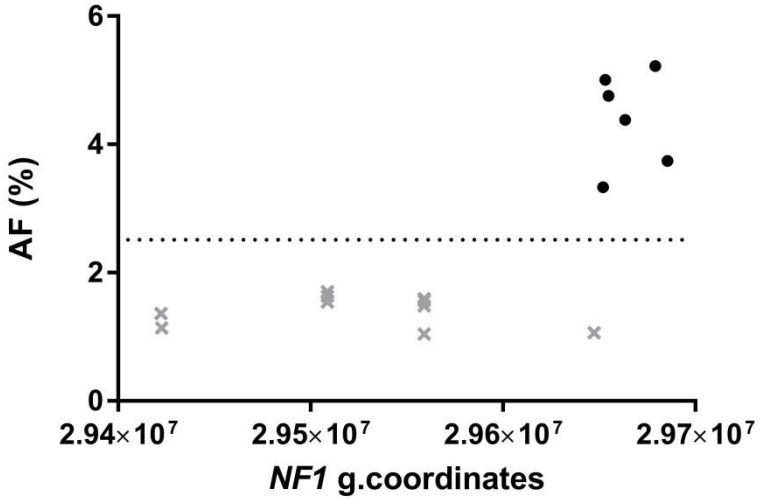


Figure 3.4. True positives and false positives identified across *NF1*. Eight known SNV at 3% allelic fraction (black circles). Grey crosses indicate variants that were not observed in either of the two initial samples used to create the dilution (false positives). *NF1* was aligned to the reference genome hg19. A) Allelic fraction (AF) of all variants observed across *NF1*. ROC analysis demonstrated 87.5% specificity and 95.7% sensitivity using a 3.1% AF cut off. B) Number of variant reads identified. ROC analysis demonstrated 75.0% specificity and 95.7% sensitivity using a 10 read cut off.

To reduce the false positive background a threshold of a minimum of 300 X total read depth was included. This reduced the 8 known 3% AF SNPs down to 6, known positives

g.29654974 and g.29678242 fell below the 300 X threshold. Using 300 X as a cut-off point was not detrimental to analysis as all exonic regions in the validation controls display >500 X. Excluding all coordinates with less than 300 X depth demonstrated a 4 fold reduction in the number of false positives identified. The 6 known SNPs in *NF1* were all identified with allelic frequencies of 3.33–5.22%. 10 false positives were also identified with allelic frequencies of 1.06–1.70%. This gave 2 distinct groups in relation to the AF% with no overlap observed between true positives and false positives resulting in 100% specificity and 100% sensitivity with a 2.5% threshold. The 2 distinct groups were also observed with the variant read depth demonstrating 100% specificity and 100% sensitivity with a 14.5 read cut off point Figure 3.5.

A)



B)

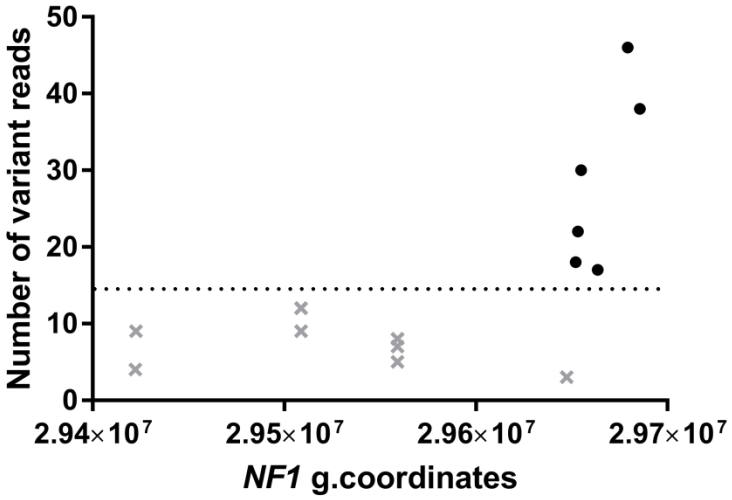
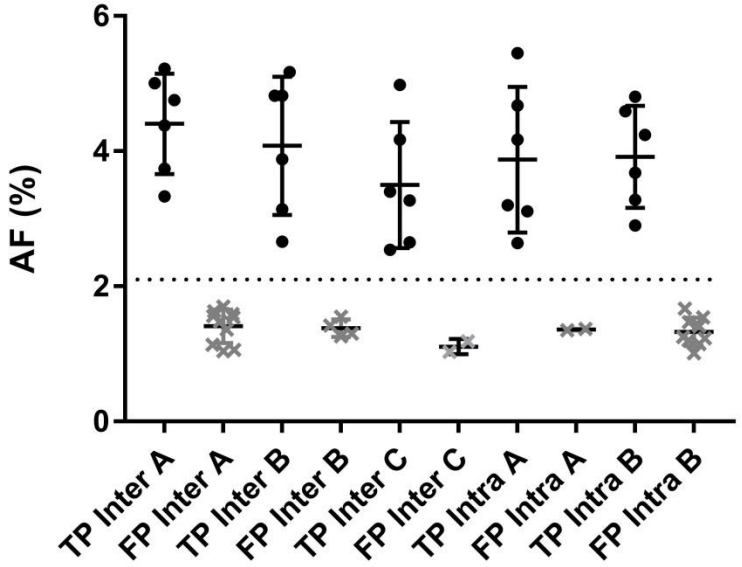


Figure 3.5. True positives and false positives identified across with a > 300 X total read depth. Six known SNV spiked at 3% allelic fraction (black circles). Grey crosses indicate false positives. *NF1* aligned to reference genome hg19. A) Allelic fraction (AF) of all variants observed across *NF1*. ROC analysis demonstrated 100% specificity and 100% sensitivity with a 2.5% AF cut off. B) Number of variant reads identified. ROC analysis demonstrated 100% specificity and 100% sensitivity using a 14.5 variant read cut off.

To assess the reproducibility of the library preparation, sequencing, and data analysis, three inter batch runs and two intra batch runs using different sequencing indexes were performed using the same 3% AF dilution. The 6 known SNPs in *NF1* were identified in all inter / intra batches with allelic frequencies of 2.54–5.45%, false positives ranged between 1.03–1.70%, demonstrating 100 % specificity and 100 % sensitivity with a 2.1% cut-off, Figure 3.6. However, when analysing the variant reads of the independent batches the inter C batch displayed an overlap of false positives and true positives. When considering all inter / intra batch analysis this resulted in 90% specificity and 100% sensitivity using a 14.5 variant read cut-off shown in Figure 3.6 . No pseudogene variants were identified. 7 false positives were observed in more than one inter batch run, the most frequent been a SNV g.29647386 C>T, identified in 4/5 of reproducibility batches Table 3.4.

A)



B)

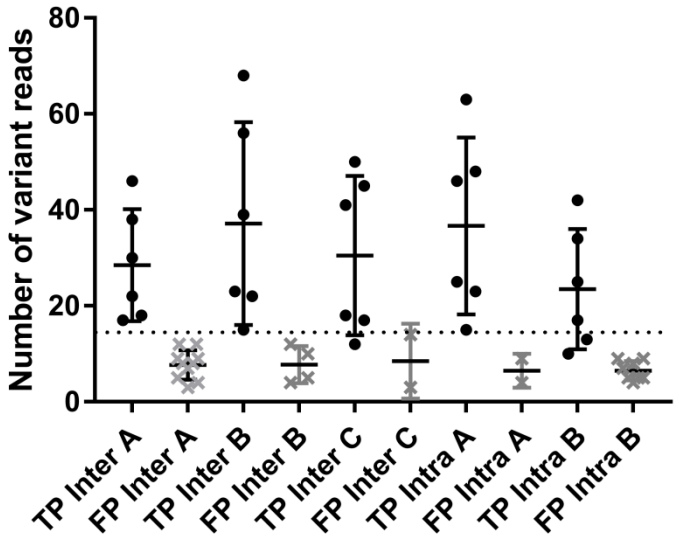


Figure 3.6. Inter / intra batch true positives and false positives identified across *NF1* with a > 300 X total read depth Grey crosses indicate false positives. Black circles represent true positives spiked at 3%. g.coordinates in reference to hg 19. A) Allelic fraction (AF) of all variants observed across *NF1*. ROC analysis demonstrated 100% specificity and 100% sensitivity with a 2.1% AF cut off. B) Number of variant reads identified. ROC analysis demonstrated 90% specificity and 100% sensitivity using a 14.5 read cut off.

To determine if using GIAB for the matched tumour had any influence over variant identification the bioinformatics was repeated using 00107 as the normal matched. 100% concordance was observed with all true positives and false positives.

Table 3.4. Allelic Frequency and read depth of false positives and true positives

Genomic coordinates	Inter A		Inter B		Inter C		Intra A		Inter B	
	VR	VAF %	VR	VAF %	VR	VAF %	VR	VAF %	VR	VAF %
***NF1* TP**										
17:29652147	21	3.33	57	5.17	23	2.54	25	3.11	10	3.28
17:29653293	27	5.01	30	3.88	22	3.4	23	4.68	13	2.90
17:29654876	39	4.76	31	2.62	56	4.98	48	5.45	25	4.24
17:29663624	21	4.38	19	3.14	13	2.65	15	2.64	17	4.59
17:29679246	62	5.22	68	4.82	71	4.17	46	3.20	42	4.81
17:29685660	48	3.74	89	4.82	64	3.27	63	4.17	34	3.68
***NF1* FP**										
17:29422184	4	1.36	NC	NC	NC	NC	NC	NC	NC	NC
17:29422543	9	1.14	NC	NC	NC	NC	NC	NC	NC	NC
17:29508788	NC	NC	NC	NC	NC	NC	NC	NC	8	1.23
17:29508855	NC	NC	NC	NC	NC	NC	NC	NC	4	1.20
17:29557853	NC	NC	NC	NC	14	1.19	NC	NC	NC	NC
17:29559045	8	1.56	NC	NC	NC	NC	NC	NC	NC	NC
17:29559086	NC	NC	NC	NC	NC	NC	NC	NC	5	1.15
17:29559090	7	1.48	NC	NC	NC	NC	NC	NC	NC	NC
17:29559093	NC	NC	NC	NC	NC	NC	NC	NC	5	1.01
17:29559131	NC	NC	1.26	12	NC	NC	NC	NC	9	1.54
17:29647382	NC	NC	4	1.30	NC	NC	NC	NC	NC	NC
17:29508768	12	1.63	NC	NC	NC	NC	NC	NC	NC	NC
17:29508790	12	1.7	NC	NC	NC	NC	NC	NC	9	1.47
17:29508811	9	1.54	NC	NC	NC	NC	NC	NC	7	1.41
17:29559035	8	1.59	10	1.55	NC	NC	NC	NC	7	1.67
17:29559057	5	1.04	NC	NC	NC	NC	9	1.35	5	1.25
17:29647386	3	1.06	5	1.42	3	1.03	4	1.37	NC	NC

Variant reads (VR). Variant allelic frequency (VAF). NF1True positives (NF1TP). NF1 false positives (NF1FP). No call (NC) indicates no variant identified.

3.5 ddPCR low allelic frequencies

To test the limits of the ddPCR protocol and its ability to detect copy number changes at low AF, a 3 *NF1* copy number control was manufactured as described in section 2.2.8.2. *RPP30* was used as the endogenous reference as reports of amplification and deep deletions are only observed in 0.4% of ADC and SQCC cases (Campbell et al., 2016). A series of validation dilutions standards was created to represent a 3 copy number change of *NF1* in 100, 80, 60, 40, 20 and 0% tumour content samples. These were analysed in duplicate on each batch, Figure 3.7.

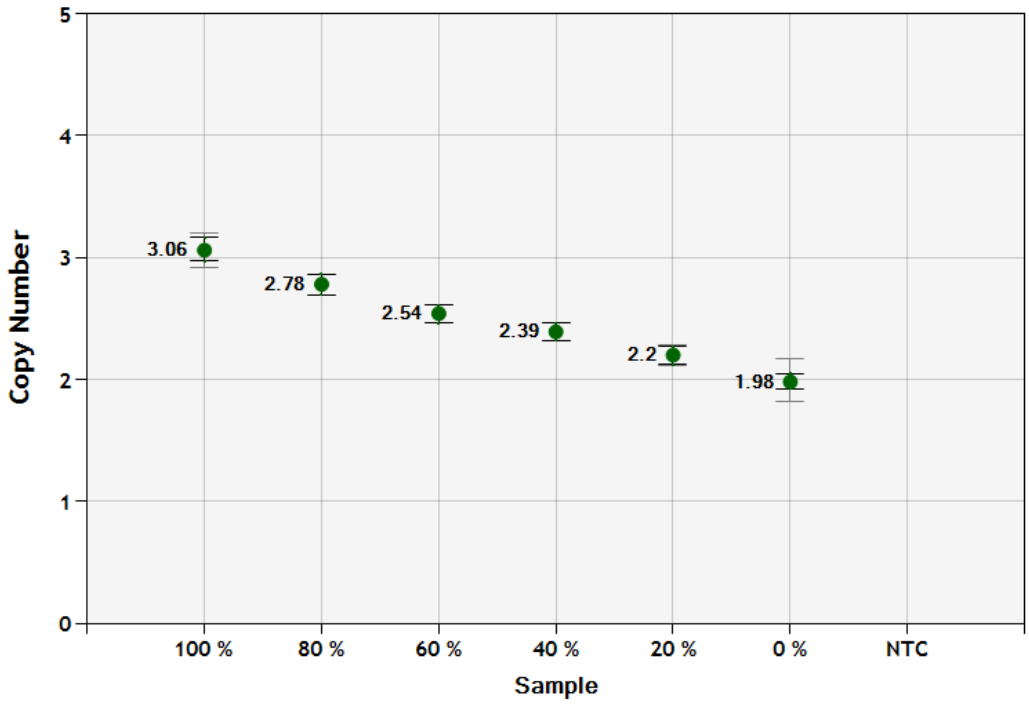


Figure 3.7. ddPCR Intra batch analysis. Dilutions standards representative of a 3 copy number change of *NF1* in 100, 80, 60, 40, 20 and 0% tumour content samples. The graph shows both intra batch replicates merged.

The copy number of the validation standards were measured in three independent inter batch runs shown in Table 3.5. In addition, commercially sourced DNA from the cell line SW48 was analysed. SW48 is reported to be diploid at the 17q11.2 and 10q23.31, the loci of *NF1* and *RPP30* (Knutsen et al., 2010). A *NF1* homozygote loss control (*NF1* 1 copy) provided by SCH with supporting array comparative genomic hybridization data was also analysed. In addition to this, one low quality FFPET sample NF1-115 was tested (DIN 2.4, QFI 3.4%) in two inter batch runs.

Table 3.5. ddPCR Intra batch reproducibility.

	Batch 1		Batch 2		Batch 3				
	R1	R2	R1	R2	R1	R2	Mean	SD	CV
100%	3.13	2.99	2.92	3.13	3.06	3.06	3.05	0.08	2.68
80%	2.77	2.78	2.84	2.77	2.78	2.88	2.80	0.05	1.64
60%	2.54	2.54	2.6	2.55	2.59	2.58	2.57	0.03	1.04
40%	2.39	2.39	2.3	2.38	2.45	2.4	2.39	0.05	2.03
20%	2.16	2.24	2.22	2.27	2.24	2.09	2.20	0.07	3.02
0%	1.91	2.09	2	2.01	2.06	2.02	2.02	0.06	3.06
SW40	2.03	2.03	2.07	1.96	2.00	2.05	2.02	0.04	1.92
FFPET	2.03	2.1	2.13	2.10	N/A	N/A	2.09	0.04	2.03
NF1 1 copy	0.98	0.99	1.02	1.02	0.98	0.99	1.00	0.02	1.66
NTC	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-

Replicate 1 (R1), replicate 2 (R2)

The 3 copy number standards and controls showed good reproducibility across the three inter batch runs with all CVs below 3.06% shown in Table 3.5. SW40, *NF1* loss, and the tumour content standards all came up as expected. Whilst there is a clear threshold between 0% and ≥40% standards there was an overlap between the 0% and 20% standards show in Figure 3.7. ROC analysis looking at the assays ability to detect a 3 copy number change in a 20% tumour content sample displayed 100% sensitivity and 83.3% specificity using a 2.08 copy number cut off point, with a AUC of 0.986 (95% CI, 0.932 to 1.000).

3.6 Discussion

FFPET pathological analysis is a primary method used to differentiate and diagnose solid tumours. This also makes it a valuable primary and readily accessible source of genetic material. Unfortunately, FFPET is a far from ideal method for preserving and storage of nucleic acids. The process induces multiple types of DNA damage including; DNA fragmentation, DNA nicks, protein-DNA crosslinks, nucleotide transitions, oxidised bases, and ultimately yields low volumes of DNA during extraction (Srinivasan et al., 2002).

Here we demonstrate that different extraction methods have significant effects on DNA yield returned. Of the extraction methods we assessed, the truXTRAC and QIAamp methods showed a >3 fold increase in DNA yield compared to the EZ1 method. We postulate the reason for this is a lack of a de-paraffinisation step prior to the EZ1 protocol being run. The mean DNA concentration extracted from the truXTRAC and QIAamp FFPE showed no significant difference ($p = 0.556$) shown in Figure 3.2. DNA quality as assessed by Agilent's DIN score demonstrated no statistical differences between the three extraction methods. This suggests all three extraction methods have the same effect on DNA damage.

Based on these data, the truXTRAC and the QIAamp FFPE kits are the optimal methods of extracting DNA from FFPET. However, the truXTRAC suffers from more random variance than the QIAamp FFPE method shown in Figure 3.2. Other drawbacks of the truXTRAC include that it is labour intensive, taking approximately 2-3 hours longer than the other methods. Furthermore, loading FFPET scrolls into the truXTRAC microtubes is challenging and thus increases the risk of contamination introduced by excess sample handling. For these reasons we chose to move forward using the QIAamp FFPE method, which proved the more robust method of extraction with reduced sample handling from FFPE tissue.

Downstream genetic analysis can be further impacted by the variance in quantification methodology. We have shown that there is a significant difference between the NanoDrop measurement of replicate DNA, compared to the Qubit and TapeStation. The NanoDrop demonstrated a 2 to 3 fold increase over the other two methods shown in Figure 3.1. Similar observations have been described in other studies (Mathot et al., 2013). One explanation is the NanoDrop measures all nucleic acids that absorb at 260 nM (RNA, ssDNA

and nucleotides), whilst the Qubit uses a dsDNA binding dye and only measures dsDNA. Simbolo and colleagues demonstrated this, showing the Qubit is insensitive to RNA contamination when measuring DNA concentration (Simbolo et al., 2013).

NGS of FFPET derived DNA is challenging. Initial attempts to perform NGS from FFPET using Illumina's TruSight cancer panel, Agilents QXT, and XT kits proved unreliable (data not shown). This was due to a number of reasons including; the enzymatic fragmentation step used in Illumina TruSight and Agilents QXT kit could not be optimised for highly degraded FFPET. Agilent's XT kit had issues with DNA loss due to its multiple clean up steps, making it unsuitable for low yield samples. However, new NGS library preparations kits, such as Agilent's XT HS released in October 2017 are more suitable for sequencing of FFPET due to their low DNA input requirement and removal of DNA clean up steps. This makes it possible to sequence and identify variants even in samples with very low tumour content.

Early NSCLC genomic profiling studies by (Ding et al., 2008, Hammerman et al., 2012, Collisson et al., 2014) only sequenced samples with >60% tumour content. Here we were able to identify variants with a theoretical 6% tumour content (3% allelic fraction). We have demonstrated, using a minimum read depth of 300 X, the specificity and sensitivity was increased to 100% with a 2.1% AF cut off, and 90% and 100% respectively using a 14.5 read cut off. This is reflective of early work by Kerick and colleagues who demonstrated that increasing read depth increased variant concordance between fresh frozen and matched FFPET samples (Kerick et al., 2011).

Whilst many papers have acknowledged the presence of *NF1* pseudogenes when analysing genetic sequence using NGS, few have addressed potential issues of how to resolve these pseudogenes mapping back to the functional gene (Pasmant et al., 2015, Cunha et al., 2016). Whilst no pseudogene reads were observed in our validation controls we have provided a further measure of security against this issue.

It is important to note that we earlier described false positives as variants not seen in either of the two samples used to create the 3% variant control. However, several false positive variants were observed in multiple inter / intra batches shown in Table 3.4. This suggests they are not just sequencing or PCR artefacts, which would be randomly distributed (Oh et al., 2015, Spencer et al., 2013). As no potential pseudogene variants were identified, this

suggests the false positives are either low level variants not identified in the initial sequencing of the two samples, contamination, or FFPET artefacts. FFPET artefacts can originate from degradation of cytosine to uracil, which post PCR results in a C>T and G>A transition in the sequence product distributed (Oh et al., 2015, Spencer et al., 2013). The potential false positive 17:29647386 observed in 4/5 of the batches was a C>T transition and could have originated from this degradation process. However, based on our data it is not possible to confirm this. While this highlights the challenges when sequencing and identifying low level variants, it is beyond the scope of our work to investigate this further.

Unlike NGS, ddPCR has shown to be more tolerant of FFPET samples as source material for analysis (Heredia et al., 2013, Zhang et al., 2016a). However, contaminating surrounding tissue is an issue that needs to be addressed when investigating copy number changes. Here we show ddPCR is highly reproducible at detecting these changes of the *NF1* locus. We have shown that a three copy number change can be detected in samples with as little as 20% tumour content with 100% sensitivity and 83.3% specificity using a 2.08 copy number cut off point. This will enable us to move forward without the need for macro-dissection in any samples with >20% tumour content.

In summary this set of validation data and pre analytical requirements enable generation of high quality sequence from low quality FFPET samples. These validation data will give confidence in the data generated from actual analysis of patient samples.

Chapter 4
Screening Patients for *NF1*, *BRAF*, *KRAS* and *EGFR* Variants

4.1 Introduction

Genomic anomalies are the defining traits of cancer. These traits lead to characteristics which are the hallmarks of the disease (Hanahan and Weinberg, 2011). Over the last two decades our ability to identify these genetic anomalies has grown almost exponentially, largely due to advances in genetic screening technology. This has led to ADC and SQCC subtypes being further characterised based on their genotype of recurrent aberrations. In limited cases the genotype can also be used to predict pathological response to targeted therapeutics. Genetic profiling studies of NSCLC highlighted and confirmed many of the recurrent aberrations of potential drivers and tumour suppressor genes (Ding et al., 2008, Hammerman et al., 2012, Collisson et al., 2014, Campbell et al., 2016). The same studies also demonstrated the high mutation burden that the ADC and SQCC subtypes carry. This makes the separation of potential driver events in relation to benign passenger anomalies a challenge.

Oncogenic variants which activate the MAPK pathway are well established drivers of ADC. These are largely somatic SNV or small indels which result in constitutive activation of this pathway. These drivers account for up to 60% of all ADC cases and are predominately found in *KRAS*, *EGFR* and *BRAF.* The drivers for the remaining 40% of cases have still to be identified (Collisson et al., 2014). Conversely, SQCC lacks many of the known *MAPK* driver mutations seen in ADC. Somatic copy number aberrations are frequently observed in SQCC. Copy number increase of the 3q26-28 locus is the most common variation, observed in 30-40% of all SQCC cases. This region includes potential drivers *PIK3CA* and *SOX2*. Somatic mutations in the *PI3K/AKT* pathway account for 47% of cases (Hammerman et al., 2012). These results have also been confirmed in more recent studies, suggesting SQCC drivers are largely MAPK independent (Campbell et al., 2016).

Whilst ADC and SQCC share few potential genetic drivers, they do share common somatic mutations of tumour suppressor genes. The most prevalent are somatic SNVs of *TP53*, which are observed in 86% and 54% of SQCC and ADC cases respectively. Deletion of chromosome 9p21.3, the locus of *CDKN2A* and *CDKN2B,* is another common anomaly, with deep deletions observed in 27% of SQCC and 17% of ADC cases (Campbell et al., 2016). The functional consequence of these somatic aberrations has been well characterised (Kathryn et al., 2014, Kim et al., 2016).

Another recurrent event in SQCC and ADC is that *NF1* aberrations are observed in 8-12% of both populations (Collisson et al., 2014, Hammerman et al., 2012, Ding et al., 2008). No study has addressed the functional relevance of *NF1* somatic variants in clinical NSCLC cases. *NF1* somatic mutations are in part responsible for Neurofibromatosis pathogenesis, leaving individuals more susceptible to development of benign and malignant tumours of the peripheral nerve sheaf (Upadhyaya et al., 2004). Knock down of *NF1* can drive cellular proliferation through activation of the MAPK pathway in ADC cell based models, which is enough to confer resistance to TKI treatment (de Bruin et al., 2014a). More recently Selumetinib, a MEK inhibitor, was granted orphan drug designation by the US Food and Drug Administration for treatment of Neurofibromatosis. These findings all suggest *NF1* loss of function could contribute to the upregulation of the MAPK pathway.

The first step to investigate these aberrations of *NF1* and their ability to drive the MAPK pathway was to recruit and consent NSCLC patients. This enabled access to their archived tumour tissue and non-identifiable clinical data. We used a combination of approaches to screen patients for somatic changes to *NF1* and key activators of the MAPK pathway. The NGS methods described in Chapter 3 produced a robust workflow with the capabilities we required to screen for genetic variants in FFPET. The validation data demonstrates its sensitivity and specificity at identifying somatic variants at low allelic fractions, which is to be expected in low content tumour and heterogenic cases.

In addition to this we have also utilised ddPCR to investigate *NF1* copy number variation. This platform is more commonly known for its ability to detect SNV at low AF. However, it has demonstrated potential in its tolerance to degraded FFPET samples (Zhang et al., 2016a), and our validation data demonstrates its ability to detect copy number changes in samples with as little as 20% tumour content.

We demonstrated that both of these methods of analysis have the potential for identifying variation in cases with low tumour content. However, the methods have yet to be fully established in consideration to the full variability of quality and quantity of FFPE derived DNA from clinical cases. We will also continue to investigate the effect that the various methods of measuring quality and quantity of degraded DNA can have on downstream analysis using these workflows.

4.2 Patient Recruitment and Sample Acceptance

We recruited 86 of a planned 100 patients from WPH lung cancer clinics from 2015 to 2018 using the eligibility criteria described in section 2.2.1.2. After reviewing the first cohort of patient samples it became apparent many were insufficient for downstream analysis. We required a minimum of 10 ng of DNA for screening of *NF1, EGFR, BRAF* and *KRAS* and 100 ng of total RNA for downstream gene expression analysis. The lower limit of starting material for ddPCR analysis was not known at this point; theoretically the assay is based on a simple ratio of endogenous reference to target. We therefore assumed 2 ng of DNA, which is equivalent to over 600 haploid copies, should be sufficient to calculate this ratio.

Based on this, we defined criteria for pre-analytical sample rejection: samples with less than 4mm^2 area of surface tissue based on the H+E section would not be viable for downstream analysis. Another factor considered was exhausting available tissue, as further diagnostic results could be required to aid future patient treatments. Based on these criteria 23 samples were not viable for analysis. Rejected samples were either bronchial brushing or bronchoscopies, none of which matched their respective size as described in histology reports. The reports described these initially rejected samples as between 2 mm to 20 mm in size, however, we observed samples with no visible tissue to 1 mm. Based on this first cohort we decided to only recruit patients with surgical tissue available. Four samples were never received from STH histology. With only 59 samples remaining, we requested to access a further 40 samples from the ReSoLuCENT study.

Post extraction analytical quality checks for the 99 clinical samples included quantification via Nanodrop, qPCR, and Qubit. In addition to this, two methods of DNA quality assessment were utilised; the DIN score was calculated using the Agilent's TapeStation 2200 and the QFI score based on the Qubit concentration and qPCR as described in section 2.2.3. A_{260} / A_{280} ratios for all samples as determined using the NanoDrop were >1.88. Samples with <5 ng DNA yield as assessed by Qubit or QPCR had the extraction process repeated. 7 samples failed this QC step and it is important to note DNA yield was not reflective of tissue size with 3 surgical samples, all with a tissue surface area of >100 mm^2 failing at this stage.

4.3 Sample quantity and quality

As with the validation samples described in the previous chapter, Nanodrop quantification of the samples was significantly greater than that of the Qubit. In addition, the qPCR method of quantification is significantly lower than the Qubit. Both of these differences are highlighted in the non-overlapping 95% confidence intervals shown in Table 4.1 Bivariate Pearson correlation coefficients (r) demonstrated a strong linear correlation between the methods of measurement; NanoDrop vs. Qubit 0.835 ($p < 0.01$), qPCR vs. Qubit 0.823 ($p < 0.01$), and NanoDrop vs. qPCR 0.590 ($p < 0.01$). When assessing the two different methods of determining DNA quality via QFI and the DIN score, Pearson correlation coefficient again demonstrated a strong linear correlation between the methods of quality assessment 0.759 ($p < 0.01$) shown in Table 4.2. When assessing our cohort of samples using the DIN score 45% (n = 41) fell below the lower limit of the 95% CI (3.3). Similarly utilising the QFI score, 49% (n = 45) fell below the lower limit of the 95% CI (24.8%). The range of yield and quality observed by the various methods of measurement all agree there is a large degree of variability across the samples.

Table 4.1. Quantification range and correlation of various methods of DNA quantification.

	NanoDrop (r)	Qubit (r)	Range (ng/µl)	Mean (ng/µl)	95% CI
NanoDrop	1	0.835	(17 – 1706)	446.0	381.6 – 510.4
Qubit	0.835	1	(1.6 – 780)	131.8	103.0 – 159.8
qPCR	0.590	0.823	(0 – 434)	48.8	32.8 – 64.9

Pearson's Correlation of quantification values (r) values. Statistical evidence for a linear relationship was observed between all methods of quantification ($p < 0.01$).

Table 4.2. Range and Correlation DIN score vs. QFI score.

	(r)	P value	Range	Mean	95% CI
DIN Score	0.759	P < 0.01	1.3 – 6.7	3.6	3.3 – 3.9
QFI Score			0.1 – 101 %	29.8 %	24.8 – 34.6

Pearson's Correlation of quantification values (r) values. Statistical evidence for a linear relationship was observed between DIN and QFI scores.

4.4 Variant Screening

4.4.1 Library preparation

92 samples passing initial criteria had library preparations undertaken using the methods described in section 2.2.4. Starting DNA material ranged from 1.6 ng to 300 ng as determined by qPCR, or 64 ng to 5171 ng by Qubit. Libraries were prepared for two of the samples (NF1-230 and NF1-229) with insufficient DNA, as quantified by qPCR. However, utilising the Qubit concentrations these should have been adequate to sequence, with 2015 ng and 726 ng of starting material respectively. These and two other samples (NF1-235 and NF1-236) were rejected at the pre-hybridisation QC step, based on poor recovery, Table 4.3.

Table 4.3. Samples rejected pre-hybridisation

Sample	Starting amount		QFI	DIN	Pre hyb (ng)
	Qubit (ng)	qPCR (ng)			
NF1-229	726	7.8	1.1	2.8	1.3
NF1-230	2015	1.6	0.08	1.5	3.2
NF1-235	1510	11.1	0.74	1.7	2.4
NF1-236	2155	50.0	2.3	2.6	19.6

Agilent's minimum recommended DNA yield to continue to the hybridisation is 500 ng

NF1-229 and NF1-236 had surplus tissue available; DNA was re-extracted using more starting tissue and library preparation repeated. Whilst low pre-hybridisation was still observed we continued with the hybridisation, in order to test lower limits of input material for this step. 34% (n = 29) of our samples fell below Agilent's minimum recommendation of 500 ng. Of these, 6 samples had insufficient post-hybridisation yield for sequencing, shown in Table 4.4. Previously rejected samples (NF1-229 and NF1-236) failed again, in despite of increase starting material. All cases with >170 ng resulted in sufficient post-hybridisation yield to sequence.

Table 4.4. Samples which demonstrated the lowest pre-hybridisation yield

Sample ID	Starting amount		QFI (%)	DIN	Pre-hyb (ng)	Post-hyb (pmol/l)
	Qubit (ng)	qPCR (ng)				
NF1-236	5171	120.0	2.3	2.6	62.88	16
NF1-226	1489	150.0	10.1	2.5	74.4	255
NF1-215	306	130.6	42.7	3.8	92.16	5050
NF1-213	220	62.5	28.3	2.7	110.28	5220
NF1-229	2342	25.0	1.1	2.8	133.2	572
NF1-174	173	20.7	11.9	2.5	135.6	6470
NF1-219	1546	50.0	3.3	2.1	151.2	231
NF1-231	335	100.0	29.8	2.2	170.4	4880
NF1-238	474	173.5	36.6	3.3	219.6	7780
NF1-237	588	105.5	17.9	4.6	223.2	8510

Samples shown in red resulted in insufficient post- hybridisation yield to sequence.

4.4.2 Patient Samples and Demographics

The age at biopsy for the remaining cases ranged from 20 to 86 years with a mean age of 62.5 years (95% CI, 60.1 to 65.0). Unfortunately, not all fields of patient data collected for the ReSoLuCENT trial correlated with data fields for our study, and some areas had missing data. Fields for which ReSoLuCENT data differed included; smoking status excluded and several NSCLC subtypes were classified as not otherwise specified (NOS). As the samples were provided in an anonymised manner mortality status was unobtainable. However, data we could collate included; gender, age and stage at biopsy, and whether this was a biopsy of primary tumour or metastatic tissue, Table 4.5. Disease stage was translated from TNM grading following the American Joint Committee on Cancer Staging Manual (Edge and Compton, 2010).

Table 4.5. Summary of patient pathology, gender, stage and smoking status

Variable	Cases (n)	NSCLC subtype		
		ADC (n)	SQCC (n)	NOS (n)
Gender				
Male	45	23	17	5
Female	41	25	13	3
Disease Stage				
Stage 1	4	1	3	0
Stage 2	25	16	9	0
Stage 3	42	23	15	4
Stage 4	15	8	3	4
Smoking Status				
Current or ex	43	25	18	0
Never Smoked	9	8	1	0
Unknown	34	15	11	8
Biopsy Site				
Primary	69	40	26	3
Metastatic	17	8	4	5

NOS (not otherwise specified)

4.4.3 Pathological Review

Pathological review of our cases revealed tumour content from 0 to 90%, based on nuclei ratio to normal non-tumour nuclei, shown in Table 4.6. Based on this, four of our remaining cohort had no visible tumour content. NF1-234 was a surgical resection, NF1-203 and NF1-204 were bronchial biopsies, and NF1-232 was a lymph node biopsy. As patients had been diagnosed based on these samples and pathological review is subjective, we decided to continue with analysis of these cases. For future reference all NF1-1XX samples were recruited in this study and were surgical resections of primary tumours. NF1-2XX cases originated from the ReSoLuCENT study and are comprised of surgical resections, bronchoscopies, or metastatic biopsies.

Table 4.6. Pathology review of tomour content in all 86 sequanced samples

Sample ID	Tumour Content (%)	Sample ID	Tumour Content (%)	Sample ID	Tumour Content (%)	Sample ID	Tumour Content (%)
NF1-107	30	**NF1-151**	70	**NF1-174**	50	**NF1-214**	60
NF1-115	50	**NF1-152**	60	**NF1-175**	30	**NF1-215**	30
NF1-117	50	**NF1-154**	50	**NF1-176**	10	**NF1-216**	40
NF1-119	50	**NF1-155**	50	**NF1-177**	70	**NF1-217**	60
NF1-120	30	**NF1-156**	50	**NF1-178**	30	**NF1-218**	60
NF1-122	60	**NF1-157**	50	**NF1-179**	50	**NF1-220**	40
NF1-126	50	**NF1-158**	70	**NF1-180**	60	**NF1-221**	60
NF1-130	40	**NF1-159**	90	**NF1-181**	60	**NF1-222**	70
NF1-132	60	**NF1-160**	30	**NF1-182**	50	**NF1-223**	10
NF1-134	40	**NF1-161**	30	**NF1-183**	30	**NF1-224**	60
NF1-136	40	**NF1-162**	20	**NF1-184**	30	**NF1-225**	70
NF1-138	10	**NF1-163**	30	**NF1-185**	70	**NF1-227**	50
NF1-139	30	**NF1-164**	30	**NF1-201**	70	**NF1-231**	50
NF1-140	80	**NF1-165**	40	**NF1-203**	0	**NF1-232**	0
NF1-141	10	**NF1-166**	50	**NF1-204**	0	**NF1-233**	70
NF1-142	30	**NF1-167**	20	**NF1-205**	10	**NF1-234**	0
NF1-144	10	**NF1-168**	50	**NF1-206**	50	**NF1-237**	60
NF1-145	80	**NF1-169**	30	**NF1-207**	40	**NF1-238**	80
NF1-146	90	**NF1-170**	40	**NF1-210**	50	**NF1-239**	20
NF1-147	60	**NF1-171**	40	**NF1-211**	20	**NF1-240**	50
NF1-148	70	**NF1-172**	50	**NF1-212**	80		
NF1-150	60	**NF1-173**	60	**NF1-213**	30		

4.4.4 Sequencing Metrics

To ensure the sequencing of the 86 cases and GIAB had passed initial criteria for downstream analysis, we considered the following sequencing metrics and QC checks. Cluster density for the pooled samples was 926 K/mm^2, this demonstrated initial sample dilution to 11 pM for HiSeq loading was within recommended limits for cluster generation (800–1100 K/mm^2). Illumina Q-scores showed 97% of clusters passing the filter were >Q30. This was generated from 324 million independent reads which resulted in a total yield of 70.4 Gbases. We used the de-multiplexed sample read percentage to confirm equimolar pooling. Individual samples ranged from 0.5-2.0% shown in Figure 4.1.

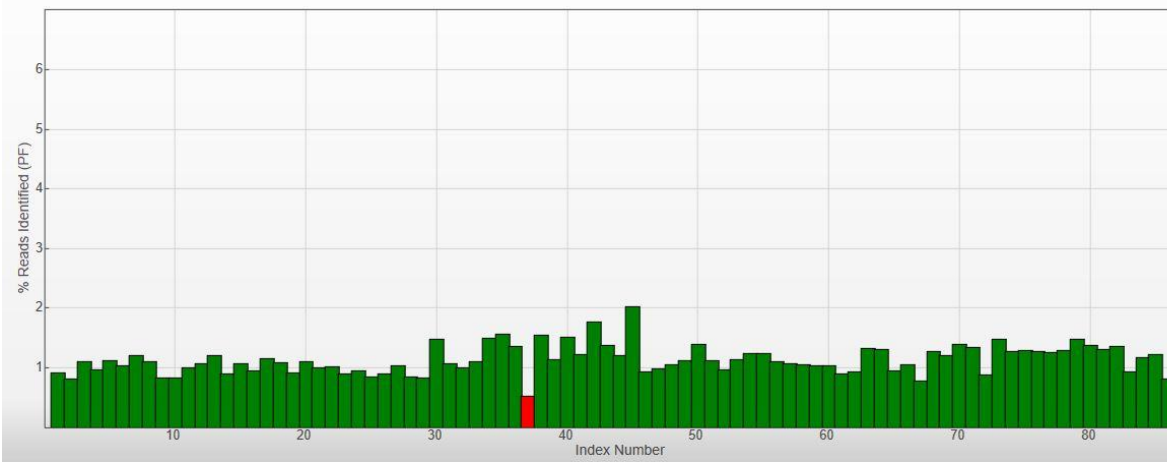


Figure 4.1. Percentage of reads per sample. Lowest sample NF1-165 highlighted in red. Range 0.53 - 2.03, mean 1.12% per sample.

4.4.5 Assessment of Sequencing Depth

As described in the previous chapter, a minimum read depth of 300 X is required to call variants in a sample with 6% tumour content based on diploid genome, with a specificity of 100% and 90%, using a 2.1% AF cut off or a 14.5 read cut off limit respectively. After discarding optical duplicates the average read depth for *NF1, EGFR, BRAF and KRAS* was 1650 X, ranging from 230 X to 3200 X, (95% CI, 1530 to 1771). These data demonstrate multiplexing 87 samples with a 275 kb capture region is adequate for deep coverage in order to confidently call low AF variants for this study. The average on target reads across all samples was 76%, ranging from 54% to 88%, (95% CI, 74.9 to 77.0). NF1-233 was the only sample to have <500 X mean coverage, but as the tumour content was 70% and we still had an average of 230 X, any ubiquitously expressed somatic mutation should be observed at higher than 3% AF. Whilst NF1-165 took the lowest percentage of reads from the pooled samples at 0.53% it still demonstrated a mean coverage of 1107 X.

4.5 Variant Identification and Annotation

Following de-multiplexing, the individual BCL files were run through the bioinformatics pipeline, as described in section 2.2.5. The workflow for variant annotation or exclusion is shown in Figure 4.2.

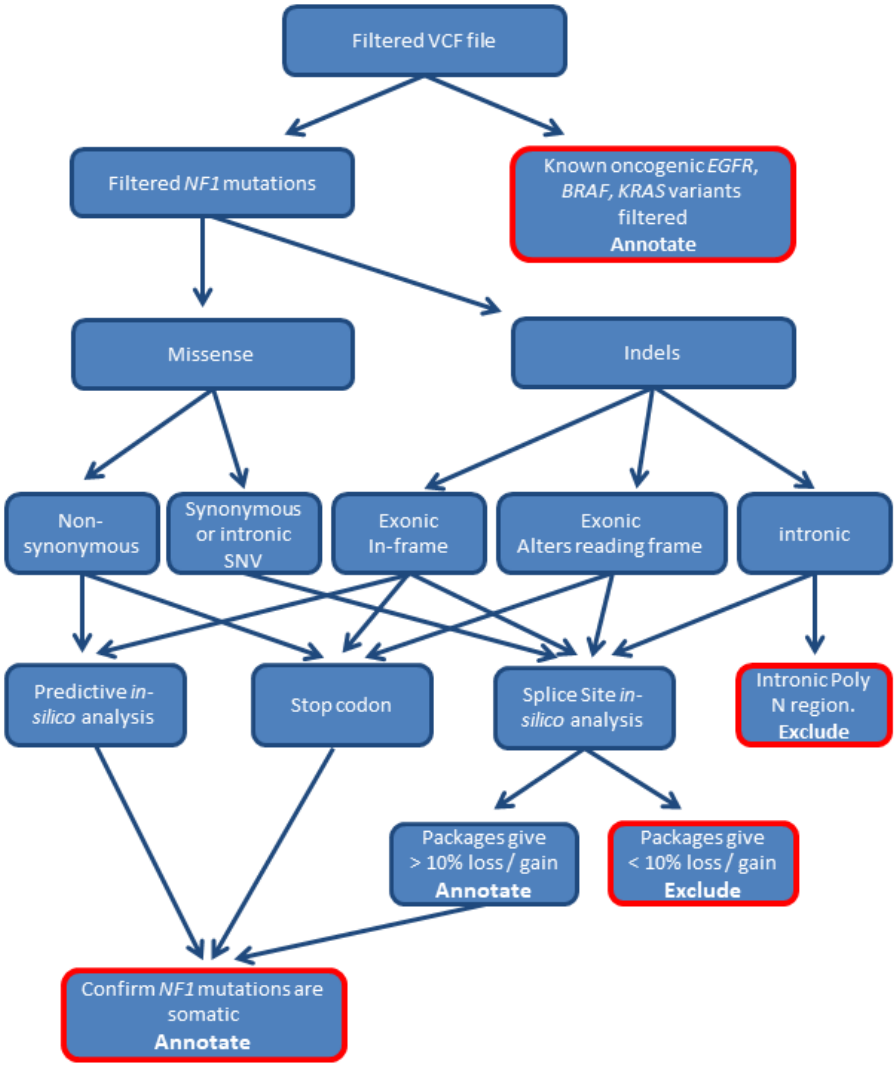


Figure 4.2. Overview of NGS variant annotation or exclusion work-flow.

4.5.1 MAPK Oncogenic Variant Identification

We used the VCFs outputted by VarScan2, described in section 2.2.5 and filtered for *EGFR*, *BRAF*, and *KRAS* regions. Variant analysis of the 86 samples using the workflow shown in Figure 4.2 and the criteria described in section 2.2.5.3 resulted in the identification of known oncogenic mutations in 29% (n = 25) of the samples. This included *EGFR* mutations in 8% (n = 7) of the cohort, shown in Table 4.7. *KRAS* mutations were identified in 17% (n = 15), shown in Table 4.8. Finally, *BRAF* mutations were identified in 3% (n = 3) of the cohort and are shown in Table 4.9. These oncogenic mutations were all mutually exclusive in all cases.

Table 4.7. *EGFR* oncogenic mutations

Sample ID	Allelic Depth	Allelic Fraction (%)	RS number	cDNA	Protein
NF1-119	192	16.77	rs121434568	c.2573T>G	p.Leu858Arg
NF1-142	1806	65.91	rs121434568	c.2573T>G	p.Leu858Arg
NF1-151	510	34.25	rs727504233	c.2236_2250del GAATTAAGAGAA GCA	p.Glu746_Ala 750del
NF1-157	434	29.21	rs121434568	c.2573T>G	p.Leu858Arg
NF1-164	270	18.54	rs121913438	c.2240_2257del TAAGAGAAGCAA CATCTC	p.Leu747_Pro 753delinsSer
NF1-170	406	29.59	rs121434568	c.2573T>G	p.Leu858Arg
NF1-205	12	2.88	rs727504232	c.2253_2276del ATCTCCGAAAGCC AACAAGGAAAT	p.Ser752_Ile 759del

All *EGFR* variants such as exon 21 substitutions and exon 19 deletions as seen in our cohort are reported as oncogenic (Lynch et al., 2004). Both exon 21 and exon 19are recognised predictive biomarkers of response to target specific tyrosine kinase inhibitors. The *EGFR* variants from our study shown in Table 4.7 were more prevalent in ADC cases (7/7), females (5/7), whilst we only had 9 known never smokers in our cohort of 86, 3 of them displayed oncogenic *EGFR* variants. All variants in these patients displayed high variants read depths with the exception of NF1-205.

Whilst NF1-205 was below the detection criteria of >3% AF and 15 variant reads, the chance of a deletion of this size being observed in multiple independent reads is below that of a SNV. The patient's clinical information had no reports on *EGFR* status. It did show the female patient with stage 3 disease was treated with Erlotinib in 2005, with a response being observed. The low pathology assessment of 10% tumour content also supports the 6% tumour contents based on a 3% AF, in a diploid genome. This does suggest the *EGFR* result obtained is correct and not a false positive.

For the patients with *EGFR* p.Leu858Arg variants, both NF1-119 and NF1-157 had the details reported in their clinical diagnostic history. The cases were both female and never-smokers. NF1-119 had stage 2 disease and suffered disease recurrence 3 years after surgery. The patient was treated with Gefitinib for a year before disease progression. NF1-157 had stage 3 disease which was not fully excised. The patient responded to Gefitinib for 4 years before disease progression. NF1-142 and NF1-170 had not been tested for *EGFR* variants, both were males with stage 2 disease, no disease recurrence has been reported for 2 and 1 years respectively. The cases with exon 19 deletions, NF1-151 and NF1-164 had the variants reported in their clinical history. Both cases were female, former smokers, with stage 2 disease. NF1-151 had disease recurrence after 1 year, Gefitinib treatment is ongoing with stable disease. NF1-164 had disease recurrence after 2 years; treatment with Gefitinib was stopped after 2 cycles.

Table 4.8. *KRAS* oncogenic mutations

Sample ID	Allelic Depth	Allelic Fraction (%)	RS number	cDNA	Protein	NSCLC subtype
NF1-107	669	34.08	rs121913530	c.34G>T	p.Gly12Cys	ADC
NF1-130	771	47.68	rs121913530	c.34G>T	p.Gly12Cys	SQCC
NF1-132	389	15.58	rs121913529	c.35G>A	p.Gly12Asp	SQCC
NF1-163	186	8.35	rs121913529	c.35G>T	p.Gly12Val	SQCC
NF1-165	203	19.08	rs121913529	c.35G>T	p.Gly12Val	ADC
NF1-173	333	17.11	rs121913529	c.35G>C	p.Gly12Ala	ADC
NF1-175	299	25.53	rs121913530	c.34G>T	p.Gly12Cys	ADC
NF1-183	141	8.32	rs121913530	c.34G>T	p.Gly12Cys	ADC
NF1-184	172	11.42	rs121913530	c.34G>C	p.Gly12Ala	SQCC
NF1-185	323	21.55	rs121913530	c.34G>T	p.Gly12Cys	ADC
NF1-206	336	23.40	rs121913529	c.35G>T	p.Gly12Val	ADC
NF1-207	270	30.86	rs121913530	c.34G>T	p.Gly12Cys	ADC
NF1-210	135	9.04	rs121913530	c.34G>T	p.Gly12Cys	ADC
NF1-218	159	23.18	rs121913530	c.34G>T	p.Gly12Cys	ADC
NF1-231	87	5.31	rs121913530	c.34G>T	p.Gly12Cys	ADC

AF range from 5.31 to 47.68%. Female patients highlighted in red, males in black

KRAS mutations such as, p.Gly12, as seen in our cohort are oncogenic, with compelling evidence to its role in many cancers and for being the most prevalent driver mutation in ADC (Garrido et al., 2017). Our cohort displayed *KRAS* mutations in 23% of the ADC cases, which is reflective of what has been reported in European cohorts (Boch et al., 2013). However, we have a larger than reported prevalence in SQCC cases, 13% in total compared to the 6% Boch and colleagues reported. It is important to note that although the non-specified subtype cases did not display any *KRAS* variants, these would still have influence over the final prevalence observed in our cohort. Whilst smoking status is not known for the ReSoLuCENT samples, 90% of the *KRAS* positive patients recruited to our *NF1* in NSCLC study were known smokers. Of the patients with *KRAS* positive disease, 60% were female (n = 9) and 40% (n = 6) male. *KRAS* mutations are not routinely tested for diagnostically, as they are not currently predictive for any treatments, therefore we cannot correlate our results with clinical diagnostic data.

Table 4.9. *BRAF* oncogenic mutations

Sample ID	Allelic Depth	Allelic Fraction (%)	RS number	cDNA	Protein	NSCLC subtype
NF1-148	572	28.57	rs121913348	c.1391G>T	p.Gly464Val	SQCC
NF1-174	367	18.62	rs121913355	c.1406G>T	p.Gly469Val	ADC
NF1-237	839	33.49	rs1057519720	c.1405_1406 delGGinsTT	p.Gly469Leu	SQCC

AF range from 18.6 to 33.5%. Female patient highlighted in red, males in black

These *BRAF* mutations such as exon 11 SNV and di nucleotide variants leading to p.Gly464Val, p.Gly469Val, and p.Gly469Leu as seen in our cohort, are potential drivers of the MAPK pathway (Gautschi et al., 2013) *BRAF* p.Gly469 is the most prevalent locus for substitutions in this gene, with Glycine to Valine being the most prevalent substitution. p.Gly464Val is observed to a lesser extent in NSCLC. p.Gly469 mutations are observed in 1.5% of ADC cases and <1% of SQCC cases (Campbell et al., 2016). Our cohort displayed *BRAF* mutations in <2% (n = 1) of the ADC cases and in 6% (n = 2) SQCC cases. It is important to note NF1-237 was reported in the VCF as c.1405G>T and c.1406G>T, as these were on the same allele they would have resulted in a p.Gly469Leu substitution. As with *KRAS*, *BRAF* is not routinely tested for diagnostic purposes

4.5.2 *NF1* Variant Filtering, Identification, and Annotation

Unlike known hotspots for the oncogenic gain of function mutations which can be targeted, *NF1* potential loss of function mutations are sporadic in nature. As described in section 2.2.5.3 we employed VCFMiner to filter the individual VCF files. We excluded all GIAB only variants. After which we limited the VCF files to the defined genomic regions. All variants with >1% population frequency based on gnomAD r2.0.2 were excluded, along with variants demonstrating <15 reads and with <3% AF. At this point we had potential *NF1* variants in 99% of the samples (n = 85). We then used the workflow shown in Figure 4.2 to further filter variants. 1 of 2 specific intronic poly N indels was observed at the same genomic coordinates in all but 1 sample. c.730+32dupT was reported in 85% (n = 73) of all cases, or c.730+32delT in 15% (n = 12) of all cases. These were reported to be deletions or

insertions in an intronic 16 polyT region. All observed at AF of 10% to 38%. Other recurring poly N indels included c.61-14delT observed in 7 % (n = 6) and c.1186-13delT in 2% (n = 2), both deletions were in a region of a 10 polyT repeats, AF ranged from 3% to 5%.

4.5.3 Pseudogene Matching Variants

We flagged 10 patients with potential *NF1* pseudogene variants using the pseudo list as described in section 3.3, variants shown in Table 4.10. Two of these variants were of great interest as both had been reported in Neurofibromatosis patients. The variant observed in NF1-120, c.1466A>G, (p.Tyr489Cys) had been described as a germline variant in two related patients, both of whom also carried somatic mutations of *NF1* (Laycock-van Spyk et al., 2011). The second was c.3721C>T, (p.Arg1241*) observed in NF1-168 and reported as a somatic mutation in a Neurofibromatosis patient (Upadhyaya et al., 2004). Based on these reports both were classified as pathogenic. One particular variant c.3114-7T>C was observed in 6 different samples at low AF. This variant was recorded on dbSNP from the Exome Aggregation Consortium with an unassessed clinical significance. Another intronic variant in NF1-146, c.4725-13A>C was flagged as potential pseudogene origin. This had not previously been reported. All 7 of the intronic variants had no predicted effect on splicing if originating from the functional gene. The final potential pseudogene variant c.1018T>G, (p.Ser340Ala) observed in NF1-220 had not previously been reported.

Table 4.10. *NF1* pseudogene matching variants

Sample ID	Allelic Depth	Allelic Fraction (%)	RS number	cDNA	Protein
NF1-119	36	4.01	rs754931967	c.3114-7T>C	-
NF1-120	78	9.55	rs137854557	c.1466A>G	p.Tyr489Cys
NF1-146	28	6.26	N/A	c.4725-13A>C	-
NF1-157	83	7.76	rs754931967	c.3114-7T>C	-
NF1-168	427	14.65	rs137854562	c.3721C>T	p.Arg1241*
NF1-211	35	3.19	rs754931967	c.3114-7T>C	-
NF1-220	23	3.07	N/A	c.1018T>G	p.Ser340Ala
NF1-220	40	13.07	rs754931967	c.3114-7T>C	-
NF1-225	20	3.14	rs754931967	c.3114-7T>C	-
NF1-232	15	4.50	rs754931967	c.3114-7T>C	-

Variants in NF1-120 and NF1-168 reported in Neurofibromatosis patients and classed as pathogenic

4.5.4 *NF1* Promoter Variants

We identified 5 potential mutations upstream of the *NF1* transcription start site shown in Table 4.11. These have the potential to effect transcription of *NF1* mRNA and therefore cause loss of function. No reports in the literature were found, but 4 variants were reported on dbSNP as indicated by the RS number. NF1-155 did have a corresponding entry on ClinVar and was deemed of uncertain clinical significance.

Table 4.11. *NF1* mutations upstream of the transcription start site.

Sample ID	Allelic Depth	Allelic Fraction (%)	RS number	cDNA
NF1-132	639	46.27	rs17879128	c.-459A>C
NF1-148	273	19.63	rs922504932	c.-356C>T
NF1-155	34	60.71	rs886052790	c.-209C>A
NF1-172	54	6.38	N/A	c.-93C>G
NF1-220	186	46.62	rs144759836	c.-468A>T

4.5.5 Potential *NF1* loss of function variants

After filtering all the previously mentioned variants, 20% (n = 17) of cases displayed 22 potential *NF1* somatic mutations, shown in Table 4.12. 5 cases displayed multiple *NF1* variants. The total variants included; 9 non-synonymous (2 = non-sense, 7 = missense), 6 synonymous, 3 indels (2 = deletions, 1 insertion), 1 exonic di-nucleotide substitution, and 3 exonic SNV.

Table 4.12. *NF1* mutations identified

Sample ID	Variant Depth	Allelic Fraction (%)	cDNA Variant	Protein
NF1-115	208	24.21	c.3118A>G	p.Lys1040Glu
NF1-146	536	47.86	1392+5_1392+6del GAinsTT	-
NF1-147	1011	58.98	c.3468C>T	p.Asn1156Asn
NF1-150	63	6.65	c.5269-14C>G	-
NF1-163	75	3.03	c.1897G>C	p.Asp633His
NF1-163	158	44.26	c.-22G>C	-
NF1-168	201	11.35	c.5785G>T	p.Glu1929*
NF1-170	111	5.70	c.3825C>T	p.Phe1275Phe
NF1-178	576	53.33	c.7595C>T	p.Ala2532Val
NF1-180	897	52.95	c.7245delA	p.(Leu2416Tyrfs*2)
NF1-201	109	9.00	c.435C>A	p.Leu145Leu
NF1-205	20	4.07	c.4311G>A	p.Arg1437Arg
NF1-210	51	4.68	c.169G>T	p.Gly57Cys
NF1-217	251	13.01	c.3634delG	Val1212Serfs*3
NF1-217	39	8.92	c.7339G>T	p.Glu2447*
NF1-225	761	58.72	c.6781C>T	p.His2261Tyr
NF1-227	24	4.53	c.6148-2A>T	-
NF1-233	69	35.57	c.2681T>C	p.Phe894Ser
NF1-233	19	10.50	c.2178G>A	p.Val726Val
NF1-233	54	29.83	c.2675_2676insA	p.(Ser892Argfs*14)
NF1-238	117	7.21	c.3481C>G	p.Leu1161Val
NF1-238	40	6.23	c.3249C>G	p.Leu1083Leu

NF1 variants remaining post VCF filtering. AF ranges from 3.03 to 58.98% and variant read depths of 19 to 1011 X.

The potential somatic variants were then subject to searches through literature and databases to determine if any had been previously reported. With the exception of the previously mentioned potential pseudogene variant, no others were reported in the literature and no functional studies had been undertaken. 8 of the mutations were located in online databases, but the clinical significance of these variants remains to be elucidated. Of these 8, NF1-163 (c.-22G>C) had the highest frequency population data, with gnomAD r2.0.2 reporting it as 0.38%, NF1-147 (c.3468C>T) and NF1-178 (c.7595C>T) were < 0.069% and < 0.079% respectively. All remaining variants were < 0.0057%. 15 of the variants were novel with no evidence of been previously reported. A summary of the results and potential clinical significance are reported in Table 4.13.

As no evidence of functional studies had been reported, the 22 potential somatic mutations were subjected to *in-silico* analysis to determine the probability of causing loss of *NF1* function. All computational *in-silico* algorithms are described in section 2.2.5.3. All samples with null variants (non-sense or frameshift) (n = 4) were classified as high probability of causing loss of function, shown in Table 4.13. NF1-217 had two null variants; however it is not possible to determine if these are in *cis* or in *trans*.

Of the variants with the potential to affect mRNA splicing, NF1-150 (c.5269-14C>G) was the only variant to have three out of five of the *in-silico* splicing packages reporting a 10% reduction in its acceptor site. This has the potential to result in c.5446 becoming a more favourable acceptor site. We classified NF1-150 as having an unknown significance of causing loss of *NF1* function. This suggests the cases (n = 9) with potential splice site variants are likely benign and have a low probability of causing loss of function of *NF1*, shown in Table 4.13.

Table 4.13. Reported *NF1* variants reported and potential clinical significance

Sample ID	RS number	cDNA	Reported	Clinical Significance
NF1-120*	rs137854557	c.1466A>G	Laycock-van Spyk et. al 2011	Pathogenic
NF1-168*	rs137854562	c.3721C>T	Upadhyaya M et. al 2003	Pathogenic
NF1-115	N/R	c.3118A>G	ClinVar	Uncertain
NF1-150	N/R	c.5269-14C>G	N/R	N/R
NF1-163	N/R	c.1897G>C	N/R	N/R
NF1-178	rs148154172	c.7595C>T	dbSNP / ClinVar	Uncertain
NF1-210	N/R	c.169G>T	N/R	N/R
NF1-233	N/R	c.2681T>C	N/R	N/R
NF1-225	rs750869272	c.6781C>T	dbSNP / ClinVar	Likely benign
NF1-238	N/R	c.3481C>G	N/R	N/R
NF1-168	N/R	c.5785G>T	N/R	N/R
NF1-180	N/R	c.7245delA	N/R	N/R
NF1-217	N/R	c.3634delG	N/R	N/R
NF1-217	N/R	c.7339G>T	N/R	N/R
NF1-233	N/R	c.2675_2676insA	N/R	N/R
NF1-220*	N/R	c.1018T>G	N/R	N/R
NF1-146	rs587782851	1392+5_1392+6delGAinsTT	dbSNP / ClinVar	Uncertain
NF1-147	rs147955381	c.3468C>T	dbSNP / ClinVar	Uncertain
NF1-163	rs556823296	c.-22G>C	dbSNP / ClinVar	Likely benign
NF1-170	N/R	c.3825C>T	N/R	N/R
NF1-201	N/R	c.435C>A	N/R	N/R
NF1-205	rs1060503896	c.4311G>A	dbSNP / ClinVar	Likely benign
NF1-233	rs369590240	c.2178G>A	dbSNP / ClinVar	Likely benign
NF1-227	N/R	c.6148-2A>T	N/R	N/R
NF1-238	N/R	c.3249C>G	N/R	N/R

Not previously reported (N/R),*Potential pseudogene origin. The two potential pseudogene variants reported as pathogenic in the literature. Sample IDs highlighted in red (n = 5) have strong evidence of loss of function based on the variant identified alone and are classified as high probability. NF1-217 displaying 2 null variants. Samples in green (n = 10) have little or no supporting *in-silico* evidence that the variants affect mRNA splicing and are classified as likely benign. Samples in black (n = 10) have *in-silico* analysis which suggests the variants could potentially cause *NF1* loss of function and are classed as unknown significance.

The 7 remaining non-synonymous variants were then analysed through 7 different *in-silico* algorithms to help predict potential loss of function, shown in Table 4.14. For the variant to be considered unknown significance of causing loss of function >65% of the tools had to be in agreement of damaging consequences. If >65% demonstrated neutrality the variants were classified as likely benign. NF1-178, NF1-210, NF1-225, and NF1-238 did not yield any result through the PANTHER package. Based on these criteria; 5 of the missense variants fell into the unknown significance bracket and 2 into the likely benign bracket shown in Table 4.14.

Table 4.14. *in-silico* analysis of *NF1* non-synonymous missense variants

Sample ID	PhD-SNP	PANTHER	SNPs & GO	PMut	PROVEAN	SIFT	PolyPhen2
NF1-115	Disease	Disease	Disease	Disease	Deleterious	Damaging	Possibly damaging
NF1-163	Neutral	Neutral	Disease	Neutral	Neutral	Tolerated	Possibly damaging
NF1-178	Neutral	N/A	Disease	Neutral	Neutral	Tolerated	Possibly damaging
NF1-210	Neutral	N/A	Disease	Neutral	Deleterious	Damaging	Possibly damaging
NF1-233	Disease	Disease	Disease	Disease	Deleterious	Damaging	Possibly damaging
NF1-225	Disease	N/A	Disease	Disease	Deleterious	Damaging	Possibly damaging
NF1-238	Disease	N/A	Disease	Disease	Deleterious	Damaging	Possibly damaging

The final prediction as indicated by 7 independent in-*silico* analysis tools. Samples in Red have variants predicted to be damaging >65 % of the packages used.

4.5.6 *NF1* Germline Sequencing

To determine if the 22 variants observed were somatic in nature or germline, we sequenced matched blood samples. Unfortunately, we did not have paired blood for NF1-168, NF1-170, and NF1-233. Library preparation and sequencing protocols used were as previously described with the following exceptions; we excluded the qPCR quantification step, and the samples were sequenced on the MiSeq. All DIN values were >8.3 and 200 ng of starting material was used. The sequencing metrics demonstrated 92.9% of reads were >Q30 and had a cluster density of 1152K/mm^2. For the tumour samples with paired blood available, we repeated the bioinformatics pipeline using the actual paired sample. The same variant filtering was applied with the exception that we did not have to remove GIAB variants.

Analysis determined both NF1-146 variants were germline, heterozygote in origin, alongside NF1-147, NF1-163 (c.-22G>C), NF1-178, and NF1-225. When comparing these back to the loss of function prediction, all but NF1-225 fall into the low probability bracket. This can also be seen looking at the AF of the tumour samples, all germline variants are found between 44% to 59% AF, shown in Table 4.12. Since all germline variants fall into this AF bracket, it suggests that NF1-168 and NF1-170 are somatic in nature, with the AF being 11.35% and 5.70% respectively. All variants found within the protein coding regions of *NF1* are shown in Figure 4.3.

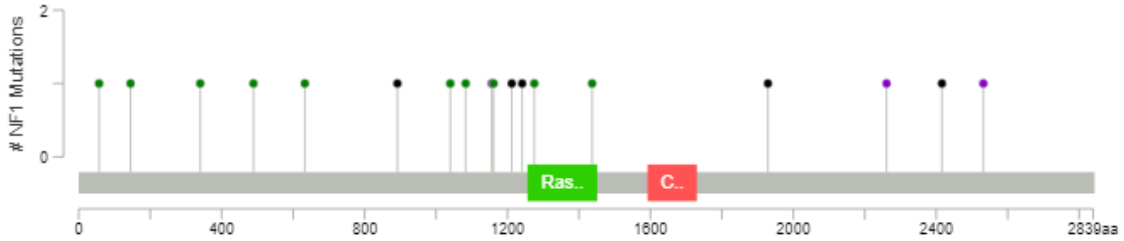


Figure 4.3. Illustration of variants and loci within *NF1* protein. Figure includes somatic variants, three possible pseudogene variants, and germline variants. Black circles; (n = 5) truncating variants. Green (n = 10); missense variants. Purple (n = 3); germline variants.

4.5.7 Prevalence of Identified Variants within Tumour Content

To further investigate if the variants we observed were expressed ubiquitously throughout the tumour we compared the sequencing AF against expected frequency. The expected frequency was based on the assumptions the tumour is diploid and tumour content is correct. As the pathology tumour content is subjective and there is variation in the sequencing AF, we set ±30% thresholds of the expected AF.

48% (n = 12) of the 25 oncogenic variants fell within ±30% of the expected frequency, shown in Table 4.15. The samples falling into the diploid bracket suggests a heterozygote genotype, consisting of wildtype and mutant allele, with this genotype being ubiquitously expressed throughout the tumour. 28% (n = 7) of the mutations fell below the 30% threshold. These frequencies are suggestive of intra-tumour heterogeneity with the oncogenic mutations being prevalent in only a percentage of the tumour. Whilst copy number loss is also a factor to consider, it is less prevalent in *BRAF*, *KRAS*, and *EGFR* and only seen in 14%, 13%, and 8% of cases respectively. 25% of the variants (n = 6) demonstrated a >30% increase in AF. This could be suggestive of copy number gain of the variant allele, the variant being homozygote, or copy number loss of the wildtype allele. Based on our data it is not possible to draw a conclusion on this. However, copy number gain is a common event for *BRAF*, *KRAS*, and *EGFR* in NSCLC and seen in 42%, 41.6%, 52% of all cases respectively.

.

Table 4.15. Observed allelic frequency of oncogenic mutations and tumour content

Sample ID	Gene	Tumour Content (%)	Expected variant AF (%)		Observed variant AF (%)
			<30%	>30%	
NF1-148	*BRAF*	70	24.5	45.5	28.57
NF1-174	*BRAF*	50	17.5	32.5	18.62
NF1-237	*BRAF*	60	21	39	33.49
NF1-119	*EGFR*	50	17.5	32.5	16.77
NF1-142	*EGFR*	30	10.5	19.5	65.91
NF1-151	*EGFR*	70	24.5	45.5	34.25
NF1-157	*EGFR*	50	17.5	32.5	29.21
NF1-164	*EGFR*	30	10.5	19.5	18.54
NF1-170	*EGFR*	40	14	26	29.59
NF1-205	*EGFR*	10	3.5	6.5	2.88
NF1-107	*KRAS*	30	10.5	19.5	34.08
NF1-130	*KRAS*	40	14	26	47.68
NF1-132	*KRAS*	60	21	39	15.58
NF1-163	*KRAS*	30	11.5	19.5	8.35
NF1-165	*KRAS*	40	14	26	19.08
NF1-173	*KRAS*	60	21	39	17.11
NF1-175	*KRAS*	30	10.5	19.5	25.53
NF1-183	*KRAS*	30	10.5	19.5	8.32
NF1-184	*KRAS*	30	10.5	19.5	11.42
NF1-185	*KRAS*	70	24.5	45.5	21.55
NF1-206	*KRAS*	50	17.5	32.5	23.40
NF1-207	*KRAS*	40	14	26	30.86
NF1-210	*KRAS*	50	17.5	32.5	9.04
NF1-218	*KRAS*	60	21	39	23.18
NF1-231	*KRAS*	50	17.5	32.5	5.31

Green sample ID indicates the observed AF was within ±30% of the expected heterozygote AF. Blue sample ID indicates the variant was below 30% of the expected. Red indicated the variant AF was greater than 30% of the expected.

Based on the AF observed and tumour content we have 18/25 samples in which the oncogenic mutations of the MAPK pathway appear to be expressed ubiquitously throughout the tumour in a clonal manner. The remaining 7 variants appear to be only expressed in certain sub-clonal populations of the tumour. Looking at the variants independently and comparing the subcolonal percentage of cases we observed similar results to the first 100 cases from TRAcking Cancer Evolution through therapy (Rx) (TRACERx) shown in Table 4.16 (Jamal-Hanjani *et al*.,2017).

Table 4.16. Percentage of patients with subcolonal MAPK driver variants in the first 100 patients of TRACERx in relation to this study (NF1 in NSCLC)

Gene	TRACERx (% of sub-clonal cases)	This study (% of sub-clonal cases)
KRAS	33	21
EGFR	29	20
BRAF	0	0

The same analysis as shown in Table 4.15 was repeated for the *NF1* variants, with the exception any cases with germline variants, were classified as 50% expected frequency shown in Table 4.17. All five germline variants fell within 10% of the expected frequency. After removing the germline variant from our cohort of *NF1* patients we were left with 20 *NF1* somatic variants, of this 80% (16/20) proved to be low level heterogenic, therefor just expressed in sub-clonal populations of the tumour analysed. Of the remaining somatic *NF1* variants 20% (4/20) were within ±30%, with only NF1-180 above this threshold, shown in Table 4.17. The three potential pseudogene variants were just below the -30% threshold, they were however within -40% of the expected AF. *NF1* has also been reported to be sub-clonal in the analysis of the first 100 lung cancer patients in the TRACERx study. The study reported 50% of NF1 variants were described as sub-clonal. When looking at the individual subtypes TRACERx reported 2/2 (100%) of the *NF1* variants in the SQCC population were sub-clonal, whereas 3/8 (38%) of the ADC cases were sub-clonal (Jamal-Hanjani et al., 2017).

Table 4.17. Observed allelic frequency of *NF1* variants and tumour content

Sample ID	Tumour Content (%)	Expected variant AF (%)		Observed variant AF (%)	Heterogeneity
		<30%	>30%		
NF1-115	50	17.5	32.5	4.21	Heterogenic
NF1-120*	30	10.5	19.5	9.55	Heterogenic
NF1-146**	90	35	65	48.05	Ubiquitous
NF1-147**	60	35	65	58.98	Ubiquitous
NF1-150	60	21	39	6.65	Heterogenic
NF1-163	30	10.5	19.5	3.03	Heterogenic
NF1-163**	30	35	65	44.26	Ubiquitous
NF1-168	50	17.5	32.5	11.35	Heterogenic
NF1-168*	50	17.5	32.5	14.65	Heterogenic
NF1-170	40	14	26	5.70	Heterogenic
NF1-178**	30	35	65	53.33	Ubiquitous
NF1-180	60	21	39	52.95	Ubiquitous
NF1-201	70	24.5	45.5	9.00	Heterogenic
NF1-205	10	3.5	6.5	4.07	Ubiquitous
NF1-210	50	17.5	32.5	4.68	Heterogenic
NF1-217	60	21	39	13.01	Heterogenic
NF1-217	60	21	39	8.92	Heterogenic
NF1-220*	40	14	26	13.07	Heterogenic
NF1-225**	70	35	65	58.72	Ubiquitous
NF1-227	50	17.5	32.5	4.53	Heterogenic
NF1-233	70	24.5	45.5	35.57	Ubiquitous
NF1-233	70	24.5	45.5	10.5	Heterogenic
NF1-233	70	24.5	45.5	29.83	Ubiquitous
NF1-238	80	28	52	7.21	Heterogenic
NF1-238	80	28	52	6.23	Heterogenic

Green sample ID indicates the observed AF was within ±30% of the expected heterozygote AF. Blue sample ID indicates the variant was below -30% of the expected. Red indicated the variant AF was greater than +30% of the expected. Black represents variants below the -30% expected. Samples were classified as heterogenic if the variant was expressed <30% of expected, as all variants above classed as ubiquitous. * Possible pseudogene origin. ** Germline in origin

Next we investigated possible relationships between these oncogenic variants observed in *BRAF*, *EGFR*, and *KRAS* and the potential loss of function variants in *NF1*. The oncoprint shown in Figure 4.4 illustrates variants in each case. Mutual exclusivity was observed for all oncogenic mutations in *BRAF*, *EGFR*, and *KRAS.* The same relationship was observed with *BRAF* and *NF1*. For the remaining we calculated the odds ratio to determine tendencies of co-occurrence or mutual exclusivity. The odds ratio was calculated as described in 2.2.6. *NF1* and *KRAS* showed a strong tendency towards mutual exclusivity with an odds ratio of 0.756, whilst *NF1* and *EGFR* showed a tendency of co-occurrence with a ratio of 2.23.

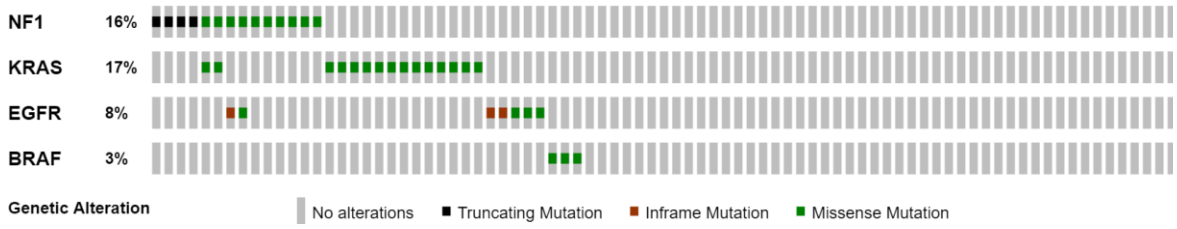


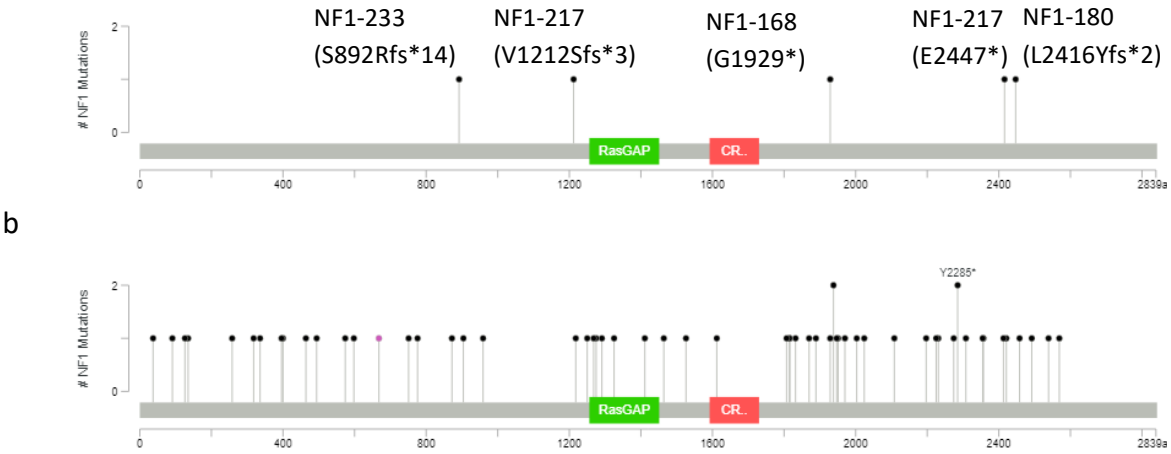
Figure 4.4. Oncoprint of all 86 NSCLC cases. Two samples shown to contain both *NF1* and *KRAS* mutations.

4.5.8 PAN-Lung Cancer TGCA Analysis

Our initial identification of 25 *NF1* variants in 20 samples was encouraging. However, many of the *in silico* prediction tools as described in section 2.2.5.3 for the *NF1* missense variants resulted in an outcome of being likely begin or unknown significance in relation to loss of NF1 function. As only the GAP domain and SEC-PH domain have had their crystal structure resolved, it was not possible to interpret functional effects of missense variants further using *in silico* tools. Our initial data demonstrated 5 *NF1* null variants in 4 cases (NF1-168, NF1-180, NF1-217, and NF1-233), shown in Table 4.12, which are highlighted in Figure 4.5 below. After taking into consideration the sub-clonal variants which are observed at low AF, only NF1-180 and NF1-233 appeared to be expressed clonally, demonstrated by their high AF in relation to tumour content, shown in Table 4.17. However, only NF1-233 would result in a truncation upstream of NF1s GAP domain. de Bruin *et al*. (2014) demonstrated that transfection of just NF1s GAP domain was enough to rescue sensitivity to erlotinib through deactivation of *KRAS* in *NF1* knock down models; therefore, it is not possible to predict the functional relevance of null variants downstream of the GAP domain. Out of all the cases with unknown significance variants only NF1-225 was clonal.

With only one *NF1* null variant clonally expressed case upstream of the GAP domain case NF1-233) in our study, we decided to investigate data from the PAN-Lung cancer trial which is accessible via The Cancer Genome Atlas. 979 samples (ADC = 498, SQCC = 481) had genetic and transcriptional data which could be accessed. Of the 979 patients 11% (111/979) displayed *NF1* variants, of which 59 were null variants resulting in NF1 truncation. All *NF1* variants were analysed as described in section 4.5.5 to predict potential loss of function, with the exception that only null variants downstream of the GAP domain were classed as high probability loss of NF1 function, as were 3 cases with homozygote *NF1* deletions. By only classing null variants downstream of NF1s GAP domain as high probability will ensure heterozygote loss of this active domain which negatively regulates *KRAS*. We also investigated *RASA1* variants which have been shown to increase both MAPK and PI3K pathway activity when co-occurring with *NF1* variants in cell based models (Hayashi et al., 2018). In silico analysis of *RASA1* was completed as described in section 4.5.5.

a



b

Figure 4.5. Lollipop illustration of NF1 variants a) 5 null variants in this studies cohort of patients. NF1-233 and NF1-217 are the only 2 cases to have a null variant downstream of the GAP domain. Only NF1-233 and NF1-180 are clonal and expressed ubiquitously throughout the tumour **b)** 59 null variants from Pan-Lung cancers TGCA data. 17 of which are downstream of the GAP domain and did not co-occur with *RASA1* variants. 16 of these null variants had a sequencing AF of ≥ 0.20. There were also 3 homozygote *NF1* deletions which were included in the NF1 high probability loss of function group. Whilst the lollipop plot shows more than 17 *NF1* variants downstream of the GAP domain, the *NF1* null variants that were not counted from the 17 co-occur with *RASA1* variants.

The total frequency of cases with *NF1* null variants from this study was 5% (4/86) compared with 6% (59/979) from the Pan-Lung study. Of the 59 *NF1* null variants from the Pan-Lung study 17 were downstream of the of the NF1s GAP domain and did not co-occur with *RASA1* variants. In total we had 10 cases with co-occurring *NF1* and *RASA1* variants which were ≥ 0.20 AF based on the PAN-Lung sequencing data.

We then grouped the Pan-Lung TGCA cases based on genotype, in addition to how we divided the *MAPK* oncogenic variants (*EGFR*, *BRAF*, *KRAS*) described in section 4.5.1, we also included *PIC3CA* and *PTEN* cases shown in Table 4.18. Furthermore, we examined *RASA1* which is another member of the GTPase family of proteins. Only 2 cases from the Pan-Lung dataset displayed *AKT* E17K variants and one of these also had a *PIK3CA* E542K substitution, therefore we did not create a *AKT* group. Of the 111 cases with *NF1* variants, 6 also displayed co-occurring *KRAS* variants, and 2 displayed co-occurring *EGFR* variants, these were included in the *EGFR* and *KRAS* groups. 13 cases demonstrated co-occurrence of *NF1* and *RASA1* co-occurring variants.

Unlike TRACERx which used multi-region sampling to generate intratumor heterogeneity status, our study and the data generated in the TGCA cohort only used single sampling, therefore the heterogeneity % is only based on the section of the tumour analysed in this study. For the intention of this study this will be utilised in the next chapter to exclude cases with low heterogeneity, as these will affect the gene expression signatures, it is not respective of overall tumour heterogeneity. The pathology reports of Pan-Lung TGCA data are unavailable for every sample, but their pathology criteria was inclusion of only samples with >60% tumour content. In table 4.18 below we have excluded samples with < 0.2 AF based on the sequencing data. Based on 60% tumour content variants ≥ 0.20 AF would translate to the variants being expressed in a minimum of 66% of the region sampled, using the method described in section 4.5.7.

Table 4.18. Oncogenic *MAPK, PIC3CA, NF1, and RASA1* variants in the Pan-Lung cohort of patients

Gene	No of cases	Number of ADC / SQCC	Number of cases with ≥ 0.20 AF
KRAS (G12, G13,and Q61)	(152, 10, 2)	163 / 4	136
EGFR (L858R, exon 19 indels)	(22, 35)	55 / 2	33
BRAF (V600E, G469, G466, G464)	(9, 9, 8, 2)	23 / 5	21
NF1 and *RASA1*	13	3 / 10	10
NF1 H (NF1 null variants upstream of the GAP domain)	20	10 /10	19
NF1 M (Unknown significance)	52	30 / 22	30
NF1 L (likely benign)	18	7 / 11	12
RASA1 H (RASA1 null variants)	26	8 / 18	18
RASA1 M(Unknown significance)	3	0 / 3	1
RASA1 L (likely benign)	2	0 / 2	2
MAPK PI3K negative	387	185 / 202	NA
PIK3CA (E545K, E542K, H1047R, H1047L)	(17, 17, 3, 0)	7 / 31	26
PIK3CA (amplification)	153	5 / 148	NA
PTEN (homozygote deletion)	16	2 / 14	NA
	Total 979		Total 864

All *PIK3CA* cases displaying both amplification and a oncogenic variant were placed in the *PIK3CA* (E545K, E542K, H1047R, H1047L) group. The MAPK PIK3CA negative group is only negative for the variants displayed in this table. For *NF1* and *RASA1* (H) represents high probability of loss of function based on *in silico* analysis, (M) represents unknown significance, and (L) represents potentially benign. The number of cases with ≥ 20AF highlighted based on sequencing frequency would suggest 66% of the tumour expresses the variant.

4.6 *NF1* Copy Number Variation

All 86 samples were then subjected to *NF1* copy number analysis using the protocol described in section 2.2.8.1. Any samples with ≤20% tumour content (n = 8) were macro dissected prior to DNA extraction. All samples were run in duplicate technical replicates and the mean used for annotation. Three independent batches were run to analyse all samples, each of which included, a NTC, and 1, 2, and 3 *NF1* copy number controls with 30 ng of starting material. All NTC had the expected 0 copy number change. The controls for 1 and 2 *NF1* copy number were calculated as expected, with a mean copy number of 0.99, range 0.038 and 1.98, range 0.14 respectively across the 6 intra batch replicates. The 3 copy number control did demonstrate more variance, with an intra batch mean of 3.15 and range of 0.26 across the 6 replicates. The greatest variance observed in the validation set reported in section 3.5 was 0.21 for both intra and inter replicates, again seen in the 3 copy number control.

Starting amounts of DNA ranged from 1.6 ng to 260 ng, quantified by the Qubit and 0.2 ng to 136 ng by qPCR. 88% (n = 76) of the samples demonstrated a SD of <0.1 difference between technical replicates, representing <0.15 copy number change. 7 % (n = 6) of the samples demonstrated a SD >0.149, which represents a >0.21 copy number difference. These cases had differences between technical replicates ranging from 0.23 to 0.37 copies. 4 of the 6 samples with a poor SD >0.149 all had low starting material of 0.4 ng to 0.7 ng based on qPCR, which equates to 121 to 212 functional haploid copies. *NF1* copies across all cases ranged from 0.87 to 3.3 copies without factoring in the tumour content. The full sample copy numbers are shown in Table 4.19.

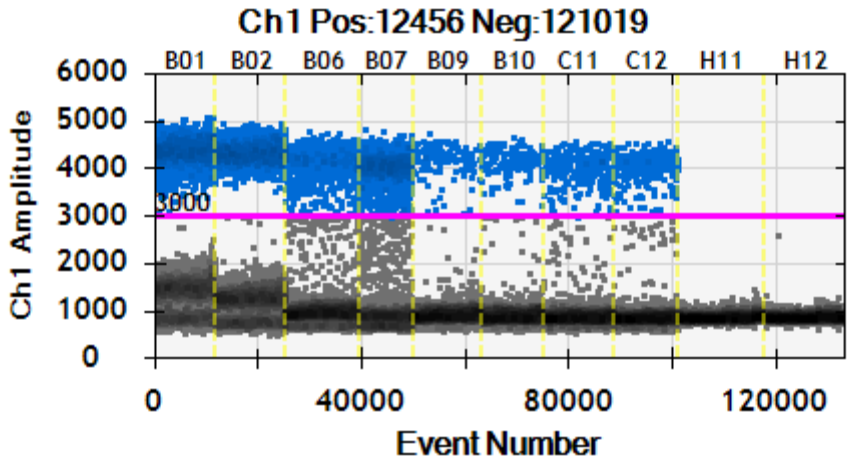
Table 4.19. *NF1* copy number analysis

Sample ID	Copy Number	Sample ID	Copy Number	Sample ID	Copy Number	Sample ID	Copy Number
NF1-107	1.91	**NF1-151**	2.06	**NF1-174**	1.86	**NF1-214**	0.87
NF1-115	2.23	**NF1-152**	1.91	**NF1-175**	1.60	**NF1-215**	1.63
NF1-117	1.77	**NF1-154**	2.15	**NF1-176**	1.73	**NF1-216**	1.42
NF1-119	1.92	**NF1-155**	1.93	**NF1-177**	1.51	**NF1-217**	1.95
NF1-120	1.51	**NF1-156**	1.50	**NF1-178**	1.90	**NF1-218**	1.87
NF1-122	3.35	**NF1-157**	1.72	**NF1-179**	2.00	**NF1-220**	1.50
NF1-126	2.64	**NF1-158**	2.51	**NF1-180**	1.84	**NF1-221**	1.37
NF1-130	2.01	**NF1-159**	1.54	**NF1-181**	1.55	**NF1-222**	1.28
NF1-132	1.76	**NF1-160**	1.75	**NF1-182**	2.09	**NF1-223**	1.21
NF1-134	1.85	**NF1-161**	1.99	**NF1-183**	2.10	**NF1-224**	1.32
NF1-136	2.27	**NF1-162**	2.16	**NF1-184**	1.84	**NF1-225**	1.63
NF1-138	1.69	**NF1-163**	1.95	**NF1-185**	2.09	**NF1-227**	1.32
NF1-139	1.69	**NF1-164**	2.31	**NF1-201**	1.98	**NF1-231**	1.07
NF1-140	1.99	**NF1-165**	2.16	**NF1-203**	1.30	**NF1-232**	1.25
NF1-141	1.29	**NF1-166**	2.07	**NF1-204**	0.99	**NF1-233**	1.57
NF1-142	2.18	**NF1-167**	1.75	**NF1-205**	0.90	**NF1-234**	2.13
NF1-144	1.22	**NF1-168**	2.00	**NF1-206**	1.53	**NF1-237**	2.37
NF1-145	1.76	**NF1-169**	1.67	**NF1-207**	1.15	**NF1-238**	1.60
NF1-146	1.49	**NF1-170**	1.88	**NF1-210**	1.53	**NF1-239**	1.33
NF1-147	2.61	**NF1-171**	1.94	**NF1-211**	1.19	**NF1-240**	1.51
NF1-148	2.19	**NF1-172**	1.80	**NF1-212**	2.29		
NF1-150	1.66	**NF1-173**	1.86	**NF1-213**	1.35		

NF1 copy number analysis. Not corrected for tumour content. Sample IDs shown in red were macro dissected prior to analysis. Four cases with no visible tumour content highlighted in green.

Initial analysis showed 3 out of 4 of the samples with no visible tumour content (NF1-203, NF1-204, and NF1-232) showed copy number loss. Initially we investigated the clusters and thresholds to determine possible errors which may have occurred during analysis. However, all clusters were clearly defined, NF1-203 showed the greatest interference from rain (droplets which appear to fall from the positive cluster back to the negative) as shown in Figure 4.6. Adjusting the threshold had little effect on the result of this sample, demonstrating the majority of the positive events were in the cluster.

A



B

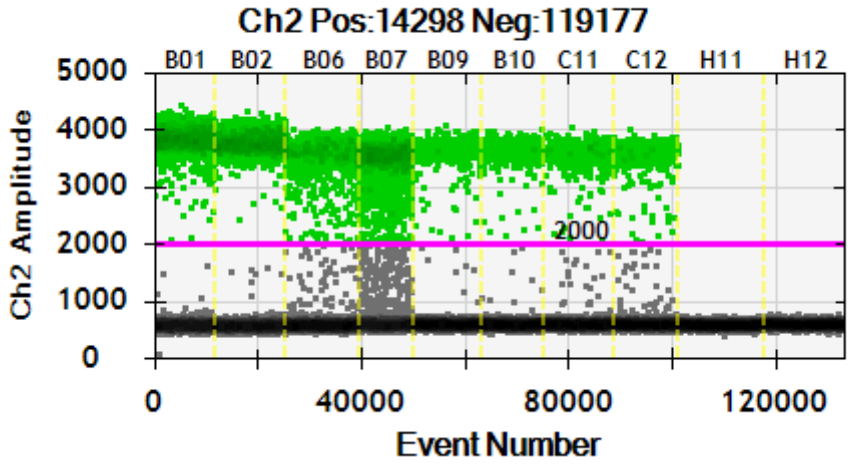


Figure 4.6. 1D plot of ddPCR analysis. A) FAM florescent amplitude for *NF1* probes, positive droplets shown in blue, negative droplets show in black, NTC. **B)** HEX florescent amplitude for *PRR30* probes. Positive droplets shown in green, negative droplets show in black. Wells B01-02, 2 copy number control, B06-07 NF1-203, C11-12 NF1-204, and H11-12 NTC.

We then investigated the starting amount of DNA to determine any correlations between this and copy number change. The three samples with no tumour content which should

have a 2 copy number ratio all had <1.6 ng of starting material. We know from the validation data and the controls 30 ng of 100% functional DNA produces accurate reproducible results. We also demonstrated reproducibility with a low quality FFPE sample with a 7 ng of functional starting material. To investigate this, cases were split based on qPCR amount of starting material using 5 independent threshold groups with high and low arms; <30> ng, <20> ng, <10> ng, <5> ng, and <2.5> ng. Within each one of these groups we compared high and low starting material and *NF1* copy number using an independent t-test. All five groups displayed significant differences in *NF1* copy number in relation to high and low starting material ($p < 0.04$) as shown in Table 4.20. A decrease in *NF1* copy number mean was observed in relation to reduced starting material.

Table 4.20. Independent t-test based on starting material and copy number variation

	Number of Samples	Standard Deviation	Mean	Sig. (2-tailed)
CNV <30 ng	69	0.449	1.72	0.038
CNV >30 ng	17	0.170	1.95	
CNV <20 ng	53	0.454	1.67	0.009
CNV >20 ng	33	0.308	1.91	
CNV <10 ng	38	0.490	1.61	0.002
CNV >10 ng	48	0.309	1.88	
CNV <5 ng	22	0.370	1.44	<0.0001
CNV >5 ng	63	0.372	1.88	
CNV <2.5 ng	11	0.333	1.28	<0.0001
CNV >2.5 ng	74	0.379	1.82	

To determine if the decreasing trend was significant, we performed an ANOVA across the high and low arms of each threshold group. No significant difference ($p = 0.664$) was observed in *NF1* copy number across the high arms (>30, 20, 10, 5, and 2.5 ng). A difference in *NF1* copy number was observed between the low arms ($p = 0.007$). A Tukey HSD *post hoc* analysis demonstrated no significant difference between the <30 ng group and <20 and <10 ng ($p > 0.796$). However, the <5 ng and <2.5 ng group did show a significant difference ($p = 0.39$ and $p = 0.021$) respectively, shown in Figure 4.7 . This does support

the observation that reduced starting material does have an effect on copy number analysis using this method. As <30 ng to <5 ng groups formed a homogeneous subset we define 5 ng starting material as a lower limit for analysis. A final Independent t-Test between the high and low arms using 30 ng as the threshold demonstrated no significance ($p = 0.411$) with all <5 ng samples removed.

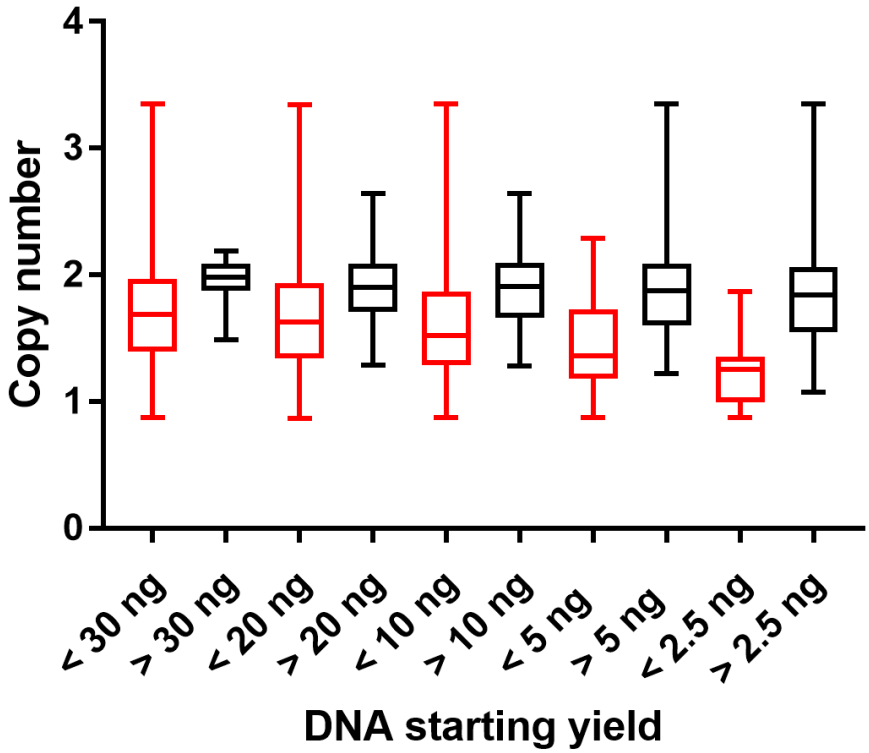


Figure 4.7. Comparison of starting material and *NF1* copy number analysis. Whiskers represent high and low. The *NF1* copy number data was analysed using five different thresholds (30 to 2.5 ng) dividing the copy number data into high and low arms. Red represents the high arms. Black represents the low arms. Corresponding threshold groups all displayed significant differences in CNA between the high and low arms as determined by independent t-tests ($p < 0.04$). ANOVA demonstrated no significant difference in CNV between the high arms of the various threshold groups ($p = 0.664$). A significant different was observed between the low arms ($p = 0.007$). Tukey HSD *post hoc* analysis determined a significant difference between <30 ng and <5 ng low arms ($p = 0.01$). 30 ng to 5 ng low arms were identified as a homogeneous subset.

Employing the 5 ng cut off limit for starting material excluded 22 cases from *NF1* copy number analysis. 20 of the cases were from the ReSoLuCENT cohort. Lower yield was to be expected of these samples as all were smaller in physical size compared to the surgical specimens. Included in the 22 excluded cases were the three with no visible tumour content, all of which had <1.6 ng starting material. With 64 samples remaining we calculated the actual copy number changes based on tumour content. The following calculation was used and is based on the assumptions that pathology tumour content is correct and all non-tumour tissue is diploid.

$$\frac{(CN \times 100) - (2 \times NT)}{TC} = \text{Tumour copy number}$$

Where CN is the raw copy number, NT is percentage of normal tissue, TC is tumour content percentage.

As the highest observed variance between replicates in the clinical samples remaining was 0.34, we implemented this as a threshold from the expected 2 copy number (<1.66 copy number loss, >2.34 copy number gain). 45% of the samples (n = 29) were calculated as diploid, 39% (n = 25) showed some degree of *NF1* copy number loss, and 16% (n = 10) with copy number gain. Campbell and colleagues work demonstrated 47% of their cohort to be diploid, 19% copy number loss and 33% copy number gain (Campbell et al., 2016).

4.6.1 *NF1* Copy Number Change and Clinical Information

To investigate any correlations in *NF1* copy number between subtype, gender, and stage we used a 3 way ANOVA. The two remaining non-specified subtypes were removed from analysis. Stage 1 and 2 were grouped together and stage 3 and 4 patients to form stage 1-2, and stage 3-4 groups. No significant 3 way interaction was observed between gender, subtype, or stage (p = 226). Further analysis using the 3 way ANOVA demonstrated no further significance when comparing all possible two way interaction ($p \geq 0.571$). Plots for these independent groups are shown in Figure 4.8, in addition to a primary vs metastatic *NF1* copy number plot.

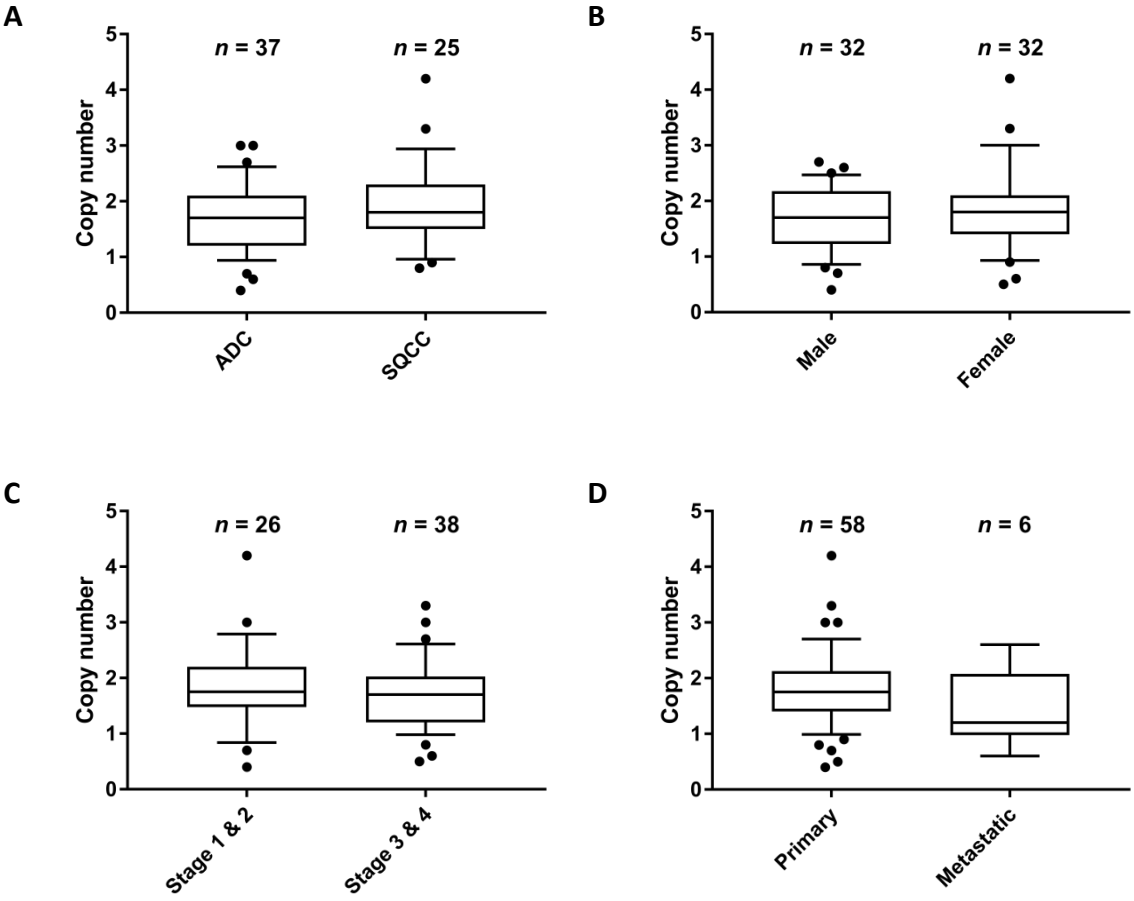


Figure 4.8. *NF1* copy number change NSCLC sub-type, gender, primary or metastatic sites. Whiskers represent 10 to 90 percentile. **A)** Adenocarcinoma and Squamous cell carcinoma. **B)** Gender. **C)** Stage 1 - 2 and 3 and 4. **D)** Primary site biopsy and metastatic site biopsy

To briefly investigate *NF1* copy number change, we analysed the 14 germline cases we sequenced in the samples previously in this chapter. Starting concentrations as assessed using the Qubit ranged from (13.8 to 150 ng). All of the cases demonstrated a copy number ratio between 1.88 and 2.01 as shown in Figure 4.9.

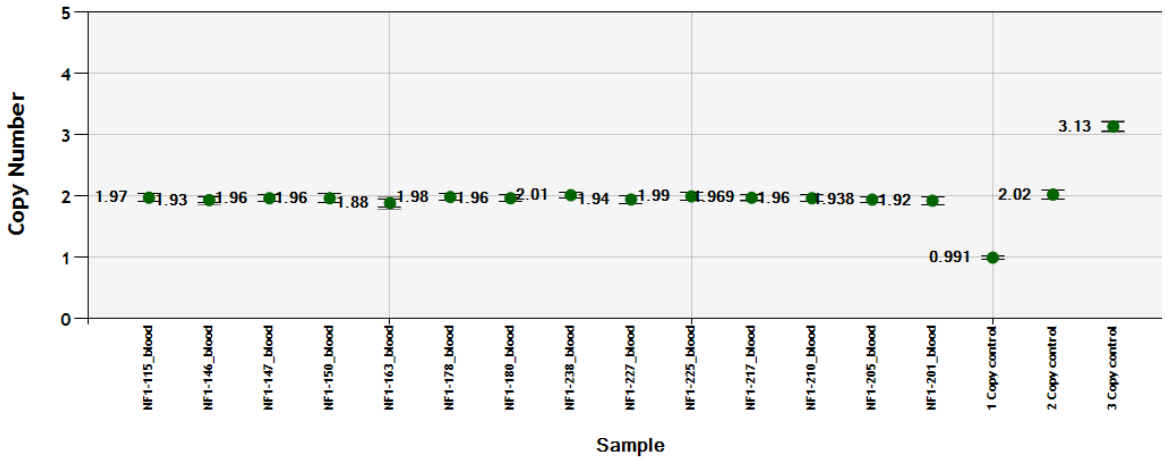


Figure 4.9. ddPCR copy number analysis of *NF1* in the paired germline cases. Replicates merged to present the mean *NF1* copy number.

4.7 Discussion

We successfully recruited 86 out of our target of 100 patients from one site in less than three years. This was a high recruitment rate when compared to larger multisite studies such as ReSoLuCENT. We demonstrated that qPCR, for quantification of DNA derived from FFPET, is an essential requirement to get the most from the tissue / DNA available. Using this, in combination with the NGS protocol described in section 2.1.4, provides a robust platform for analysis of highly degraded DNA. We demonstrated that this technology is highly efficient at identifying clinically important genetic information, even when only expressed at low allelic fractions. Finally, we demonstrated that ddPCR has great potential for future applications, including copy number analysis. However, before this can be used in a clinical setting further validation work to fully define its limitations is required.

FFPET is challenging to work with for genetic analysis, but remains the primary source of material available for the identification of genetic variations in solid tumours. 23 of our samples did not match their reported size in histology reports and proved to have insufficient tissue available for analysis. The samples which were deemed sufficient demonstrated a high degree of variability in quality of functional DNA recovered. This variability is largely dependent on how anoxic the excised tissue has become in the timeframe between patient surgery and formalin fixation, and the actual fixation process itself. Findings in this study demonstrate this variability is not site or laboratory dependent; it fluctuates on a case by case basis. Our cohort of surgical samples (n = 59) were excised and fixed in the same hospital and laboratory. These samples demonstrated DIN scores ranging from 1.9 to 6.7 and QFI scores of 3.3 to 98%. We observed large fluctuations in DNA yield and quality, even in samples excised and processed within a 24 hours of each other. NF1-158 and NF1-185, both were similar sized specimens, NF1-158 had a DNA concentration of 8.4 ng/µl, DIN score of 2.6, and QFI of 3.3%, whilst NF1-185 had a concentration of 51.5 ng/µl, DIN score of 5.6, and QFI of 51.6%. These differences are sufficient to determine whether or not a reliable genetic result can be obtained.

However, over the last decade with the exponential growth of technologies such as NGS, genetic methods are being optimised to enable confident analysis of DNA derived from FFPET. Here we initially demonstrated the differences in quantification methods of FFPET derived DNA. Whilst the three methods used in this study all display linear relationships to

each other they are actually measuring different nucleic acid states. This leads to the NanoDrop and Qubit overestimating the actual amount of functional DNA, which can have significant consequences on downstream genetic analysis. Similar findings were reported by Sah and colleagues (Sah et al., 2013). qPCR quantification is not perfect, it is largely based on the assumption that the genome is degraded uniformly in relation to the small locus you are actually quantifying. However, for NGS and ddPCR we have shown it to be the most reliable method of quantification to enable successful downstream analysis. Another benefit of qPCR analysis is its ability to determine the actual level of degradation in relationship to actual dsDNA as determined by the Qubit. We demonstrated the QFI score has a linear relationship to the DIN score; however, as with DNA quantification these are measuring different aspects in order to determine the level of degradation.

All samples which had adequate starting material as determined by both Qubit and qPCR but still failed NGS library preparation, had low QFI scores of ≤10.1%. This affects the underlying kinetics of library preparation as over 89% of the DNA is actually non-functional and has an inhibitory effect. Whether this is due to enzyme saturation /exhaustion in the pre-PCR steps, or just overloading the PCR reaction remains to be determined.

Using the methods describe in section 2.2.4 we were able to obtain quality sequence for 86 of the 92 cases which were not rejected at the QC steps. We also demonstrated that using these methods it is possible to successfully hybridise and enrich for target regions with >170 ng of pre-hybridisation PCR material. This enabled the synthesis of sequence from 29 of the samples which would have been rejected following the manufactures' protocol. Only one sample did not meet the minimum sequence depth of 300 X required for calling variants at 3% AF.

No oncogenic variants were identified in the four samples with no reported tumour content in the pathology reports. Overall a total of 25 oncogenic mutations known to activate the MAPK pathway were identified: 7 *EGFR* oncogenic variants, 4 of which were previously reported by orthogonal methods of detection in the patients' notes. The remaining 3 cases were not tested for *EGFR* variants elsewhere. Included in these was an *EGFR* exon 19 deletion, at just below 3% AF in a highly degraded sample (DIN 1.9 and QFI

14.4%). This variant would have not been identified it we had relied on the bioinformatics pipeline alone to filter variants based on our criteria of ≥3% AF and ≥ 15 supporting reads.

The overall percentage of identified *EGFR* variants in our cohort 8% (7/86) was below what is generally reported for overall NSCLC prevalence (14%) in the European population. However, just analysing the ADC population we observed 15% (7/48) *EGFR* positive cases which is slightly below meta-analysis reported frequency (17.5 to 21.8%) (Zhang et al., 2016b). Our data also agreed with observations that these variants are more frequently observed in females and non-smokers (Zhang et al., 2016b).

17% (15/86) of our total cohort displayed *KRAS* (G12) activating mutations which is the same frequency as reported by Pan-Lung TGCA data 17% (187/1144). 23% (11/48) of our *KRAS* G12 variants were observed in ADC cases and 13% (4/30) in SQCC cases, which differs to the previously published report of 28% (187/660) and <1% (2/484) respectively in the Pan-Lung study (Campbell et al., 2016, Boch et al., 2013). 90% of patients recruited to our study with *KRAS* mutations were current or ex-smokers which is reflective of these variants been more prevalent in smokers (Paik et al., 2012). No *KRAS* G13 or Q61 variants were identified in our study, Campbell et al. (2016) reported <2% G13 and <1% Q61 variants, so this difference could be due to the small population used in this study.

BRAF variants (G464, G466, G469) were observed in 3% (3/86) of our total cohort which is slightly above the 2% (22/1144) level that the Pan-Lung study reported, just counting these specific variants Campbell *et al*. (2016). Considering just the ADC *BRAF* cases our data 1% (1/86) is just below the reported frequency of 2% (18/660) in the Pan-Lung study. Of the remaining *BRAF* variants observed in the SQCC cases we observed 2% (2/86) frequency which is above that of the Pan-Lung study which reported <1% (4/484) (Campbell et al., 2016).

Whilst our results looking at the frequency of the MAPK oncogenic variants was similar to that observed in the Pan-Lung study, we did see an increase of *BRAF* and *KRAS* variants in the SQCC population. One potential reason could be the small cohort of 86 cases used in this study compared to the 1144 cases analysed in the Pan-Lung study. Excluding any potential sample mix up or errors in documentation, another potential reason for this discrepancy could be mixed histology. Adenosquamous carcinoma is the most common

mixed-histology lung cancer pattern and accounts for up to 3.4% of all NSCLC cases (Deng et al., 2013). Therefore, it is possible the two *BRAF* variants in our SQCC cases, which were both clonal, could potentially be due to mixed histology for NF1-148 and NF1-238. Mixed histology could also partially explain the increase in frequency of *KRAS* variants in our SQCC cases. Another potential cause for the *KRAS* variants observed in our SQCC population could be the increased sensitivity observed in our study. 3 of the *KRAS* variants were reported at low AF, these would not have been detected using our variant detection criteria when only sequencing to a depth of 100X, which is the mean depth observed in the Pan-Lung study (Campbell *et al*., 2016), compared to deep sequencing used in this study which had a mean of 1560X across the G12 genomic loci.

Screening for *NF1* potential loss of function variants, we identified 22 variants in 17 cases. We also identified 10 possible cases with potential pseudogene variants mapping back to *NF1*. Furthermore, 5 variants were identified in the promoter region. Of the potential pseudogenes variants two were of great interest, p.Tyr489Cys and p.Arg1241*. Both of these were reported as pathogenic in Neurofibromatosis patients (Laycock-van Spyk et al., 2011, Upadhyaya et al., 2003). Based on our current data it is not possible to determine if these are *NF1* variants or pseudogene in origin. This would require longer sequence reads to identify any up or downstream variants which match the pseudogene or functional *NF1*. As they are reported in the literature as pathogenic we classed them as high potentials for *NF1* loss of function. The other 7 potential pseudogenes variants were intronic and would have no predicted effect on mRNA splicing, if they are actual *NF1* variants. The final potential pseudogene variant was only reported at 3% AF in a sample with an expected 20% AF based on tumour content and classed as likely benign.

Five cases exhibited variants upstream of the transcription start site, within the promoter region. Whilst we did not sequence the matched blood samples to determine if these were somatic or germline line in nature, the low AF of NF1-148 and NF1-172 does suggest these are somatic, with the remaining being germline. These variants have the potential to downregulate or indeed upregulate *NF1* expression.

Of the remaining 22 *NF1* variants 14 were novel with no previous reports in the literature or databases. We classified variants; high potential *NF1* loss of function, unknown

significance, and likely benign based on *in-silico* analysis. 9 synonymous, intronic SNVs, or dinucleotide variants, which had no predicted effect on splicing, were classified as low. NF1-150 was the only case with a variant with the potential to effect splicing and was classified as unknown significance. Of the non-synonymous variants 2 were classified as likely benign and 5 unknown significance, based on analysis using 7 different predictive tools. Finally, we identified 5 null variants in 4 cases which were all classified as high potential. When we compared the frequency of null variants in our cohort 5% (5/86) with the cases from the Pan-Lung TGCA data 6% (59/979) we observed similar frequencies even with our relatively small cohort.

Our cohort of *NF1* variants in NSCLC was above 12% reported in the literature, 22% (n = 19) cases displaying a total of 25 variants. However, variants reported in these studies only investigated the exonic regions ±2bp. Removing the intronic variants beyond ±2bp in our cohort resulted in 17% (n = 16) still displaying *NF1* variants. To investigate this further we interrogated the Pan-Lung sequencing data (Campbell et al., 2016). The mean sequencing depth of the variants identified in *NF1* was 98 X, ranging from 16 to 542 X. 37% of the 137 *NF1* variants identified had less than 15 supporting reads which ranged from 4 to 91. Whilst our variant calling criteria was more stringent than Campbell and colleagues, our greater sequencing depth enabled more sensitivity at picking up low level variants. We tested this using Campbell's 4 reads minimum to call variants and reducing our read depth to 100 X and adjusting the number of variant reads respectively. This resulted in 13% (n = 11) of cases displaying *NF1* variants.

We then had to consider the AF in relation to tumour content and its potential to null any effect on the MAPK pathway the variants could have. Cases with low AF and high tumour content suggest low level intra-tumour heterogeneity of the variant population. We could not select for this limited population when enriching for tumour content for downstream gene expression signature analysis. 20 (n = 4) of the *NF1* variants have supporting AF that suggests they could be ubiquitously expressed throughout the tumour. The remaining 80% (n = 16) were only found in a sub population of the tumour and represents intra-tumour heterogeneity. Of the ubiquitously expressed, only four matched variants which resulted in *NF1* medium or above loss of function potential. This is lower than we anticipated, based on previous reports of inter tumour heterogeneity in NSCLC (de Bruin et al., 2014b). The

two potential pseudogene reads reported as pathogenic fell just below the ±30% threshold. This leaves a reduced potential *NF1* loss population when considering the gene expression signatures.

When we compared intra-tumour heterogeneity in the *KRAS*, *EGFR*, and *BRAF* positive cases against that of the first 100 samples to be analysed from the TRACERx study we observed similar degrees of correlation shown in Table 4.16. However, there was a significant difference in intra-tumour heterogeneity of the *NF1* variants with TRACERx reporting 50% to be only observed in sub-colonal populations, in relation to on 80% reported in this study. It is important to note the reasons for investigating this and the method of doing so is different for both studies. TRACERx is tracking tumour evolution through genetic changes leading to intratumour heterogeneity, whereas this studies goal was to determine if the *MAPK* and *NF1* variants were expressed throughout the section of the tumour analysed for potential downstream effects when comparing the gene expression signatures. Because of this, different methods of analysis were used; TRACERx used multi-region (2 – 8 regions) of sample collection from each tumour, whereas our study just used single region sampling. TRACERx, therefore has a more comprehensive view of the tumours mutational landscape as a whole. This explains why the clonal variants (*KRAS*, *EGFR*, and *BRAF*) between the studies are of similar frequency, however with more sub-clonal variants such as *NF1* will experience more variability in single region sampling.

NF1 Copy Number Analysis

Our investigation of ddPCRs ability to measure *NF1* copy number displayed issues. Whilst all the controls functioned as expected, three samples with no visible tumour content displayed unexpected *NF1* copy number variation. Our previous validation work was based largely on none FFPET derived DNA and the lower limits of starting material were not assessed. We did not validate for the variance observed in quality and low yield of our cohort.

Analysis suggested one potential cause of false positive *NF1* copy number changes was low level starting material. When comparing the low starting material with *NF1* copy number

a trend was observed. Using specific upper and lower threshold limits we determined a significant decline in copy number at 5 ng starting material as assessed by qPCR. Using this as a threshold excluded 22 cases, including the three with no visible tumour content. Whilst this unforeseen copy number loss in these three samples did give us reason to question the data, low starting material gave us a possible explanation. We therefore continued with copy number analysis of the remaining 64 cases.

Taking tumour content into consideration and setting a threshold based on the assay variance, our data suggested 45% of our remaining cases were diploid, 39% displaying some level of copy number loss, and 16% with *NF1* gain. Campbell reported 19% copy number loss and 33% gain. However, earlier work suggests the diploid prevalence is 93% and 73% in ADC and SQCC respectively, with low levels of copy number gain / loss (Hammerman et al., 2012, Collisson et al., 2014). Different methods are also a consideration, as total concordance is rarely observed between platforms. However, data from the previously mentioned genetic profiling studies did not correlate on *NF1* copy number prevalence using the same Affymetrix Genome-Wide Human SNP Array 6.0. This could suggest *NF1* copy number change is highly variable in NSCLC cases. Copy number increases are more commonly observed in oncogenes as they give the cell a distinct growth advantage. While copy number loss is more common in TSG as, this would give the cell a growth advantage. As *NF1* is understood to be a TSG there would be no known advantage for the cancer with *NF1* copy number increase. One potential reason for the *NF1* copy number increase observed in NSCLC is that *NF1* copy number increase could just be a passenger in close chromosomal proximity to a gene which would confer a growth advantage such as *ERBB2.* Campbell *et al*. (2016) data did show the amplification of *ERBB2 and NF1* co-occurred in 66% of the *NF1* amplification cases, which supports this hypothesis. Both *NF1* and *ERBB2* are located on adjacent bands the q arm of chromosome 17, suggesting the growth advantage could come from *ERBB2* amplification and *NF1* is just a passenger.

Our analysis of the different clinical and demographic groups including; subtype, gender, and disease stage, demonstrated that neither of the groups had any influence over *NF1* copy number changes. ddPCR demonstrated great promise for analysis of copy number variation. The initial validation work was highly accurate and reproducible. The analysis of

the germline samples provided expected results. However, before this method can be transferred confidently to FFPET there are areas which require addressing. It requires the lower starting limit defining, and the effects on inhibition from non-functional DNA assessing. Another key area is the requirement of running at least 2 reference genes to compensate for potential reference copy number changes. Whilst other options are available for CNA including; low pass whole NGS approaches which have been shown to detect copy number changes in low AF cases (Silva et al., 2018), or more established methods, such as FISH. We explored ddPCR for its low cost and potential to reduce analysis and turnaround time over NGS and FISH.

To conclude we have identified oncogenic variants and potential *NF1* loss of function variants which correlate with previous reports. However, not all of these are expressed ubiquitously throughout the tumour. This has to be a major consideration in the next stage of analysis. Another consideration is *NF1* loss in a subset of patients. This could also have influence over the MAPK mRNA gene expression signatures, which will be explored in the next chapter. Whilst mRNA expression and copy number changes are not always reflective of each other, it is something which will be further explored.

Chapter 5
Functional relevance of NF1 in NCSLC

5.1 Introduction

To investigate the functional relevance of *NF1* variants in NSCLC we required a method capable of measuring MAPK pathway activation in clinical FFPET samples. One option which has demonstrated promising results in degraded FFPET is the use of gene expression signatures. There have been several publications over the last decade reporting gene expression signatures relating to activation of the MAPK pathway. In 2010 two different signatures, a MEK signature and a RAS signature were reported, both demonstrating promising results in multiple cancer cell lines (Dry et al., 2010a, Loboda et al., 2010b). The RAS pathway signature comprised 147 genes, 105 up-regulated and 42 down-regulated in response to RAS activation. The MEK signature consisted of 18 genes which were up-regulated during MAPK activation. Both signatures were developed from publically available microarray expression datasets. The signatures shared some target genes including; DUSP4, *DUSP6*, S100 calcium binding protein A6 (*S100A6*), serpin family B member 1 (*SERPINB1*), and zinc finger protein 106 (*ZFP106*).

The RAS signature was suggested to be a more effective predictor to MEK inhibition and MAPK dependence than *KRAS* mutations. Work in cell lines demonstrated that the RAS signature score was directly related to MEK phosphorylation. It further demonstrated that the score decreased in *KRAS* positive cell lines when treated with a MEK inhibitor (Loboda et al., 2010b). The RAS signature was shown to have a high sensitivity (>90%) for detecting *KRAS* mutations in various cancer cell lines and lung cancer clinical samples. However, it lacked specificity (50%), with high signature scores also found in many cases with wildtype *KRAS* status. One explanation, which was not previously investigated, that could account for this low specificity is potential oncogenic upstream or downstream effectors of MAPK signalling.

The second gene signature relating to the *MAPK* pathway was proposed by Dry *et al*. (2010). Work in breast, colon, melanoma, and lung cancer cell lines suggested the RAS signature could predict sensitivity to MEK inhibition. Dry and colleagues demonstrated knock down and transfection of MEK in these models correlated with decrease or increase in this signature respectively. Dry and colleagues suggested the signature was more consistent across cell lines at predicting sensitivity to selumetinib compared to measuring pERK or using *BRAF* and *RAS* status.

Whilst both signatures were promising for predicting MAPK activation, both Loboda's and Dry's data was generated via microarray analysis in cell models or fresh frozen tissue. Microarrays are widely established platforms for gene expression profiles, however they do require high quality RNA which cannot be guaranteed from FFPET (Wu Thomas et al., 2009). Dry *et al*. (2010) investigated transferring the MEK signature to reverse transcription (RT) qPCR platform, but noted that not all genes were detectable in 10% of cases. While 18 genes is manageable using RT-qPCR, the analysis of Loboda and colleagues 147 gene panel is more challenging. Another drawback of RT-qPCR is the inhibition of the enzymatic reaction by degraded nucleic acid. We have already demonstrated a large degree of variability of nucleic acid quality in our samples in Chapter 3.

The Nanostring nCounter is an alternative to the two afore-mentioned platforms. This platform demonstrated high correlation to the Affymetrix GeneChip (R^2 = 0.79) and RT-QPCR (R^2 = 0.95) when measuring gene expression (Geiss et al., 2008). The Nanostring nCounter has also proved to be more tolerant to FFPET samples than microarrays or RT-qPCR methods, demonstrating a Pearson correlation of r = 0.94 between fresh tissue and FFPET samples (Reis et al., 2011). A further advantage is that the RAS and MEK signatures have been transferred to the Nanostring platform (Brant et al., 2017). Brant and colleagues undertook validation work to compare the NanoString vs. microarray data. A R^2 coefficient between specific genes measured by the platforms ranged from 0.75 to 0.96, demonstrating a good degree of correlation. They also acknowledged the large degree of variability observed when working with FFPET and established the optimum starting amount of RNA to be 100 ng. The MEK signature was initially developed using multiple mixed cancer cell lines. Brant *et al*. (2017) went on to refine this signature, making it disease specific for NSCLC. This refinement condensed the 18 gene signature down to 6 genes; *DUSP4*, *DUSP6*, pleckstrin homology like domain family A member 1 (*PHLDA1)*, *ETS* variant gene 4 (*ETV)* 4, *ETV5*, and sprouty like homolog 2 (*SPRY2*).

In this chapter we aim to test the ability of these three signatures to predict *KRAS* positive samples. We know *NF1* is a negative regulator of *KRAS*, thereby *NF1* loss and its potential effects on RAS signalling should be quantifiable using these signatures.

5.2 Gene Expression Analysis via the NanoString nCounter

To determine the ability of the NanoString CodeSet to detect complementary mRNA targets we used a commercially sourced RNA human reference control. The control had a RNA integrity number (RIN) of 7.5 with both 18s and 28s ribosomal bands being clearly observed, Figure 5.2. The control was analysed on 3 independent occasions, using 100 ng of RNA. All Nanostring QC criteria described in section 2.2.9.3 were satisfied. Initial analysis of the RAW gene count highlighted multiple genes within the RAS signature which were below the mean +2SD of the internal negative controls. These included *DRD4* and *RFPL3S,* which belong to the down-regulated arm of the RAS signature and *CXCL5*, and *IL13RA2*, which belong to the upregulated arm. All other genes within the MEK and RAS signatures were counted above the internal and external (RNA free water) negative controls. The external negative controls displayed > 99.9 % of all genes in the mRNA signatures to have lower counts than the mean +2SD of the internal negative controls. The positive control read counts ranged from 9 for *IL13RA2*, to 168740 for the endogenous reference glyceraldehyde 3-phosphate dehydrogenase (*GAPDH)* (95% CI, 2184 to 5321). We plotted the RAW Log_2 gene expression data of the 3 inter batch positive control replicates against each other. All three demonstrated R^2 coefficients of > 0.994 demonstrating a high degree of inter batch reproducibility observed in Figure 5.1.

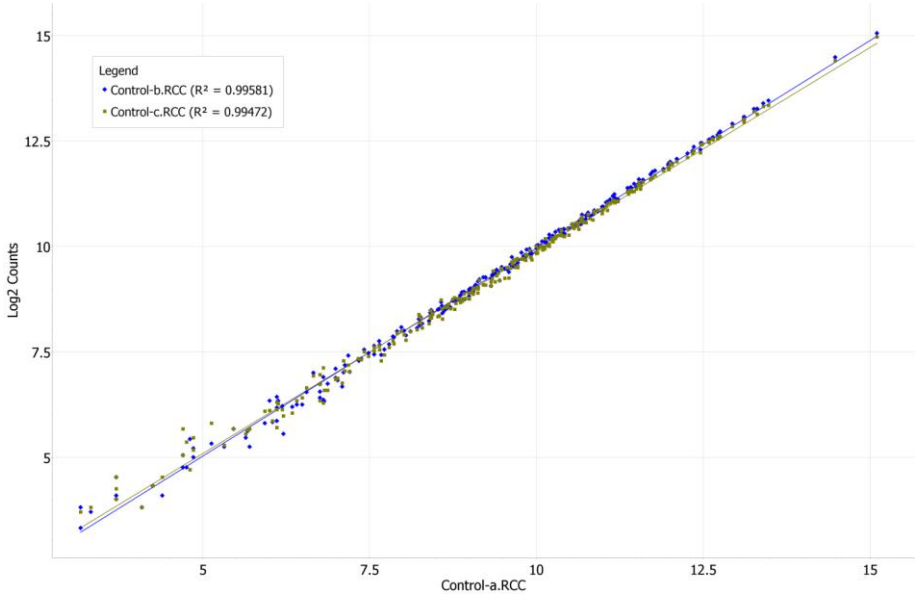


Figure 5.1. RAW Log_2 gene expression of the NanoString CodeSet. Three RNA replicates (Control-a/b/c) across three independent batches. Control-b and c plotted against control-a.

5.3 Reproducibility of the Gene expression Signatures in Degraded Samples

Reproducibility in the positive controls is not representative of what was observed in the FFPET samples, as demonstrated earlier in the ddPCR data. To address this we utilised five FFPET controls, three had an *EGFR*, *KRAS* or *BRAF* oncogenic variant, one had a loss of function *NF1* variant, and the final had no known *MAPK* activating variant. These cases were then set up in duplicate, in three independently prepared batches. The RIN score for these FFPET controls ranged from 1.2 to 1.4, the average fragment size for all of the controls was <200 bp, and the ribosomal bands were not visible demonstrating a high level of degradation, shown in Figure 5.2. 100 ng starting material was used for all replicates.

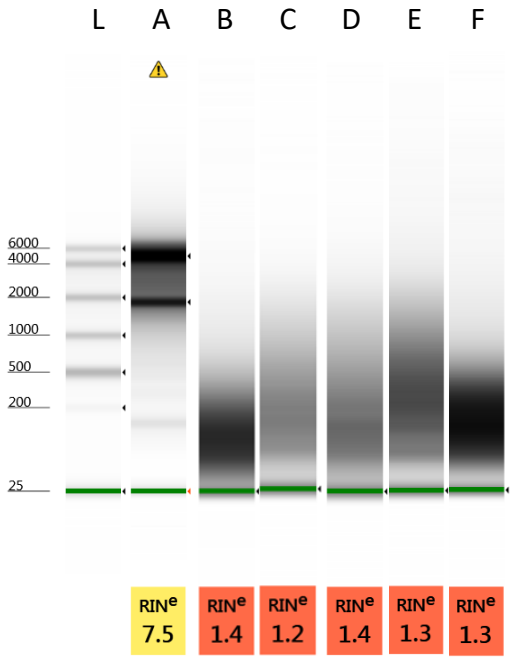


Figure 5.2. Agilent's TapeStation genomic tape pseudo-gel RNA image.L) 6000 kb RNA ladder. **A)** Total RNA human reference control with both 18s and 28s ribosomal bands being clearly observed. Yellow warning sign indicate the sample is out of range for quantification. **B)** *EGFR* positive FFPET RNA. **C)** *KRAS* positive FFPET RNA. **D)** *BRAF* positive FFPET RNA. **E)** *NF1* positive RNA. **F)** No known *MAPK* variant. None of the 5 FFPET controls (B–F) display evidence of the 18s and 28s ribosomal bands.

All Nanostring QC criteria described in section 2.2.9.3 were satisfied. Compared to the positive control used previously the RAW data demonstrated *DRD4* was expressed consistently in the no *MAPK* and the *BRAF* samples, *RFPL3S* in the no *MAPK* sample, *IL13RA2* in the *EGFR* sample, and *CXCL5* expressed in all but the *BRAF* sample. This does suggest the RNA control is not expressing these genes and it is not an issue with the CodeSet. Each of the 6 total replicates for the five FFPET samples were normalised together using nSolver 4.0 as describe in section 2.2.9.3. Negative and positive controls were removed prior to normalisation. We utilised AstraZeneca's in house normalisation tool to determine if any of the 21 housekeeping genes were unsuitable, shown in Figure 5.3. *G6PD* was unsuitable based on the criteria described in section 2.2.9.4 and excluded from normalisation.

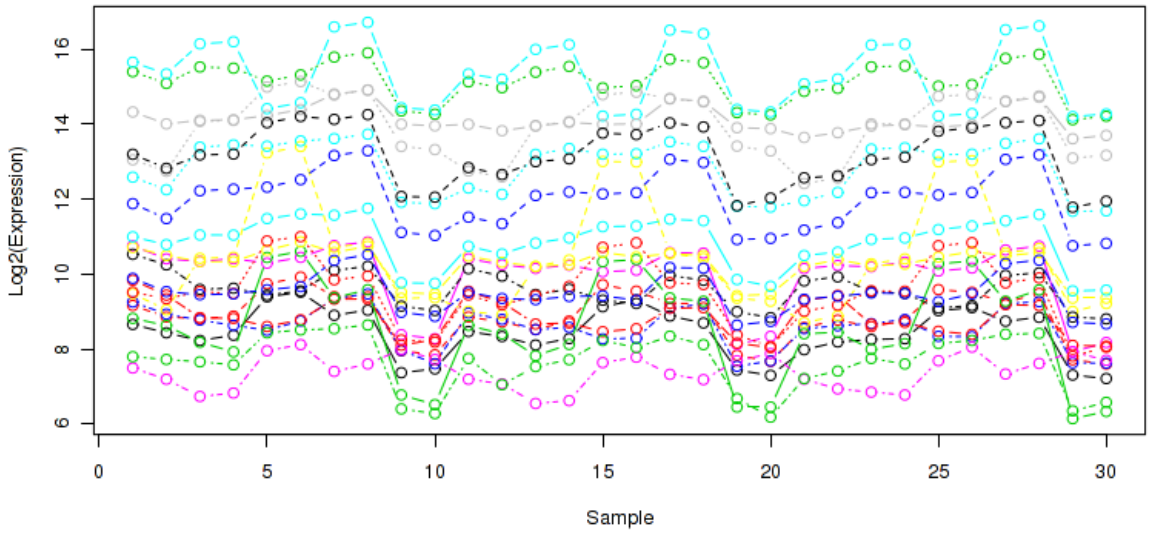


Figure 5.3. RAW Log2 expression for the 21Housekeeping genes across the three independent batches. Intra batch replicates we run in adjacent lanes for the cartridge. *EGFR* variant replicates (1, 2, 11, 12, 22, 21, and 22), *KRAS* (3, 4, 13, 14, 23, and 24), No known *MAPK* variants (5, 6, 15, 16, 25, and 26), *NF1* variant (7, 8, 17, 18, 27, and 28), *BRAF* (9, 10, 19, 20, 29, and 30). *G6PD* shown in yellow was excluded from normalisation.

The normalised gene expression Log_2 data based on the 20 remaining housekeeping genes was plotted against intra and inter batch technical replicates, Figure 5.4. As with the commercial control we observed a high level of reproducibility. Of the 5 independent samples run across three batches the average coefficient was $R^2 = 0.99$ and ranged from 0.977 to 0.996. This demonstrates the CodeSet was again highly reproducible even with degraded FFPET derived total RNA.

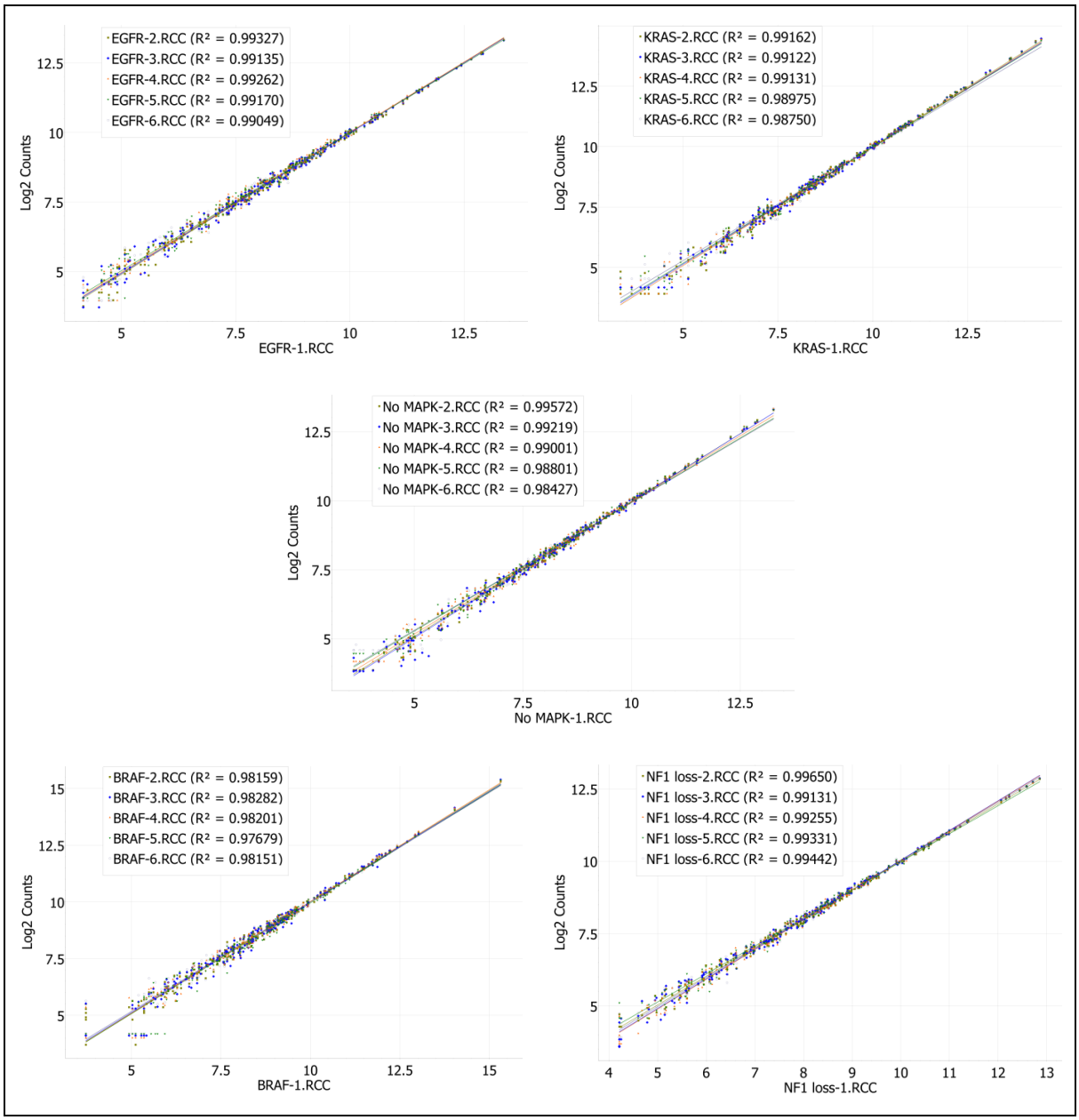


Figure 5.4.Normalised Log_2 gene expression of the NanoString CodeSet. Six FFPET RNA replicates run in three independent batches. Top left *EGFR* replicates. Top right, *KRAS* replicates. Middle, No known *MAPK* oncogenic variants. Bottom left, *BRAF* replicates. Bottom right, *NF1* loss replicates.

We then calculated the RAS signature and the 2 MEK signatures for all six technical replicates across the inter and intra batch runs, shown in Table 5.1. All three different signatures demonstrated a high degree of reproducibility with the greatest range between replicates being observed in the MEK 6 and MEK 18 gene signature of 0.13 (SD = 0.03). The MEK 18 and MEK 6 signature demonstrated a higher score for all the three controls with known *MAPK* oncogenic variants in relation to the *NF1* loss and no *MAPK* controls. *KRAS* had the highest score using the RAS signature with the no *MAPK* controls demonstrating the lowest score. Unlike the MEK signatures the *BRAF* signature score was almost half of that observed in the *KRAS*, with the *EGFR* signature score being barely indistinguishable from the *MAPK* negative score.

Table 5.1. Reproducibility of the mRNA signatures across independent batches.

	Batch 1		Batch 2		Batch 3			
MEK 18	**R1**	**R2**	**R1**	**R2**	**R1**	**R2**	**Range**	**SD**
EGFR +	9.00	8.91	8.96	8.97	8.96	8.94	0.09	0.03
KRAS +	9.05	9.06	9.07	9.07	9.07	9.07	0.02	0.01
BRAF +	8.98	8.98	8.99	9.04	8.91	9.01	0.13	0.04
NF1 +	8.35	8.35	8.39	8.39	8.36	8.40	0.06	0.02
MAPK -	7.90	7.89	7.86	7.88	7.96	7.92	0.10	0.03
MEK 6								
EGFR +	9.17	9.12	9.15	9.14	9.18	9.11	0.06	0.03
KRAS +	8.71	8.71	8.79	8.77	8.77	8.76	0.07	0.03
BRAF +	8.58	8.65	8.64	8.66	8.61	8.68	0.07	0.04
NF1 +	8.05	8.08	8.08	8.10	8.07	8.14	0.06	0.03
MAPK -	7.68	7.61	7.61	7.56	7.69	7.63	0.13	0.05
RAS								
EGFR +	0.11	0.11	0.12	0.12	0.13	0.11	0.02	0.01
KRAS +	0.56	0.55	0.55	0.56	0.56	0.53	0.03	0.01
BRAF +	0.24	0.25	0.24	0.24	0.22	0.23	0.02	0.01
NF1 +	0.31	0.30	0.33	0.31	0.30	0.31	0.03	0.01
MAPK -	0.09	0.09	0.10	0.09	0.09	0.09	0.02	0.01

5.4 Sample Analysis

Two FFPET cases were returned to STH for further diagnostic analysis and were not accessible for this stage of analysis. The 84 remaining samples were all macro-dissected to enrich the tumour content. Three cases were below the recommended starting amount of 100 ng, NF1-216 (86 ng), NF1-223 (68 ng), and NF1-240 (41 ng). The RIN score of our samples ranged from (1 to 2.5) with 85% of them being ≤1.5. No linear correlation was observed between the DIN and RIN scores with a R^2 value of 0.076. The 84 samples were divided into 9 independent batches, each including a negative and positive control. These batches were hybridised, prepared, and had their reads counted on the NanoString digital analyser independently of each other. All the 9 independent RCC files generated by the digital analyser from the batches were normalised together in one experiment. During initial QC analysis all but 2 samples passed nSolver's internal QC checks. NF1-211 and NF1-223 displayed an mRNA content flag. We continued with analysis of these two samples as both still had >63% of the probes above the threshold. Analysis of the external negative controls displayed >99.9 % of all genes in the mRNA signatures to have less counts than the mean +2SD of the internal negative controls. The positive controls, as described earlier had low or extremely low counts for *DRD4*, *RFPL3S*, *CXCL5*, and *IL13RA2*. This trend was also observed in the majority of the samples. As these genes are spread evenly between the RAS up and down arms we did not exclude them from the RAS signature analysis. Plotting the RAW Log_2 expression data from the batch independent positive controls against each other demonstrated all R^2 coefficients to be >0.993 demonstrating a high level of reproducibility between the batches. Analysis of the 21 housekeeping genes again flagged *G6PD* as unsuitable for normalisation. The full housekeeping gene plot is shown in Figure 5.5.

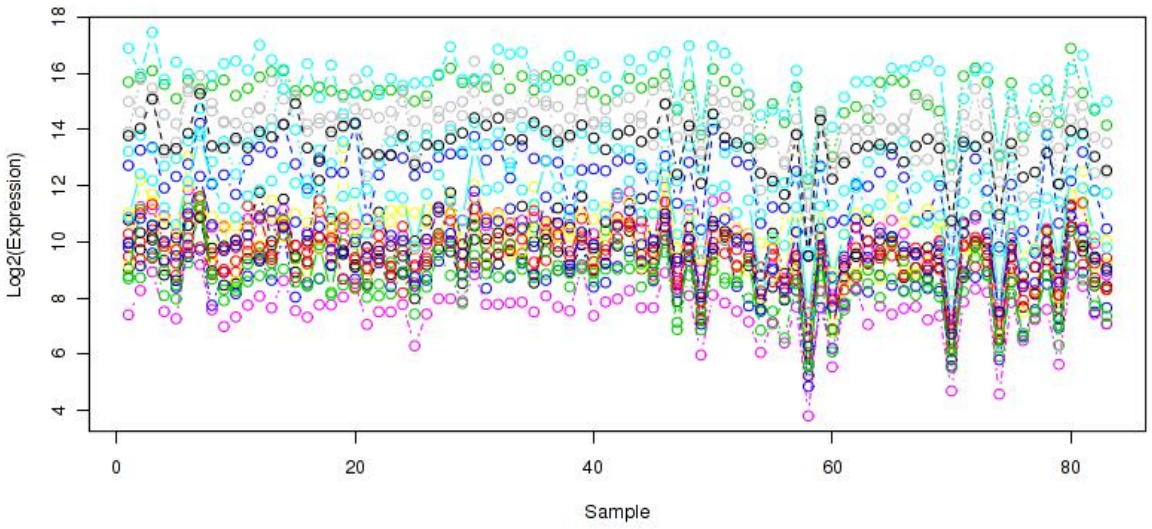


Figure 5.5. RAW Log_2 expression for the 21 housekeeping genes across nine independent batches. Samples 1 to 84. *G6PD* shown in yellow was excluded from normalisation.

5.5 *NF1* Gene Expression

In the NanoString codeSet we designed three *NF1* probes which in theory could differentiate between the two most common mRNA transcripts. The probe NF1-A and NF1-B would detect all three transcripts (NM_001042492.2, NM_000267.3, and NM_001128147.2). NF1-C would only detect the transcript (NM_000267.3). We used an ANOVA to compare the differences in expression measured between the three different probes across all the samples. This revealed a significant difference in transcription measured between the three probes ($p < 0.001$). A Tukey HSD *post hoc* analysis demonstrated there was a significant difference between all the *NF1* probes ($p < 0.001$), as shown in Figure 5.6.

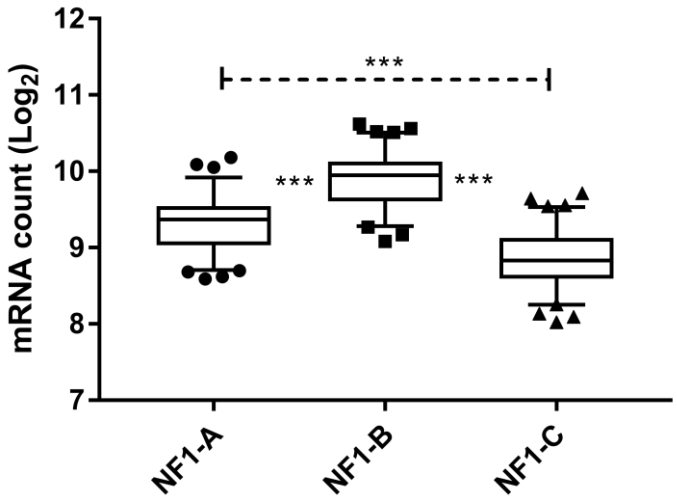


Figure 5.6. Expression of *NF1* in 84 NSCLC cases. Whiskers represent 5–95 percentiles. NF1-A spans exon 7-8 and NF1-B spans exon 22-23 both probes detect the two most common *NF1* mRNA transcripts. NF1-C spans exon 30-32 and only identifies *NF1* transcript 2 (NM_000267.3). An ANOVA followed by a Tukey HSD analysis demonstrated significant differences between expression identified by all three probes.

We expected to see a lower expression of *NF1* transcript 2 (NM_000267.3) targeted by NF1-C, as it is said to be predominately expressed in neurons (Hinman et al., 2014). However, the difference found between NF1-A and NF1-B was not anticipated. One potential cause for this being observed in all samples could be simply a difference in affinity of the capture probes to their mRNA targets. However, based on our data alone this is something we cannot investigate further. Another potential cause could be the transcription of a pseudogene. *NF1P9* and *NF1P10* have been reported by Emsembl as having the potential to be transcribed, and both contain exon 22-23 homologs as shown in Table 3.3.

We compared the expression of the three probes on a sample-by-sample basis including the controls. Whilst not obvious in the grouped analysis shown in Figure 5.6, all samples showed greatest expression of NF1-B (exon 21-22) and lowest of NF1-C (exon 30-32), with NF1-A falling in the middle, Figure 5.7. We used NF1-A (all transcripts) to compare against the transcription of NF1-C which only detects *NF1* transcript 2. By considering the differences in expression between them, it became apparent that *NF1* transcript 2 is the most prevalently expressed. The fold change measured across all the samples demonstrated a mean increase from transcript 2 to transcript 1 of 0.44 fold. The total range of this across all the samples was 0.0 (100% transcript 2) to 1.0 (50% transcript 2). The four NSCLC samples with no visible tumour content as well as the commercially sourced standards also fell within this range. Figure 5.7 also highlighted the fact that the RNA controls had the lowest *NF1* expression compared to the NSCLC samples. An independent T test demonstrated significant decrease ($p < 0.001$) in expression of all three probes in the controls compared to the NSCLC samples.

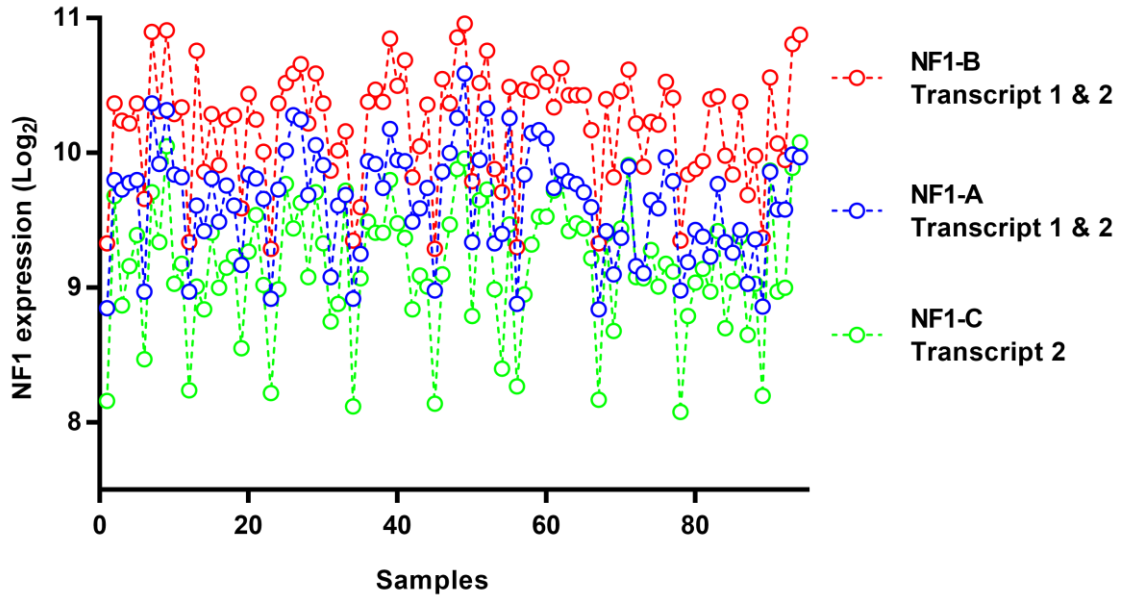


Figure 5.7. Expression of *NF1* in 84 NSCLC cases and the RNA controls. NF1-A & B detect *NF1* transcript 1 & 2. NF1-C detects transcript 2. The 9 lowest inverted peaks represent the controls. An Independent Student T test demonstrated all controls had reduced expression in all probes compared to the NSCLC samples (p = 0.001).

5.5.1 *NF1* Copy Number and NF1 mRNA Transcription

When investigating the relationship between *NF1* copy number and *NF1* mRNA expression initial analysis demonstrated no clear linear relationship with a R^2 coefficient of 0.064. As there appeared to be a monotonic relationship, a Spearman correlation assessment was utilised. The correlation coefficient (r_s) did demonstrate a slight positive correlation (r_s = 0.210), however this was not significant (p = 0.101).

5.5.2 *NF1* copy number and promoter variants

To investigate the possibility that the variants observed within the promoter region could have an effect on *NF1* transcription we looked at the expression of the samples and where they fell within the total percentile range. Only NF1-220 fell below the 5th percentile with all other cases with promoter variants being expressed at similar levels. This does suggest c.-468A>T could function as an inhibitor of *NF1* transcription; however this is only circumstantial and cannot be confirmed using this datas*et al*one.

Table 5.2. *NF1* mutations upstream of the transcription start site.

Sample ID	RS number	cDNA	NF1-A expression (Log2)	NF1 expression percentile
NF1-132	rs17879128	c.-459A>C	9.49	50 -75
NF1-148	rs922504932	c.-356C>T	8.94	10 -25
NF1-155	rs886052790	c.-209C>A	9.3	25 -50
NF1-172	N/A	c.-93C>G	9.55	50 -75
NF1-220	rs144759836	c.-468A>T	8.77	5

5.6 *NF1* Copy Number Change and Clinical Information

To investigate any correlation in *NF1* transcription between subtype, gender, and stage we used a 3 way ANOVA. Stage 1 and 2 were grouped together and stage 3 and 4 patients to form stage 1-2, and stage 3-4 groups. A significant 3 way interaction was observed between gender, subtype, and stage ($p = 0.014$). Further analysis using a 2 way ANOVA demonstrated there was a statistically significant simple two-way interaction between stage and subtype ($p = 0.018$) for males. Males with ADC and stage 1 or 2 disease demonstrated an increase in *NF1* expression compared to other groups.

5.7 Are the MEK and RAS Gene Expression Signatures Predictive of *BRAF* and *EGFR* Variants?

Whilst the RAS and MEK signatures have both been shown to be predictive of *KRAS* oncogenic drivers, it has yet to be established whether these signatures are predictive of *EGFR* or *BRAF* drivers (Dry et al., 2010a, Ahn et al., 2017, Brant et al., 2017, Loboda et al., 2010a). Our gene expression signature reproducibility work suggested that the MEK signatures have the ability to predict *EGFR* and *BRAF* mutation status. The RAS signature demonstrated that the threshold between *KRAS* and the *MAPK* negative controls has distinct margins, but the threshold between *EGFR* and the *MAPK* negative controls were almost indistinguishable.

To further investigate this, we divided our patients into *EGFR* positive, *KRAS* positive, *BRAF* positive, and *MAPK* negative groups. Comparing the signature scores across the groups would indicate if they could be predictive of not only *KRAS*, but also *EGFR* and *BRAF* variants. The grouped samples included *BRAF* (n = 3), *EGFR* (n = 7), *KRAS* (n = 15), and no known activating mutations (n = 44). The four samples which displayed both *MAPK* and *NF1* variants were placed in the *MAPK* positive group. In addition, to determine the effect that low allelic fraction of variants have on the gene expression signature, we repeated the analysis but excluded oncogenic samples with <30% of the expected AF in relation to tumour content. This reduced the *EGFR* (n = 5) and the *KRAS* group (n = 10) in size. Three of the samples which displayed both *MAPK* and *NF1* variants were in the excluded group. Initial analysis via boxplots determined we had one extreme outlier in the MEK 18 and 6 signatures from the same sample within the *MAPK* negative group. NF1-126 demonstrated a MEK 18 signature score of 7.11 and a MEK 6 score of 6.21. As this violated an assumption for ANOVA analysis we replaced the low outliers with the second lowest score, 7.11 was replaced with 7.97 and 6.21 was replaced with 7.40. To confirm this had minimal effect on the results described in this chapter we repeated the analysis keeping the outlier. No changes to any of the overall significant differences or similarities based on the ANOVA results were observed (data not shown).

5.7.1 *EGFR*, *BRAF* and *KRAS* Analysis via the MEK 18 Gene Expression Signature

We analysed the MEK 18 signature score for all *BRAF*, *EGFR*, *KRAS*, and *MAPK* negative groups based on genotype alone with a one way ANOVA. A Levene's test showed that homogeneity of variances was not assumed ($p < 0.05$). The Welsh ANOVA demonstrated significant differences between groups ($p = 0.011$). A *post hoc* Games-Howell test determined two significant differences. One between the *MAPK* negative and the *KRAS* groups ($p < 0.001$) and the other between the *MAPK* negative and *EGFR* groups ($p = 0.47$), shown in Figure 5.8A.

The analysis was repeated excluding the samples with variants at low AF. A Levene's test showed that homogeneity of variances was not assumed ($p < 0.05$). The Welsh ANOVA demonstrated significant differences between groups ($p = 0.008$). A *post hoc* Games-Howell test determined two significant differences. One between the *MAPK* negative and the *KRAS* groups ($p < 0.001$) and the final between the *KRAS* and *EGFR* group ($p = 0.033$) shown in Figure 5.8B. No significance was observed between *MAPK* negative and *EGFR* ($p = 0.132$)

For both sets of analysis with and without low level variants using the MEK 18 signature demonstrated the *KRAS* cases to have an increased signature score over the *MAPK* negative. Removing the low AF variants did not correlate with the low signature score within the groups. The two *EGFR* variants excluded were the ones with the highest MEK 18 signature score. This resulted in the significance displayed when including the low AF samples between *MAPK* negative and *EGFR* groups being lost. The five removed from the *KRAS* group did include the one with the lowest signature score, the rest were mid-level (9.25–9.54) signature scores. Both sets of analysis suggest the MEK 18 signature score is predictive of *KRAS* but not *BRAF* oncogenic variants. Whilst initially *EGFR* did show an increased score compared to the *MAPK* negative, removing the low AF variants resulted in a significant difference between the *EGFR* and *KRAS* being observed.

5.7.2 *EGFR*, *BRAF* and *KRAS* Analysis via the MEK 6 Gene Expression Signature

We analysed the MEK 6 signature for *BRAF*, *EGFR*, *KRAS*, and *MAPK* negative groups based on genotype alone using a one way ANOVA. A Levene's test showed that homogeneity of variances was not assumed ($p < 0.05$). The Welch ANOVA demonstrated significant differences between the groups ($p = 0.005$). A *post hoc* Games-Howell test determined two significant differences. One between *MAPK* negative and the *KRAS* group ($p < 0.001$), shown in Figure 5.9A, and the second between the *MAPK* negative and the *EGFR* group ($p = 0.001$).

The analysis was repeated excluding the samples with variants at low AF. A Levene's test showed that homogeneity of variances was not assumed ($p < 0.05$). The Welch ANOVA demonstrated significant differences between the groups ($p = 0.004$). A *post hoc* Games-Howell test determined two significant differences. One between *MAPK* negative and the *KRAS* group ($p < 0.001$), shown in Figure 5.9B, and the second between the *MAPK* negative and the *EGFR* group ($p = 0.001$).

For both sets of analysis with and without low level variants using the MEK 6 signature demonstrated identical groupings. The MEK 6 signature showed an increased significance in predicting not only *KRAS* drivers but also *EGFR* drivers over the MEK 18 signature. Unlike the MEK 18 signature scores, three out of the four *KRAS* samples excluded due to low AF were the ones with the lowest signature scores. As with the MEK 18 signature the *EGFR* samples removed because of low AF were the ones with the highest scores.

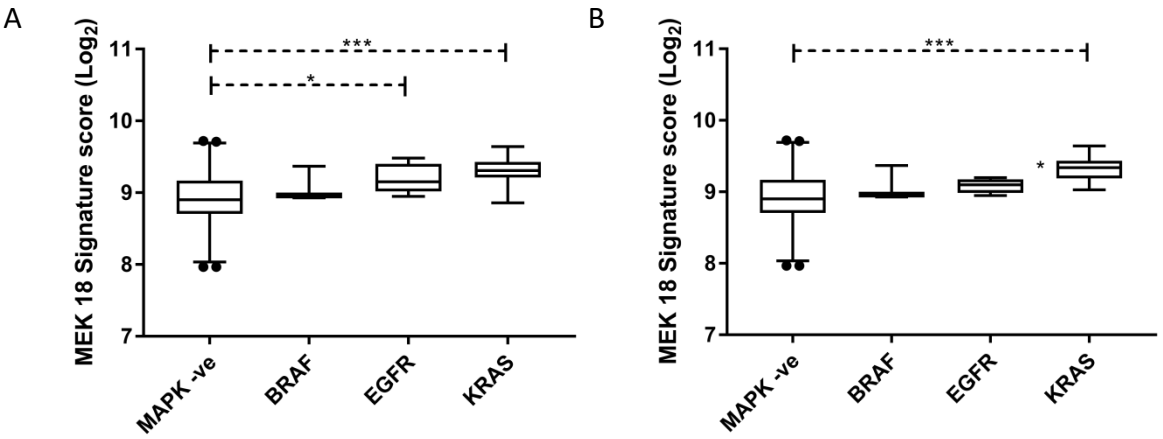


Figure 5.8. MEK 18 signature scores. A) Groups based on genotype alone. **B)** Groups bases on genotype with low allelic fraction samples excluded. Both A and B demonstrate the overall signature score is increased in the *KRAS* group.

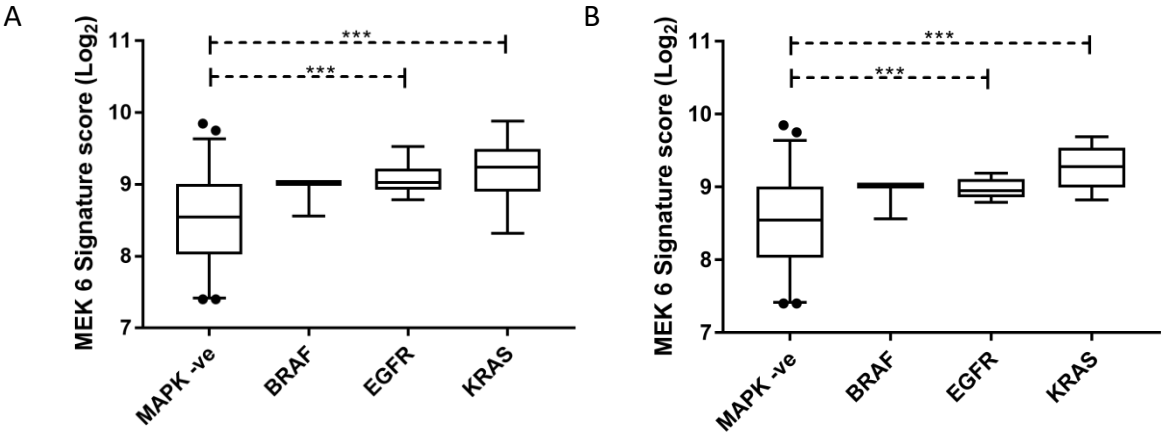


Figure 5.9. MEK 6 gene expression signature scores. **A)** Groups based on genotype alone. **B)** Groups bases on genotype with low allelic fraction samples excluded. Both A and B demonstrate identical grouping. The cases with *KRAS* and *EGFR* variants displaying significant differences to the *MAPK* negative group.

5.7.3 *EGFR*, *BRAF,* and *KRAS* Analysis via the RAS Gene Expression Signature

We analysed the *BRAF*, *EGFR*, *KRAS*, and *MAPK* negative groups based on genotype alone using the RAS signature using a one way ANOVA. A Levene's test showed that homogeneity of variances was assumed ($p > 0.05$). The ANOVA demonstrated significant differences between the groups ($p = 0.012$). A *post hoc* Tukey HSD test determined the one significant difference between *MAPK* negative and the *KRAS* group ($p < 0.023$), shown in Figure 5.10A.

The analysis was repeated excluding the samples with variants at low AF. A Levene's test showed that homogeneity of variances was not assumed ($p < 0.05$). The Welch ANOVA demonstrated significant differences between the groups ($p = 0.012$). A *post hoc* Games-Howell test determined two significant differences, between *MAPK* negative and the *KRAS* group ($p = 0.003$), shown in Figure 5.10B, and the other between the *KRAS* and the *EGFR* group ($p = 0.003$).

As with both the MEK 18 and MEK 6 signatures both sets of analysis with and without low level variants using the RAS signature demonstrated the *KRAS* group to have an increased score over the *MAPK* negative group. However, the analysis with the low AF variants excluded resulted in a significant difference also being observed between *KRAS* and *EGFR*. Unlike the MEK 18 and 6 signatures the *EGFR* variant excluded due to low AF was the one with the lowest signature score. *KRAS* lost the highest and the lowest signature scores along with 2 others.

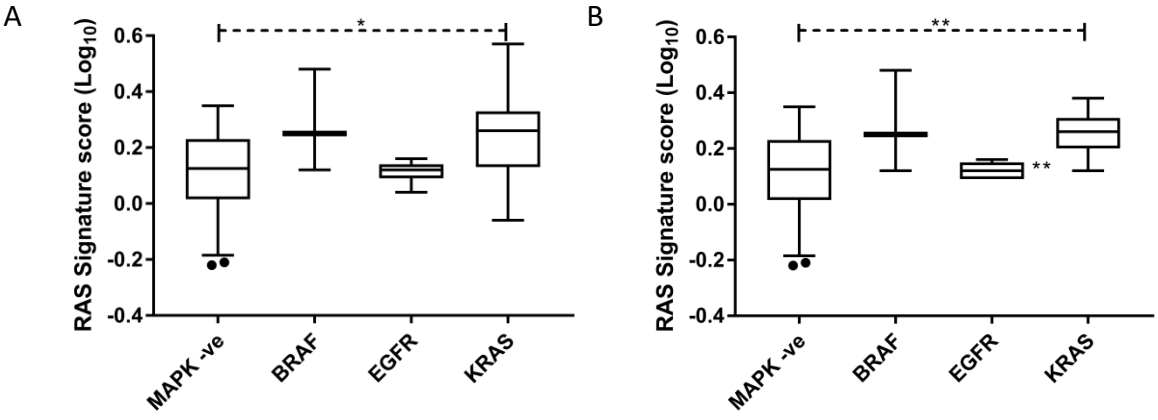


Figure 5.10.RAS gene expression signature score. A) Groups based on genotype alone. **B)** Groups bases on genotype but low allelic fraction samples have been excluded.

5.8 Gene Expression Signatures, Sensitivity, and Specificity

Our initial data demonstrated that all three signatures are predictive of *KRAS* positive cases in clinical samples. It also demonstrates that samples with low AF oncogenic variants have minimal influence over the signature scores. None of the signatures were able to predict *BRAF* status, however it has to be noted that the number of *BRAF* cases in our cohort was limited. Unlike the RAS and the MEK 18 signature the MEK 6 demonstrated the ability to be predictive of *EGFR* variants.

To give further weight to these signature scores and their ability to predict *MAPK* activation we performed a ROC analysis. We initially investigated all the signatures potential to predict *KRAS* mutations in our patient samples. In addition our data suggested the MEK signatures have the potential to predict *EGFR* oncogenic variants. We therefore included them in the ROC analysis. The *MAPK* negative group included cases with no known *MAPK* oncogenic variants and no *NF1* variants. The ROC analysis was performed excluding the oncogenic variants with low AF.

All signatures provided an acceptable level of discrimination (AUC > 0.7) at predicting *KRAS* oncogenic variants, shown in Table 5.3. The MEK 18 and MEK 6 however provided an excellent level of discrimination (AUC > 0.8) (Hosmer et al., 2013). The RAS signature displayed the lowest specificity which correlated with previous reports using the signature (Loboda et al., 2010a). Of the MEK signatures the MEK 6 provided an increased sensitivity and specificity over the MEK 18. In addition to this the MEK 6 signatures provided an excellent level of discrimination (AUC > 0.8) for detecting both *EGFR* and *KRAS* oncogenic variants in the cohort.

Based on this we proceeded with both MEK signatures to determine *NF1s* functional relevance. Whilst both signatures have the ability to predict *EGFR* oncogenic variants we only used *KRAS* samples as the *MAKP* positives during the next stages of analysis.

Table 5.3. Receiver operator characteristics for the three signatures

Signature	AUC	95% CI	Cut off	Sensitivity (%)	Specificity (%)
MEK 18	0.809	0.730 to 0.934	9.04	90	61
MEK 18**	0.758	0.640 to 0.877	9.04	80	61
MEK 6	0.855	0752. to 0.957	8.82	100	68
MEK 6**	0.820	0.717 to 0.924	8.82	93	68
RAS	0.799	0.674 to 0.923	0.125	90	50

*Signature's ability to predict *KRAS* variants ** Signature's ability to predict *EGFR* and *KRAS* variants

5.9 Functional Relevance of *NF1* Variants in NSCLC

To determine if the *NF1* variants in our patient samples were functionally relevant and could upregulate MAPK activity we compared the cases with *NF1* variants against the *MAPK* positive group and the *MAPK* negative group. This resulted in 44 *MAPK* negative samples for all signatures. As *NF1* copy number demonstrated no correlation to *NF1* gene expression we did not take copy number into consideration. The MEK 18 and 6 signatures had 10 *KRAS* positives as the *MAPK* positive group. Of the 19 samples with *NF1* variants there were 9 which were predicted to be likely benign, 5 of which had high probability of *NF1* loss of function, and 6 with unknown significance, shown in Table 5.4. The four cases which displayed co-occurrence *MAPK* drivers and *NF1* variants were excluded from the *NF1* group, leaving 15 *NF1* cases. However, when cross-referenced with heterogeneity of the variants the number of high probability samples was reduced to 2 (NF1-180 and NF1-233) and the unknown significance to one (NF1-225) which were ubiquitously expressed throughout the tumour.

Table 5.4. *NF1* variants and predicted clinical significance

Sample ID	cDNA	Protein	Heterogeneity	Predicted Significance
NF1-115	c.3118A>G	p.Lys1040Glu	Heterogenic	Uncertain significance
NF1-120*	c.1466A>G	p.Tyr489Cys	Heterogenic	Uncertain significance
NF1-146**	1392+5_1392 +6delGAinsTT	-	Ubiquitous	Likely benign
NF1-147**	c.3468C>T	p.Asn1156Asn	Ubiquitous	Likely benign
NF1-150	c.5269-14C>G	-	Heterogenic	Uncertain significance
NF1-163	c.-22G>C	-	Heterogenic	Likely benign
NF1-163**	c.1897G>C	p.Asp633His	Ubiquitous	High significance
NF1-168	c.5785G>T	p.Glu1929*	Heterogenic	High significance
NF1-168*	c.3721C>T	p.Arg1241*	Heterogenic	Uncertain significance
NF1-170	c.3825C>T	p.Phe1275Phe	Heterogenic	Likely benign
NF1-178**	c.7595C>T	p.Ala2532Val	Ubiquitous	Likely benign
NF1-180	c.7245delA	p.(Leu2416Tyrfs*2)	Ubiquitous	High significance
NF1-201	c.435C>A	p.Leu145Leu	Heterogenic	Likely benign
NF1-205	c.4311G>A	p.Arg1437Arg	Ubiquitous	Likely benign
NF1-210	c.169G>T	p.Gly57Cys	Heterogenic	Uncertain significance
NF1-217	c.3634delG	Val1212Serfs*3	Heterogenic	High significance
NF1-217	c.7339G>T	p.Glu2447*	Heterogenic	High significance
NF1-220*	c.1018T>G	p.Ser340Ala	Heterogenic	Likely benign
NF1-225	c.6781C>T	p.His2261Tyr	Ubiquitous	Uncertain significance
NF1-227	c.6148-2A>T	-	Heterogenic	Likely benign
NF1-233	c.2178G>A	p.Val726Val	Heterogenic	Likely benign
NF1-233	c.2681T>C	p.Phe894Ser	Ubiquitous	Uncertain significance
NF1-233	c.2675_2676in sA	p.(Ser892Argfs*14)	Ubiquitous	High significance
NF1-238	c.3481C>G	p.Leu1161Val	Heterogenic	Uncertain significance
NF1-238	c.3249C>G	p.Leu1083Leu	Heterogenic	Likely benign

*Potential pseudogene origin. **Germline origin. Samples shown in red are ubiquitously expressed throughout the tumour and of high or unknown significance of causing *NF1* loss of function. Samples in green have oncogenic MAPK variants.

5.9.1 Functional Relevance of *NF1* using the MEK 18 Gene Expression Signatures

To investigate the functional relevance of NF1 variants we compared samples grouped into *MAPK* positive, *MAPK* negative, and the 15 samples with *NF1* variants using a one way ANOVA. As many of the *NF1* variants were either, likely benign or low level heterogenetic, we expected to see no difference between the *MAKP* negative group and the *NF1* group.

A Levene's test showed that homogeneity of variances was not assumed ($p < 0.05$). A Welch ANOVA demonstrated significant differences in the signature scores between the groups ($p < 0.001$). A *post hoc* Games-Howell test determined a significant increase in the *MAKP* positive group with a signature score mean of 9.32 $\log_2$ or 639 individual RNA counts compared to the *MAPK* negative group 8.91 (478 RNA counts) ($p < 0.001$). The same increase was also observed with the *MAPK* positive group and the *NF1* loss group with a mean of 8.8 (452 RNA counts) ($p < 0.001$). As expected no significance was observed between the *MAPK* negative vs. *NF1* loss groups ($p = 0.464$), shown in Figure 5.11.

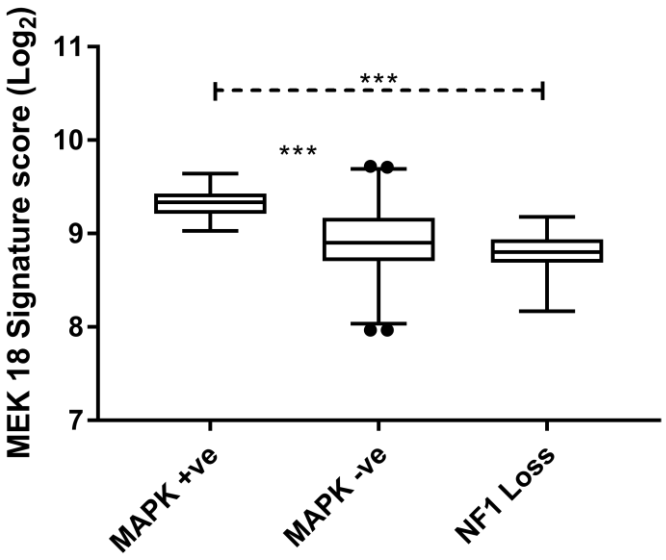


Figure 5.11. MEK 18 gene expression signature. Whiskers represent 5–95 percentiles. Cases with *KRAS*, oncogenic variants (*MAPK* +ve). Cases with no known MAPK activating variants (*MAPK* –ve). Cases which have a potential *NF1* loss of function variant (*NF1* loss). Significant differences observed between the *MAPK* positive group vs. *MAPK* negative ($p < 0.001$) and the *NF1*-loss group ($p < 0.001$). No significance was observed between the *MAPK* negative and the *NF1* loss groups.

Investigating the samples individually, one of the likely benign samples NF1-178 (p.Ala2532Val) with a score of 9.18 (580 counts) was above the 9.02 (519 counts) cut off

limit (90% sensitivity and 62% specificity), shown in Figure 5.12. NF1-178 carried a germline *NF1* variant which was ubiquitously expressed. The ubiquitously expressed high probability samples NF1-180 and NF1-233 had a score of 8.69 (413 counts) and 8.91 (481 counts) with the unknown significance case NF1-225 scoring 8.78 (440 counts). The previously reported pathogenic variants, NF1-168 and NF1-120, both of which were just below the –30% expected AF, were below the 9.02 cut off limit.

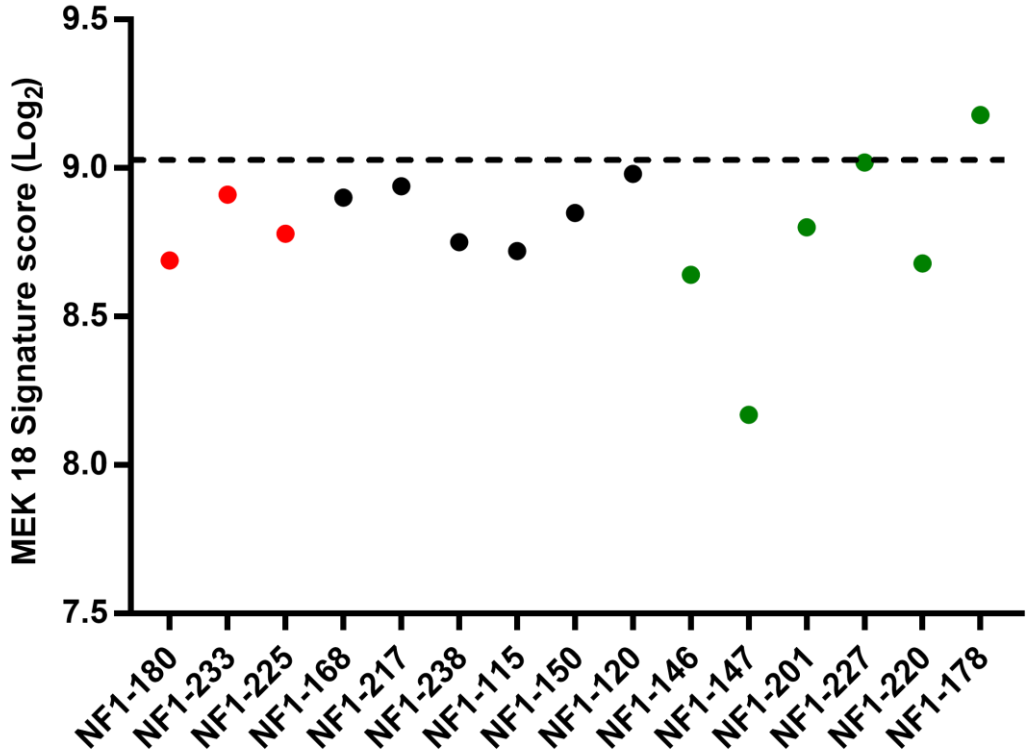


Figure 5.12. MEK 18 gene expression signature of *NF1* samples. Dotted line represents the 9.02 cut off with a sensitivity of 90% and a specificity of 61%. Red ubiquitously expressed two high probability *NF1* loss cases and one unknown significance *NF1* variant. Black unknown significance *NF1* variants. Green likely benign *NF1* variants.

5.9.2 Functional Relevance of *NF1* using the MEK 6 Gene Expression Signature

As with the MEK 18 signature we compared the grouped samples using a one way ANOVA. A Levene's test showed that homogeneity of variances was not assumed ($p < 0.05$). A Welch ANOVA demonstrated significant differences in the signature scores between the groups ($p < 0.001$). A *post hoc* Games-Howell test determined a significant increase ($p < 0.001$), in the *MAPK* positive group with a signature score mean of 9.24 or 605 counts compared to the vs. *MAPK* negative group 8.49 (360 counts). The same significant increase ($p = 0.001$) was also observed with the *MAPK* positive group compared to the *NF1* loss groups with a mean score of 8.54 (372 counts). No significance was observed between the *MAPK* negative vs. *NF1* loss groups ($p = 0.942$), shown in Figure 5.13.

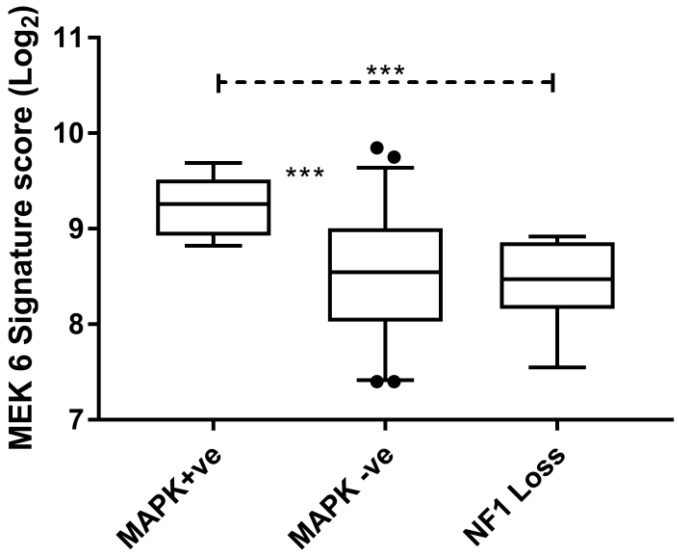


Figure 5.13. MEK 6 gene expression signature. Whiskers represent 5–95 percentiles. Cases with *KRAS*, oncogenic variants (*MAPK* +ve). Cases with no known MAPK activating variants (*MAPK* –ve). Cases which have a potential *NF1* loss of function variant (*NF1* loss). Significant differences observed between the *MAPK* positive group vs. *MAPK* negative ($p < 0.001$) and the *NF1*-loss group ($p = 0.001$). No significance was observed between the *MAPK* negative and the *NF1* loss groups.

Utilising the MEK 6 signature, 4 samples were above the 8.82 (452 counts) cut off limit (100% sensitivity and 68% specificity). Two of the variants were in the likely benign group including, NF1-178 with a signature score of 8.86 (465 counts) also identified by the MEK 18 signature and NF1-201 with 8.90 (478 counts). The MEK 6 signature also identified NF1-115 and NF1-217 from the unknown significance group as just above the threshold. NF1-217 was a high probability case with 2 null variants (Val1212Serfs*3 and p.Glu2447*), however, it was low level AF and calculated to be found in <30% of the tumour. NF1-115 was also a low level AF sample calculated to be found in <20% of the tumour. As with the MEK 18 signature the ubiquitously expressed high probability samples NF1-180 and NF1-233 had a score of 8.70 (416 counts) and 8.18 (290 counts) retrospectively, with the unknown significance NF1-225 having a score of 8.68 (396 counts). The previously reported pathogenic variants, NF1-168 and NF1-120 were also below the 8.82 cut off.

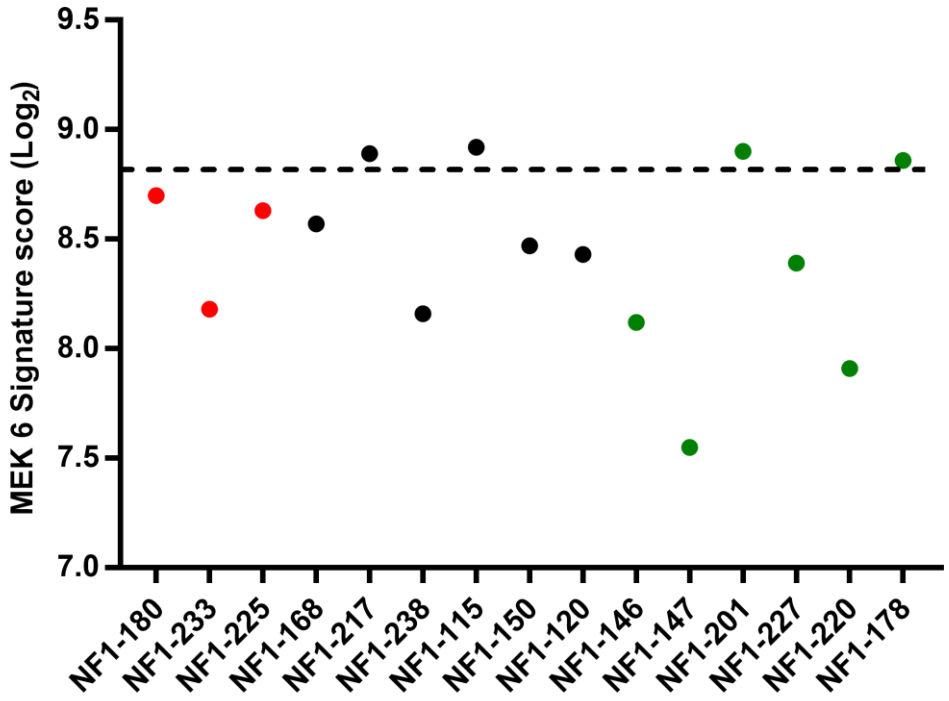


Figure 5.14. MEK 6 gene expression signature of *NF1* samples. Dotted line represents the 8.82 cut off with a sensitivity of 100% and a specificity of 68%. Red ubiquitously expressed two high probability *NF1* loss cases and one unknown significance *NF1* variant. Black unknown significance *NF1* variants. Green likely benign *NF1* variants.

5.9.3 MEK 6 Gene Expression Signature in the TGCA Pan-Lung Data Set: *EGFR*, *BRAF* and *KRAS* Analysis via the MEK 6 Gene Expression Signature

We analysed the MEK 6 signature for *BRAF*, *EGFR*, *KRAS*, and *MAPK* PIK3CA negative groups using a one way ANOVA. A Levene's test showed that homogeneity of variances was not assumed ($p < 0.05$). The Welch ANOVA demonstrated significant differences between the groups ($p < 0.001$). A *post hoc* Games-Howell test determined a significant increase in the MEK 6 signature in the *KRAS*, *EGFR*, and *BRAF* groups compared to the MAPK PIK3CA negative group Figure 5.15. The *KRAS* group demonstrated the largest increase ($p < 0.001$) with mean of 10.4 Log_2 (1351 counts) compared to the 9.5 Log_2 (724 counts) of the MAPK PIK3CA negative group demonstrating a 0.9 fold increase.

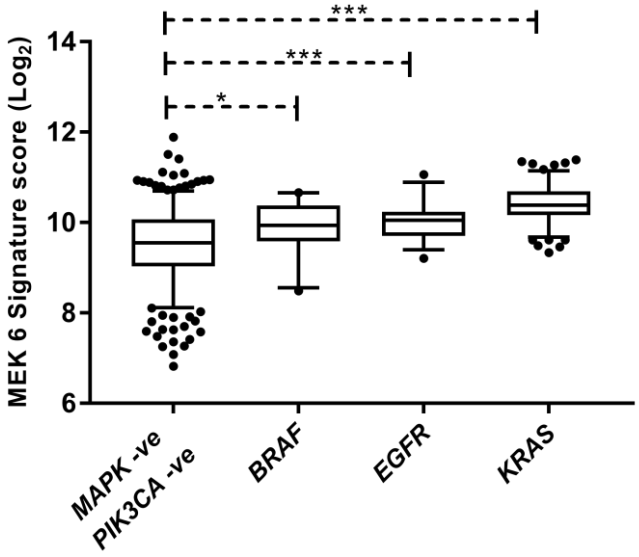


Figure 5.15. Pan-Lung MAPK positive MEK 6 gene expression signature scores. Groups bases on genotype with low allelic fraction samples excluded. *MAPK PIK3CA* negative (387), *KRAS* (136), *EGFR* (33), *BRAF* (21).

5.9.4 Pan-Lung Data Set: PI3K/AKT/mTOR pathway Analysis via the MEK 6 Gene Expression Signature

Using the Pan-Lung dataset we investigated if the MAP 6 signature was influenced by the presence of *PIK3CA* oncogenic variants (*E545K, E542K, H1047R, H1047L*), *PIK3CA* amplification and *PTEN* homozygote loss, in relation to the MAPK PI3K negative group. A Levene's test showed that homogeneity of variances was assumed ($p > 0.05$). A one way ANOVA demonstrated significant differences between the groups ($p = 0.006$). A post hoc Tukey HSD test determined significant reduction in the MEK 6 signature score between the PIK3CA oncogenic variants group ($p = 0.041$), and the PIK3CA amplification group ($p = 0.046$) in relation to the MAPK PIK3CA negative group. No difference was observed between the MAPK PIK3CA negative group and the *PTEN* loss group. This and the results in section 5.10.1 suggest an increase in the MEK 6 signature is predictive of MAPK activation, but not PI3K/AKT/mTOR activationas all group demonstrated a reduction in the signature score.

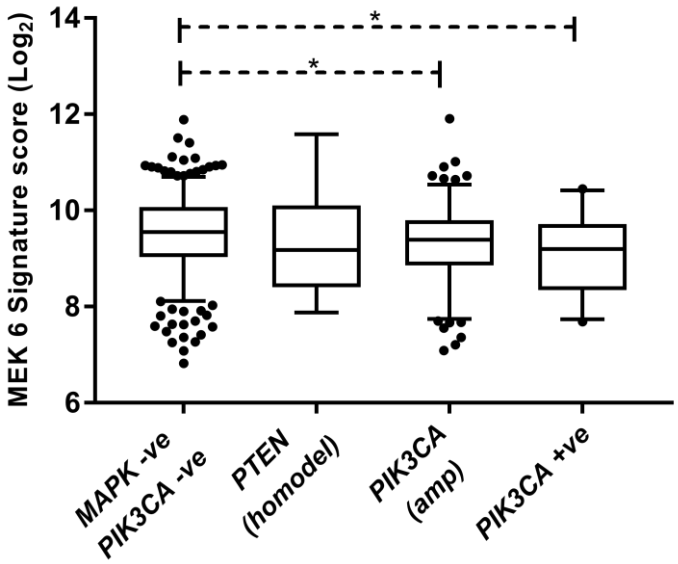


Figure 5.16. MEK **6 gene expression signature scores**. Groups bases on genotype with low allelic fraction samples excluded. *MAPK PIK3CA* negative (n =387),*PIK3CA +ve (E545K, E542K, H1047R, H1047L))(n = 26), PIK3CA amp (n = 153), PTEN homodel (n = 16).*

5.9.5 Pan-Lung Data Set: MEK 6 Gene Expression Signature, Sensitivity and Specificity

Analysis of the Pan-Lung data with the MEK 6 signature demonstrated a significant increase in the *KRAS*, *EGFR*, and *BRAF* groups in relation to the *MAPK PIK3CA* negative group. We next performed a ROC analysis to see if it correlated with the data from this study. The MEK 6 signatures provided an acceptable level of discrimination (AUC > 0.7) at predicting *KRAS+EGFR+BRAF* oncogenic variants as one group, shown in Table 5.5. The comparison of *MAPK PIK3CA* negative group and the *KRAS* group, the signature provided an excellent level of discrimination (AUC > 0.8) (Hosmer et al., 2013).

The AUC demonstrated a high level of reproducibility for predicting *KRAS* oncogenic variants between our dataset (AUC 0.855) shown in Table 5.3 and the Pan-Lung data set (AUC 0.841) shown below. The MEK 6 signature also provided similar results at predicting *KRAS* and *EGFR* oncogenic variants; our study (AUC 0.820) compared to the Pan-Lung study (AUC of 0.797). The *BRAF* variants in our study did not show a significant difference to the MAPK negative group, but in the Pan-Lung dataset did demonstrate an acceptable level of discrimination with an AUC of 0.780 when comparing the *KRAS*, *EGFR*, and *BRAF*, cases with the *MAPK PIK3CA* negative group.

Table 5.5. Receiver operator characteristics for the MEK 6 signatures

MEK 6 Signature	AUC	95% CI	Cut off	Sensitivity (%)	Specificity (%)
KRAS	0.841	0.808 to 0.875	9.8	90	62
KRAS EGFR	0.797	0.762 to 0.832	9.8	85	62
KRAS EGFR BRAF	0.780	0.744 to 0.816	9.8	81	62

Signature's ability to predict *KRAS*, *KRAS+EGFR*, and *KRAS+EGFR+BRAF* variants

5.9.6 Pan-Lung Data Set: Functional Relevance of *NF1* and *RASA1* using the MEK 6 Gene Expression Signature

Next to determine if the *NF1* and *RASA1* variants in our patient samples were functionally relevant in regulation of the MAPK pathway activity based on the MEK 6 signature we analysed the *NF1, RASA1* groups, and *MAPK* PIK3CA negative groups using a one way ANOVA. A Levene's test showed that homogeneity of variances was not assumed ($p < 0.05$). The Welch ANOVA demonstrated significant differences between the groups ($p < 0.001$). A *post hoc* Games-Howell test determined a significant increase in the MEK 6 signature in the *NF1 RASA1* group ($p = 0.002$) compared to the *MAPK PIK3CA* negative group which had a mean signature 9.5 (724 reads), Figure 5.17. All other *NF1* and *RASA1* groups did not demonstrate a significant increase in the MEK signature in relation to the *MAPK PIK3CA* negative group. *MAPK PIK3CA* negative, *NF1* (H, M, L), and *RASA1* groups form a homogenous subset $p = 0.947$ with mean signature scores ranging from 9.4 – 9.6 (675 – 776 reads). As shown previously the *KRAS* group demonstrated the largest mean score of 10.4 (1351 reads) followed by the cases with co-occurring *NF1* and *RASA1* variants, mean 10.2 (1176 reads) again forming a homogenous subset $p = 0.936$. Between our data set and the Pan-Lung dataset the difference between the MAPK negative group and the KRAS group was identical at 0.9 Log_2.

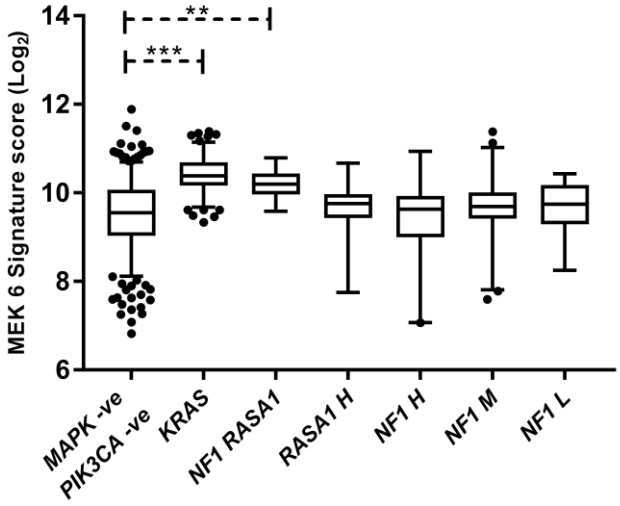


Figure 5.17. Pan-Lung NF1 And RASA1 positive MEK 6 gene expression signature scores. Groups bases on genotype with low allelic fraction samples excluded. *NF1* groups; high probability loss of function (H) (n = 19), unknown significance (M)(n = 30), and likely benign groups (L)(n = 12), MAPK PIK3CA negative (n = 387), KRAS (n = 135), NF1 RASA1 (n = 10), RASA1 (n = 21)

The *NF1* high probability loss of function group in the Pan-Lung dataset which included null variants downstream of the GAP domain and 3 homozygote deletions of *NF1* showed no significance of the MEK 6 signature in relation to the MAPK PIK3CA negative group (p = 0.998). This closely correlates with our data shown in section 5.9.2 (p = 0.942). Investigation of the 3 *NF1* homozygote deletions demonstrated a signature scores ranging from 8.31 – 10.94 Log_2 (317 – 1965 reads). This wide range suggests other factors could be involved in regulation of the MAPK pathway and *NF1* loss alone is not enough to drive the this pathway in NSCLC, based on the MEK 6 signature score. The *NF1* homozygote deletion with the lowest signature score also had a co-occurring *PTEN* homozygote deletion and all of the *NF1* high probability group co-occurred with *TP53* variants. The suggestion that other factors are required to drive the MAPK pathway in addition to *NF1* null variants is also observed when looking at the cases with co-occurring *NF1* and *RASA1* variants. This group demonstrated a significant increase in the MEK 6 signature in relation to the MAPK PIK3CA negative group. However, unlike the KRAS group the *NF1* and *RASA1* did not show any significant difference with any of the *NF1* groups or the *RASA1* group.

5.9.7 Pan-Lung Data Set: Functional Relevance of *NF1* and *RASA1* using a PI3K/AKT/mTOR Gene Expression Signiture

Next to investigate if *NF1* or *RASA1* variants could have an effect in regulation of the PI3K/AKT/mTOR pathway we looked a related gene expression signature. The PI3K/AKT/mTOR (CMAP) signature was initially proposed by (Creighton et al., 2010)) Craighton et al. 2010, but had more recently been refined in (Zhang et al., 2016b) Pan-Cancer proteogenomic study of PI3K/AKT/mTOR pathway alterations. The signature has been shown to correlate to phosphorylation status of key members of the PI3K/AKT/mTOR pathway and correlates to inhibition of the same pathway. However, unlike the MEK 6 signature, its power to predict genotypic changes has not been investigated. Normalised RNA seq gene expression was downloaded as described in method section 2.2.10. Cases with < 0.2 AF as shown in

Table 4.18. Oncogenic *MAPK, PIC3CA, NF1, and RASA1* variants in the Pan-Lung cohort of patients

4.18 were removed prior to analysis.

Initial analysis of the individual genotypic groups using the 190 gene CRAS signature via a heatmap did demonstrate a visually observable difference in the 136 gene, up regulated arm, Figure 5.18. This arm demonstrated reduced expression in the *MAPK* positive groups, in relation to increased expression in the *PI3K* positive group. Visual analysis of the 56 gene down arm was less clear.

The *MAPK PI3K* negative group displayed a mix of both increased and decreased expression in the up regulated arm of the CMAP signature as observed in Figure 5.17. Over half of the MAPK PI3K negative group demonstrated an increase in the expression similar to what is observed in the PI3K/AKT/MTOR groups, whilst the remaining demonstrated a decrease similar to the *MAPK* positive cases. When investigating the 36 gene down regulated arm there was less visual difference between any of the groups. The combination of increased and reduced expression in the *MAPK PI3K* negative group could represent potentially unknown drivers regulating MAPK and PI3K/AKT/mTOR pathways.

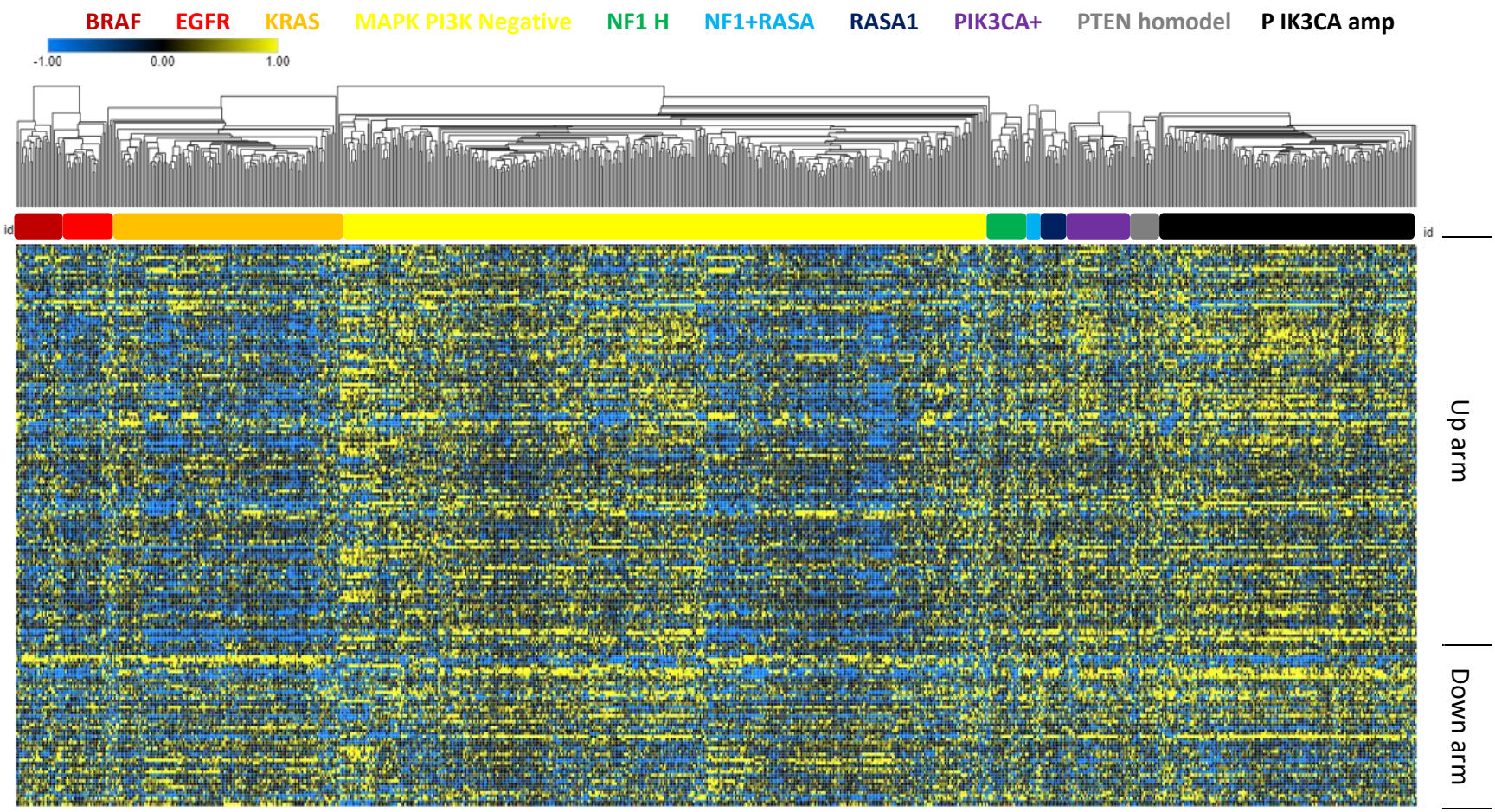


Figure 5.18. Pan-Lung CMAP 190 gene expression signature. Groups based on genotype with low allelic fraction samples excluded. Blue represents lower expression (relative to the mean of the specific gene across all cases), yellow represents higher expression. 136 genes in the up regulated arm and 56 genes in the down regulated arm. Bright yellow/blue denotes change of ≥ 1 $\log_2$ from the mean. *BRAF (n = 21), EGFR* (n = 33), KRAS (n = 135), MAPK PIK3CA negative (n = 387), *NF1* high probability loss of function (H) (n = 19), *NF1+RASA1* (n = 10), *RASA1* (n = 18), *PIK3CA* +ve (*E545K*, *E542K*, *H1047R*, *H1047L*)(n = 26), *PIK3CA* amp (n = 153), *PTEN* homodel (n = 16).

To further investigate this signature we analysed the mean of all the genes in the up regulated arm to determine if this could distinguish between groups in a similar manner to the MEK and RAS signatures used previously.

A Levene's test showed that homogeneity of variances was not assumed ($p < 0.05$). The Welch ANOVA demonstrated significant differences between the groups ($p < 0.001$). A post hoc Games-Howell test determined significant increase in the CMAP up regulated 136 gene signature in the *PTEN* homozygote deletion group ($p = 0.024$) and *PIK3CA* amplification group ($p < 0.001$) in relation to the *MAPK PI3K* negative group. No significant difference was observed between the *MAPK PI3K* negative group and the *PIK3CA* positive group. Unexpectedly, the CMAP 136 up regulated signature did demonstrate a significant reduction in the signature score between the MAPK positive groups (*BRAF* $p = 0.037$, *EGFR* $p = 0.001$, and *KRAS* $p < 0.001$) in relation to the MAPK PI3K negative group. No significant difference was observed between the *MAPK PI3K* negative group and the *NF1* H, *NF1*+*RASA1*, and *RASA1* groups shown in Figure 5.19.

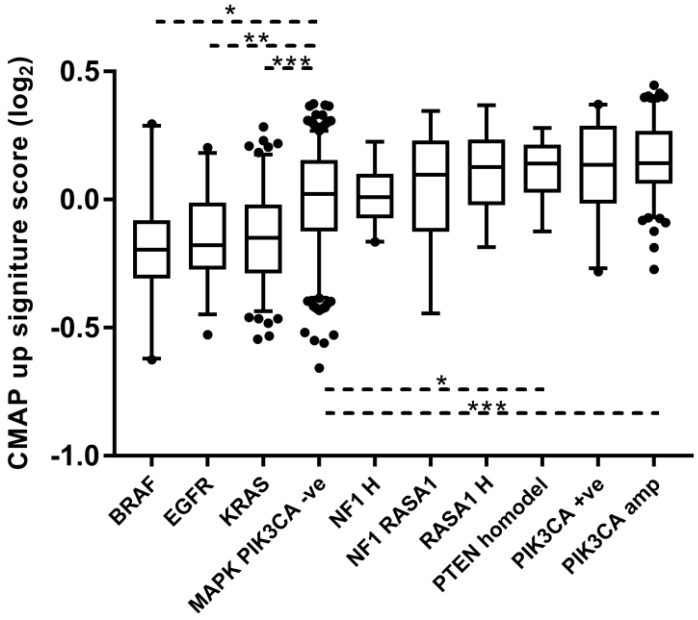


Figure 5.19 Pan-Lung CMAP upregulated 156 gene expression signature scores based on individual groups. Group's bases on genotype with low allelic fraction samples excluded. The up regulated are of the CMAP signature (relative to the mean of the specific gene across all cases). *NF1* groups; high probability loss of function (H) (n = 19), unknown significance (M)(n = 30), and likely benign groups (L)(n = 12), MAPK PIK3CA negative (n = 387), KRAS (n = 135), NF1 RASA1 (n = 10), RASA1 (n = 21).

As the *MAPK PI3K* negative group could contain unknown drivers of both the MAPK and the PI3K/AKT/mTOR pathways, as we only excluded the main common drivers of this pathway described in Table 4.18, we decided to investigate the *MAPK* positive groups compared the *PI3K/AKT/mTOR* positive groups using the CMAP signature. Whilst there is still overlap between the range of the samples within the groups, the signatures resulted in two distinct subsets, with *EGFR*, *KRAS*, and *BRAF* forming one group ($p = 0.999$), and *NF1* H, *RASA1* H, *NF1*+*RASA1*, *PIK3CA* amp, *PIK3CA* +ve, *PTEN* homodel, forming another group ($p = 0.161$). The *MAPK* positive groups were down regulated in relation to the up regulated PI3K/AKT/mTOR positive groups and *NF1* and *RASA1* groups. Surprisingly the *NF1*+*RASA1* group was the only one not to show a significant increase in relation to the MAPK positive groups. A Welch ANAOA followed by a Games-Howell test determined up regulation of all groups in relation to the MAPK positive groups (*EGFR, KRAS, BRAF*) ($p < 0.021$), with the exception of the *NF1*+*RASA1* group, which was not significant ($p > 0.259$).

Next we grouped all *MAPK* positive samples together (*EGFR*, *KRAS*, and *BRAF*) and all of the PI3K/AKT/mTOR positive samples together (*PIK3CA* +ve, *PIK3CA* amp *PTEN* homodel) as both formed distinct sub-groups to examine the *NF1* and *RASA1* groups using the up arm of the CMAP signature. A Levene's test showed that homogeneity of variances was not assumed ($p < 0.05$). The Welch ANOVA demonstrated significant differences between the groups ($p < 0.001$), shown in Figure 5.20. A post hoc Games-Howell test determined significant increase in the CMAP up regulated 136 gene signature in relation to the MAPK positive group for the *NF1* H ($p < 0.001$), and *RASA1* group ($p < 0.001$). However, whilst the *NF1*+*RASA1* group did visually display an increase in relation to the MAPK positive group, this was not deemed statically significant ($p = 0.133$). This is due to the increased range observed of the CMAP signature score and the overlapping 1-3 quartile range of the *NF1*+*RASA1* and the MAPK positive group. One potential reason for this could be the relatively small number of co-occurring *NF1*+*RASA1* cases. The Pan-Lung data only shows co-occurrence for *NF1* and *RASA1* variants in 1% of cases of NSCLC.

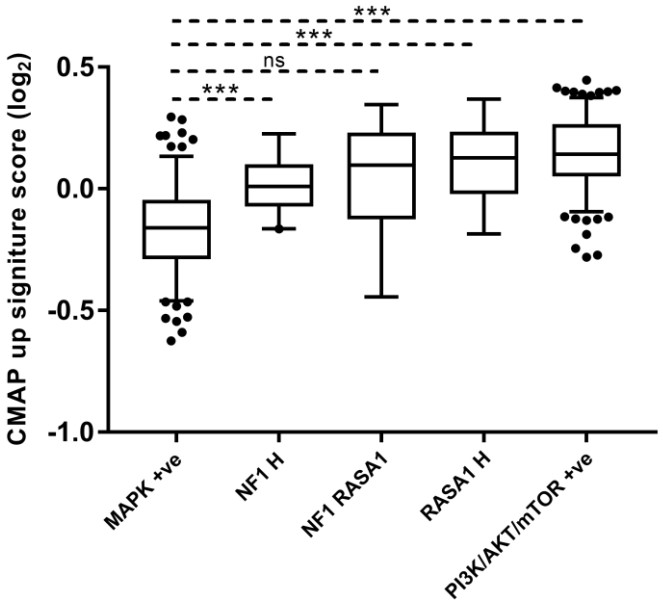


Figure 5.20. Pan-Lung CMAP 156 up regulates gene expression signature scores. Groups based on genotype with low allelic fraction samples excluded. Upregulated arm of the 136 gene CMAP signature (relative to the mean of the specific gene across all cases). MAPK+ve group (n = 190), NF1 H (n = 19), NF1+RASA1 (n = 10), RASA1 H (n = 18), and PI3K+ve (n = 195).

The up regulated arm of the CMAP signature does demonstrate the *NF1* H and *RASA1* H groups appear to have a functional relevance in up regulation of the PI3K/AKT/mTOR pathway in comparison to the MAPK positive group based on this signature. The log_2 difference between the means of the MAPK positive and PI3K/AKT/mTOR groups was 0.3. When looking at the original counts from the RNA seq data, the CMAP signature mean score for the up arm was 973 reads for the MAPK positive group, compared to up the upregulated counts of 1195 for the PI3K/AKT/mTOR positive group, 1121 for *NF1* H, 1098 for *RASA1* H and 1121 for the *NF1*+RASA1 group. Whist this fold change is less than what is observed in the MEK 6 signature there are still significant differences suggesting both NF1 and RASA1 regulate the PI3K/AKT/mTOR pathway.

Next we repeated the analysis shown in Figure 5.20 using the 56 gene down arm of the CMAP signature. A Levene's test showed that homogeneity of variances was assumed ($p > 0.05$). An ANOVA demonstrated significant differences between the groups ($p < 0.001$). A Tukey HSD post hoc test only demonstrated one significant difference with a slight increase in the PI3K/AKT/mTOR positive group in relation to the *MAPK* positive group, shown in Figure 5.21. However, the difference between the means of these groups was only 0.05 Log_2. When looking at the original reads from the RNA seq data, the CMAP signature mean score for the up arm was 825 counts for the MAPK positive group, compared to 854 counts for the PI3K/AKT/mTOR positive group, 800 for *NF1* H, 850 for RASA1H and 878 for the *NF1*+*RASA1* group. This was unexpected as we should have seen a reduction in the PI3K/AKT/mTOR positive group in relation to the MAPK positive group, suggesting the down arm of the CMAP signature still requires further optimisation.

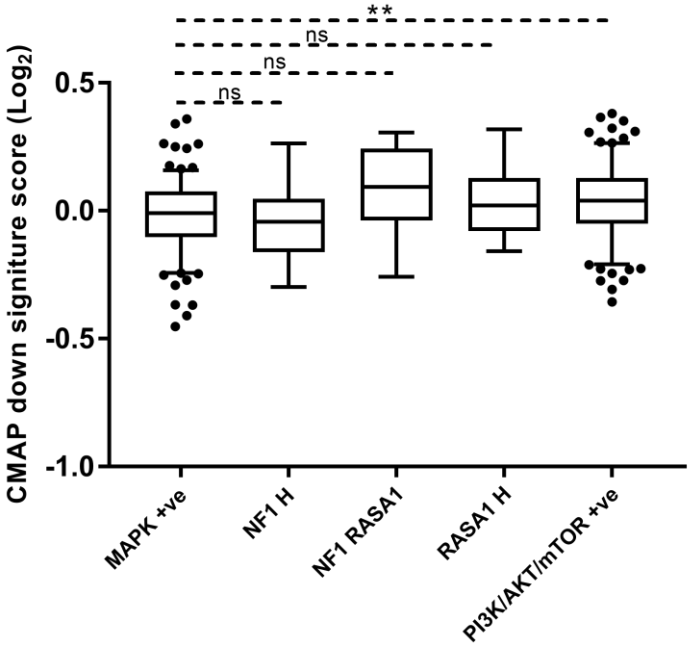


Figure 5.21. Pan-Lung CMAP 56 down regulated gene expression signature scores. Groups based on genotype with low allelic fraction samples excluded. Up-regulated arm of the 136 gene CMAP signature. *MAPK+ve* group (n = 190), *NF1* H (n = 19), *NF1*+*RASA1* (n = 10) *RASA1* H (n = 18), and *PI3K*+ve (n = 195).

5.10 Gene Expression signatures and Clinical Information

To investigate any correlations in the signature scores between subtype, gender, and stage we used a 3 way ANOVA. Stage 1 and 2 were grouped together and stage 3 and 4 patients to form stage 1-2, and stage 3-4 groups. Both the MEK 18 and MEK 6 signatures did not demonstrate any significant 3 way interaction ($p > 0.288$). However, a significant difference in subtype was noted using both signatures, as shown in Figure 5.22.

The MEK 18 demonstrated the ADC group (n = 47) to have the highest mean score over the SQCC group (n = 29) and NOS (n = 8). A Tukey HSD *post hoc* test demonstrated a significance between ADC and SQCC ($p = 0.015$). No difference was observed between ADC vs. NOS ($p = 0.068$) or AQCC vs. NOS ($p = 0.844$).

A Tukey HSD post-hoc test on the MEK 6 signature demonstrated a significance between ADC and NOS ($p = 0.004$) and SQCC and NOS ($p = 0.023$). No difference was observed between ADC vs. SQCC ($p = 0.689$).

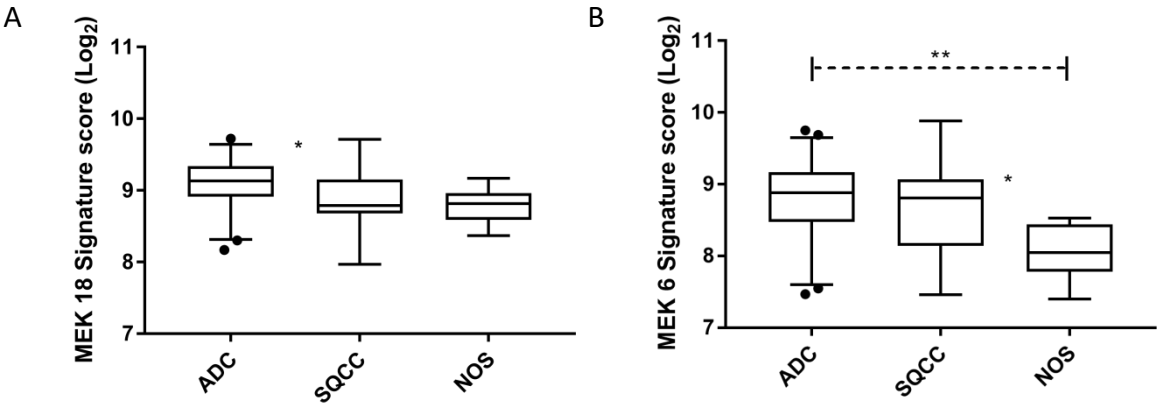


Figure 5.22. MEK 18 and MEK 6 gene expression signature score across subtypes.
A) MEK 18 gene expression signature score across subtypes B) MEK 6 gene expression signature score across subtypes

5.11 Discussion

Here it was demonstrated that the Nanostring nCounter is a highly robust platform for quantitative analysis of mRNA from degraded clinical FFPET samples. The platform's reproducibility with both high quality commercial standards and with highly degraded RNA was assessed. The Nanostring CodeSet which included over 200 genes demonstrated high levels of reproducibility both inter batch and intra batch. Of the 84 samples analysed, there were no failures or repeats required based on the internal or external QC measures. It was determined that the *NF1* copy number changes seen at the genetic level do not influence *NF1* mRNA expression. Furthermore, it was determined that *NF1* transcript 2 was the predominately expressed transcript in NSCLC. Previous reports of the relevance and predictive power of these signatures in relation to *KRAS* oncogenic variants through activation of the MAPK pathway were confirmed (Dry et al., 2010a, Brant et al., 2017, Loboda et al., 2010a). In a direct comparison of the three signatures that relate to the MAPK activation it was shown that the MEK 6 signature has increased sensitivity and specificity over the MEK 18 and RAS signature in NSCLC. It was also demonstrated that the MEK 6 signature is predictive of *EGFR* oncogenic variants. Finally, this signature was used to demonstrate that the heterozygote loss of function *NF1* variants in NSCLC do not directly impact MAPK regulation.

Whilst FFPET is a challenge to work with for genetic analysis, methods of detection and measurement are evolving to take this into consideration. One such platform is the NanoString nCounter, which provides a non-enzymatic method of counting the mRNA transcripts in the sample using the hybridisation of fluorescent barcodes. Here it is shown that many of the issues, such as degradation of nucleic acids, which inhibit other methods such as microarray analysis and RT-qPCR are not an issue with this platform.

The platforms reproducibility using commercially sourced RNA human reference control has been demonstrated. Analysing the RAW mRNA Log_2 counts and comparing across 3 independent batches demonstrated a high level of reproducibility with a R^2 coefficient of >0.994 measuring over 200 genes. Utilising five highly degraded controls with RIN values ≤1.4 this level of reproducibility was transferable to highly degraded FFPET derived mRNA. Analysis of the Log_2 gene count across technical duplicates run across 3 independent batches resulted in R^2 coefficients which ranged from 0.977 to 0.996. Use of this validation

s*et al*so enabled the analysis of the suitability of housekeeping genes utilising AstraZeneca's in house normalisation tool. Only one housekeeping gene (*G6PD*) out of 21 genes proved unsuitable due to deviations in the overall trend in relation to the other housekeeping genes. *G6DP* was flagged as unsuitable in both analyses of the validation samples and the actual 84 patient samples.

Whilst this gave confidence in the CodeSet overall, it did not provide any information on the reproducibility of the gene expression signatures. Using the Log_2 normalised read count, the MEK 18 and MEK6 gene expression signatures were calculated for the *EGFR*, *KRAS*, *BRAF*, *NF1*, and *MAPK* negative controls as described in section 2.1.6. The replicate gene expression signatures across the three batches demonstrated a high level of reproducibility with all 5 biological controls demonstrating ≤0.05 standard deviation (fold change ≤0.13) across the 6 technical replicates. This validation also demonstrated that *MAPK* oncogenic positive samples have an increased signature score in relation to *NF1* positive and *MAPK* negative controls. Calculating the RAS signature using the Log_{10} normalised read count across all replicates produced similar results with the standard deviation of the five controls being 0.01. Whilst the RAS signature appeared to be selective for *KRAS* and *BRAF*, this was not observed with *EGFR* which was almost indistinguishable from the *MAPK* negative score.

During analysis of the 84 patient samples three had below the recommended 100 ng of starting material as suggested by Brant and colleagues (2016). These samples ranged from 41 ng to 86 ng. The internal QC using nSolver 4.0 highlighted two of the 84 samples to have low mRNA content; one of these was from NF1-223 which had 41 ng of starting material.

NF1 expression was investigated using three probes which either target *NF1* mRNA transcript 1 and 2 or just transcript 2. Initial investigations demonstrated significant differences between the probes NF1-A (exon 7-8) and NF1-B (exon 22-23) which target both *NF1* transcripts. NF1-B demonstrated a mean increase across all the samples of 0.6 fold in relation to NF1-A. This could potentially be the variance in the probes affinity to the target as it was observed consistently in all samples, shown in Figure 5.7. Whilst differences in probe batches have been investigated in the literature, the variability of different probes on the same target gene has not (Brant et al., 2017). A further possible explanation is

partially transcribed *NF1* pseudogenes. Whilst there are no reports in the literature of transcription of *NF1* pseudogenes, the genome browser and database Ensembl listed both *NF1P9* and *NF1P10* as having the potential to be transcribed. Our earlier work mapping the pseudogenes back to *NF1* demonstrated that both of these pseudogenes contain homologs of exon 22-23 but not exon 7-8. One way to resolve this would be to compare multiple probes across *NF1* to determine where any discrepancy lies.

We determined *NF1* transcript 2 to be predominantly expressed in all samples, including the four samples with no visible tumour content and positive controls. *NF1* transcript 2 is most commonly expressed in neurons and has a higher GTPase activity than transcript 1 (Hinman et al., 2014, Brant et al., 2017). This is due to the 21 amino acid exon 31 being excluded from the GAP related domain in transcript 2. The predominance of transcript 1 has been linked to breast cancer and epithelial ovarian tumours, however this is not observed in out NSCLC cohort (Iyengar et al., 1999, Marrero et al., 2012). The fact this was observed in the four samples with no tumour contents suggest this ratio of transcript 1 and 2 is normal for lung tissue. As transcript 2 has 10 X increased GTPase activity over transcript 1 it could be considered advantageous for down regulation of the MAPK pathway in NSCLC. Expression of these two transcripts was detected in the positive controls at the same ratio observed in the NSCLC samples. However, overall expression of *NF1* was increased by 1-fold in the NSCLC samples compared to the positive controls. As the RNA controls originate from mixed cell lines a difference was to be expected as *NF1* expression is tissue specific.

When investigating *NF1* copy number analysis against *NF1* mRNA transcription a Spearman correlation demonstrated a slight positive relationship ($r_s = 0.210$), however this was not significant ($p = 0.101$). Studies investigating the relationship between copy number and mRNA transcription in lung cancer only report significant relationships with a correlation coefficient of $R^2 > 0.7$ in 1.6% of genes (Jabs et al., 2017). Jabs and colleagues work confirmed our reports that *NF1* copy number does not correlate with mRNA expression.

In the analysis of the 84 patient samples using the three gene expression signatures the cases were grouped into *KRAS*, *EGFR*, *BRAF* and *MAPK* negative groups. This analysis was performed in parallel, one set just with the oncogenic variants which were ubiquitously

expressed and the other set removing the low AF samples. For the *KRAS* group with the low AF cases excluded an increase in mean over the *MAPK* negative groups across all the signatures was observed. This was not observed with the *EGFR* group, as two of the sample with low AF variants had the greatest signature scores. The two *EGFR* samples which were excluded based on low AF (NF1-119 and NF1-205) were only just below the minus 30% of expected frequency cut off limit. Based on the positive response for *KRAS* group we continued with analysis using only the oncogenic variants which were ubiquitously expressed.

The three signatures all demonstrated significant differences between the *KRAS* groups and the *MAPK* negative groups ($p > 0.003$). This is reflective of previous work using these signatures (Brant et al., 2017, Dry et al., 2010a, Loboda et al., 2010a). When we compared the three signatures against each other in their abilities to predict *KRAS* status the MEK 6 signature displayed an increased AUC (0.855) with an increased specificity and sensitivity over the other two signatures. The RAS signature demonstrated the lowest with an AUC of 0.799. This was expected as Brant and colleagues optimised this signature for NSCLC, whereas the MEK 18 and RAS were developed using multiple cancer cell lines (Brant et al., 2017). Whilst the exact sensitivity and specificity was unreported in the literature for the MEK signatures, the RAS signature correlated exactly with Loboda and colleagues previous reports (Loboda et al., 2010a).

To confirm the MEK 6 signatures ability to predict *KRAS* status further, we analysed 864 cases from the Pan-Lung TGCA dataset (Campbell et al. 2016). The sensitivity and specificity closely correlated to the result seen in this study's data, with the Pan-Lung dataset demonstrating an AUC of 0.841, compared to our 0.855 shown in Table 5.5. The AUC for both our study and analysis using the Pan-Lungs TGCA data demonstrated the MEK 6 signature to provide an excellent level of discrimination at predicting *KRAS* oncogenic variants in relation to the *MAPK PI3K* negative group ($AUC > 0.8$) (Hosmer et al., 2013). This also validated the MEK 6 signatures predictive power using an orthogonal analytical platform in NSCLC, as the Pan-Lung RNA transcription analysis was generated via RNA seq in relation to our Nanostring analysis.

None of the signatures demonstrated any significant differences between the *MAPK* negative and the *BRAF* groups in data generated in this study. However, as previously mentioned the *BRAF* group only consisted of three cases. Analysis of the Pan-Lung TGCA data which had 21 *BRAF* positive cases did demonstrate a significant difference of the *BRAF* group in relation to the *MAPK PI3K* negative group (p = 0.011).The *EGFR* group demonstrated a significant difference against the *MAPK* negative group using the MEK 6 signature and with the MEK 18 signature before removing the two potentially low AF samples. As the *EGFR* samples were just below the low AF cut off limit, which was calculated based on the pathologist subjective assessment, we decided to investigate the MEK 18 and MEK 6 signatures abilities to predict *KRAS* and *EGFR* driver variants. The MEK 18 signature did see a reduction of the AUC including the *EGFR* variants compared to just *KRAS* of AUC = 0.809 to 0.758. Whilst this is still described as an acceptable level of discrimination it did see the sensitivity drop by 10%. The MEK 6 signature also displayed a reduction from AUC = 0.855 to 0.820 when detecting *KRAS* and *EGFR* drivers. Even with the *EGFR* variants included this provides an excellent level of discrimination (Hosmer et al., 2013). This was also confirmed using data from the Pan-Lung trial, in which the MEK 6 signature demonstrated a AUC of 0.797 which is an acceptable level of discrimination at predicting *KRAS* and *EGFR* oncogenic variants in relation to the *MAPK PI3K* negative group (AUC > 0.8) (Hosmer et al., 2013).

The analysis of the *NF1* loss group which included all samples with *NF1* variants demonstrated no significant difference from the *MAPK* negative group, but did display a significant difference from the *MAPK* positive *KRAS* group. This was confirmed using both the MEK signatures shown in Figure 5.11 and 5.13. This was expected, as many of the samples had low AF and likely benign variants as discussed in the previous chapter. This left two high probability *NF1* loss of function samples (NF1-180 and NF1-233) and the unknown significance sample (NF1-225) which were ubiquitously expressed. Investigating the *NF1* cases individually using the MEK 18 and MEK 6 signatures, NF1-180, NF1-233, and NF1-225 were all below the cut off limit for both signatures, defined in table 3 suggesting low MAPK activity. Furthermore, the 2 potential pseudogene variants, reported as pathogenic in neurofibromatosis patients, which were just below the -30% expected AF were below the cut off for both signatures.

Of the two null variants (NF1-233 and NF1-180) which were expressed throughout the region of the sample analysed, only NF1-233 was downstream of NF1s GAP domain and would result in heterozygote loss of function of *NF1* GAP domain. The cut-off limit based on the MEK 6 signatures ability to predict *KRAS* variants was 8.82 which is equivalent to 452 mean counts of the genes in the MEK 6 signature which demonstrated 100% sensitivity and 68% specificity at detecting *KRAS* variants. The signature score for NF1-223 was 8.18 (290 mean counts), 162 counts below the cut-off limit. However, NF1-180 which had a null variant upstream of the NF1s GAP domain had a signature score of 8.7 (416 mean counts), whilst still below the cut-off limit the difference was only 36 counts.

As this study only had one NF1 null variant downstream of NF1s GAP domain we repeated the analysis using the dataset from the Pan-Lung TGCA trial. Of the 19 NF1 cases which were classed as high probability of loss of NF1 function, 16 cases had null variants downstream of NF1s GAP domain and 3 were classed as NF1 homozygote deletions. The 19 cases in the NF1 high probability group did not demonstrate any significate difference considering them as a group in relation to the *MAPK PI3K* negative group ($p = 0.999$) shown in Figure 5.17. However, using the cut-off limit of 9.8 (891 mean RNA reads) based on the MEK 6 signatures ability to predict *KRAS* variants in the Pan-Lung dataset, 31% (5/16) of the *NF1* null variants cases downstream of the GAP domain were above this threshold. Interestingly, of the 3 homozygote *NF1* deletions only 33% (1/3) was above the threshold. The sensitivity of the MEK 6 signature was calculated at 90%, therefore it was expected that 10% of cases above this threshold would be false positives. As greater than 10% of the NF1 cases were above this threshold, it does suggest based on the MEK 6 signature, that *NF1* heterozygote or *NF1* homozygote deletions alone is not sufficient to drive the MAPK pathway, and further unknown factors could be involved in the 31% of cases which were above the threshold. Similar findings have been observed in neurofibromatosis, where further mutations in addition to *NF1* are involved in the transition of tumours from benign to malignant (Sohier et al., 2017).

The observation based on the MEK 6 signature that NF1 loss alone is not sufficient to drive the MAPK pathway is also supported by the cases with co-occurring *RASA1* and *NF1* variants. Of the grouped analysis, shown in Figure 5.17, the *NF1*+*RASA1* cases are the only group to demonstrate a significant increase in up regulation of the MEK 6 signature in

relation to the *MAPK PI3K* negative group (p = 0.002). The analysis also highlighted the *KRAS* and *NF1*+*RASA1* groups to form a homogenous a subset (p = 0.936). This combination of co-occurring variants in NSCLC has also been described in cell based models (Hayashi et al., 2018). Hasyashi and colleagues demonstrated the EPLC272H cell line with co mutated *NF1* and *RASA1* genes was more sensitive to MEK and PI3K inhibitors, as single agents or in combination, compared to cell lines only harbouring *NF1* or *RASA1*.

In data generated in this study a number of unexpected *NF1* positive samples were observed with signature scores which indicate MAPK activation. Both MEK signatures demonstrated NF1-178 (p.Ala2532Val) to be above the cut off limits, suggesting increased MAPK activation. NF1-178 had a germline *NF1* variant which was predicted as likely benign by in-silico analysis. The MEK 6 signature identified another 3 *NF1* to be above the 8.82 threshold. NF1-201 was in the likely benign group, NF1-115 and NF1-217 were from the unknown significance group. Of these, NF1-217 was of interest, as this was the only sample to display 2 *NF1* null variants and therefore has the potential to cause homozygote loss of *NF1*. Unfortunately the calculated AF of these NF1-217 variants was calculated to be seen in only 30% of the tumour. When considering the 68% specificity the MEK 6 signature has demonstrated, 5 potential false positives would be predicted, which is exactly what was observed in the *NF1* cases.

Using the Pan-Lung TGCA dataset we investigated the effect *NF1* variants have on regulation of the PI3K/AKT/mTOR pathway. Analysis using the 136 gene up regulated arm of the CMAP signature demonstrated the *MAPK PI3K* negative group could contain unknown drivers of both the MAPK and the PI3K/AKT/mTOR pathways, as shown in Figure 5.18. Using the MAPK positive group as a PI3K/AKT/mTOR negative control we observed a significant increase in the CMAP mean signature score for the PI3K/AKT/mTOR positive group shown in Figure 5.20 ($p < 0.001$). The *NF1* and *RASA1* independent groups also demonstrated a significant increase in the CMAP signature score in relation to the MAPK positive group ($p < 0.001$). This does suggest cases with either a *NF1* or *RASA1* loss of function variant would function in up-regulation of the PI3K/AKT/mTOR pathway. Interestingly, the group with co-occurring *NF1* and *RASA1* variants did not show a significant difference in relation to the *MAPK* positive group, using the CMAP signature. However, based on the median and the quartile distribution of the *NF1*+*RASA1* group, the

large range of the CMAP scores, and the small number of cases (n = 10) shown in Figure 5.19, we strongly believe increasing the population of this group would result in a significant difference being observed. *NF1* variants have been described to drive the PI3K/AKT/mTOR pathway in malignant peripheral nerve sheath tumours (MPNSTs) in neurofibromatosis models (Malone et al., 2014, Johannessen et al., 2005). Malone and colleagues demonstrated that NF1 positive MPNSTs were sensitive to mTOR1 inhibition in mouse models, but this only induced a cytostatic effect. However, combining an mTOR1 inhibitor with a MEK inhibitor resulted in tumour regression.

Our data does suggest, based on the MEK 6 and CMAP signatures that NF1 or RASA1 loss alone is not enough to drive the MAPK pathway, but these variants are sufficient to upregulate the PI3K/AKT/mTOR pathway. We postulate further genetic variations in addition to *NF1* are required for upregulation of the MAPK pathway. This was observed in the Pan-Lung data analysed in this study and was demonstrated by the increase of the MEK 6 signature score in cases with co-occurring *NF1* and *RASA1* variants. This combination of co-occurring *NF1* and *RASA1* variants has also been described by Hayashi et al., (2018) in lung cancer models and could represent a combination biomarker for PI3K and MEK inhibition. mTOR1 and MEK inhibitors have also been shown to reduce MPNSTs in neurofibromatosis mouse models (Malone *et al.*, 2014). It is important to note 30% of *NF1* cases in the Pan-Lung dataset were above the *KRAS* positive threshold described in Table 5.5, this does indicate further genetic variants could also be functioning alongside *NF1* in upregulation of the MAPK pathway.

This analysis investigating the different clinical and demographic groups including; subtype, gender, and disease stage, demonstrated a significant difference between the ADC and SQCC subtypes using the MEK 18 signature, and between the ADC and NOS groups using the MEK 6 signature. Whilst the significance was not enough to be predictive of the various subtypes, it does suggest the MAPK pathway is active in many SQCC cases, which are thought to be largely driven through MAPK independent mechanisms (Hammerman et al., 2012)

Chapter 6
Final Discussion

Background

Previous investigations into the genetic landscape of NSCLC identified a subpopulation of patients that harbour somatic variants in *NF1* (Ding et al., 2008, Collisson et al., 2014, Hammerman et al., 2012). The majority of these patients only demonstrate aberrations in one allele. It is known from Neurofibromatosis pathogenesis that somatic variants in *NF1* leave the individual more susceptible to the development of benign and malignant neurofibromas (Riccardi and Lewis, 1988). In Neurofibromatosis disease development is subject to Knudson's two-hit hypothesis, where bi-allelic inactivation of *NF1* is required (Yang et al., 2008). It has been shown in lung cancer cell models that knock down of *NF1* drives cellular proliferation through activation of the MAPK pathway and causes resistance to TKI (de Bruin et al., 2014a). This study aimed to address whether *NF1* variants in NSCLC are sufficient to upregulate MAPK activation in the absence of any other MAPK drivers.

At the early stages of this study it was apparent that the variability observed in the quantity and quality of patients archived FFPET was going to be a limiting factor. To minimise the impact of this, an attempt was made to optimise and validate DNA extraction, quantification, and quality assessment, as detailed in Chapter 3. This included optimising the NGS protocol, validating the lower limits of detection, and setting up a customised bioinformatics pipeline. The use of ddPCR was explored in order to determine its potential for measuring *NF1* copy number. This provided a solid foundation for the work described in Chapter 4. In this chapter the patients' clinical FFPET samples were screened for somatic *NF1* variants and known MAPK oncogenic drivers. The results of this directed the division of the patients based on genotype and moved forward the investigation of the gene expression signature as detailed in Chapter 5. The signatures were analysed to determine their ability to predict MAPK activation based on *KRAS* status. Once established, the scores from these signatures were then utilised to determine the effects of *NF1* variants on MAPK activation.

Study Findings

During patient screening a great degree of variability in both the quality and quantity of the DNA recovered from the clinical samples was observed. Use of conventional methods of quantification such as the Nanodrop or Qubit would have resulted in a greater number of samples failing at the library preparation stage of sequencing. In line with previous reports the qPCR and QFI score provided a reproducible pre-sequencing QC check to determine the quantity and quality of starting material for NGS (Sah et al., 2013). The *NF1* pseudogene mapping, explored in Chapter 3, provided a method of flagging potential pseudogene variants, which should be treated with caution before reporting. This adds a further level of security rather than relying on the sequencing metrics alone, as previously described in section 3.1 (Cunha et al., 2016).

The genomic profiling studies, which originally identified the prevalence of *NF1* variants in NSCLC, were restricted by sequencing depth, which was dictated by the size of the whole genome or exome being sequenced (Ding et al., 2008, Collisson et al., 2014, Hammerman et al., 2012). This limited the ability to assess low level tumour content samples with sample with < 60% tumour content been excluded from the studies. Use of a relatively small targeted NGS panel enabled the sequencing of *NF1* and other key MAPK genes at >10-fold the depth of the NSCLC Pan-Lung (Campbell et al., 2016) as described in Chapter 3. The advantage of the current study was the ability to confidently identify variants with a high sensitivity at low allelic fractions, which enabled the identification of ubiquitously expressed variants in samples with only 6% tumour content.

Of the 86 samples that were successfully sequenced, 25 previously reported MAPK oncogenic variants which are known to activate the MAPK pathway were identified. Of these, the 7 *EGFR* variants correlated with demographics and prevalence observed for ADC in the wider literature (Zhang et al., 2016b). In this study the prevalence of *KRAS* patients, was slightly below that seen in ADC and above that for SQCC patients (Campbell et al., 2016, Boch et al., 2013). The two *BRAF* variants in ADC cases correlated with previous reports, but the one SQCC case was above reported prevalence (Campbell et al., 2016). However, based on the size of the cohort of patients this level of fluctuation in prevalence was to be expected.

Use of deep sequencing made it possible to determine whether these oncogenic variants were observed ubiquitously in the region of the tumour sampled, or in limited sub-populations only. This builds on the information reported in large genomic profiling studies (Ding et al., 2008, Collisson et al., 2014, Hammerman et al., 2012). 72% of the 25 oncogenic variants were found ubiquitously throughout the region of tumour sampled, in data generated in this study, shown in Table 4.15. This was expected, with known oncogenic drivers out-competing less dominant sub-clones, these would therefore be expected to be found in most regions of the tumour sampled, using either multi region or single region analysis. This was observed in our study's data when compared to TRACERx's whose data reported 33% of *KRAS* cases to be sub-clonal, compared to our 21%.

25 *NF1* variants were identified in 20 patients, including the three flagged as potential pseudogene in origin, two of which have been reported as pathogenic (Upadhyaya et al., 2003, Laycock-van Spyk et al., 2011), shown in Table 4.17. Of these 25 variants 15 were novel with no evidence of been previously reported in the literature or in databases. *In-silico* analysis predicted the effect these variants could have on *NF1* loss of function; 5 were classed as null variants, 10 defined as of unknown significance, and 10 likely benign.

22% of patients displayed *NF1* variants, close to double the 12% previously reported (Campbell et al., 2016). However, using the same criteria as the Pan-Lung study, which excluded intronic variants and those not identified using 100 X read depth, 13% of patients in this study displayed *NF1* variants.

Unlike the known oncogenic drivers which demonstrated similar frequencies to what was observed in TRACERx, after excluding germline *NF1* variants, only 20% (4/20) of our cases expressed *NF1* variants ubiquitously throughout the region of tumour analysed. It was not possible to correlate this with the Pan-Lung TGCA data as exact tumour content of individual samples was unknown. However, mining the Pan-Lung allelic frequency sequencing data for *NF1* showed that 76% of their samples had <40% variant reads in relation to reference reads. Of the four patients that displayed co-expression of *MAKP* oncogenic and *NF1* variants in this study, three had *NF1* AF at half the level of the AF of the known oncogenic driver. This does demonstrate *NF1* variants commonly found in NSCLC patients are largely sub-clonal using single region sampling. TRACERx reported *NF1*

variants to be sub-clonal in 50% of patients based on multi-region sampling; whereas our analysis showed only 20% of *NF1* variants to be expressed throughout the region of tumour analysed. If a *NF1* variants are sub-clonal in 50% of cases, it would be expected to see a greater difference between multi-region and single region analysis.

In Chapter 5 the NanoString nCounter was shown to be tolerant to highly degraded RNA when quantifying gene expression. Use of this platform enabled not only the exploration of the gene expression signatures, but also investigation of *NF1* mRNA expression.

Analysis of *NF1* transcription demonstrated that it had no correlation to *NF1* copy number, in agreement with previously reported results (Jabs et al., 2017). The level of expression of the different transcripts in lung cancer has not previously been reported but an increase in expression of transcript 1 has been reported in several other cancers including, breast cancer and epithelial ovarian tumours (Iyengar et al., 1999, Marrero et al., 2012). Transcript 1 has been shown to have a 10-fold reduction in its GTPase activity over transcript 2 leading to an increase of MAPK activation through its inhibited ability to deactivate KRAS. The results from this study indicate that both *NF1* mRNA transcripts are expressed in NSCLC, with transcript 2 being predominantly expressed. As this was observed in the four samples with no visible tumour content, it suggests that this is a normal ratio for lung tissue.

Initial work with gene expression signatures determined that these provided a viable means of measuring *KRAS* oncogenic status and therefore MAPK activation. Investigations of the MEK 6 signature using both data generated in this study and the Pan-Lung TGCA dataset confirmed the findings of previous studies with the signature being predictive of *KRAS* status (Dry et al., 2010a, Brant et al., 2017, Loboda et al., 2010a). The RAS signature demonstrated the lowest sensitivity and specificity but still retained an acceptable level of discrimination. MEK 18 and MEK 6 provided an excellent level of discrimination (AUC > 0.8) (Hosmer et al., 2013). In addition to the published data it was also confirmed that both the MEK 6 signature could be used to predict *EGFR and BRAF* in addition to *KRAS* status. This corroborates work in cell based models which suggests *EGFR* predominately drives the MAPK pathway (de Bruin et al., 2014a).

When investigating our cohort of patients only had one case with a null variant downstream of NF1s GAP domain. To expand on this we utilised Pan-Lung TGCA data and analysed 864 NSCLC cases. In addition to analysis with the MEK 6 signature we also investigated the CMAP signature, which is related to the PI3K/AKT/mTOR pathway. We observed an increased MEK 6 signature in the group with co-occurring *NF1* and *RASA1* similar to the *KRAS* group. This could represent a combination biomarker for PI3K and MEK inhibition, which was original proposed by Hayashi *et al*. (2017) using cell based models. Whilst the *NF1* group did not show a significant increase in the signature, as a whole, over 30% of the individual cases were over the *KRAS* threshold shown in Table 5.5. We also demonstrated an increase in the CMAP signature for the *NF1* and *RASA1* groups, which suggests these variants, can cause upregulation of the PI3K/AKT/mTOR pathway. Malone *et al*. (2014) demonstrated *NF1* can cause upregulation of the PI3K/AKT/mTOR pathway in pre-clinical models. Whilst the *NF1* and co-occurring *RASA1* group did not show a significant increase in the CMAP signature, we suspect this was due to the low number of cases in the cohort.

In summary, the prevalence of *NF1* variants in NSCLC has been confirmed. A method of flagging pseudogene reads by mapping back to the functional *NF1* gene during realignment in the bioinformatics pipeline has been identified. It has been shown that *NF1* transcript 2 is predominantly expressed in NSCLC and the potential of MEK signatures for predicting *EGFR* and *BRAF* status has been further demonstrated. We have demonstrated co-occurring *NF1* and *RASA1* variants are enough to drive the MAPK pathway based on the MEK 6 signature, while cases with individual *NF1* or *RASA1* variants can increase PI3K/AKT/mTOR activity. Finally, a third of all *NF1* cases had an increased MEK 6 signature score above the *KRAS* calculated threshold, which warrants further investigations into co-drivers.

Study Limitations

The primary limitation in this study was the variability of patient samples. 23 were rejected based on insufficient tissue prior to any analytical assessment. Based on this, the pre-sample analysis establishment and validation was undertaken as described in Chapter 3,

to maximise the utility of the available tissue and potentially mitigate for the degree of variability observed.

Time was also a constraint. There were significant challenges in in the transfer of archived patient material from Sheffield Teaching Hospitals Trust to Sheffield Children's Hospital Trust. Many cases took in excess of 8 months from being requested to being received. This resulted in the majority of the samples not being accessible until March 2018, making it impossible to repeat any analysis. Hence it was not possible to further validate the lower limits of starting material required after observing the erroneous copy number results, or to macro-dissect samples prior to copy number analysis.

Whilst the initial group of patients was divided adequately based on genotype, with 25 *MAPK* oncogenic variants, 44 *MAPK* negative and 15 *NF1* patients, the finding of such a small cohort of patients with ubiquitous *NF1* variants was not expected. For further statistical power it would have been advantageous to have more ubiquitous *NF1* patients. This was considered early on in the study. Using the knowledge that *KRAS* and *NF1* are largely mutually exclusive, by excluding *KRAS* patients it would be possible to enrich the *NF1* patients. Unfortunately, unlike *EGFR*, *KRAS* is not routinely tested therefore it could not be utilised for patient screening and selection.

Future Work

The MEK 6 and the CMAP signatures have been shown to be predictive of key drivers of each both the MAPK and PI3K/AKT/mTOR pathway. These signatures can be used as functional tools to investigate pathway activity and help provide evidence establishing the functional relevance of potential new genetic biomarkers. Whilst this study has focused on NF1 in NSCLC, these signatures should be assessed further in different cancers to uncover potential unknown drivers of the MAPK and PI3K/AKT/mTOR pathways. New potential biomarkers should then be transferred back to preclinical models to investigate there therapeutic potential. Finally, the gene expression signatures need to be further investigated and refined in future clinical trials to develop their clinical potential. The signatures could add an extra dimension to current genetic biomarkers and could identify alternate patient groups who could respond to targeted therapeutics.

There is also the potential to follow up patients using their clinical data, to look at overall survival rates and response to treatment. This study only considered *NF1's* ability as a GTPase activating protein and its effect on the negative regulation of *KRAS* and the MAPK pathway. However, the GAP related domain makes up less than 10% of *NF1*. It was briefly mentioned in Chapter 1 that *NF1* can function as a tumour suppresser gene using MAPK independent mechanisms. This is an alternate route which is open to investigation and could determine further functional implications for the variants observed in our study.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. and Sunyaev, S. R. (2010) 'A method and server for predicting damaging missense mutations', *Nat Methods: Vol. 4*. United States, pp. 248-9.

Ahmadian, M., Zor, T., Vogt, D., Kabsch, W., Selinger, Z., Wittinghofer, A. and Scheffzek, K. (1999) 'Guanosine triphosphatase stimulation of oncogenic Ras mutants', *Proceedings Of The National Academy Of Sciences Of The United States Of Ame,* 96(12), pp. 7065-7070.

Ahmadian, M. R., Kiel, C., Stege, P. and Scheffzek, K. (2003) 'Structural fingerprints of the Ras-GTPase activating proteins neurofibromin and p120GAP', *J Mol Biol: Vol. 4*. England, pp. 699-710.

Ahn, S., Brant, R., Sharpe, A., Dry, J. R., Hodgson, D. R., Kilgour, E., Kim, K., Kim, S. T., Park, S. H., Kang, W. K., Kim, K. M. and Lee, J. (2017) 'Correlation between MEK signature and Ras gene alteration in advanced gastric cancer', *Oncotarget,* 8(64), pp. 107492-107499.

Anastasaki, C. and Gutmann, D. H. (2014) 'Neuronal NF1/ RAS regulation of cyclic AMP requires atypical PKC activation', *Human molecular genetics,* 23(25), pp. 6712.

Andersen, L. B., Ballester, R., Marchuk, D. A., Chang, E., Gutmann, D. H., Saulino, A. M., Camonis, J., Wigler, M. and Collins, F. S. (1993a) 'A conserved alternative splice in the von Recklinghausen neurofibromatosis (NF1) gene produces two neurofibromin isoforms, both of which have GTPase-activating protein activity', *Mol Cell Biol,* 13(1), pp. 487-95.

Andersen, L. B., Fountain, J. W., Gutmann, D. H., Tarle, S. A., Glover, T. W., Dracopoli, N. C., Housman, D. E. and Collins, F. S. (1993b) 'Mutations in the neurofibromatosis 1 gene in sporadic malignant melanoma cell lines', *Nat Genet,* 3(2), pp. 118-21.

Apolline, I., Eric, P., Audrey, S., Armelle, L., Magali, S., Philippe, G., Hélène, B., Ingrid, L., Salah, F., Michel, V., Stéphane, P., Christine, B.-C., Dominique, V., Pierre, W. and Béatrice, P. (2015) 'NF1 single and multi- exons copy number variations in neurofibromatosis type 1', *Journal of Human Genetics*.

Arreaza, G., Qiu, P., Pang, L., Albright, A., Hong, L. Z., Marton, M. J. and Levitan, D. (2016) 'Pre-Analytical Considerations for Successful Next-Generation Sequencing (NGS): Challenges and Opportunities for Formalin-Fixed and Paraffin-Embedded Tumor Tissue (FFPE) Samples', *Int J Mol Sci,* 17(9).

Barron, V. A. and Lou, H. (2012) 'Alternative splicing the Neurofibromatosis type I pre-mRNA', *Biosci Rep,* 32(2), pp. 131-8.

Barron, V. A., Zhu, H., Hinman, M. N., Ladd, A. N. and Lou, H. (2010) 'The neurofibromatosis type I pre-mRNA is a novel target of CELF protein-mediated splicing regulation', *Nucleic Acids Res,* 38(1), pp. 253-64.

Beauchamp, E., Woods, B. A., Dulak, A., Tan, L., Xu, C., Gray, N., Bass, A., Wong, K., Meyerson, M. and Hammerman, P. S. (2014) 'Acquired Resistance to Dasatinib in Lung Cancer Cell Lines Conferred by DDR2 Gatekeeper Mutation and NF1 Loss', *Mol. Cancer Ther.,* 13(2), pp. 475-482.

Boch, C., Kollmeier, J., Roth, A., Stephan-Falkenau, S., Misch, D., Gruning, W., Bauer, T. T. and Mairinger, T. (2013) 'The frequency of EGFR and KRAS mutations in non-small cell lung cancer (NSCLC): routine screening data for central Europe from a cohort study', *BMJ Open,* 3(4).

Brant, R., Sharpe, A., Liptrot, T., Dry, J. R., Harrington, E. A., Barrett, J. C., Whalley, N., Womack, C., Smith, P. and Hodgson, D. R. (2017) 'Clinically Viable Gene Expression Assays with Potential for Predicting Benefit from MEK Inhibitors', *Clin Cancer Res,* 23(6), pp. 1471-1480.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A. and Jemal, A. (2018) 'Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA Cancer J Clin*.

Brinckmann, A., Mischung, C., Bassmann, I., Kuhnisch, J., Schuelke, M., Tinschert, S. and Nurnberg, P. (2007) 'Detection of novel NF1 mutations and rapid mutation prescreening with Pyrosequencing', *Electrophoresis,* 28(23), pp. 4295-301.

Brown, J., Gianino, S. M. and Gutmann, D. (2010) 'Defective cAMP Generation Underlies the Sensitivity of CNS Neurons to Neurofibromatosis-1 Heterozygosity', *J. Neurosci.,* 30(16), pp. 5579-5589.

Brown, T., Darnton, A., Fortunato, L. and Rushton, L. (2012) 'Occupational cancer in Britain - Respiratory cancer sites: Larynx, lung and mesothelioma', *British Journal of Cancer,* 107(1), pp. S56-S70.

Cadranel, J., Ruppert, A. M., Beau-Faller, M. and Wislez, M. 2013. Therapeutic strategy for advanced EGFR mutant non- small- cell lung carcinoma.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. and Casadio, R. (2009) 'Functional annotations improve the predictive score of human disease-related mutations in proteins', *Hum Mutat,* 30(8), pp. 1237-44.

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., Shukla, S. A., Guo, G., Brooks, A. N., Murray, B. A., Imielinski, M., Hu, X., Ling, S., Akbani, R., Rosenberg, M., Cibulskis, C., Ramachandran, A., Collisson, E. A., Kwiatkowski, D. J., Lawrence, M. S., Weinstein, J. N., Verhaak, R. G., Wu, C. J., Hammerman, P. S., Cherniack, A. D., Getz, G., Artyomov, M. N., Schreiber, R., Govindan, R. and Meyerson, M. (2016) 'Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas', *Nat Genet,* 48(6), pp. 607-16.

Capriotti, E., Calabrese, R. and Casadio, R. (2006) 'Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information', *Bioinformatics,* 22(22), pp. 2729-34.

Castellano, E. and Downward, J. (2011) 'RAS Interaction with PI3K: More Than Just Another Effector Pathway', *Genes & cancer,* 2(3), pp. 261.

Cawthon, R. M., Andersen, L., Buchberg, A., Xu, G., Oconnell, P., Viskochil, D., Weiss, R., Wallace, M., Marchuk, D., Culver, M., Stevens, J., Jenkins, N., Copeland, N., Collins, F. and White, R. (1991) 'CDNA SEQUENCE AND GENOMIC STRUCTURE OF EV12B, A GENE LYING WITHIN AN INTRON OF THE NEUROFIBROMATOSIS TYPE- 1 GENE', *Genomics,* 9(3), pp. 446-460.

Cawthon, R. M., Weiss, R., Xu, G. F., Viskochil, D., Culver, M., Stevens, J., Robertson, M., Dunn, D., Gesteland, R., O'Connell, P. and et al. (1990) 'A major segment of the neurofibromatosis type 1 gene: cDNA sequence, genomic structure, and point mutations', *Cell: Vol. 1*. United States, pp. 193-201.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C. and Schultz, N. (2012) 'The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data', *Cancer Discov: Vol. 5*. United States: 2012 Aacr., pp. 401-4.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. and Chan, A. P. (2012) 'Predicting the functional effect of amino acid substitutions and indels', *PLoS One,* 7(10), pp. e46688.

Choudhary, A., Mambo, E., Sanford, T., Boedigheimer, M., Twomey, B., Califano, J., Hadd, A., Oliner, K. S., Beaudenon, S., Latham, G. J. and Adai, A. T. (2014) 'Evaluation of an integrated clinical workflow for targeted next-generation sequencing of low-quality tumor DNA using a 51-gene enrichment panel', *BMC Med Genomics,* 7, pp. 62.

Collisson, E. A. and Campbell, J. D. and Brooks, A. N. and Berger, A. H. and Lee, W. and Chmielecki, J. and Beer, D. G. and Cope, L. and Creighton, C. J. and Danilova, L. and Ding, L. and Getz, G. and Hammerman, P. S. and Hayes, D. N. and Hernandez, B. and Herman, J. G. and Heymach, J. V. *et al*. (2014) 'Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network', *Nature,* 511(7511), pp. 543-550.

Creighton, C. J., Fu, X., Hennessy, B. T., Casa, A. J., Zhang, Y., Gonzalez-Angulo, A. M., Lluch, A., Gray, J. W., Brown, P. H., Hilsenbeck, S. G., Osborne, C. K., Mills, G. B., Lee, A. V. and Schiff, R. (2010) 'Proteomic and transcriptomic profiling reveals a link between the PI3K pathway and lower estrogen-receptor (ER) levels and activity in ER+ breast cancer', *Breast Cancer Res,* 12(3), pp. R40.

CRUK (2018) *Lung Cancer Key Stats*. Cancer Research UK: Cancer Research UK. Available at: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer (Accessed: 05/02/2018 2018).

Cunha, K. S., Oliveira, N. S., Fausto, A. K., de Souza, C. C., Gros, A., Bandres, T., Idrissi, Y., Merlio, J. P., de Moura Neto, R. S., Silva, R., Geller, M. and Cappellen, D. (2016) 'Hybridization Capture-Based Next-Generation Sequencing to Evaluate Coding Sequence and Deep Intronic Mutations in the NF1 Gene', *Genes (Basel),* 7(12).

D'Angelo, I., Welti, S., Bonneau, F. and Scheffzek, K. (2006) 'A novel bipartite phospholipid-binding module in the neurofibromatosis type 1 protein', *EMBO Rep,* 7(2), pp. 174-9.

Dawson, J. P., Berger, M. B., Lin, C., Schlessinger, J., Lemmon, M. A. and Ferguson, K. (2005) 'Epidermal growth factor receptor dimerization and activation require ligand-induced conformational changes in the dimer interface', *Mol. Cell. Biol.,* 25(17), pp. 7734-7742.

de Bruin, E., McGranahan, N., Salm, M., Wedge, D., Mitter, R., Yates, L., Matthews, N., Stewart, A., Campbell, P. and Swanton, C. (2014b) 'Intra-tumour heterogeneity in early-

stage lung cancer inferred by multi-region sequencing', *European Journal Of Cancer,* 50, pp. S4-S4.

de Bruin, E. C., Cowell, C., Warne, P. H., Jiang, M., Saunders, R. E., Melnick, M. A., Gettinger, S., Walther, Z., Wurtz, A., Heynen, G. J., Heideman, D. A. M., Gomez-Roman, J., Garcia-Castano, A., Gong, Y., Ladanyi, M., Varmus, H., Bernards, R., Smit, E. F., Politi, K. and Downward, J. (2014a) 'Reduced NF1 Expression Confers Resistance to EGFR Inhibition in Lung Cancer', *Cancer Discovery,* 4(5), pp. 606-619.

Deng, P., Hu, C., Zhou, L., Li, Y. and Huang, L. (2013) 'Clinical characteristics and prognostic significance of 92 cases of patients with primary mixed-histology lung cancer', *Mol Clin Oncol,* 1(5), pp. 863-868.

Desmet, F. O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M. and Beroud, C. (2009) 'Human Splicing Finder: an online bioinformatics tool to predict splicing signals', *Nucleic Acids Res,* 37(9), pp. e67.

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B., Fulton, L., Fulton, R. S., Zhang, Q., Wendl, M. C., Lawrence, M. S., Larson, D. E., Chen, K., Dooling, D. J., Sabo, A., Hawes, A. C., Shen, H., Jhangiani, S. N., Lewis, L. R., Hall, O., Zhu, Y., Mathew, T., Ren, Y., Yao, J., Scherer, S. E., Clerc, K., *et al*. (2008) 'Somatic mutations affect key pathways in lung adenocarcinoma', *Nature,* 455(7216), pp. 1069-1075.

Dry, J. R., Pavey, S., Pratilas, C. A., Harbron, C., Runswick, S., Hodgson, D., Chresta, C., McCormack, R., Byrne, N., Cockerill, M., Graham, A., Beran, G., Cassidy, A., Haggerty, C., Brown, H., Ellison, G., Dering, J., Taylor, B. S., Stark, M., Bonazzi, V., Ravishankar, S., Packer, L., Xing, F., Solit, D. B., Finn, R. S., Rosen, N., Hayward, N. K., French, T. and Smith, P. D. (2010a) 'Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244)', *Cancer Res,* 70(6), pp. 2264-73.

Dry, J. R., Runswick, S., Chresta, C., Byrne, N., Cockerill, M., Graham, A., Beran, G., Cassidy, A., Haggerty, C., Brown, H., French, T., Smith, P. D., Harbron, C., McCormack, R., Ellison, G., Hodgson, D., Pavey, S., Stark, M., Bonazzi, V., Ravishankar, S., Packer, L., Hayward, N. K., Pratilas, C. A., Taylor, B. S., Xing, F., Solit, D. B., Rosen, N., Dering, J. and Finn, R. S. (2010b) 'Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244)', *Cancer Research,* 70(6), pp. 2264-2273.

Easton, D. F., Ponder, M. A., Huson, S. M. and Ponder, B. A. J. (1993) 'An analysis of variation in expression of neurofibromatosis (NF) type 1 (NF1): Evidence for modifying genes', *American Journal of Human Genetics,* 53(2), pp. 305-313.

Edge, S. B. and Compton, C. C. (2010) 'The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM', *Ann Surg Oncol,* 17(6), pp. 1471-4.

Ferrer-Costa, C., Orozco, M. and de la Cruz, X. (2004) 'Sequence-based prediction of pathological mutations', *Proteins,* 57(4), pp. 811-9.

Friedman, J. M. (1999) 'Epidemiology of neurofibromatosis type 1', *Am J Med Genet: Vol. 1*. United States: 1999 Wiley-Liss, Inc., pp. 1-6.

Fukuoka, M., Yano, S., Giaccone, G., Tamura, T., Nakagawa, K., Douillard, J. Y., Nishiwaki, Y., Vansteenkiste, J., Kudoh, S., Rischin, D., Eek, R., Horai, T., Noda, K., Takata, I., Smit, E.,

Averbuch, S., Macleod, A., Feyereislova, A., Dong, R. P. and Baselga, J. (2003) 'Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non- small- cell lung cancer', *Journal of Clinical Oncology,* 21(12), pp. 2237-2246.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C. and Schultz, N. (2013) 'Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal', *Sci Signal: Vol. 269*. United States, pp. pl1.

Garrido, P., Olmedo, M. E., Gomez, A., Paz Ares, L., Lopez-Rios, F., Rosa-Rosa, J. M. and Palacios, J. (2017) 'Treating KRAS-mutant NSCLC: latest evidence and clinical consequences', *Ther Adv Med Oncol,* 9(9), pp. 589-597.

Gautschi, O., Peters, S., Zoete, V., Aebersold-Keller, F., Strobel, K., Schwizer, B., Hirschmann, A., Michielin, O. and Diebold, J. (2013) 'Lung adenocarcinoma with BRAF G469L mutation refractory to vemurafenib', *Lung Cancer,* 82(2), pp. 365-7.

Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., Fell, H. P., Ferree, S., George, R. D., Grogan, T., James, J. J., Maysuria, M., Mitton, J. D., Oliveri, P., Osborn, J. L., Peng, T., Ratcliffe, A. L., Webster, P. J., Davidson, E. H., Hood, L. and Dimitrov, K. (2008) 'Direct multiplexed measurement of gene expression with color-coded probe pairs', *Nat Biotechnol,* 26(3), pp. 317-25.

Gupta, S., Ramjaun, A. R., Haiko, P., Wang, Y., Warne, P. H., Nicke, B., Nye, E., Stamp, G., Alitalo, K. and Downward, J. (2007) 'Binding of Ras to Phosphoinositide 3-Kinase p110α Is Required for Ras- Driven Tumorigenesis in Mice', *Cell,* 129(5), pp. 957-968.

Haaland, B., Tan, P. S., de Castro, G., Jr. and Lopes, G. (2014) 'Meta-analysis of first-line therapies in advanced non-small-cell lung cancer harboring EGFR-activating mutations', *J Thorac Oncol,* 9(6), pp. 805-11.

Hammerman, P. S., Sos, M., Ramos, A., Xu, C., Dutt, A., Zhou, W., Brace, L., Woods, B. A., Lin, W., Zhang, J., Deng, X., Lim, S. M., Heynck, S., Peifer, M., Simard, Jr., Lawrence, M. S., Onofrio, R. C., Salvesen, H., Seidel, D., Zander, T., Heuckmann, J., Soltermann, A., Moch, H., Koker, M.,*et al.* (2011) 'Mutations in the DDR2 Kinase Gene identify a Novel therapeutic target in squamous cell lung cancer', *Cancer Discov.,* 1(1), pp. 78-89.

Hammerman, P. S. and Voet, D. and Lawrence, M. S. and Voet, D. and Jing, R. and Cibulskis, K. and Sivachenko, A. and Stojanov, P. and McKenna, A. and Lander, E. S. and Getz, G. and Imielinski, M. and Helman, E. and Hernandez, B. and Pho, N. H. and Meyerson, M., *et al*. (2012) 'Comprehensive genomic characterization of squamous cell lung cancers', *Nature,* 489(7417), pp. 519-525.

Hanahan, D. and Weinberg, Robert A. (2011) 'Hallmarks of Cancer: The Next Generation', *Cell,* 144(5), pp. 646-674.

Hartman, Z., Zhao, H. and Agazie, Y. (2013) 'HER2 stabilizes EGFR and itself by altering autophosphorylation patterns in a manner that overcomes regulatory mechanisms and promotes proliferative and transformation signaling', *Oncogene,* 32(35), pp. 4169-4180.

Hayashi, T., Desmeules, P., Smith, R. S., Drilon, A., Somwar, R. and Ladanyi, M. (2018) 'RASA1 and NF1 are Preferentially Co-Mutated and Define A Distinct Genetic Subset of

Smoking-Associated Non-Small Cell Lung Carcinomas Sensitive to MEK Inhibition', *Clin Cancer Res,* 24(6), pp. 1436-1447.

Heredia, N. J., Belgrader, P., Wang, S., Koehler, R., Regan, J., Cosman, A. M., Saxonov, S., Hindson, B., Tanner, S. C., Brown, A. S. and Karlin-Neumann, G. (2013) 'Droplet Digital PCR quantitation of HER2 expression in FFPE breast cancer samples', *Methods,* 59(1), pp. S20-3.

Hinman, M. N., Sharma, A., Luo, G. and Lou, H. (2014) 'Neurofibromatosis Type 1 Alternative Splicing Is a Key Regulator of Ras Signaling in Neurons', *Mol Cell Biol,* 34(12), pp. 2188-97.

Hodis, E., Watson, I. R., Kryukov, G. V., Arold, S. T., Imielinski, M., Theurillat, J. P., Nickerson, E., Auclair, D., Li, L., Place, C., Dicara, D., Ramos, A. H., Lawrence, M. S., Cibulskis, K., Sivachenko, A., Voet, D., Saksena, G., Stransky, N., Onofrio, R. C., Winckler, W., *et al.* (2012) 'A landscape of driver mutations in melanoma', *Cell: Vol. 2*. United States: 2012 Elsevier Inc, pp. 251-63.

Hosmer, D., Lemeshow, S. and Sturdivant, R. (2013) *Applied Logistic Regression. Wiley Series in Probability and Statistics* Third Edition edn. New Jersey: John Wiley & Sons, Inc.

Hwang, S.-J., Cheng, L., Lozano, G., Amos, C., Gu, X. and Strong, L. (2003) 'Lung cancer risk in germline p53 mutation carriers: association between an inherited cancer predisposition, cigarette smoking, and cancer risk', *Hum Genet,* 113(3), pp. 238-243.

Imielinski, M., Berger, Alice H., Hammerman, Peter S., Hernandez, B., Pugh, Trevor J., Hodis, E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A., Sougnez, C., Auclair, D., Lawrence, Michael S., *et al*. (2012) 'Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing', *Cell,* 150(6), pp. 1107-1120.

Iyengar, T. D., Ng, S., Lau, C. C., Welch, W. R., Bell, D. A., Berkowitz, R. S. and Mok, S. C. (1999) 'Differential expression of NF1 type I and type II isoforms in sporadic borderline and invasive epithelial ovarian tumors', *Oncogene,* 18(1), pp. 257-62.

Jabs, V., Edlund, K., Konig, H., Grinberg, M., Madjar, K., Rahnenfuhrer, J., Ekman, S., Bergkvist, M., Holmberg, L., Ickstadt, K., Botling, J., Hengstler, J. G. and Micke, P. (2017) 'Integrative analysis of genome-wide gene copy number changes and gene expression in non-small cell lung cancer', *PLoS One,* 12(11), pp. e0187246.

Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B. K., Veeriah, S., Shafi, S., Johnson, D. H., Mitter, R., Rosenthal, R., Salm, M., Horswell, S., Escudero, M., Matthews, N., Rowan, A., Chambers, T., Moore, D. A., Turajlic, S., Xu, H., Lee, S. M.,*et al*. (2017) 'Tracking the Evolution of Non-Small-Cell Lung Cancer', *N Engl J Med,* 376(22), pp. 2109-2121.

Janne, P. A., Shaw, A. T., Pereira, J. R., Jeannin, G., Vansteenkiste, J., Barrios, C., Franke, F. A., Grinsted, L., Zazulina, V., Smith, P., Smith, I. and Crino, L. (2013) 'Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell lung cancer: a randomised, multicentre, placebo-controlled, phase 2 study', *Lancet Oncol,* 14(1), pp. 38-47.

Janne, P. A., van den Heuvel, M. M., Barlesi, F., Cobo, M., Mazieres, J., Crino, L., Orlov, S., Blackhall, F., Wolf, J., Garrido, P., Poltoratskiy, A., Mariani, G., Ghiorghiu, D., Kilgour, E., Smith, P., Kohlmann, A., Carlile, D. J., Lawrence, D., Bowen, K. and Vansteenkiste, J. (2017) 'Selumetinib Plus Docetaxel Compared With Docetaxel Alone and Progression-

Free Survival in Patients With KRAS-Mutant Advanced Non-Small Cell Lung Cancer: The SELECT-1 Randomized Clinical Trial', *Jama,* 317(18), pp. 1844-1853.

Johannessen, C. M., Reczek, E. E., James, M. F., Brems, H., Legius, E. and Cichowski, K. (2005) 'The NF1 tumor suppressor critically regulates TSC2 and mTOR', *Proc Natl Acad Sci U S A,* 102(24), pp. 8573-8.

Katayama, R., Khan, T. M., Benes, C., Lifshits, E., Ebi, H., Rivera, V. M., Shakespeare, W. C., Iafrate, A. J., Engelman, J. A., Shaw, A. T. and Vogt, P. K. (2011) 'Therapeutic strategies to overcome crizotinib resistance in non-small cell lung cancers harboring the fusion oncogene EML4- ALK', *Proceedings of the National Academy of Sciences of the United States of America,* 108(18), pp. 7535-7540.

Kathryn, T. B., Stephano Spano, M. and Laura, D. A. (2014) 'Unravelling mechanisms of p53-mediated tumour suppression', *Nature Reviews Cancer,* 14(5), pp. 359.

Kerick, M., Isau, M., Timmermann, B., Sültmann, H., Herwig, R., Krobitsch, S., Schaefer, G., Verdorfer, I., Bartsch, G., Klocker, H., Lehrach, H. and Schweiger, M. R. (2011) 'Targeted high throughput sequencing in clinical cancer Settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity', *BMC Med Genomics,* 4, pp. 68.

Kim, N., Song, M., Kim, S., Seo, Y., Kim, Y. and Yoon, S. (2016) 'Differential regulation and synthetic lethality of exclusive RB1 and CDKN2A mutations in lung cancer', *Int J Oncol,* 48(1), pp. 367-75.

Kleinerman, R., Tarone, R., Abramson, D., Seddon, J., Li, F. and Tucker, M. A. (2000) 'Hereditary retinoblastoma and risk of lung cancer', *Journal Of The National Cancer Institute,* 92(24), pp. 2037-2039.

Knutsen, T., Padilla-Nash, H. M., Wangsa, D., Barenboim-Stapleton, L., Camps, J., McNeil, N., Difilippantonio, M. J. and Ried, T. (2010) 'Definitive molecular cytogenetic characterization of 15 colorectal cancer cell lines', *Genes Chromosomes Cancer,* 49(3), pp. 204-23.

Laycock-van Spyk, S., Thomas, N., Cooper, D. N. and Upadhyaya, M. (2011) 'Neurofibromatosis type 1-associated tumours: their somatic mutational spectrum and pathogenesis', *Hum Genomics,* 5(6), pp. 623-90.

Loboda, A., Nebozhyn, M., Klinghoffer, R., Frazier, J., Chastain, M., Arthur, W., Roberts, B., Zhang, T., Chenard, M., Haines, B., Andersen, J., Nagashima, K., Paweletz, C., Lynch, B., Feldman, I., Dai, H., Huang, P. and Watters, J. (2010a) 'A gene expression signature of RAS pathway dependence predicts response to PI3K and RAS pathway inhibitors and expands the population of RAS pathway activated tumors', *BMC Med Genomics,* 3, pp. 26.

Loboda, A., Paweletz, C., Nagashima, K., Andersen, J., Haines, B., Chenard, M., Zhang, T., Roberts, B., Arthur, W., Chastain, M., Frazier, J., Klinghoffer, R., Nebozhyn, M., Lynch, B., Feldman, I., Dai, H., Huang, P. and Watters, J. (2010b) 'A gene expression signature of RAS pathway dependence predicts response to PI3K and RAS pathway inhibitors and expands the population of RAS pathway activated tumors', *BMC Medical Genomics,* 3(1).

Luijten, M., Redeker, S., Minoshima, S., Shimizu, N., Westerveld, A. and Hulsebos, T. J. (2001) 'Duplication and transposition of the NF1 pseudogene regions on chromosomes 2, 14, and 22', *Hum Genet,* 109(1), pp. 109-16.

Luijten, M., Wang, Y., Smith, B. T., Westerveld, A., Smink, L. J., Dunham, I., Roe, B. A. and Hulsebos, T. J. (2000) 'Mechanism of spreading of the highly related neurofibromatosis type 1 (NF1) pseudogenes on chromosomes 2, 14 and 22', *Eur J Hum Genet,* 8(3), pp. 209-14.

Lynch, T. J., Bell, D. W., Sordella, R., Gurubhagavatula, S., Okimoto, R. A., Brannigan, B. W., Harris, P. L., Haserlat, S. M., Supko, J. G., Haluska, F. G., Louis, D. N., Christiani, D. C., Settleman, J. and Haber, D. A. (2004) 'Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non– Small- Cell Lung Cancer to Gefitinib', *The New England Journal of Medicine,* 350(21), pp. 2129-2139.

Maemondo, M., Inoue, A., Kobayashi, K., Sugawara, S., Oizumi, S., Isobe, H., Gemma, A., Harada, M., Yoshizawa, H., Kinoshita, I., Fujita, Y., Okinaga, S., Hirano, H., Yoshimori, K., Harada, T., Ogura, T., Ando, M., Miyazawa, H., Tanaka, T., Saijo, Y., Hagiwara, K., Morita, S. and Nukiwa, T. (2010) 'Gefitinib or Chemotherapy for Non– Small- Cell Lung Cancer with Mutated EGFR', *The New England Journal of Medicine,* 362(25), pp. 2380-2388.

Maertens, O., Johnson, B., Hollstein, P., Frederick, D., Cooper, Z., Messiaen, L., Bronson, R., McMahon, M., Granter, S., Flaherty, K., Wargo, J., Marais, R. and Cichowski, K. (2013a) 'Elucidating distinct roles for NF1 in melanomagenesis', *Cancer Research,* 73(8).

Maertens, O., Johnson, B., Hollstein, P., Frederick, D. T., Cooper, Z. A., Messiaen, L., Bronson, R. T., McMahon, M., Granter, S., Flaherty, K., Wargo, J. A., Marais, R. and Cichowski, K. (2013b) 'Elucidating Distinct Roles for NF1 in Melanomagenesis', *Cancer Discovery,* 3(3), pp. 338-349.

Malone, C. F., Fromm, J. A., Maertens, O., DeRaedt, T., Ingraham, R. and Cichowski, K. (2014) 'Defining key signaling nodes and therapeutic biomarkers in NF1-mutant cancers', *Cancer Discov,* 4(9), pp. 1062-73.

Marrero, D., Peralta, R., Valdivia, A., De la Mora, A., Romero, P., Parra, M., Mendoza, N., Mendoza, M., Rodriguez, D., Camacho, E., Duarte, A., Castelazo, G., Vanegas, E., Garcia, I., Vargas, C., Arenas, D., Jimenez, F. and Salcedo, M. (2012) 'The neurofibromin 1 type I isoform predominance characterises female population affected by sporadic breast cancer: preliminary data', *J Clin Pathol,* 65(5), pp. 419-23.

Matallanas, D., Birtwistle, M., Romano, D., Zebisch, A., Rauch, J., von Kriegsheim, A. and Kolch, W. (2011) 'Raf family kinases: old dogs have learned new tricks', *Genes Cancer,* 2(3), pp. 232-60.

Mathot, L., Wallin, M. and Sjoblom, T. (2013) 'Automated serial extraction of DNA and RNA from biobanked tissue specimens', *Bmc Biotechnology,* 13.

Mattocks, C., Baralle, D., Tarpey, P., Ffrench-Constant, C., Bobrow, M. and Whittaker, J. (2004) 'Automated comparative sequence analysis identifies mutations in 89% of NF1 patients and confirms a mutation cluster in exons 11- 17 distinct from the GAP related domain', *Journal Of Medical Genetics,* 41(4).

Mebratu, Y. and Tesfaigzi, Y. (2009) 'How ERK1/2 activation controls cell proliferation and cell death is subcellular localization the answer?', *Cell Cycle,* 8(8), pp. 1168-1175.

Michael, R. S., Peter, J. C. and Futreal, P. A. (2009) 'The cancer genome', *Nature,* 458(7239), pp. 719.

Miller, V. A., Kris, M. G., Shah, N., Patel, J., Azzoli, C., Gomez, J., Krug, L. M., Pao, W., Rizvi, N., Pizzo, B., Tyson, L., Venkatraman, E., Ben-Porat, L., Memoli, N., Zakowski, M., Rusch, V. and Heelan, R. T. (2004) 'Bronchioloalveolar pathologic subtype and smoking history predict sensitivity to gefitinib in advanced non-small-cell lung cancer', *J Clin Oncol,* 22(6), pp. 1103-9.

Moch, H., Seidel, D., Soltermann, A., Weiss, J., et al. and Sos, M. L. 2010. Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. American Association for the Advancement of Science (AAAS).

Mukhopadhyay, D., Anant, S., Lee, R. M., Kennedy, S., Viskochil, D. and Davidson, N. O. (2002) 'C→U Editing of Neurofibromatosis 1 mRNA Occurs in Tumors That Express Both the Type II Transcript and apobec-1, the Catalytic Subunit of the Apolipoprotein B mRNA–Editing Enzyme', *Am J Hum Genet,* 70(1), pp. 38-50.

NCT00890825 (2018) *AZD6244 in Combination With Docetaxel Versus Docetaxel Alone in KRAS Mutation Positive NSCLC Patients*. Clinical Trials Gov: U.S. National Library of Medicine. Available at: https://clinicaltrials.gov/ct2/show/NCT00890825 (Accessed: 19/10/2018 2018).

NCT01514864 (2015) *Trial of Dasatinib in Patients With Advanced Cancers Harboring DDR2 Mutation or Inactivating B-RAF Mutation*. Clinical Trials Gov: U.S National Library of Medicine. Available at: https://clinicaltrials.gov/ct2/show/NCT01514864 (Accessed: 18/10/2018 2018).

NCT01933932 (2018) *Assess Efficacy & Safety of Selumetinib in Combination With Docetaxel in Patients Receiving 2nd Line Treatment for v-Ki-ras2 Kirsten Rat Sarcoma Viral Oncogene Homolog (KRAS) Positive NSCLC (SELECT-1)*. Clinical Trials Gov: U.S. National Library of Medicine Available at: https://clinicaltrials.gov/ct2/show/NCT01933932 (Accessed: 19/10/2018 2018).

NCT02664935 (2018) *National Lung Matrix Trial: Multi-drug Phase II Trial in Non-Small Cell Lung Cancer*. Clinical Trial Gov: U.S. National Library of Medicine Available at: https://clinicaltrials.gov/ct2/show/NCT02664935 (Accessed: 19/10/2018 2018).

NCT02965378 (2017) *Lung-MAP: AZD4547 as Second-Line Therapy in Treating FGFR Positive Patients With Recurrent Stage IV Squamous Cell Lung Cancer*. Clinical Trials Gov: U.S. National Library of Medicine. Available at: https://clinicaltrials.gov/ct2/show/NCT02965378 (Accessed: 18/10/2018 2018).

NICE (2014) *Afatinib for treating epidermal growth factor receptor mutation-positive locally advanced or metastatic non-small-cell lung cancer [TA310]*. The National Institute for Clinical Excellence. Available at: https://www.nice.org.uk/guidance/ta310 (Accessed: 04/03/2015 2015).

Oh, E., Choi, Y. L., Kwon, M. J., Kim, R. N., Kim, Y. J., Song, J. Y., Jung, K. S. and Shin, Y. K. (2015) 'Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples', *PLoS One,* 10(12).

Ozawa, T., Araki, N., Yunoue, S., Tokuo, H., Feng, L., Patrakitkomjorn, S., Hara, T., Ichikawa, Y., Matsumoto, K., Fujii, K. and Saya, H. (2005) 'The neurofibromatosis type 1 gene product neurofibromin enhances cell motility by regulating actin filament dynamics via the Rho-ROCK-LIMK2-cofilin pathway', *J Biol Chem,* 280(47), pp. 39524-33.

Paik, P. K., Johnson, M., Angelo, S., Sima, C. S., Ang, D., Dogan, S., Miller, V. A., Ladanyi, M., Kris, M. G. and Riely, G. (2012) 'Driver mutations determine survival in smokers and never-smokers with stage IIIB/IV lung adenocarcinomas', *Cancer,* 118(23), pp. 5840-5847.

Pao, W. and Chmielecki, J. 2010. Rational, biologically based treatment of EGFR- mutant non- small- cell lung cancer. *Nat. Rev. Cancer.*

Pao, W., Miller, V., Zakowski, M., Doherty, J., Politi, K., Sarkaria, I., Singh, B., Heelan, R., Rusch, V., Fulton, L., Mardis, E., Kupfer, D., Wilson, R., Kris, M. and Varmus, H. (2004) 'EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib', *Proc Natl Acad Sci U S A,* 101(36), pp. 13306-11.

Parkin, D. M. (2011) 'Tobacco-attributable cancer burden in the UK in 2010', *British Journal Of Cancer,* 105, pp. S6-S13.

Pasmant, E., Parfait, B., Luscan, A., Goussard, P., Briand-Suleau, A., Laurendeau, I., Fouveaut, C., Leroy, C., Montadert, A., Wolkenstein, P., Vidaud, M. and Vidaud, D. (2015) 'Neurofibromatosis type 1 molecular diagnosis: what can NGS do for you when you have a large gene with loss of function mutations?', *Eur J Hum Genet,* 23(5), pp. 596-601.

Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X. J., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., Reymond, A., Hubbard, T. J., Harrow, J. and Gerstein, M. B. (2012) 'The GENCODE pseudogene resource', *Genome Biol,* 13(9), pp. R51.

Perou, C. (2012) 'Comprehensive molecular portraits of human breast tumors', *Nature,* 490(7418), pp. 61-70.

Pertea, M., Lin, X. and Salzberg, S. L. (2001) 'GeneSplicer: a new computational method for splice site prediction', *Nucleic Acids Res,* 29(5), pp. 1185-90.

Rasmussen, S. A., Overman, J., Thomson, S. A., Colman, S. D., Abernathy, C. R., Trimpert, R. E., Moose, R., Virdi, G., Roux, K., Bauer, M., Rojiani, A. M., Maria, B. L., Muir, D. and Wallace, M. R. (2000) 'Chromosome 17 loss-of-heterozygosity studies in benign and malignant tumors in neurofibromatosis type 1', *Genes Chromosomes Cancer,* 28(4), pp. 425-31.

Ratner, N. and Miller, S. J. (2015) 'A RASopathy gene commonly mutated in cancer: the neurofibromatosis type 1 tumour suppressor', *Nat Rev Cancer,* 15(5), pp. 290-301.

Reese, M. G., Eeckman, F. H., Kulp, D. and Haussler, D. (1997) 'Improved splice site detection in Genie', *J Comput Biol,* 4(3), pp. 311-23.

Reis, P. P., Waldron, L., Goswami, R. S., Xu, W., Xuan, Y., Perez-Ordonez, B., Gullane, P., Irish, J., Jurisica, I. and Kamel-Reid, S. (2011) 'mRNA transcript quantification in archival samples using multiplexed, color-coded probes', *BMC Biotechnol,* 11, pp. 46.

Riccardi, V. M. and Lewis, R. A. (1988) 'Penetrance of von Recklinghausen neurofibromatosis: A distinction between predecessors and descendants', *American Journal of Human Genetics,* 42(2), pp. 284-289.

Rojewski, A. M., Baldassarri, S., Cooperman, N. A., Gritz, E. R., Leone, F. T., Piper, M. E., Toll, B. A. and Warren, G. W. (2016) 'Exploring Issues of Comorbid Conditions in People Who Smoke', *Nicotine Tob Res,* 18(8), pp. 1684-96.

Rosell, R., Carcereny, E., Gervais, R., Vergnenegre, A., Massuti, B., Felip, E., Palmero, R., Garcia-Gomez, R., Pallares, C., Sanchez, J., Porta, R., Cobo, M., Garrido, P., Longo, F., Moran, T., Insa, A., De Marinis, F., Corre, R., Bover, I., Illiano, A., Dansin, E., de Castro, J., *et al*. (2012) 'Erlotinib versus standard chemotherapy as first- line treatment for European patients with advanced EGFR mutation- positive non- small- cell lung cancer (EURTAC): a multicentre, open- label, randomised phase 3 trial', *Lancet Oncol.,* 13(3), pp. 239-246.

Roskoski, R., Jr. (2012) 'MEK1/2 dual-specificity protein kinases: structure and regulation', *Biochem Biophys Res Commun,* 417(1), pp. 5-10.

Ross, J. S., Ali, S. M., Wang, K., Palmer, G., Yelensky, R., Lipson, D., Miller, V. A., Zajchowski, D., Shawver, L. K. and Stephens, P. J. (2013) 'Comprehensive genomic profiling of epithelial ovarian cancer by next generation sequencing-based diagnostic assay reveals new routes to targeted therapies', *Gynecol Oncol,* 130(3), pp. 554-9.

Sah, S., Chen, L., Houghton, J., Kemppainen, J., Marko, A. C., Zeigler, R. and Latham, G. J. (2013) 'Functional DNA quantification guides accurate next-generation sequencing mutation detection in formalin-fixed, paraffin-embedded tumor biopsies', *Genome Med,* 5(8), pp. 77.

Scheffler, M., Bos, M., Gardizi, M., König, K., Michels, S., Fassunke, J., Heydt, C., Künstlinger, H., Ihle, M., Ueckeroth, F., Albus, K., Töpelt, K., Nogova, L., Zander, T., Merkelbach-Bruse, S., Heukamp, L. C., Büttner, R., Wolf, J., Serke, M., Gerigk, U., Schulte, W., *et al*. (2015) 'PIK3CA mutations in non-small cell lung cancer (NSCLC): Genetic heterogeneity, prognostic impact and incidence of prior malignancies', *Oncotarget,* 6(2), pp. 1315-1326.

Siegelin, M. D. and Borczuk, A. C. (2014) 'Epidermal growth factor receptor mutations in lung adenocarcinoma', *Laboratory Investigation,* 94(2), pp. 129-137.

Sikorsky, J. A., Primerano, D. A., Fenger, T. W. and Denvir, J. (2007) 'DNA damage reduces Taq DNA polymerase fidelity and PCR amplification efficiency', *Biochem Biophys Res Commun,* 355(2), pp. 431-7.

Silva, S., Danson, S., Teare, D., Taylor, F., Bradford, J., McDonagh, A. J. G., Salawu, A., Wells, G., Burghel, G. J., Brock, I., Connley, D., Cramp, H., Hughes, D., Tiffin, N. and Cox, A. (2018) 'Genome-Wide Analysis of Circulating Cell-Free DNA Copy Number Detects Active Melanoma and Predicts Survival', *Clin Chem,* 64(9), pp. 1338-1346.

Simbolo, M., Gottardi, M., Corbo, V., Fassan, M., Mafficini, A., Malpeli, G., Lawlor, R. and Scarpa, A. (2013) 'DNA Qualification Workflow for Next Generation Sequencing of Histopathological Samples', *Plos One,* 8(6).

Sohier, P., Luscan, A., Lloyd, A., Ashelford, K., Laurendeau, I., Briand-Suleau, A., Vidaud, D., Ortonne, N., Pasmant, E. and Upadhyaya, M. (2017) 'Confirmation of mutation landscape of NF1-associated malignant peripheral nerve sheath tumors', *Genes Chromosomes Cancer,* 56(5), pp. 421-426.

Spencer, D. H., Sehn, J. K., Abel, H. J., Watson, M. A., Pfeifer, J. D. and Duncavage, E. J. (2013) 'Comparison of Clinical Targeted Next-Generation Sequence Data from Formalin-Fixed and Fresh-Frozen Tissue Specimens', *J Mol Diagn,* 15(5), pp. 623-33.

Srinivasan, M., Sedmak, D. and Jewell, S. 2002. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. *Am. J. Pathol.*

Starinsky-Elbaz, S., Faigenbloom, L., Friedman, E., Stein, R. and Kloog, Y. (2009) 'The pre-GAP-related domain of neurofibromin regulates cell migration through the LIM kinase/cofilin pathway', *Molecular and Cellular Neuroscience,* 42(4), pp. 278-287.

Suzuki, H., Takahashi, K., Kubota, Y. and Shibahara, S. (1992) 'MOLECULAR- CLONING OF A CDNA CODING FOR NEUROFIBROMATOSIS TYPE-1 PROTEIN ISOFORM LACKING THE DOMAIN RELATED TO RAS GTPASE-ACTIVATING PROTEIN', *Biochem. Biophys. Res. Commun.,* 187(2), pp. 984-990.

Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., Vandergriff, J. A. and Doremieux, O. (2003) 'PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification', *Nucleic Acids Res,* 31(1), pp. 334-41.

Tonsgard, J. H. (2006) 'Clinical manifestations and management of neurofibromatosis type 1', *Semin Pediatr Neurol,* 13(1), pp. 2-7.

Tutar, Y. (2012) 'Pseudogenes', *Comp Funct Genomics,* 2012.

Upadhyaya, M., Han, S., Consoli, C., Majounie, E., Horan, M., Thomas, N. S., Potts, C., Griffiths, S., Ruggieri, M., von Deimling, A. and Cooper, D. N. (2004) 'Characterization of the somatic mutational spectrum of the neurofibromatosis type 1 (NF1) gene in neurofibromatosis patients with benign and malignant tumors', *Hum Mutat,* 23(2), pp. 134-46.

Upadhyaya, M., Majounie, E., Thompson, P., Han, S., Consoli, C., Krawczak, M., Cordeiro, I. and Cooper, D. N. (2003) 'Three different pathological lesions in the NF1 gene originating de novo in a family with neurofibromatosis type 1', *Hum Genet,* 112(1), pp. 12-7.

Vallée, B., Doudeau, M., Godin, F., Gombault, A., Tchalikian, A., de Tauzia, M. L. and Bénédetti, H. (2012) 'Nf1 RasGAP Inhibition of LIMK2 Mediates a New Cross-Talk between Ras and Rho Pathways', *PLoS One,* 7(10).

Veldman-Jones, M. H., Brant, R., Rooney, C., Geh, C., Emery, H., Harbron, C. G., Wappett, M., Sharpe, A., Dymond, M., Barrett, J. C., Harrington, E. A. and Marshall, G. (2015) 'Evaluating Robustness and Sensitivity of the NanoString Technologies nCounter Platform to Enable Multiplexed Gene Expression Analysis of Clinical Samples', *Cancer Res,* 75(13), pp. 2587-93.

Viskochil, D., Carey, J., Cawthon, R., Connell, P., Xu, G., Stevens, J., White, R. and Culver, M. (1991) 'The Gene Encoding the Oligodendrocyte- Myelin Glycoprotein Is Embedded within the Neurofibromatosis Type 1 Gene', *Molecular and Cellular Biology,* 11(2), pp. 906-912.

Vojtek, A. B., Hollenberg, S. M. and Cooper, J. A. (1993) 'Mammalian Ras interacts directly with the serine/threonine kinase raf', *Cell,* 74(1), pp. 205-214.

Welti, S., Kuhn, S., Angelo, I., Brugger, B., Kaufmann, D. and Scheffzek, K. (2011) 'Structural and Biochemical Consequences of NF1 Associated Nontruncating Mutations in the Sec14- PH Module of Neurofibromin', *Hum. Mutat.,* 32(2), pp. 191-197.

Whittaker, S. R., Theurillat, J.-P., Van Allen, E., Wagle, N., Hsiao, J., Cowley, G. S., Schadendorf, D., Root, D. E. and Garraway, L. A. (2013) 'A Genome-Scale RNA Interference Screen Implicates NF1 Loss in Resistance to RAF Inhibition', *Cancer Discovery,* 3(3), pp. 350-362.

Wu Thomas, D., Guerrero, S., Fedorowicz, G. and Modrusan, Z. (2009) 'Microarray analysis of RNA extracted from formalin-fixed, paraffin-embedded and matched fresh-frozen ovarian adenocarcinomas', *BMC Medical Genomics,* 2(1).

Xu, G., Lin, B., Tanaka, K., Dunn, D., Wood, D., Gesteland, R., White, R., Weiss, R. and Tamanoi, F. (1990) 'The catalytic domain of the neurofibromatosis type 1 gene product stimulates ras GTPase and complements ira mutants of S. cerevisiae', *Cell,* 63(4), pp. 835-841.

Yamamoto, H., Shigematsu, H., Nomura, M., Lockwood, W. W., Sato, M., Okumura, N., Soh, J., Suzuki, M., Wistuba, I. I., Fong, K. M., Lee, H., Toyooka, S., Date, H., Lam, W. L., Minna, J. D. and Gazdar, A. F. (2008) 'PIK3CA mutations and copy number gains in human lung cancers', *Cancer research,* 68(17), pp. 6913.

Yang, F.-C., Chen, S., Yuan, J., Li, X., Yang, X., Knowles, S., Horn, W., Li, Y., Yang, Y., Robertson, K., Clapp, D. W., Ingram, D. A., Hutchins, G., Zhang, S., Vakili, S. T., Yu, M., Zhu, Y., Parada, L. F. and Burns, D. (2008) 'Nf1- Dependent Tumors Require a Microenvironment Containing Nf1 +/- - and c- kit- Dependent Bone Marrow', *Cell,* 135(3), pp. 437-448.

Yeo, G. and Burge, C. B. (2004) 'Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals', *J Comput Biol,* 11(2-3), pp. 377-94.

Zarich, N., Oliva, J. L., Martinez, N., Jorge, R., Ballester, A., Gutierrez-Eisman, S., Garcia-Vargas, S. and Rojas, J. M. (2006) 'Grb2 is a negative modulator of the intrinsic Ras- GEF activity of hSos1', *Mol. Biol. Cell,* 17(8), pp. 3591-3597.

Zhang, M. Q. (1998) 'Statistical features of human exons and their flanking regions', *Hum Mol Genet,* 7(5), pp. 919-32.

Zhang, Y., Tang, E. T. and Du, Z. (2016) 'Detection of MET Gene Copy Number in Cancer Samples Using the Droplet Digital PCR Method', *PLoS One,* 11(1).

Zhang, Y. L., Yuan, J. Q., Wang, K. F., Fu, X. H., Han, X. R., Threapleton, D., Yang, Z. Y., Mao, C. and Tang, J. L. (2016) 'The prevalence of EGFR mutation in patients with non-small cell lung cancer: a systematic review and meta-analysis', *Oncotarget,* 7(48), pp. 78985-78993.

Appendices

Appendix 1. Project Protocol

Next-Generation Sequencing into practice: Identifying functionally important aberrations in neurofibromin-1 (NF-1) in non-small cell lung cancer (NSCLC)

Investigators:

Professor Sarah Danson, Academic Unit of Clinical Oncology, University of Sheffield

Professor Ann Dalton , Director, Sheffield Diagnostic Genetics Service, Sheffield Children's NHS Trust

Miranda Durkie Lead Clinical Scientist, Sheffield Diagnostic Genetics Service, Sheffield Children's NHS Trust

Dr Matthew Parker Lead Bioinformatician, Sheffield Diagnostic Genetics Service, Sheffield Children's NHS Trust

Dr Marianne Ratcliffe, AstraZeneca

Dr Alexander Kohlmann, AstraZeneca

Greg Wells, Academic Unit of Oncology and Metabolism , University of Sheffield

Chief Investigator/Coordinator

Professor Sarah Danson

Academic Unit of Oncology
University of Sheffield
Weston Park Hospital
Whitham Road
Sheffield, S10 2SJ
Phone: 0114 226 5235
Fax: 0114 226 5364
Email: S.Danson@sheffield.ac.uk

STH project reference number: 18741

IRAS Project ID: 165821

Summary:

Lung cancer kills more people than any other cancer and NSCLC is the commonest type of lung cancer. Mainstay treatment options for advanced disease include chemotherapy, which offers limited benefits, or, for a small percentage of patients, new targeted agents offer a better outlook. This percentage could increase with further targeted therapies currently in clinical trials. In order to maximise the benefits these new agents will bring, we need to understand the underlying mechanisms and identify biomarkers which will allow successful stratification and selection of patients.

Neurofibromatosis is a common inherited disorder affecting approximately 1 in 3000 individuals. The disease is the result of mutations in the tumour suppressor gene NF-1. These mutations leave the individual with increased susceptibility to the development of specific benign and malignant tumours. NF-1 functions to inhibit the activation of cellular signalling pathways which are widely associated with cancer. It has recently been observed that sporadic NF-1 mutations can be found in many different cancers. Mutations in NF-1 are observed in 8-12% of NSCLC tumours. Inactivation of NF-1 has already been linked to possible acquired resistance to targeted therapeutics in models for NSCLC. However, a link between mutations found in NF-1 and functional effects on cell signalling pathways is lacking.

This study aims to address this, by conducting genetic analysis of NF-1 mutations and linking these to cell signalling pathway activation, commonly related to NSCLC. Next generation sequencing will enable rapid analysis of NF-1 and further establish this platform as a diagnostic tool. If successful, this could lead to NF-1 status being used as a biomarker to predict response to cancer treatments, such as inhibitors of this pathway that are currently involved in several clinical trials.

Background:

Cancer research UK statistics showed there were 43,463 new cases of lung cancer in 2011, making this the second most commonly diagnosed cancer in the UK. Of these 87% were classified as NSCLC (CRUK 2014). Prognosis with lung cancer is poor, with a 5 year survival rate of less than 10% (CRUK 2014). This is largely due to late stage at diagnosis and lack of effective treatments. 68% of cases were stage III and IV (CRUK 2014). At this late stage curative measures such as surgical intervention and radical radiotherapy are not possible. Systemic chemotherapy, such as platinum doublet regimens carry a median survival of 9 months (Rossi et al. 2014). In recent years targeted agents have become available and approved by NICE as first line therapies for tumours with specific genetic characteristics. These targeted agents include Gefitinib, Erlotinib, and Afatinib and have demonstrated significant survival benefits for patients harbouring epidermal growth factor receptor (EGFR) mutations (Maemondo et al. 2010, Rosell et al. 2012, Haaland et al. 2014). Other targeted therapies include Crizotinib which is used as a second line chemotherapy with ALK positive tumours. However, these targeted options are only available for a minority of patients. Unfortunately, the tumours eventually develop resistance to these inhibitors and the disease progresses. Substantial research is being undertaken to characterise these resistance mechanisms, and develop further targeted agents which circumvent these mechanisms. New combinations of targeted agents are in current clinical trials and these could dramatically increase the numbers of NSCLC patients who respond to these therapies (Giaccone 2014). However, a greater insight into causes of NSCLC dysfunctional cellular signalling is required in order to advance therapy.

Neurofibromin 1 (NF-1) is most commonly known for its role in Neurofibromatosis, otherwise known as von Recklinghausen's disease. This disorder is inherited as an autosomal dominant disease and is associated with loss of function mutations of the NF-1 gene on chromosome 17 (Yap et al. 2014). This increases susceptibility to a wide range of malignant and benign tumours of the peripheral nerve sheath (Yap et al. 2014). NF-1 functions as a negative regulator of Ras which is involved in regulation of several cell signalling pathways, dysregulation of which is associated to various types of cancer. Common genetic mutations in the genes encoding these pathways results in their permanent activation, thus enabling cancerous cells to grow in an unregulated manner.

The two most common histological types of NSCLC are lung adenocarcinoma and squamous cell carcinoma. Recent genetic profiling studies of these have revealed somatic mutations of NF-1 in 8-12% (Collisson et al. 2014, Imielinski et al. 2012, Hammerman et al. 2012, Ding et al. 2008). Furthermore, there is evidence that inactivation of NF-1 is associated with acquired resistance to targeted agents in lung cancers (Beauchamp et al. 2014: de Bruin et al. 2014). Bruin et al (2014)

demonstrated a dual therapy regime of EGFR inhibitor and a downstream MEK inhibitor overcomes this resistance. Currently clinical trials are underway investigating this combination of therapeutic agents (Giaccone 2014). However, the functional relevance of these NF-1 somatic mutations and their link to cell signalling remains poorly investigated and understood.

Advances in genomic analysis technologies have greatly improved our capabilities to detect genomic mutations and linking then to functional outcomes. This study aims to link NF-1 somatic mutations observed in the NSCLC population to the regulation of cellular signalling pathways, such as Ras/Raf/MEK/ERK activation. Knowledge of this could be highly relevant for future treatment strategies and monitoring disease progression in NSCLC. For example; recognising the use of MEK targeted inhibitors to reduce ERK activation if NF-1 is non-functional. Currently genetic tests used to aid diagnosis and stratify patients for specific treatments for NSCLC do not include NF-1 mutations. Characterisation of NF-1 mutations could lead to them becoming a valuable diagnostic marker and possible prognostic factor enabling improved disease management increasing patient wellbeing.

Study Aim:

To identify NF-1 variants observed in the non-small cell lung cancer population and determine their effect on cellular signalling. Specific somatic mutations associated with NSCLC known to activate the Ras/Raf/MEK/ERK pathway will also be analysed. In addition, cell free circulating tumour DNA (cfDNA) will be analysed in order to determine if the NF-1 genetic variants and key NSCLC related mutations can be detected in patient blood samples.

Study Objectives:

1. To recruit 100 patients with non-small cell lung cancer over 4 years, obtain a single set of blood samples, and access their archived tumour.
2. To investigate whether somatic NF-1 variants are present in archived tumour tissue and whether NF-1 status correlates with histological subtype, stage, response to treatment or survival.
3. To determine if there is any relationship between observed NF-1 variants and Ras/Raf/MEK/ERK activation. Known specific mutations in EGFR, BRAF, and KRAS commonly associated with NSCLC and activation of the Ras/Raf/MEK/ERK pathway will also be analysed from the tumour samples.
4. To assess if somatic NF-1 variants and other NSCLC related mutations present in archived tumour tissue can be detected in cell free circulating tumour DNA (cfDNA) extracted from a blood sample.

Study design:

Non-small cell lung cancer population-based study which aims to collect detailed data on NF-1 genetic variants and measure their effect on cellular signalling. The status of known activating mutations in other significant genes (EGFR, BRAF, and KRAS) involved in NSCLC aberrant cellular signalling will also be collected to rule out their effects on pathway activation in relation to NF-1 status.

Study participants:

Eligibility: patients

- Pathological (cytological or histological) confirmation of non-small cell lung cancer
- Archived tissue available
- Willing to have a blood test
- Able to give written informed consent

Study procedures:

Over a period of 4 years, we aim to recruit 100 patients with NSCLC. Patients will be recruited at Weston Park Hospital. Eligible patients will be identified by their own clinicians who are members of the Weston Park Lung Cancer Group or by screening notes by the research team. We aim to identify and recruit all eligible patients seen in our clinics. The patient will be approached in their clinic if their clinician feels that it is appropriate. If they are interested then they will be given a patient information leaflet and consented then or a plan will be made for when the study will be next discussed e.g. with a phone call or at next scheduled clinic visit. If the patient wishes to proceed then written informed consent will be obtained for blood tests, access to medical records, and access to archived tumour tissue. This will be the end of involvement in the study for the patient.

ReSoLuCENT Samples

If we do not archive or target of 100 patients, the number of sample will be made up to 100 using samples remaining from the ReSoLuCENT study (STH13872 05/MRE07/72). Patients in the ReSoLuCENT study have provided consent for the samples to be used for future research. These samples and non-identifiable data will be provided in an anonymised manner. ReSoLuCENT samples will not be subject to cfDNA analysis.

Study visit:

The patient will only be involved in one visit, to give consent and have blood samples taken. Patients will be asked for permission for their archived tumour to be accessed and analysed for DNA and mRNA studies.

All samples and data will be pseudo-anonymised (labelled with a study identification number) before leaving STH. Data and samples will be then transferred to Sheffield Children's Hospital in person; all data will be stored on a password-protected computer. Data will be transferred using an encrypted USB drive, as soon as transfer is complete the data will be deleted from the USB drive.

ReSoLuCENT Patients:

ReSoLuCENT samples and none identifiable information will be provided in an anonymised manner so patients will not be subject to any visits in relation to the study.

Sample processing/storage:

At the visit, 50ml of blood will be collected in EDTA vacutainers on ice. 2 whole blood EDTA samples will be stored at –20°C prior to leucocyte DNA extraction for candidate gene analysis. With the other EDTA samples, within an hour of collection, plasma will be separated from nucleated cells and stored at –80°C for cfDNA analysis. Leukocyte and cfDNA will only be screened for NF-1 mutations in patients with NF-1 tumour mutations. FFPE samples will be stored at ambient temperature.

FFPE archived tumour tissue, blood and derived plasma samples will be stored at Sheffield Children's Hospital (SCH) within the Sheffield Genetics Diagnostics Services department. FFPE tissue will be returned to STH after DNA and RNA extraction within 3 to 9 months. Blood and plasma samples will be disposed of after DNA extraction, or 1 year after completion of the study whichever is sooner. RNA and FFPE archived tumour tissue will be transferred to, and analysed at AstraZeneca. These samples will be returned to SCH after analysis. DNA and RNA will be stored at SCH archive facility for 5 years after study completion.

Samples stored at SCH will be tracked using StarLIMS throughout the study. Samples will only be accessible by members of this department and only identifiable by their pseudo-anonymised identification and a bar code added at SCH. RNA stored at AstraZeneca during sample analysis will be

tracked using AstraZeneca's UK Biobank and only identifiable by their pseudo-anonymised identification and barcode.

Clinical and Research Data Storage:

Clinical and personal data will be stored at the SCH for 6 to 12 months after study completion. Research data will be stored at SCH, STH, and AstraZeneca for 20 years after study completion. The pseudo-anonymised identification link will remain at STH to allow Professor Sarah Danson to monitor patient status after study completion, providing further data related to the significance of NF-1 as a prognostic factor in NSCLC.

Sample analysis:

DNA analysis:

The study aims to screen for somatic mutations of NF-1 in patients diagnosed with NSCLC and determine if these mutations correlate with activation of the Ras/Raf/MEK/ERK pathway. Genomic DNA will be extracted from formalin fixed paraffin embedded archived tumour samples. In NF-1 mutation positive samples, cell free DNA will be extracted from plasma using the QIAamp circulating nucleic acid kit, and genomic DNA extracted from leukocytes in the blood samples. Agilent's SureSelect custom panel will be used to enrich and amplify exonic loci including NF-1, creating a DNA library using a method validated at Sheffield Children's Hospital. The DNA libraries will then be screened for genetic mutations using next generation sequencing. Raw data generated will be DNA sequence, this will be analysed using bioinformatics software based at SCH,the University of Sheffield, and AstraZeneca. Somatic variants of NF-1 be determined and annotated. Only key activating mutations known to influence the MAPK pathway in NSCLC in will be analysed in EGFR, KRAS, and BRAF genes. These mutations are not found in germline DNA, therefore only the tumour DNA will be analysed.

FFPE tumour DNA will also be subject to copy number variation analysis to determine if there are multiple copies or deletions of exonic loci in the NF-1 gene. This can be accomplished using the data generated by next generation sequencing, multiplex Ligation-dependent probe amplification or digital droplet polymerase chain reaction.

RNA analysis:

Activation of the Ras/Raf/MEK/ERK pathway will be investigated using a MEK gene activating signature assay. mRNA will be extracted from formalin fixed paraffin embedded tumour samples using the Qiagen RNeasy FFPE extraction kit. Gene expression of specific genes associated with MEK activation including DUSP4, ETV4, ETV5, PHLDA1 and SPRY2 will be analysed in order to determine the pathways activation status (Dry et al, 2010). NF1 gene expression could also be analysed comparatively across all samples to determine if transcription of NF1 also correlates to MEK activation.

Immunohistochemistry analysis:

FFPE archived tumour tissue slides will be analysed via immunohistochemistry at AstraZeneca. NF1 will be targeted to investigate protein expression and see it this correlates with NF1 mRNA expression and has any influence over the MAPK pathway.

Incidental Findings:

Initially the study will only be screening tumour tissue for *de novo* NF1 mutations which are observed in 10% of NSCLC tumours. *De novo* mutations without the germline mutation are of no clinical significance to the patient or their family with regards to neurofibromatosis and so we do not propose to inform the patient of this result.

If NF-1 mutations are found in the tumour samples, DNA from leukocytes within the blood sample will be screened for NF-1 mutations to confirm the tumour mutations are *de novo*, not germline. Germline mutations in the NF-1 gene are linked to the heredity disease Neurofibromatosis. Neurofibromatosis is highly penetrant and easily diagnosed so it is not anticipated that germline NF-1 mutations will be detected without a previous diagnosis. Patients enrolling into the study will be informed of this link in the patient information sheet and given the option on the consent form if they wish to be informed of any germline NF-1 mutations taken from the blood samples. Should the situation occur where a germline mutation is found and the patient did not have a previous diagnosis of neurofibromatosis then the patient will be informed and a referral would be made to Clinical Genetics.

Statistical considerations:

For the primary outcome the NSCLC population of the study will be arranged into three groups for further statistical analysis: NF-1 loss of function genetic mutation, MAPK associated wildtype, and MAPK known activating mutations. The NF-1 genetic mutation group will be assigned actual mutation identification. Based on current literature it is expected approximately 10% of the NSCLC population have NF-1 mutations. Links between the groups and epidemiological and clinical data will be measured to see if any trends are observed.

A MEK gene activation signature score will be calculated based on the activation of specific genes regulated by this pathway. This score should have direct correlation to activation of MEK. Parametric or non-parametric statistical analysis will be used depending on the distribution of the MEK gene

X

activation signature score of the groups. Status of other significant genes known to be associated with MEK activation will also be considered in order to rule out their effect on MEK activation.

If genetic NF-1 variants are observed in tumour samples cell free circulating tumour DNA would be analysed to determine it the NF-1 variants could be detected from plasma. This would be qualitative data comparing NF-1 genetic mutations in individuals. This will be calculated as a percentage as only 10 patients samples are estimated to contain mutations. Methods of statistical analysis may be adapted as the project progresses.

Project management:

Timeline:

Proposed study start date:	May 2015
Proposed study end date:	May 2018 (for recruitment)
	December 2018 (for analysis)

Years 1,2,3, & 4:	Recruitment, DNA extraction, genetic analysis of NF-1, data entry, and generation of descriptive statistics of the cohort.
Year 3 & 4:	Analysis of genomic copy number profiles. MEK gene activation signature score determined, mRNA NF-1 gene expression and statistical analysis between groups.

Managerial responsibilities:

Professor Sarah Danson will have overall responsibility for coordinating the study. Dr Danson or clinical delegate will have responsibility for patient recruitment and consent. Clinical data collection will be carried out by Dr Sarah Danson or clinical delegate. Dr Gill Wilson will assist with project management and supervise laboratory analysis. Miranda Durkie will supervise methods and laboratory analysis. Dr Matthew Parker will be responsible for supervision of the statistical and bioinformatics analysis. Dr Marianne Ratcliffe and Dr Alexander Kohlmann will assist with project management and supervise laboratory analysis during analysis at AstraZeneca.

Ethical considerations

The study involves genetic research and this will be made clear to the patient from the outset.

Issues may include:

1. Confidentiality: all samples and data will be pseudo-anonymised and protected.

2. Medical insurance: individual results will not be available so there will not be these implications.

3. Use of diagnostic material: any archived material used will be surplus to requirements.

The study will be subject to approval by the appropriate research ethics committee and registered under research governance procedures.

Dissemination of research results

Results will be published in peer reviewed scientific journals, presented at scientific conferences, and published in the form of a thesis at White Rose eThesis Online.

Intellectual Property

Intellectual property will rest between Sheffield Teaching Hospitals, Sheffield Children's Hospital, The University of Sheffield and the relevant funding organisations.

References

Beauchamp, E. M. (2014) 'Acquired Resistance to Dasatinib in Lung Cancer Cell Lines Conferred by DDR2 Gatekeeper Mutation and NF1 Loss', Molecular Cancer Therapeutics, 13(2), 475-482.

CRUK.(2014). Cancer Research UK. [Online]. Available from: http://www.cancerresearchuk.org/cancer-info/cancerstats/keyfacts/lung-cancer. [Accessed 13th December 2014]

Collisson, E. A., Campbell, J. D., Brooks, A. N., et al. (2014) 'Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network', Nature, 511(7511), 543-550.

de Bruin, E. C., Cowell, C., Warne, P. H., Jiang, M., Saunders, R. E., Melnick, M. A., Gettinger, S., Walther, Z., Wurtz, A., Heynen, G. J., Heideman, D. A. M., Gomez-Roman, J., Garcia-Castano, A., Gong, Y., Ladanyi, M., Varmus, H., Bernards, R., Smit, E. F., Politi, K. and Downward, J. (2014) 'Reduced NF1 Expression Confers Resistance to EGFR Inhibition in Lung Cancer', Cancer Discovery, 4(5), 606-619.

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., et al. (2008) 'Somatic mutations affect key pathways in lung adenocarcinoma', Nature, 455(7216), 1069-1075.

Dry J.R., Pavey S., Pratilas C.A., et al. (2010) 'Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244)', Cancer Research, 70(6), 2264-2273.

Haaland, B., Tan, P. S., de Castro, G., Jr. and Lopes, G. (2014) 'Meta-analysis of first-line therapies in advanced non-small-cell lung cancer harboring EGFR-activating mutations', J Thorac Oncol, 9(6), 805-11.

Hammerman, P. S., Voet, D., Lawrence, M. S., Voet, D., Jing, R., et al. (2012) 'Comprehensive genomic characterization of squamous cell lung cancers', Nature, 489(7417), 519-525.

Giaccone, G,. (2014) Randomized Phase II Study of AZD6244 MEK-Inhibitor and/or Erlotinib in KRAS Wild Type or Mutant KRAS Advanced Non-Small Cell Lung Cancer (NCI-10-C-0218) [Online]. Available from: http://www.cancer.gov/clinicaltrials/featured/trials/NCI-10-C-0218 [Accessed 13th December 2014]

Imielinski, M., Berger, Alice H., Hammerman, Peter S., et al. (2012) 'Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing', Cell, 150(6), 1107-1120.

Maemondo, M., Inoue, A., Kobayashi, K., Sugawara, S., Oizumi, S., Isobe, H., Gemma, A., Harada, M., Yoshizawa, H., Kinoshita, I., Fujita, Y., Okinaga, S., Hirano, H., Yoshimori, K., Harada, T., Ogura, T., Ando, M., Miyazawa, H., Tanaka, T., Saijo, Y., Hagiwara, K., Morita, S. and Nukiwa, T. (2010) 'Gefitinib or Chemotherapy for Non–Small- Cell Lung Cancer with Mutated EGFR', The New England Journal of Medicine, 362(25), 2380-2388.

Nissan, M. H., Pratilas, C. A., Jones, A. M., Ramirez, R., Won, H., Liu, C., Tiwari, S., Kong, L., Hanrahan, A. J., Yao, Z., Merghoub, T., Ribas, A., Chapman, P. B., Yaeger, R., Taylor, B. S., Schultz, N., Berger, M. F., Rosen, N. and Solit, D. B. (2014) 'Loss of NF1 in cutaneous melanoma is associated with RAS activation and MEK dependence', Cancer Research, 74(8), 2340-2350.

Rosell, R., Carcereny, E., Moran, T., et al. (2012) 'Erlotinib versus standard chemotherapy as first- line treatment for European patients with advanced EGFR mutation- positive non- small- cell lung cancer (EURTAC): A multicentre, open-label, randomised phase 3 trial', The Lancet Oncology, 13(3), 239-246.

Rossi, A., Gridelli, C., Chiodini, P., Gallo, C., Sun, J.-M., Park, K., Brien, M. E. R., Popat, S., von Plessen, C., Barata, F., Bergman, B., Parente, B., Perrone, F. and Di Maio, M. (2014) 'Six versus fewer planned cycles of first- line platinum- based chemotherapy for non- small- cell lung cancer: A systematic review and meta- analysis of individual patient data', The Lancet Oncology, 15(11), 1254-1262.

Whittaker, S. R., Theurillat, J.-P., Van Allen, E., Wagle, N., Hsiao, J., Cowley, G. S., Schadendorf, D., Root, D. E. and Garraway, L. A. (2013) 'A Genome-Scale RNA Interference Screen Implicates NF1 Loss in Resistance to RAF Inhibition', Cancer Discovery, 3(3), 350-362.

Yap, Y.-S., McPherson, J. R., Ong, C.-K., Rozen, S. G., Teh, B.-T., Lee, A. S. G. and Callen, D. F. (2014) 'The NF1 gene revisited - from bench to bedside', Oncotarget, 5(15), 5873-92.

Appendix 2. Patient Information Sheet

The University Of Sheffield.

Department Of Oncology.

Academic Unit of Clinical Oncology
Head: Professor Penella Woll
Cancer Clinical Trials Centre
Broomcross Building
Weston Park Hospital
Whitham Road, Sheffield S10 2SJ
Telephone: +44 (0) 114 226 5235
Fax: +44 (0) 114 226 5364

INFORMATION SHEET FOR PATIENTS

Next generation sequencing in non-small cell lung cancer

Thank you for reading this.

You are being invited to take part in a research study. Before you decide, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with friends, relatives, your GP and others if you wish. Ask us if there is anything that is not clear, or if you would like more information. Take time to decide whether or not you wish to take part.

1. What is the purpose of the study?
This study is designed to collect together important clinical information, one set of blood samples and any spare tissue samples from patients with non-small cell lung cancer.

Many factors contribute to whether or not an individual develops non-small cell lung cancer. About 1 in 10 people with non-small cell lung cancer have a tumour with a mutation in the NF-1 gene. We are investigating whether this mutation is linked to how the cancer behaves and looking at whether NF-1 could be targeted or monitored as a way to treat lung cancer. We are using a new laboratory test, Next Generation Sequencing for this work.

2. Why have I been chosen?
We are inviting individuals who we know have non-small cell lung cancer to take part in this study.

Version 2.0: 17th May 2015

3. Do I have to take part?
It is up to you to decide whether or not to take part. A member of the research team will talk to you about this study and give you time to think about whether or not you would like to take part. If you do decide to take part you will be given this information sheet to keep and will be asked to sign a consent form. If you decide to take part you are still free to withdraw at any time and without giving a reason. Whether you take part or not will not affect your treatment or the standard of care you receive.

4. What will happen to me if I take part in the study?
If you agree to take part, we will answer any questions and then ask you to sign a consent form and provide one set of blood samples.

At this visit, you will be asked to give a single 50ml blood sample. This should take about 5 minutes. Neither you nor your doctors will receive a result from this blood sample because it is purely for research. The exception to this is if you wish to be informed of any non-tumour related mutations found in the NF-1 gene. This sample will be carefully processed to allow us to study the DNA (substance that makes up genes) and proteins in it.

We will also analyse stored spare biopsy material and pathology specimens. Such specimens will have been taken when you had your biopsy or surgery, for examination by the hospital pathologist, to determine your diagnosis. We would like to use any tissue that is left over from this procedure to extract DNA to look for any genetic changes within the tissue itself.

We will obtain information on your disease and treatment from the medical records, to link in with the work described above. This information will include information about any cancer and the treatment received.

All this information will be stored anonymously (i.e. labelled with a code number and not your name) on a secure computer database.

5. What are the possible disadvantages and risks of taking part?
You will have the discomfort of a blood test but then your involvement in the study will end.
You would be able to continue with any other treatments you require (such as surgery, chemotherapy or radiotherapy) whilst on this study. With the exception of non-tumour NF-1 mutations, individual results from this research study will not be fed back to you, so there are no implications for your treatment.

Version 2.0: 17th May 2015

6. What if an NF-1 mutation is found in my blood sample?
The study is looking for NF-1 mutations in the tumour, using your stored spare biopsy material. If a mutation is found we will need to confirm this is caused by the tumour and is not present in normal non-tumour DNA extracted from your blood sample. Tumour DNA contains a far greater number of mutations than normal non-tumour DNA. New NF-1 mutations are found in 10% of non-small cell lung cancer tumours and are of no clinical significance to the patient or family with regards to Neurofibromatosis. Mutations in the NF-1 gene in non-tumour DNA are linked to the heredity disease Neurofibromatosis. These non-tumour NF-1 mutations can be passed on to offspring. Neurofibromatosis results in a variety of clinical symptoms ranging from, freckles in the arm pit and groin to skeletal abnormalities. Neurofibromatosis is easily diagnosed and it would be extremely unlikely that you carry a mutation in your non-tumour DNA without a previous diagnosis. In the unlikely event that a NF-1 mutation is found in your blood sample you will have an option in the consent form as to whether you wish to be notified or not.

7. What are the possible benefits of taking part?
There is no immediate clinical benefit to you in taking part. This research may lead to the development of a new genetic test or treatment that may benefit those with non-small cell lung cancer at some point in the future.

8. What if new information becomes available?
If any important new information becomes available that may affect your health during the period of the study, the investigators will contact you to tell you about it.

9. What happens when the research study stops?
The biopsy samples that you give us will be returned to the Sheffield Teaching Hospital. The blood samples will be destroyed within 6 – 12 months after the completion of the study. The samples will only be used for research purposes and not for any financial gain. The samples will be treated as a gift from you to us. Additional studies will be planned in the future and may use your samples. Any such studies would have to be reviewed and gain approval from a research ethics committee.

10. Will my taking part in the study be confidential?
Yes. Anything you say will be treated in confidence, no names will be mentioned in any reports of the study and care will be taken so that individuals cannot be identified from details in reports of the results of the study. Everyone who takes part in the study will be assigned a code number, and all of the data relating to each person will be held on a computer database and will only be linked to that code number, and not to people's names or addresses.

Version 2.0: 17th May 2015

11. What will happen to the results of the study?
The results will be presented at scientific conferences and published in scientific journals. You will not be identified in any reports or publications. If you would like a summary of the study results, please contact a member of the study team (see contact details below).

12. Who is organising and funding this study?
The study is being organised by Dr Sarah Danson at Weston Park Hospital, Sheffield, in collaboration with researchers from the Sheffield Children's Hospital and AstraZeneca. The study is funded by the Biotechnology and Biomedical Sciences Research Council (BBSRC), AstraZeneca and the University of Sheffield.

13. Who has reviewed the study?
This research protocol has been reviewed in the hospital where patients will be seen and by the University of Sheffield. It has also been scientifically reviewed by external reviewers, and by Sheffield Research Ethics Committee.

14. What if something goes wrong?
If you are harmed by your participation in this study, there are no *special* compensation arrangements. We will only be taking a blood sample, which carries negligible risk. If you are harmed due to someone's negligence, then you may have grounds for a legal action. If you have *any* cause to complain about *any* aspect of the way in which you have been approached or treated during the course of this study, the normal National Health Service complaints mechanisms are available to you and are not compromised in any way because you have taken part in a research study.

If you have any complaints or concerns please contact in the first instance the project organiser, Dr Sarah Danson at Weston Park Hospital, Whitham Road, Sheffield, S10 2SJ. Tel: 0114 226 5000.

Or contact Dr David Throssell, Medical Director, Sheffield Teaching Hospitals NHS Foundation Trust, 8 Beech Hill Road, Sheffield, S10 2SB. Tel: 0114 271 2178.

15. Contact for further information.
If you need more information or have any questions concerning this study, please contact Dr Sarah Danson, Weston Park Hospital, Sheffield, S10 2SJ. Tel: 0114 226 5000. Email: s.danson@sheffield.ac.uk

Thank you for reading this and for taking an interest in this research study.

Version 2.0: 17th May 2015

Appendix 3. Patient Consent Form

The University Of Sheffield.

Sheffield Teaching Hospitals NHS
NHS Foundation Trust

Department Of Oncology

Academic Unit of Clinical Oncology
Head, Professor Penella J Woll
Cancer Clinical Trials Centre
Weston Park Hospital
Whitham Road, Sheffield S10 2SJ

Tel: 0114 226 5000/5208 Fax: 0114 226 5678

PATIENT CONSENT FORM

Next generation sequencing in non-small cell lung cancer

Name of Researcher:……Identification Number for this study:……

	Please initial box
1. I confirm that I have read and understand the patient information sheet dated 17th May 2015 (Version2.0) for the above study and have had the opportunity to ask questions	
2. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason, without my medical care or legal rights being affected.	
3. I give permission for one set of blood samples to be obtained from me. I understand that DNA (the substance that genes are made of) and proteins will be extracted from the blood sample and analysed to identify genes which may influence the development of non-small cell lung cancer.	
4. I give consent for the storage of clinical information about me on a computer database. I understand that the laboratory and clinical information will be labelled with a code number and that no other personal information will be held with the DNA or clinical information (i.e. it will not have my name and address on it).	
5. I agree to the use of data about my health provided by the Office of National Statistics.	
6. I understand that sections of any of my medical notes may be looked at by responsible individuals from Sheffield Teaching Hospitals or the University of Sheffield or from regulatory authorities where it is relevant to my taking part in research. I give permission for such individuals directly involved in this study to have access to my records.	
7. I give permission for surplus biopsy material and pathology specimens to be used in this study.	

8.I understand that all data (personal and clinical) collected will be treated in accordance with European and national laws for the protection of data.	
9 I agree to the use of my samples and data for future research.	
10. I understand that I will not benefit financially if this research leads to the development of a new treatment or medical test, but that a proportion of any profits may go towards further research in this field.	
11. I do not wish to be informed if NF-1 disease causing mutations are identified in my normal non-tumour DNA taken from my blood sample.	
12. I agree to take part in the above study.	

Name of Patient_____ Date_____ Signature_____

Researcher _____ Date_____ Signature_____

Appendix 4. Project Approval Letter

NHS
Health Research Authority
NRES Committee Yorkshire & The Humber - Sheffield
Jarrow Business Centre
Viking Business Park
Rolling Mill Road
Jarrow
Tyne and Wear
NE32 3DT

Telephone: 0191 4283564

03 June 2015

Dr Sarah Danson
Academic Unit of Clinical Oncology
Weston Park Hospital
Whitham Road
Sheffield
S10 2SJ

Dear Dr Danson

Study title:	**Next-Generation Sequencing into practice: Identifying functionally important aberrations in neurofibromin-1 (NF-1) in non-small cell lung cancer (NSCLC)**
REC reference:	**15/YH/0105**
IRAS project ID:	**165821**

Thank you for your letter of 23 May 2015, responding to the Committee's request for further information on the above research and submitting revised documentation.

The further information has been considered on behalf of the Committee by the Chair.

We plan to publish your research summary wording for the above study on the HRA website, together with your contact details. Publication will be no earlier than three months from the date of this favourable opinion letter. The expectation is that this information will be published for all studies that receive an ethical opinion but should you wish to provide a substitute contact point, wish to make a request to defer, or require further information, please contact the REC Manager, Miss Kathryn Murray, nrescommittee.yorkandhumber-sheffield@nhs.net. Under very limited circumstances (e.g. for student research which has received an unfavourable opinion), it may be possible to grant an exemption to the publication of the study.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

Conditions of the favourable opinion

The favourable opinion is subject to the following conditions being met prior to the start of the study.

Management permission or approval must be obtained from each host organisation prior to the start of the study at the site concerned.

Management permission ("R&D approval") should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements.

Guidance on applying for NHS permission for research is available in the Integrated Research Application System or at http://www.rdforum.nhs.uk.

Where a NHS organisation's role in the study is limited to identifying and referring potential participants to research sites ("participant identification centre"), guidance should be sought from the R&D office on the information it requires to give permission for this activity.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of approvals from host organisations

Registration of Clinical Trials

All clinical trials (defined as the first four categories on the IRAS filter page) must be registered on a publicly accessible database. This should be before the first participant is recruited but no later than 6 weeks after recruitment of the first participant.

There is no requirement to separately notify the REC but you should do so at the earliest opportunity e.g. when submitting an amendment. We will audit the registration details as part of the annual progress reporting process.

To ensure transparency in research, we strongly recommend that all research is registered but for non-clinical trials this is not currently mandatory.

If a sponsor wishes to request a deferral for study registration within the required timeframe, they should contact hra.studyregistration@nhs.net. The expectation is that all clinical trials will be registered, however, in exceptional circumstances non registration may be permissible with prior agreement from NRES. Guidance on where to register is provided on the HRA website.

It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

Ethical review of research sites

NHS sites

The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion" below).

Approved documents

The final list of documents reviewed and approved by the Committee is as follows:

Document	*Version*	*Date*
Covering letter on headed paper		21 May 2015
Participant consent form	2.0	28 April 2015
Participant information sheet (PIS)	2.0	17 May 2015
REC Application Form [REC_Form_20022015]		20 February 2015
Research protocol or project proposal	2.0	17 May 2015
Summary CV for Chief Investigator (CI)		06 February 2015
Summary CV for student	1	11 February 2015
Summary CV for supervisor (student research)	1	11 February 2015

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

After ethical review

Reporting requirements

The attached document *"After ethical review – guidance for researchers"* gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study

The HRA website also provides guidance on these topics, which is updated in the light of changes in reporting requirements or procedures.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website:
http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/

HRA Training

We are pleased to welcome researchers and R&D staff at our training days – see details at http://www.hra.nhs.uk/hra-training/

15/YH/0105	**Please quote this number on all correspondence**

With the Committee's best wishes for the success of this project.

Yours sincerely

pp.
Professor Basil Sharrack
Chair

Email:nrescommittee.yorkandhumber-sheffield@nhs.net

Enclosures: "After ethical review – guidance for researchers" *[SL-AR2]*

Copy to: *Ms Rachel Simpson, Sheffield Teaching Hospitals*

A Research Ethics Committee established by the Health Research Authority

Appendix 5. NGS Indexes

NGS indexes for the individual samples

Sample ID	Index	Sample ID	Index	Sample ID	Index
NA12878	ATGCCTAA	NF1-159	ACATTGGC	NF1-205	ATAGCGAC
NF1-107	AACGTGAT	NF1-160	CAATGGAA	NF1-206	GAATCTGA
NF1-115	AGATGTAC	NF1-161	TGAAGAGA	NF1-207	ATTGGCTC
NF1-117	CATACCAA	NF1-162	GCGAGTAA	NF1-210	AATGTTGC
NF1-119	CCACACGA	NF1-163	ACGTATCA	NF1-211	AACGCTTA
NF1-120	ACGCTCGA	NF1-164	AGATCGCA	NF1-212	GTACGCAA
NF1-122	TGGAACAA	NF1-165	ACAAGCTA	NF1-213	GAACAGGC
NF1-126	TAGGATGA	NF1-166	TATCAGCA	NF1-214	CCGAAGTA
NF1-130	ACTATGCA	NF1-167	AGCCATGC	NF1-215	CTCAATGA
NF1-132	GGAGAACA	NF1-168	CGGATTGC	NF1-216	CAAGGAGC
NF1-134	TGGCTTCA	NF1-169	GATAGACA	NF1-217	ACAGCAGA
NF1-136	ACCTCCAA	NF1-170	CGACACAC	NF1-218	TAGAGTCG
NF1-138	AAGAGATC	NF1-171	ACACAGAA	NF1-220	GTGCGACA
NF1-139	CGACTGGA	NF1-172	GCTCGGTA	NF1-221	AAGACGGA
NF1-140	AGTGGTCA	NF1-173	TTCACGCA	NF1-222	AACTCACC
NF1-141	CCTAATCC	NF1-174	ACACGACC	NF1-223	CAAGACTA
NF1-142	GGTGCGAA	NF1-175	GTCTGTCA	NF1-224	GTATCAAC
NF1-144	CCAGTTCA	NF1-176	CAGCGTTA	NF1-225	CCTCCTGA
NF1-145	CTAAGGTC	NF1-177	CCATCCTC	NF1-227	GAGCTGAA
NF1-146	CATCAAGT	NF1-178	ATTGAGGA	NF1-231	TCTTCACA
NF1-147	CACCTTAC	NF1-179	CAACCACA	NF1-232	TCGACACT
NF1-148	CCGTGAGA	NF1-180	CGAACTTA	NF1-233	GATCCGCT
NF1-150	ATCCTGTA	NF1-181	AGCAGGAA	NF1-234	AGGCTAAC
NF1-151	AGTCACTA	NF1-182	AAACATCG	NF1-237	ATCATTCC
NF1-152	GTCGTAGA	NF1-183	ACCACTGT	NF1-238	GCCACATA
NF1-154	TGGTGGTA	NF1-184	GACTAGTA	NF1-239	AAGGTACA
NF1-155	GAGTTAGC	NF1-185	GTGTTCTA	NF1-240	GCTAACGA
NF1-156	CTGAGCCA	NF1-201	CAGATCTG		
NF1-157	CGCATACA	NF1-203	AGTACAAG		
NF1-158	AGAGTCAA	NF1-204	AAGGACAC		

Appendix 6. Gene Expression Signatures and Probes

RAS Gene Expression Signature, Down Arm (Loboda *et al*. 2010)

Gene	Probe
ABCC5	CTTCCAAACACCAGCACCCAGTGGACAATGCTGGGCTTTTTTCCTGTATGACTTTTTCGTGG CTTTCTTCTCTGGCCCGTGTGGCCCACAAGAAGGGGGA
ACAP2	TGCTTTGATAACAAAGACTCCAAATATTCTGGAGAACCTGGATAAAAGTTTGAAGGGCTA GATTGGGATTTGAAGACAAAATTGTAGGAAATCTTACATT
ARMC8	TGAAAACTGAATGTGCAGTGGTGTTGGGAAGTCTTGCTATGGGTACTGAAAACAATGTCA AGTCTCTACTGGACTGCCATATTATCCCTGCCTTATTGCA
ATPAF1	GACAGCCAAGGACCTACAAGACAGTTGACTTGATTTTGCACAGTGTAACAGCGCAGTTGC ATTCTGGCCACTTTGACCTTATAGCTCCCAAATGATGAGT
AUTS2	TGCAGGGACAGTGACAGTGAAAGTGCCAGTGGAGAATCCAAGGGCTTCCACCGGAGCAG CTCTCGGGAAAGGCTCAGTGATAGTTCAGCTCCTTCCAGCT
CCSAP	TATAAGAGATAATGGTGCTCTGCCTATGCCAGCTTGGCACCCGAAGATGTGTGAGTGGAC GTGAGGCTGAGTATTACCTTAGTATTTTTCTCTGGGTCTT
CEP57L1	AAAGTTTGGAGCACTGCCTTTTGTGGCTGAAAAGATGAGGCAACATCGTGACCCACATAT CCTTCAGAAACCTTTTAACGTGACTGAGACTAGATGTCTC
CELSR2	ATCTCAGCCACTGATCGTGATTCTGGACTTAATGGCAGGGTCTTCTACACCTTCCAAGGAG GCGACGATGGAGACGGTGACTTTATTGTTGAGTCCACGT
COQ7	TCGGGCCAGTCATTCAGAAAATGTGGGATCAAGAAAAGGACCATTTGAAAAAGTTCAATG AGTTGATGGTTACGTTCAGGGTCCGGCCAACAGTTCTGAT
DRD4	AGGAGGCGGCGTGCCAAGATCACCGGCCGGGAGCGCAAGGCCATGAGGGTCCTGCCGG TGGTGGTCGGGGCCTTCCTGCTGTGCTGGACGCCCTTCTTCG
ENAH	AACAGACTTTCCAGAGCTCGATGAAGTGAAGAAGCCCTTCTTTCAATTCCTGGGCAGAATT CAGTGGTTTTCATATCATCTGAACCTGCTGTGCTGCTGT
HNRNPU	GTGGTCCTCATGAATTGCATTAAGACTCTGTACTGCTCATATTACACTCCATCCTCTCTGTA GTTTGCTGGGTAGTGGAGGGGGTAAGCTAAATCATAGT
HTATSF1	CTTTGATAGGCACCCAGATGGTGTGGCCTCTGTGTCCTTTCGGGATCCAGAGGAAGCTGA TTATTGTATTCAGACTCTCGATGGAAGATGGTTTGGTGGC
ID4	GCTTGCTACCAAAGGACAAACTCTTGGAAATGAACACTTTCTGCTTTCCTTCCTCCAAAGAA TTAATAGGCAACAGTGGGAGAAAAAAAAGGCATAATGG
ITSN1	TAAATTACAAAAGGCACAGTCATTTGATGTGGCCAGTGTCCCACCAGTGGCAGAGTGGGC TGTTCCTCAGTCATCAAGACTGAAATACAGGCAATTATTC
KDM4C	ATTCTCCACCCAATGCCTTCCTTGAAGAGGATGGAACAAGTCTCCTTATTTCCTGTGCAAAG TGCTGCGTACGGGTTCATGCAAGTTGTTATGGTATTCC
GREB1L	CACTTAAAGTACTACCTAGTCAGAAGCTCCCAGGGTGTACTGTCTAAAGGACCTCTTATCT GCTGGAAAGAATGTAGAAGCCGACAATCCTCTGCTTCTT
MIB1	GAACATCTTGGACATACAATCCAGCAGCAGTTTCCAAGGTGGCATCTGCAGGATCAGCCA TTAGCAATGCATCTGGTGAAAGACTCTCACAACTCCTGAA
MRPS14	GGGATAGCTGTCCTGTTAGAATCAGAAATCGGTGTGTTATGACGTCCCGTCCGCGTGGTG TGAAGCGGCGCTGGAGGCTTAGTCGTATAGTCTTCCGTCA
MSI1	AAACCCCACCTTTGGAGAGTTAATTGTCTGTGTGAGGTGCTTAACCTATCAGCCCTGAGAA CACAAAGCAATAATCTTTGTTACTGAGATGCGCGGCTGT
MSI2	AAACGCTCCAGAGGCTTCGGTTTCGTCACGTTCGCAGACCCAGCAAGTGTAGATAAAGTA TTAGGTCAGCCCCACCATGAGTTAGATTCCAAGACGATTG
NUP133	AGGGATGCACCTATGGATTCCATTGAATGGGCTGAAGTGGTGATCAATGTGAACAATATT CTCAAGGATATGCTGCAGGCTGCTAGTCATTATCGCCAAA

RAS Gene Expression Signature, Down Arm (Loboda *et al*. 2010)

Gene	Probe
OGN	TCTGCAAGGCTAATGACACCAGTTACATCCGGGACCGCATTGAAGAGATACGCCTGGAGG GCAATCCAATCGTCCTGGGAAAGCATCCAAACAGTTTTAT
PARP1	AAGGTTTGGGCAAAACTACCCCTGATCCTTCAGCTAACATTAGTCTGGATGGTGTAGACGT TCCTCTTGGGACCGGGATTTCATCTGGTGTGAATGACAC
PIAS1	AGTTAAGTGCAGGAGGCAGTACTTCTCTGCCAACCACCAATGGAAGCAGTAGTGGCAGTA ACAGCAGCCTGGTTTCTTCCAACAGCCTAAGGGAAAGCCA
RASL10B	GAGTCTGGGGGAAGGGTCGTGGGTGGGGAATTTATCACCAACATCCATTGTAGGGGGAA TCTATGATTCTGCTTCCCCAGCGGATTCCCACTCTGTCCAC
RFPL3S	CCCCTTCTCTGTTACCAAGGTGACCCCAAGGAACACAGTAAATGTGGCGGCTTATTTGGCC TCCCCAGGACGGACTGGAGCATCAGTAGTGCCTGAGTTC
RTN3	CCTTCAGGTTCTCACTCATAGTCATGATCCTTCAGAGGGAATATGCACTGGCGAGTTTAAA GTAAGGGCTATGATATTTGATGGTCCCAAAGTACGGCAG
SEC63	CGGAACCGTGAAGAAAGGGAGTTTCGTGCTCCAACTTTGGCATCCCTAGAAAACTGCATG AAGCTTTCTCAGATGGCCGTTCAGGGACTTCAGCAATTTA
SUGP1	CTGAACAATCCCAGACGGCCTTACTACTGAGTGTTCTGGAAATACATACTTTCTGAATGAC CAACCGTCCCTGGACTGTGGAATGTTCCGGCCTGCATTT
SH3GL2	TGCAGCAAGTCACGGTCAGACTGGAAGAAAGAATAAGACAGGCTTCATCTCAGCCTAGAA GGGAATATCAACCTAAACCACGAATGAGCCTGGAGTTTCC
SMAD9	TGCAATTGGAAAAGCAGGACTTTGGTGCCTGTGCTATAAGCAGCAGATTTTGTGGGAGGA AACACTTGAGAGGCGATATTGTCAACAGTATTTGAAGGGT
STARD7	TGAAATTCGGTATCTCACTGAGCTAATCTGGAACAAACCTCTCACCTCAGGCCAGAAGGG GATGACCTCCATTTGCTTCTCTGAGTAGTTTCCTCTGCTG
TBC1D24	CCTCCAGGAGCAGGGAAGTCATATCCGTGGACCATTTGTAGGTTTCCCTCTGCAAACTTCC TGAACTGGGTGGTGATTTTTCTGGCGAGTTAAGTCCAGC
MSANTD3	GTGGCTCTCTTGTAATTAATCTGTGTTGGCAAAGAATGTCTGGAACATGGACTTGGCGGTC AGTAACCTGTAACAGAGCTACAACTAGGAAAATTAGAGT
TTC28	CAACTTGGGCCTAGCATTCAAGGCTCTGCTGAATTTCAGTAAAGCTGAAGAGTGTCAGAA GTACCTACTGTCCCTAGCCCAGTCTCTGAATAATTCCCAG
ERP44	TGAGGAAGTATTACTTAGCATTCATGCATATTGGGCTTAGGCTCTAGCCCTGCCACTATCA TTGTCTTCTCTGGACTGTGAAGTCACTGAGGACAAGGAA
ZFP36	CTGGAACCTCTCCTGAGGGGGAATCCTGGTGCTCAAATTACCCTCCAAAAGCAAGTAGCC AAAGCCGTTGCCAAACCCCACCCATAAATCAATGGGCCCT
ZNF292	GTTCCACCGCAGTCCGACCTAAGTAATTCATTAGGAACTCCATCAGTGCCTCCAAAAGCTC CAGTTCAGAAATTCAGCTGCCAGGTCGAGGGATGTACTC
ZNF441	TGGGAAAGCAGTGAGTGTACAGACGTCCTCATGGGTCGTTCATCTCTTAATTGCTACGTTA GAGTTGACAGTGAACACAAACCATGTGAGTATCAAGAAT
ZNF493	CATGTGGCAAAGCTTTTAGGCGGTCTTCACATCTTAGTAGACATAAGATAATTCATATTGG AATTCATACTGAAGAGACTGTACAAAAGTGAAGAATGTG
ZNF669	GAATGTAAGAAATGCGGTAAAGCCTACACTCGTTCCAGTCACCTTACTCGCCATGAAAGA AGTCATGATATAGAGGCTGGGTGTAGTGACTCAGCCTATA

RAS Gene Expression Signature, Up Arm (Loboda *et al*. 2010)

Gene	Probe
ADAM8	TTAGGAATCGCCTTTATGGAAAGGGCTATGTGGGAGAGTCAGCTATCTTGTCTGGTTTTCT TGAGACCTCAGATGTGTGTTCAGCAGGGCTGAAAGCTTT
ADRB2	CGCAGGTCTTCTTTGAAGGCCTATGGGAATGGCTACTCCAGCAACGGCAACACAGGGGAG CAGAGTGGATATCACGTGGAACAGGAGAAAGAAAATAAAC
ANGPTL4	CACCTGCAGCCATTCCAACCTCAACGGCCAGTACTTCCGCTCCATCCCACAGCAGCGGCAG AAGCTTAAGAAGGGAATCTTCTGGAAGACCTGGCGGGGC
ARNTL2	TCCGTAAAATCTGGGGCACAGGTACAAGAAATAGCCAATATTTAGTTCCCAGACCATGTTT AGTAGTGTCCAGTTTCAGATCATGCTGCCAAGAGGTATC
MYDGF	GTCCCTGAAAGGGCCAGCACATCACTGGTTTTCTAGGAGGGACTCTTAAGTTTTCTACCTG GGCTGACGTTGCCTTGTCCGGAGGGGCTTGCAGGGTGGC
FERMT1	CCCAGAAGCACTTGCGGATATGTACCAGCCTCGGTCTCTGGTTGATAAAGCCAAGCTCAAT GCAGGTTGGCTAGACTCCTCACGCTCCCTTATGGAACAA
CALM2	TGGAGTTGTAACTCTGCGTGGACTATGGACAGTCAACAATATGTACTTAAAAGTTGCACTA TTGCAAAACGGGTGTATTATCCAGGTACTCGTACACTAT
CALU	CAAGTGCTGTGAATGGTGCCACTTAGGAGTAGAAACTGATGATTGTTGAAATGGCCTTTG GGTGGTCTGAAATAGACCAGTGGAAACAGTAGCTTTGTAG
CAPZA1	ACTTCTGGTACTTCAGTCGTTTTTACTGATCCACCAACACCTAAAGAGGCTATGCTACAGTC TCTAGCTAAATGGAAGACACATTCATCCTTCTCCCTCT
CCL20	ATCTGTTCTTTGAGCTAAAAACCATGTGCTGTACCAAGAGTTTGCTCCTGGCTGCTTTGATG TCAGTGCTGCTACTCCACCTCTGCGGCGAATCAGAAGC
CD274	GGATACTTCTGAACAAGGAGCCTCCAAGCAAATCATCCATTGCTCATCCTAGGAAGACGG GTTGAGAATCCCTAATTTGAGGGTCAGTTCCTGCAGAAGT
CDCP1	GCCCTCGACATCGTTGTTGCCTACCCTCAACAGAACTTTCATCTGGGATGTCAAAGCTCATA AGAGCATCGGTTTAGAGCTGCAGTTTTCCATCCCTCGC
CLCF1	ACCGAGCTGGGGAGGAGGTACAGTAGGCCCTGTCCTGTCCTGTTTCTACAGGAAGTCATG CTCGAGGGAGTGTGAAGTGGTTCAGGTTGGTGCAGAGGCG
CSNK1D	GTGGTCTTCAGTCTGTCGTGCACCGATGAGAACTCTCCTTATTGCTGTGAAGGGCAGACAA TGCATGGCTGATCTACTCTGTTACCAATGGCTTTACTAG
CXCL1	TATGTTAATATTTCTGAGGAGCCTGCAACATGCCAGCCACTGTGATAGAGGCTGGCGGAT CCAAGCAAATGGCCAATGAGATCATTGTGAAGGCAGGGGA
CXCL2	ATCACATGTCAGCCACTGTGATAGAGGCTGAGGAATCCAAGAAAATGGCCAGTGAGATCA ATGTGACGGCAGGGAAATGTATGTGTGTCTATTTTGTAAC
CXCL3	TCCCTGCCCTTACCAGAGCTGAAAATGAAAAAGAGAACAGCAGCTTTCTAGGGACAGCTG GAAAGGACTTAATGTGTTTGACTATTTCTTACGAGGGTTC
CXCL5	AGAGAGCTGCGTTGCGTTTGTTTACAGACCACGCAAGGAGTTCATCCCAAAATGATCAGT AATCTGCAAGTGTTCGCCATAGGCCCACAGTGCTCCAAGG
DENND2C	CTACCCATTGAAGAGGTGCTGATAGTTGATCTCTGTGCAGACAAGTTCTTACAGGAGGTAT CTGATGAGGATGAAATTCTACCACCAAAACTTCAAGCTG
DUSP1	TCAAGAATGCTGGAGGAAGGGTGTTTGTCCACTGCCAGGCAGGCATTTCCCGGTCAGCCA CCATCTGCCTTGCTTACCTTATGAGGACTAATCGAGTCAA
DUSP5	GTGGATGTAAAACCCATTTCACAAGAGAAGATTGAGAGTGAGAGAGCCCTCATCAGCCAG TGTGGAAAACCAGTGGTAAATGTCAGCTACAGGCCAGCTT
DUSP6	ATGTGACAACAGGGTTCCAGCACAGCAGCTGTATTTTACCACCCCTTCCAACCAGAATGTA TACCAGGTGGACTCTCTGCAATCTACGTGAAAGACCCCA
EFNB1	CGAGGCTTCGGGGGCGCAAACTAATGGGACTGGCTCGCTCGGCAGCATCTCCCCGCTCTT CTAAGTACACTGAGCAGGGCCCGCGCTGAAGTAGAAGCTG
EGR1	GAGGCATACCAAGATCCACTTGCGGCAGAAGGACAAGAAAGCAGACAAAAGTGTTGTGG CCTCTTCGGCCACCTCCTCTCTCTCTTCCTACCCGTCCCCG

RAS Gene Expression Signature, Up Arm (Loboda *et al*. 2010)

Gene	Probe
EHD1	TACCTTCCTTCCTCCTCTGTTTAGCAAAGGAGGGCAGCTCACTTGGATGTCCTTACAACGCC CCTGGCCCCCAGGTTGAGCAATAAGAAACCAGAACCTT
ELK3	GGCGTCCGCCTTCCTGGCCTCGTCCGTCTCGGCCAAGATCTCCTCTTTAATGTTGCCAAACG CTGCCAGTATTTCATCCGCCTCACCCTTCTCATCTCGG
EREG	GAGAGTCCAGTGATAACTGCACAGCTTTAGTTCAGACAGAAGACAATCCACGTGTGGCTC AAGTGTCAATAACAAAGTGTAGCTCTGACATGAATGGCTA
FOS	ACTCAAGTCCTTACCTCTTCCGGAGATGTAGCAAAACGCATGGAGTGTGTATTGTTCCCAG TGACACTTCAGAGAGCTGGTAGTTAGTAGCATGTTGAGC
FOXQ1	ACAGCTTTATTAGTGGTTCTCTAACTGTGGTCTCCTTGGGCCAAGCAATTTCTTTAAAGGAA AAGTTGATTATGTATGTGGGGTGCCAGGACCACTGCCT
G0S2	TTTGGACTTAACTTCAGAGAAACCGCTGACATCTAGAACTGACCTACCACAAGCATCCACC AAAGGAGTTTGGGATTGAGTTTTGCTGCTGTGCAGCACT
GDF15	ACTCCAGATTCCGAGAGTTGCGGAAACGCTACGAGGACCTGCTAACCAGGCTGCGGGCCA ACCAGAGCTGGGAAGATTCGAACACCGACCTCGTCCCGGC
GLTP	TAAGATCCAAGAAAACGCCTCACTGCCTTAACCTTAACTGTTCTTCCTGGCGCTAAAAAGA GCTGTATTTTTTAAAGTGCTGGGGCAAACAAAGCAACCC
HBEGF	TGAGAGTCACTTTATCCTCCAAGCCACAAGCACTGGCCACACCAAACAAGGAGGAGCACG GGAAAAGAAAGAAGAAAGGCAAGGGGCTAGGGAAGAAGAG
IER3	TCAACTCCGTCTGTCTACTGTGTGAGACTTCGGCGGACCATTAGGAATGAGATCCGTGAG ATCCTTCCATCTTCTTGAAGTCGCCTTTAGGGTGGCTACG
IL13RA2	AGATGGGTTTGATCTTAACAAGGGCATTGAAGCGAAGATACACACGCTTTTACCATGGCA ATGCACAAATGGATCAGAAGTTCAAAGTTCCTGGGCAGAA
IL1A	ACTCCATGAAGGCTGCATGGATCAATCTGTGTCTCTGAGTATCTCTGAAACCTCTAAAACA TCCAAGCTTACCTTCAAGGAGAGCATGGTGGTAGTAGCA
IL1B	GGGACCAAAGGCGGCCAGGATATAACTGACTTCACCATGCAATTTGTGTCTTCCTAAAGA GAGCTGTACCCAGAGAGTCCTGTGCTGAATGTGGACTCAA
CXCL8	ACAGCAGAGCACACAAGCTTCTAGGACAAGAGCCAGGAAGAAACCACCGGAAGGAACCA TCTCACTGTGTGTAAACATGACTTCCAAGCTGGCCGTGGCT
ITGA2	CAACGGGTGTGTGTTCTGACATCAGTCCTGATTTTCAGCTCTCAGCCAGCTTCTCACCTGCA ACTCAGCCCTGCCCTTCCCTCATAGATGTTGTGGTTGT
ITPR3	TCCTGGAAATCCTTCAGTTCATCCTCAATGTCCGCCTGGATTACCGCATATCCTACCTGCTG TCTGTCTTCAAGAAGGAGTTTGTGGAGGTGTTTCCCAT
KCNK1	GTGCTCCTTGGGTTTGTCACTGTGTCCTGCTTCTTCTTCATCCCGGCCGCTGTCTTCTCAGTC CTGGAGGATGACTGGAACTTCCTGGAATCCTTTTATT
KCNN4	GCGTCCTGCTCAACGCTTCCTACCGCAGCATCGGCGCTCTCAATCAAGTCCGCTTCCGCCA CTGGTTCGTGGCCAAGCTTTACATGAACACGCACCCTGG
KLF5	AACGTCTTCCTCCCTGACATCACTCACCTGAGAACTGGCCTCTACAAATCCCAGAGACCGT GCGTAACACACATCAAGACAGAACCTGTTGCCATTTTCA
KLF6	CGGCGCCTAAGCCTTTGCCGTGAGCATGCACACTGAGAATGCTAATGGTTGGGTTGATTG TATGTTGAGGATCTATTACTGACCGTATGATGAGGCCAAC
LAMA3	GAGGACTGGTGTTTCACACGGGCACTAAGAACTCCTTTATGGCTCTTTATCTTTCAAAAGG ACGTCTGGTCTTTGCACTGGGGACAGATGGGAAAAAATT
LDLR	TTTCTGAAATCGCCGTGTTACTGTTGCACTGATGTCCGGAGAGACAGTGACAGCCTCCGTC AGACTCCCGCGTGAAGATGTCACAAGGGATTGGCAATTG
LHFPL2	AGTCAAGTGGAGAACTAACCTTTACCTACCAAAGCCACGTTCCACGGCCCGAGGCTTAAA CAGGACCAATGAGAGGCCACATCCAGCTACGCAAAGTTAC

RAS Gene Expression Signature, Up Arm (Loboda *et al*. 2010)

Gene	Probe
LIF	GGGATGGAAGGCTGTCTTCTTTTGAGGATGATCAGAGAACTTGGGCATAGGAACAATCTG GCAGAAGTTTCCAGAAGGAGGTCACTTGGCATTCAGGCTC
MALL	ACAAGTACATGCCACGATTGTTTCTGAGAAACTGCTGGACCCAAGAATTTACTACATTAAT TCGGCAGCCTCGTTCTTCGCCTTCATCGCCACGCTGCTC
MAP1LC3B	AGAGCCGTACGCTCTTTACAGATACTAATGTCAAGAGTTAAACCTCCTCAGGTTCAACCTG TGATAAAAGACTAGTGCTTCCCAGTACTTGCATGGGGTT
MAST4	AATTGCTTGGTCAAACGCCCTGTGTGTCCAAATGCTGGGAGAACATCACCCCTTGGATGAA TTGCCACCACATTAAATAAAACATATCCAAAGCTCAAAA
MAST4	CACATTTTGTAGTAGTCCTGCACCAGCCCTGCTCTTGAATAAAAAGGAAAATTACTGGCTG CACCCAAATCTTCTAGTACTTGAGTAATAAGCATCAGGC
MMP14	GACAAGATTGATGCTGCTCTCTTCTGGATGCCCAATGGAAAGACCTACTTCTTCCGTGGAA ACAAGTACTACCGTTTCAACGAAGAGCTCAGGGCAGTGG
MXD1	GAGAATAAAGCTGCAGGACAGTCACAAGGCGTGTCTTGGTCTCTAAGAGAGTGGGCACT GCGGCTGTCTCCTTGAAGGTTCTCCCTGTTGGTTCTGATTA
NAV3	CCAGCACTTCTTCTCTTTACTCTACAGCTGAAGAAAAGGCTCATTCAGAGCAAATCCATAA ACTGCGGAGAGAGCTGGTTGCATCACAAGAAAAAGTTGC
NDRG1	CGCCTACATCCTAACTCGATTTGCTCTAAACAACCCTGAGATGGTGGAGGGCCTTGTCCTT ATCAACGTGAACCCTTGTGCGGAAGGCTGGATGGACTGG
NFKBIZ	ATGTTGCTGCCAGCTTGCAGTATCGGTTGACACAATTAGATGCTGTCCGCCTGTTGATGAG GAAGGGAGCAGACCCAAGTACTCGGAACTTGGAGAACGA
NIPAL1	CCCAGGCTTGGTGTGAGATCACAAATGTGTCACAGCTGCTGGCTTCTCCTGTGCTCTACAC GGACCTGAATTACAGCATAAACAACTTGAGCATTTCAGC
NT5E	ATTCGGGTTTTGAAATGGATAAACTCATCGCTCAGAAAGTGAGGGGTGTGGACGTCGTGG TGGGAGGACACTCCAACACATTTCTTTACACAGGCAATCC
OXSR1	TCCTCCCATGAAAATCAGTTCGCCTGGCCTCCAAGTCGTGAGGAAATGGGTATGCAAGGC TGAGATTTCTACAGCAATAAAGGAGACACACACTGGGCCA
NAMPT	CCTGTCCTCCGGCCCGAGATGAATCCTGCGGCAGAAGCCGAGTTCAACATCCTCCTGGCCA CCGACTCCTACAAGGTTACTCACTATAAACAATATCCAC
PHLDA1	ACCAGGACAGATGCTACTTGGGTTTAAATGGAGCCATAGATGATACAAAGTCCTCTTGGG GCTGAAAATCACTTCCTATTTGCATGGCTTTACTAACTGG
PHLDA2	TGGAACGCGGCCATCGCGCTGGCGCTCATCGATTTCCAGAACCGCCGCGCCCTGCAGGAC TTTCGCAGCCGCCAGGAACGCACCGCACCCGCCGCACCCG
PI3	CCATGTTGAATCCCCCTAACCGCTGCTTGAAAGATACTGACTGCCCAGGAATCAAGAAGTG CTGTGAAGGCTCTTGCGGGATGGCCTGTTTCGTTCCCCA
PIK3CD	TGACACTCATTGATTCTAAAGCATCTTTAATCTGCCAGGCGGAGGGGGCTTTGCTGGTCTT TCTTGGACTATTCCAGAGAGGACAACTGTCATCTGGGAA
PIM1	CTTCATCATGAGTTCTGCTGAATGCCGCGATGGGTCAGGTAGGGGGGAAACAGGTTGGG ATGGGATAGGACTAGCACCATTTTAAGTCCCTGTCACCTCT
PLAUR	GAGAAGACCAACAGGACCCTGAGCTATCGGACTGGCTTGAAGATCACCAGCCTTACCGAG GTTGTGTGTGGGTTAGACTTGTGCAACCAGGGCAACTCTG
PNMA2	CCTGAGTTCTTGCCACTGAGTAGGCCAGGGTCATTTGTCCAGAAAACTTTGTGACTGTCTT TGAGTGACCTAGTCTGGGACCCATTCATTGGTGGGTTCT
PPP1R15A	CTGGGGCTGAAAACCAGCAGTTCCCTTCCTGAAGCCTGGGGACTTTTGGATGATGATGAT GGCATGTATGGTGAGCGAGAGGCAACCAGTGTCCCTAGAG
PRNP	GGGAGGCGGTATCCACCTGCAGCCCTTTTAGTGGTGGTGTCTCACTCTTTCTTCTCTCTTTG TCCCGGATAGGCTAATCAATACCCTTGGCACTGATGGG

RAS Gene Expression Signature, Up Arm (Loboda *et al*. 2010)

Gene	Probe
PTGS2	GCTACAAAAGCTGGGAAGCCTTCTCTAACCTCTCCTATTATACTAGAGCCCTTCCTCCTGTG CCTGATGATTGCCCGACTCCCTTGGGTGTCAAAGGTAA
PTHLH	TTGGGTCTGATGATGAGGGCAGATACCTAACTCAGGAAACTAACAAGGTGGAGACGTAC AAAGAGCAGCCGCTCAAGACACCTGGGAAGAAAAAGAAAGG
PTPRE	CCTTTGTTTGACTGAAGGAGAAGAGAGGGAAGTAAGCTTCTGTTCAGCACCTGAACCCTA GAAAAAGAGCCAGTTTGCTACGATGAAGGTGACATTTCTC
PTX3	ATATCTGGGATAGTGTTCTTAGCAATGAAGAGATAAGAGAGACCGGAGGAGCAGAGTCT TGTCACATCCGGGGGAATATTGTTGGGTGGGGAGTCACAGA
PVR	GCTGCGGAATGCCTCGCTGAGGATGTTCGGGTTGCGCGTAGAGGATGAAGGCAACTACA CCTGCCTGTTCGTCACGTTCCCGCAGGGCAGCAGGAGCGTG
MAP7D1	CACCTGTAGTATTTGCCTTGATTTGGTGGGGTACAGTGGATGTGAATACTGTAAATAGCTT GTGCTCAGACTCCTCTGCGTGGAGAGGGTGGGTGCAGGA
S100A6	TTCCTGGGGGCCTTGGCTTTGATCTACAATGAAGCCCTCAAGGGCTGAAAATAAATAGGG AAGATGGAGACACCCTCTGGGGGTCCTCTCTGAGTCAAAT
SDC1	TGAAATTCTCCTGGAGGTCGGTAGGTTCAGCCAAGGTTTTATAAGGCTGATGTCAATTTCT GTGTTGCCAAGCTCCAAGCCCCATCTTCTAAATGGCAAA
SDC4	ACTGTTCATTCCTTTGTGCAGAGTGTATATCTCTGCCTGGGCAAGAGTGTGGAGGTGCCGA GGTGTCTTCATTCTCTCGCACATTTCCACAGCACCTGCT
SEMA4B	CGCTCTTTGTGCTGGCCGTGCTGCTCCCAGTTTTATTCTTGCTCTACCGGCACCGGAACAGC ATGAAAGTCTTCCTGAAGCAGGGGGAATGTGCCAGCGT
SERPINB1	TTACACTCTCAACTCCGACCTCGCCCGCCTAGGTGTGCAGGATCTCTTTAACAGTAGCAAG GCTGATCTGTCTGGCATGTCAGGAGCCAGAGATATTTTT
SERPINB2	CTGTGGGTTCATGCAGCAGATCCAGAAGGGTAGTTATCCTGATGCGATTTTGCAGGCACA AGCTGCAGATAAAATCCATTCATCCTTCCGCTCTCTCAGC
SERPINB5	GCCTGTTCCTTTTCCACGCATTTTCCAGGATAACTGTGACTCCAGGCCCGCAATGGATGCC CTGCAACTAGCAAATTCGGCTTTTGCCGTTGATCTGTTC
SESN2	GCTTGTGTGTGATGTGCAGTCCCGAAGCCACACCCTCCCTTTTCCTCACTGGAATGGACAG TTCATTGCACTGACTCTGGGATCTCAGCCCTGCTCCTGG
SFN	TGGGCCTGGCCCTGAACTTTTCCGTCTTCCACTACGAGATCGCCAACAGCCCCGAGGAGGC CATCTCTCTGGCCAAGACCACTTTCGACGAGGCCATGGC
SLC16A3	GGCCCAGCGGATCGTCGCCCGATCAGTGTTTTGAGGGGGAAGGTGGCGGGGTGGGAAC CGTGTCATTCCAGAGTGGATCTGCGGTGAAGCCAAGCCGCAA
SLC2A14	GAGGAAGCAGTACCTTGAAGAGAAATTGGAGAGGGAGTCAATTCCTAGGATAGCAGAGA GATGGACAACAGACAGAATAGATGGAGTTTCACAATGGTGG
SLC2A3	GGCTGGGGGCTTGTCGCCCTTTCAGGCTCCACCCTTTGCGGAGATTATAAATAGTCATGAT CCCAGCGAGACCCAGAGATGCTGTAATGGTAAGACTTTG
SLC9A1	TGGCTGTGAAGAAAAAGCAAGAGACGAAGCGCTCCATCAACGAAGAGATCCACACACAG TTCCTGGACCACCTTCTGACAGGCATCGAAGACATCTGTGG
SPRY4	ATAATGACTTGTCCAAGATGGCACACGTGGAAAGTTGATCTGCACCAGAACCCGGATGAC TGTCACCTTGAAGCATCCTGTTCTCCTTCTGTGCTGTCCC
TFPI2	GCAACGCCAACAATTTCTACACCTGGGAGGCTTGCGACGATGCTTGCTGGAGGATAGAAA AAGTTCCCAAAGTTTGCCGGCTGCAAGTGAGTGTGGACGA
TGFA	TGCCACAGACCTTCCTACTTGGCCTGTAATCACCTGTGCAGCCTTTTGTGGGCCTTCAAAAC TCTGTCAAGAACTCCGTCTGCTTGGGGTTATTCAGTGT
TIMP1	CGTGGGGACACCAGAAGTCAACCAGACCACCTTATACCAGCGTTATGAGATCAAGATGAC CAAGATGTATAAAGGGTTCCAAGCCTTAGGGGATGCCGCT

RAS Gene Expression Signature, Up Arm (Loboda *et al*. 2010)

Gene	Probe
TMEM45B	CAACAGCACTGAAATGGCTGTAAGGACTCCTGAGATATGTGTCCAGCAAGGAGTTTACA GTCAAACAGGAGAGACATGCCTGTAGTTACATCCAGTGTGA
TNFRSF10A	CACAACGAGATTCTGAGCAACGCAGACTCGCTGTCCACTTTCGTCTCTGAGCAGCAAATG GAAAGCCAGGAGCCGGCAGATTTGACAGGTGTCACTGTAC
TNFRSF10B	GCATCTCCTGCAAATATGGACAGGACTATAGCACTCACTGGAATGACCTCCTTTTCTGCT TGCGCTGCACCAGGTGTGATTCAGGTGAAGTGGAGCTAAG
TNFRSF12A	GCCCAGGGTTCAGGGGAACCTTCCAAGGTGTCTGGTTGCCCTGCCTCTGGCTCCAGAAC AGAAAGGGAGCCTCACGCTGGCTCACACAAAACAGCTGACA
TNS4	GTCTGATGTCAGCTATATGTTTGGAAGCAGCCAGTCCCTCCTGCACTCCAGCAACTCCAG CCATCAGTCATCTTCCAGATCCTTGGAAAGTCCAGCCAAC
TOR1AIP1	CCGGTGGTGGCTACTTCCTCTGATAGCTGCTCTTGCCTCTGGGAGTTTTTGGTTCTTTAGT ACTCCTGAGGTAGAAACCACTGCTGTTCAAGAGTTCCAG
TSC22D1	AAGAATTGTTCGATAGCGTGCATGTGTTATAAAGTGGTGACACGGGCATCCTGTTGAAA TGATGGATGGCTCACTGCCATAGGCTGATAGCAGTTGTCAT
TUBA1B	GAGGAAGGCGAGTTTTCAGAGGCCCGTGAAGATATGGCTGCCCTTGAGAAGGATTATG AGGAGGTTGGTGTGGATTCTGTTGAAGGAGAGGGTGAGGAAG
UAP1	CAAAGTTTCTATGGCTCCAGATGGGAATGGTGGTCTTTATCGGGCACTTGCAGCCCAGA ATATTGTGGAGGATATGGAGCAAAGAGGCATTTGGAGCATT
UPP1	GATACCTGCTTCAAGGCAGAGTTTGAGCAGATTGTCCTGGGGAAGCGGGTCATCCGGA AAACGGACCTTAACAAGAAGCTGGTGCAGGAGCTGTTGCTGT
VEGFA	TATTTGACTGCTGTGGACTTGAGTTGGGAGGGGAATGTTCCCACTCAGATCCTGACAGG GAAGAGGAGGAGATGAGAGACTCTGGCATGATCTTTTTTTT

MEK 18 Gene Expression Signature (Dry *et al*. 2010)

Gene	Probe
DUSP4	GCACCGTAGCATGCAGATGTCAAGGCAGTTAGGAAGTAAATGGTGTCTTGTAGATATGTG CAAGGTAGCATGATGAGCAACTTGAGTTTGTTGCCACTGA
DUSP6	ATGTGACAACAGGGTTCCAGCACAGCAGCTGTATTTTACCACCCCTTCCAACCAGAATGTA TACCAGGTGGACTCTCTGCAATCTACGTGAAAGACCCCA
ELF1	TAAGGGGAATGCTTTATTATGGCTGCTGTTGTCCAACAGAACGACCTAGTATTTGAATTTG CTAGTAACGTCATGGAGGATGAACGACAGCTTGGTGATC
ETV4	CCTGCAGCTGTGGCAATTTCTGGTGGCCTTGCTGGATGACCCAACAAATGCCCATTTCATT GCCTGGACGGGCCGGGGAATGGAGTTCAAGCTCATTGAG
ETV5	CCCGCCTTGTTGCCTGCTTTGTGCTTTCTGCACCAGACAACCTGATGGAACATTTGCACCTG AGTTGTACATTTTTGAAGTGTGCAGGGCAGCCTGGACA
FXYD5	TGAACACACCCTCCGGAAACGGGGGCTGTTGGTCGCAGCTGTGCTGTTCATCACAGGCAT CATCATCCTCACCAGTGGCAAGTGCAGGCAGCTGTCCCGG
KANK1	GAGGAGAACATGAACGACATCGTCGTGTACCACAGAGGCTCCAGGTCCTGTAAGGATGC AGCTGTAGGGACACTTGTTGAGATGAGAAATTGTGGGGTCA
LGALS3	CACGGTGAAGCCCAATGCAAACAGAATTGCTTTAGATTTCCAAAGAGGGAATGATGTTGC CTTCCACTTTAACCCACGCTTCAATGAGAACAACAGGAGA
LZTS1	CCACCAGACGTCAGGCCCTGACTCCTCTGGCTTTCCCAGGAGATGGGTCCAGGGGTCTGT CTGCTTTGGTTAAGGGCTCCCTAAACTTTGGCCTTTGTTC
MAP2K3	GAACCTGGACTCCCGGACCTTCATCACCATTGGAGACAGAAACTTTGAGGTGGAGGCTGA TGACTTGGTGACCATCTCAGAACTGGGCCGTGGAGCCTAT
PHLDA1	ACCAGGACAGATGCTACTTGGGTTTAAATGGAGCCATAGATGATACAAAGTCCTCTTGGG GCTGAAAATCACTTCCTATTTGCATGGCTTTACTAACTGG
PROS1	TGAAGACCTTCAAAGACAACTTGCCGTCTTGGACAAAGCAATGAAAGCAAAAGTGGCCAC ATACCTGGGTGGCCTTCCAGATGTTCCATTCAGTGCCACA
S100A6	TTCCTGGGGGCCTTGGCTTTGATCTACAATGAAGCCCTCAAGGGCTGAAAATAAATAGGG AAGATGGAGACACCCTCTGGGGGTCCTCTCTGAGTCAAAT
SERPINB1	TTACACTCTCAACTCCGACCTCGCCCGCCTAGGTGTGCAGGATCTCTTTAACAGTAGCAAG GCTGATCTGTCTGGCATGTCAGGAGCCAGAGATATTTTT
SLCO4A1	CCTCTACACGCTGGGCGTCACCTACCTGGATGAGAACGTCAAGTCCAGCTGCTCGCCCGTC TACATTGCCATCTTCTACACAGCGGCCATCCTGGGCCCA
SPRY2	ACTGTTGTACACGATGGTCAGCCATGGGTGTCATGTCCCTCTTTTTGCCTTGTTTATGGTGT TACCTTCCAGCCAAGGGTTGCCTTAAATTGTGCCAGGG
TRIB2	AGAGAATGATCCTTTTAAGGCTTGTAAGGCCCTCTGGTTTGGACAAAAACCCTCAGTAGA GACAAGCGGGAAGGATAATTAGCTGAAAGCTATGATGATA
ZNF106	GGTTTAAGAGAAGTGGTCTAATAGACAGGAATAGAAGTTGTGGTGGAGGTGATTTGGGA TAGACTAGGTTTCCTTATGCAAGTGGATATGAGCTCCTCAT
MEK 6 Signature highlighted in red (Brant *et al*. 2017)	

Housekeeping Genes

Gene	Probe
ABCF1	CTAAACAAACAAGAGGTGACCACCTTATTGTGAGGTTCCATCCAGCCAAGTTTATGTGGCC TATTGTCTCAGGACTCTCATCACTCAGAAGCCTGCCTCT
ABCF1	GATGTCCTCCCGCCAAGCCATGTTAGAAAATGCATCTGACATCAAGCTGGAGAAGTTCAG CATCTCCGCTCATGGCAAGGAGCTGTTCGTCAATGCAGAC
ALAS1	GGGGATCGGGATGGAGTCATGCCAAAAATGGACATCATTTCTGGAACACTTGGCAAAGCC TTTGGTTGTGTTGGAGGGTACATCGCCAGCACGAGTTCTC
B2M	CGGGCATTCCTGAAGCTGACAGCATTCGGGCCGAGATGTCTCGCTCCGTGGCCTTAGCTG TGCTCGCGCTACTCTCTCTTTCTGGCCTGGAGGCTATCCA
CLTC	GGGTATCAACCCAGCAAACATTGGCTTCAGTACCCTGACTATGGAGTCTGACAAATTCATC TGCATTAGAGAAAAAGTAGGAGAGCAGGCCCAGGTGGTA
G6PD	ACAACATCGCCTGCGTTATCCTCACCTTCAAGGAGCCCTTTGGCACTGAGGGTCGCGGGG GCTATTTCGATGAATTTGGGATCATCCGGGACGTGATGCA
GAPDH	CACTCCTCCACCTTTGACGCTGGGGCTGGCATTGCCCTCAACGACCACTTTGTCAAGCTCAT TTCCTGGTATGACAACGAATTTGGCTACAGCAACAGGG
GUSB	CGGTCGTGATGTGGTCTGTGGCCAACGAGCCTGCGTCCCACCTAGAATCTGCTGGCTACT ACTTGAAGATGGTGATCGCTCACACCAAATCCTTGGACCC
GUSB	CCGATTTCATGACTGAACAGTCACCGACGAGAGTGCTGGGGAATAAAAAGGGGATCTTCA CTCGGCAGAGACAACCAAAAAGTGCAGCGTTCCTTTTGCG
HPRT1	TGTGATGAAGGAGATGGGAGGCCATCACATTGTAGCCCTCTGTGTGCTCAAGGGGGGCT ATAAATTCTTTGCTGACCTGCTGGATTACATCAAAGCACTG
LDHA	AACTTCCTGGCTCCTTCACTGAACATGCCTAGTCCAACATTTTTTCCCAGTGAGTCACATCC TGGGATCCAGTGTATAAATCCAATATCATGTCTTGTGC
PGK1	GCAAGAAGTATGCTGAGGCTGTCACTCGGGCTAAGCAGATTGTGTGGAATGGTCCTGTG GGGGTATTTGAATGGGAAGCTTTTGCCCGGGGAACCAAAGC
POLR1B	GGAGAACTCGGCCTTAGAATACTTTGGTGAGATGTTAAAGGCTGCTGGCTACAATTTCTAT GGCACCGAGAGGTTATATAGTGGCATCAGTGGGCTAGAA
POLR2A	TTCCAAGAAGCCAAAGACTCCTTCGCTTACTGTCTTCCTGTTGGGCCAGTCCGCTCGAGAT GCTGAGAGAGCCAAGGATATTCTGTGCCGTCTGGAGCAT
TBP	ACAGTGAATCTTGGTTGTAAACTTGACCTAAAGACCATTGCACTTCGTGCCCGAAACGCCG AATATAATCCCAAGCGGTTTGCTGCGGTAATCATGAGGA
TUBB	TTCTAAGTATGTCCATTTCCCATCTCAGCTTCAAGGGAGGTGTCAGCAGTATTATCTCCACT TTCAATCTCCCTCCAAGCTCTACTCTGGAGGAGTCTGT
TUBB	TGGTGGATCTAGAACCTGGGACCATGGACTCTGTTCGCTCAGGTCCTTTTGGCCAGATCTT TAGACCAGACAACTTTGTATTTGGTCAGTCTGGGGCAGG
ACTB	TGCAGAAGGAGATCACTGCCCTGGCACCCAGCACAATGAAGATCAAGATCATTGCTCCTC CTGAGCGCAAGTACTCCGTGTGGATCGGCGGCTCCATCCT
RPL19	CCAATGCCCGAATGCCAGAGAAGGTCACATGGATGAGGAGAATGAGGATTTTGCGCCGG CTGCTCAGAAGATACCGTGAATCTAAGAAGATCGATCGCCA
RPLP0	CGAAATGTTTCATTGTGGGAGCAGACAATGTGGGCTCCAAGCAGATGCAGCAGATCCGCA TGTCCCTTCGCGGGAAGGCTGTGGTGCTGATGGGCAAGAA
SDHA	TGGAGGGGCAGGCTTGCGAGCTGCATTTGGCCTTTCTGAGGCAGGGTTTAATACAGCATG TGTTACCAAGCTGTTTCCTACCAGGTCACACACTGTTGCA

Appendix 7. CMAP PI3K/AKT/mTOR gene expression signature

CMAP upregulated 136 gene signature					
ACLY	GTF2E2	POLE2	MLEC	RPIA	ENOPH1
ATP1B1	HDAC2	PPIC	TATDN2	RRP1B	ATG3
ATP6V0B	HIF1A	RFC5	MRPL19	JMJD6	ZDHHC6
CALU	HMGCR	RPN1	LPGAT1	RRS1	ACD
CASP7	HMGCS1	SPR	PARP2	ARFIP2	TMEM109
CBR1	HMOX1	DYNLT3	G3BP1	RAI14	CCDC86
CCND3	HPRT1	TFDP1	TXNDC9	TES	CHAC1
CSK	HSPA1A	TPD52L2	SLC35B1	PHGDH	CRELD2
CSRP1	HSPA4	TXNRD1	HYOU1	TRAPPC3	CHMP6
DDX10	HSPD1	UGDH	DRAP1	DNTTIP2	NUP85
DFFA	DNAJB1	MAPKAPK3	EBP	GMNN	DHDDS
DUSP3	ICAM3	FOSL1	WASF3	ATP6V1D	GRWD1
FDFT1	ITGAE	BHLHE40	UTP14A	PRKAG2	TUBB6
FHL2	KIF5C	PSMG1	BLCAP	UBE2J1	MICALL1
FKBP4	MAT2A	IER3	TCERG1	GFOD1	H2AFV
GABPB1	MCM3	DNAJA3	MLLT11	KCTD5	RPL39L
GALE	PNP	USP14	CCDC85B	EXOSC4	
GLI2	NRAS	PRPF4	TOPBP1	DDIT4	
GLRX	NUCB2	CCNE2	CORO1A	YTHDF1	
GNA11	NUP88	VAPB	SLC2A6	FKBP14	
GRB10	PAFAH1B1	UBE2L6	NISCH	HEATR1	
MSH6	PMAIP1	ITGB1BP1	ECD	ANO10	
GTF2A2	PMM2	HERPUD1	PUF60	TRIB3	

CMAP downregulated 56 gene signature					
CBLB	LAMA3	DUSP11	HMG20B	HSD17B11	TRAPPC6A
CDKN1B	MYC	RNMT	CRTAP	EVL	ARID5B
CETN3	PIK3CA	INPP4B	ST6GALNAC2	TXNL4B	PSRC1
CSNK2A2	PPOX	VGLL4	KIAA0907	ANKRD10	HIST1H2BK
DDB2	KDM5A	KIAA0355	WDTC1	FAM63A	WIPF2
DUSP6	RTN2	NCAPD2	KIAA1033	KDM3A	SPRED2
ERBB3	TLE1	HDAC6	KLHDC2	EAPP	
GRB7	TOP2A	ARL4C	CHIC2	CABC1	
ID2	TP53	TRIM13	SESN1	SCYL3	
IGF1R	HIST2H2BE	TRIB1	VPS28	POLD4	