Mutation detection in normal mucosa and early lesions of colorectal cancer

by

Kate Marie Sutton

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

The University of Leeds

School of medicine

Leeds Institute of Cancer & Pathology

February 2014

Intellectual property and publication statement

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others. The serial dilution of mutant KRAS data from Chapter 3 has been included in the following publications:

CHAMBERS, P. A., STEAD, L. F., MORGAN, J. E., CARR, I. M., SUTTON, K. M., WATSON, C. M., CROWE, V., DICKINSON, H., ROBERTS, P., MULATERO, C., SEYMOUR, M., MARKHAM, A. F., WARING, P. M., QUIRKE, P. & TAYLOR, G. R. 2013. Mutation Detection by Clonal Sequencing of PCR Amplicons and Grouped Read Typing is Applicable to Clinical Diagnostics. Human mutation, 34, 248-254.

For this paper I prepared NGS libraries. PC contributed to library preparation and analysis and wrote the paper as did GT. LS and GT analysed the data. All other authors contributed to data generation and writing of the manuscript.

STEAD, L. F., SUTTON, K. M., TAYLOR, G. R., QUIRKE, P. & RABBITTS, P. 2013. Accurately Identifying Low-Allelic Fraction Variants in Single Samples with Next-Generation Sequencing: Applications in Tumor Subclone Resolution. Human mutation.

For this paper I prepared NGS libraries. LS and GT analysed the data and LF wrote the manuscript. All other authors contributed to the writing of the manuscript. For this thesis, I reanalysed the data from the libraries generated.

Acknowledgements

I would like to express my extreme gratitude to Professor Phil Quirke for not only the opportunity to undertake this project but also for his continuing motivation, unwavering support and constant guidance throughout. It has been a huge privilege to work with him and I am very grateful for the unique opportunity he has given me.

I am extremely grateful to Dr. Lucy Stead for not only her help with bioinformatics but also for her continuing guidance and support. I would also like to greatly thank Dr. Phil Chambers for undertaking the pyrosequencing work for this project and his general assistance and guidance throughout. I am also very grateful to Miss Morag Taylor for her assistance with pyrosequencing and laboratory training. I would also like to offer my sincerest thanks to Dr. Michelle Cummings for all her help and guidance. I am also very thankful for being provided with laboratory training from Mr. Mike Shires.

I am also very grateful to Dr. Ian Carr, Professor Graham Taylor and Dr. Sally Harrison for all of their help and guidance with next generation sequencing as well as Dr. Stefano Berri and Dr. Henry Wood for their support and teaching in bioinformatics.

I would like to offer my sincerest thanks to all in the section of Pathology, Anatomy and Tumour Biology at LICAP for their support and advice throughout including Miss Gemma Hemmings, Dr. Susan Richman and Dr. Nic Orsi. I am very thankful to have been granted the University of Leeds Mary and Alice Smith memorial scholarship for funding my PhD. Also I would like to acknowledge Yorkshire Cancer Research for the funding of this project.

Finally, I would like to thank my wonderful family and especially my parents and Aryeh for being so supportive throughout. Their unwavering patience and love has helped me to complete this thesis and I am very grateful for all of their encouragement.

Abstract

Colorectal cancer (CRC) has a relatively poor prognosis when detected at a later stage, therefore understanding its development to allow prevention or early detection is key to improving outcomes. Bowel cancer screening allows for the detection of tumours and precursor lesions. Even earlier changes could potentially be detected; genetic aberrations within the normal bowel before phenotypic changes occur. Early lesions may develop independently throughout the bowel or through a cancer field effect. FAP adenomas are a useful model due to adenomas developing within the same environment and all incurring the same first “hit” within APC. Using next generation sequencing the genetic profiles of FAP adenomas can be compared to better understand the development of these lesions.

Firstly the sensitivity of pyrosequencing and NGS was determined as was the value of PCR-based mutant enrichment techniques. Samples of carcinoma, adenoma and their associated normals alongside normal mucosa from patients with normal colonoscopies were tested for mutations in commonly mutated genes in CRC using NGS. Multiple FAP adenomas from four patients were also tested with this mutation panel. Alongside this, copy number analysis was performed. The mutational and copy number data was used to ascertain the pattern of adenoma development throughout the bowel.

Mutations were detected in carcinoma associated normal in APC, KRAS, CTNNB1 and SMAD4. The KRAS mutations in carcinoma associated normal differed to the KRAS mutation in the matched tumour. No mutations were detected in oncogenes in adenoma associated normal or normal mucosa from patients with normal colonoscopies. Studying mutations and copy number aberrations in FAP adenomas revealed that some adenomas shared specific lesions, indicating that they were clonally related.

These results have confirmed previously findings of KRAS mutations in carcinoma associated normal mucosa as well as describing mutations in APC, CTNNB1 and SMAD4. This combined with the large amount of similarity in terms of mutations and copy number seen in adenomas from the same patient provides evidence for the cancer field theory.

Contents

List of figures

List of tables

List of abbreviations

°C	degrees Celsius
µl	micro litre
ACF	aberrant crypt foci
ACB-PCR	allele-specific competitive block polymerase chain reaction
APC	adenomatous polyposis coli
APS	adenosine 5'-phosphosulphate
ARMS	amplification-refractory mutant system
ATP	adenosine triphosphate
BAM	binary sequence alignment/map
BCSP	bowel cancer screening programme
bp	base pair
BRAF	V-raf murine sarcoma viral oncogene homolog B1
BWA	burrows wheeler alignment
CIMP	CpG island methylator phenotype
CIN	chromosomal instability
CMF	calcium and magnesium free
COLD-PCR	co-amplification at lower denaturation temperature polymerase chain reaction
COSMIC	catalogue of somatic mutations in cancer
CpG	cytosine - phosphate - guanine
CRC	colorectal cancer
CTNNB1	catenin beta-1
DGGE	denaturation gradient gel electrophoresis
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide
DSH	dishevelled receptor
EB	elution buffer
EDTA	ethylenediaminetetraacetic acid
EGFR	epidermal growth factor receptor
ERK	extracellular-signal-related kinase
FAP	familial adenomatous polyposis
FASTA	Fast-all
FASTQ	Fast-all with quality scores
FBXW7	F-box/WD repeat-containing protein 7
FCS	fetal calf serum
FFPE	formalin fixed paraffin embedded
FIT	faecal immunochemical testing
FOBT	faecal occult blood testing
Frz	frizzled
Fwd	forward
G1	growth 1 phase
GAIIe	genome analyser IIe
GATK	genome analysis toolkit
gDNA	genomic deoxyribose nucleic acid
H&E	haematoxylin and eosin

HBSS	Hank's buffered salt solution
HNPCC	hereditary non polyposis colorectal cancer
HP	hyperplastic polyp
Ice-COLD PCR	improved and complete enrichment co-amplication at lower denatuer temperate polymerase chain reaction
IQR	interquartile range
kb	kilo bases
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
LCM	laser capture microscopy
LRP	lipoprotein receptor-related protein
MCF7	Michigan cancer foundation-7
MD	mechanical dissociation
MEGA	molecular evolutionary genetics analysis software
MEK	mitogen-activated protein kinase kinase
$MgCl_2$	magnesium chloride
mins	minutes
ml	millilitre
MLH1	mutL homolog 1
MP	mixed polyp
MSH2	mutS protein homolog 2
MSH6	mutS homolog 6
MSI	microsatellite instability
MSI-H	microsatellite instable - high
MSI-L	microsatellite instable - low
MSS	microsatellite stable
mTOR	mammalian target of rapamycin
NEB	New England Biolabs
ng	nanograms
NGS	next generation sequencing
NHS	national health service
NRAS	neuroblastoma rat angiosarcoma viral oncogene homolog
NRES	national research ethics service
OCT	optimal cutting temperature
p	chromosome short arm
PBS	phosphate buffered saline
PEN	polyethylene napthalate
Phred	Phil's read editor
PI3K	phosphatidylinositide 3-kinase
PMS2	postmeiotic segregation increased 2
Ppi	pyrophosphate
PTEN	phosphatase and tensin homolog
q	chromosome long arm
RAF	v-raf murine leukaemia viral oncogene homolog
RAS	rat angiosarcoma gene
REC	research ethics committee
Rev	reverse
RFLP	restriction fragment length polymorphism

RMPI	Roswell Park Memorial Institute medium
SAM	sequence alignment/map
sec	seconds
SMAD4	mothers against decapentaplefic homolog 4
SNP	single nucleotide polymorphism
SOS	sons of sevenless
SSL	sessile serrated lesion
TALC	targeted amplicon library creation
T_c	critical temperature
TCF7L2	transcription factor 7-like 2
TE	tris-ethylenediaminetetraacetic acid buffer
TGFβ	transforming growth factor beta
TNM	tumour node metastases
TP53	tumour protein 53
TSA	traditional serrated adenoma
TSG	tumour suppressor gene
U	unit
UK	United Kingdom
USA	United States of America
UV	ultra violet
VCF	variant caller format
Wnt	wingless-int
WT	wild type

1 Introduction

Pathology, genetics and early lesions of colorectal cancer

1.1 Epidemiology of colorectal cancer

1.1.1 Incidence

There are over 1.2 million new cases of colorectal cancer (CRC) diagnosed each year worldwide which makes it the third most common cancer in men and second most frequent in women (GLOBOCAN, 2008). The incidence of CRC varies geographically, with more cases seen in economically developed countries. This is due to associations with diet, obesity, sedentary lifestyle and inadequate fibre intake alongside other risk factors (Cunningham et al., 2010). Furthermore, the incidence has increased in countries that have started to adopt a more westernised diet such as Japan and other Asian countries (Cancer Research UK, 2011). In the UK, CRC is the third most common cancer type in both men and women with around 110 new cases diagnosed every day. It is a disease that affects the older population with 86% of cases being diagnosed in those patients aged 60 or over as illustrated in Figure 1 (Cancer Research UK, 2010).

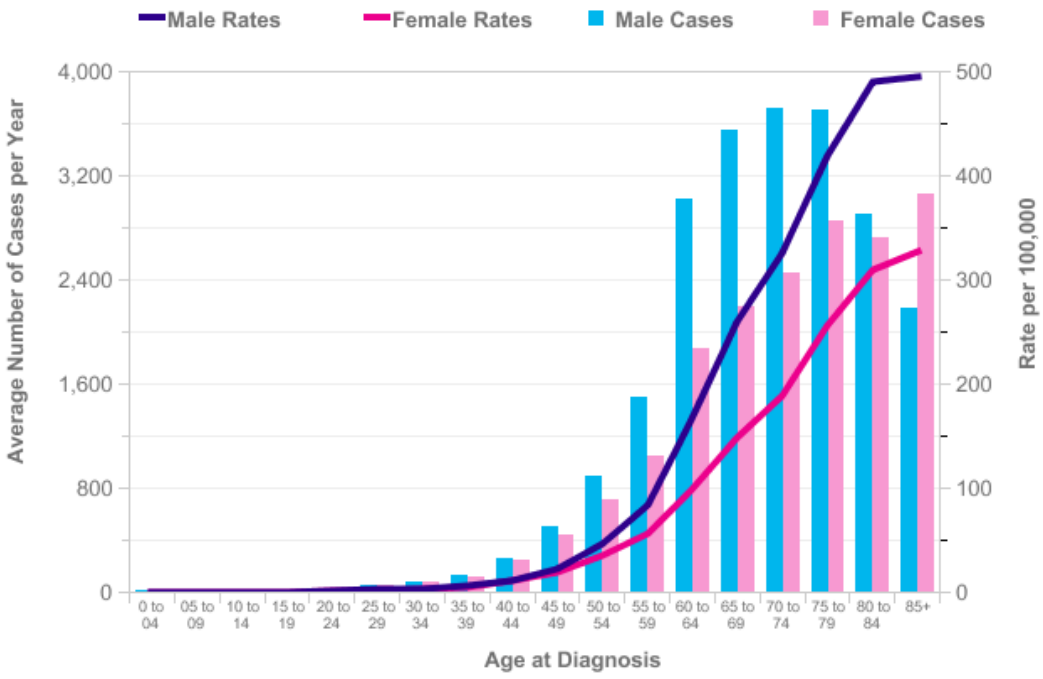


Figure 1. Colorectal cancer incidence. Number of new cases per year and age-specific incidence rates per 100,000 population, UK (Cancer Research UK, 2011).

1.1.2 Mortality rate and 5-year survival

Worldwide, CRC is the fourth most common cause of cancer death. In the UK, it is the second highest cause of cancer-death and the five-year survival rates for CRC are 54.2% and 55.6% for men and women respectively (2005-2009) (Cancer Research UK, 2012b). Patient prognosis for CRC is heavily reliant upon the stage at which it is diagnosed (Figure 2). Approximately one-third of CRCs are diagnosed as Dukes' stage A and B (early stage disease) and one-third are categorised as Dukes' stage C and D (later stage, more extensive disease) with the remaining third being unknown (Figure 3). Dukes' A stage has a 5-year relative survival of 93.2% compared to 6.6% for stage D (Figure 2) (Cancer Research UK, 2012b) and therefore diagnosing cancer at an earlier stage is a key part of improving patient outcomes.

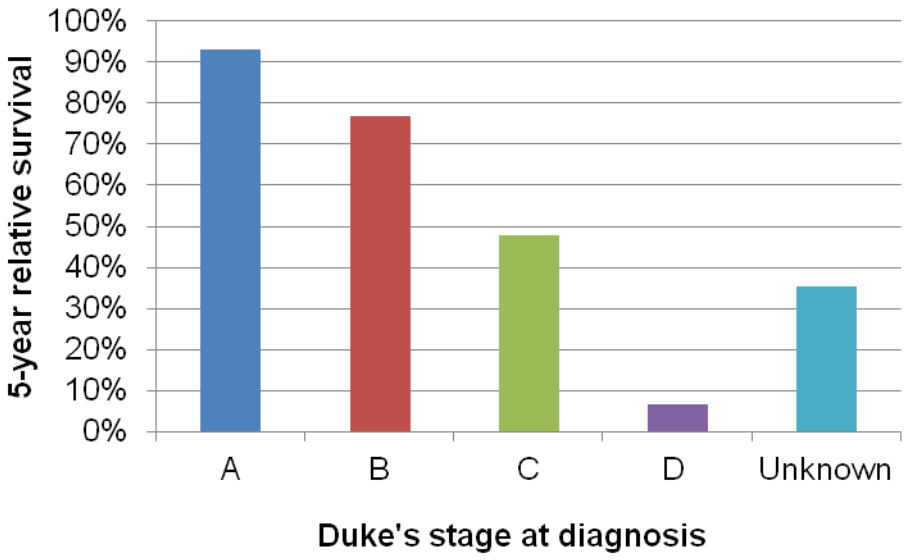


Figure 2. Five-year relative survival rates by Dukes' stage at diagnosis. England, 1996-2002 (Cancer Research UK, 2012b).

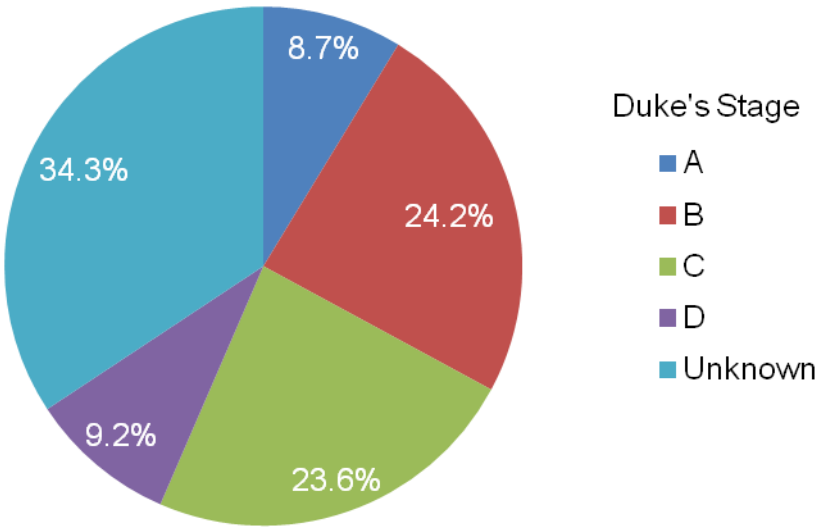


Figure 3. Percentage of cases by Dukes' stage at diagnosis, England 1996-2002 (Cancer Research UK, 2012b)

1.2 Screening

Population screening for CRC is an effective method to reduce mortality rates due to disease being detected at an earlier stage (Mandel et al., 1993). The Bowel Cancer Screening Programme (BCSP) in England was initiated in 2006 and was fully rolled out by 2009. This consists of faecal occult blood testing (FOBT) for the population aged 60 every 2 years and in 2011 the BCSP introduced flexible sigmoidoscopy from age 55 (Atkin et al., 2010). The FOBT is used to detect tiny amounts of blood in the stool. Those with positive tests are then subject to further investigation which normally involves colonoscopy to visualise the large bowel. By this method, mortality can be reduced by up to 25% (Logan et al., 2012).

An alternative method for the detection of lesions from stool is faecal immunohistochemical testing (FIT) which tests specifically for the globin protein. This differs from FOBT which detects the haem component (Park et al., 2010). Typically, stool tests such as FOBT and FIT tend to detect cancer as well as incidental adenomas. Visual examinations (flexisigmoidoscopy, colonoscopy etc.) allow for the direct detection of both cancer and also premalignant lesions (Quintero et al., 2012). There is also the potential to develop mutation testing. If genetic mutations were to be detected by a sensitive technique they might either be used as the definitive test or as a marker to identify those who are more "at risk" of developing colorectal cancer.

1.3 Anatomy and histology of the colorectum

1.3.1 Anatomy of the large bowel

The large bowel is the region of alimentary canal between the terminal ileum and the anus. The colon begins at the caecum which is situated below the ileo-caecal valve. The ascending colon stretches up to the hepatic flexure, at which point it becomes the transverse colon. This runs to the splenic flexure where it becomes the descending colon. This joins to the loop of colon known as the sigmoid which then meets the straight rectum before finally ending at the anal canal. The majority of CRCs occur in the sigmoid and rectum (Cancer Research UK, 2012a).

A clinical division is often made at a point two-thirds along the transverse colon to form the right and left colon. This is due to the different embryological origins; the right colon is derived from the embryological midgut and the left colon is derived from the hindgut. Reflecting this division, the left and right colon have different vasculature.

1.3.2 Normal histology of the large bowel

The gastrointestinal tract is subdivided into five distinct functional layers: mucosa, submucosa, muscularis propria, sub serosa and serosa. The mucosa from the caecum to rectum consists of epithelial cells that lie on a basement membrane with supporting lamina propria beneath. Underneath lies the muscularis mucosae which then forms the boundary with the submuscosa. The epithelial cells are folded into crypts which provide the large surface area required for the large bowel's absorptive functions. The basic architecture of the colon is illustrated in Figure 4.

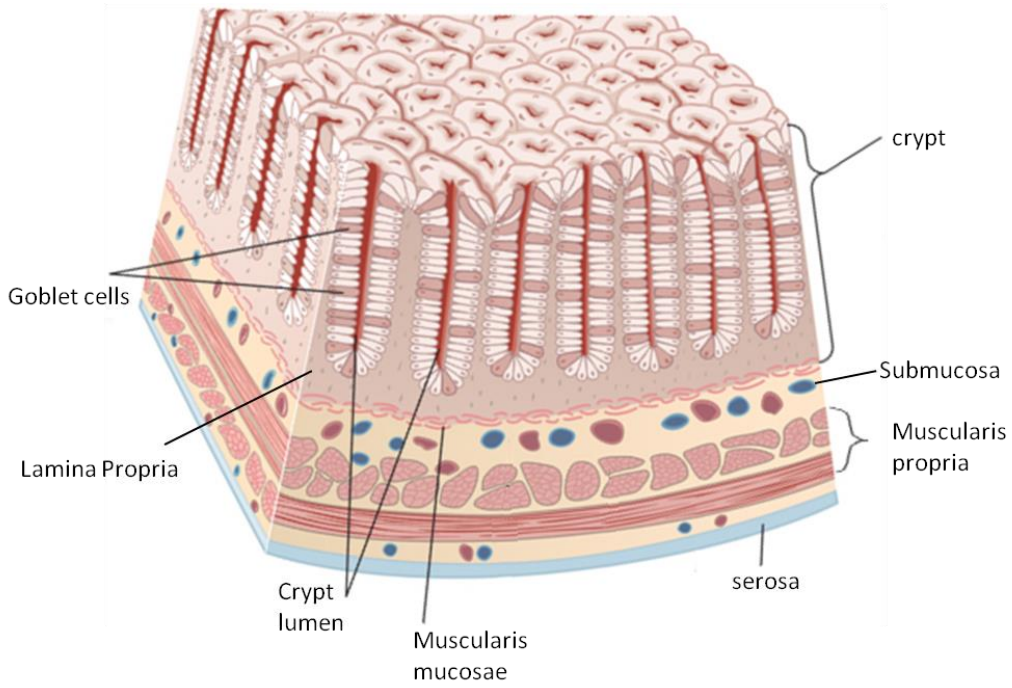


Figure 4. The internal structure of the colon. Adapted from (Encyclopædia Britannica Online, 2003).

The slide in Figure 5 shows the histological features of normal colorectal mucosa. The epithelium-lined crypts and surrounding lamina propria clearly define the colon mucosa with the thin muscle layer of the muscularis mucosae separating it from the submucosa.

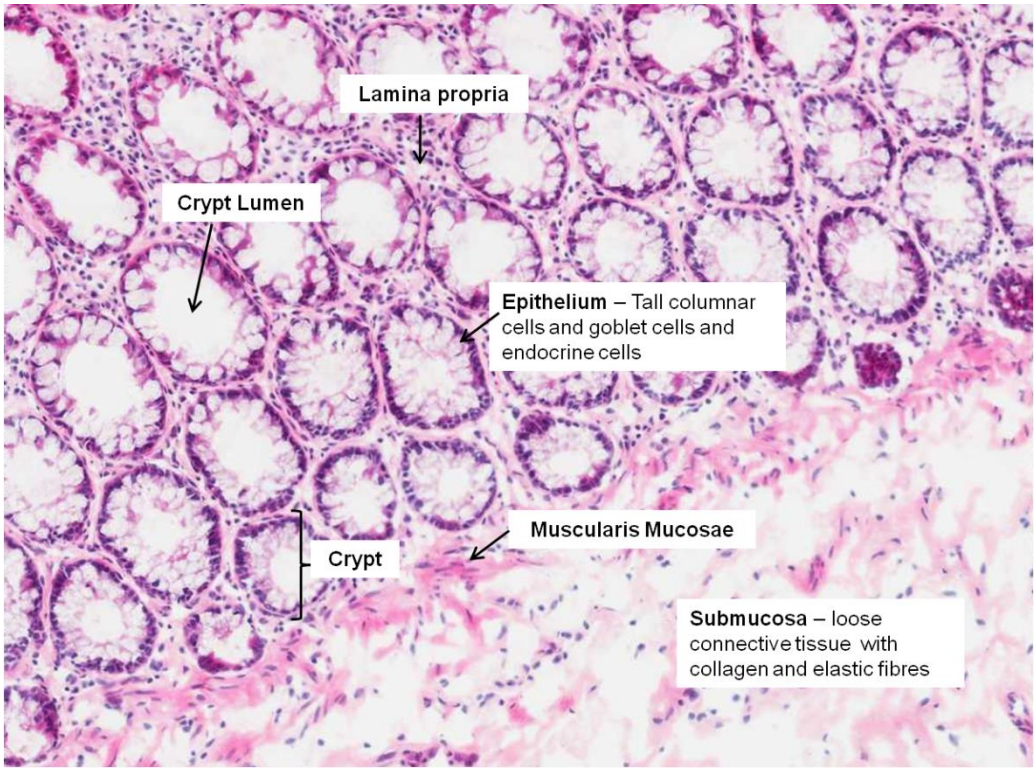


Figure 5. Frozen section of normal mucosa stained with haematoxylin and eosin.

1.4 Staging of colorectal cancer

1.4.1 Dukes' Staging

Staging CRC has a number of advantages. It allows for better prediction of survival, enables international comparisons of outcome, helps to determine treatment and streamlines clinical trials. The Dukes' classification system indicates the stage, or extent of spread of disease. There are 4 categories: A,B,C1 and C2 described by Dukes and category D added by (Turnbull Jr et al., 1967) . Table 1 outlines the criteria for each particular stage.

Stage	Criteria
A	No invasion beyond muscularis propria.
B	Invasion beyond muscularis propria, but no lymph node involvement.
C1	Regional lymph node involvement, but not the highest node
C2	Lymph node involvement at high surgical tie
D	Distant metastases.

Table 1. Dukes' classification system for colorectal cancer (Turnbull Jr et al., 1967)

1.4.2 Tumour, Node, Metastases staging

TNM staging is a more preferred method of classification due to the higher amount of information it provides. Cancers are categorised according to the size of the primary tumour (T), the number of lymph nodes involved (N) and the presence of any metastases (M) (Table 2). There are various versions of TNM staging and version 5 is preferred in the UK rather than version 7 (Quirke et al., 2011).

Category	Criteria
T1	No invasion beyond muscularis propria.
T2	Invasion into muscularis propria.
T3	Invasion into subserosa or neighbouring tissues.
T4	Invasion into other organs and peritoneal invasion.
N0	No regional lymph node involvement.
N1	Metastasis in 1 to 3 regional lymph nodes.
N2	Metastasis in 4 or more regional lymph nodes.
M0	No distant metastasis.
M1	Distant metastasis.

Table 2. TNM classification of colorectal cancer (Sobin and Fleming, 1997)

1.5 Lesions of colorectal cancer

The initial model of carcinoma development proposed a step-wise progression from normal mucosa to adenoma to adenocarcinoma (Vogelstein et al., 1988). This sequence is characterised by activation of oncogenes and loss of tumour suppressor genes (TSGs) at different points, contributing to the stages of tumourigenesis. However, subsequent classification of early colorectal lesions and alternative pathways of tumour development have led to substantial modification of this model (Bozic et al., 2010).

1.5.1 Aberrant crypt foci

Aberrant crypt foci (ACF) are defined as an area of colorectal mucosa with two or more adjacent enlarged crypts with structural changes (Orlando et al., 2008, Anderson et al., 2012). ACF progress to small adenoma via the accumulation of molecular damage and are considered the intermediate between normal epithelium and early adenoma. They have been shown to have increased proliferative activity and often carry a KRAS mutation (section 1.7.2) (Takayama et al., 1998). ACFs can be further classified by the degree of dysplasia they show (Sancho et al., 2004). Figure 6 shows an example of a dysplastic ACF.

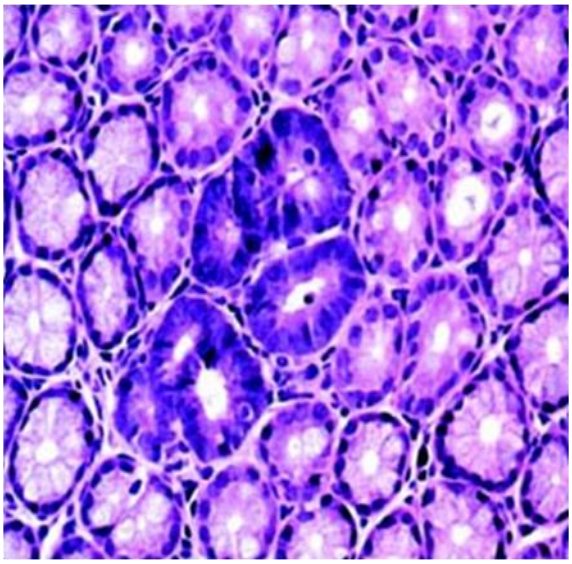


Figure 6. Dysplastic aberrant crypt foci. Reproduced with permission from (Redston, 2001).

1.5.2 Serrated lesions

This serrated category refers to a distinct histological appearance that is observed and encompasses both neoplastic and non-neoplastic lesions. The surface epithelium of the crypts displays a saw-tooth appearance, with increased levels of mucin in the cytoplasm (Rex et al., 2012). Serrated lesions are associated with microsatellite instability (section 1.6.2) and BRAF mutations (section 1.7.2) and are thought to develop to carcinoma through this pathway (Mesteri et al., 2013). There are four main types of lesion that are included within the serrated category: hyperplastic polyps (HPs), sessile serrated lesions (SSL), traditional serrated adenomas (TSA) and mixed polyps (Cunningham and Riddell, 2006, Noffsinger, 2009, Rex et al., 2012)

HPs are benign polyps with negligible malignant potential (Figure 7A). They are most often found in the left colon and are typically less than 5mm in size. Crypts within HP are elongated with serrated epithelium in the upper half of the crypt and increased proliferation in the basal portion of the crypt (Rex et al., 2012). SSLs are characterized by the presence of abnormal proliferation but without the presence of mucosal neoplasia (Figure 7B). Serration is seen down to the crypt base. TSAs are dysplastic with a villiform/filliform appearance with foci of microcrypt formation and display mucosal neoplasia, with associated serration (Figure 7C). They are often found on the left-side of the colon (Lash et al., 2010). Mixed polyps are lesions that contain a combination of histological types (Figure 7D) and they may include neoplastic changes (Rex et al., 2012).

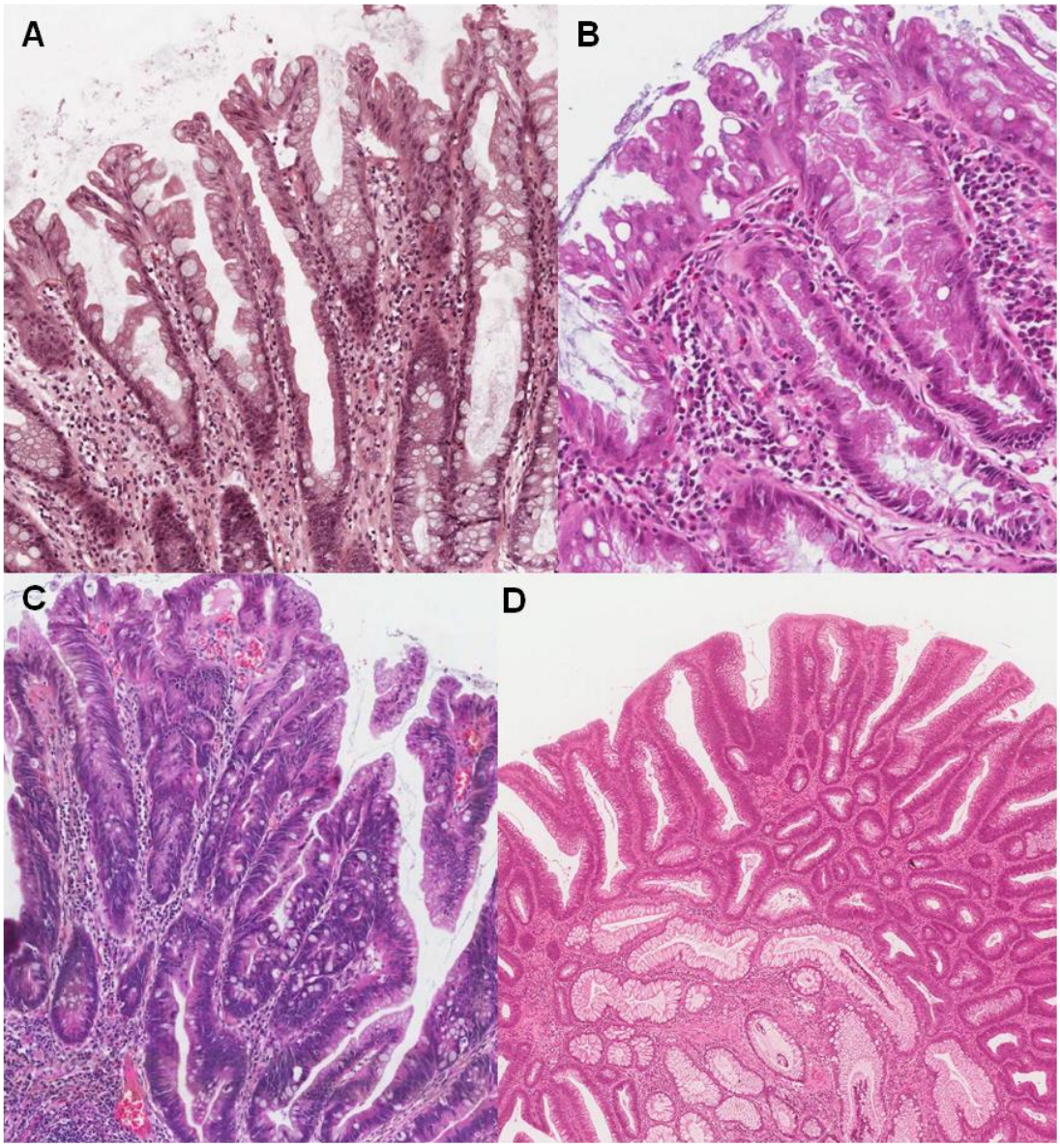


Figure 7. Serrated Lesions. A = Hyperplastic polyp. B = Sessile serrated lesion. C = Traditional serrated adenoma. D = Mixed polyp.

1.5.3 Adenoma and carcinoma

Colorectal adenomas are defined by the presence of epithelial neoplasia (Quirke et al., 2011). They are divided according to their morphological appearance into tubular, villous and tubulo-villous. A tubulo-villous adenoma must contain between 20% and 80% villous component to be classed as such. They are also categorised according to the Vienna grading system according to the degree of structural and cytological abnormality (Lanza et al., 2011). Colorectal adenocarcinomas are defined by the invasion of neoplastic cells through the muscularis mucosae and into the submucosa (Bosman et al., 2010). The cells of both adenomas and carcinomas display nuclear pleomorphism, increased mitotic activity and disruption of normal mucosal architecture, however only carcinomas invade.

Category	Criteria
1	No neoplasia
2	Indefinite from neoplasia
3	Mucosal low-grade neoplasia
	Low grade adenoma
	Low grade dysplasia
4	Mucosal high-grade neoplasia
4.1	High grade adenoma/dysplasia
4.2	Non-invasive carcinoma (carcinoma in-situ)
4.3	Suspicious for invasive carcinoma
4.4	Intramucosal carcinoma
5	Submucosa invasion by carcinoma

Table 3. Revised Vienna Classification of neoplasia grading (Dixon, 2002).

1.6 Genetic instability

Colorectal cancer is increasingly classified according to molecular pathogenesis. The underlying genetic alterations are caused by various mechanisms of genetic instability. There are three main forms of genetic instability which are described; chromosomal instability (CIN), microsatellite instability (MSI) and CpG Islands Methylator Phenotype (CIMP). These pathways cause the subsequent inactivation of tumour suppressor genes (TSGs) and activation of oncogenes, thereby resulting in tumour formation.

1.6.1 Chromosomal instability

The most common form of genetic instability seen in colorectal cancer is chromosomal instability (CIN) and is observed in up to 85% of tumours (Grady and Carethers, 2008). It is defined by an accelerated rate of loss and gain of whole sections of genetic material. This leads to aneuploidy, genetic amplification, translocations and loss of heterozygosity. (Watanabe et al., 2012)

The most common feature of CIN is the gain or loss of large genetic sequences that code for oncogenes and TSGs respectively. Chromosomal gains are most often found at 8q, 20, 7 and 13. Loss of genetic segments most often occurs at 5q, 8p, 17p and 18q which may result in loss of heterozygosity (Pino and Chung, 2010, Ashktorab et al., 2010).

There are multiple causes of CIN, many of which are still not fully understood. Mutations in the adenomatous polyposis coli gene (APC) (section 1.7.1), as well as its tumourigenic effects, are a suggested initiator of CIN (Pino and Chung, 2010). Similarly there appears to be a causative link between Kirsten rat sarcoma (KRAS) mutation and aneuploidy (Gordon et al., 2012) (section 1.7.2). Amplification of chromosome 7 has been noted in some colon adenomas and also appears to play a role in colorectal cancer progression. The epidermal growth factor receptor gene is located on chromosome 7 and therefore its expression may be affected (Sartore-Bianchi et al., 2012). However, it is unknown whether changes in TSGs and oncogenes are due to CIN or the driving force behind it.

1.6.2 Microsatellite instability

Microsatellite instability (MSI) is observed in approximately 15-20% of sporadic colorectal cancers and over 95% of tumours from patients with hereditary non polyposis coli (HNPCC) (section 1.8.2) (Boland and Goel, 2010). It is a form of genetic instability characterized by the accumulation of abnormally sized short nucleotide repeats (microsatellites) throughout the tumour genetic sequence. MSI has been shown to be due to errors in DNA mismatch repair (MMR) genes such as MSH2, MLH1, PMS2 and MSH6 (Tian et al., 2012). There are two main mechanisms by which MMR genes can lose their function; somatic inactivation of both alleles of MMR genes or through germline MMR gene mutations followed by inactivation of the other copy as seen in HNPCC.

MMR proteins are enzymes that normally function to recognize nucleotide mismatches and facilitate their repair. Loss of this function prevents errors that occur during cell replication from being corrected. Throughout the genome microsatellites of a fixed length are present that are prone to error during replication with deficient MMR. Therefore repeat units of DNA of incorrect lengths are accumulated throughout the genome resulting in MSI (Markowitz and Bertagnolli, 2009). Microsatellites in TSGs disrupt the reading frame, thereby rendering the protein truncated and inactive. This causes the loss of TSG function and subsequent adenoma development. Key TSGs thought to be affected by microsatellites include TGFβ receptor type II and components of the Wnt signalling pathway (section 1.7.1) (Soreide et al., 2009, Alhopuro et al., 2012).

Tumours can be further classified according to the degree of MSI observed. The majority of tumours display either large amounts of MSI or none at all (microsatellite stable, MSS) (Ogino and Goel, 2008, de la Chapelle and Hampel, 2010). This gives rise to the terms MSI-high (MSI-H) and MSI-low (MSI-L) to further classify MSI tumours. More recently the term “hypermutated” has been used to refer to tumours with a high mutation rate (over 12 mutations per 10,000,000 bases) and 77% of these are reported as MSI-H (The Cancer Genome Atlas Network, 2012).

1.6.3 Aberrant DNA methylation

TSGs become inactivated not only by genetic mutations and partial or whole gene loss but also through epigenetic changes; the most common mechanism being aberrant DNA methylation. In sporadic CRC, the MLH1 gene is epigenetically silenced resulting in MSI (Markowitz and Bertagnolli, 2009). Within the promoter region of a gene can be found an area referred to as a CpG island. CpG islands are defined as segments of DNA which have a high frequency of the CpG dinucleotide that are often found upstream to the gene's transcription start site (Soreide et al., 2009). CpG repeats located elsewhere within the genome are normally silenced by methylation, however promoter CpG islands remain unmethylated (Issa, 2004). In CRC promoter CpG islands in TSGs undergo methylation thereby preventing transcription of the gene. Approximately 15% of CRCs display the CpG island methylator phenotype, (CIMP) (Markowitz and Bertagnolli, 2009). As well as tumours, abberant DNA methylation has been observed in normal mucosa that lies adjacent to colorectal tumours (Kim et al., 2010).

Tumours that are classified as CIMP-high have been shown to display distinct pathological and molecular features. There are associations with female sex, proximal tumour location and high BRAF mutation rates. CIMP-low as an independent category is not as strongly defined as CIMP-high, however these tumours have been shown to hold a stronger association with KRAS mutations (Kawasaki et al., 2008)

1.7 Molecular pathogenesis and genetics of colorectal cancer

1.7.1 Wnt Signalling (APC, CTNNB1, TCF7L2)

The wingless related integration (wnt) signalling pathway regulates cell fate determination, cell migration and cell polarity. Wnt target genes are involved in tumour formation and invasion and therefore activation of wnt signalling promotes tumour growth. (Anastas and Moon, 2012). As illustrated by Figure 8, when the wnt pathway is in the “off” state there is an absence of a wnt signal and the frizzled (Frz) receptor is free of ligand. This leads to β-catenin binding to a destruction complex consisting of adenomatous polylposis coli (APC), axin, casein kinase 1 (CK1) and glycogen synthase kinase 3β (GSK3β) protein. This results in the cytoplasmic degradation of β-catenin via proteosomes. In the nucleus, target wnt genes remain silenced due to the interaction between the repressor groucho and DNA-bound T cell factor 4 (TCF4). To change this pathway to the “on” state, firstly the wnt signal has to bind to the frizzled receptor (Frz) and co-receptor low-density lipoprotein receptor-related protein (LRP). This causes the LRP tail to become phosphorylated by CK1, GSK3β and the dishevelled (DSH)-dependent recruitment of axin. This results in β-catenin no longer becoming phosphorylated and levels accumulate. β-catenin then translocates to the nucleus where it stimulates the transcription of wnt target genes by displacing groucho and recruiting B cell lymphoma 9 (Bcl-9) and Pygopus through TCF4.

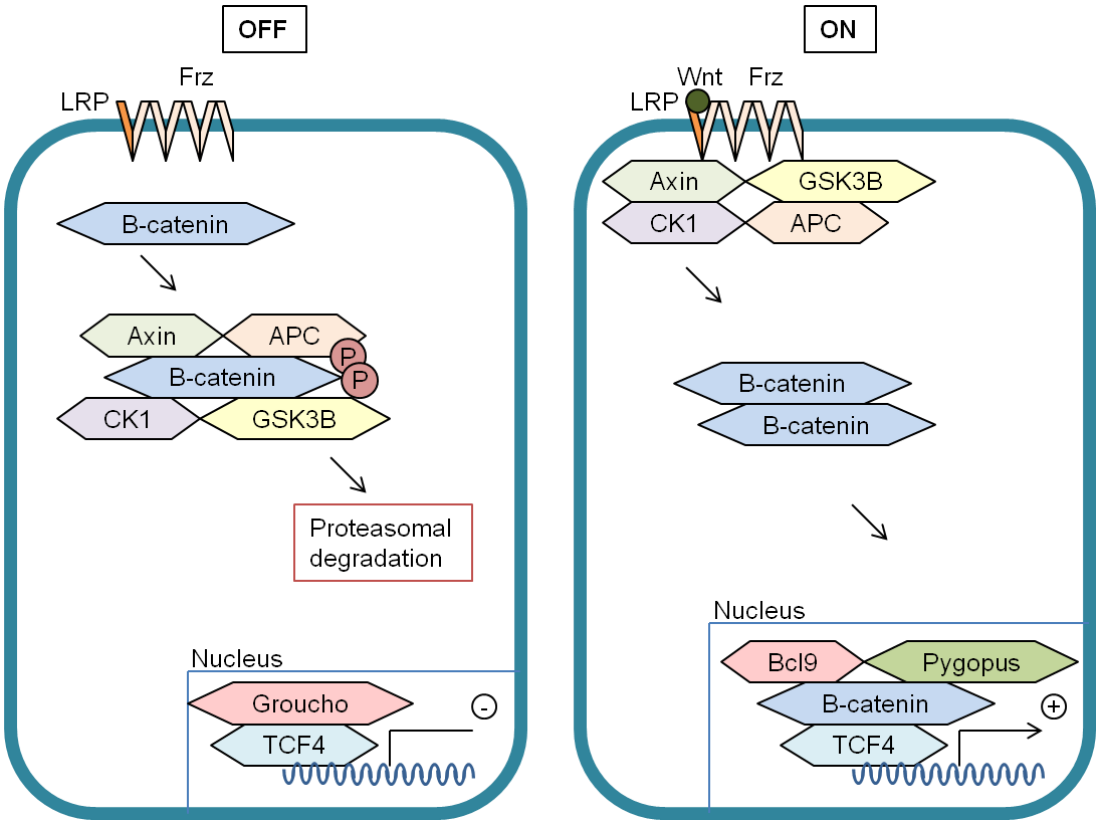


Figure 8. Wnt Signalling. LRP = lipoprotein receptor-related protein. Frz = frizzled receptor. APC = adenomatous polyposis coli. P = Phosphate. CK1 = Casein kinase 1. GSK3B = Glycogen synthase kinase 3.TCF4 = T cell factor 4. Bcl9 = B-cell CLL/lymphoma 9.

The APC tumour suppressor gene (APC) is found inactivated in 80-95% of sporadic colorectal cancers (Markowitz and Bertagnolli, 2009). The APC protein normally functions to bind to β-catenin and degrade it thereby inactivating the wnt signalling pathway. Mutated APC protein contains structural deformities in its β-catenin binding site, leading to unopposed activation of the wnt signalling pathway. β-catenin accumulates in the cytoplasm and is translocated to the nucleus. Here it stimulates the transcription of multiple genes that are involved in tumour formation and invasion. Inactivation of both APC genes is found in most sporadic CRCs and this loss of APC function is a major step in the initiation of tumourigenesis.

The β-catenin protein is encoded by the CTNNB1 gene and CTNNB1 mutations are found in around 5% of sporadic CRCs and in up to 50% of tumours with intact APC (Pino and Chung, 2010, The Cancer Genome Atlas Network, 2012). Activating mutations in the CTNNB1 gene allow for the β-catenin protein to be stabilized, switching on wnt signalling. TCF7L2 mutations are found in around 9% of sporadic CRCs (The Cancer Genome Atlas Network, 2012). The TCF7L2 gene codes for TCF4 protein and mutation can affect its binding properties allowing for wnt target gene transcription (Bass et al., 2011).

1.7.2 RAS-RAF-MEK-ERK pathway

This pathway controls cell proliferation through the activation of epidermal growth factor receptors (Figure 9). Growth factors bind to cell surface receptors which then trigger activation of sons of sevenless (SOS). This causes the exchange of GDP for GTP on the RAS protein, thereby converting it to the active state. This, in turn, stimulates the RAF-MEK-ERK pathway which results in the transcription of genes responsible for cell proliferation.

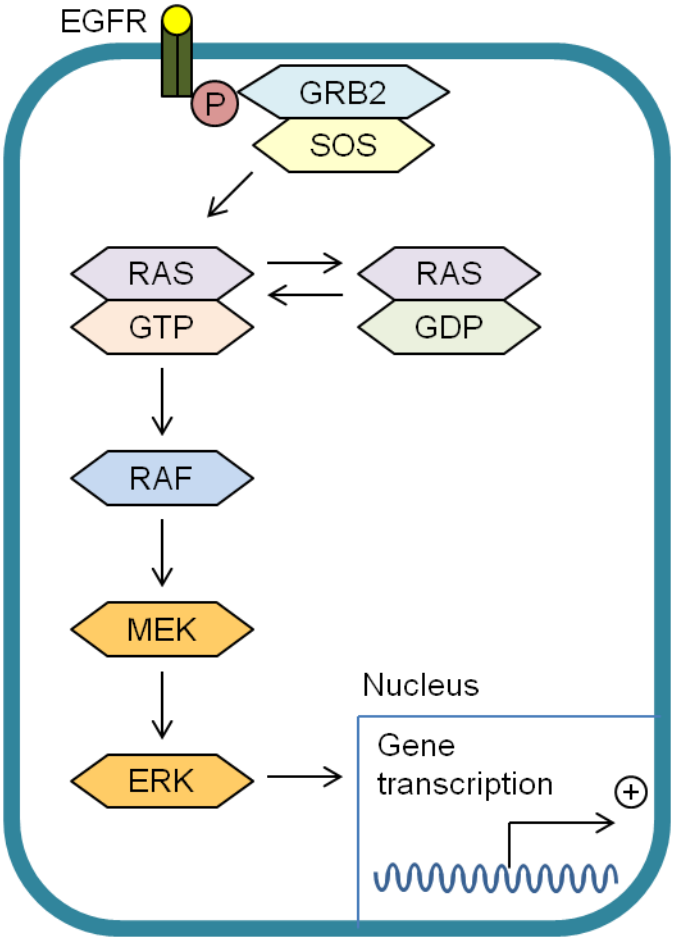


Figure 9. RAS-RAF-MEK-ERK signalling pathway. EGFR = Epidermal growth factor receptor. P = Phosphate. GRB2 = Growth factor receptor-bound protein 2. SOS = Sons of sevenless. RAS = Rat sarcoma. GTP = Guanosine triphosphate. GDP = Guanosine diphosphate. RAF = Rapidly accelerated fibrosarcoma. MEK = Mitogen-activated protein kinase kinase. ERK = Extracellular signal-regulated kinase.

The RAS protein family include Kirsten-RAS (KRAS) and neuroblastoma-RAS (NRAS) in which mutations are commonly implicated in CRC. Activating mutations in hotspots of these oncogenes result in a protein that is permanently in the active state with GTP bound. Alternatively, this pathway may be permanently activated due to mutations in the RAF gene and most commonly the BRAF gene. Approximately 50% of CRCs contain a KRAS mutation, 9% contain a BRAF mutation and 5% contain an NRAS mutation (Bamford et al., 2004, Pino and Chung, 2010). Table 4 shows a breakdown in the prevalence of RAS and RAF mutations in CRC.

Mutation	Prevalence (%)
KRAS codon 12	33.0
KRAS codon 13	9.5
KRAS codon 61	2.1
KRAS codon 146	1.9
NRAS codon 12	1.0
NRAS codon 13	0.1
NRAS codon 61	1.8
BRAF codon 600	8.8

Table 4. Frequency of KRAS, NRAS and BRAF mutations found in colorectal cancer (Bamford et al., 2004, Pino and Chung, 2010, The Cancer Genome Atlas Network, 2012).

1.7.3 Phosphoinositide 3-kinase pathway

The Phosphatidylinositide 3-kinase (PI3K)/ Ak-thymoma (AKT) pathway plays an important role in apoptosis and cell survival (Liao et al., 2012). Activated EGFR receptors and consequently SOS protein causes activation of the PI3K, AKT and mammalian target of rapamycin (mTOR) signalling cascade. This results in transcription of target genes (Figure 10). Phosphatase and tensin (PTEN) protein acts on AKT to inhibit its signalling and therefore inhibits the pathway at this point.

The PI3K protein is coded for by the PIK3CA gene which is found mutated in approximately 15-20% of CRCs (Liao et al., 2012). Mutations in the PIK3CA oncogene are found in hotspots; at codons 542, 545-547 and 1047 which result in gain of function (Bamford et al., 2004) and permanent activation of the pathway. PTEN acts as a tumour suppressor and mutations in this gene result in a loss of inhibition of the pathway. PTEN mutations are found in 4% of CRCs.

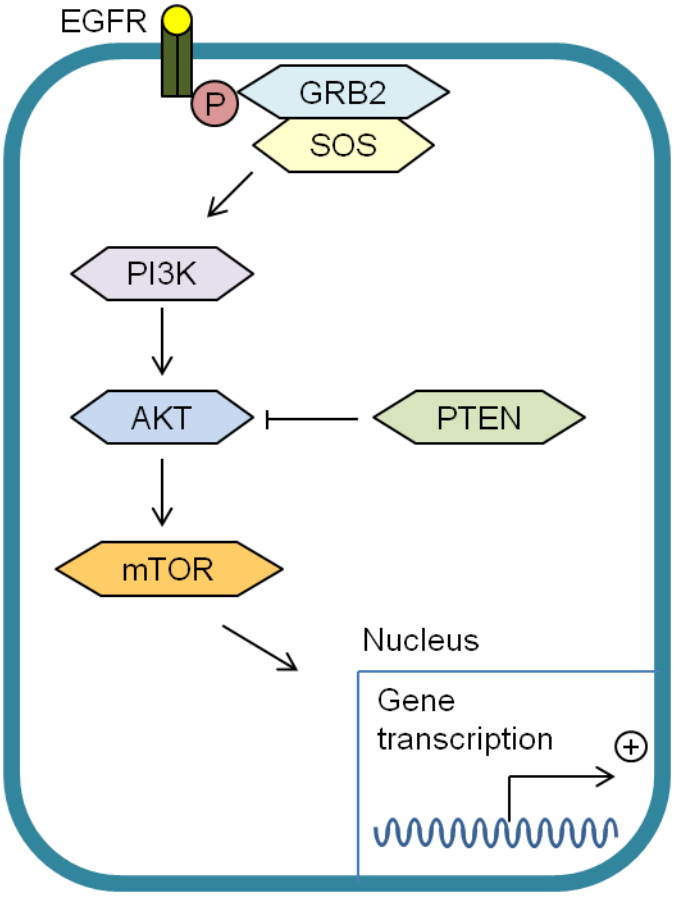


Figure 10. PIK-AKT-mTOR pathway. . EGFR = Epidermal growth factor receptor. P = Phosphate. GRB2 = Growth factor receptor-bound protein 2. SOS = Sons of sevenless. PI3K = Phosphatidylinositide 3-kinase. AKT = Ak-thymoma. PTEN = Phosphatase and tensin. mTOR = mammalian target of rapamycin.

1.7.4 TP53

The tumour protein 53 (TP53) is responsible for Gap 1 (G1) phase cell-cycle arrest; allowing for DNA repair to occur during replication, or alternatively induce apoptosis. Therefore TP53 acts as a tumour suppressor gene and loss of this protein results in increased DNA damage, resistance to apoptosis and loss of other tumour suppressor effects (Moran, 2010). The TP53 gene is located on the short arm (p) of chromosome 17. It is found mutated in up to 60% of sporadic CRCs (The Cancer Genome Atlas Network, 2012). It has been identified as playing an important role in the transition from adenoma to carcinoma. TP53 mutation frequency increases with the progression of the adenoma-carcinoma sequence (Moran, 2010). Alternatively TP53 can be lost through deletion of 17p which occurs frequently in CRC (Pino and Chung, 2010).

1.7.5 SMAD4

Mothers against decapentaplegic 4 (SMAD4) is a protein involved in transforming growth factor β (TGFβ) signalling. This pathway is responsible for neoplastic cell proliferation, especially at the later stages of the malignant process (Labelle et al., 2011). TGFβ binds to receptors on the cell surface which results in activation of SMAD2/3 proteins. These translocate to the nucleus where they interact with SMAD4 and cofactor to initiate gene transcription of target genes associated with tumour progression.

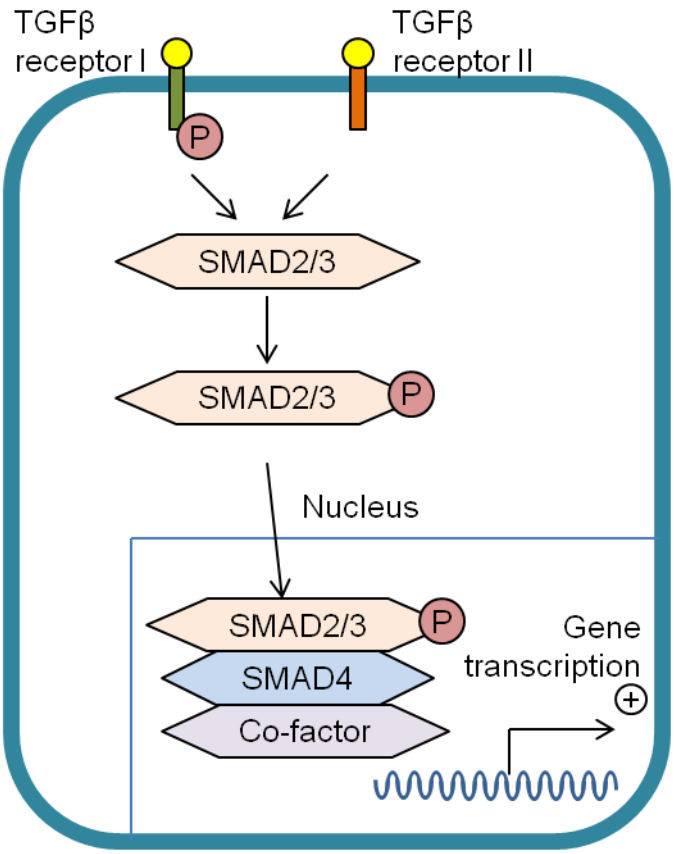


Figure 11. TGFβ signalling. TGFβ = Transforming growth factor β. SMAD = Mothers against decapentaplegic.

SMAD4 is located on the long arm of chromosome 18 (18q) and is found mutated in 10% of CRCs (The Cancer Genome Atlas Network, 2012). Its function is also lost through deletion of 18q which is observed in approximately 70% of tumours (Pino and Chung, 2010).

1.7.6 FBXW7

F-box/WD repeat-containing protein 7 (FBXW7) plays a role in the major regulatory pathways of protein degradation and is involved in cellular division (Akhoondi et al., 2007). Mutated FBXW7 is thought to interact with c-Myc which is a transcription factor. Therefore it is thought that FBXW7 mutation leads to the unregulated expression of many genes implicated in cancer growth (King et al., 2013). Mutations in the FBXW7 TSG have been found in approximately 11% of CRCs (The Cancer Genome Atlas Network, 2012).

1.8 Hereditary colorectal cancer

Inherited colorectal cancer syndromes account for around 5% of all CRCs. The two most common are familial adenomatous polyposis (FAP) and hereditary non-polyposis colorectal cancer (HNPCC).

1.8.1 Familial adenomatous polyposis

FAP is an autosomal dominant colorectal cancer syndrome and is caused by a germline APC mutation. Hundreds to thousands of adenomas form throughout the colon and rectum and progression to carcinoma is 100% by the age of 35-40 if left untreated (Half et al., 2009). The most common type of APC mutation involves the premature introduction of a stop codon resulting in the formation of a truncated APC protein. The majority of germline mutations are found in exon 15, known as the mutation cluster region (MCR) (Galiatsatos and Foulkes, 2006).

1.8.2 Hereditary non-polyposis colorectal cancer

HNPCC, also referred to as Lynch syndrome is characterised by an inherited germline mutation of the MMR genes (MSH2, MLH1, MSH6, PMS2) followed by subsequent somatic inactivation of the unaffected allele. It is an autosomal dominant cancer predisposition syndrome, resulting often in the early development of colorectal cancer. Patients with Lynch syndrome carry an 80% lifetime risk of developing colorectal cancer by on average, 45 years of age (Markowitz).

1.9 The development of early lesions from normal mucosa

By studying the pathways involved in colorectal cancer development we have begun to understand the events that occur in order for a cancer to develop. A very important part of this process is to understand how molecular events effect the transition from normal bowel mucosa into adenoma. It has been hypothesised that the somatic mutations found in tumours occur before the tumour has developed and this has led to the development of the clonal evolution model.

1.9.1 Clonal evolution and mutations in normal mucosa

Cancer is thought of as an evolutionary process; a cell acquires a mutation which gives it a selective growth advantage, allowing it to clonally expand (Greaves and Maley, 2012). This clone can then acquire genetic aberrations which allow it to develop as an adenoma and further changes to become carcinoma. Alternatively some of these clones may not survive and therefore not contribute to the tumour. Multiple clones can occur simultaneously and these may evolve separately or interact with each other. If a clone is present in the bowel for a long enough time to become “fixed” before the next clone starts to evolve, then clones will develop periodically. However, if the time for a new clone to emerge is shorter than the amount of time it takes for the first clone to become “fixed”, the clones will interfere with each other (Baker et al., 2013). Figure 12 is a schematic of clone development within the bowel.

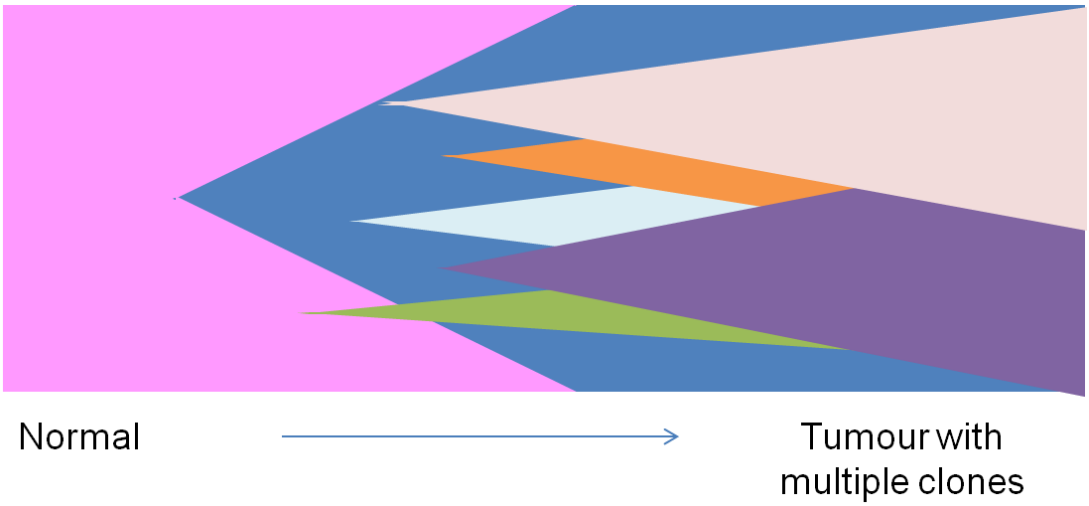


Figure 12. Clonal expansion of the colorectum. The pink on the left of the diagram represents normal bowel mucosa. The expanding triangles represent different clones expanding and either dying or surviving. Adapted from (Baker et al., 2013) and (Tomasetti et al., 2013).

1.9.2 Mutations in normal colorectal mucosa

Mathematical models have shown that a large proportion of the somatic mutations present in tumours occurred before phenotypic changes have occurred (Tomasetti et al., 2013). It has also been observed that low level mutations are present in histologically normal mucosa in KRAS (Parsons et al., 2010). Therefore mutations from abnormal clones in normal mucosa may be detectable with a highly sensitive technique.

1.9.3 Cancer field change

The term field cancerisation was originally used in 1953 to describe histologically abnormal tissue that surrounded oral squamous-cell carcinoma (Slaughter et al., 1953). It has since been used as a theory for multiple cancer types to refer to clonal patches of abnormal epithelium (Ha and Califano, 2003, Heaphy et al., 2009). The cells within these patches may share histological or genetic features, indicating that they are of clonal origin. This theory has also been applied to the colorectum to describe a field of abnormal bowel mucosa from which cancer may arise (Graham et al., 2011b, Galandiuk et al., 2012).

1.9.4 Clonality in the colon and rectum

The bowel epithelium is organised into functional units described as crypts whereby the epithelium forms invaginations into the underlying lamina propria. Stem cells that reside at the base of the crypt divide to maintain the stem cell niche and form progenitors. If a stem cell acquires a genetic “hit” such as a mutation or larger structural genetic change that confers a growth advantage, it will become dominant and take over the stem cell niche. This process is known as niche succession (Figure 13A) (Zeki et al., 2011). The progenitors of the stem cell replace terminally differentiated cells that are shed from the top of the crypt (Humphries and Wright, 2008, Gutierrez-Gonzalez et al., 2009). This process whereby stem cells divide and differentiate to dominate the crypt is known as monoclonal conversion (Figure 13B) (Zeki et al., 2011). Whole crypts can then further divide through bifurcation of the crypt into two new crypts in a mechanism known as crypt fission (Figure 13C) (Baker et al., 2013). As a result, the bowel is a dynamic structure, constantly undergoing micro-structural changes as well as reacting to the surrounding environment.

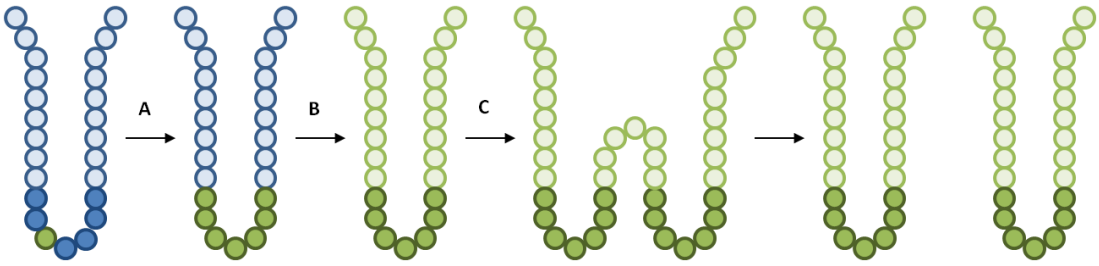


Figure 13. Crypt dynamics with normal cells (blue) and mutant (green). Darker colouring denotes stem cells. A: niche succession. B: monoclonal conversion. C: crypt fission. Adapted from (Graham et al., 2011a).

It is therefore thought that in order for a mutation to have an effect, it must occur within the stem cell niche in order to be maintained within the crypt. After a mutation in the stem cell has occurred, it will eventually be present within all cells in the crypt forming a clone. These clones can then spread throughout the bowel via crypt fission. Eventually, there is a variable replacement of the normal epithelial cells throughout the bowel with mutant clone cells and this can be widespread. These cells still appear histologically normal, but their genetic aberrations predispose them to tumour development. This process is referred to as field cancerisation (Galandiuk et al., 2012, Graham et al., 2011b). By this theory, a carcinogen can induce a mutation at a single site that has the potential to spread throughout the bowel.

1.10 Concluding remarks

Colorectal cancer is a significant cause of cancer death both world-wide and in the UK. It has a particularly bad prognosis when detected at a later stage; less than 50% 5-year survival for Dukes' C and Dukes' D. Therefore key to improving outcomes is the ability to detect cancer and precancerous changes as early as possible. This has already been seen with the introduction of bowel cancer screening (Mandel et al., 1993). In the bowel cancer screening programme in the UK, the introduction of flexisigmoidoscopy as well as FOBT has allowed for cancerous and pre-cancerous lesions to be detected at an even earlier stage in their development (Atkin et al., 2010). It could further be explored how to detect changes at an even earlier stage in the pathway of tumour progression; genetic changes, before phenotypic changes have occurred. This could have the potential to be used as a definitive test or as a marker to identify those who are more "at risk" of developing colorectal cancer.

1.11 Aims

For this thesis, the overall aims were as follows:

- Establish the sensitivity of mutation detection with pyrosequencing and next generation sequencing
- Investigate the presence of KRAS mutations in normal mucosa as well as other commonly mutated genes in colorectal cancer.
- Investigate how FAP adenomas develop and relate to each other according to their mutational profiles and copy number changes to better understand the development of early adenomas.

2 Assessing sensitivity of mutation detection

2.1 Introduction

2.1.1 KRAS mutations in normal colon epithelium

It has been observed in numerous studies that KRAS mutations may occur in histologically normal mucosa (Ronai, 1994, Minamoto et al., 1995, Ronai and Minamoto, 1997, Zhu et al., 1997, Zhang et al., 1998, Dieterle et al., 2004, Yamada et al., 2005, Kraus et al., 2006, Parsons et al., 2010). All of these studies reported mutations in cancer-associated normal mucosa and Ronai et al. also reported mutations in non-neoplastic associated normal mucosa (Ronai, 1994). These studies employed polymerase chain reaction (PCR) based methods to detect KRAS mutations in codons 12 and 13. The reported sensitivities of these methods are 0.01% to 1% mutant DNA within a wild-type background. Table 5 summarises the studies detecting KRAS mutations in normal mucosa and the reported sensitivities of the methods used.

Study	Method	Reported Sensitivity	Sample type	Results	
(Ronai, 1994)	RFLP	0.01%	Normal mucosa (normal patients)	9%	(1/11)
			Normal mucosa (CRC patients)	15%	(2/13)
			Tumour	33%	(14/42)
(Minamoto et al., 1995)	RFLP	0.01%	Normal mucosa (CRC patients)	20%	(14/70)
			Tumour	51%	(36/70)
(Zhu et al., 1997)	RFLP & DGGE	0.01% - 0.001%	Normal mucosa (CRC patients)	0%	(0/4)
			Normal mucosa adjacent – tumour	54%	(7/13)
			Tumour	100%	(22/22)
(Zhang et al., 1998)	ARMS	1%	Normal mucosa (CRC patients)	0%	(0/106)
			Normal mucosa adjacent – tumour	6%	(4/69)
			Tumour	28%	(41/149)
(Dieterle et al., 2004)	RFLP	0.01%	Normal mucosa (CRC patients)	20%	(3/15)
			Tumour	8%	(74/199)
(Yamada et al., 2005)	ARMS	1%	Normal mucosa (normal patients)	0%	(0/15)
			Normal mucosa (CRC patients)	31%	(20/65)
			Tumour	57%	(37/65)
(Kraus et al., 2006)	RFLP	0.01%	Normal mucosa (CRC patients)	15%	(21/144)
			Tumour	40%	(34/84)
(Parsons et al., 2010)	ACB-PCR	0.001%	Normal mucosa (normal patients)	17%	(1/6)
			Normal mucosa (CRC patients)	19%	(6/31)
			Normal mucosa adjacent – tumour	57%	(8/14)
			Tumour	36%	(8/22)

Table 5. Summary of studies detecting KRAS mutations in normal mucosa

2.1.2 Isolation of colonic crypts

If mutations are present in normal colon epithelial cells, they may be at low frequency. DNA from whole mucosa would include wild-type (WT) DNA from the surrounding lamina propria which could dilute the mutant signal from the epithelium. In order for relatively pure collections of colon epithelial cells to be subject to genomic analysis, it is preferable for them to be isolated from the surrounding tissue architecture. There are two main methods for colonic crypt isolation: laser-capture microdissection (LCM) (Espina et al., 2006, Humphries et al., 2011, Galandiuk et al., 2012) and mechanical dissociation (MD) in the presence of a chelating agent (Whitehead et al., 1987).

For LCM, sections of tissue are mounted onto polyethylene napthalate coated slides (PEN). A high intensity laser beam obliterates the tissue and membrane surrounding the crypt and then catapults the entire intact crypt into a collecting tube for downstream processing (Espina et al., 2006). Alternatively, fresh mucosa can be first incubated with a calcium chelating agent such as ethylenediaminetetraacetic acid (ETDA). This removes calcium from the cell-adhesion molecules and thereby releases epithelial cells from the surrounding lamina propria. Whole crypts can then be dislodged through mechanical dissociation (Cheng et al., 1984, Goodlad et al., 1991, Liu et al., 2012).

2.1.3 PCR-based mutation detection techniques

The PCR-based techniques used for detecting these mutations in normal mucosa were restriction fragment length polymorphism (RFLP) and amplification refractory mutation system (ARMS). RFLP-PCR is a 2 step PCR. The first round of PCR uses primers that introduce a base change just before the mutation site of interest, thereby creating an artificial restriction site. This is followed by incubation with a restriction enzyme, allowing for enzymatic digestion of wild-type. If a mutation is present, the restriction site will not be recognised by the enzyme and the PCR produce will evade digestion. Finally a subsequent round of PCR results in the enrichment of mutant DNA which can be visualised via gel electrophoresis (Haliassos et al., 1989). This process is schematically represented in Figure 14.

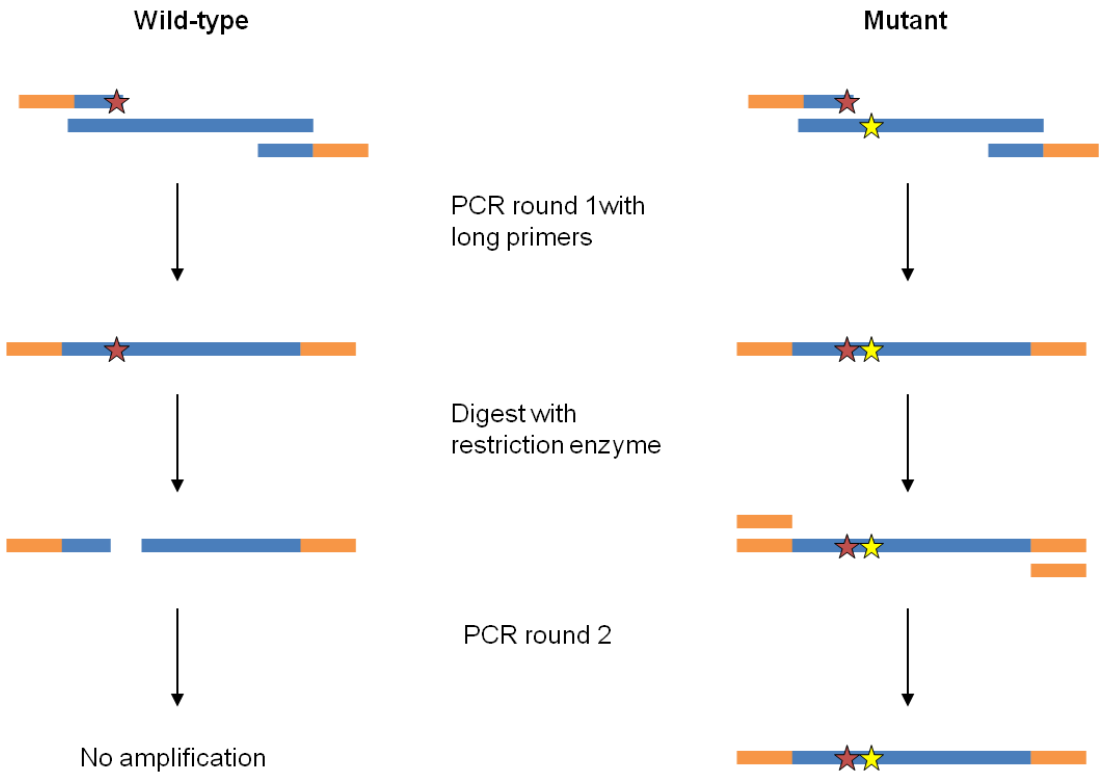


Figure 14. Restriction fragment length polymorphism polymerase chain reaction (RFLP-PCR). The yellow star denotes the base change in mutant DNA. The red star denotes the base change incorporated by the long primers used in the first round of PCR to create a restriction site.

The ARMS method consists of two complementary PCR reactions to test the DNA of interest for a specific mutation (Newton et al., 1989). Each PCR reaction uses a common reverse primer; however the forward primer is designed to specifically bind to either mutant or WT. One reaction contains a forward primer that is WT to the mutation site of interest, whilst the other uses a primer that differs at its 3' residues so that it will be complementary to mutant DNA. After amplification the PCR products are visualised with gel electrophoresis. If the sample is WT then only the WT primer reaction will show a band on the gel. If the sample is homozygous mutant, then only the mutant reaction will show a band. Finally if the sample is heterozygous mutant then both reactions will produce a band on the electrophoresis gel. A schematic of ARMS is outlined in Figure 15.

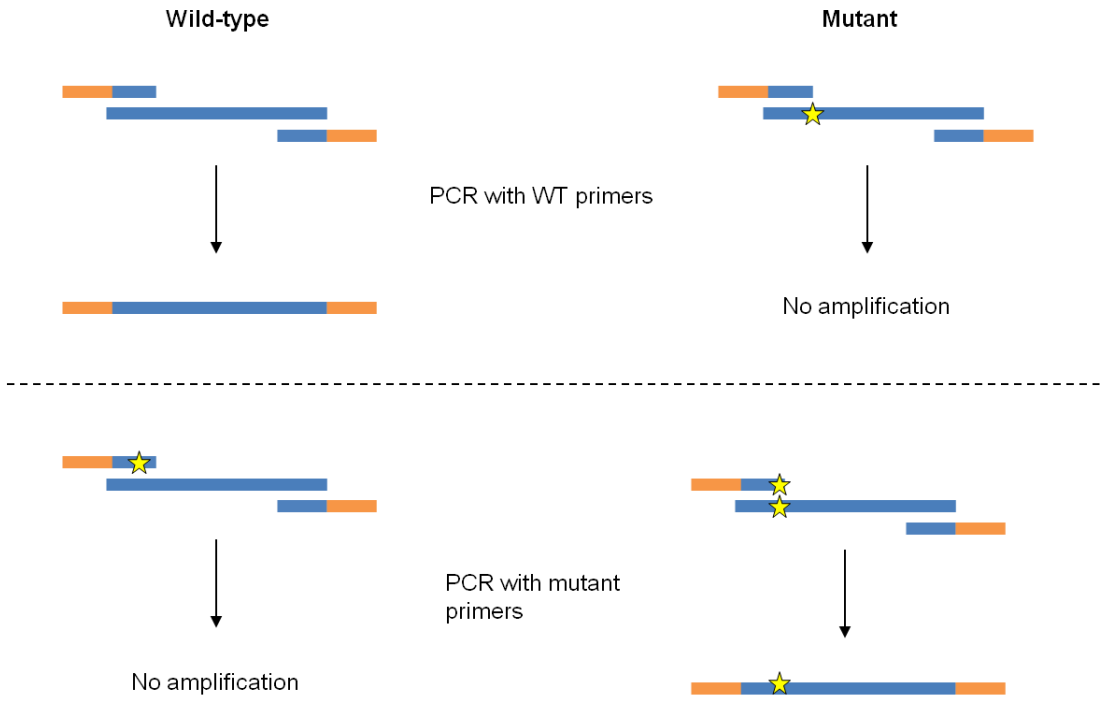


Figure 15. Amplification refraction mutation system (ARMS).

Mutations can be detected through direct sequencing of PCR products to confirm the presence of any mutation at a defined site. Sanger sequencing, the original technology for DNA sequencing has a reported detection limit of around 20% and is poorly quantitative (Tsiatis et al., 2010, Davidson et al., 2012). Pyrosequencing however, is reported to be more sensitive and allows for more accurate quantification of mutant allele frequencies (Tsiatis et al., 2010). In order to better detect minority mutant alleles, a PCR mutant enrichment technique such as Co-amplification at lower denaturation temperature (COLD-PCR) and Improved and complete enrichment co-amplification at lower denaturation temperature (Ice-COLD PCR) have been developed. These methods can have a reported limit of detection of 0.2% and 0.1% respectively (Milbury et al., 2012, Milbury et al., 2011).

2.1.4 Co-amplification at lower denaturation temperature – PCR

Co-amplification at lower denaturation temperature (COLD-PCR) is a method used to preferentially amplify the mutated sequence over WT (Li et al., 2008b). It exploits the difference in temperature that a single nucleotide mismatch causes in a heteroduplex of mutant strand and WT compared to a homoduplex. An extra step is added in the PCR whereby all the strands of DNA are heated to denature and then cooled to 70°C to encourage cross-hybridisation and the formation of heteroduplexes between the mutant strands and WT. The reaction is then heated to a critical temperature, Tc, which is higher than the melting temperature of the mutant heteroduplex but below that of the WT. This results in only the mutant heteroduplexes separating and being available for primers to anneal and thus there is preferential amplification of mutant strands over the WT which remains bound. This process is schematically represented in Figure 16.

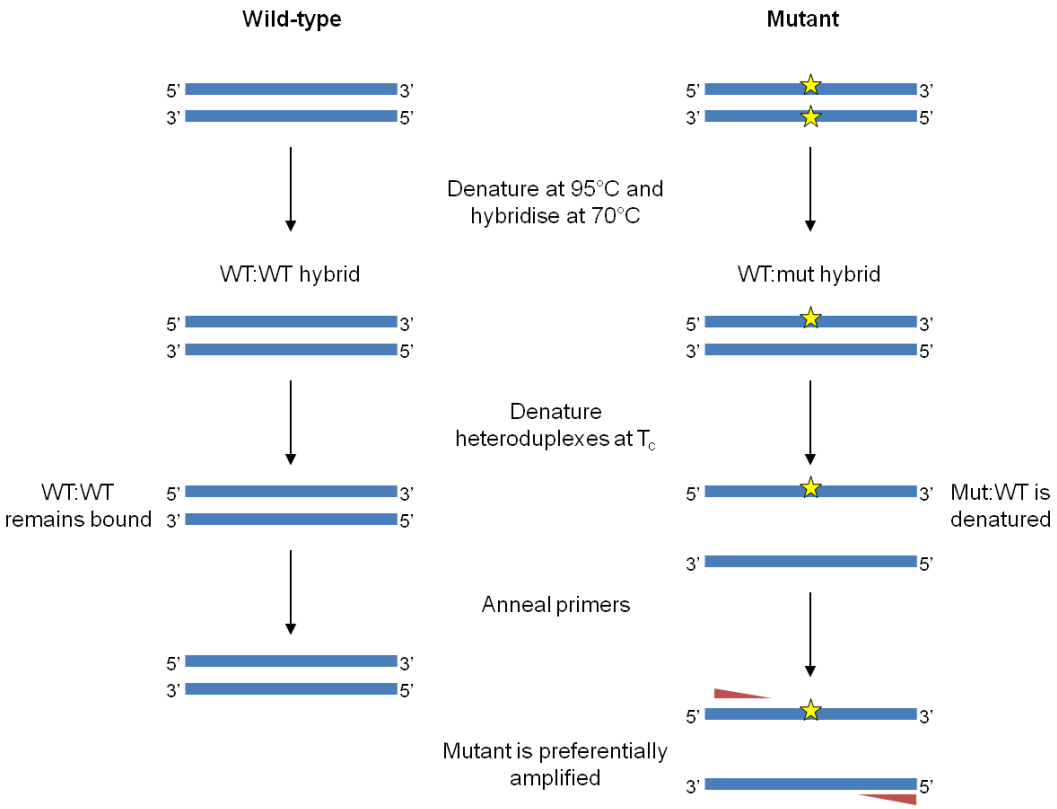


Figure 16. COLD-PCR. The yellow star represents the mutation, creating a mismatch when heteroduplexes are formed. This results in the availability of mutant template when the reaction reaches a critical temperature (T_c).

2.1.5 Improved and complete enrichment co-amplification at lower denaturation temperature (Ice-COLD) - PCR

Improved and complete enrichment co-amplification at lower denaturation temperature (Ice-COLD PCR) is a technique that has similarities with COLD-PCR, except it exploits the difference in melting temperature between heteroduplexes of the mutant and a non-extendable reference sequence (RS) oligonucleotide (Milbury et al., 2011). The RS, which is identical to the wild-type anti-sense strand, is added to the PCR reaction in excess. After heating to denature, the reaction is cooled to 70°C to encourage the formation of heteroduplexes between both the mutant strand and RS and also WT strand and RS. The heteroduplexes with the mutant DNA strand contained a mismatch and therefore have a lower melting temperature than the WT heteroduplexes. The reaction is then heated to a critical temperature Tc; higher than the melting temperature of the mutant heteroduplex, but below that of the wild-type. Mutant template was then more widely available for primers to anneal and be preferentially amplified over the WT as shown in Figure 17.

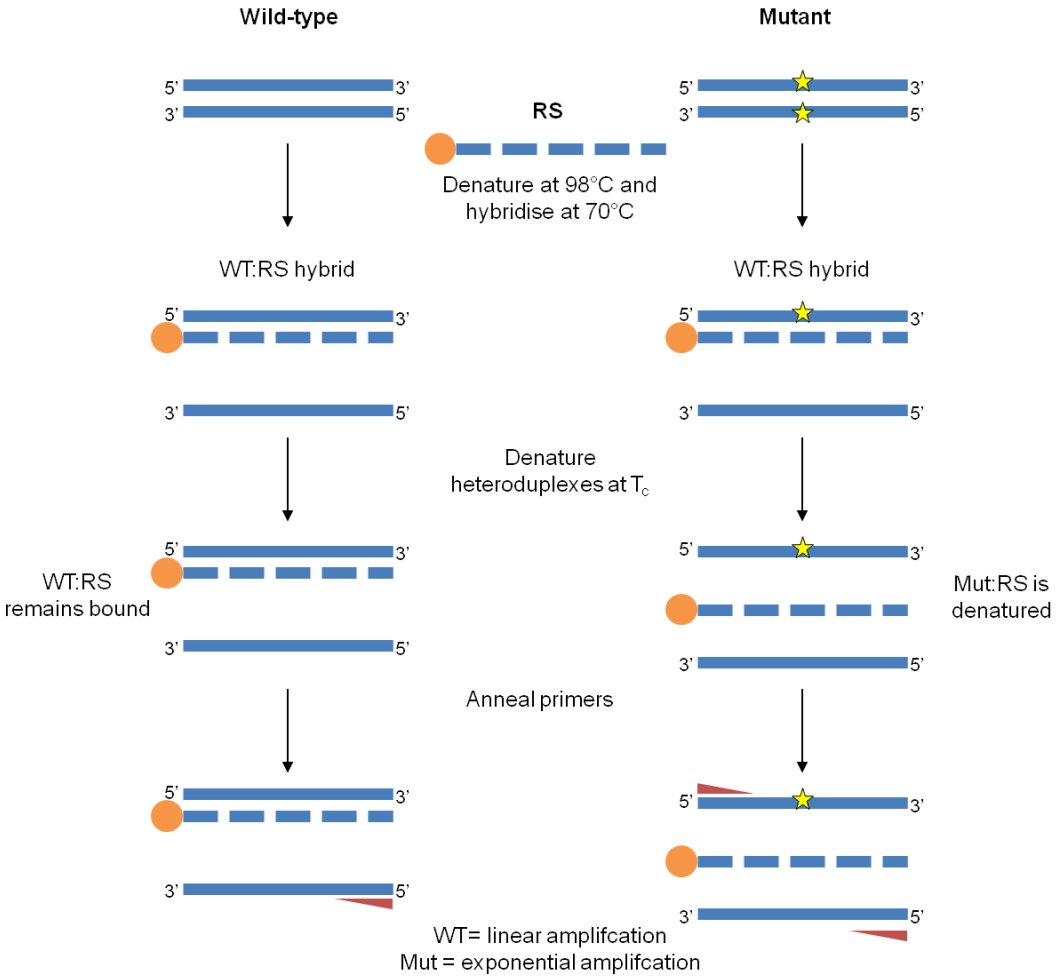


Figure 17. Ice-COLD-PCR. The yellow star represents the mutation, creating a mismatch when heteroduplexes are formed with the reference sequence (RS).

2.1.6 Pyrosequencing

Pyrosequencing is a sequencing-by-synthesis technology that works on the detection of light emitted as the result of nucleotides being incorporated into the DNA sequence. As each base is added through the actions of the polymerase, pyrophosphate (Ppi) is released in equimolar amounts. This generates adenosine triphosphate (ATP) from adenosine 5´-phosphosulfate (APS), which then in turn converts luciferin to oxyluciferin via the actions of the luciferase enzyme. This final reaction causes a measurable burst of light which is directly proportional to the amount of nucleotide incorporated into the DNA sequence. Unincorporated nucleotides and excess ATP is broken down by the apyrase enzyme which allows the process to repeat for the next nucleotide in the sequence. The light generated by the enzyme reactions from the sequencing is detected by a camera, interpreted by the computer and displayed visually as a pyrogram which can then be used to read the DNA sequence and detect changes (Ronaghi, 2001). This overall sequence is illustrated in Figure 18.

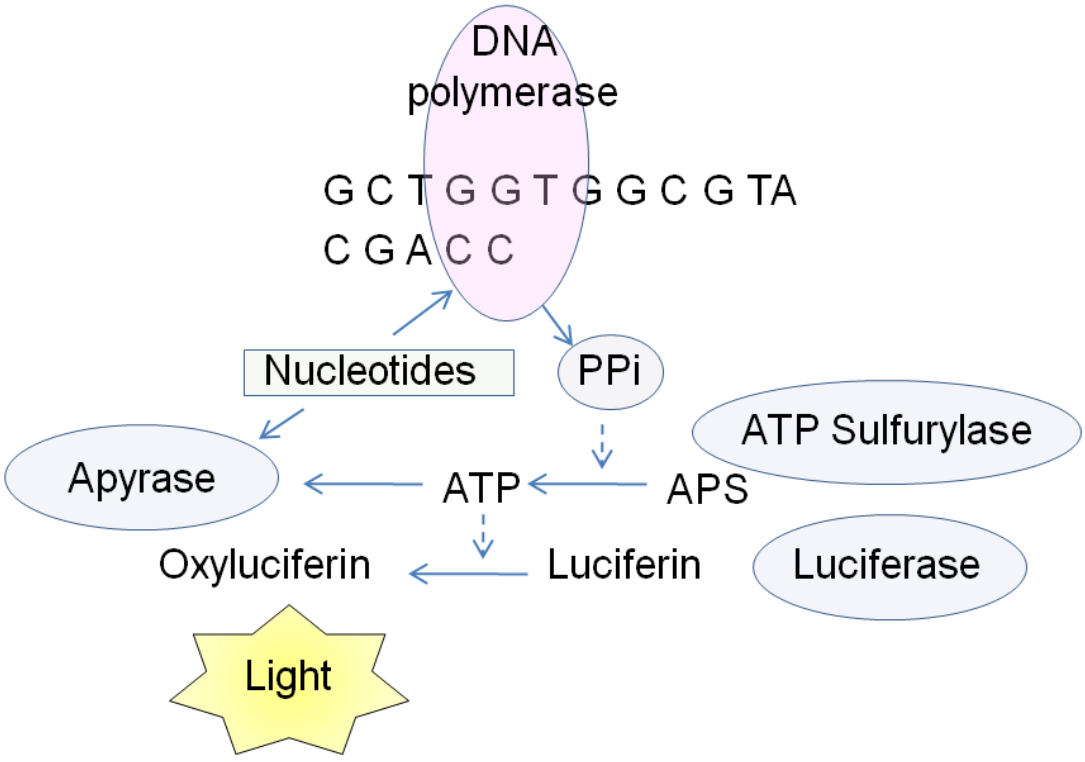


Figure 18. The reactions involved in pyrosequencing adapted from (Ronaghi, 2001).

2.2 Chapter aims

The aims of this chapter are as follows:

- The development of methods for the isolation of DNA from normal colonic mucosa.
- To investigate the value of PCR based mutant enrichment techniques
- To determine the sensitivity of pyrosequencing

2.3 Chapter Methods

2.3.1 Preparation of fresh-frozen tissue

Specimens were collected from anonymous surgical resections for colorectal cancer and samples of the tumour and normal tissue from the resection margin were taken. These samples were then flash frozen in liquid nitrogen, mounted on optimal cutting temperature embedding medium (OCT) (Leica, Milton Keynes, UK) and stored at -80°C. Between 6 and 10 cryosections of 7µm thickness were taken for DNA extraction from both tumour and normal samples from each case. Two methods of colonic crypt isolation were investigated: LCM and MD in comparison to DNA extraction from whole macrodissected mucosa.

2.3.2 Ethical approval

Ethical approval for the collection, storage and use of patient material for these samples was obtained from the South Yorkshire Research Ethics Committee REC reference: 10/H1310/61.

2.3.3 Spot counting

In order to determine the proportion of mucosa consisting of epithelial cells Spot counting analysis (West et al., 2010) was undertaken as outlined in section 2.3.3. For each sample to undergo spot counting, 3 cryosections were taken and mounted onto glass sides, stained with H&E and then were digitally scanned at 20x magnification. Using ImageScope v10 software (Aperio Technologies Inc., Vista, CA) the mucosa was identified and a 300-point grid was applied to this area (Figure 19). A code was applied to each point that was counted as seen in Table 6. The epithelial component percentage was calculated from the percentage of points coded as “1” amongst the total number of informative points counted. A final average was taken from all 3 sections from each sample.

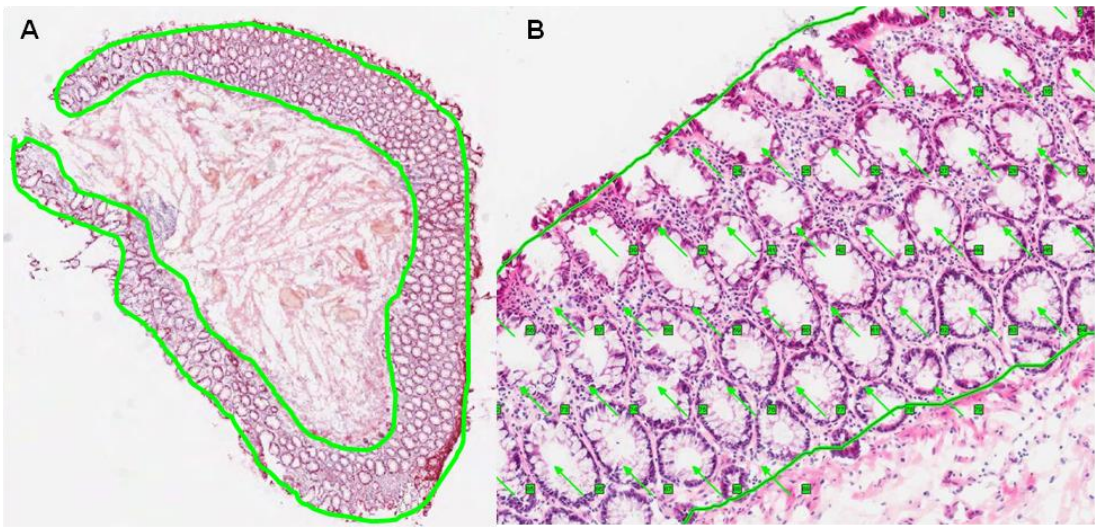


Figure 19. Spot-counting analysis on normal mucosa. A: The area of mucosa is selected by freehand. B: Random spots applied to selected area for counting.

Component	Numerical code
Non-informative	0
Epithelium	1
Lamina propria	2
Necrosis	3
Vessels	4
Inflammation	5
Crypt lumen	6
Mucin	7
Muscle	8

Table 6. Coding of tissue components for spot-counting analysis

2.3.4 Laser-capture microdissection (LCM)

Cryosections of 7µm were taken onto glass slides that were not superfrost in order to reduce adhesion onto the slide and improve capture of material. Sections were also taken onto membrane-coated polyethylene naphthalate (PEN) slides for comparison of laser-capture methods. For downstream analysis a minimum yield of 20ng/µl gDNA required in a volume of 10-20µl (a minimum of 200ng of gDNA in total). Using the assumption that each diploid cell contains approximately 6pg of DNA:

$(200x10^{-9}) / (6x10^{-12}) \approx 30{,}000$ cells

It was visually estimated that a cross sectional crypt contains around 50 cells therefore at least 600 crypts are required to obtain 200ng of gDNA. It is expected that material will be lost at every stage in the extraction process and therefore 1000 crypts per sample were captured to ensure that a sufficient amount of DNA could be obtained.

Cryosections were stained before LCM with the following protocol:

- 2 min Mayer's haematoxylin
- Rinse in distilled H_2O
- 2 min Scott's tap water
- Rinse in distilled H_2O
- 1 min eosin
- Rinse in distilled H_2O
- 1 min 70% ethanol

2.3.5 Laser settings

Two laser-capture functions were tested on the sections. The first used a defocused UV laser to fire multiple pulses over a pre-defined area on a glass-mounted slide. This catapults the tissue at discrete points into the collecting tube and destroys the tissue architecture as shown in Figure 20. However maintaining tissue integrity is not a priority for the purposes of DNA extraction.

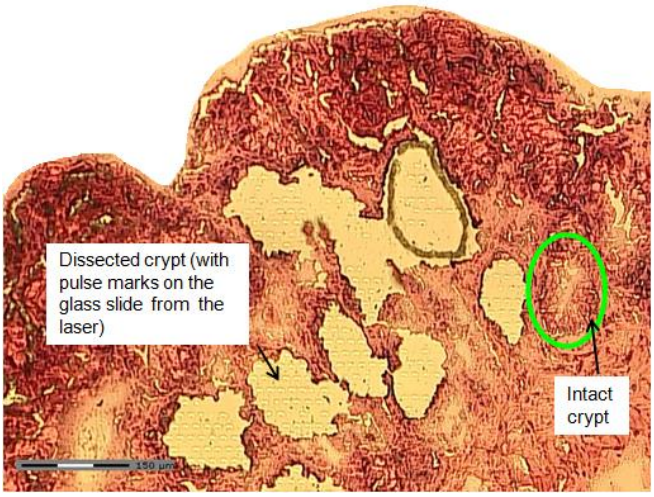


Figure 20. Laser pressure catapulting

The second method tested was used with PEN slides and utilises a high intensity beam that obliterates the tissue and membrane around the crypt and then catapults the intact crypt adhered to the membrane, into the collecting tube. This allows for the crypts to be visualised in the cap of the collecting tube to confirm it has been captured and the tissue architecture is maintained (Figure 21).

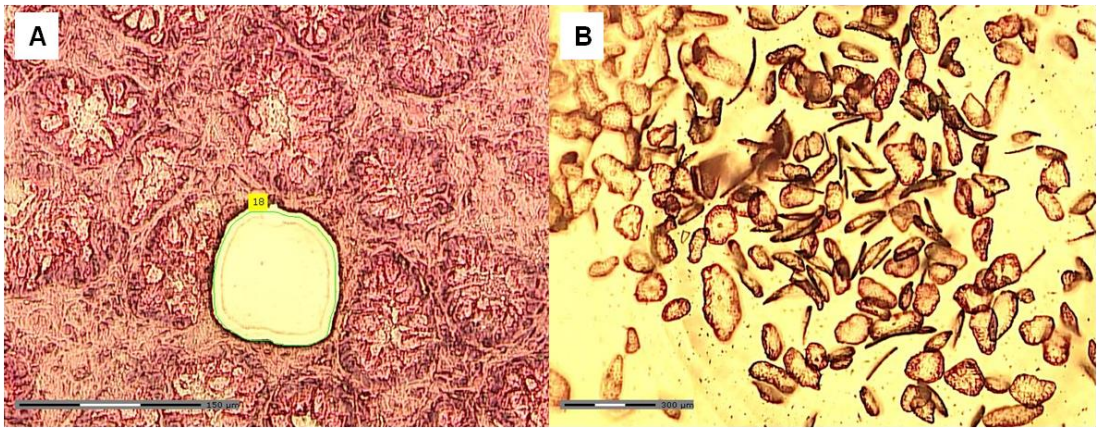


Figure 21. A: Laser-dissected crypt from PEN slide. B: Collecting tube cap showing captured crypts.

2.3.6 Mechanical dissociation of crypts

Colon epithelial cells may also be isolated from the surrounding tissue through incubation with a calcium chelating agent followed by dislodging crypts through mechanical dissociation (Cheng et al., 1984, Goodlad et al., 1991). The following protocol was adapted from the method as described by Cheng et al. (Cheng et al., 1984)

Frozen samples were thawed and washed in Hank's buffered salt solution (HBSS) (Invitrogen, UK) to remove OCT. The samples were subsequently transferred to cold (4°C) HBSS calcium and magnesium-free (CMF) (Invitrogen, UK) and washed twice. After the washing steps they were transferred to fresh HBSS CMF with EDTA and incubated for 45 mins at 37°C (see appendix for recipe). Next the samples were washed again in HBSS-CMF before being transferred to fresh HBSS-CMF and underwent mechanical disruption by use of a magnetic stirrer for 2 hours at 4°C. Finally the solution was spun at 750g for 5 mins at 4°C and the pellet removed. A sample was taken for visual confirmation that crypts had been isolated (Figure 22) and the remaining pellet was used for DNA extraction with Qiagen QIAamp DNA Mini Kit (QIAGEN, UK). The remaining tissue was fixed in formaldehyde and embedded in paraffin wax, before sections were taken and stained with H&E (section 2.3.7) to visualise the crypt dissociation (Figure 23).

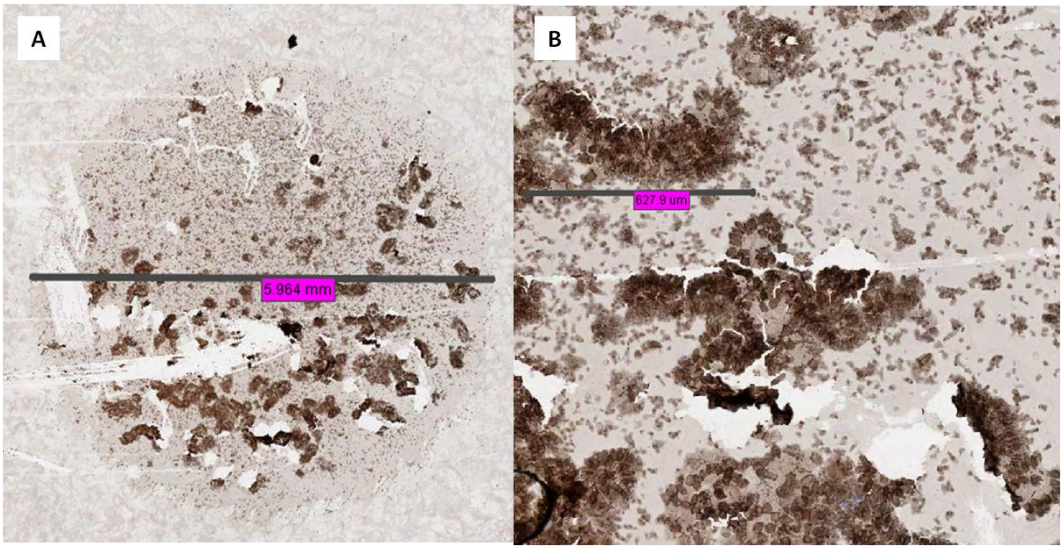


Figure 22. Cytokeratin immunohistochemistry stain to confirm epithelial cells. A: Cytospin of pellet from mechanical dissociation. B: Isolated colonic crypt

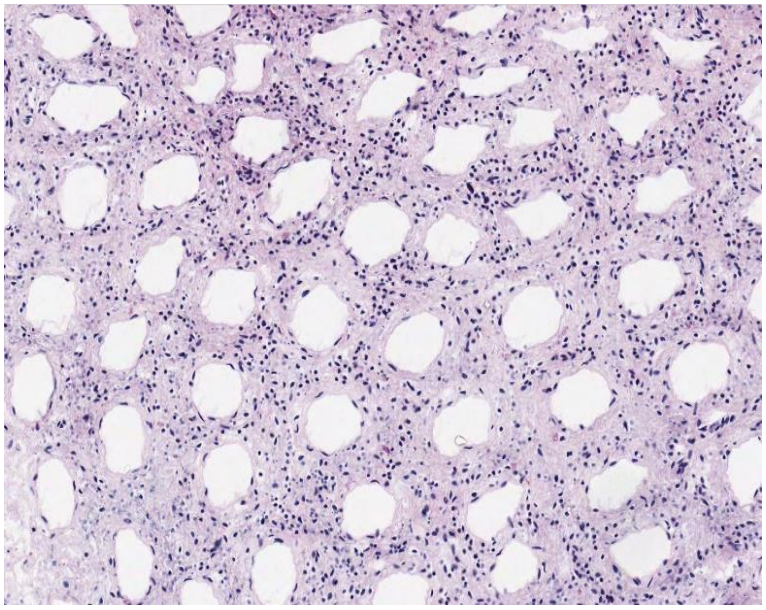


Figure 23. H&E section of mucosa after mechanical dissociation of crypts

2.3.7 Preparation of formalin fixed paraffin embedded tissue

10-20 5µm sections were taken from formalin fixed paraffin embedded (FFPE) tissue blocks for DNA extraction. The sections were de-waxed for 12 mins in xylene, followed by 12 mins in ethanol and finally 9 mins in graded ethanols before being transferred into distilled water for extraction. A further section from each sample was cut and stained to guide macro-dissection for DNA extraction. The staining protocol used was as follows:

- 2 mins Mayer's haematoxylin
- 1 min rinse in water
- 2 mins Scott's tap water
- 1 min rinse in water
- 2 mins eosin
- 1 min rinse in water
- 12 mins ethanol
- 9 mins xylene

Finally cover slips were mounted onto the slides with Di-N-butyl phthalate in xylene (DPX) (CellPath, Powys, UK).

2.3.8 Macrodissection of normal mucosa

H&E sections for each case to undergo DNA extraction were assessed using light microscopy to identify an area of histologically normal mucosa. These areas were then annotated on the H&E slide. De-waxed sections were then macrodissected using the annotated H&E slide as guidance for the area of tissue to be used for extraction. Macrodissection was performed using a needle and/or scalpel blade for each section using a dissecting microscope for visual guidance. Figure 24 shows an example of a section before and after macrodissection.

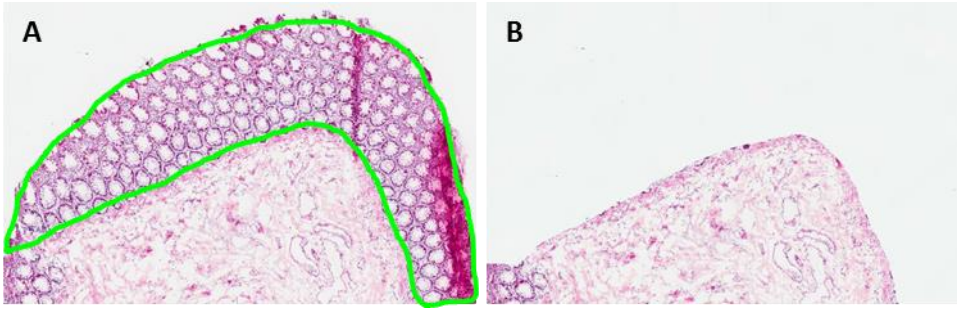


Figure 24. Macrodissection of H&E section. A Before macrodissection with area to be removed for extract highlighted in green. B section after macrodissection.

2.3.9 DNA extraction

For samples where FFPE sections had been taken, they were macro-dissected before undergoing DNA extraction with the Qiagen QIAamp Micro Kit (QIAGEN, Crawley, UK). This was done in accordance with the standard protocol with the modification of an over-night digestion step and a final elution volume of 25µl H_2O.

DNA from crypts isolated from fresh-frozen tissue by MD was extracted with the Qiagen QIAamp DNA Mini Kit (QIAGEN, Crawley, UK) in accordance with the standard manufacturer's protocols with an elution volume of 50µl. This was due to the higher expected yield of DNA from this method.

All extracted DNA was tested on the nanodrop spectrophometer (Thermo Scientific, Washington, USA). This provided an estimate of the DNA concentration as well as the quality by providing absorbance ratios. The 260:280 ratio which measures the amount of absorbance of light at wavelengths of 260nm and 280nm was used as a measure of DNA purity. A value in the range of 1.8-2.0 was taken as optimal.

2.3.10 Cell line DNA

DNA from cell line SW48 of known concentration was provided by Horizon Discovery (Horizon Discovery, Cambridge, UK) containing an engineered heterozygous KRAS mutation. Both wild-type (WT) DNA and mutant DNA with one of 6 KRAS 12 and 13 mutations were provided as shown below in Table 7.

Mutation	cdna notation	Amino acid change
G12A	c.35G>C	glycine – alanine
G12C	c.34G>T	glycine – cysteine
G12D	c.35G>A	glycine – aspartic acid
G12R	c.34G>C	glycine – arginine
G12V	c.35G>T	glycine – valine
G13D	c.38G>A	glycine – aspartic acid

Table 7. KRAS codon 12 and 13 mutant DNA provided by Horizon Discovery (Horizon Discovery, Cambridge, UK) for testing sensitivities of mutation detection technologies.

2.3.11 Polymerase chain reaction amplification of KRAS

Primers were designed using pyrosequencing assay design software (Biotage AB, Uppsala, Sweden) as outlined by Dr. Phil Chambers et al. (Chambers et al., 2013). Primers were designed for KRAS codons 12 and 13 to create an 80bp amplicon:

Fwd: 5'GGCCTGCTGAAAATGACTGA3'

Reverse: 5'AGCTGTATCGTCAAGGCACTCT3'

KRAS Pyrosequencing primer: 5'Biotin-AAACTTGTGGTAGTTGGA 3'

For standard PCR reactions, HotStarTaq Master Mix from Qiagen (QIAGEN, Crawley, UK) was used to create a reaction volume of 25µl or 50µl for downstream sequencing (Table 8).

Component	Volume per reaction		Final concentration
	25µl reaction	50µl reaction	
HotStarTaq Master Mix	12.5µl	25µl	I unit HotStarTaq DNA polymerase 1xPCR buffer* 200µM of each dNTP
Forward Primer (25µM)	0.2µl	0.4µl	0.2µM
Reverse Primer (25µM)	0.2µl	0.4µl	0.2µM
$MgCl_2$ (25mM)	0.5µl	1µl	2mM
H_2O	9.6 µl	19.2µl	
gDNA (10ng/ul)	2µl	4µl	

Table 8. PCR reaction composition with the use of Taq polymerase enzyme. * buffer contains 1.5mM $MgCl_2$.

The thermocycling conditions for KRAS primers with the use of HotStarTaq were as follows:

- 95°C for 12 mins
- 40 cycles of:
- 94°C for 10 secs
- 55°C for 20 secs
- 72°C for 20 secs

- Hold at 4°C

All PCR products amplified were visualised by gel electrophoresis. A 2% agarose gel was made by dissolving agarose in Tris-EDTA buffer (TE) (Promega, Madison, USA) by heating and allowing it to set. 4µl of PCR product with 1µl of Orange G loading buffer were loaded and the gelrun for 30 mins at 120mV before visualisation. The remaining PCR product was cleaned up using the Qiagen MinElute kit (QIAGEN, Crawley, UK), eluting in 20µl. Finally quantification of the PCR products was performed by use of a Quant-iT PicoGreen assay (Invitrogen, Paisley, UK) following the standard protocol.

Pyrosequencing was performed by Dr. Phil Chambers and Miss Morag Taylor on a PyroMark Q96 ID (QIAGEN, Crawley, UK). All sequencing was conducted according to the manufacturer's protocols.

2.3.12 Restriction fragment length polymorphism PCR

An adapted protocol was developed in order to enrich mutations in KRAS codon 12 to be visualised with gel electrophoresis as well as subsequent pyrosequencing to confirm the mutation (Ronai and Minamoto, 1997). The primers used were the same as those in the original protocol (Ronai and Minamoto, 1997) and outlined in Table 9.

.

PCR stage	Primer Sequences
Round 1	Fwd: 5'**GCGGTTGGGGCTTAATTGCA**TATAAACTGAATATAAACTTGTGGTAGTTGGACCT 3´
	Rev: 5'**GCTGTTGTCATAGTAATGAT**TCAAAGAATGGTCCTGCACCAG
Round 2	Fwd: 5' GCGGTTGGGGCTTAATTGCA
	Rev: 5' GCTGTTGTCATAGTAATGAT

Table 9. Primers used in wild-type restriction digest PCR. The underlined sequence of the first round primers binds to the gDNA target. The shorter primers used in the second round bind to the long tail of the first round PCR primers, indicated by sequences in bold font.

A 25µl PCR reaction with Qiagen HotStarTaq (section 2.3.11) underwent a first round of PCR with the following thermocycling protocol:

- 95°C for 12 mins
  20 cycles of:
- 94°C for 1min
- 55°C for 1min
- 72°C for 1min

- Hold at 4°C

Following the initial round of PCR, PCR products were incubated with one unit of ScrFI enzyme (New England Biolabs, UK) at 37°C with 10xNEB buffer 3 and Dnase free H2O. The restriction sites for ScrFI are shown in Figure 25.

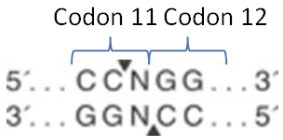


Figure 25. Restriction site for ScrFI enzyme. N= G,T,A or C.

After the initial digestion step samples were then subject to purification using QIAquick PCR Purification Kit (QIAGEN, UK) to remove previous PCR and digestion buffers. A 12µl aliquot was used as the template for the second round of PCR. The same reagents and reaction conditions were used for the second round of PCR. After this stage a second digestion step was performed to ensure that all wild-type KRAS 12 had undergone digestion. The digestion products were then run on a 2% agarose gel and were gel purified to isolate the 196bp product for sequencing.

2.3.13 COLD-PCR method

The reaction mix for COLD-PCR was identical to that used for standard PCR (Table 5) with taq polymerase (section 2.3.11) underwent the following thermocycling protocol:

- 95°C for 12 mins

10 cycles of:

- 95°C for 10sec
- 55°C for 10sec
- 72°C for 10sec

30 cycles of:

- 95°C for 10sec
- 70°C for 8 mins
- T_c for 10sec
- 55°C for 10sec
- 72°C for 10sec

- Hold at 4°C

The critical temperature, T_c varies for each primer paired used and was determined by producing melting curves of mutant:wild-type heteroduplexes (see section 2.4.5).

2.3.14 Ice-COLD-PCR method

Again, the same reaction mix as before (Table 8 section 2.3.11) for Taq polymerase was used for Ice-COLD-PCR. The RS was used as described by Milbury et al. (Milbury et al., 2011):

5'CTCTTGCCTACGCCACCAGCTCCAACTACCACAAGTTTATATTCAG-phosphate3'

The phosphate group at the 3' end prevents extension by the polymerase during the PCR. The thermocycling protocol for Ice-COLD-PCR was as follows:

- 98°C for 30 sec

5 cycles of:

- 98°C for 10 sec
- 59°C for 20 sec
- 72°C for 10 sec

- 35 cycles of:
- 95°C for 10 sec
- 70°C for 30 secs
- T_c for 10 sec
- 59°C for 10 sec
- 72°C for 10 sec

- Hold at 4°C

The critical temperature, T_c, was determined by the melting curves of mutant:wild-type heteroduplexes (see section 2.4.6).

2.3.15 Analysis of pyrograms

Pyrosequencing produces the peak heights for each nucleotide that is added to the sequence and can therefore produce quantitative results. By comparing the intensity of the mutant peak heights to the other peak intensities in the pyrogram the percentage of mutant in the sample can be calculated. However, the calculated mutant percentage has to be interpreted in the context of the background peaks seen in the wild-type, particularly at low mutant allele frequencies. Each of the possible 6 KRAS codon 12&13 mutations has to be interpreted individually and the formulae were as follows:

G12A c.35G>C mutation

$$\frac{\text{C peak height [position 35]x 2}}{\text{G peak height [position 34]} + \text{C peak height [position 35]}}$$

G12C c.34G>T mutation

$$\frac{\text{T peak height [positions 33 and 24]} - (\text{(average of all single peak heights)x2})}{\text{G peak heights [positions 34 and 35]} + (\text{T peaks heights [positions 33 and 34]} - \text{(average ofl single peak heights)}}$$

G12D c.34G>T mutation

$$\frac{\text{(A peakheight [position35]x 0.9)x 2)}}{\text{G peak height [position 34]} + \text{A peak height [position 35]}}$$

G12R c.34G>C mutation

$$\frac{\text{C peak height [position 34]x 2}}{\text{C peak height [position 34]} + \text{G peak height [position 35]}}$$

G12V c.35G>T mutation

$$\frac{\text{T peak heights [positions 35 and 36} - (\text{(average of single peak heights) x2})}{\text{G peak height [positions 24 and 35]} + (\text{T peak height [positions 35 and 36]} - \text{average of all single peak heights}}$$

G13D c.38G>A mutation

$$\frac{\text{(A peak height [position 38]x 0.9)x 2}}{\text{G peak height [position 37 and 38]} + \text{A peak height [position 38]}}$$

2.3.16 Statistical methods

To compare yields of DNA obtained through different methods of crypt isolation, the median and interquartile ranges were calculated as the data was not assumed to fit a normal distribution. P values were calculated by the Mann-Whitney U test to compare yields from different colonic crypt isolation methods. All statistics were carried out using the Statistical Package for the Social Sciences (SPSS version 20, Chicago, USA).

2.4 Results

2.4.1 Spot counting analysis

Spot-counting analysis was performed on 3 samples of sections of normal mucosa in order to determine the percentage of the mucosa that consists of epithelium. A median percentage of 51.7% was determined as epithelial cells (Table 10) (for full data see appendix section 6.2).

Sample	Number of sections	Number of spots counted	Number of spots on epithelium	Median % epithelium	Interquartile range (%)
1	5	1437	712	49.8	47.3 – 51.5
2	5	1377	745	51.8	40.5 – 54.0
3	4	1086	523	46.2	40.1 – 52.9
Overall median percentage epithelium				51.7	47.9 – 54.1

Table 10. Summary of spot counting analysis

2.4.2 Crypt isolation techniques

A comparison of DNA yield was made between DNA extraction by microdissection of whole mucosa and two colonic crypt isolation techniques: laser-capture microscopy (LCM) and mechanical dissociation (MD).

For whole mucosa DNA extraction was performed on 5 different frozen samples of normal colorectal mucosa. Each sample had six 7µm sections taken for extraction. Extraction buffer without any tissue was used as a negative control. Scraping off approximately 5mm^2 of mucosa per section with the aid of a dissection microscope from H&E stained frozen sections produced a consistently high yield of DNA from extraction. The median yield obtained was 149ng/µl (interquartile range 131.66ng/µl – 158.97ng/µl) compared to the negative control 3.56ng/µl (interquartile range 3.02ng/µl – 7.43ng/µl). This difference was statistically significant (Mann-Whitney p value=0.008). The yield of DNA obtained from this method is shown in Table 11.

Sample	Quantity of DNA extracted (ng/µl)	
	Normal mucosa	Negative control
1	149.00	7.43
2	131.66	9.95
3	188.97	3.02
4	158.97	2.74
5	130.53	3.56
Median	149.00	3.56
Interquartile range	131.66 – 158.97	3.02 – 7.43

Table 11. DNA yield from extraction of whole mucosa

LCM was performed on 5 samples of normal mucosa. 3-4 sections of each sample were dissected until approximately 1000 crypts had been captured. The median DNA yield after extraction was 8.58ng/µl compared to a negative control of 2.81 (Mann-Whitney p value=0.095) (Table 12).

Sample	Quantity of DNA extracted (ng/µl)	
	Normal mucosa	Negative control
1	3.06	1.84
2	22.9	2.81
3	3.56	1.66
4	8.58	4.16
5	33.11	4.65
Median	8.58	2.81
Interquartile range	3.56 – 22.90	1.84-4.16

Table 12. DNA yield from extraction of crypts isolated by laser-capture microscopy

Finally 5 samples of normal colon mucosa of dimensions ~10mm x 5mm x 0.45mm underwent MD to isolate the colonic crypts. A median yield of 522.48ng/µl was obtained via this method which was statistically significant from the negative control of 3.02ng/µl (Mann-Whitney p value<0.0001) (Table 13).

Sample	Quantity of DNA extracted (ng/µl)	
	Normal mucosa	Negative control
1	151.95	3.02
2	133.52	2.18
3	1194.46	9.61
4	1046.15	1.53
5	522.48	7.78
Median	522.48	3.02
Interquartile range	151.95 – 1064.15	2.18 – 7.78

Table 13. DNA yield from extraction of crypts isolated by mechanical dissociation

Table 14 gives a summary of the DNA yields that were obtained from the two crypt isolation techniques in comparison to extraction from whole mucosa. This data is also displayed in Figure 26 to show the range and IQR for the different techniques. MD produced a significantly higher yield of DNA than LCM: a median yield of 149.00ng/µl compared to 8.58ng/µl for LCM, $p<0.05$.

Method	Median DNA yield (ng/µl)	Negative control (ng/µl)	IQR	Range
Whole mucosa	149.00	3.56	131.66 – 158.97	130.53 – 188.97
	p = 0.008			
LCM	8.58	2.81	3.56 – 22.9	3.06 – 33.11
	p = 0.095			
MD	522.48	3.02	151.95 – 1064.15	131.52 – 1194.46
	p <0.0001			

Table 14. Comparison of median DNA yield and interquartile range (IQR) from crypt isolation techniques laser-capture microscopy (LCM) and mechanical dissociation (MD) compared to whole mucosa extraction. P values from Mann-Whitney statistical analysis.

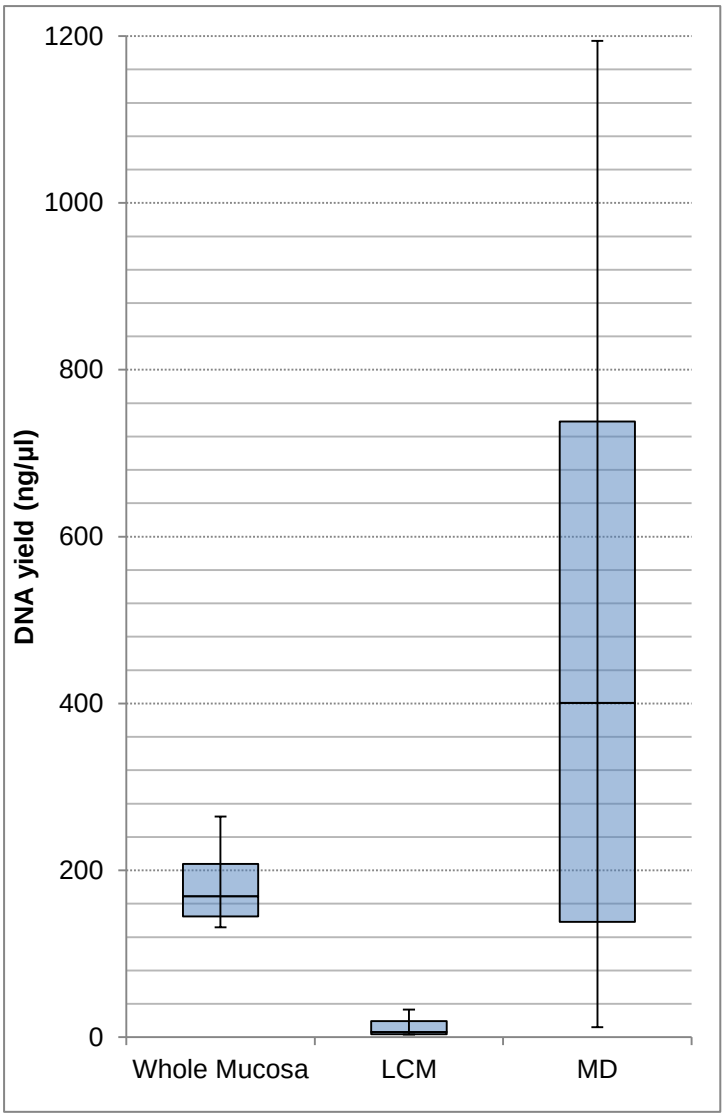


Figure 26. Box and whisker plot comparing medians and interquartile ranges of DNA yield from crypt isolation techniques laser-capture microscopy (LCM) and mechanical dissociation (MD) and whole mucosa extraction.

At low DNA concentrations, nanodrop readings can be subject to interference from the extraction buffers. However, the amount of DNA obtained through MD was significantly higher than the negative control, indicated by $p<0.0001$. This was confirmed by visualising the DNA on an agarose gel (Figure 27).

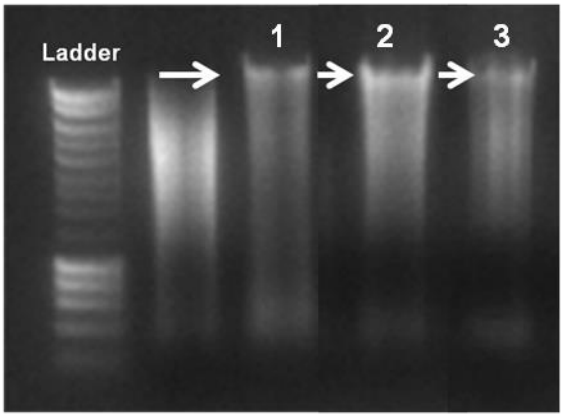


Figure 27. 1.5% agarose gel confirming extraction of DNA through MD from three samples of fresh normal mucosa. The recovery of high molecular weight DNA is indicated by arrows. The first lane shows a more fragmented sample for comparison.

For the three samples shown in Figure 27 the proportion of DNA recovered at high molecular weight (10,000kb and above) was 38%, 28% and 31% respectively.

2.4.3 Pyrosequencing of serial dilutions of mutant *KRAS*

Pyrosequencing serial dilutions of 6 different KRAS 12 and 13 mutations found the limit of detection of the pyrosequencer to lie between 5 and 25% (Table 15) as there was variation in sensitivity with each run. Two runs were performed on all 6 mutations and an extra run was performed on just 2 of the mutations (12A c.35G>C and 12C c.34G>T) due to limited amounts of control mutant DNA available. Figure 28 shows the pyrograms from one run and repeat runs produced comparable programs (see appendix).

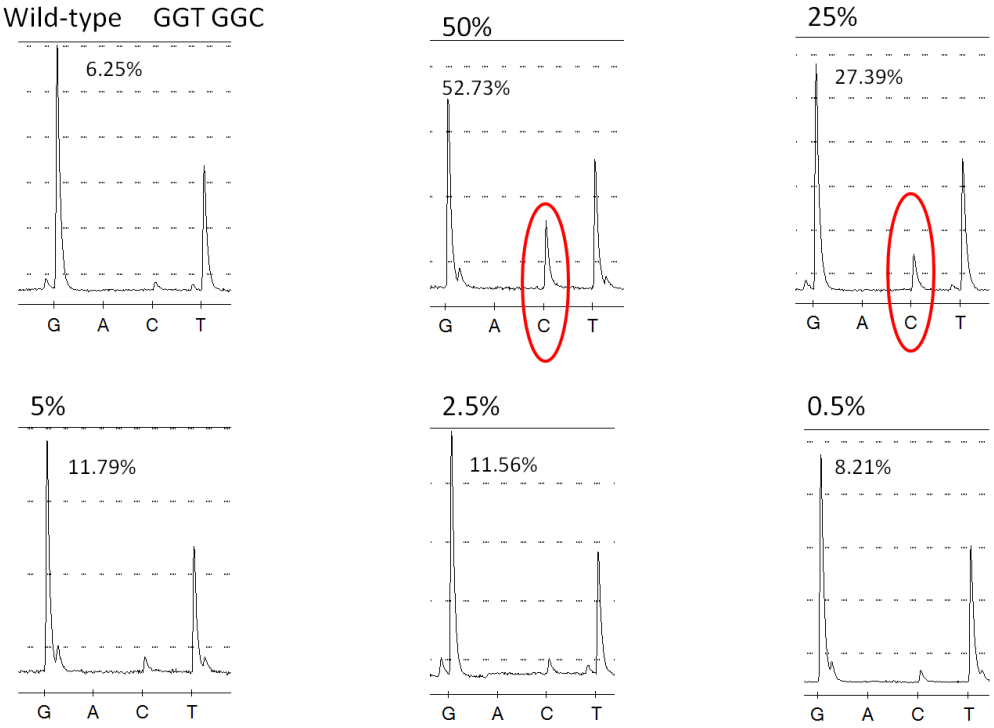
	Limit of Detection		
Mutation	**Run1**	**Run 2**	**Run3**
G12A c.35G>C	5%	25%	5%
G12C c.34G>T	25%	25%	25%
G12D c.35G>A	5%	25%	-
G12R c.35G>C	5%	5%	-
G12V c.35G>T	5%	25%	-
G13D c.38G>A	25%	5%	-

Table 15. Limit of detection of Pyrosequencing for 6 different KRAS codon 12 and 13 mutations

Figure 28. A-F: Pyrograms of serial dilutions of 6 KRAS codon 12 and 13 mutations. The percentage shown is calculated from the peak height intensities. Solid circled peaks denotes detected mutation and dotted-line circles denote peaks on the borderline of detection.

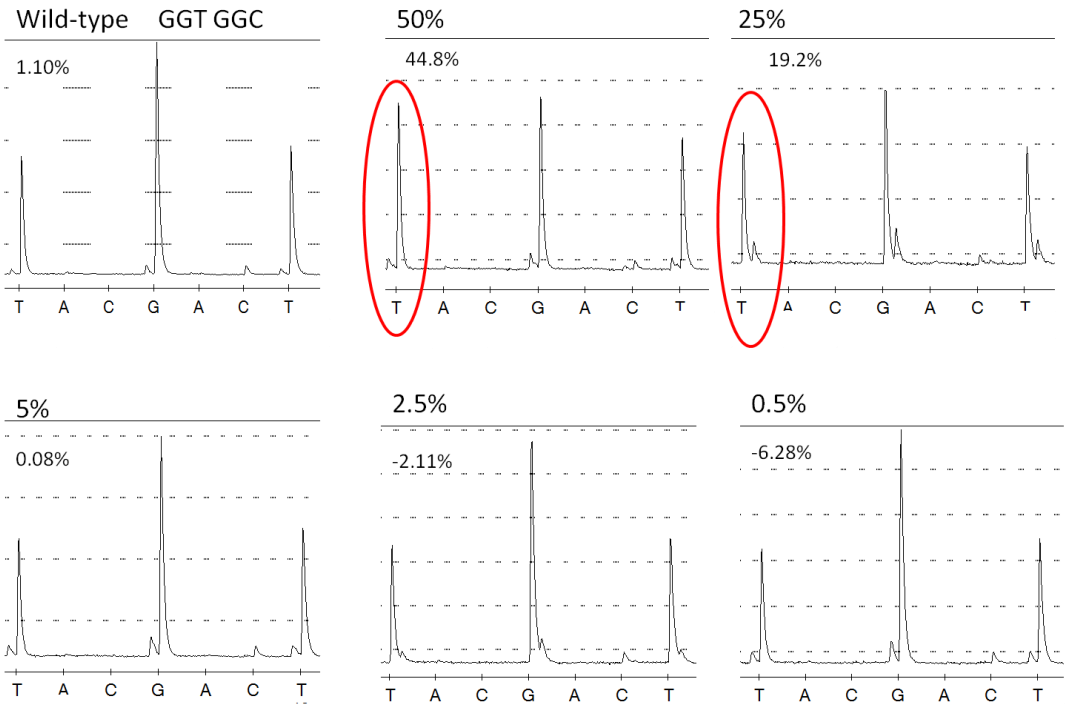
A.

G12A c.35G>C GCT GGC



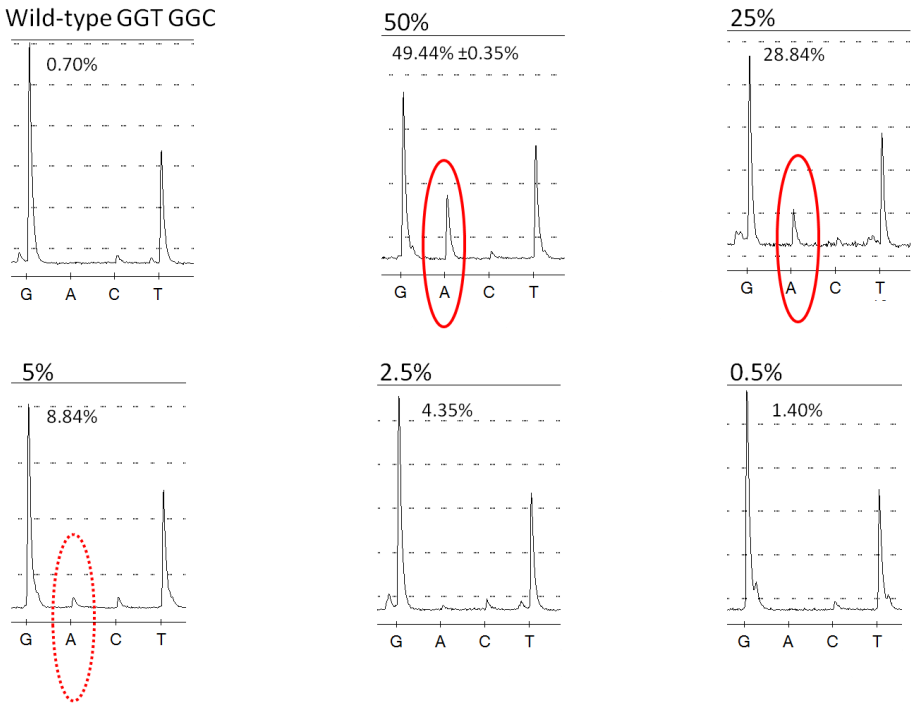
B.

G12C c.34G>T TGT GGC



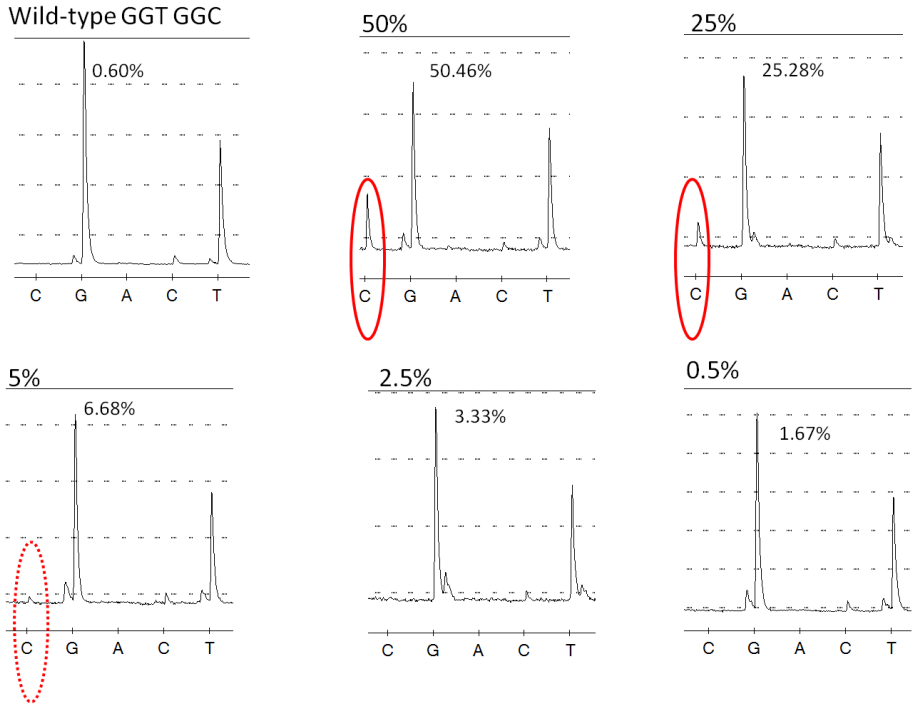
C.

G12D c.35G>A GAT GGC



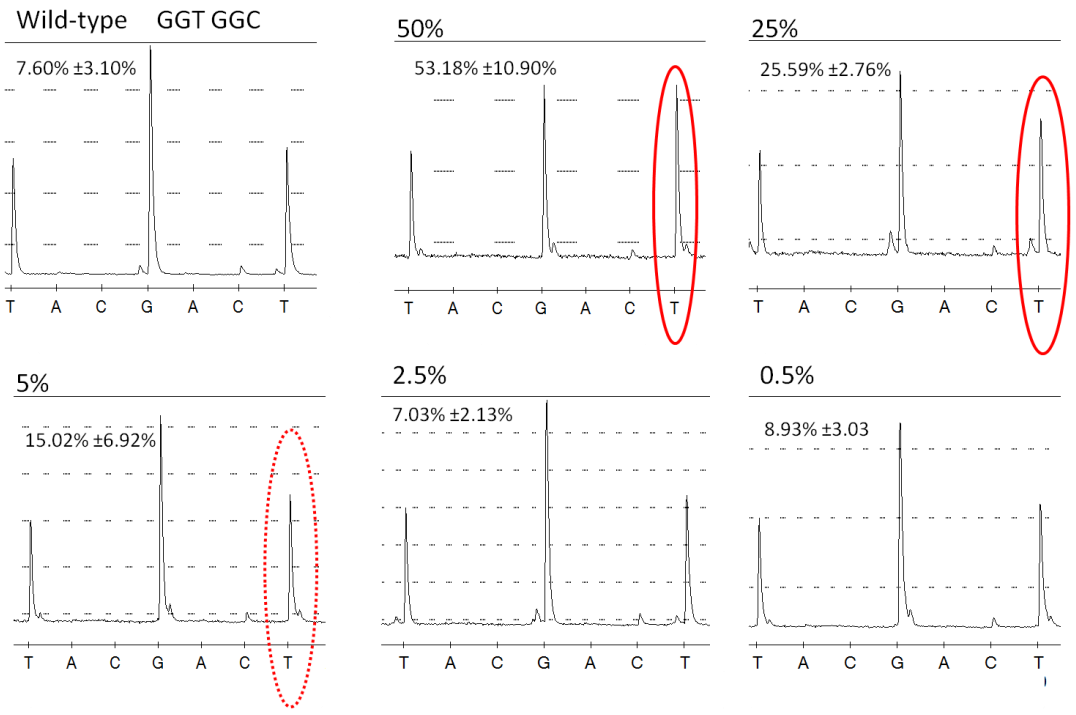
D.

G12R c.34 G>C CGT GGC



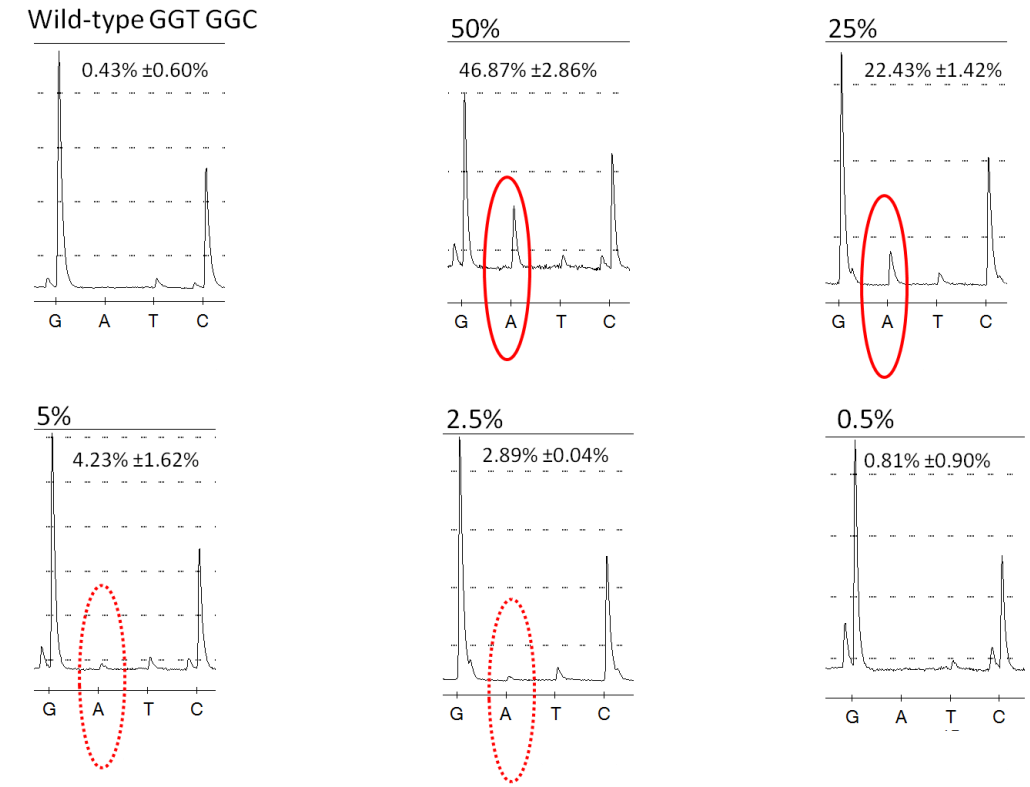
E.

G12V c.35G>T GTT GGC



F.

G13D c.38G>A GGT GAC



2.4.4 RFLP

After the initial round of PCR, the modified primers successfully amplified the 196bp KRAS amplicon in both mutant and WT samples (Figure 29).

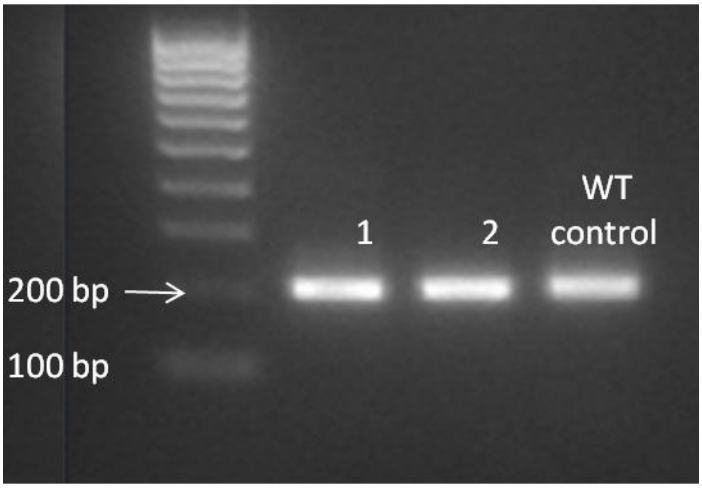


Figure 29. Amplifciation of 196 base pair KRAS amplicon in mutant and wild-type

After incubation with BstNI enzyme, WT samples were digested and then used as template for the second round of PCR, which successfully amplified the 196bp mutant whilst no amplification was seen in WT samples (Figure 30).

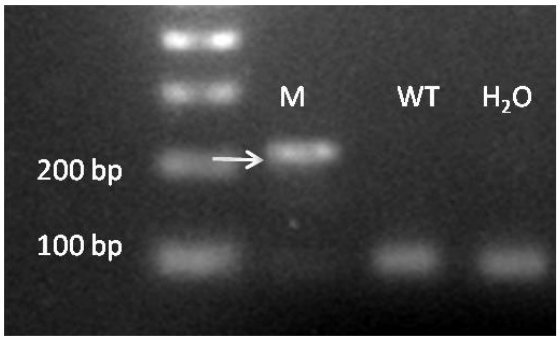


Figure 30. Second-round amplification of restriction digest products. Only the mutant sample is amplified. The 196 base pair KRAS amplicon is indicated by the arrow.

In order to determine the limit of detection of this technique, WT-restriction digest was performed on serial dilutions of mutant KRAS DNA and were subsequently sequenced by pyrosequencing. Through visual inspection of agarose gel electrophoresis, mutant DNA could be detected down to 5% (Figure 31). Pyrosequencing was performed in order to confirm that the correct mutation was present and that the band was not a non-specific product. Through pyrosequencing, mutations could be detected down to a level of 2.5% (Figure 32).

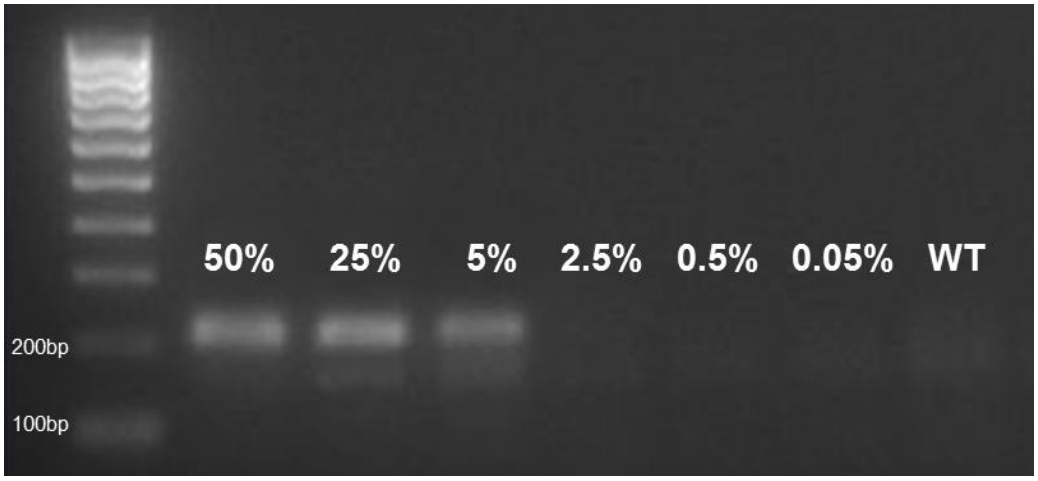
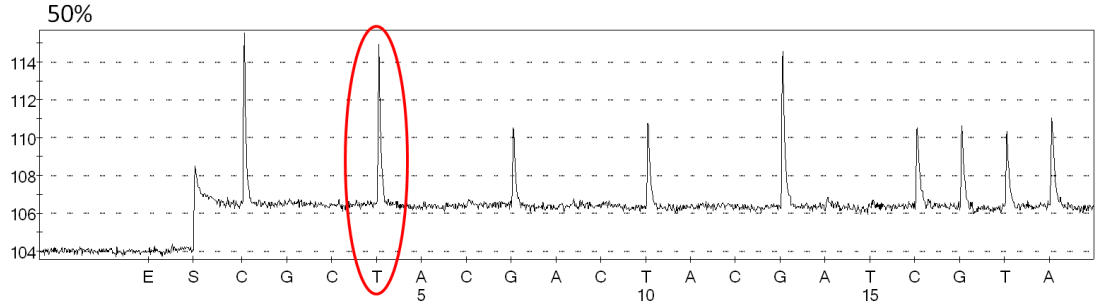
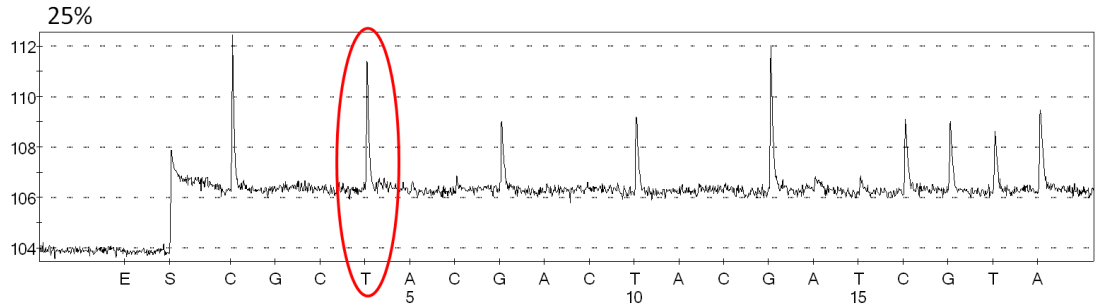


Figure 31. Serial dilutions of mutant KRAS after WT restriction digest.

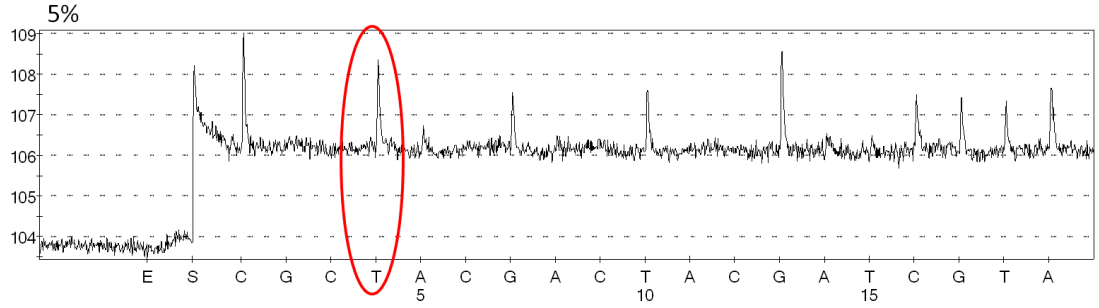
50%


114
112
110
108
106
104
E S C G C T A C G A C T A C G A T C G T A
5
10
15

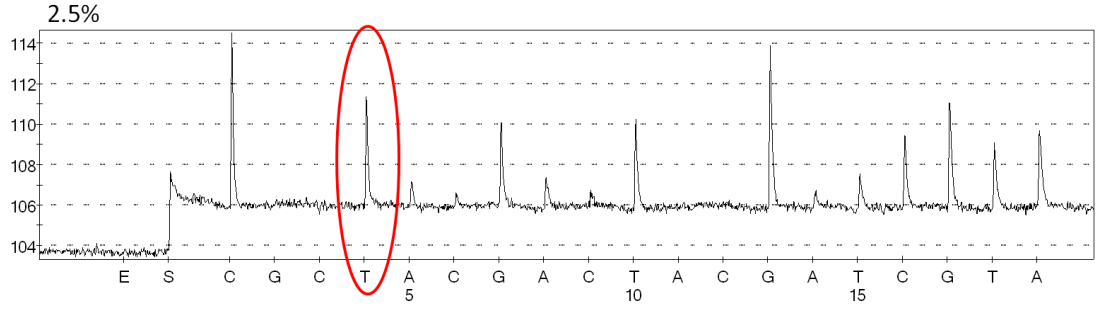
25%


112
110
108
106
104
E S C G C T A C G A C T A C G A T C G T A
5
10
15

5%


109
108
107
106
105
104
E S C G C T A C G A C T A C G A T C G T A
5
10
15

2.5%


114
112
110
108
106
104
E S C G C T A C G A C T A C G A T C G T A
5
10
15

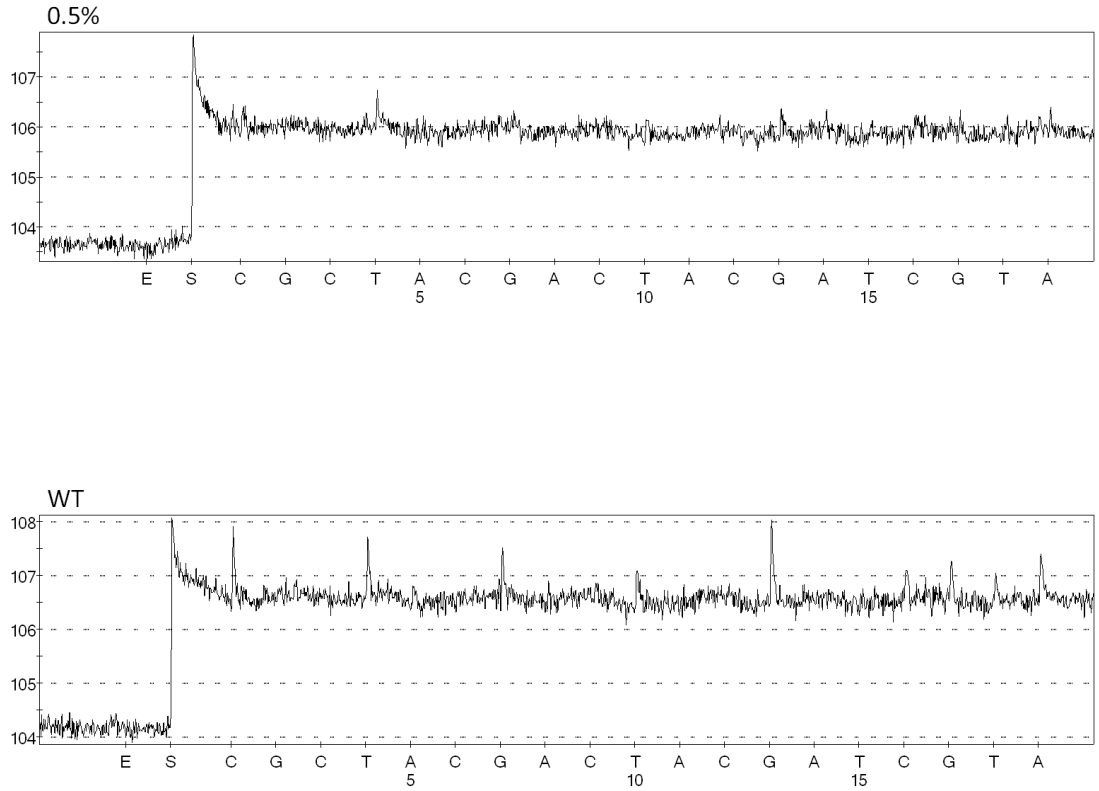


Figure 32. Pyrograms of serial dilutions of G12C c.34G>T after WT restriction digest showing a detection of mutant allele down to 2.5%.

2.4.5 COLD-PCR

In order to determine the critical temperature (T_C) for COLD-PCR, firstly melting curve analysis was performed on the mutant-wildtype heteroduplexes (Figure 33). The melting temperature of the WT:WT heteroduplexes was found to fall just below 86°C whereas the mutant-WT melted at around 85°C. Therefore a range from 84.1°C – 85.5°C was tested as the critical temperature.

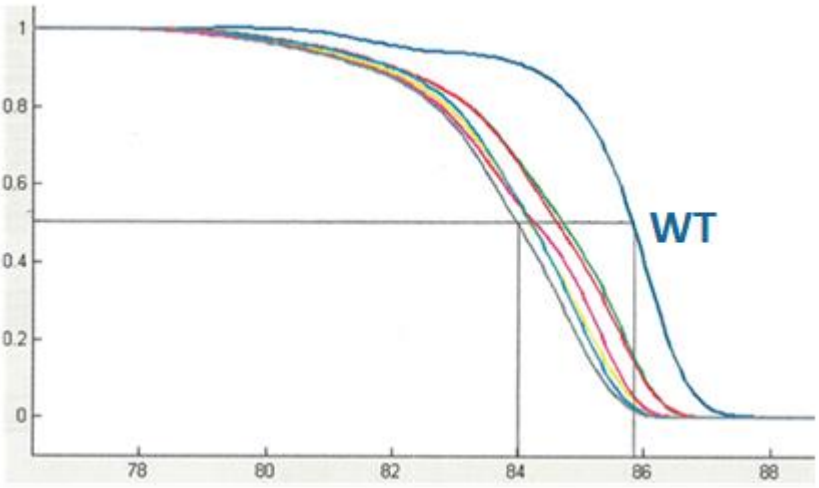


Figure 33. Melting curves of Wildtype:Wildtype duplexes (blue line) and Wildtype:mutant heteroduplexes

A 2.5% dilution of the KRAS G12D c.35G>A mutation was amplified by COLD-PCR (section 2.1.4) as well as standard PCR amplification (section 2.3.11) and then pyrosequenced. Enrichment was determined by subtracting the WT background peaks from both the COLD-PCR sample and a standard PCR sample, then comparing the calculated mutant allele frequencies (section 2.3.15). A maximum enrichment of 2.1-fold was found at 85.5°C and so this was taken as the critical temperature (Table 16, Table 17 and Figure 34).

T_c (°C)	Calculated % from mutant peak heights	Enrichment (fold)
84.1	5.64	1.38
85.1	6.97	1.83
85.3	7.26	1.93
85.5	7.76	2.10

Table 16. Enrichment after COLD-PCR for critical temperatures 84.1°C – 85.5°C on the 2.5% dilution of 12Dc.35G>A mutant.

T_c (°C)	Calculated % from mutant peak heights	Enrichment (fold)
85.5	6.15	1.62
85.7	5.16	1.06
85.9	6.17	1.45
86.1	4.87	1.03

Table 17 Enrichment after COLD-PCR for critical temperatures 85.5°C – 86.1°C on the 2.5% dilution of 12Dc.35G>A mutant.

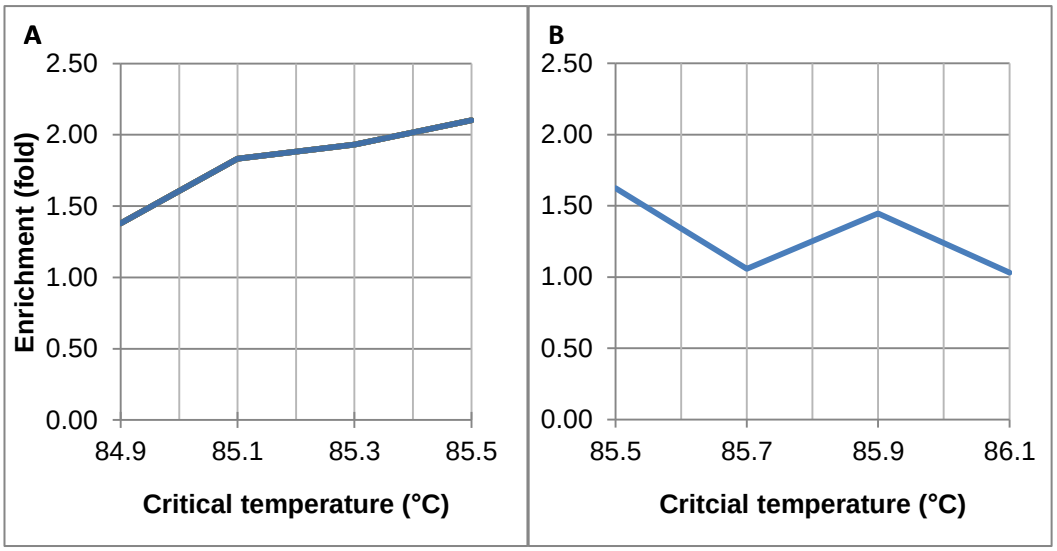


Figure 34 A&B. Enrichment after COLD-PCR for critical temperatures 84.9°C – 86.1°C on the 2.5% dilution of 12Dc.35G>A mutant.

2.4.6 Ice-COLD-PCR

As with the COLD-PCR technique, melting curve analysis was first performed but with the RS:mutant and RS:WT heteroduplexes as shown in Figure 35. From this analysis, a temperature range of 83.5°C – 84.5°C was tested as the critical temperature. 50% and 25% dilutions of KRAS G12Dc.35G>A were tested and enrichment was only seen at 84°C in the order of 1.15-fold (Table 18 and Table 19).

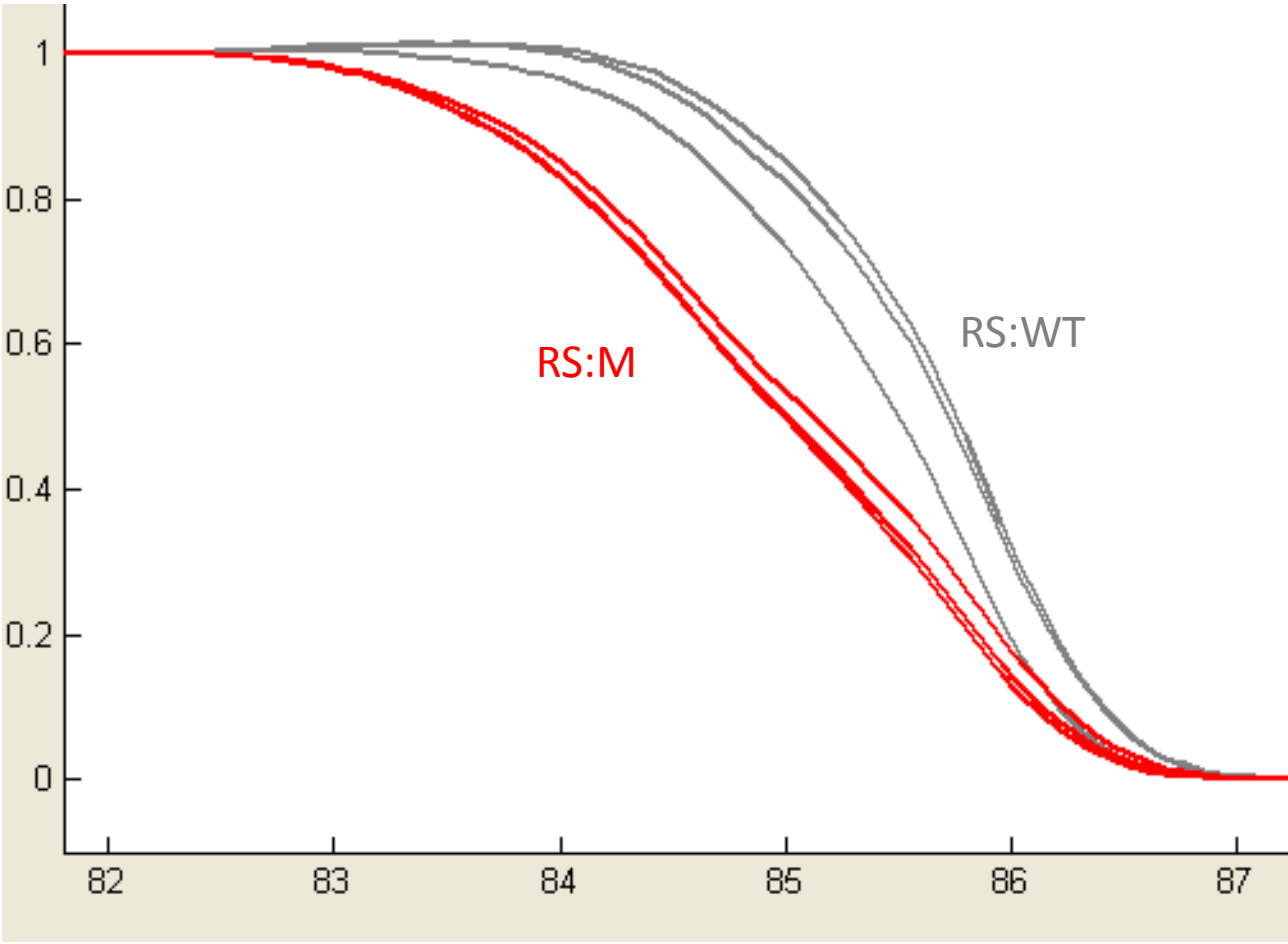


Figure 35. Melting curves of RS:wildtype (RS:WT) heteroduplexes and RS:mutant (RS:M) heteroduplexes.

T_c (°C)	Calculated % from mutant peak heights	Enrichment (fold)
83.5	53.06	1.08
84.0	56.63	1.16
84.5	51.64	1.06

Table 18. Critical temperatures tested on 50% 12Dc.35G>A mutant template

T_c (°C)	Calculated % from mutant peak heights	Enrichment (fold)
83.5	24.32	0.84
84.0	32.82	1.14
84.5	28.26	0.98

Table 19. Critical temperatures tested on 25% 12Dc.35G>A mutant template

In order to improve the amount of enrichment seen, the concentration of RS was optimised. A range of concentrations (0nM – 50nM) was tested with 25% dilutions of two KRAS mutants, G12Dc.35G>A and G13Dc.38G>A. Enrichment was seen at the 25nM RS concentration for the 12Dc.35G>A (Table 20) and at both 20M and 50nM for 13Dc.38G>A (Table 21). Therefore an optimal concentration of 25nM was chosen for further experiments.

RS Concentration (nM)	Enrichment (fold)
0	-
1	0.95
5	0.98
20	0.94
25	1.46
50	0.37

Table 20. Enrichment from a range of RS concentrations for KRAS 12D c.35G>A mutant at 25% dilution

RS Concentration (nM)	Enrichment (fold)
0	-
1	0.96
5	0.98
10	0.95
20	1.23
50	1.21

Table 21. Enrichment from a range of RS concentrations for KRAS 13D c.38G>A mutant at25% dilution

Finally Ice-COLD-PCR with a T_c of 84°C and RS concentration of 25nM was performed on serial dilutions of 3 different KRAS mutants; 12D c.35G>A, 12R c.34G>C and 12V c.35G>T. Enrichment was seen at the lower concentrations (2.5% - 0.05%) for 12D c.35G>A (Table 22). Enrichment was seen at the 0.5% dilution for both 12R c.34G>C and 12V c.35G>T (Table 23 and Table 24), but only at 5% and 25% for the 12V c.35G>T mutation (Table 24). The maximum enrichment achieved was 2.24-fold.

Dilution (%)	Enrichment (fold)
50	1.15
25	0.40
5	0.84
2.5	1.14
0.5	2.44
0.05	1.81

Table 22. Enrichment from Ice-COLD-PCR on serial dilutions of KRAS 12D c.35 G>A mutation

Dilution (%)	Enrichment (fold)
25	1.07
5	1.00
0.5	2.24

Table 23. Enrichment from Ice-COLD-PCR on serial dilutions of KRAS 12R c.34 G>C mutation

Dilution (%)	Enrichment (fold)
25	1.23
5	1.30
0.5	1.36

Table 24. Enrichment from Ice-COLD-PCR on serial dilutions of KRAS 12V c.35 G>T mutation

2.4.7 Pyrosequencing of clinical samples

The extracted DNA from crypts isolated by mechanical dissociation (section 2.3.6) in normal mucosa samples from 20 patients with colorectal cancer were pyrosequenced. DNA from the matched tumour was also pyrosequenced (Table 25). No mutations were detected in any of the histologically normal mucosa. There were 6 different KRAS codon12 and13 mutations were detected in 8 (40%) of the tumour samples (Table 25). No KRAS 61 or 146 mutations or BRAF mutations were detected in any of the tumour or normal mucosa samples.

Sample Number	Mutation	Sample Number	Mutation
1	WT	11	WT
2	13D c.38G>A	12	WT
3	12V c.35G>T	13	WT
4	12D c.35G>A	14	WT
5	WT	15	WT
6	WT	16	12V c.35G>T
7	12R c.34G>C	17	12V c.35G>T
8	WT	18	WT
9	12A c.35G>C	19	WT
10	12S c.34G>A	20	WT

Table 25. Pyrosequencing of 20 tumour samples.

2.5 Discussion

2.5.1 Spot-counting analysis

The spot-counting analysis found 51.7% of the mucosa to consist of epithelial cells (section 2.4.1 Table 10). These are the cells that become mutated in CRC and therefore are the cells of interest to extract DNA from for mutational analysis (Markowitz and Bertagnolli, 2009). The mucosa contains other cell types such as fibroblasts, macrophages, lymphocytes etc. that are found predominately in the supporting lamina propria (Hunyady et al., 2000). These cells would contaminate the DNA extraction and would effectively dilute the mutant epithelial DNA. If mutations are to be detected in normal colorectal mucosa they are likely to be at a very low level, therefore enriching the epithelial cell population within the DNA extraction may increase the chance of detecting mutations within the epithelial cells if they are present. If the epithelium were to be separated, then the relative proportion of WT:mutant alleles would change, increasing the ratio of WT alleles : mutant alleles from 3:1 to 1:1. It is therefore optimal to isolate the colonic crypts in order to obtain the purest sample of epithelial DNA possible.

2.5.2 Comparison of crypt isolation techniques

Despite micro-dissecting approximately 1000 crypts that was estimated to produce a sufficient yield of DNA (section 2.4.2) LCM failed to produce enough material for DNA extraction to produce a nanodrop reading that was high enough compared to the negative control ($p=0.095$) (Table 12) and could not be visualised on an agarose gel. LCM produced the lowest amount of DNA from any of the extraction methods (Table 14). It was also the most time-intensive method (approximately 1- 2 days per sample) in contrast to MD where multiple samples can be extracted simultaneously in a few hours.

MD produced the highest median yield of DNA but also the largest range and IQR (Table 13 and Table 14). The median yield obtained was significantly higher than the median yield from the LCM method ($p<0.05$). However, this large difference is most likely caused by the difference in starting material. For whole mucosa, six 7µm sections of ~5mm^2 were taken for extraction, giving a total volume of starting material of ~0.035mm^3. In comparison, whole samples of mucosa that underwent MD measured ~22mm^3 in volume. However, due to the length of time it requires to isolate crypts via LCM it would be impractical to perform LCM on a comparably sized sample. The large range of DNA yield for MD may be due to variance in the

size of the starting material. Therefore to directly compare methods, starting samples of mucosa with set dimensions could be collected. However the advantage of a large sample would be the increased ability to identify the presence of a mutated clone. All samples extracted by MD produced a high enough amount of DNA for downstream sequencing.

The negative control for each extraction produced a reading on the nanodrop; however these always fell below 10ng/µl. This could be due to contaminating DNA. However it is worth noting that the 260:280 ratios for these readings strongly indicated that it was not DNA causing these readings. It was likely due to components of the extraction buffers that had washed through with the elution buffer. It would be useful to perform PCR on the negative control extractions to confirm if the reading was from contamination DNA.

MD is a suitable method for isolating crypts from fresh tissue, however, for FFPE tissue this method is not possible and therefore macrodissection with the aid of a dissecting microscope is the most practical method available for DNA extraction from normal mucosa. Table 11 shows that a sufficient yield of DNA was obtained from this method (section 2.3.7).

2.5.3 Limit of detection of pyrosequencing

The sensitivity of pyrosequencing was found to be between 5%-25% (Table 15) and this reflects the findings of previous studies that report a sensitivity of 5% (Ogino et al., 2005, Tsiatis et al., 2010). The pyrograms in Figure 28 also show that the sensitivity of the pyrosequencer varies according to which mutation is being detected (see appendix for full dataset). As seen in Figure 28C the G12Dc.35G>A mutation is detected at 5% and is on the limits of detection at 2.5% whereas for the 12C c.34G>T mutation (Figure 28B), the mutant could not be detected at a dilution below 25%. All the pyrograms show that there is very low background at c.34 for the A-peak and the single A peaks produce higher intensity peaks than the other nucleotides and this may contribute to the increased sensitivity observed for the 12D c35G>A mutation. Whereas for a G>T mutation, the single “T” peak becomes a double and this increase in intensity of the T peak is harder to distinguish (Figure 28B) than a new peak over background such as a G>A mutation, which is why pyrosequencing is less sensitive for the G>T mutation.

Diagnostic mutation testing in clinical samples utilises a different method to detect if a mutation is present. If a mutant peak height is three standard deviations away from the mean then it is classed as a mutation and samples that lie 2-3 standard deviations from the mean are repeated. The peaks observed on the pyrograms are also used in conjunction with the peak heights. This method is sufficient for clinical samples as tumour cell DNA concentration is usually around 20-25% of the mutant allele since half the cells are tumour cells and they only carry one mutated allele. However, for determining the sensitivity of the pyrosequencer with cell-line DNA with a known mutated DNA concentration, dilutions below 25% of the mutant allele do not fall over 3 standard deviations and peaks can be difficult to distinguish from the background on the pyrogram. Therefore the calculated percentage method can be used to determine if the pyrosequencer can detect the presence of a mutation when the starting dilution of the mutant cell-line DNA was known.

2.5.4 RFLP

The WT restriction digestion method allowed for the c.34G>C mutation to be detected at 5% when using visual detection with agarose gel electrophoresis (Figure 31). The brightness of the bands (and therefore the sensitivity) could be enhanced by blotting and probing with fluorescence or radioactivity. By using WT-restriction digest combined with pyrosequencing the c.34G>C mutation could be detected at the 2.5% dilution as shown in Figure 32. Below this dilution the template sequence could not be detected as the noise-to-peak ratio was too high as seen in the 0.5% and WT sample. From previous pyrosequencing of the 12C c.34G>C mutation the lowest dilution to be detected for this mutation was 25% (Figure 28B). Therefore this improved the sensitivity of pyrosequencing and allowed for a lower mutant concentration to be detected.

2.5.5 COLD-PCR

This method enriched the 12Dc.35G>A 2.5% mutant by up to 2-fold at the critical temperature of 85.5°C (Table 16 and Figure 34A). At the lower temperatures, less enrichment was seen due to the temperature not being high enough to melt the WT:mutant heteroduplex. If it remains bound, then primers cannot anneal to amplify the mutant sequence. As the temperature was increased and more of the WT:mutant heteroduplexes melted, there was an increased amount of template

available for the primers. This is seen by the graph in Figure 33. Once too high a critical temperature is reached, then the WT:WT duplex will melt and more WT template will be available to be amplified. This will then reduce the amount of enrichment seen by the COLD-PCR method as shown in Table 17 and Figure 34B. Although a small amount of enrichment was observed, it was not substantial enough to greatly improve the limit of detection of pyrosequencing.

2.5.6 Ice-COLD-PCR

The enrichment from Ice-COLD-PCR and pyrosequencing is reported in the literature as 5.5 - 75 fold for mutant DNA dilutions of 0.1%– 10% (Milbury et al., 2011). However, the enrichment found in comparison to the literature was much lower.

For the optimisation of critical temperature two different dilutions of the KRAS 12D c.35 G>A mutation were tested and slight enrichment of 1.14-fold and 1.16-fold was detected at the critical temperature of 84°C for both dilutions (Table 18 and Table 19). Enrichment was not detected at the 83.5°C or 84.5°C temperature for the 50% dilution which indicates that this method is reliant on an accurate temperature to ensure that the RS:M heteroduplex melts but that the RS:WT remains bound.

For RS concentration titration, the original protocol specified an optimal concentration of 25nM (Milbury et al., 2011). Maximum enrichment of 1.46-fold was seen at this concentration for the 12D c.35G>A mutation whereas enrichment was not seen at other RS concentrations (Table 20). For the RS titration with the 13D c.38G>A mutation, enrichment was seen at both 20nM (1.23-fold) and 50nM (1.21-fold) (Table 21). Enrichment was not seen at 50nM for the 12D c.35G>A mutation and therefore an optimal concentration of 25nM was chosen. The RS sequence has to be added in excess in order to encourage the formation of heteroduplexes. However there is the chance that too high a concentration may result in primers binding to the RS sequence, although this is unlikely due to the RS sequence design in that it overlaps primers by only 4bp. Conversely, if too low a concentration is added then heteroduplexes will not form and the mutant cannot be enriched.

Serial dilutions of 3 different mutations were sequenced after Ice-COLD-PCR (Table 22, Table 23 and Table 24). Enrichment was seen only at the lower dilutions (2.5%, 0.5% and 0.05%) for the 12D c.35G>A mutation (Table 22) with a maximum enrichment of 2.44-fold. For the 12R c.34G>C mutation enrichment was only found at the 0.5% dilution (Table 23) and finally for the 12V c.35G>T mutation a slight enrichment was seen for all three dilutions tested (25%, 5% and 0.5%) with maximum enrichment at 0.5% of 1.36-fold (Table 24). Enrichment was not seen at the 5% dilution for the three mutations tested, although it would be expected. Results show that the enrichment seen is moderate and varies with each run and results reported in the literature from this technique could not be duplicated.

2.5.7 Tumour Samples

No mutations were detected in any of the 20 samples of normal mucosa, but KRAS codon 12 and 13 mutations were detected in 8 of the tumour samples (Table 25) which is a prevalence of 40%. As the sensitivity of the pyrosequencer lies between 5-25% if there are mutations in the normal mucosa, then these will not be detected by pyrosequencing. Therefore it was decided to explore a next generation sequencing approach for mutation detection.

2.6 Chapter Summary

- Epithelial cells make up 51.7% of the bowel mucosa
- Mechanical dissociation is a practical and effective method of isolating colonic crypts from fresh material, which yields on average 522ng/µl of DNA.
- Pyrosequencing has a limit of detection between 5% and 25% depending on the mutation being detected
- Restriction fragment length polymorphism improved limit of detection by 10-fold (25% to 2.5%)
- COLD-PCR based mutation enrichment techniques failed to produce significant enrichment

3 Next generation sequencing detection of KRAS in normal mucosa

3.1 Introduction

3.1.1 Next generation sequencing

Next generation sequencing (NGS) is a broad term that encompasses sequencing technologies that allow for multiple DNA templates to be read simultaneously. This high output of multiple reads gives NGS its advantages of high throughput, high sensitivity and low cost per read (Mardis, 2008). There are various NGS platforms with differing biochemistries; however they all have three broad components in common; library generation, sequencing and bioinformatic analysis (Shendure and Ji, 2008, Metzker, 2010). Commonly used platforms are the Illumina sequencers including the Genome Analyzer IIe (GAIIe), Miseq and Hiseq sequencers (Illumina, San Diego, USA). Other common NGS sequencers include the ion torrent and SOLiD sequencing (Life Technologies, San Francisco, USA) and 454 sequencing platforms (Roche, Branford, USA).

Ion torrent sequencing works via the detection of Hydrogen ions released during DNA polymerisation. As nucleotides are added in a known sequence, a hydrogen ion is released if it is incorporated into the sequence from the template molecule. This results in a change in pH which can then be detected. This technology offers rapid sequencing speeds and low running costs (Quail et al., 2012). The disadvantage to this technology is that homopolymer repeats of 5 nucleotides or longer reduce accuracy (Loman et al., 2012).

Sequencing by Oligonucleotide Ligation and Detection (SOLiD) sequencing works via the ligation of probes rather than sequencing by synthesis. The template is immobilised and then four fluorescently labelled probes are added to the template. Where the probe matches the template sequence, ligation occurs and the fluorescent label is detected. The probe is then denatured from the template and the process repeats seven times with the other probes to decode the template sequence. This results in each base within the template being detected twice which leads to high accuracy and low SNP miscall probabilities. However, due to its sequencing chemistry, SOLid sequencing has reported issues with sequencing palindromic sequences (Huang et al., 2012).

454 sequencing works by pyrosequencing in parallel. Rather than an overall signal as is obtained through classic pyrosequencing, each individual template molecule is stabilized and sequenced individually through pyrosequencing. Some modified kits have enabled read lengths of 1000bp. However, it faces the same issues as pyrosequencing with accuracy of calling homopolymer repeats (Quail et al., 2012). Table 26 outlines the main advantages and disadvantages of the most commonly used NGS sequencing technologies.

Sequencer	Read length	Accuracy	Output data per run	Time per run
Illumina HiSeq 2000	Paired end 100bp	98%	600 Gb	11 days
Ion torrent	400 bp	99%	1.2-2 Gb	7.3 hours
SOLiD 4	Paired end 50bp x 25bp	99.94%	100 Gb	11-13 days
454 sequencing	Up to 1000bp	99.9997%	0.07 Gb	23 hours

Table 26. Comparison of next generation sequencing platforms. Data obtained from company websites.

3.1.2 Illumina next generation sequencing

Before sequencing can take place, libraries have to be prepared from the sample as outlined in Figure 36. These libraries must then be stabilised onto the sequencer flow cell. The flow cell contains a “lawn” of oligonucleotides; any DNA template that has complementary adaptors on its ends will anneal to oligonucleotides on the flowcell and be sequenced. The template DNA can be either gDNA, PCR generated targets or hybridization capture targets depending on the application of the sequencing, but they all must contain the Illumina adaptors to enable sequencing.

After the DNA library has been prepared, it is applied to an 8-channel flowcell followed by a series of heating and cooling steps to allow the DNA molecules to hybridize, before the molecules undergo bridge amplification in order to generate clusters. Each cluster contains approximately 1000 copies of the original DNA fragment and each flow cell can contain over 100 million clusters (Mardis, 2008).

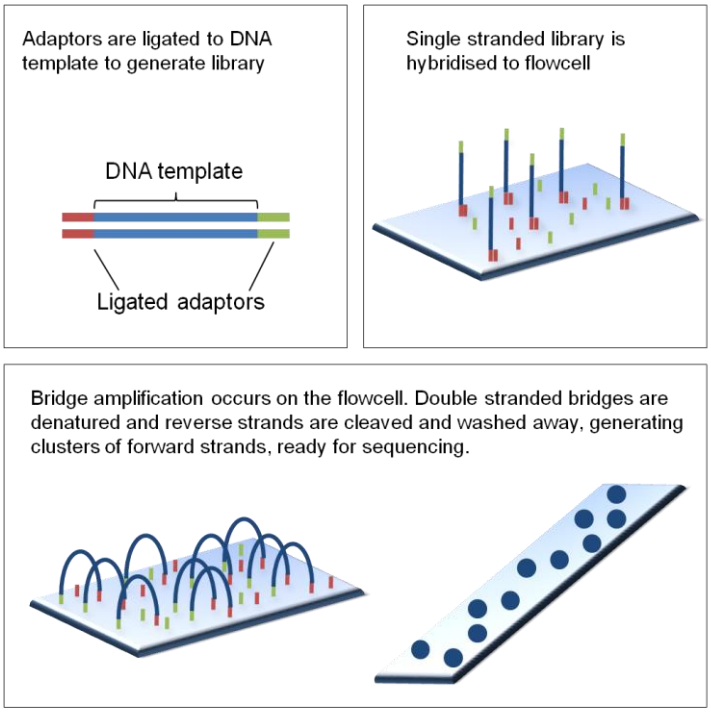


Figure 36. Library preparation and cluster generation for Illumina sequencers. Adapted from Mardis (2008).

Once clusters are generated, the forward and reverse sequencing primers are added and anneal to the adaptor sequence. With the addition of polymerase, fluorescently tagged nucleotides are added. Each base has a unique fluorescent tag and as they are incorporated into the DNA sequence, the dye is cleaved and the fluorescence is detected. An imaging step occurs after every addition of fluorescent nucleotides in order to determine the sequence from the clusters. This process repeats for varying read lengths. Each cluster can be read in both a forward and reverse direction in a process known as paired-end sequencing (Shendure and Ji, 2008).

3.1.3 Library preparation for targeted sequencing

Whole genome sequencing is a common application of NGS, however for the purposes of mutation detection, it may not be able to detect low allele frequencies due to low numbers of template covering the key cancer genes (Meyerson et al., 2010). It is therefore desirable to employ a target enrichment/capture method so that only areas of interest are sequenced and high coverage can be obtained. There are a number of approaches for target capture including PCR, molecular inversion probes, hybrid capture and in-solution capture (Mamanova et al., 2010).

PCR is a robust method for target amplification and can be modified in a number of ways for NGS library preparation. PCR amplicons of genetic targets can be created, adaptors ligated and directly sequenced as outlined in Figure 37 (Chambers et al., 2013, Peeters et al., 2013). Due to the adaptors containing a barcode sequence, multiple targets can be amplified for one patient and pooled together before barcoded adaptors are attached. This method is appropriate where the number of targets remain small, however, if larger numbers of targets are to be interrogated a parallel PCR method is favourable such as outlined in Chapter 4.

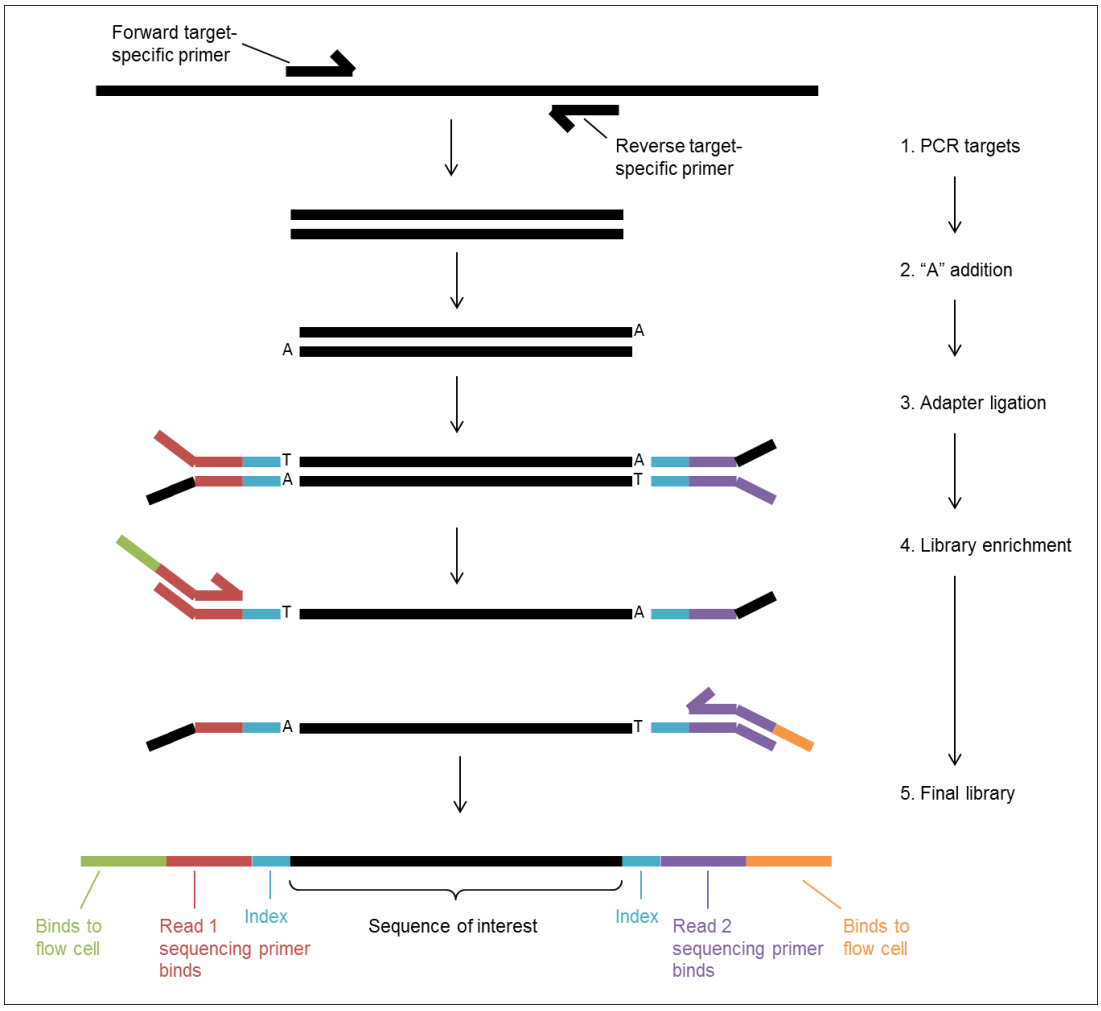


Figure 37. Schematic of library generation by ligating Illumina adaptors onto PCR amplicons

Adaptor ligation is not idea for preparing PCR amplicons for NGS sequencing for a number of reasons:

- There are many incubation and clean up steps which increase the risk of sample contamination.
- Up to 20% of the DNA is lost in every clean up step and therefore libraries require a final enrichment PCR. Also this allows for the final adaptor sequences to be added to the libraries. However, this means that PCR amplicons that have undergone 40 cycles of PCR originally have an extra 12-15 cycles that can incorporate error.
- Multiple reactions and clean up steps incur a high cost and are time consuming.

To overcome these issues, a one-step method was developed; targeted amplicon library creation (TALC). The Illumina adaptor sequences are incorporated into the targeted PCR primers as an additional tag; therefore adaptors are added as the target is amplified. Primers containing an index are also added to the PCR, therefore allowing the gDNA to be amplified for the gene of choice and indexed simultaneously. Targets are then pooled equally for each patient and a final pool of patients is purified before it is ready for sequencing. The advantage to this TALC approach is a faster method with fewer steps and purifications. There is only one single PCR step which reduces the chance of PCR error. An overall schematic is outlined in Figure 38.

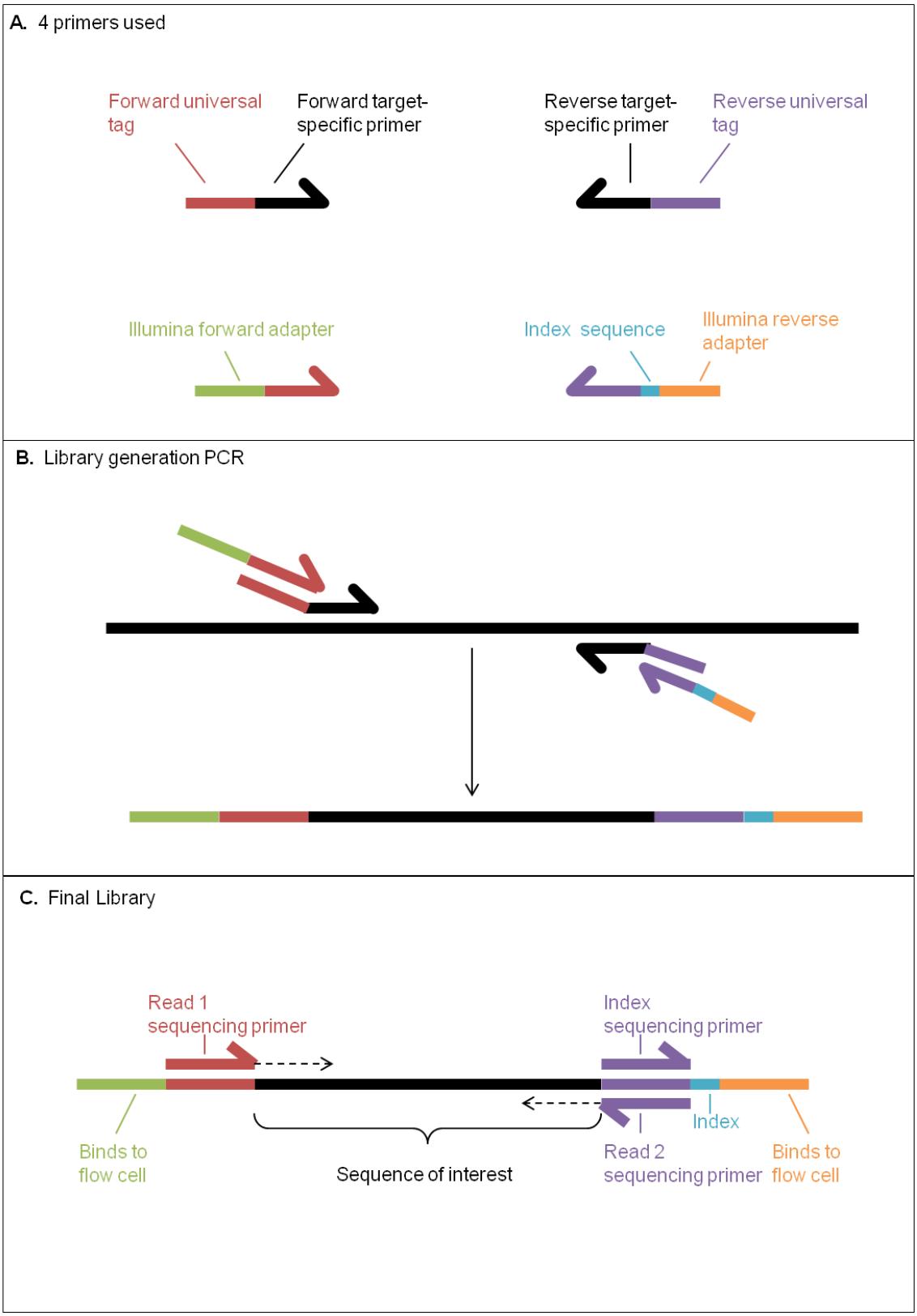


Figure 38. Schematic of targeted amplicon library creation (TALC) showing the primer sequences used to create final libraries in a single PCR reaction.

3.1.4 Bioinformatic approaches

The management and analysis of NGS data poses a number of challenges to efficiently and accurately analyse multiple samples due to the large amount of data obtained in a single NGS run. There are numerous tools to manipulate NGS data for variant calling offering different functions, but they all broadly slot into a general analysis pipeline outlined in Figure 39. Each base call from the sequencer is given an associated quality score. This is computed by analysing parameters relevant to the specific sequencing chemistry. These parameters are then compared to a large empirical data set of known quality to generate a quality score for each base call. These scores are known as Phred scores (Phil's Read Editor) and quantify the quality of the base-call based on the probability of a miscall (Ewing and Green, 1998, Voelkerding et al., 2009). Scores are encoded symbolically and fall between 0 and 40. They are calculated on a logarithmic scale where miscall probabilities of 0.1 (10%), 0.01 (1%) and 0.001 (0.1%) are given scores of 10, 20 and 30. They are calculated as $Q=-10\log_{10}P$ or $P=10^{-Q/10}$ where Q is the Phred quality score and P is the base-calling error probability (Voelkerding et al., 2009).

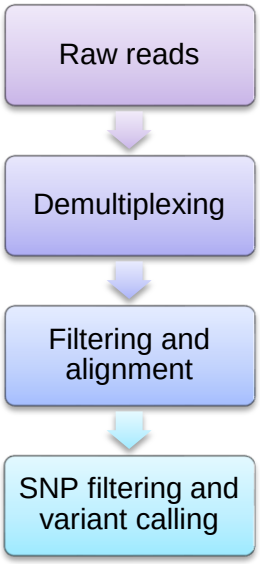


Figure 39. General pipeline of NGS data analysis showing the steps required to generate a final mutation report from raw output reads from the sequencer.

The raw reads directly outputted from the sequencer are bioinformatically demultiplexed and split into separate read files per sample. After this step the reads from each sample can be aligned to a reference genome, where reads with poor quality scores are filtered out. There is a huge range of sequence alignment software available and due to the relatively short read lengths compared to Sanger sequencing, NGS data favours short-read sequencing alignment programmes (Horner et al., 2010). Aligned data can then be interrogated for variant calls. Many programmes exist for this purpose with the aim to identify true SNPs and somatic mutations from artefactual non-reference alleles which may be incorporated through library preparation and sequencing. These variant callers are also affected by the depth of sequencing (Stead et al., 2013).

In order for data analysis to be easily automatable through a pipeline, data is encoded in a number of file formats that can be inputted into open-access software. Table 27 summarises some of the commonly used file formats used in NGS analysis pipelines.

File format	Description	Example	Format
FASTA	Simple text-based format for representing nucleotide sequences	>1 GTACGTACGGAATGTACG >2 GTACGGACGGAATGTACG	The information for each sequence runs over two lines. Firstly the header line begins with “>” followed by sequencing information. The nucleotide sequence is on the following line.
FASTQ	Text-based format for nucleotide sequences and quality scores	@SEQ_ID_1 GTACGTACGGAATGTACG + 456#$%(//$$6776%(//$$	Information is contained over 4 lines. Firstly the“@” line is followed by sequence identifier. The second line contains the sequence. The “+” line may contain more sequence identifier information. The final line contains the characters that encode the Phred score for each base.
SAM	Sequence Alignment/Map format. TAB-delimited text format for aligned reads.	@HD @SQ SN:chr1 LN249450621 1 0 chr1 3155567 37 22M * 0 0 GGCCTGCTGAAAAT BDDBB@BEE@CE@ MD:Z:22	The “@HD” line contains header information. “@SQ” lines contain the alignment reference information and are in alignment sorting order. Each subsequent line contains the aligned reads with 11 fields separated by tab characters. Each field contains alignment information and scores for how well the read aligns.
BAM	Binary format of SAM files.		These files contain compressed information from SAM files and are therefore quicker to use with programs, but they are not easily human readable.
MPILEUP	Text-based format for summarizing the base calls of aligned reads to a reference	APC_1 T 1365 .....A =AFDFFDDD	Each line contains 6 tab separated fields: sequence identifier, position in sequence, nucleotide at that position, number of reads covering that position, bases at that position, mapping qualities of those bases. “.” (dot) denotes that the base matches the reference.
VCF	Variant Call Format. File for storing gene sequence variation	#CHROM POS ID REF ALT QUAL KRAS1213 8 . T G .	Header line begins with “#” symbol and names the fields of interest. Each line shows within a target at what position a variant is found and provides information about that variant.

Table 27. Common file formats used in NGS pipelines

3.1.5 Sequencing errors

With the high sequence output of next generation sequencers, it may be possible to detect rare allele variants. However, challenges are faced where the sequencing error rate may be too high to confidently identify very low level variants. Previously reported error rates for Illumina sequencing lie between 0.05% and 1%, however this depends on multiple factors such as DNA template quality, read length, base-calling algorithms and the type and site of variant detected (Kinde et al., 2011, Chen et al., 2012). Errors can be introduced at a number of steps: tissue preparation, initial PCR amplification, library enrichment, amplification on the flowcell and errors occurring during sequencing. However, these can be addressed in a number of ways in order to minimise error and maximise the accuracy of rare-variant calling. The use of high fidelity polymerase such as Phusion minimises errors introduced during PCR amplification. Bioinformatic tools can be used to recalibrate quality scores and perform realignments to reduce potential artefacts such as the Genome Analysis Toolkit (GATK) (McKenna et al., 2010, Chen et al., 2012).

In order to call low frequency variants, it is important to keep the coverage of the target of interest high so that mutant calls can be confidently made. For single-end sequencing the error rate is not even throughout all amplicons and the quality scores decline after the first 30 bases of the read (Chen et al., 2012, Li et al., 2008a). Therefore calling mutations at any position within a gene such as TP53 is challenging as any base within the read could be mutated. This is in contrast to detecting mutations in an oncogene like KRAS where mutational hotspots are clearly defined as demonstrated in Figure 40. For this reason, it may be possible to set a low threshold for mutation detection in an oncogene and sensitively identify very low level mutations whereas higher levels may have to be set for tumour suppressor genes.

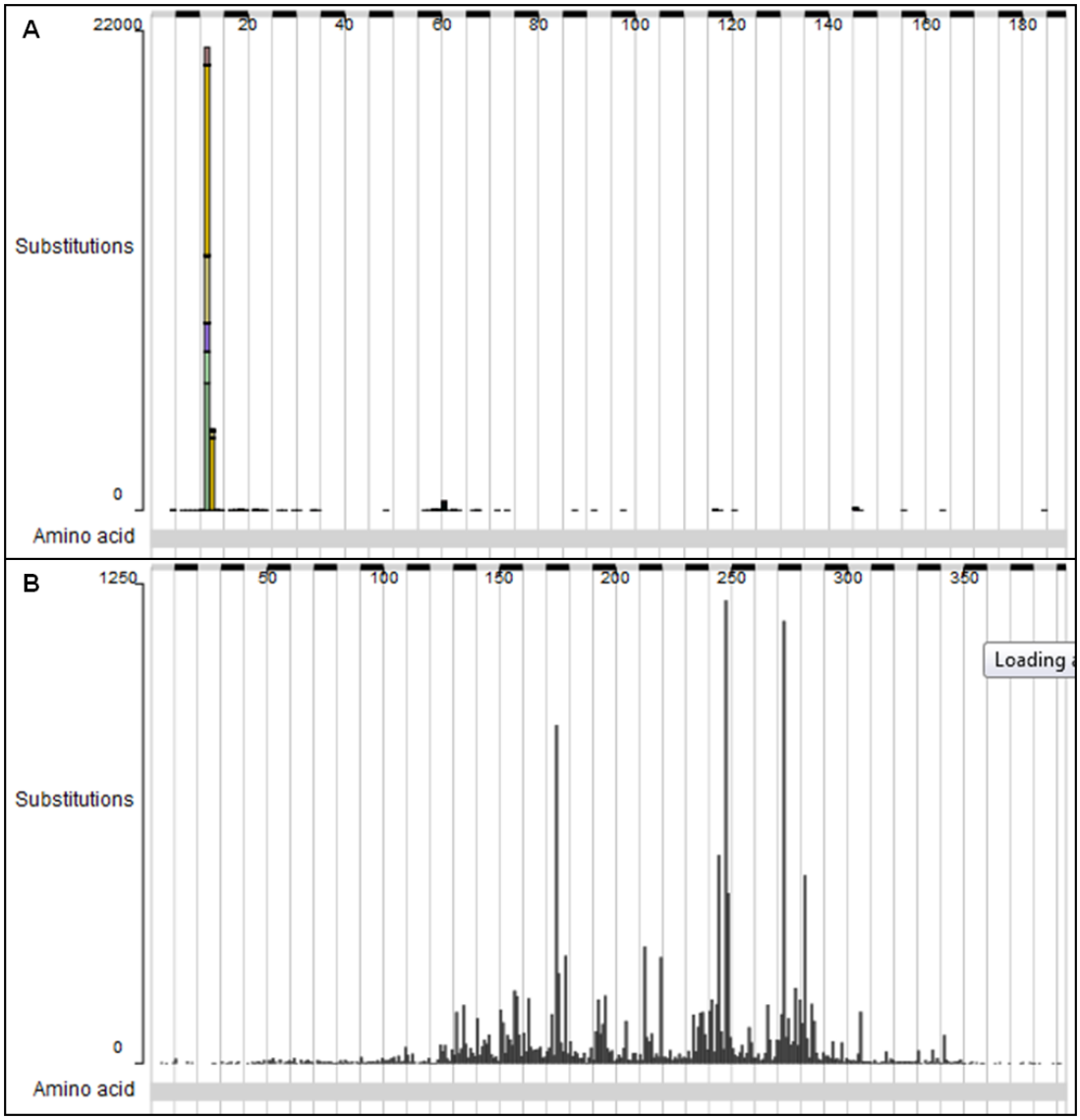


Figure 40. Mutational profiles for A. KRAS oncogene and B. TP53 tumour-suppressor gene from the COSMIC database

Separate to sequencing errors, there may be errors within the starting template DNA. This may particularly affect DNA extracted from FFPE tissue. It is reported that de-amination of cytosine bases results in C:G>T:A sequence artefacts and this is seen in FFPE tissue (Do et al., 2013). These artefacts may therefore affect the overall error rate of NGS.

3.2 Chapter Aims

The aims of this chapter are as follows:

- To investigate the sensitivity of NGS
- To investigate where mutations can be detected in cancer-associated normal mucosa, adenoma-associated normal mucosa alongside non-neoplastic normal mucosa by NGS
- The development and improvement of NGS library preparation and automated analysis

3.3 Methods

3.3.1 Generation of PCR amplicons for NGS library preparation

In order to assess the limit of detection of NGS, mutated cell line DNA dilutions were amplified for KRAS codons12 and 13 as outlined previously in Chapter 2. The taq polymerase used previously was substituted for Phusion Hot Star Flex Master Mix (New England Biolabs, Hitchin, UK) due to its improved amplification error rates and PCR accuracy. The PCR reaction conditions are described in Table 28.

Component	Volume per reaction		Final concentration
	25µl reaction	50µl reaction	
Phusion Hot Start Flex 2X Master Mix	12.5µl	25µl	1X Phusion Hot Start Flex Master Mix
Forward Primer (25µM)	0.5µl	1µl	0.5µM
Reverse Primer (25µM)	0.5µl	1µl	0.5µM
Dimethyl sulfoxide (DMSO)	0.75µl	1.5µl	3%
H_2O	8.75µl	17.5µl	
gDNA (10ng/µl)	2µl	4µl	

Table 28. PCR reaction composition with the use of Phusion polymerase enzyme.

The thermocycling conditions with the use of Phusion Hot Star Flex were as follows:

- 98°C for 30secs

- 40 cycles of:
- 98°C for 5secs
- 60°C for 10secs
- 72°C for 15secs

- Hold at 4°C

After confirmation of successful amplification on a 2% agarose electrophoresis gel, PCR products were cleaned up using Qiagen's MinElute PCR purification kit (QIAGEN, Crawley, UK) eluting in 15ul EB buffer. Quantification of PCR products was performed by use of a Quant-iT PicoGreen assay (Invitrogen, Paisley, UK) following the standard protocol.

3.3.2 RFLP and NGS

Alongside this, in order to determine the enrichment gained from RFLP and NGS, RFLP amplicons from mutant serial dilutions were prepared as previously outlined in chapter 2. At the very low mutant allele frequencies the mutant PCR product amplified by the RFLP method was present at very low quantities and had to undergo a further 15 cycles of PCR in order to gain enough to use for NGS library preparation. The mutant amplicons were then gel purified using Qiagen's gel purification kit (QIAGEN, Crawley UK) and eluted in 15ul EB buffer before undergoing standard NGS library preparation.

3.3.3 NGS amplicon library preparation

Standard Illumina Protocols were adapted for the use of PCR products as starting material. A full overview of the library prep process is shown in Figure 42. Y-shaped adaptors were ligated onto PCR products before being enriched to create the final library. The use of Y-shaped adaptors meant that PCR products with adaptors correctly ligated could easily be amplified by PCR. If the adaptors were complementary and not-Y-shaped, the PCR primers would be complementary and form primer-dimers preferentially over amplifying the final library, illustrated in Figure 41 and Figure 42. Final libraries were loaded onto the sequencer by Dr. Sally Harrison in accordance with the manufacturer's protocols on a Genome Analyser IIe, Miseq or Hiseq sequencing machine (Illumina, San Diego, USA).

Figure 41. Diagram illustrating the structure of Y-shaped adaptors and how they anneal to A tailed PCR products.

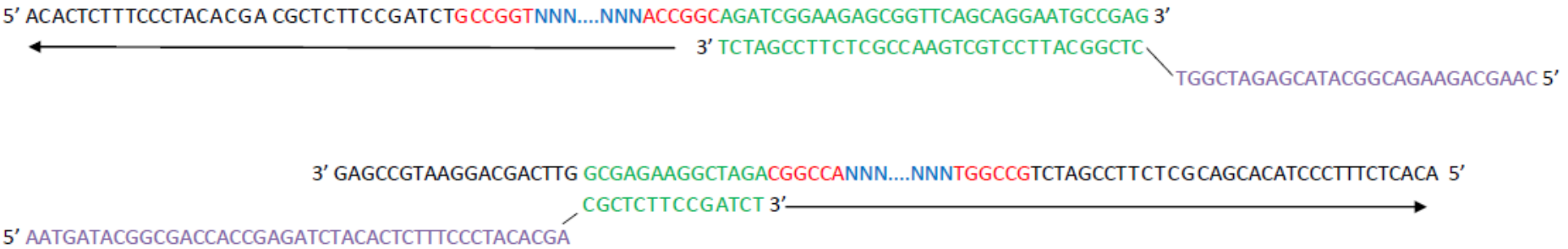
PTO Forward Primer

5’ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T 3’

PTO Reverse Primer

5’ CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3’

PCR – DENATURE LIBRARY AND AMPLIFY



End of Round 1 PCR

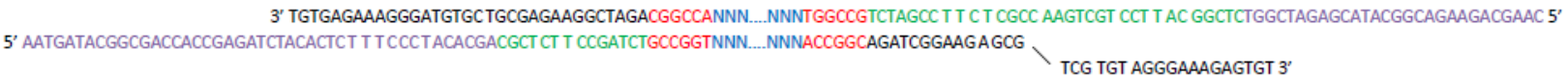


Figure 42. Schematic of NGS library enrichment from adaptor-ligated PCR products.

3.3.4 End-repair

The End-It DNA End-Repair Kit (Epicentre Biotechnologies, Madison, USA) was used to ensure that all PCR products were blunt ended to allow for the downstream steps of A-addition and ligation. A starting amount of 100-150ng of DNA was found to be optimal. The DNA was added to the reaction mix (Table 29) and left to incubate at 25°C for 45 mins. The samples were then cleaned up by a QIAquick PCR Purification Kit (QIAGEN, Crawley, UK) and eluted into 34.6µl elution buffer (EB).

Component	Volume
DNA (15ng/µl)	10µl
End-Repair 10X Buffer	5µl
dNTPs	5µl
ATP	5µl
End-Repair Enzyme Mix	1µl
H_2O	24µl
Total volume:	50µl

Table 29. Reaction mix for end-repair.

3.3.5 A-addition

The A-addition step was performed by the addition of ATP in the presence of klenow fragment, exonuclease minus (Promega, Madison, USA). This resulted in the addition of an adenosine nucleotide onto the 3' end of the blunt-ended PCR product. This A-overhang would then allow for the adaptors to ligate in the next step. The reaction mix (Table 30) was incubated at 37°C for 30 mins and then cleaned up by Qiagen MinElute PCR Purification Kit (QIAGEN, Crawley, UK), eluting in 10.2µl EB.

Component	Volume
DNA	34.45µl
Klenow 10X Buffer	5µl
DNA Polymerase I Large (Klenow) Fragment, Exonuclease Minus:	0.55µl
1mM dATP	10µl
Total volume:	50µl

Table 30. Reaction mix for adenosine (A)-addition

3.3.6 Ligation of adaptors

Custom adaptors (Morgan et al., 2010) were ligated to the end-repaired A-overhanged PCR products with the LigaFast Rapid DNA Ligation System (Promega, Madison, USA). The reaction mix (Table 31) was incubated at 25°C for 15 mins and then 65°C for 20 mins to deactivate the enzyme.

Component	Volume
DNA	10µl
2X Rapid Ligation Buffer	15µl
T_4 DNA Ligase	3µl
H_2O	1µl
Adaptor (2µM)	1µl
Total volume:	30µl

Table 31. Reaction mix for adaptor ligation

The reaction was cleaned up with Agencourt AMPure XP magnetic beads (Beckman Coulter, High Wycombe, UK). The standard clean up protocol was followed except the ratio of beads was halved to 0.9X in order to better remove unwanted adaptor-dimers of around 100bp. Therefore for a 30µl reaction volume; 27µl of Agencourt AMPure XP (Beckman Coulter, High Wycombe, UK) was added. The final elution step was with 40µl EB.

3.3.7 Enrichment PCR

The final library prep step was enrichment PCR to extend the adaptors and amplify correctly ligated libraries.

Component	Volume
Phusion High Fidelity Mastermix	25 µl
H_2O	13 µl
PE-PTO primer F (25µM)	1 µl
PE-PTO primer R (25µM)	1 µl
DNA	10 µl
Total	50 µl

Table 32. Reaction mix for PCR enrichment

The PCR reaction (Table 32) underwent the following thermocycling programme:

- 98°C for 30 sec

- 12 cycles of:
- 98°C for 10 sec
- 65°C for 30 sec
- 72°C for 30 sec

- 72°C for 5 mins

The final libraries were cleaned up with Agencourt AMPure XP (Beckman Coulter, High Wycombe, UK) following the standard protocol with a concentration of 1.8X beads and eluting into a final volume of 30µl EB. The libraries were tested by Agilent Bioanalyser DNA 1000 LabChip (Agilent, Santa Clara, USA) and Quant-iT PicoGreen assay kit (Invitrogen, Paisley, UK) to assess for library quality and concentration respectively before being pooled together in equal amounts and run on the Illumina sequencer. Figure 43 shows the expected trace from the Agilent Bioanalyser where there is a clear sharp peak at the expected final size and no adaptor-dimer peaks.

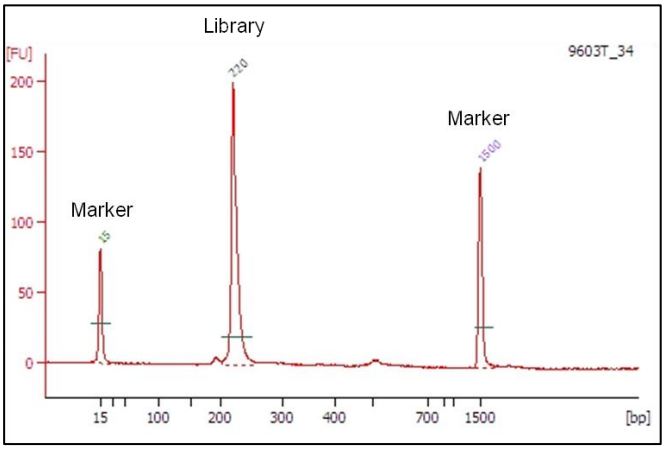


Figure 43. Bioanalyser trace of final PCR amplicon library.

3.3.8 Fixation of cells for DNA to investigate formalin effects

Cells were grown in culture in order for their DNA to be extracted as control DNA and to investigate the effects of fixation on mutation error rate. The cells grown were a breast cancer cell line, MCF7 and two colon cancer cell lines: SW48 and SW480. All cell lines were grown in Roswell Park Memorial Institute (RPMI) medium with glutaMAX and 10% fetal calf serum (FCS) (Invitrogen, Paisley, UK) and cells were grown in a T75 flask to subconfluency and counted.

The old medium was aspirated and the cells were washed with 8ml phosphate buffered saline (PBS). This was then aspirated and 2ml of trypsin was added before incubation for 2 mins at 37°C to trypsinise the cells. After incubation, 8ml of RPMI medium was added to deactivate the trypsin before the cells were transferred to a centrifuge tube. They were spun at 1000g for 3 mins to form a cell pellet before the old medium was aspirated and the cells were resuspended in 8ml PBS. This was then split into aliquots so that there were approximately $5x10^6$ cells per aliquot. The aliquots were then spun again at 1000g for 5 mins to pellet the cells, the old medium was aspirated and the cells were resuspended in 200µl PBS and transferred to a microcentrifuge tube for DNA extraction.

In order to investigate the effects of fixation on sequencing error rate, cells were fixed in differing percentages of formalin. Aliquots of cells were prepared and resupsended in 8ml of 1%, 5% or 10% formalin. They were incubated on a roller at room temperature (20°C -25°C) for 30 mins to allow the cells to fix. They were then spun at 1000g for 3 mins to pellet the cells and the formalin discarded. The cells were washed with PBS, spun as before and the PBS discarded and then resuspended in 200µl fresh PBS for DNA extraction. DNA from cells was extracted with the Qiagen QIAamp DNA Mini Kit (QIAGEN, Crawley, UK) in accordance with the standard manufacturer's protocols and eluted into a 50µl volume of H_2O.

The DNA from cells fixed at differing percentage formalin was then amplified for 5 targets: KRAS codons 12&13, KRAS codon 61, NRAS codons 12&13, NRAS codon 61 and BRAF codon 600. The primers used are outlined in Table 33 and were amplified with Phusion polymerase (section 3.3.1). Primers were designed using pyrosequencing assay design software (Biotage AB, Uppsala, Sweden) as outlined by Dr. Phil Chambers (Chambers et al., 2013). PCR amplicons were quantified by Invitrogen's Quant-iT PicoGreen dsDNA BR assay kit (Invitrogen, Paisley, UK) before all the 5 targets were pooled for each sample and underwent standard library preparation (sections 3.3.3-3.3.7).

Gene Target	Primer Sequences	Amplicon length (bp)
KRAS codons 12 & 13	Fwd: 5'GGCCTGCTGAAAATGACTGA	80
	Rev: 5'AGCTGTATCGTCAAGGCACTCT	
KRAS codon 61	Fwd: 5'AATTGATGGAGAAACCTGTCTCTT	86
	Rev: 5'TCCTCATGTACTGGTCCCTCATT	
NRAS codons 12 & 13	Fwd: 5'CTTGCTGGTGTGAAATGACTGAG	79
	Rev: 5'TGGATTGTCAGTGCGCTTTTC	
NRAS codon 61	Fwd: 5'GAAACCTGTTTGTTGGACATACTG	83
	Rev: 5'TCGCCTGTCCTCATGTATTG	
BRAF codon 600	Fwd: 5'TGAAGACCTCACAGTAAAAATAGG	91
	Rev: 5'TCCAGACAACTGTTCAAACTGAT	

Table 33. Gene targets and primers used for PCR

3.3.9 Investigating KRAS in cancer-associated normal mucosa

38 samples of FFPE normal colorectal mucosa from patients with tumours with KRAS codon12&13 mutations were processed and underwent DNA extraction and NGS sequencing as previously described (sections 2.3.9 and 3.3.3) alongside a WT control. The tumour samples had been previously confirmed as KRAS mutant by pyrosequencing. The samples of normal mucosa that were sequenced by NGS were repeated in a second independent library preparation process and sequencing run. These samples were also run alongside the tumour samples for those patients to confirm the pyrosequencing mutation calls with NGS.

3.3.10 Ethical approval

Ethical approval for the collection storage and use of patient material was obtained from the national research ethics service (NRES) committee London – Bloomsbury REC reference 12/LO/1217.

3.3.11 Targeted amplicon library creation

TALC is a one-step PCR method for generating targeted libraries developed by me to address the issues faced with adaptor ligation of PCR amplicons. There are 4 primers added to the reaction: the forward and reverse targeted primers, the universal forward primer and the indexing reverse primer. There were 8 targeted primers created for KRAS, NRAS, BRAF and PIK3CA as shown in Table 34.

Gene Target	Primer Sequences	Amplicon length (bp)	Final library length (bp)
KRAS codons 12/13	Fwd: 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAATGGTCCTGCACCAGTAAT 3'	168	290
	Rev: 5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGCCTGCTGAAAATGACTGAAT 3'		
KRAS codon 61	Fwd: 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCTTACACAAAGAAAGCCCTCCCCAG 3'	154	276
	Rev: 5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGACTGTGTTTCTCCCTTCTCAGG 3'		
KRAS codon 146	Fwd: 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGTATTTATTTCAGTGTTACTTACCTGTCT 3'	161	283
	Rev: 5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACTCTGAAGATGTACCTATGGTCCT 3'		
NRAS codons 12/ 13	Fwd: 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCTGACAAGTGAGAGACAGGATCAGG 3'	158	280
	Rev: 5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGACTGAGTACAAACTGGTGGTG 3'		
NRAS codon 61	Fwd: 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTGATGGCAAATACACAGAGGAA 3'	150	272
	Rev: 5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCCCCAGGATTCTTACAGAAAACA3'		
BRAF codon 600	Fwd: 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCCTCAATTCTTACCATCCACA 3'	163	285
	Rev: 5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACTGTTTTCCTTTACTTACTACACCTCA 3'		
PIK3CA codons 542/ 545	Fwd: 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTCAAAGCAATTTCTACACGAGAT 3'	173	295
	Rev: 5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCCAATAGGTATGGTAAAAACATGCTG 3'		
PIK3CA codon 1047	Fwd: 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTTAGATAAAACTGAGCAAGAGG 3'	170	292
	Rev: 5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGTGGAATCCAGAGTGAGCTTTC 3'		

Table 34. Targeted primers used with TALC.

The PCR mastermix was made using the FastStart High Fidelity PCR system (Roche, Basel, Switzerland) and was made for each target with forward and reverse targeted primers and the universal primer. 22µl of mastermix was added to 2µl of gDNA at 10ng/µl and 1µl of bar-coding primer at 12.5µM. The reaction mix is shown in Table 35 and the thermocyling conditions are shown in Table 36.

Reagent	Volume (µl)	Final Concentration
FastStart High Fidelity Buffer without MgCl2 (10x)	2.5	1x
$MgCl_2$ (25mM)	4.5	4.5mM
DMSO	1.25	5%
Nucleotides (10mM)	0.5	0.2µM
FastStart High Fidelity Enzyme (5U/µl)	0.25	
Forward Target Primer (25µM)	0.5	500nM
Reverse Target Primer (25µM)	0.5	500nM
Universal Primer (25µM)	0.5	500nM
H_2O	11.5	
Barcode Primer (12.5µM)	1	500nM
DNA (10ng/µl)	2	
Total	25	

Table 35. Reaction mix for TALC

Temperature (°C)	Time	Number of cycles
50	2 mins	1
70	20 mins	
95	10 mins	
95	15 sec	10
60	30 sec	
72	1 min	
95	15 sec	2
80	30 sec	
60	30 sec	
72	1 min	
95	15 sec	8
60	30 sec	
72	1 min	
95	15 sec	2
80	30 sec	
60	30 sec	
72	1 min	
95	15 sec	8
60	30 sec	
72	1 min	
95	15 sec	5
80	30 sec	
60	30 sec	
72	1 min	

Table 36. Thermocycling protocol for TALC

All samples were then run on an Agilent Bioanalyser DNA 1000 LabChip (Agilent, Santa Clara, USA) to quantify the amount of library. Firstly, targets for each patient were pooled. Finally each barcoded patient pool was combined to form the final library. This was then run on a 2.5% agarose gel and the final library band at 300bp was excised and purified using Qiagen's MinElute gel purification kit (QIAGEN, Crawley, UK) eluting in 20µl EB. 31 FFPE samples underwent TALC library prep that had previously been sequenced with adaptor-ligated libraries.

3.4 Bioinformatics

3.4.1 Demultiplexing and quality filters

Two methods of demultiplexing samples from within the run data were tested. Firstly reads were separated by barcode only using a custom perl script written by Dr. Stefano Berri. The input contained all of the reads from the sequencing run and the script outputted de-multiplexed reads into individual files with the barcode trimmed from the 5' end of the read. The second method used a programme designed by Dr. Ian Carr, AgileQualityFilter:

(http://dna.leeds.ac.uk/agile/AgileQualityFilter/) which demultiplexed and filtered the reads according to user inputted thresholds outlined in Table 37.

Option	Setting	Description
Maximum number of uncalled bases per sequence	0	Due to the high amount of coverage obtained, a high threshold was set, where any reads containing "N" bases within the sequence were filtered out
Minimum score to call a base	20	Reads containing bases with a Phred score less than 20 (miscall probability higher than 1%) were filtered out
Minimum median quality score for a read	36	Within each read, the median Phred score of all the bases had to be a minimum of 36 (miscall probability of 0.025%) or the read was filtered out

Table 37. Parameters used with AgileQualityFilter software

Filtering kept on average 89% of the reads for downstream analysis and improved the mean Phred scores of the reads by 2.52 points, which is a 1.7-fold improvement in base-call error probability from 0.05% to 0.03%. The effects of filtering on 4 independent runs from the Illumina GAIIe are shown in Figure 44. It was therefore decided to include this initial filtering step in the bioinformatics pipeline, to improve quality scores and therefore improve confidence of mutant calls downstream.

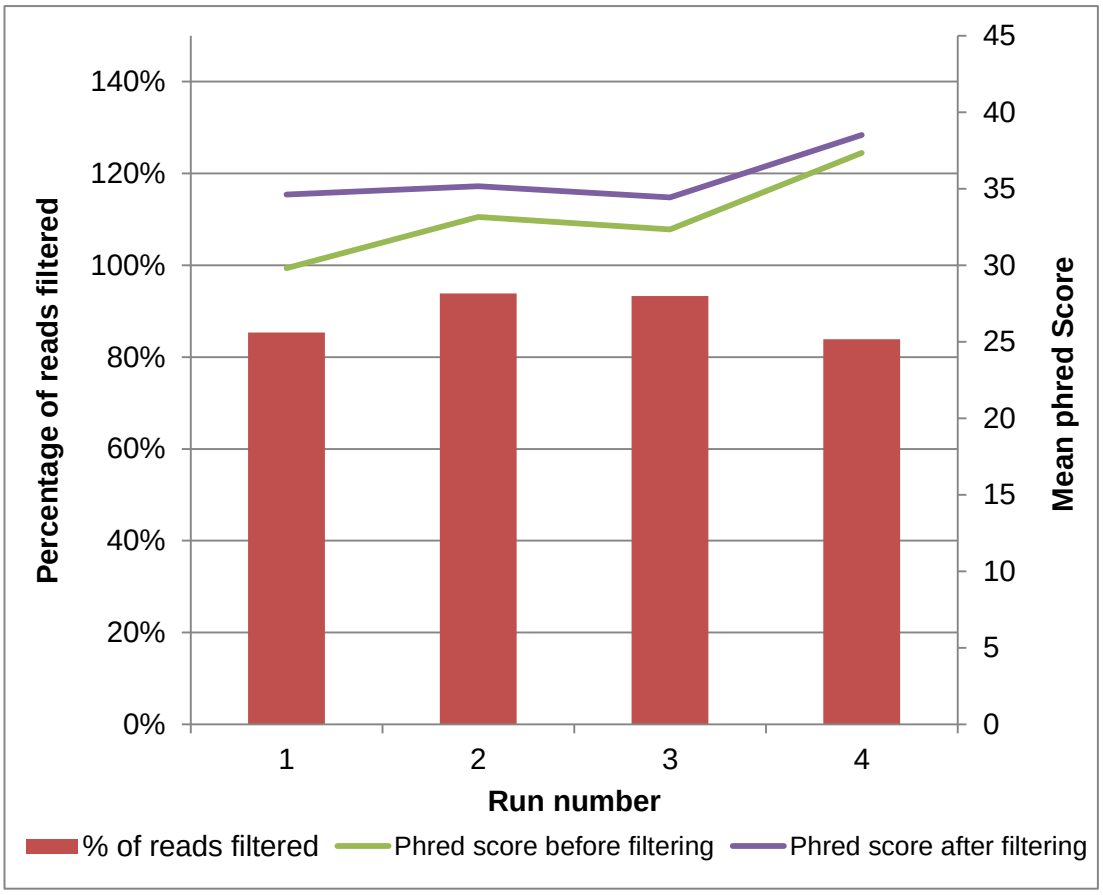


Figure 44. Effect of filtering on percentage of reads maintained and the mean Phred quality score of reads.

3.4.2 Look-up method for variant detection

This approach was used in order to produce a mutation report quickly and easily directly from filtered reads. Initially, custom perl scripts written by Dr. Graham Taylor were developed to search for a short sequence that was either WT or mutant within the reads (Chambers et al., 2013). This was further developed into a programme with a user-interface that could be run on the Windows operating system by Dr. Ian Carr, AgileVariantFastaFinder:

(http://dna.leeds.ac.uk/agile/AgileFastaVariantFinder/).

This method does not require reads to be aligned to a reference sequence, making it ideal for targeted sequencing with low numbers of targets. The alignment process can be difficult to run on a Windows platform, requiring the user to have a basic knowledge of Linux and some computer programming skills. Also alignments can be time consuming and memory-heavy for the computer. In cases where only a few target sequences are to be interrogated, full alignment is not necessary and the

coverage at each site may be in the 1000s. This programme works for analysing oncogenes where only mutational hotspots are checked for mutations. However in the case of TSGs, an alignment approach is needed due to mutations occurring at any point within the tested amplicon.

3.4.3 Alignment

Aligning reads for targeted sequencing allows for data to be passed into a pipeline using published software to make variant calls. This is especially useful when there are a large number of files with a large number of targets and allows the process to be easily automated. Also alignment is required when allele frequencies at each position within the read are to be assessed in order to estimate the error rate from the percentage of non-reference bases.

Alignments were performed using the Burrows-Wheeler Aligner (BWA) software (Li and Durbin, 2009) available at: http://bio-bwa.sourceforge.net. Reads were aligned to custom reference sequences containing the amplicons rather than the whole genome in order to speed up the alignments. The reference sequence was indexed using BWA with default settings. Alignments were performed using the BWA aln algorithm with the seed function disabled. A seed is a certain number of bases at the start of the read, set by the user, in which no mismatches are allowed. This is important for aligning whole genome or whole exome data, however for targeted PCR amplicons a seed would interfere with downstream analysis. All other options for BWA were in accordance with the default settings. After alignment, SAM files were outputted for each sample which could then be subject to further analysis.

3.4.4 Assessing allele frequencies and coverage

In order to determine the error rate of sequencing, the percentage of non-reference bases at each position within the read was determined with a custom perl script written by myself AllFreqChecker.pl (see appendix for full script). It uses SAM files for aligned reads and the fasta file of the reference sequence. The script works by checking which target the sequence aligns to and at what position. It then checks each base within the read and if it does not match the reference, it will check which base it is and keep a tally. Once this has been repeated for all reads of all targets within a file it calculates the percentages of each non-reference base within the reads for a certain file. Alongside this it produces a total of the number of reads aligned to each target for each sample which can be used to produce a coverage report.

3.5 Results

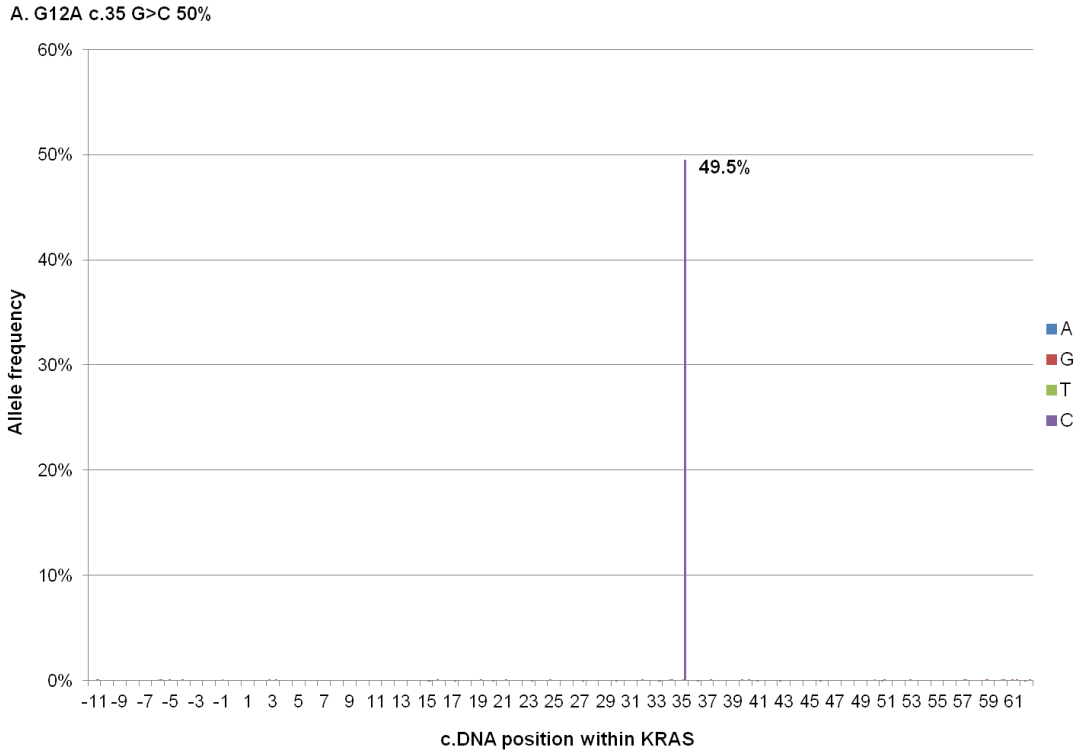
3.5.1 Limit of detection of NGS

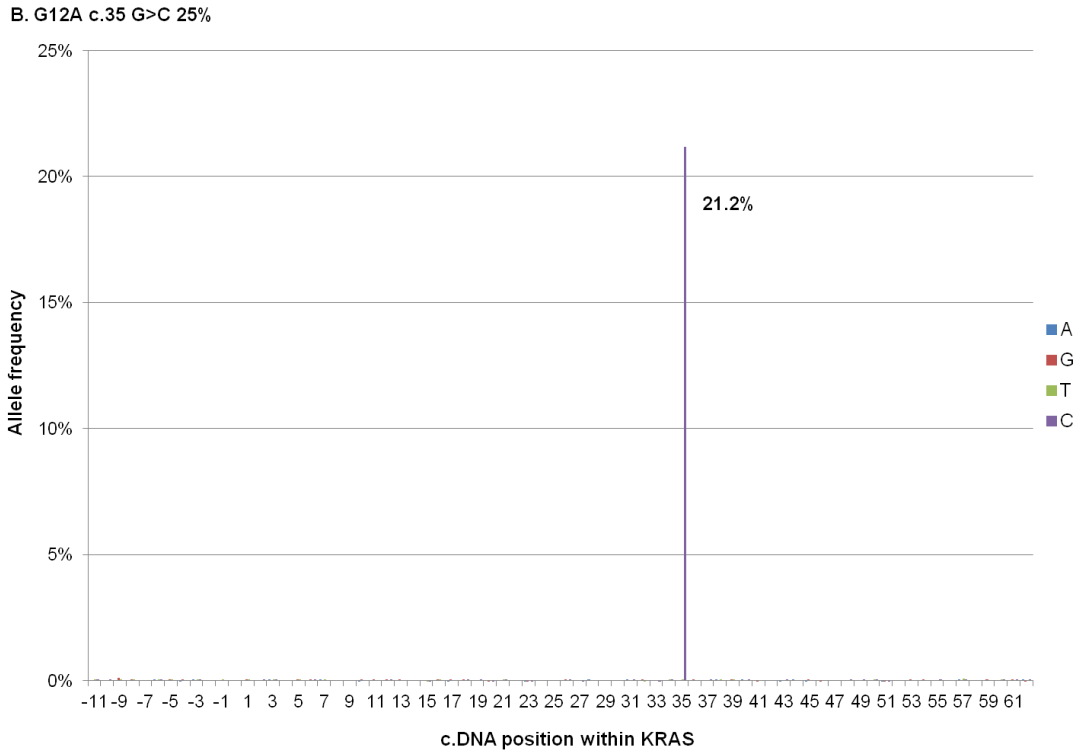
By assessing the mutant allele frequency of serial dilutions of KRAS mutant DNA, the lowest allele frequency of inputted DNA that could be detected was 0.5% as shown in Table 38. This was replicated in two different KRAS mutants: G12A c.35G>C and G12C c.34G>T and for two independent library preparations and runs. Figure 45 A-H shows the plot of non-reference bases at each position within the amplicon for one run of the G12A mutant serial dilution. Other runs produced similar looking plots.

Mutation tested and frequency		Mutation detected	Run1 allele frequency	Run2 allele frequency
c.35 G>C	50%	c.35G>C	49.53%	51.33%
	25%	c.35G>C	21.18%	19.50%
	5%	c.35G>C	4.56%	1.26%
	2.5%	c.35G>C	1.98%	1.92%
	1%	c.35G>C	0.64%	0.63%
	0.5%	c.35G>C	0.14%	0.21%
	0.05%	WT	0%	0%
	0.005%	WT	0%	0%
c.34 G>T	50%	c.34G>T	54.14%	57.73%
	25%	c.34G>T	20.04%	23.61%
	5%	c.34G>T	3.23%	2.93%
	2.5%	c.34G>T	1.50%	1.53%
	1%	c.34G>T	0.87%	0.69%
	0.5%	c.34G>T	0.31%	0.25%
	0.05%	WT	0%	0%
	0.005%	WT	0%	0%

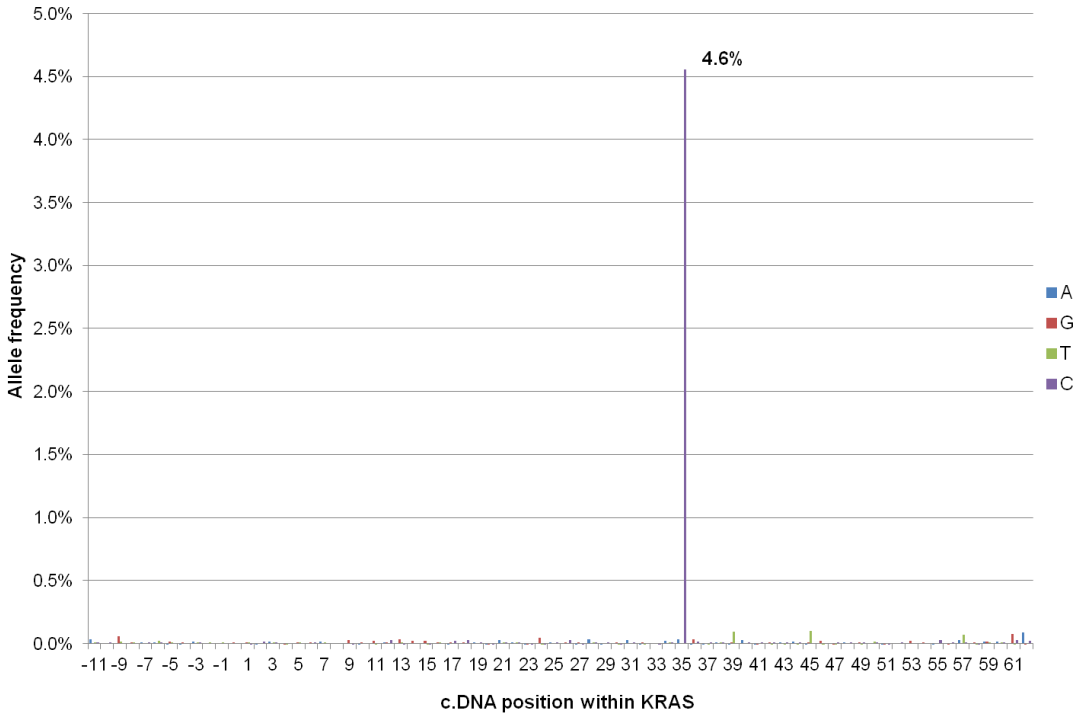
Table 38. Detected mutant allele frequencies for serial dilutions of mutant KRAS DNA

Figure 45. A-H. Non-reference allele frequencies for serial dilutions of G12A c.35G>C mutant KRAS showing minimum detection level of 0.5% mutant allele frequency.

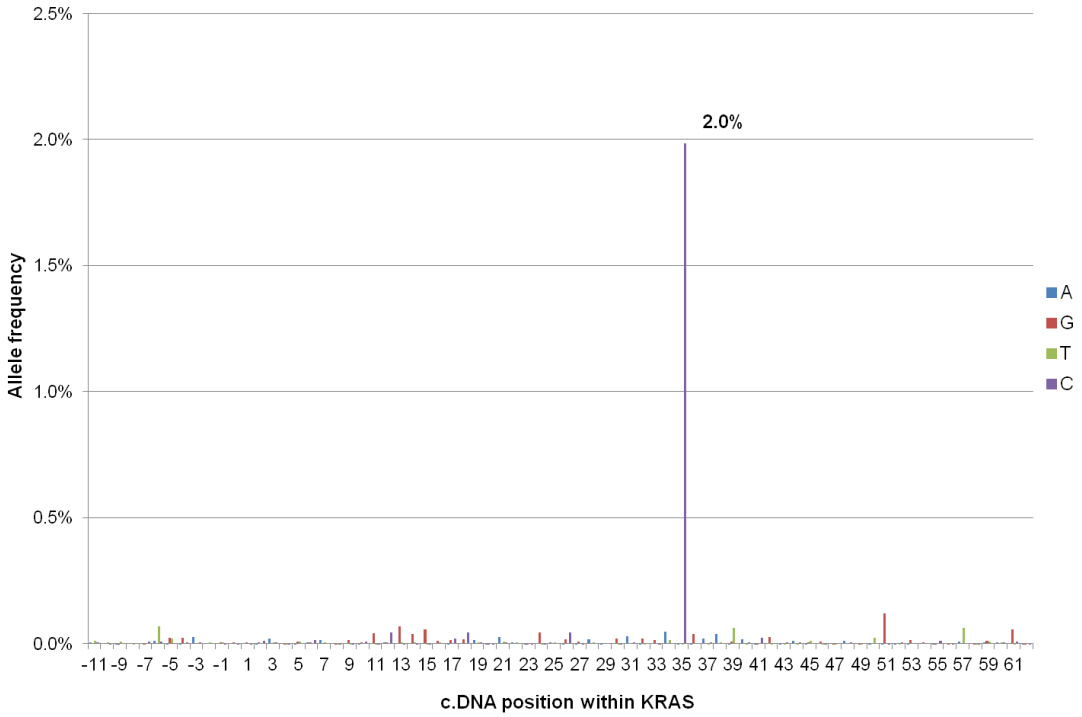




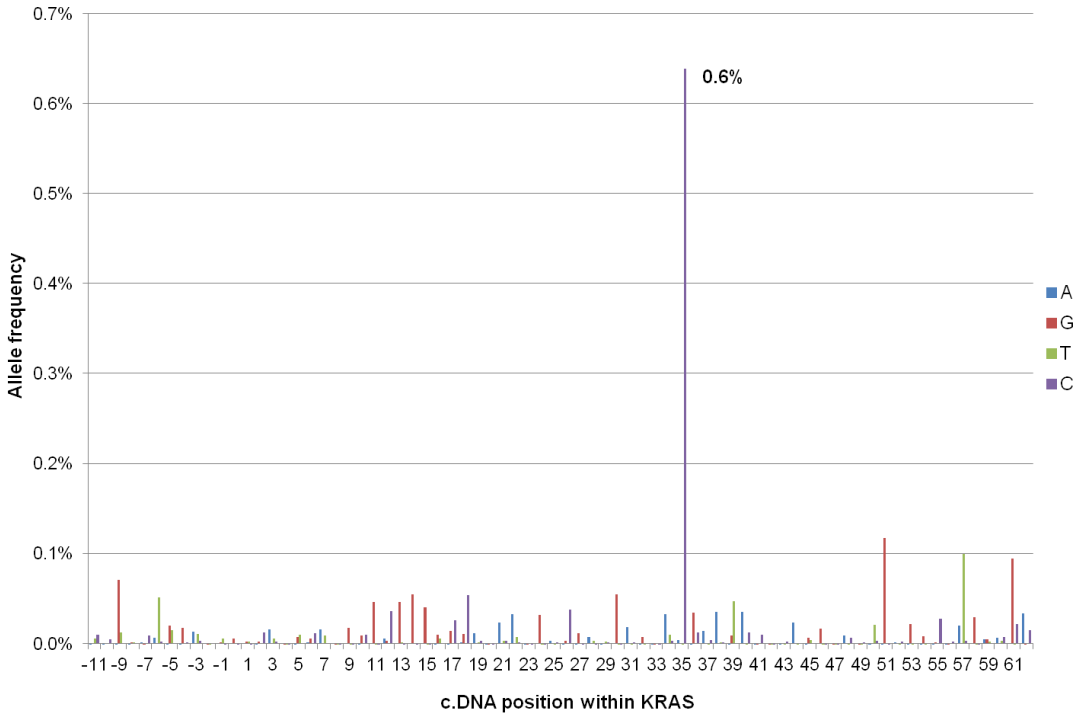
C. G12A c.35 G>C 5%



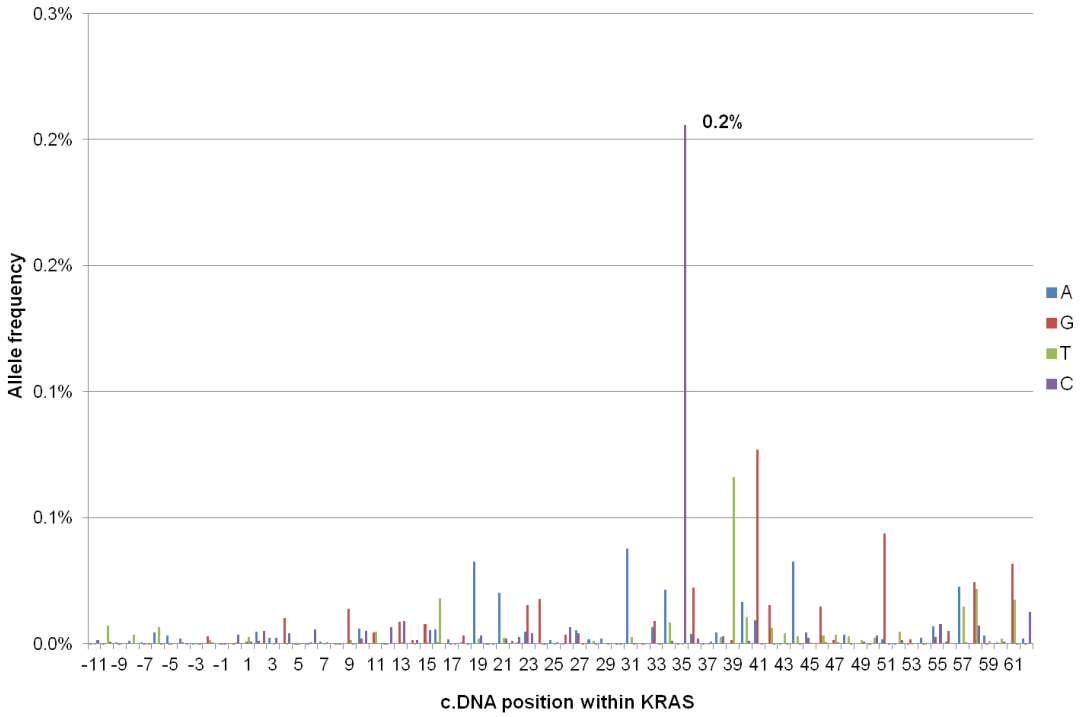
D. G12A c.35 G>C 2.5%



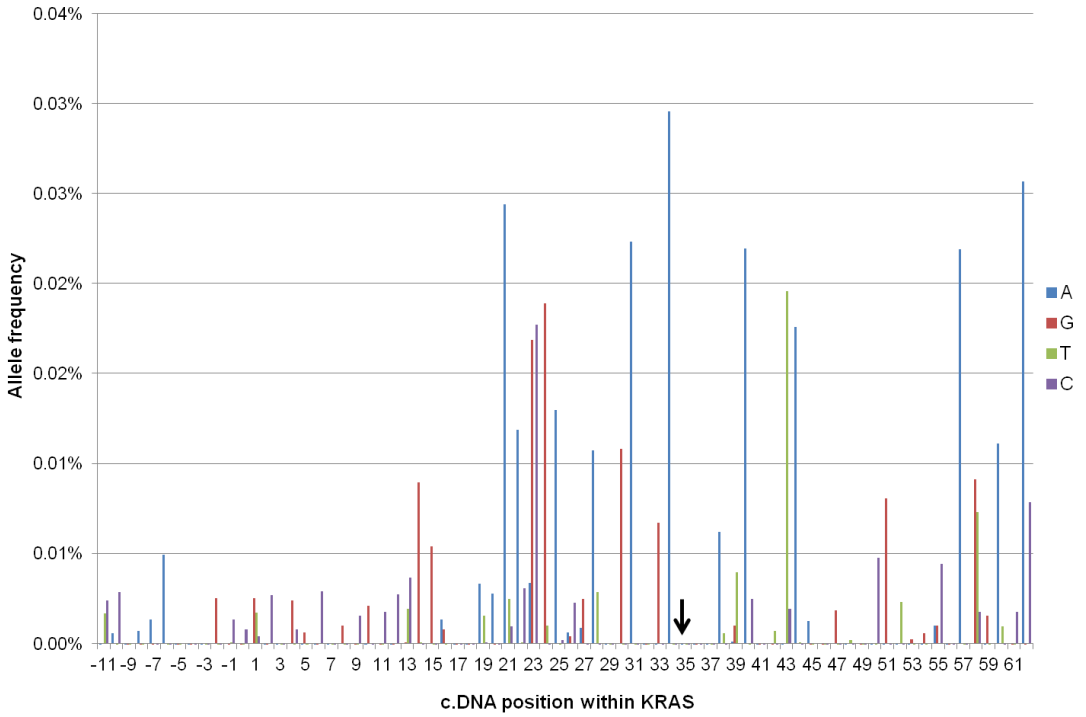
E. G12A c.35 G>C 1%



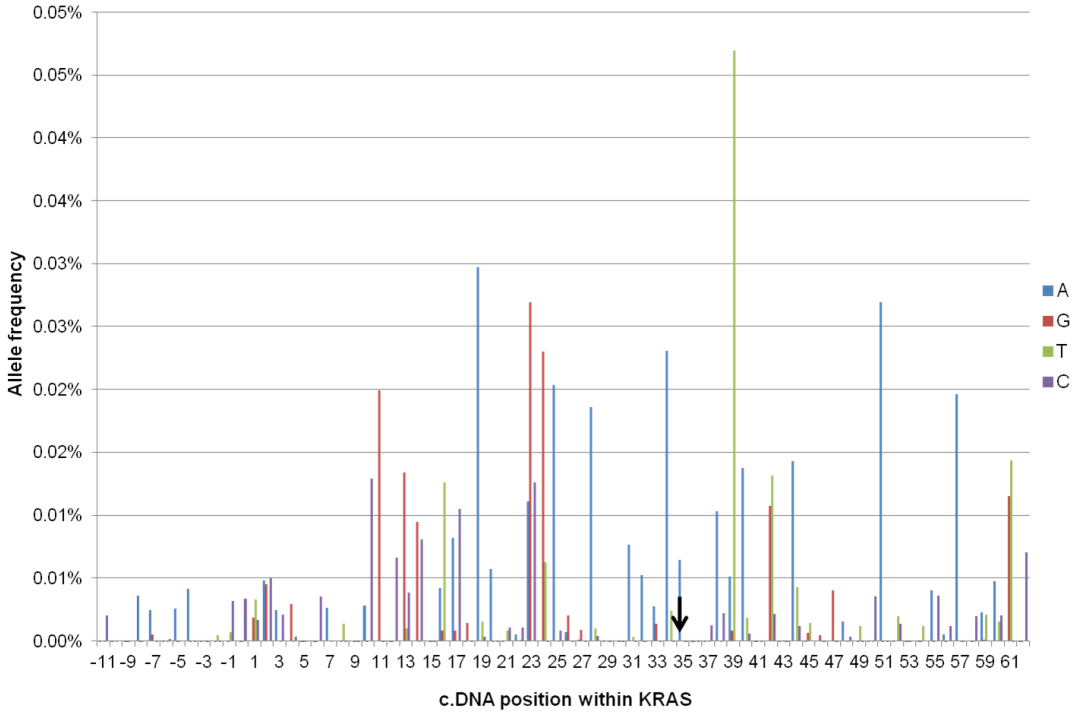
F. G12A c.35 G>C 0.5%



G. G12A c.35 G>C 0.05%



H G12A c.35 G>C 0.005%



3.5.2 Enrichment with RFLP and NGS

Table 39 shows the enriched allele frequencies detected for the G12A mutation and G12C mutation. The maximum enrichment gained was 200-fold. RFLP enabled the G12A mutation to be detected at a level of 0.05% (Figure 46) and the G12C mutation at 0.005% (Figure 47). However, there was a high level of contamination and error within the PCR product and therefore it was decided that this method was not reliable or consistent enough to use with NGS.

	Mutant dilution	Mutant allele frequency detected	Enrichment (fold)
G12A	50%	80%	1.6
	25%	88%	3.52
	5%	53%	10.6
	2.5%	28%	11.2
	1%	8%	8
	0.5%	11%	22
	0.05%	2%	40
	0.005%	0%	-
G12C	50%	39%	0.78
	25%	29%	1.16
	5%	11%	2.2
	2.5%	11%	4.4
	1%	0%	-
	0.5%	0%	-
	0.05%	9%	180
	0.005%	1%	200

Table 39. Mutant allele frequencies detected for serial dilutions of G12A and G12C KRAS mutant DNA showing the amount of enrichment attained.

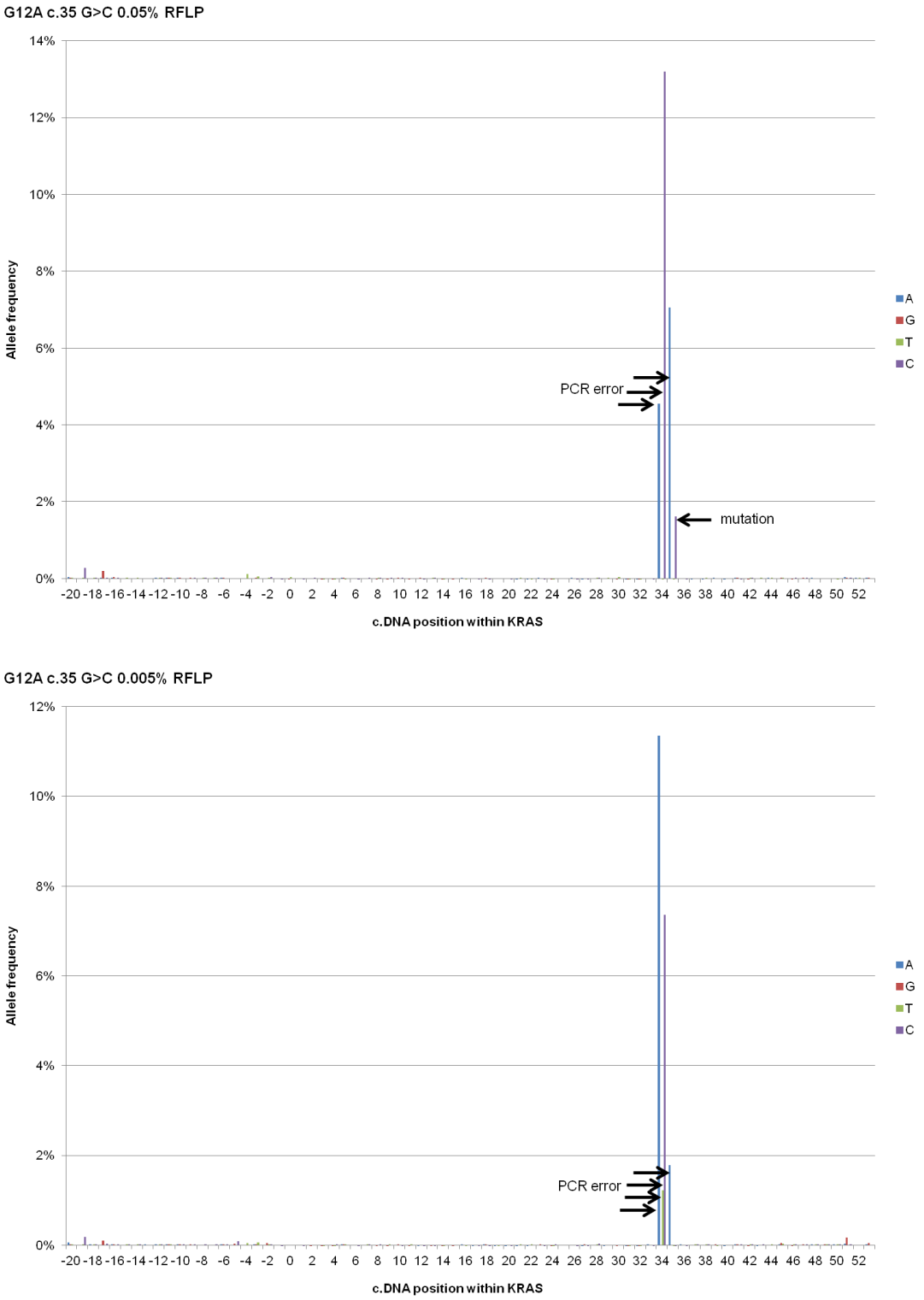


Figure 46. Non-reference allele frequencies for 0.05% and 0.005% G12A c.35G>C mutant KRAS enriched with RFLP PCR showing a minimum detection level of 0.05% of the mutant allele and a high frequency of PCR errors around the site of interest.

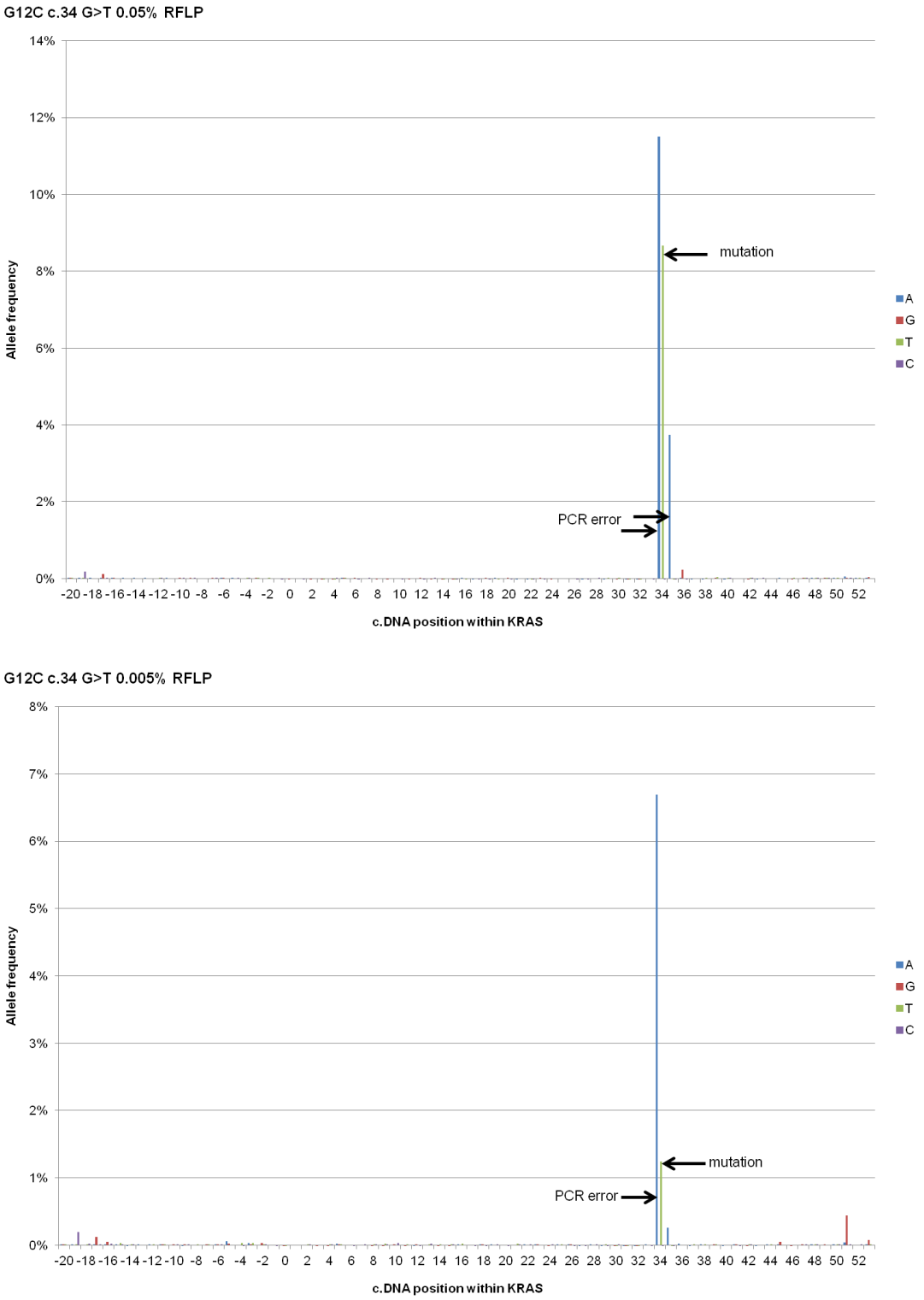


Figure 47. Non-reference allele frequencies for 0.05% and 0.005% G12C c.34G>T mutant KRAS enriched with RFLP PCR showing a minimum detection level of 0.005% of the mutant allele and a high frequency of PCR errors around the site of interest.

3.5.3 Effect of formalin of cells and error rates

Fixing cell lines in different percentages of formalin had no effect on the sequencing error rate as show in Figure 48 A-B. This was the case for 2 different cell lines that had libraries run in duplicate. The sequencing error rate was calculated by taking the median of all the non-reference allele frequencies from each position within each target. The overall median non-reference allele frequency was calculated as 0.1%.

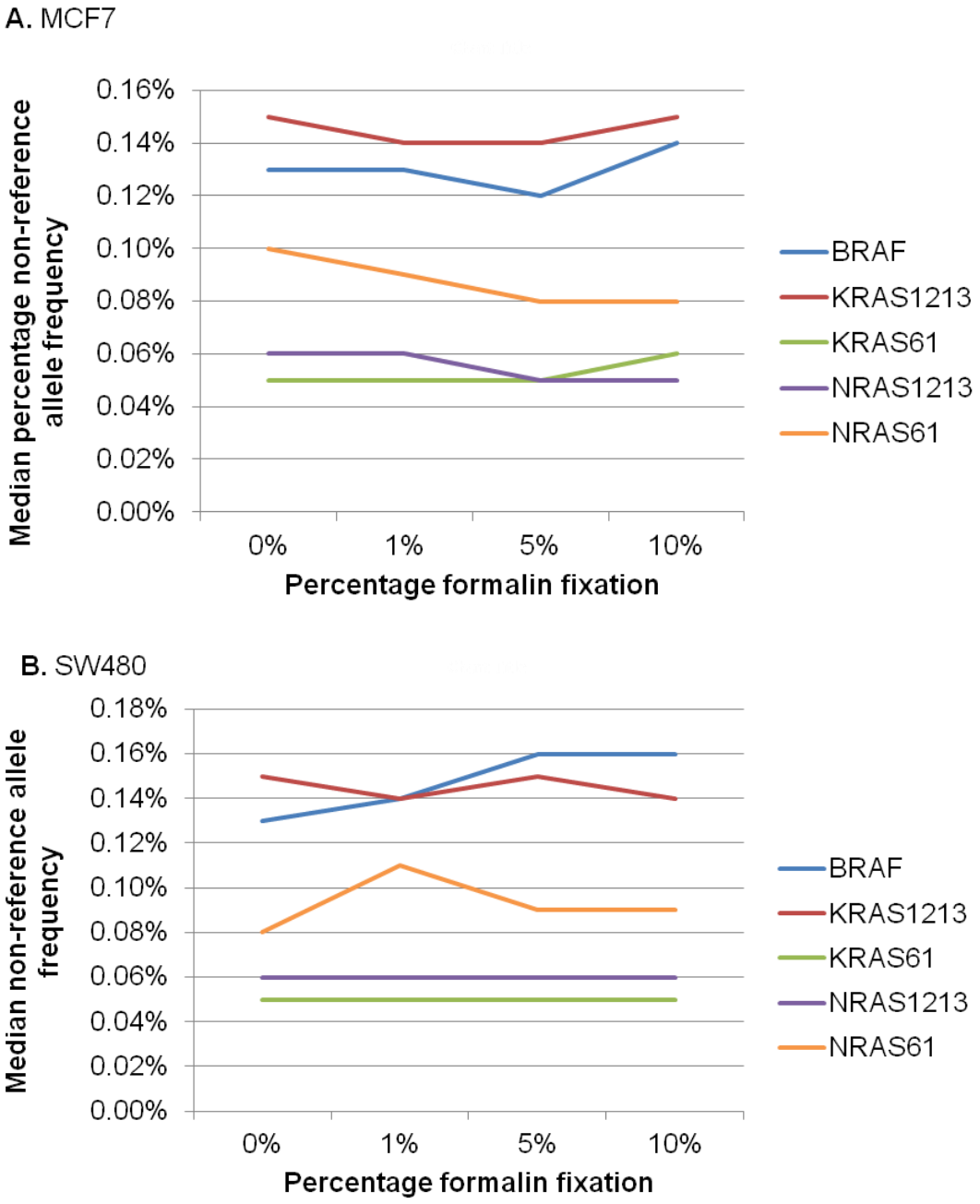


Figure 48. A &B. Median non-reference allele frequency for 5 amplicons in A. MCF7 cell line DNA and B. SW480 cell line DNA that has been fixed with 0%,1%15% and 10% formalin.

The percentage of G>A base changes increased with 5% and 10% formalin fixation for the MCF7 cell line DNA (Table 40) however this was not found to be statistically significant. No significant differences were seen in any of the other base changes. For SW480 there were no significant differences in any of the base changes for any percentage of formalin fixation (Table 41).

	MCF7 formalin fixation			
Base change	**0%**	**1%**	**5%**	**10%**
A>G	17%	17%	16%	15%
A>T	3%	6%	5%	5%
A>C	26%	26%	27%	27%
G>A	9%	11%	14%	14%
G>T	7%	5%	6%	5%
G>C	6%	1%	2%	2%
T>A	2%	2%	2%	2%
T>G	8%	9%	8%	8%
T>C	11%	11%	11%	11%
C>A	4%	6%	4%	4%
C>G	4%	1%	2%	3%
C>T	3%	4%	4%	4%

Table 40. Percentage of base changes in background noise for MCF7 cell line DNA fixed at different percentages of formalin showing an increase in the proportion of G>A changes.

	SW480 formalin fixation			
Base change	**0%**	**1%**	**5%**	**10%**
A>G	18%	17%	19%	18%
A>T	3%	3%	5%	3%
A>C	29%	31%	25%	29%
G>A	6%	7%	7%	7%
G>T	11%	7%	7%	9%
G>C	2%	2%	6%	2%
T>A	2%	2%	2%	3%
T>G	8%	9%	10%	9%
T>C	11%	11%	10%	10%
C>A	4%	4%	6%	5%
C>G	2%	4%	2%	2%
C>T	3%	4%	4%	4%

Table 41. Percentage of base changes in background noise for SW480 cell line DNA fixed at different percentages of formalin showing no significant differences.

3.5.4 Detection of KRAS mutations in cancer-associated normal mucosa

Out of the 39 samples of DNA from cancer-associated normal mucosa, 38 were successfully sequenced. Mutations were detected by AgileFastaVariantFinder in 11 (29%) ranging from 1.3% - 8.7% mutant allele frequency (Figure 49). The mutation calls from both repeats of these samples are shown in Table 42. Data files were visualised by running them through the AllFreqChecker.pl script, and an example is seen in Figure 50. These mutations were confirmed on a second independent run with an independent library preparation. Out of those 11 samples with KRAS mutations in normal, only 3 contained the same KRAS codon12&13 as the tumour sample from the same case. Figure 51 shows the frequency distribution for KRAS found in CRC according to the COSMIC database compared to the distribution found in the 11 mutated normal samples.

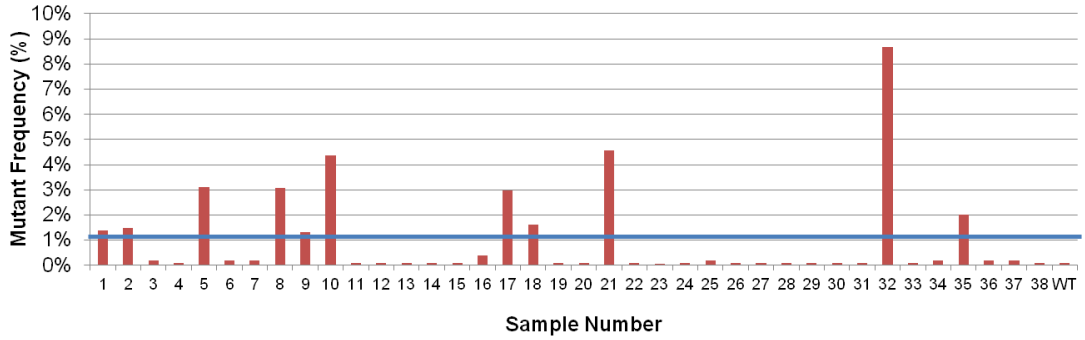


Figure 49. Mutant allele frequencies in 38 clinical samples of cancer-associated normal mucosa showing mutant allele frequency for each of the 38 samples plus a WT control, with detected mutations in 11/38.

Sample	% mutant run 1	% mutant run 2
1	1.38%	2.02%
2	1.48%	1.64%
5	3.10%	1.70%
8	3.06%	2.31%
9	1.33%	1.58%
10	4.37%	5.00%
17	2.96%	2.49%
18	1.60%	1.51%
21	4.55%	4.09%
32	8.68%	8.21%
35	2.06%	2.79%

Table 42. KRAS mutant allele frequencies from duplicate runs of 11 samples with mutations in normal mucosa.

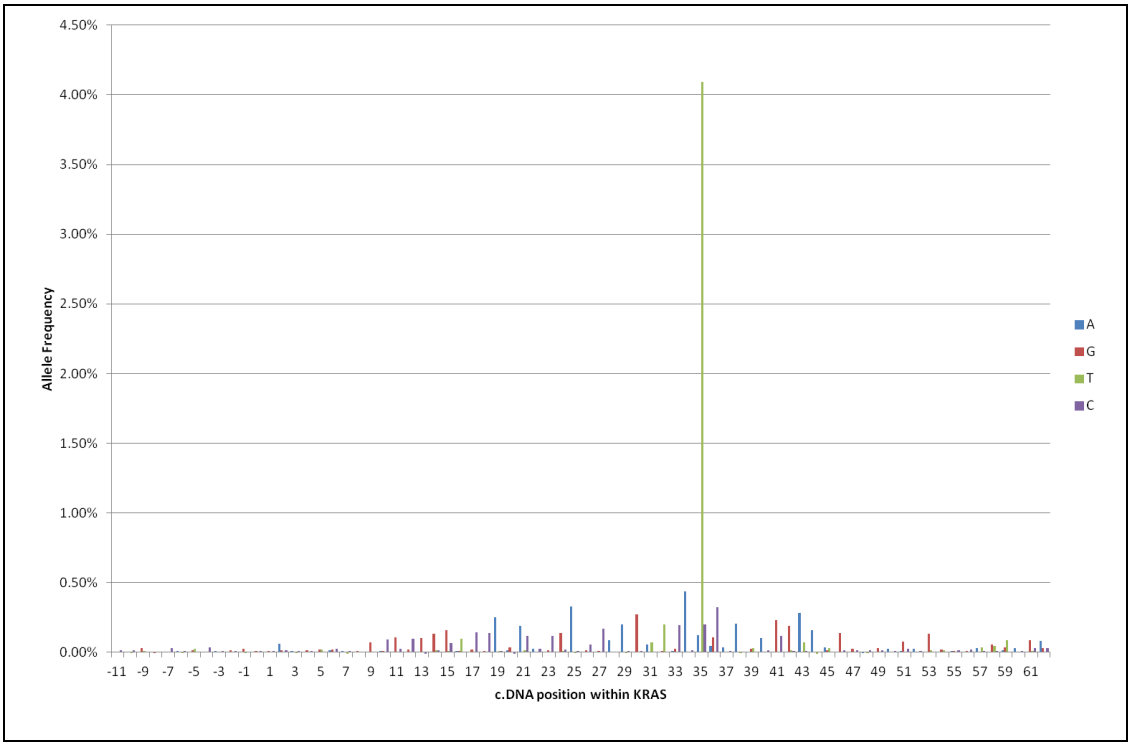


Figure 50. Example of allele frequencies for KRAS amplicon in normal mucosa sample

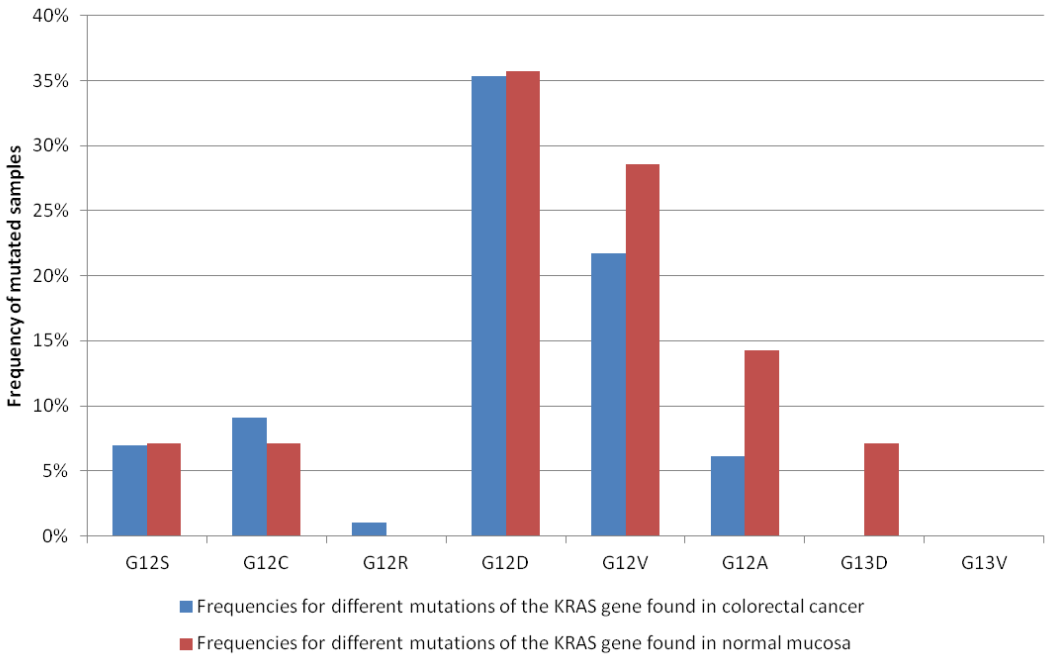


Figure 51. Frequency distribution of the different KRAS codon 12&13 mutations found in colorectal cancer from the COSMIC data base compared to those detected in normal mucosa

3.5.5 Comparison of TALC to adaptor ligation

TALC produced equivalent results compared to adaptor-ligated PCR libraries in terms of mutation calls and percentage allele frequencies. Figure 52 shows results for 23 samples prepared for sequencing by adaptor-ligation and TALC. The concordance between the two methods produced a Pearson's r value of 0.9968 showing very high correlation between the two methods. For the 23 samples prepared for sequencing by both methods the mean difference in percentage allele frequency was 1.4% (σ = 1.9%). The median error rate of the 23 samples tested for KRAS by TALC was 0.4%. There was no trend in the error rate across the amplicon as illustrated in Figure 53.

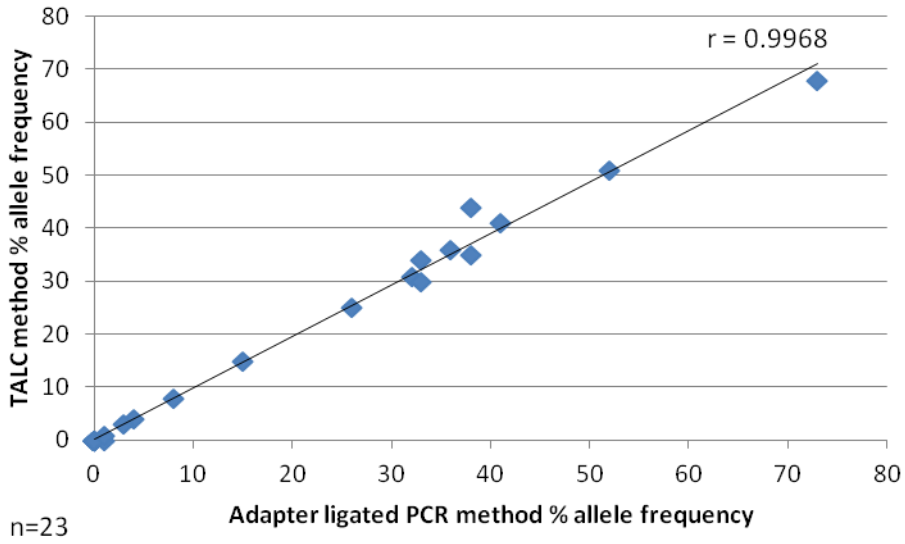


Figure 52. Correlation between adaptor-ligated PCR libraries and TALC for calling mutant allele frequencies

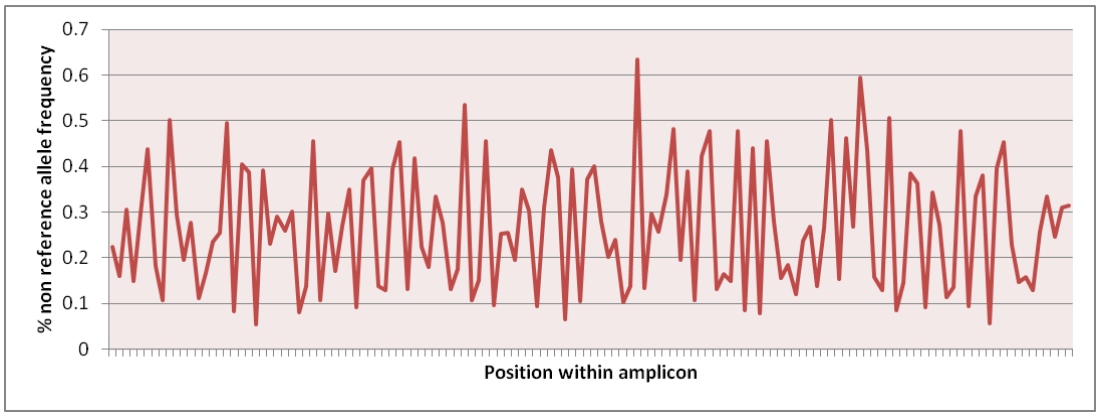


Figure 53. Error rates across the KRAS amplicon

3.6 Discussion

3.6.1 Limit of detection of NGS

There is a range of reported sensitivities from 0.1% mutant allele frequency (Flaherty et al., 2012) to 2% (Milbury et al., 2012) depending on the targets of interest, coverage and the bioinformatics approaches used. In the case of the serial dilution study of fresh cell line DNA, we were able to detect 0.5% mutant DNA within a WT background reproducibly for two different KRAS codon 12&13 mutants. However, when testing clinical FFPE samples, the background noise was higher compared to the fresh cell line DNA at 0.4% (see Figure 50 and Figure 53). For this reason it was decided that the cut-off threshold for FFPE clinical samples should be at 1% therefore allowing for greater certainty when making variant calls and reducing the chance of false positives. It is also recommended to prepare and sequence samples in duplicate to allow for greater certainty when making mutation calls (Schmitt et al., 2012).

High accuracy in mutation calling is especially important when testing clinical samples where patient's therapy is directed by mutation results. It is currently unknown what the significance of a 1% mutation allele frequency holds for patient outcome or response to therapy, but the advance of NGS in recent years has made these studies now possible. It is expected that the incidence of mutations in common cancer genes such as KRAS, NRAS, BRAF and PIK3CA will rise with the increased implementation of high coverage targeted NGS testing due to the higher sensitivity (Hadd et al., 2013). Clinical trials are now stratified according to a tumour genotype that can be targeted with treatment instead of morphology. The high sensitivity of NGS may help to answer questions about response to therapy in future clinical trials. Previously, if a patient's tumour contains low level mutated clones at a frequency below that which could be detected, it may be assumed that their tumour is wild-type. This would result in therapy being directed towards the EGFR-RAS-RAF signalling pathway and in certain cases drug-resistance may develop. The advance of NGS would allow for the identification of these low-level mutant clones which may be selected for by treating with anti EGFR agents and therefore grow, become dominant within the tumour and cause treatment resistance (Moch et al., 2012). This is dependent on the time the clone takes to expand and may be addressed by treatment with multiple agents. If tumour evolution is similar to bacterial or viral evolution then three agents may be necessary as found in mycobacteria tuberculosis or human immunodeficiency virus.

3.6.2 Use of RFLP with NGS

RFLP enabled a 0.05% dilution of mutant G12A c.35 G>C KRAS to be detected and a 0.005% dilution of G12C c.34 G>T mutant. However, the amount of PCR error seen at the site of interest (c.DNA positions 34 and 35) was very high. At low mutant allele frequencies the PCR error was present at an even higher allele frequency. The extra steps of PCR involved in this method means that error is far more likely to become incorporated into the amplicon and be amplified further. Due to the worryingly high level of PCR error seen with this method and the sufficient baseline sensitivity of NGS without enrichment, it was decided not to use RFLP with clinical samples.

3.6.3 Effect of formalin on error rates

For targeted sequencing of KRAS codons12 and 13 the median percentage of non-reference bases across the 80bp amplicon was 0.1% for fresh cell line DNA. After fixing cells in formalin there was no effect on the error rate as shown in Figure 48. This was the case for two independent cell lines that were fixed in increasing percentages of formalin at 1%, 5% and 10% and were tested in duplicate. The error rate for the BRAF and KRAS 12/13 and NRAS 61 amplicons was higher in both cell lines across all levels of formalin fixation. This could be due to the efficiency of the PCR being sequence dependent which may affect PCR error rate. Also the reads will be affected by the amplification of pseudogenes contaminating the final genes. Although the majority of pseudogene reads can be filtered out, some will remain and affect the analysis. The KRAS12/13 and BRAF amplicons contain the most bases differences to the pseudogene amplicons. The NRAS12/13 and NRAS61 primers do not amplify a pseudogene so the final reads will be free of this contamination. This follows the same pattern seen in error rates for the amplicons (Figure 48). Therefore pseudogene contamination will ultimately contribute to the error rate and could partly explain the differences seen in error rate between the amplicons.

There was an increase in the proportion of G>A base changes in the MCF7 cell line DNA that was fixed at 5% and 10% compared to non-fixed DNA (Table 40). Although this difference was not found to be statistically significant this may be due to the small sample size. It is known that formalin induces G>A changes and an increasing trend as the percentage of formalin fixation increases can be seen (Williams et al., 1999). There were no changes seen in the proportion of C>T changes for either cell line and SW48 showed no significant differences in any of the base changes from fixation.

The no change in error rate or the proportion of G>A and G>C base changes between fresh and fixed could be due to one of two reasons. Firstly, it may be possible that formalin fixation has no effect on error rate; however this is highly unlikely due to the median error rates of runs of FFPE 0.4% in contrast to that found for fresh samples (0.1%) in both our studies and the literature (Hadd et al., 2013). It is likely that the experiment used to model the effects of formalin fixation was not an accurate representation of tissue fixation and did not take into account other processes that FFPE tissue undergoes such as hot-plating. When tissue is fixed, it can be immersed in 10% formalin for a number of hours, often overnight, to allow for the formalin to perfuse through the tissue and sometimes a number of days for large resections. Therefore cells are slowly exposed over a long time period. In the case of the cell-line experiment, cells were left in formalin for only half an hour as the cells are free in solution and fixation is instantaneous. As a result, the formalin may not have had the same effect on the cell-line DNA as it would with the DNA in a tumour sample. The error rates observed for the fixed cell-line DNA were the same as those for un-fixed cell line DNA and this suggests that the formalin did not have any effect on the cells in this particular case. However, it is reported that FFPE tumour samples still produce an error rate that is double for fresh (Hadd et al., 2013, Yost et al., 2012) and this was also seen in tumour FFPE samples tested (Figure 50).

3.6.4 KRAS in cancer-associated normal

Next generation sequencing was able to detect mutations in 11 of 38 (29%) cancer-associated normal mucosa samples. These mutations were confirmed on a second independent NGS run (Table 42). It was found that the majority of these mutations did not match those found in the tumour and Figure 51 shows that the mutation distribution matches that seen for KRAS in colorectal cancer. This is a strong indication that these are true mutations that are being detected and not DNA from exfoliated tumour cells co-extracted with the normal mucosa that are being sequenced. The range of mutant allele frequencies found in normal was 1.3% - 8.7% that would usually fall below the detection limit of pyrosequencing.

It is known that KRAS and BRAF mutations are present in hyperplastic polyps and serrated lesions, the latter which are thought to be early cancer precursor lesions (Chan et al., 2003). Therefore if KRAS mutation is an early event, its presence in normal mucosa from cancer patients, whose bowels have been exposed to carcinogens, is not unlikely. The detection of these low level clones suggests that genetic changes are more frequent than thought previously and support the hypothesis that cancer is an evolutionary process where mutations offer a cell a growth advantage. Multiple clones may develop with one becoming dominant which expands and forms the tumour phenotype (Baker et al., 2013). However, other low-level background clones may still be present which would explain why the mutations detected in normal mucosa differed to those observed in the tumour. It was noted in the sample of normal mucosa with the highest rate of KRAS mutant at 8.7% that there were an excess number of bifid crypts. It would be expected that patients with inflammatory bowel disease (IBD) would have a higher crypt turnover rate and therefore have more bifid crypts, however the patient did not have IBD and the pathology report described the samples as normal. An example of the bifid crypts seen in this case is shown in Figure 54.

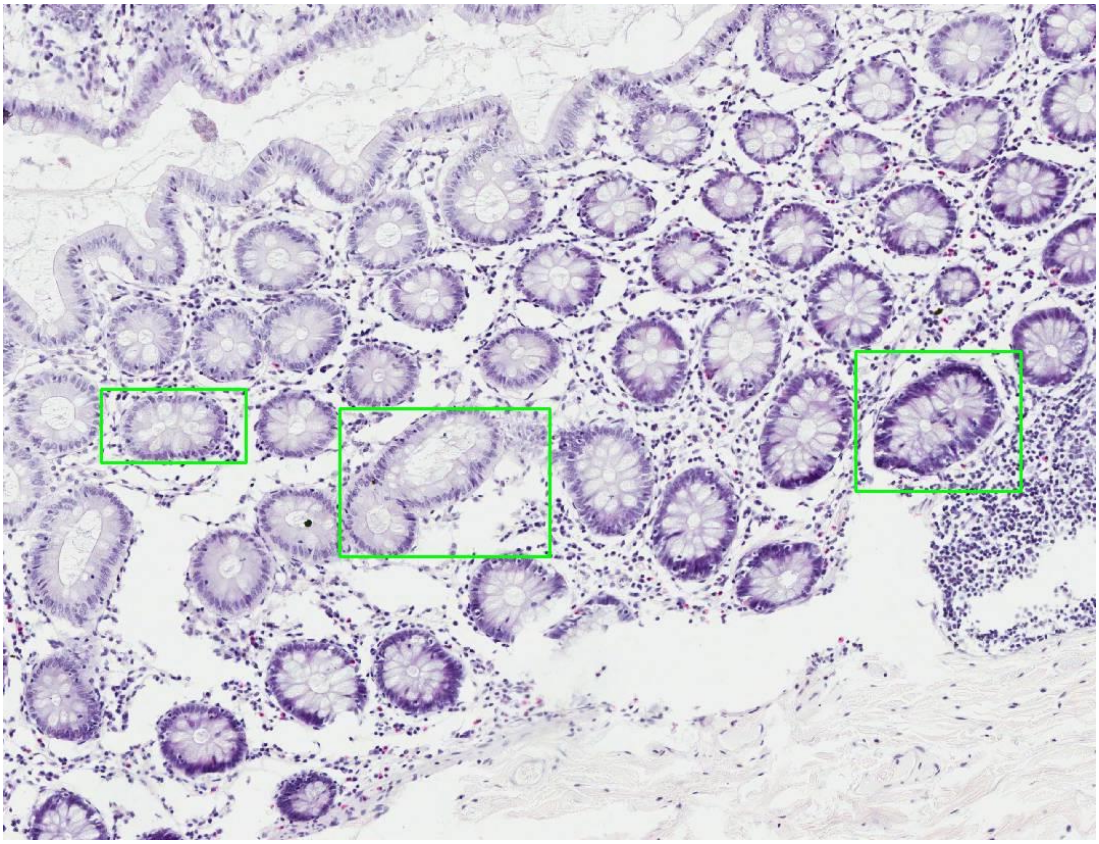


Figure 54. Sample of normal mucosa with high number of bifid crypts, annotated in green.

3.6.5 Usability of TALC

Targeted amplicon library creation (TALC) allows for the generation of indexed, targeted libraries in a one-step PCR. There are several distinct advantages to this method over the traditional adaptor-ligation version of library generation.

1. There is a significant time and cost reduction in generating libraries.

For adaptor-ligation, before library preparation can begin, all samples have to undergo PCR amplification, clean-up, quantification and pooling of targets. The majority of the workflow for TALC is equivalent to this initial PCR amplification step. This also results in a significant reduction in the cost of reagents.

2. The number of preparation steps is dramatically reduced.

This eliminates opportunities for sample contamination and human-error in library preparation.

3. Adaptor-dimers can interfere with the final library during sequencing for adaptor-ligated libraries.

The final gel purification of the pooled library allows for a clean library without adaptor contamination. Adaptor peaks can be difficult to eliminate with adaptor-ligated prepared libraries and reduce the number of useful sequencing runs.

4. Reduction in the number of PCR cycles.

Adaptor ligated libraries undergo a final enrichment step of 15 PCR cycles which increases the risk of introducing artefacts. TALC only requires the initial 35 cycles of amplification and may therefore reduce this artefact risk, which is an important factor in calling variants.

TALC libraries were sequenced with paired-end sequencing and therefore the error rate did not change across the amplicon (Figure 53). The quality scores of the bases reduce along the read and therefore for single-end sequencing it is expected that the error rate will increase as a result. However, paired-end sequencing compensates for this by reading the amplicon in both directions, averaging out the error rate across the read.

3.7 Chapter Summary

- NGS has a limit of detection of 0.5% for fresh DNA and 1% for FFPE DNA.
- RFLP with NGS is not reliable for specific mutation enrichment.
- FFPE DNA samples have a background noise that is higher than fresh DNA.
- Fixing cell-lines directly has no effect on background sequencing noise.
- Mutations can be detected in cancer-associated normal mucosa by NGS.
- Targeted amplicon library creation is an improved methodology for targeted library creation due to reduced costs and time preparation, less contamination probability, and fewer PCR cycles.
- One non-inflammatory bowel disease sample with a very high KRAS mutation rate showed many bifid crypts correlating a mutational change to a phenotype.

4 High throughput mutation detection and copy number changes in early lesions

4.1 Introduction

4.1.1 FAP adenomas as a model of field cancerisation

Field cancerisation refers to the theory that there are multiple clonal patches of pre-cancerous disease that may share genetic aberrations. Familial adenomatous polyposis (FAP) provides a useful model of field cancerisation. Patients affected by this condition inherit a mutated copy of the APC gene in all cells. They also share the same genetic background and modifiers as well as the same micro-biome and thus are an excellent model for studying the development of neoplastic lesions. As the bowel receives a “second hit” within APC, multiple adenomas develop throughout the colon (Will et al., 2010). The second hit is known to be different between different adenomas. Also there has been substantial progress in identifying and characterising the “second hit” of APC and how it relates to the inherited germline mutation (Lamlum et al., 1999, Crabtree et al., 2003, Will et al., 2010). Germline mutations between codons 1285 and 1398 are associated with allelic loss of APC. A germline mutation before codon 1285 is associated with a somatic mutation after codon 1285 whereas a germline mutation after codon 1399 is associated with a somatic mutation before 1284 (Lamlum et al., 1999) Crabtree et al., 2003)

However, other gene mutations that may occur and how these adenomas evolve are less thoroughly investigated. FAP adenomas have been previously reported as polyclonal which has led to the hypothesis of multiple fields of mutant crypts interacting (Thirlwell et al., 2010). A study by Obrador-Hevia et al. revealed that 2 out of 10 adenomas from an FAP patient contained a KRAS codon 12 mutation, however these two mutations were different (Obrador-Hevia et al., 2010). Similarly a study by (Jones et al., 2007) of 22 adenomas from 5 FAP patients revealed that different copy number changes were observed between FAP adenomas. This indicates that although FAP adenomas share a similar genetic background and are within the same environment, it appears individual adenomas may evolve in different ways. There are however no large studies of this process.

4.1.2 High-throughput targeted NGS library creation

Small numbers of genetic targets can be interrogated by NGS by a simple PCR approach such as that outlined in Chapter 2. However, for the purposes of investigating a much broader panel, a multiplexed PCR approach is more suitable. High-throughput multiplex parallel PCR has been claimed to be performed on a microfluidic platform such as the Access Array by Fluidigm (Fluidigm, San Francisco, USA). Each Access Array chip contains 48 sample inlets and 48 assay inlets. PCR primers can be multiplexed within these inlets to create a maximum of 480 amplicons for each of the 48 samples. The Access Array chip contains integrated fluidic circuits (IFC) whereby each sample is mixed with the primers in micro-PCR chambers, which allows for nanolitre reaction volumes. The resulting amplicons can then be barcoded in a simple PCR reaction and then pooled together on one NGS run (Halbritter et al., 2012). An overall schematic of this process is outlined in Figure 55.

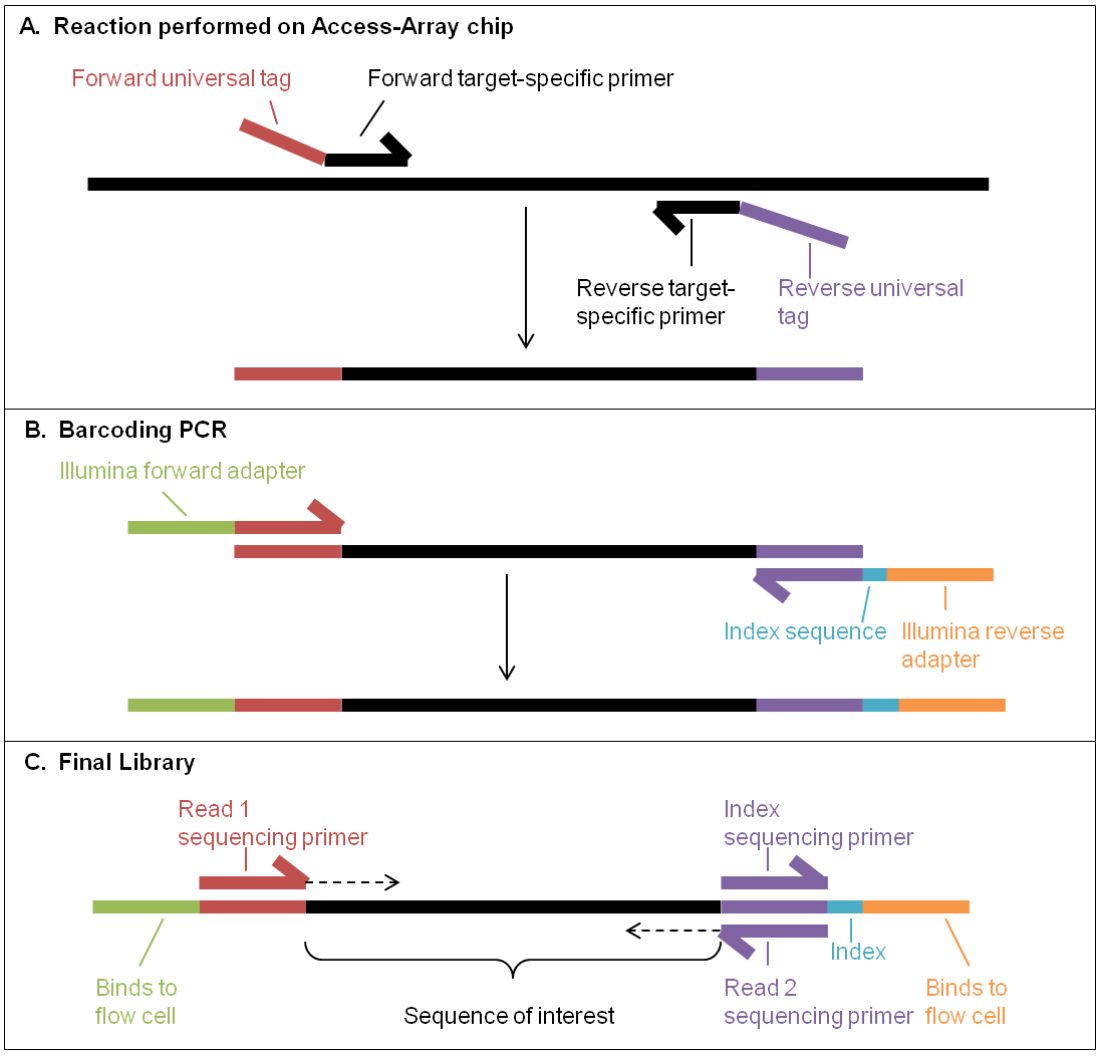


Figure 55. Schematic of library generation by Fluidigm Access Array (Fluidigm, San Francisco, USA). Adapted from (Halbritter et al., 2012).

4.1.3 Copy-number analysis and NGS

Characterising larger genetic structural lesions is important in understanding the link between mutation genotype and overall phenotype. It has long been established that aberrant chromosomal copy number is a characteristic of chromosomal instable cancers (CIN) (Rajagopalan et al., 2003, Pino and Chung, 2010). Many studies have reported chromosomal gains or losses in CRC and some have linked these to clinical outcome. (Poulogiannis et al., 2010, Xie et al., 2012, Janssen and Medema, 2013, Burrell et al., 2013).

CIN is a common feature of carcinoma as the genomic damage accumulates over time. Copy number aberrations are thought to be a late feature of colorectal progression. Commonly seen aberrations seen in adenoma include gains of chromosome 7, 13, 19 and 21 and losses in 1p, 9q, 17, 19 and 22 (Jones et al., 2007, Flora et al., 2012, Yamamoto et al., 2012)

Techniques for detecting copy-number variation (CNV) vary from large genome-wide assays such as array comparative genome hybridization (aCGH) and NGS to more targeted approaches such as fluorescent in-situ hybridization (FISH), multiplex ligation-dependant probe amplification (MLPA) and other molecular techniques (Aradhya et al., 2013). NGS is an advantageous approach for CNV due to its broad range across the genome combined with resolution down to single nucleotides. This is achievable for small quantities of low quality DNA and resolution can be adjusted to allow for greater multiplexing of samples (Hayes et al., 2013). Algorithms for determining CNV from NGS data vary depending on their intended application. They work in principle by comparing the number of reads that once aligned, map throughout the genome in comparison to a normal control. CNAnorm is an ideal analysis tool for determining CNV in tumours due to it adjusting for possible contamination of normal cells within the tumour (Gusnanto et al., 2012).

4.2 Chapter aims

The aims of this chapter are as follows:

- To develop a targeted panel of CRC genes to be used with the Fluidigm Access Array (Fluidigm, San Francisco, USA).
- To assess the performance of the Fluidigm Access Array (Fluidigm, San Francisco, USA).
- To investigate mutations in carcinoma, adenoma and their associated “normal” mucosa alongside non-neoplastic mucosa using the Fluidigm Access Array (Fluidigm, San Francisco, USA).
- To investigate mutations in a cohort of fresh carcinomas and their matched normals as well as a cohort of FFPE carcinomas and their matched normal.
- To investigate mutations in multiple adenomas from FAP patients.
- To investigate copy-number variation in multiple adenomas from FAP patients.

4.3 Methods

4.3.1 Samples

For 32 cases of CRC FFPE blocks were selected for the tumour and corresponding normal mucosa from the resection margin. Similarly 32 cases of sporadic adenoma and their matched normal mucosa were selected. For adenomas, biopsy blocks were used and the normal mucosa selected was at least 5cm from the tumour as recorded in the pathology report. Also 32 biopsies of non-dysplastic mucosa from patients with normal colonoscopies were selected. All blocks were sectioned and DNA extracted as previously outlined in Chapter 2.

Blocks from four FAP patients were also selected for DNA extraction. Patient 1 = 22 adenomas, patient 2 = 40 adenomas, patient 3 = 7 adenomas and patient 4 = 7 adenomas. For each patient, a sample of muscle was also selected to act as a normal control. NGS libraries were generated from these FAP samples alongside the carcinoma, adenoma and normal samples using the Fluidigm Access Array (Fluidigm, San Francisco, USA). This produced 240 samples in total to be sequenced on a single run on the Illumina Hiseq (Illumina, San Diego, USA).

Blocks of 6 carcinomas and their matched normal were repeated for mutational detection with Fluidigm alongside 6 samples of fresh frozen carcinoma and normal. The libraries were created in duplicate in order to determine the amount of error from formalin fixation and PCR to improve confidence in mutation calling.

Alongside mutational analysis, the 80 FAP samples also underwent copy number variation analysis using NGS. All 80 samples were sequenced on a single Hiseq run.

4.3.2 Fluidigm system of high-throughput target enrichment

A panel of key colorectal gene targets was designed by collating mutation frequencies from the literature (The Cancer Genome Atlas Network, 2012) and the Sanger database COSMIC v.66 (Bamford et al., 2004). When designing the panel, the number of targets had to be balanced with the target size and number of samples intended to assess in order to have high enough coverage of each target whilst maintaining sensitivity. Table 43 shows the targets chosen to create a design of 157 amplicons. Amplicons were designed to be 188bp or less in order for the whole target to be covered with a paired end read and a 6bp indexing read. The Illumina Hiseq (illumina, San Diego, USA) has a minimum output of 400 million reads in a paired-end run and 80% of these are of high quality. Therefore for 157 amplicons, 240 samples could be multiplexed to produce a coverage of 8,500 for each target.

Gene	Mutation Frequency	Target size (bp)	Number of amplicons
APC	81%	8532	106
TP53	60%	413	6
KRAS	43%	12	3
PIK3CA	18%	18	2
FBXW7	11%	9	3
SMAD4	10%	1659	28
NRAS	9%	9	2
TCF7L2	9%	48	2
CTNNB1	5%	42	1
BRAF	8%	3	1
PTEN	4%	9	3
	Total	10754	157

Table 43. Targets chosen for Fluidigm assay

Libraries were prepared in accordance with the standard Fluidigm protocol for the use of multiplex amplicon tagging with the access array. The pre-amplification step suggested for FFPE tissue was omitted due to high yields of PCR product generated without this step. The purification of libraries protocol was modified to normalise the quantities of each library added to the final pool due to uneven amplification during PCR from differing template qualities. Mutations present in KRAS 12 &13 were validated with pyrosequencing as outlined Chapter 2.

4.3.3 Copy-number library preparation

In order to determine the chromosomal copy-number of the 80 FAP adenoma lesions, each sample of gDNA was fragmented, adapters ligated, libraries amplified and then pooled together and sequenced on a single lane of the Illumina Hiseq (illumina, San Diego, USA). In contrast to the previous library preparation approach as outlined in chapter 3 section 3.4.1, the index for each sample was read in a separate indexing read rather than incorporating the barcode into the first 6 bases of the read. This allowed for the NEBNext adaptor for illumina to be used for every sample. This adaptor contains a loop due to a uracil residue. After initial ligation of the adaptor, the loop is removed by adding Uracil-Specific Excision Reagent (USER) enzyme. The 6bp is then added in the final PCR enrichment step by using a different index primer for every sample.

4.3.4 Shearing of gDNA

gDNA was quantified by Quant-iT PicoGreen assay (Invitrogen, Paisley, UK) and 200ng of each sample was diluted with TE buffer to a volume of 250µl and transferred to glass shearing tubes. Shearing was performed with the Covaris S2 system (Covaris, Massachusetts, USA) for 25 cycles per sample to produce fragmented DNA of an average size of 200bp. Samples were then cleaned up and concentrated with the Qiagen MinElute kit (QIAGEN, Crawley, UK) eluting in a final volume of 11µl. 1µl of each sample was run on an Agilent Bioanalyser DNA 1000 LabChip (Agilent, Santa Clara, USA) to ensure that DNA had been successfully sheared to the required size fragments.

4.3.5 End-repair

An End-repair step was performed to produce blunt-ended DNA fragments from the uneven-ended fragments produced from the shearing step. This End-repair was performed using the NEB End-repair kit (New England Biolabs, Hitchin, UK) as outlined in Table 44.

Component	Volume
NEBNext End Repair Reaction Buffer (10x)	5µl
NEBNext End Repair Enzyme Mix	2.5µl
dH2O	33.5µl
gDNA	9µl
Total volume:	50µl

Table 44. Reaction mix for CNV end-repair

The reaction mix was incubated at room temperature (20-25°C) for 30 mins. The reaction was then cleaned up using Qiagen MinElute kit (QIAGEN, Crawley, UK) clean-up eluting in 21µl EB.

4.3.6 A-addition

After the ends of the DNA fragments had been blunted, an adenosine-addition step was performed to create an “A” overhang. This would allow for ligation of adapters containing a “T” overhang. The reaction mix is outlined in Table 45.

Component	Volume
NEBNext dA-Tailing Reaction Buffer (10x)	2.5µl
Klenow Fragment (3´→ 5´ exo–)	1.5µl
End Repaired, Blunt DNA	21µl
Total volume:	25µl

Table 45. Reaction mix for CNV A-addition

The reaction mix was incubated at 37°C for 30 mins. The reaction was then cleaned up using Qiagen MinElute kit (QIAGEN, Crawley, UK) eluting in 12.5µl EB.

4.3.7 Adaptor ligation

Adaptor ligation was performed using the NEBNext Adaptor (New England Biolabs, Hitchin, UK) with a T4 ligase enzyme. The reaction is outlined below in Table 46.

Component	Volume
Quick Ligation Reaction Buffer (5x)	5µl
NEBNext Adaptor	2.5µl
Quick T4 ligase	2.5µl
dH_2O	2.5µl
DNA	12.5µl
Total volume:	25µl

Table 46. Reaction mix for CNV adaptor ligation

The reaction was incubated at 20°C for 15 mins. 3µl of USER enzyme was added to each sample and then incubated for a further 15 mins at 37°C to remove the uracil residue loop from the adaptors. The reaction was then cleaned up using Qiagen PCR purification kit (QIAGEN, Crawley, UK) eluting in 50µl EB.

4.3.8 Size-selection

The resulting DNA fragments have a wide spread in size around the 200bp average. Therefore a size-selection step was performed using Agencourt AMPure XP magnetic beads (Beckman Coulter, High Wycombe, UK). This works as a two-step process, to firstly remove unwanted large fragments and then smaller fragments. The polyethylene glycol (PEG) within the bead buffer is the key agent to precipitate DNA and bind it to the magnetic beads. It is the ratio of this buffer to the DNA volume that affects the size of DNA fragments captured.

Firstly 40µl of beads was added to each 50µl sample, resulting in a 0.8 bead buffer to DNA ratio and incubated at room temperature for 5 minutes. This results in the capture of larger fragments onto the beads (300bp and higher) and smaller fragments remain in the supernatant. The samples were then moved to a magnetic rack and the beads were allowed to separate from the supernatant. The supernatant containing smaller fragments was transferred to a new tube whilst the larger fragments on the beads were discarded. 10µl of beads were then added to each sample. This volume, plus the 40µl already added, resulted in a 1:1 bead buffer to DNA ratio. This allowed for the capture of fragments 200bp in size and larger onto

the magnetic beads whilst small fragments remained in the supernatant. After separation on the magnetic rack, the supernatant was removed and this time discarded. The fragments left on the beads underwent two washes with 80% ethanol before they were finally eluted into 20µl of EB.

4.3.9 Enrichment PCR

The final enrichment step enabled samples to be barcoded with an individual 6bp index. This is done by performing a PCR with a different indexing forward primer for each sample. The reaction mix can be seen in Table 47.

Component	Volume
NEB High Fidelity PCR master mix (2x)	12.5µl
Universal PCR primer	1.25µl
Indexing PCR primer	1.25µl
DNA	10µl
Total volume:	25µl

Table 47. Reaction mix for CNV Indexing PCR enrichment

The PCR reaction (Table 47) underwent the following thermocycling programme:

- 98°C for 30 sec

- 15 cycles of:
- 98°C for 10sec
- 65°C for 30sec
- 72°C for 30sec

- 72°C for 5 mins

These final libraries were then cleaned up using Agencourt AMPure XP magnetic beads (Beckman Coulter, High Wycombe, UK) following the standard protocol with a 1x bead ratio and eluting into 30µl of EB. The libraries were visualised using an Agilent Bioanalyser DNA 1000 LabChip (Agilent, Santa Clara, USA). Figure 56 shows an example trace from the Agilent Bioanalyser where there is a clear peak at the expected size range with no adaptor-dimers.

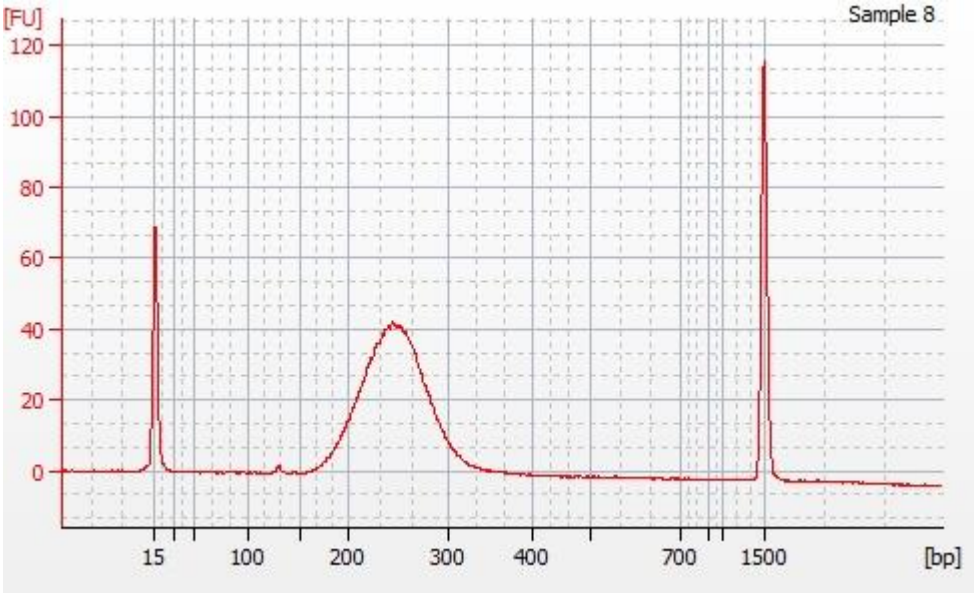


Figure 56. Bioanalyser trace of final gDNA CNV library.

4.4 Bioinformatic pipelines

4.4.1 Bioinformatic pipeline for mutation calling

For analysis of data generated from Fluidigm libraries, an automated pipeline was set up using published software. For file formats please refer to Chapter 3, Table 27. Reads were firstly aligned using BWA alignment software, with default settings allowing for a 4% mismatch rate within the read. This generated SAM files which were then converted to sorted and indexed BAM files using SAMtools with default settings alongside the –q60 option to remove reads with a poor mapping quality (Li et al., 2009). Reads then underwent local realignment and base quality recalibration with the GATK (Li and Durbin, 2009). Local realignment looks for indels within the reads and realigns around them to allow for more accurate variant calls downstream. The base quality recalibration step looks for known SNPs within the reads and recalibrates the Phred scores to allow for SNPs to be discovered and more reliable variant calls.

The output from GATK was then fed back into SAMtools to generate mpileup files. When using Samtools mpileup, the –d1000000 option was applied (maximum read depth) to allow the programme to work with high coverage. Also the –C50 option was used to reduce the effects of reads containing excessive mismatches. The final step in the pipeline to call variants was performed using VarScan version2 (Koboldt et al., 2012). Where matched tumour and normal samples were available, the VarScan –somatic command was used and VarScan –mpileup2snp for other cases. The settings were adjusted to call for a minimum allele frequency of 5% and a minimum coverage of 100 in Fluidigm libraries.

4.4.2 Clustering analysis of mutation profiles

In order to call mutations in adenomas from patients with FAP, the base changes in each tumour sample was compared to a normal muscle control. This was done using VarScan version2 with the “somatic” algorithm. In this way, SNPs and any PCR artefacts that were shared between adenomas and the control were filtered out from the analysis. This means that the germline APC mutation could not be identified. The remaining sites that were mutated in more than one adenoma were then compared for each sample to compile a matrix using a custom perl script mut_matrix.pl (see appendix). Heatmaps to display all shared mutations across the

samples were created with TM4 software (Saeed et al., 2003). Dendrograms were drawn to display clustering of adenomas that shared mutations by hierarchal clustering.

4.4.3 Bioinformatic pipeline for copy number variation

Samples sequenced for CNV analysis underwent an indexing read as part of the sequencing process and therefore the raw output of the sequencer were compressed fastq files for each individual sample. Therefore a demultiplexing step was not required. Sequencing was performed with the illumina Hiseq (illumina, San Diego, USA) on two lanes in “rapid mode” which resulted in single-end sequencing of 50bp read length. This amount of sequence had been previously validated by Dr. Henry Wood to be shown as sufficient for determining copy number (unpublished). All samples were decompressed and aligned to the genome using BWA “aln” and “samse” options with default settings. The output SAM files were then converted to BAM files using default settings with samtools. Due to sequencing being performed on two lanes of the Hiseq, two files were produced for each sample and therefore these BAM files were merged for each sample with the samtools “merge” algorithm.

In order to determine copy number ratios, the CNAnorm pipeline was used. Firstly, BAM files were converted to tab delimited files, to provide compatible input for the CNAnorm programme. This was done using the bam2windows.pl perl script with default settings. These .tab files were then inputted to the CNAnorm programme which compared each sample to a normal control and produces plots of copy number ratio along the genome every 2Mb. Plots were generated for individual chromosomes, for example Figure 57 as well as a general overview of the whole genome for each sample as seen in Figure 58.

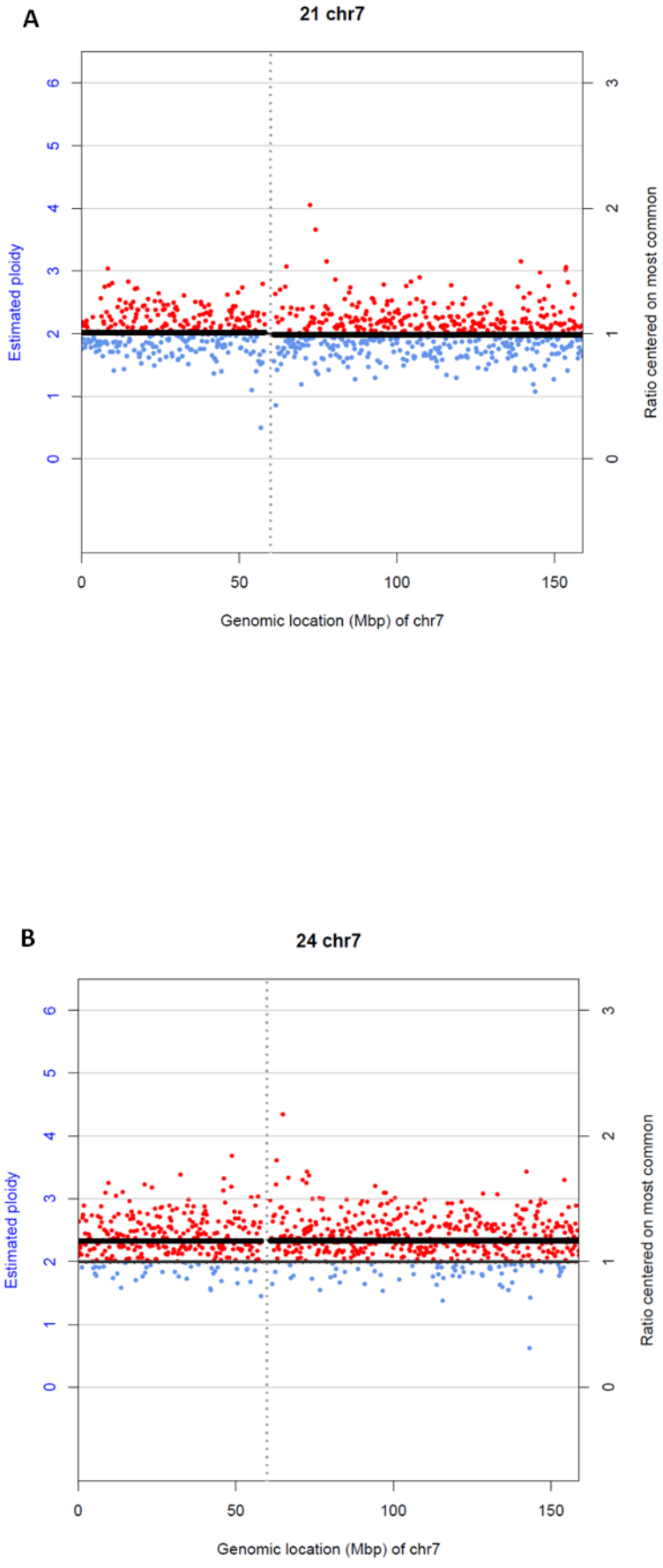


Figure 57. CNAnorm output CNV plots for chromosome 7 in two adenomas A: sample 21 and B: sample 24 from the same patient.

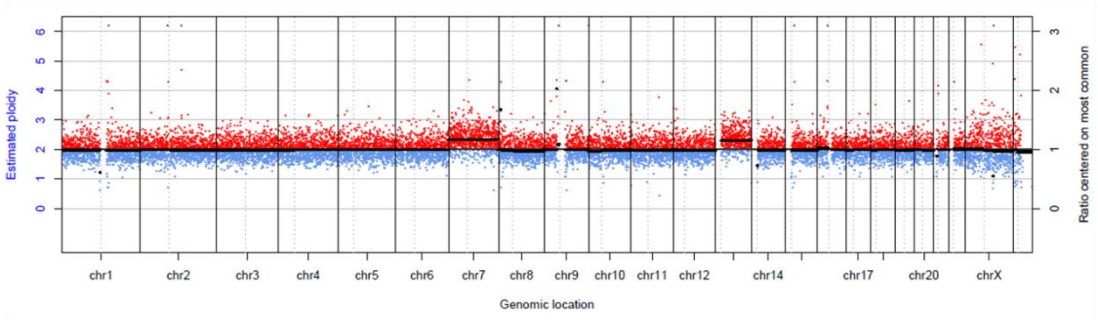
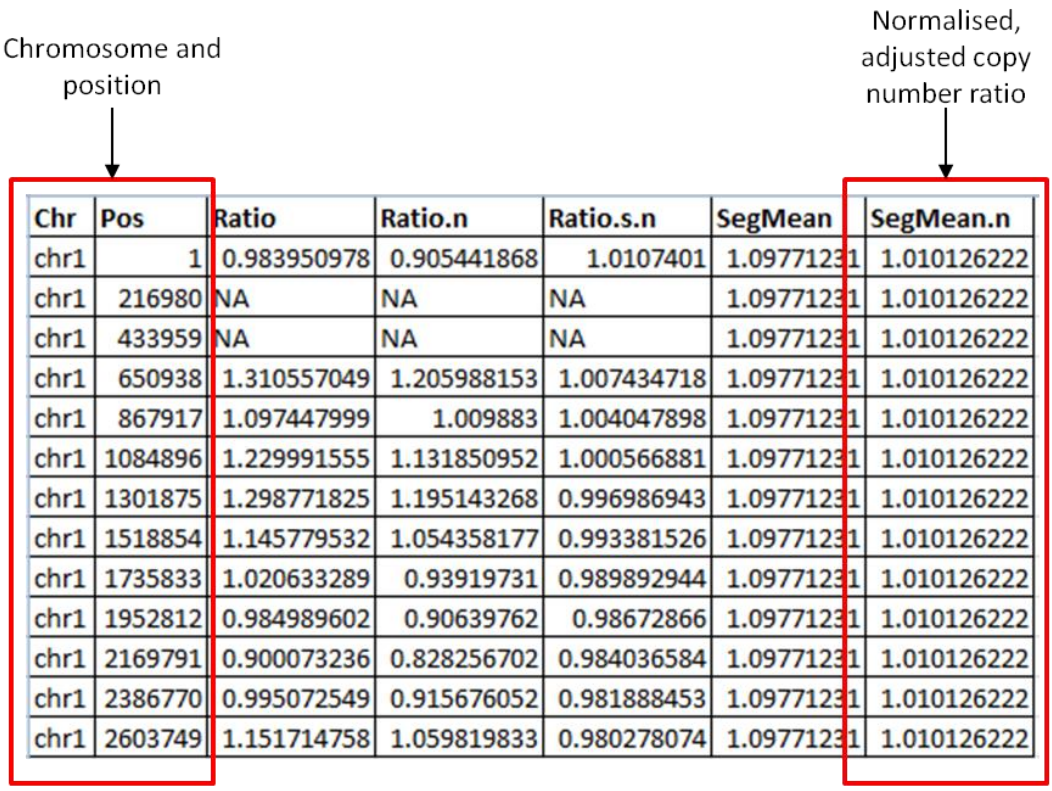


Figure 58. Copy number variation profile overview for the entire genome of a single adenoma showing increased copy number of chromosomes 7 and 13.

4.4.4 Comparison of copy number profiles

In order to compare the profiles of lesions from the same patient across the genome, a matrix approach was adopted. The raw output from the CNAnorm produced text files which showed the position in the genome at around 200Mb intervals and the adjusted, normalised copy number ratio (Figure 59). Two custom perl scripts, matrix1.pl and matrix2.pl were used to code these profiles (see appendix for script). For each 200Mb window, a copy number above 1.2 was coded as a “1” whilst a copy number below 0.8 was coded as “2”. A normal copy number was coded as “0”. This enabled shared aberrations to have the same entry within the matrix for those adenomas that had those shared aberrations, whilst unique aberrations were coded differently in the adenoma matrix. In this way, small aberrations could be compared between adenomas simultaneously to larger chromosomal gains and losses.



Chr	Pos	Ratio	Ratio.n	Ratio.s.n	SegMean	SegMean.n
chr1	1	0.983950978	0.905441868	1.0107401	1.09771231	1.010126222
chr1	216980	NA	NA	NA	1.09771231	1.010126222
chr1	433959	NA	NA	NA	1.09771231	1.010126222
chr1	650938	1.310557049	1.205988153	1.007434718	1.09771231	1.010126222
chr1	867917	1.097447999	1.009883	1.004047898	1.09771231	1.010126222
chr1	1084896	1.229991555	1.131850952	1.000566881	1.09771231	1.010126222
chr1	1301875	1.298771825	1.195143268	0.996986943	1.09771231	1.010126222
chr1	1518854	1.145779532	1.054358177	0.993381526	1.09771231	1.010126222
chr1	1735833	1.020633289	0.93919731	0.989892944	1.09771231	1.010126222
chr1	1952812	0.984989602	0.90639762	0.98672866	1.09771231	1.010126222
chr1	2169791	0.900073236	0.828256702	0.984036584	1.09771231	1.010126222
chr1	2386770	0.995072549	0.915676052	0.981888453	1.09771231	1.010126222
chr1	2603749	1.151714758	1.059819833	0.980278074	1.09771231	1.010126222

Figure 59. Text output from CNAnorm programme showing copy number ratio across the genome.

Due to nucleotide repeats at the telomeres and centromere, reads that map to these regions may not be aligned accurately. Therefore these positions within the genome are not as reliable for determining copy number. Ratios from the telomeres and centromere were removed from the analysis to improve the accuracy of the data. Once matrices had been generated for multiple samples from one patient they were inputted into

TM4 software (Saeed et al., 2003) was used to create heatmaps of regions of gains and losses as well as cluster the adenomas using hierarchical clustering to produce dendrograms. These dendrograms allowed for the adenomas to be examined for how related they were to each other according to their chromosomal copy aberrations.

4.5 Results

4.5.1 Mutations in carcinoma and carcinoma-associated normal duplicates

12 samples of carcinoma and their associated normal mucosa from the resection margin were sequenced in duplicate with the Fluidigm Access Array (Fluidigm, San Francisco, USA) library generation system. The P53 gene did not amplify evenly and did not obtain enough coverage to be included in the analysis. The percentage of samples containing mutations in the tested genes are shown in Figure 60 in comparison to the COSMIC database (Bamford et al., 2004).

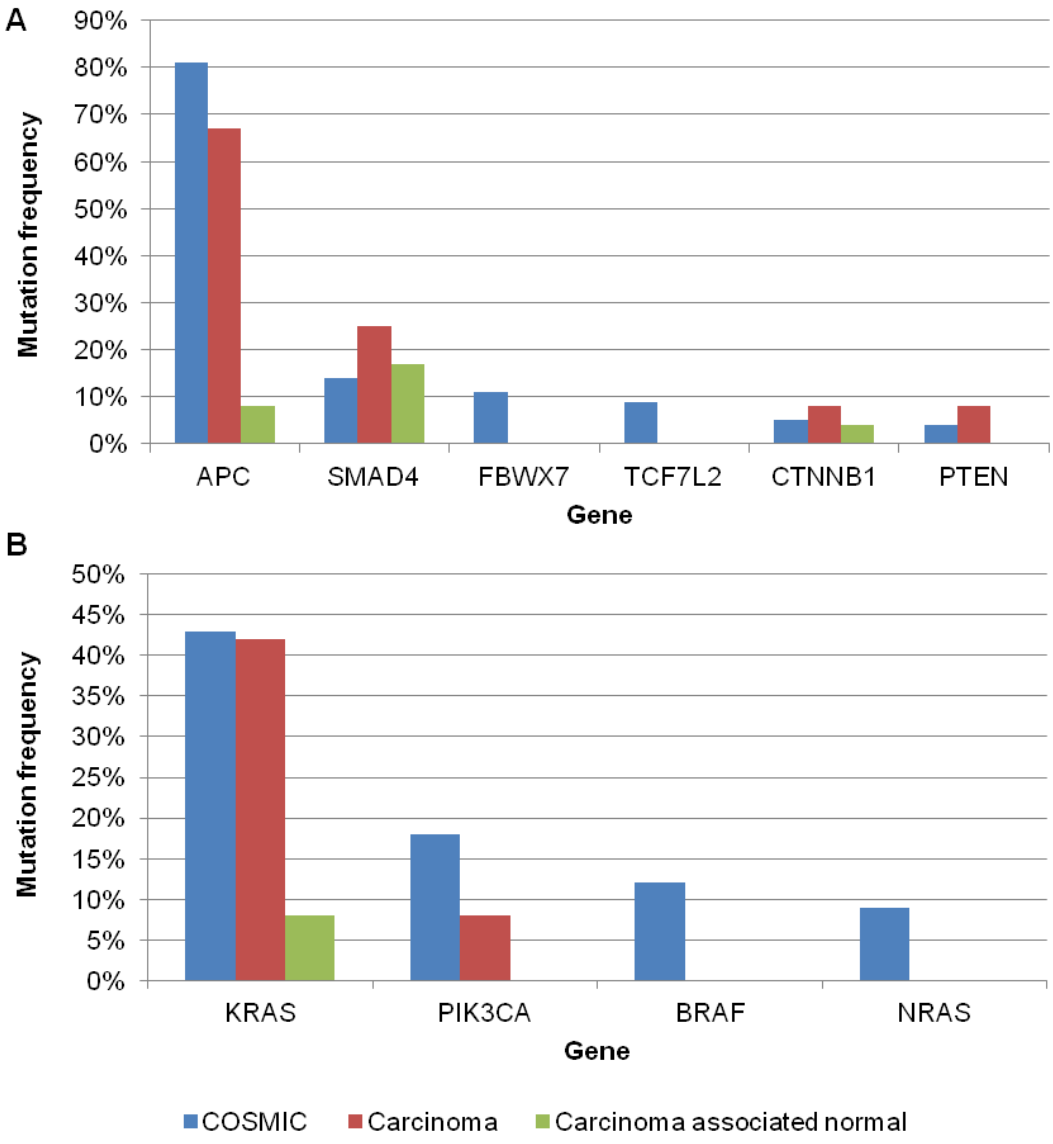


Figure 60. Mutation frequencies for carcinoma and carcinoma-associated normal compared to the cosmic reference database for colorectal cancer (Bamford et al., 2004) with a 5% detection threshold. A: whole genes tested. B: genes where hotspots were tested.

8 of the 12 carcinomas contained mutations in APC with an allele frequency range of 11%-35%. There were mutations in KRAS in 5 of the 12 carcinomas; all were located in codons 12&13 with an allele frequency range of 32%-48%. One sample contained a KRAS codon 12 mutation in the normal mucosa with an allele frequency of 9%. Only one carcinoma contained a 30% PIK3CA mutation at codon 545 and this carcinoma also had a KRAS mutation. There were SMAD4 mutations in 3 of the 12 carcinomas with a mutant allele frequency range of 35-59%. Two of the carcinoma associated normal samples contained a SMAD4 mutation (7.06% and 19%), and these cases did not contain SMAD4 mutations in their corresponding tumour. Two of the carcinomas had CTNNB1 mutations (15% and 26%) and one of these had a mutation at a different location in CTNNB1 in the normal mucosa.

The analysis was repeated for a minimum allele frequency of 1%, the distribution of frequencies in the tested genes is shown in Figure 61. By lowering the minimum cut-off the number of mutations called increased by 1.7-fold. There were more mutations seen in the carcinoma-associated normal mucosa, however, there were no new genes that were implicated. Figure 62 and Figure 63 display the distribution of mutations found in APC for carcinoma and carcinoma normal at a 5% and 1% minimum allele frequency.

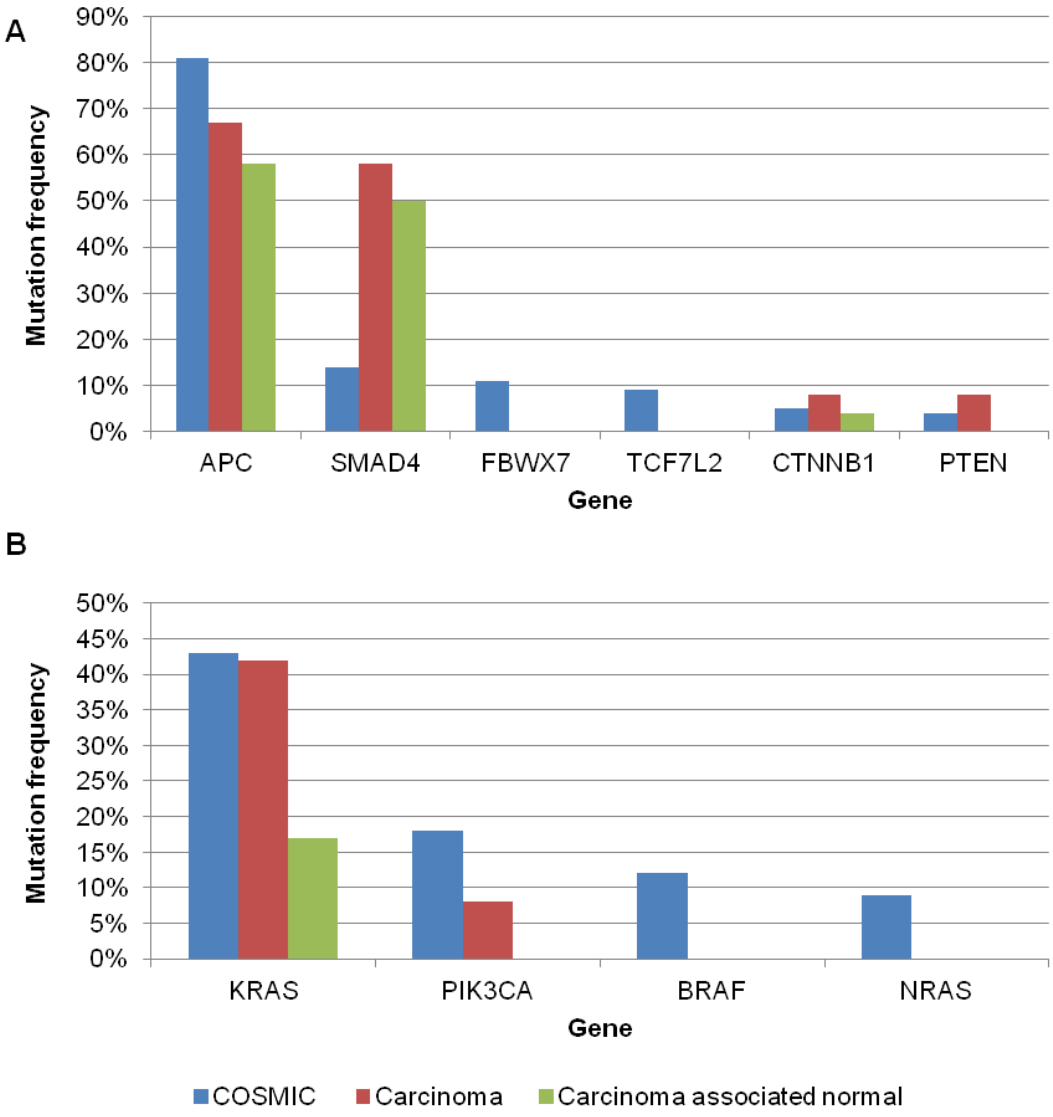


Figure 61. Mutation frequencies for carcinoma and carcinoma-associated normal compared to the cosmic reference database for colorectal cancer (Bamford et al., 2004) with a 1% detection threshold. A: whole genes tested. B: genes where hotspots were tested.

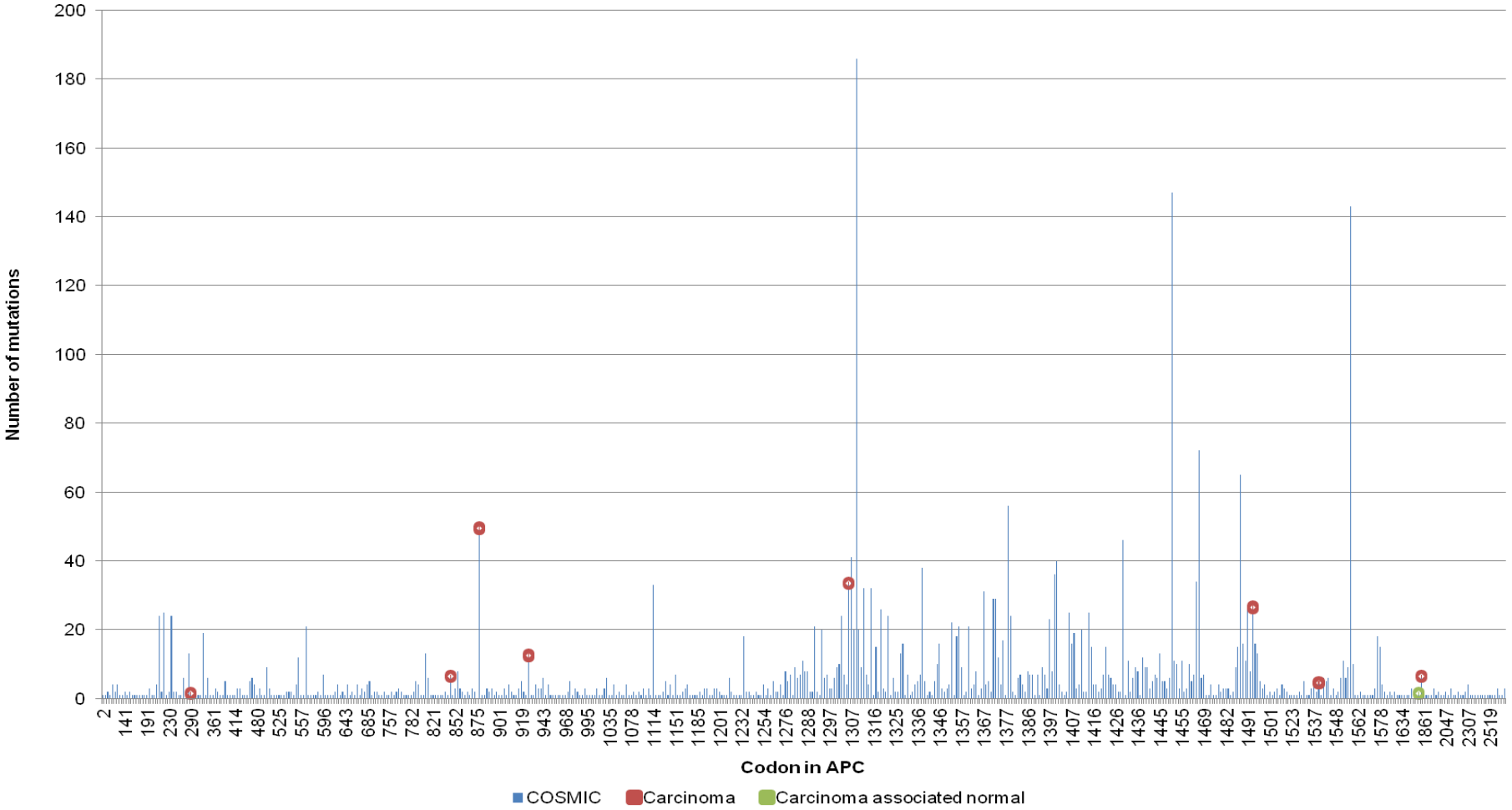


Figure 62. Distribution of mutations found in carcinoma and carcinoma associated normal in APC at a 5% minimum allele frequency threshold compared to the mutation distribution from the COSMIC database (Bamford et al., 2004). The red and green points show the locations of APC mutations in individual carcinoma and normal samples compared to the frequency of mutations at that position according to COSMIC (blue bars).

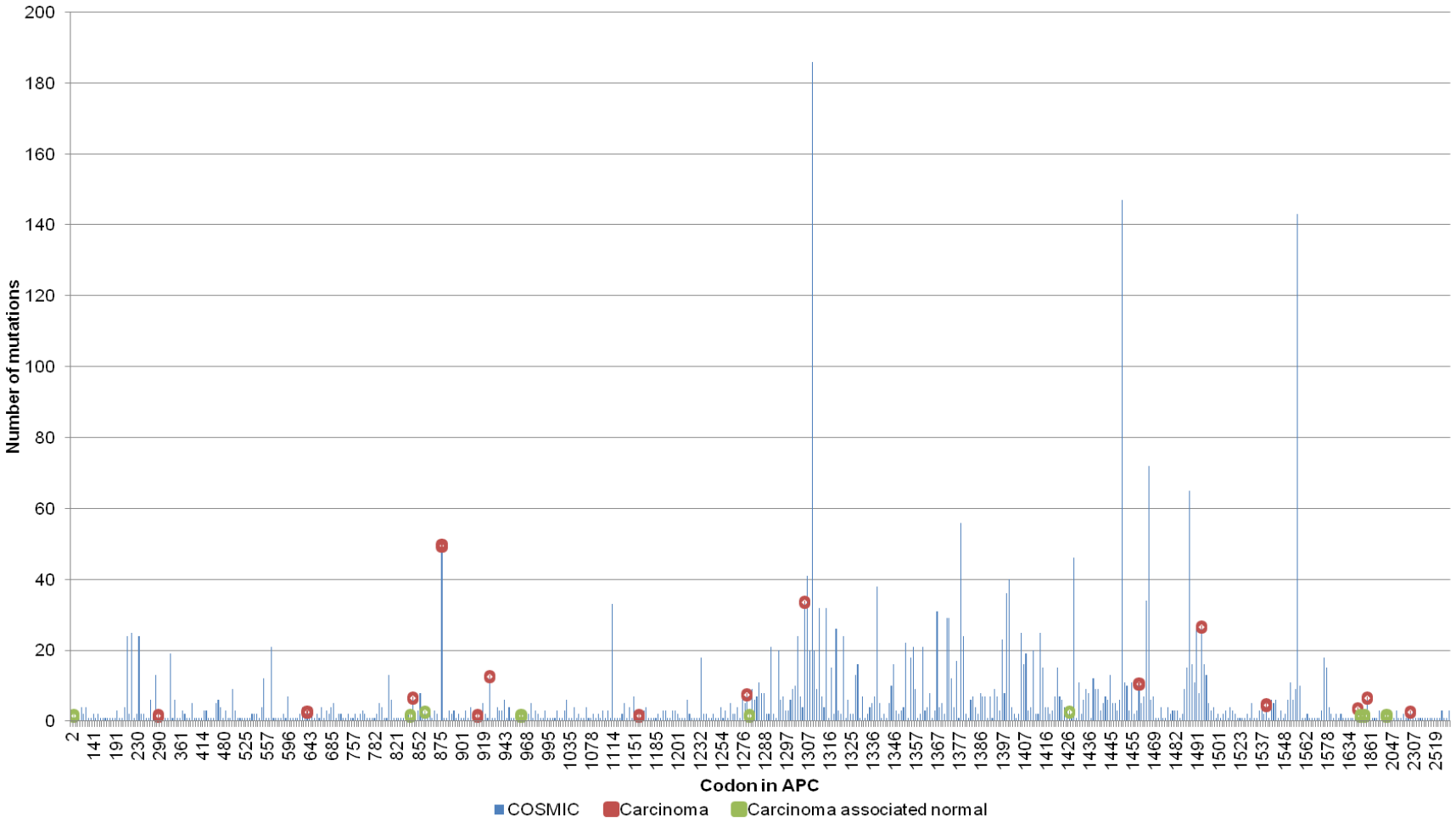


Figure 63. Distribution of mutations found in carcinoma and carcinoma associated normal in APC at a 1% minimum allele frequency threshold compared to the mutation distribution from the COMIC database (Bamford et al., 2004). The red and green points show the locations of APC mutations in individual carcinoma and normal samples compared to the frequency of mutations at that position according to COSMIC (blue bars).

4.5.2 Validation of KRAS codon 12&13 mutations

The 5 carcinomas with KRAS 12 &13 mutations alongside a WT carcinoma were validated by pyrosequencing. The results are as follows in Table 48. The mutant allele frequency of lesions 3, 4 and 5 fell below the sensitivity of pyrosequencing and therefore could not be detected.

Sample	NGS		Pyrosequencing	
1	c.38 G>A	21.36%	c.38G>A	32.7%
2	c.35 G>T	28%	c.35G>T	29.7%
3	c.38 G>A	9.2%	WT	-
4	c.35 G>T	6.8%	WT	-
5	c.35 G>A	5%	WT	-
6	WT	0%	WT	-

Table 48. Comparison of sequencing of KRAS codon 12 and 13 mutated tumours with NGS and pyrosequencing.

4.5.3 Agreement between duplicates

For 6 fresh tumours and their matched normals, when libraries were created and sequenced in duplicate using the Fluidigm Access-Array, a total of 195 mutations were called in carcinoma associated normal and 273 in carcinoma (Figure 64). These mutations were called with a minimum coverage of 300 and 5% minimum allele frequency threshold. Of these mutations, 80 were detected in both duplicates in carcinoma associated normal and 95 in carcinoma, resulting in an overall agreement rate between the duplicates of 81%.

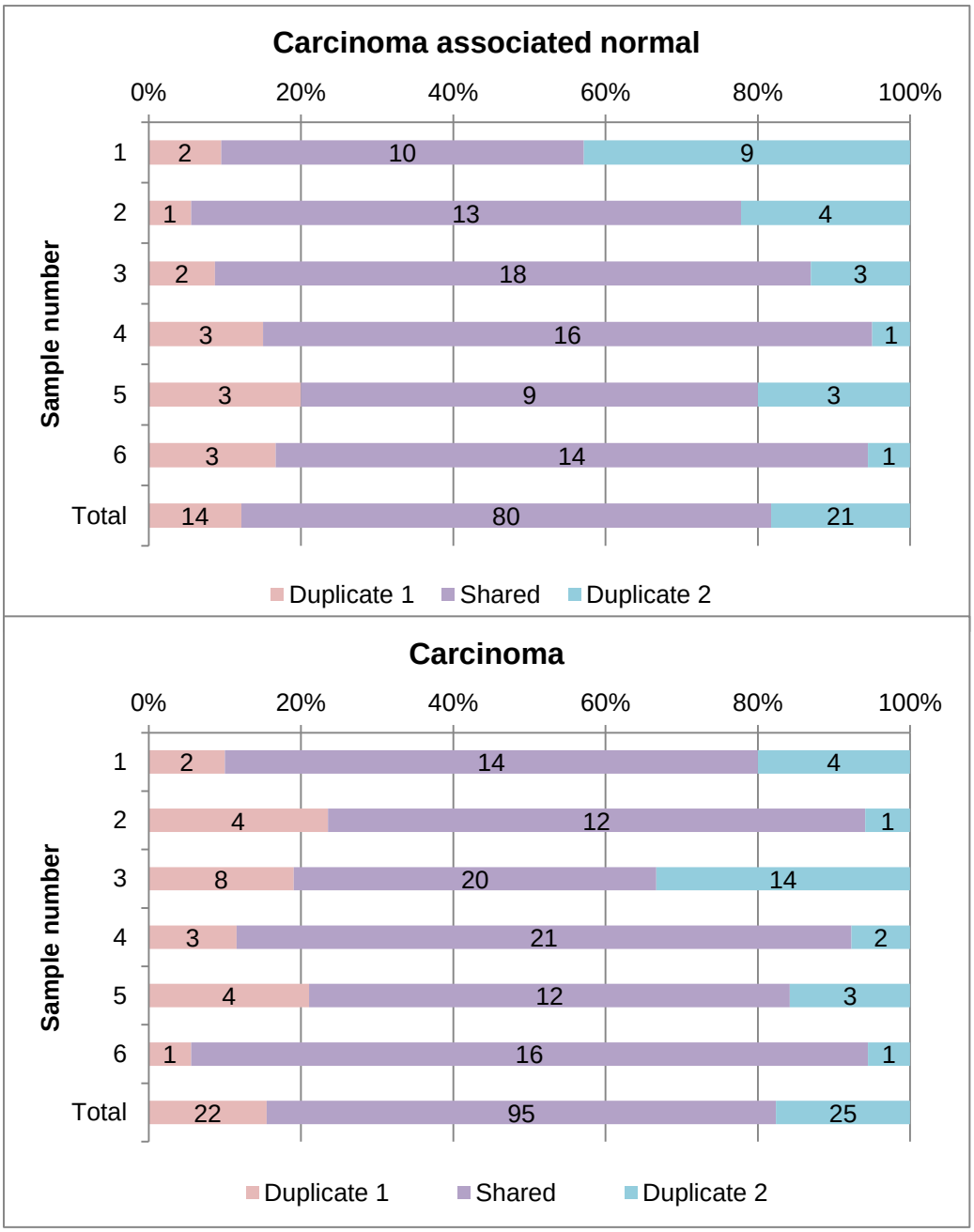


Figure 64. Proportion of mutations shared between duplicates 1 and 2 for fresh tumours and matched normals.

For 6 FFPE tumours and matched normals a total of 888 mutations were called in carcinoma associated normal and 910 in carcinoma. Only 39 of these were detected in both library duplicates for carcinoma associated normal and 58 in carcinoma (Figure 65), equating to an overall agreement rate between the duplicates of 11%.

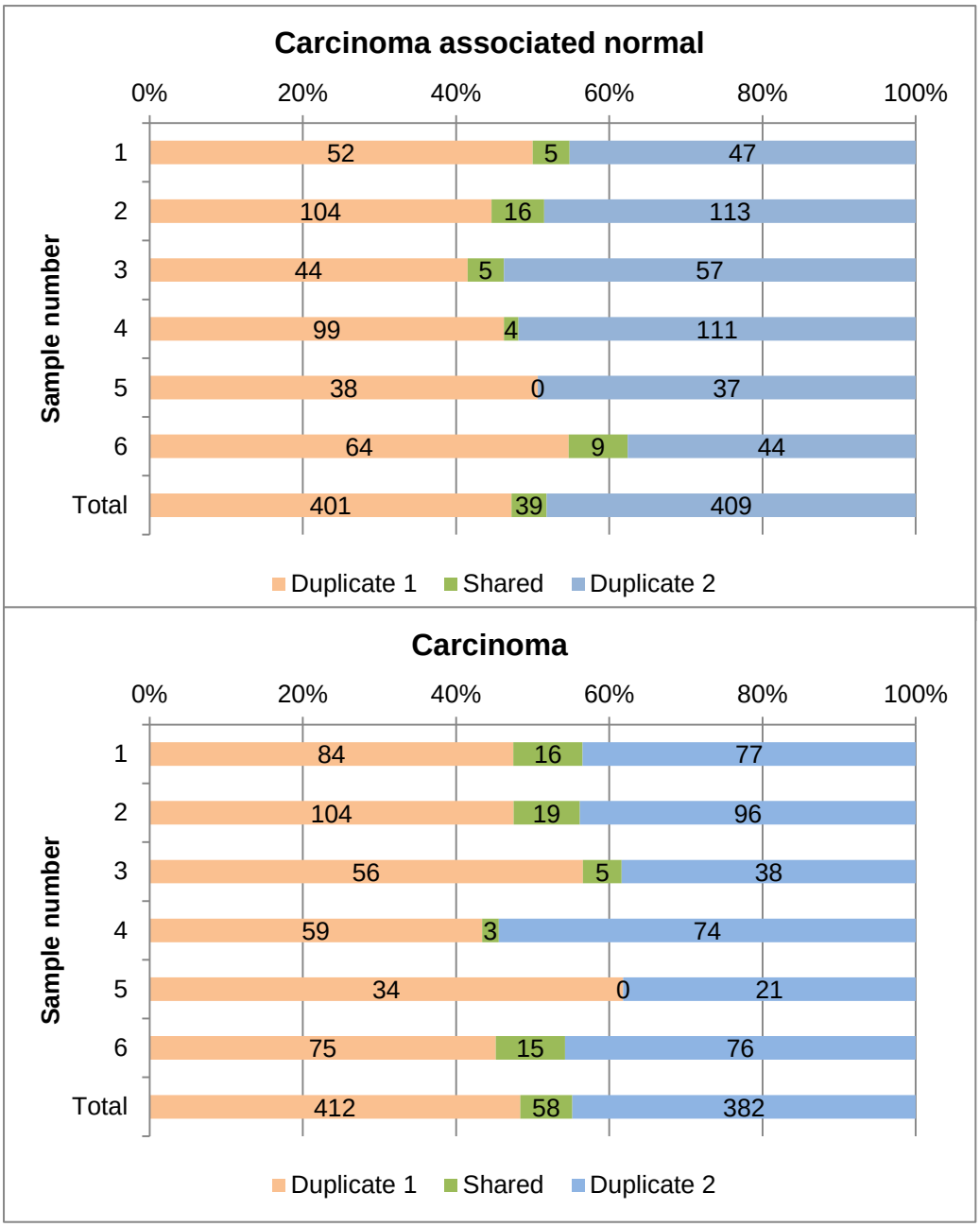


Figure 65. Proportion of mutations shared between duplicates 1 and 2 for FFPE tumours and matched normals.

This analysis was repeated for a range of minimum allele frequency thresholds as shown in Table 49. When plotted (Figure 66), a clear relationship could be seen for fresh tissue however there was no obvious trend for FFPE and the agreement rate between duplicates remained from 10% to 21%.

Allele frequency threshold	Agreement between duplicates	
	Fresh	FFPE
1%	33%	13%
2%	68%	11%
5%	85%	10%
10%	89%	13%
20%	97%	21%

Table 49. Percentage of reads that are random PCR error for fresh and FFPE tissue for a range of different minimum allele frequency cut-offs.

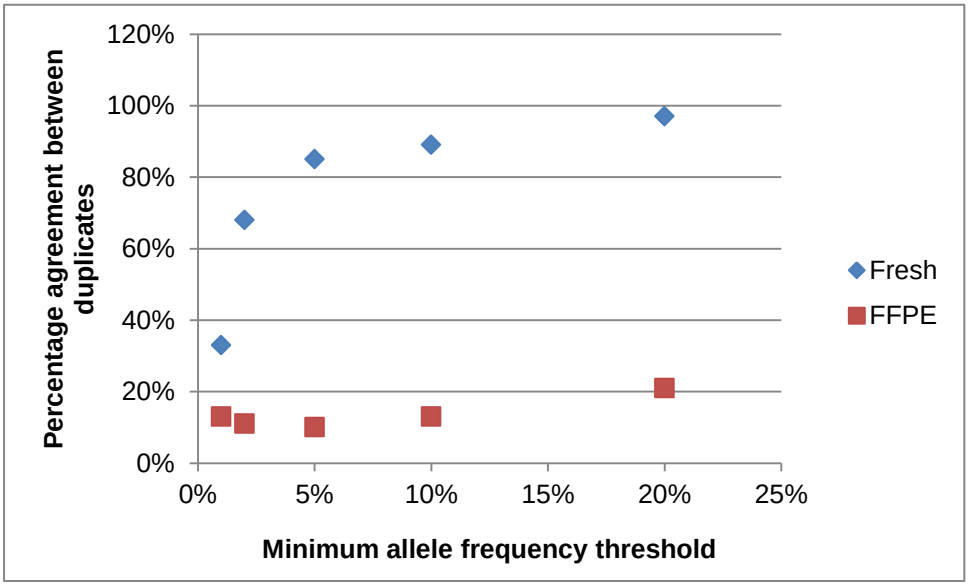


Figure 66. Plot of percentage agreement between library duplicates against minimum allele frequency threshold for a series of fresh and FFPE samples

For the cohort of fresh samples, the 195 mutations detected in 2 duplicates from 6 carcinoma associated normal samples equated to an average of 16 mutations called in each single duplicate per sample. Similarly, for carcinoma, the 273 mutations in total averaged out to 23 mutations for a single duplicate. The amount of sequence interrogated was 26,767 bases and so this equates to a mutation rate of 0.06% and 0.09% for carcinoma associated normal and carcinoma respectively in fresh tissue. This equates to between 0.1 and 0.2 bases per amplicon of an average size of 170bp.

In FFPE there was an average of 74 mutations called in carcinoma associated normal and 76 in carcinoma in each single duplicate. This gives a mutation rate of 0.3%. For an average amplicon of 170bp, this equates to 0.5 mutated bases in each amplicon. For the oncogene amplicons: KRAS, NRAS, BRAF and PIK3CA, this would equate to 4 mutated bases within the 8 amplicons. However, only 36bp of sequence are analysed across these 8 amplicons, as only mutations in the hotspots were of interest in this study. This would result in 0.1 mutated bases across the oncogenes (0.3% of 36bp). For a threshold of 5%, only 10% of mutations were detected in both duplicates (Table 49). Therefore, across the oncogenes, if a sample is tested in duplicate, the background error is 0.09 bases per 36bp (0.025%) which was considered low enough in order to be able to interrogate FFPE samples for the oncogenes where they were not tested in duplicate (section 4.5.4).

There was a significantly greater proportion of G>A and C>T changes in the FFPE cohort compared to fresh (Table 50). When mutations were called from duplicate libraries, 41% were C>T changes and 35% were G>A for FFPE. For mutations called in duplicate in the fresh cohort, 36% were C>T changes and 18% were G>A. Comparing the FFPE and fresh cohort mutations called in duplicate, the differences in the proportion of C>T and G>A changes was not significant.

Base change	Proportion of mutations		p-value
	Fresh	FFPE	
G>A	31%	46%	<0.0001
C>T	11%	36%	<0.0001
A>G	9%	8%	0.4973
T>C	10%	5%	<0.0001
G>T	6%	1%	<0.0001
T>G	4%	1%	<0.0001
C>A	12%	1%	<0.0001
T>A	6%	1%	<0.0001
A>C	5%	1%	<0.0001
G>C	0%	0%	
C>G	6%	0%	<0.0001
A>T	0%	0%	

Table 50. Proportion of base changes for fresh and FFPE cohorts. p values calculated from z-test statistic.

4.5.4 Mutations in carcinoma, adenoma and carcinoma-associated normal

For carcinoma, adenoma and their associated normals run in single-plex, mutations were detected in the hotspots oncogenes in carcinoma, adenoma and carcinoma-associated normal mucosa as shown in Figure 67 and Figure 68. No mutations were detected in adenoma-associated normal mucosa or in normal mucosa from patients with normal colonoscopies in the oncogenes. Mutation calls in the tumour suppressor genes could not be determined due to poor coverage and high false positive background noise from FFPE material as previously discussed. Also for the normal mucosa from patients with normal colonoscopies there was no other control tissue available such as blood in order to the determine patients' SNPs and therefore these could not be removed from the data making mutations undeterminable.

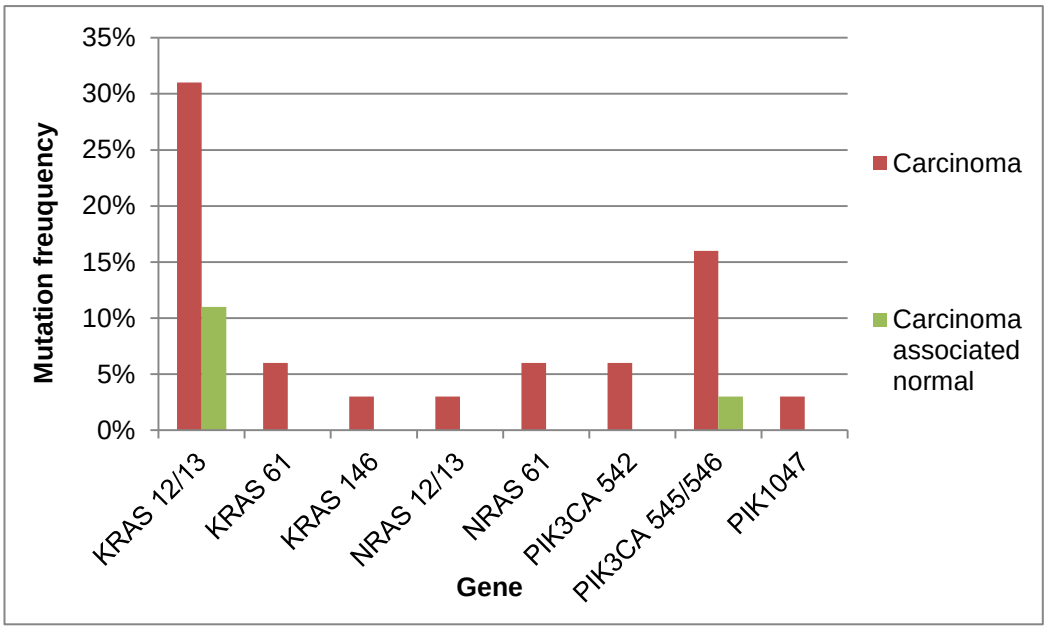


Figure 67. Mutations present in hotspots of oncogenes for 32 cases of carcinoma (red) and their associated normal mucosa (green).

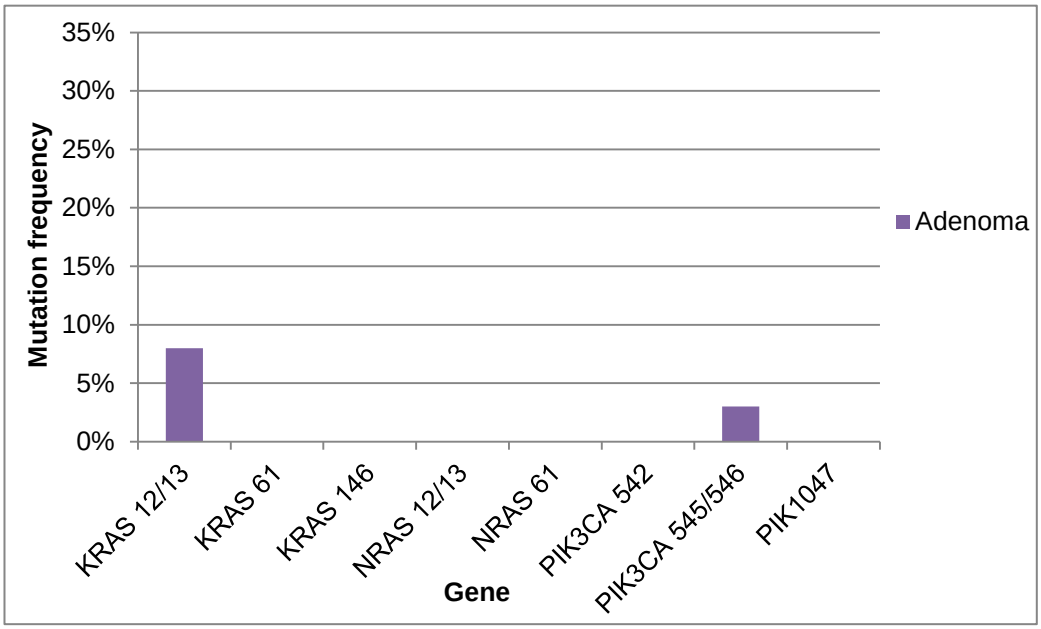


Figure 68. Mutations present in hotspots of oncogenes for 32 cases of adenoma. No mutations in these oncogenes were detected in the associated normal mucosa.

4.5.5 Mutations in FAP adenomas

Heat maps were generated to display the mutations from each adenoma. Only mutations present in more than one adenoma were included in the analysis in order to remove mutation calls that may be error incorporated by the PCR cycles. A black cell represents that the adenoma is WT at the position of interest and a red cell signifies a mutation. Those mutations of an allele frequency of 20% and higher have a stronger intensity colour. Hierarchical clustering was performed to analyse similarities between adenomas based on their mutational profiles.

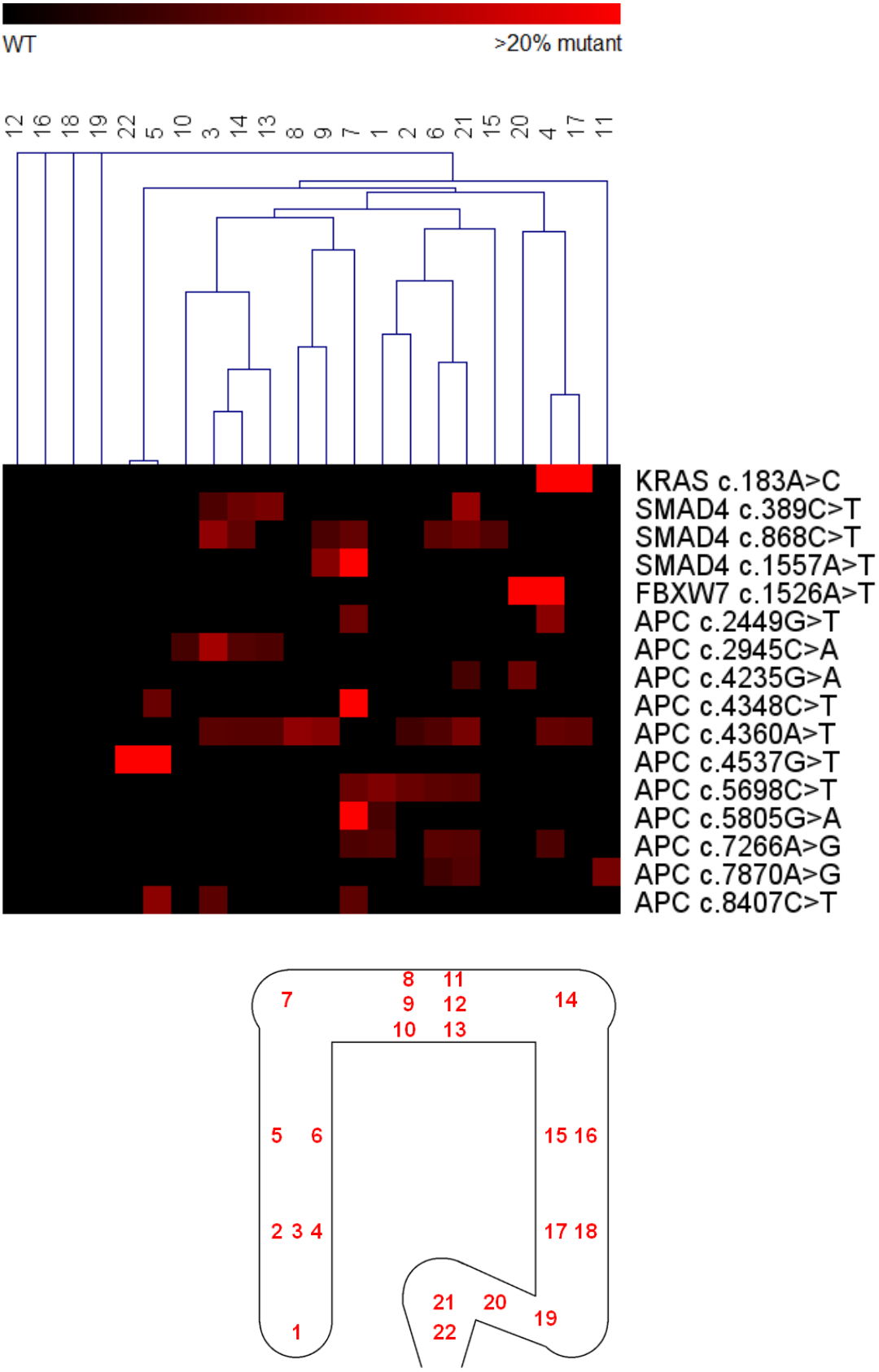


Figure 69. Heatmap with dendrogram for mutations called in 22 adenomas from patient 1 and their location within the bowel.

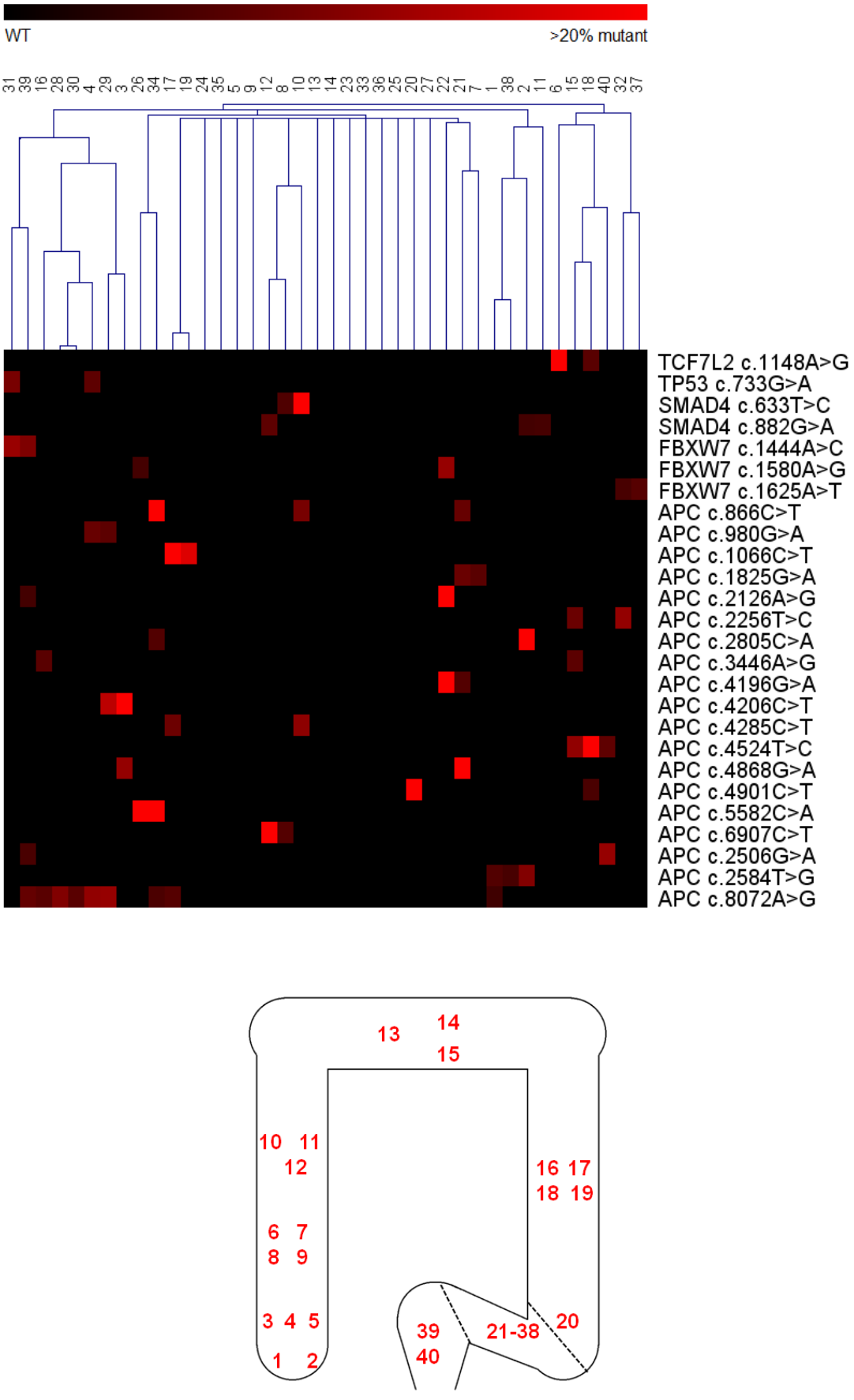


Figure 70. Heatmap with dendrogram for mutations called in 40 adenomas from patient 2 and their location within the bowel.

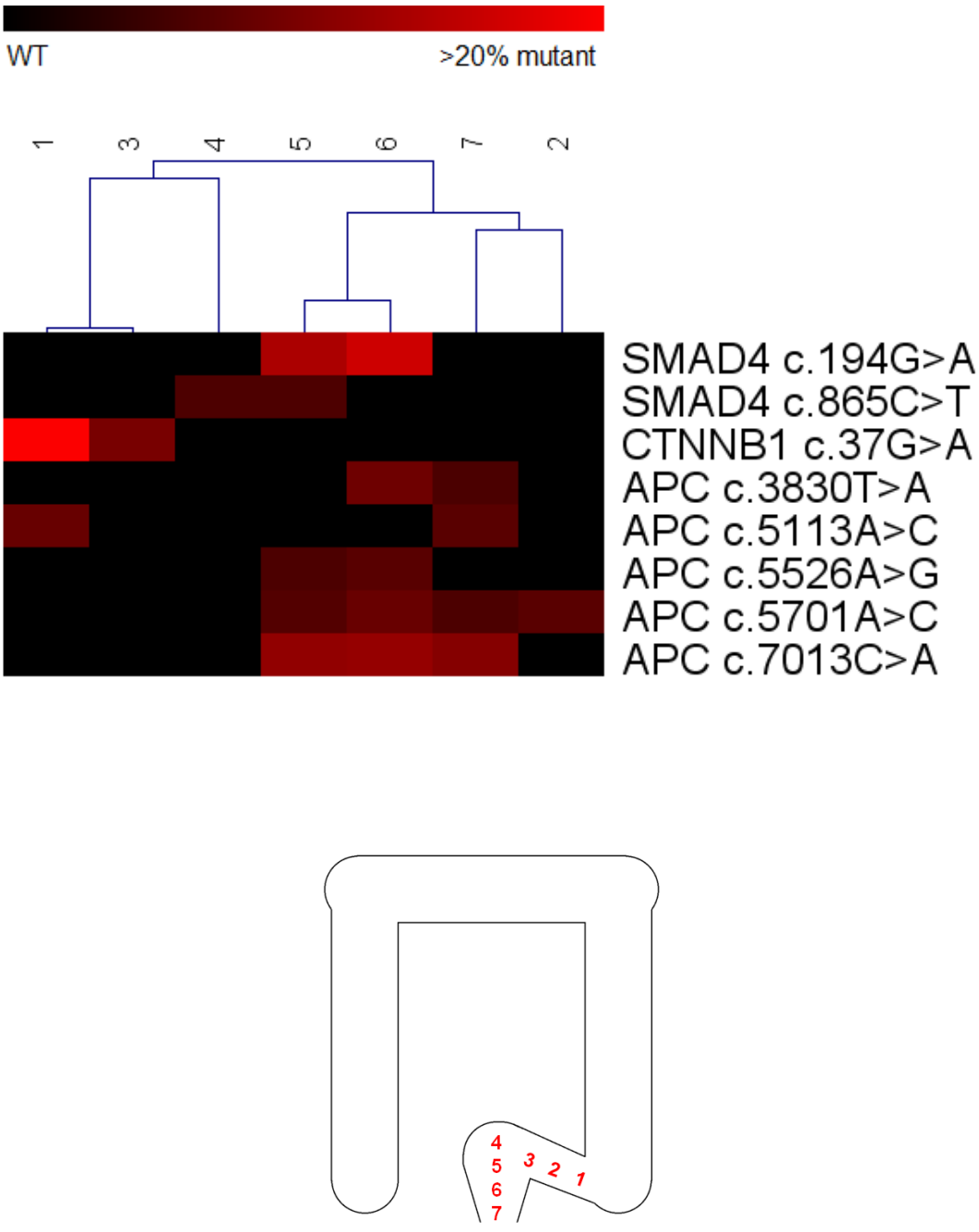


Figure 71. Heatmap with dendrogram for mutations called in 7 adenomas from patient 3 and their location within the bowel.

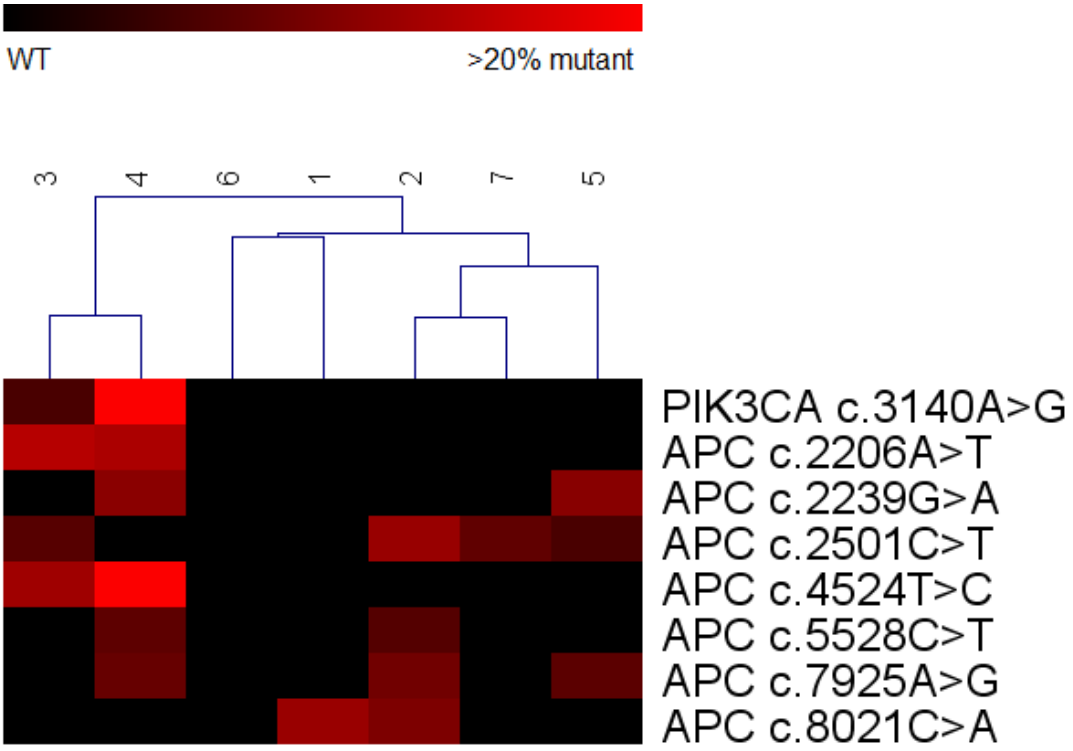


Figure 72. Heatmap with dendrogram for mutations called in 7 adenomas from patient 4.

4.5.6 Copy number profiles of FAP adenomas

Copy number profiles were generated for 75 out of 76 (99%) of the tested adenomas. Patterns of losses and gains in adenomas from the same patient showed that chromosomal aberrations were shared, but the amount of similarity varied between patients. The most common aberrations seen were gains in chromosomes 7,8 and 13.

Figure 77 -

Figure **82** show examples of lesions from the same patient. The percentage of the genome with abnormal copy number was calculated for each patient by dividing the number of windows across the genome that had an abnormal copy number (below 0.8 and above 1.2) by the total number of windows. This was then plotted against the maximum diameter of each adenoma to see if there was correlation. For patient 1 and 2, there was no correlation between the size of the lesion and the percentage of abnormal copy number across the genome (Figure 73 and Figure 74), however there was slight positive correlation seen for patient 3 and patient 4 (Figure 75 and Figure 76).

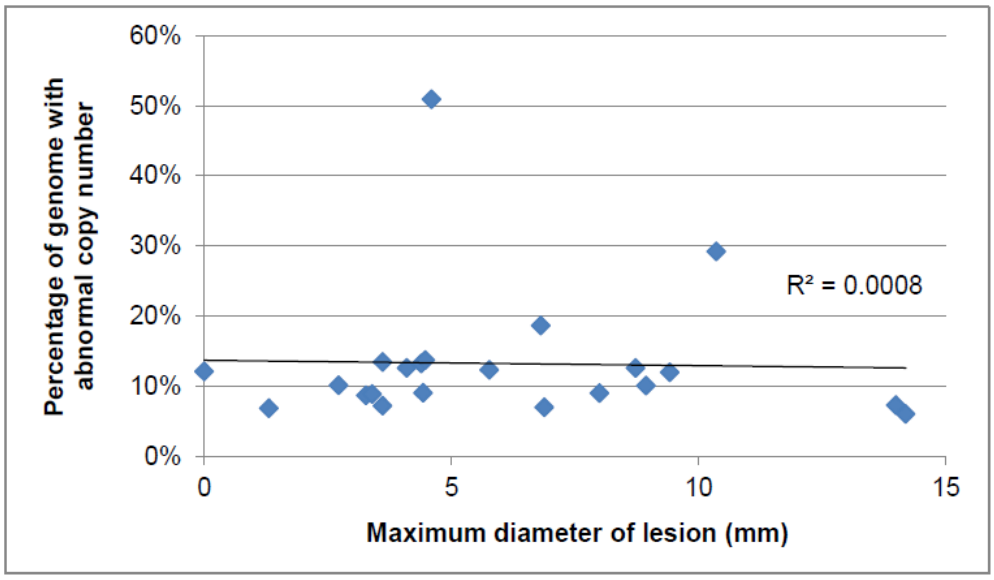


Figure 73. Percentage of abnormal copy number vs size of lesion for patient 1.

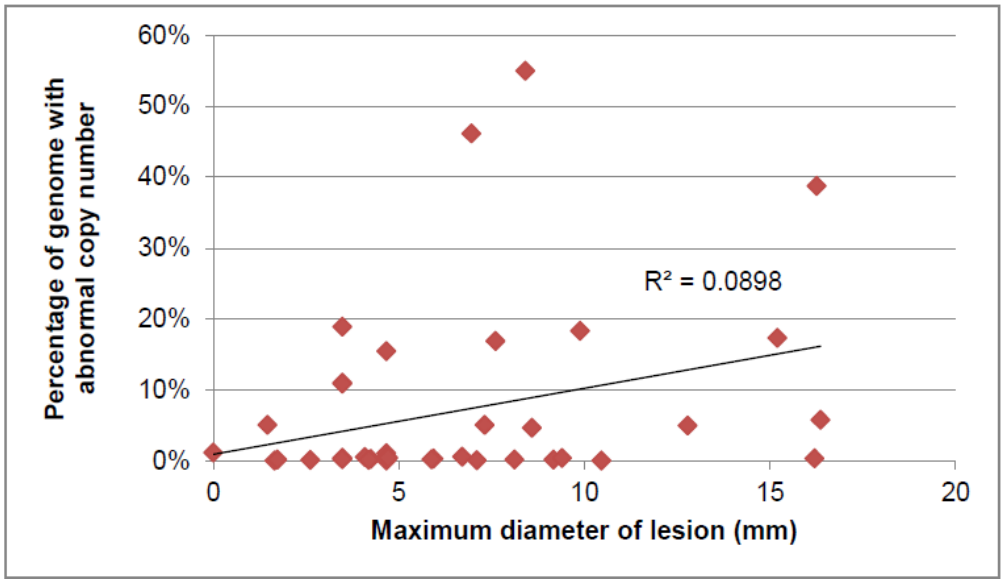


Figure 74. Percentage of abnormal copy number vs size of lesion for patient 2.

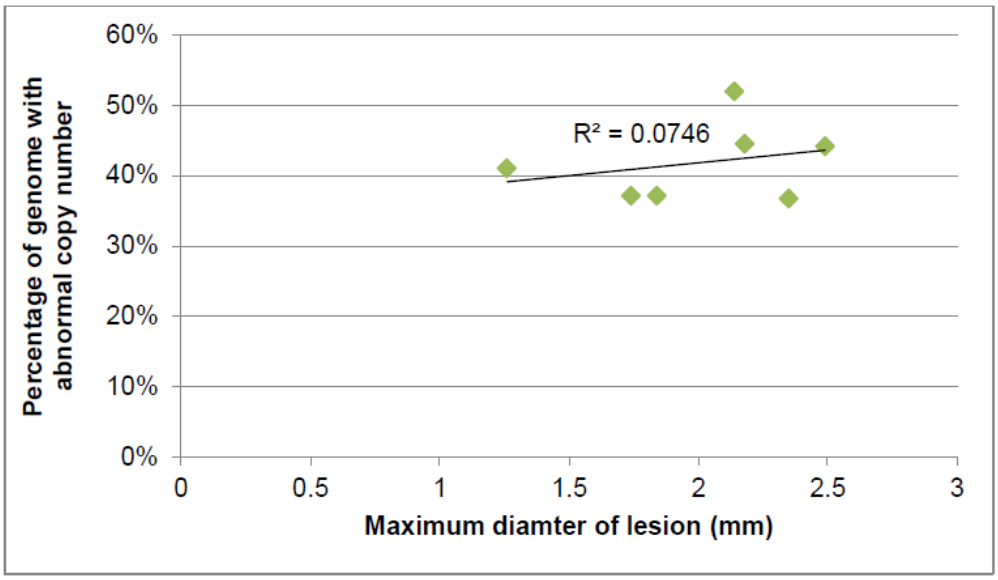


Figure 75. Percentage of abnormal copy number vs size of lesion for patient 3.

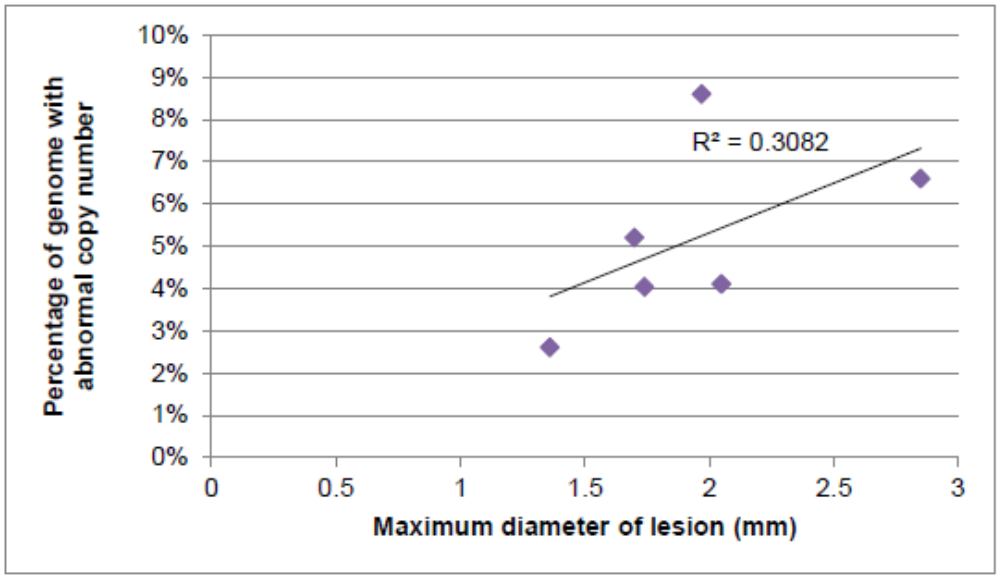


Figure 76. Percentage of abnormal copy number vs size of lesion for patient 4.

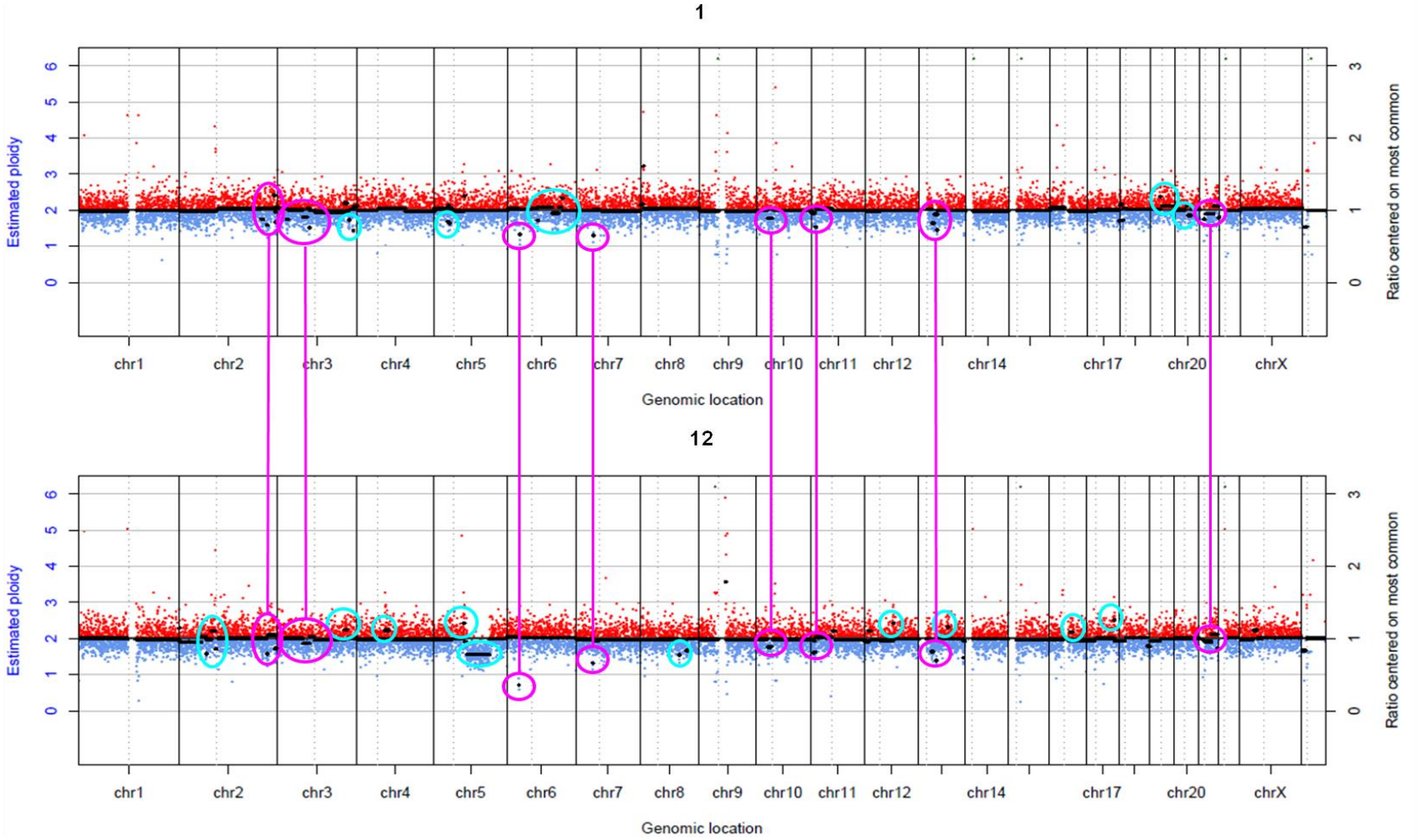


Figure 77. Copy number profiles of two lesions from patient1. Unique copy number changes for each adenoma are circled in blue and shared aberrations in purple. This shows a clear relationship between the two in 8 sites with at least 15 differences between them showing divergent evolution.

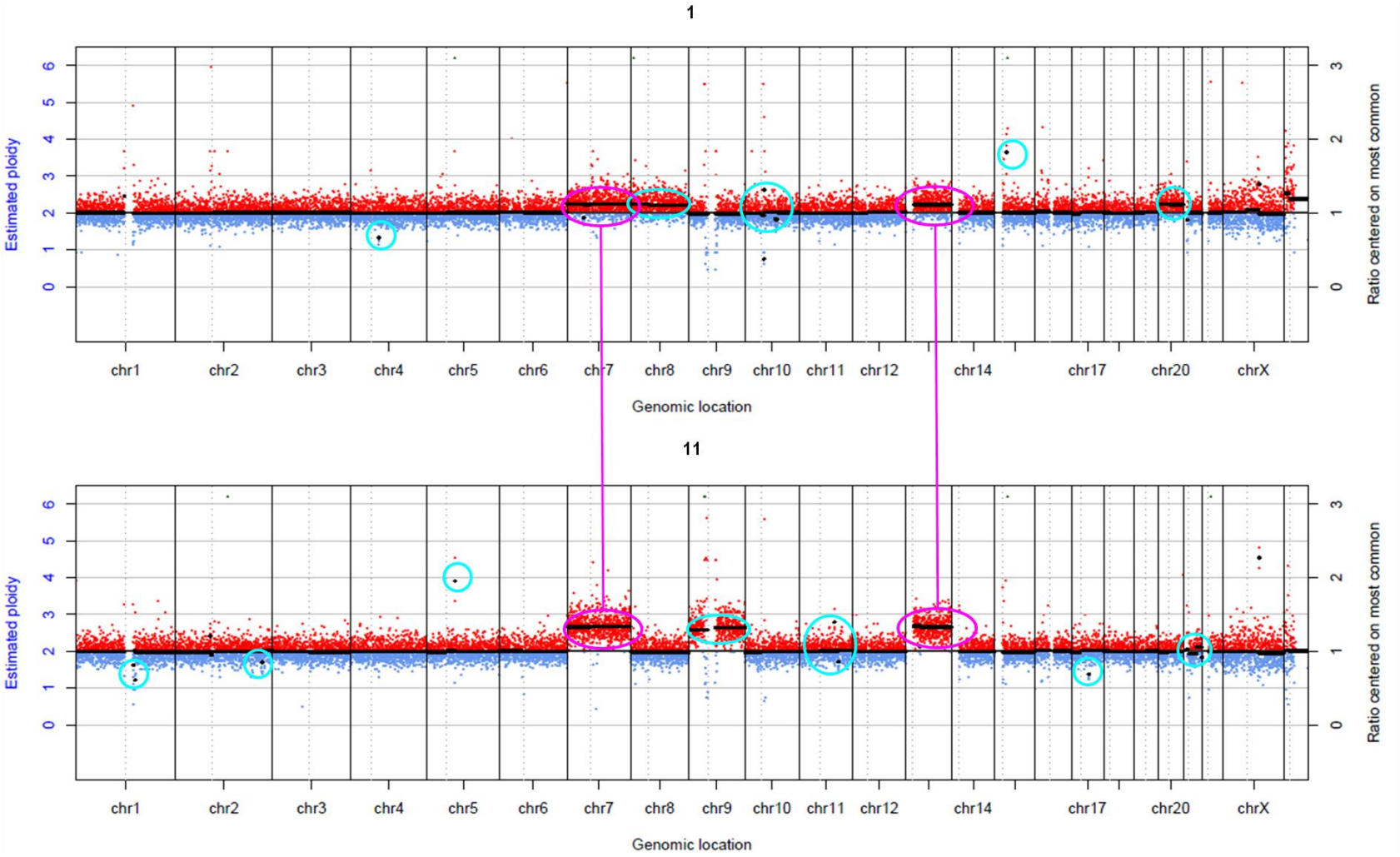


Figure 78. Copy number profiles of two lesions from patient 2. Unique copy number changes for each adenoma are circled in blue and shared aberrations in purple showing shared whole gains of chromosomes 7 and 13 as well as unique gains of 8 and 9.

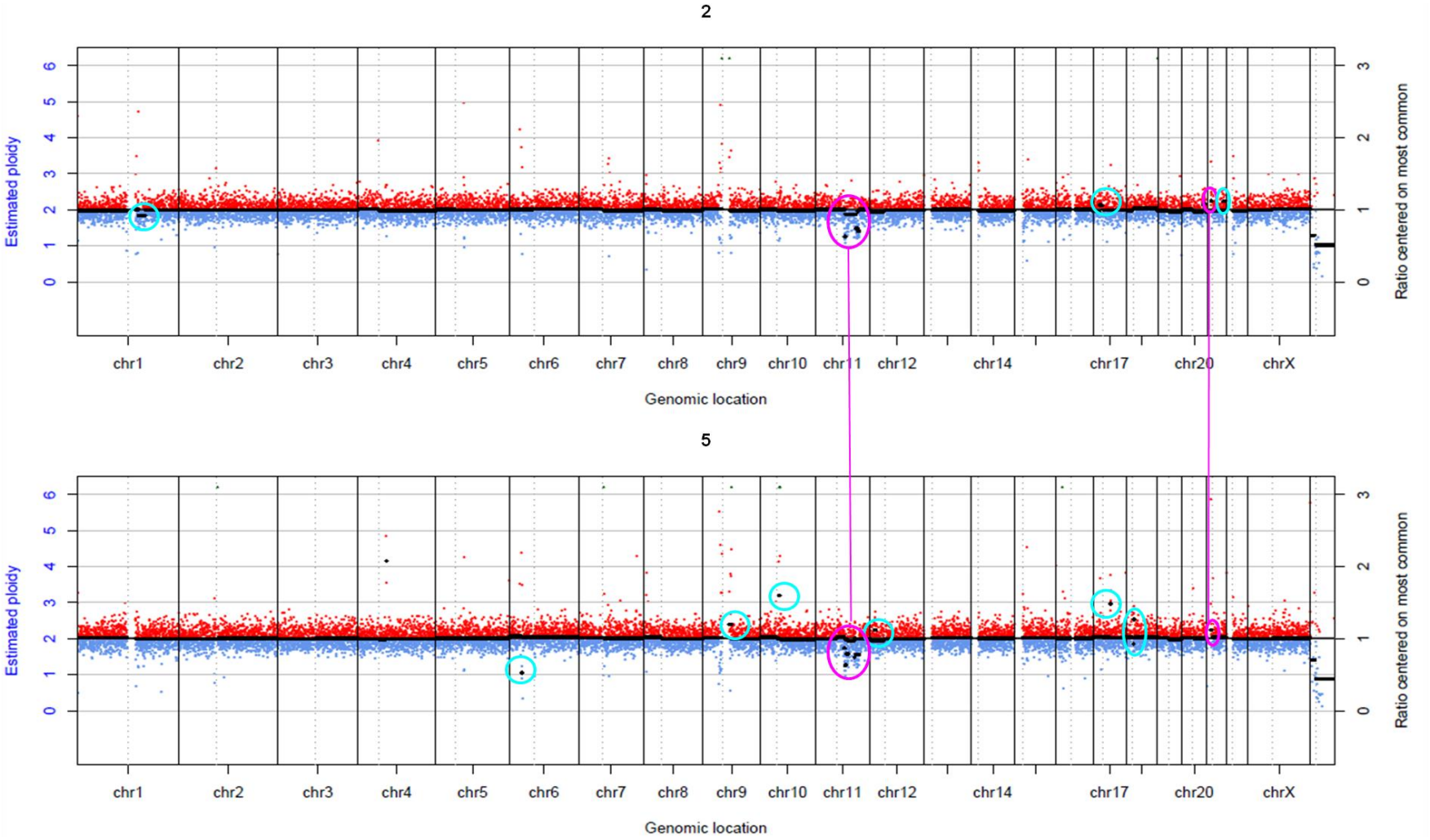


Figure 79. Copy number profiles of two lesions from patient 3. Unique copy number changes for each adenoma are circled in blue and shared aberrations in purple. These adenomas show whole chromosome gains of 7 and 13 which is frequent in adenomas but also many other changes.

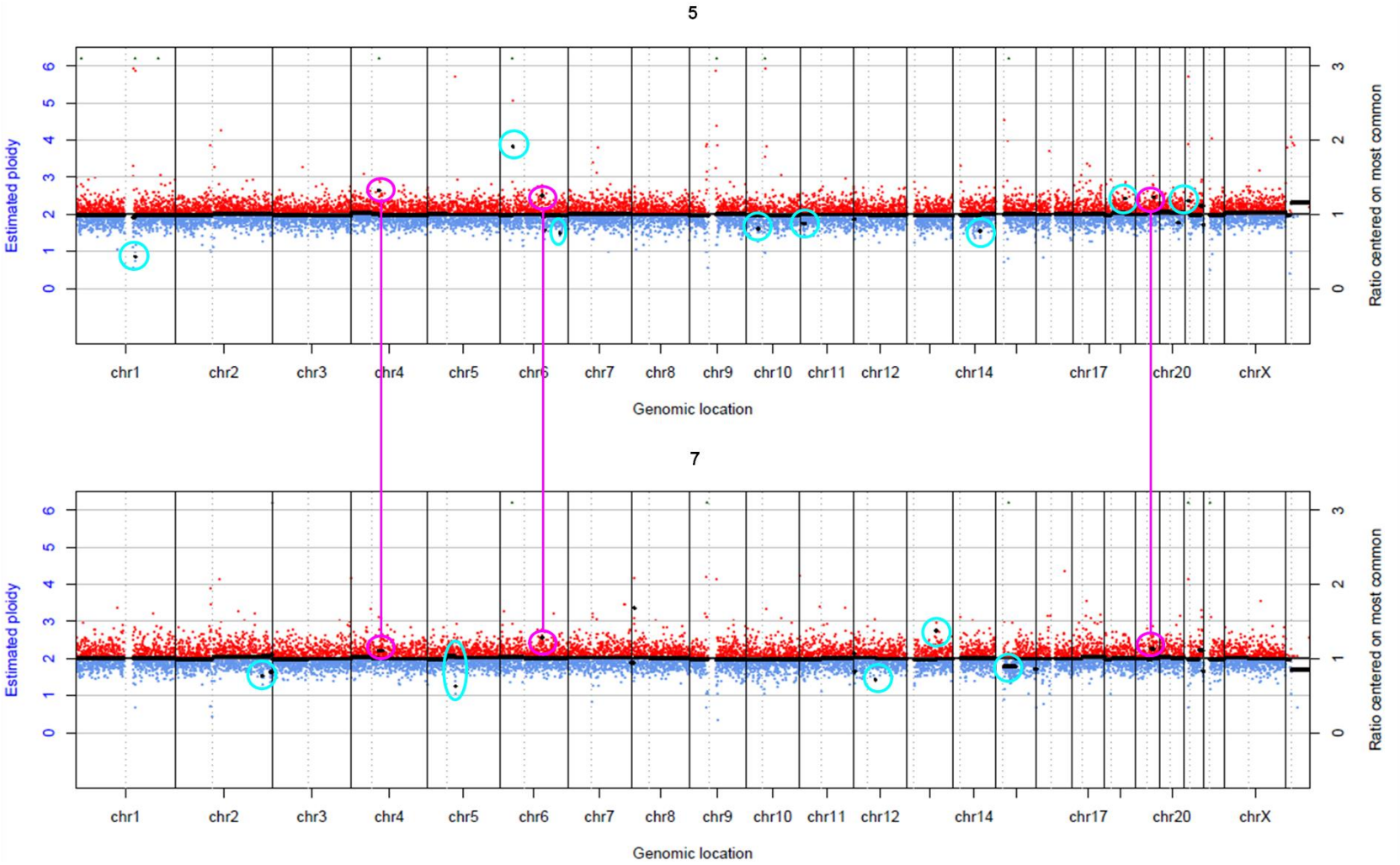


Figure 80. Copy number profiles of two lesions from patient 4. Unique copy number changes for each adenoma are circled in blue and shared aberrations in purple showing 3 small shared aberrations.

For patient 3, all adenomas contained a small aberration in chromosome 11. These can be viewed at greater resolution in Figure 81.

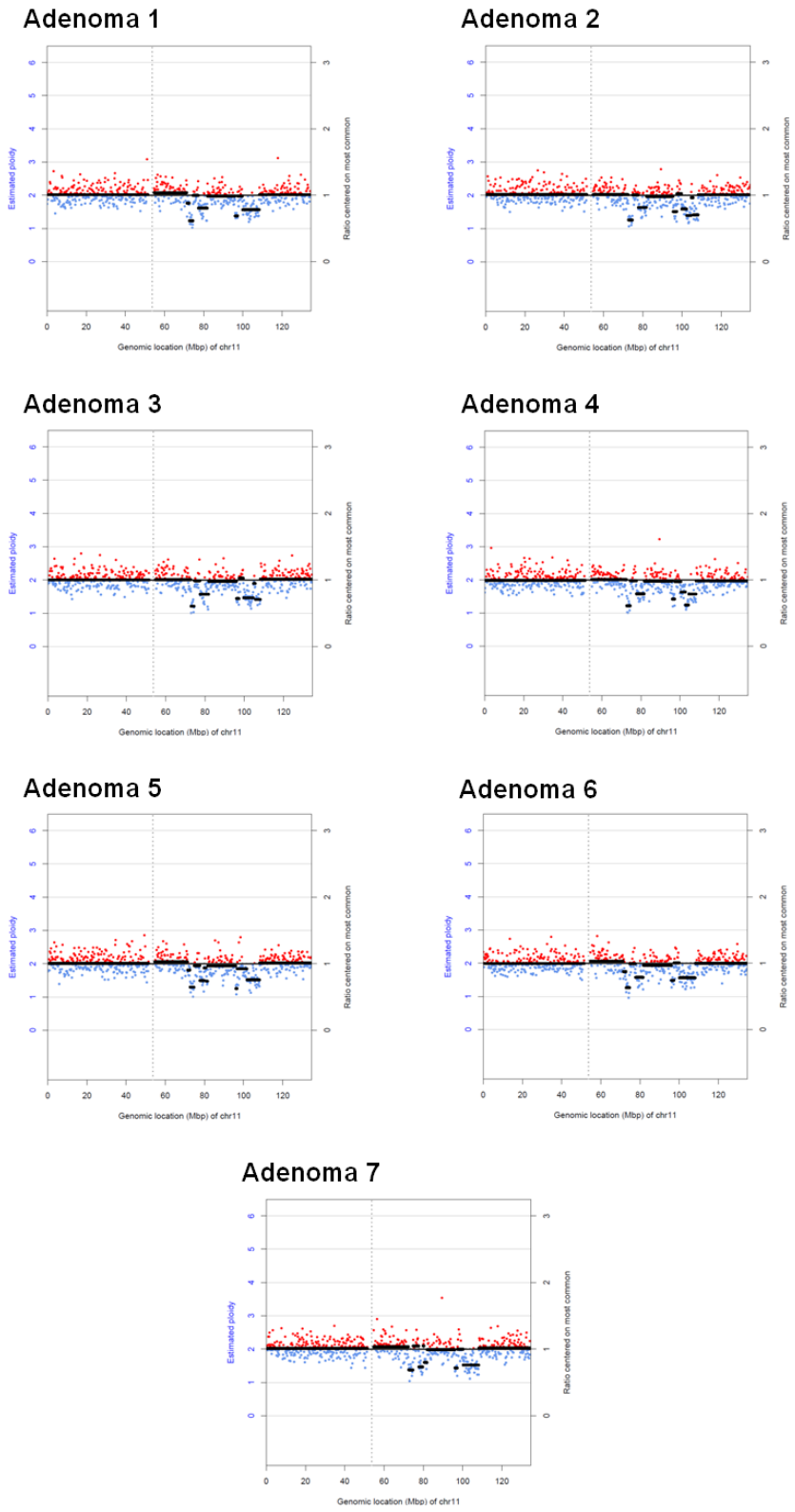


Figure 81. Plot of chromosome 11 for all adenomas from patient 3 to show different patterns of loss.

4.5.7 Clustering analysis

As before for the mutational profiles, heat-maps were drawn from the copy number matrices generated from each adenoma to show areas of gain (red) and loss (green). Positions where copy number was normal for all samples were omitted from the sample to allow for only areas of abnormality to be analysed. Hierarchical clustering was performed to assess similarity between adenomas in each patient. Adenomas from all patients were numbered in the order proximal-distal.

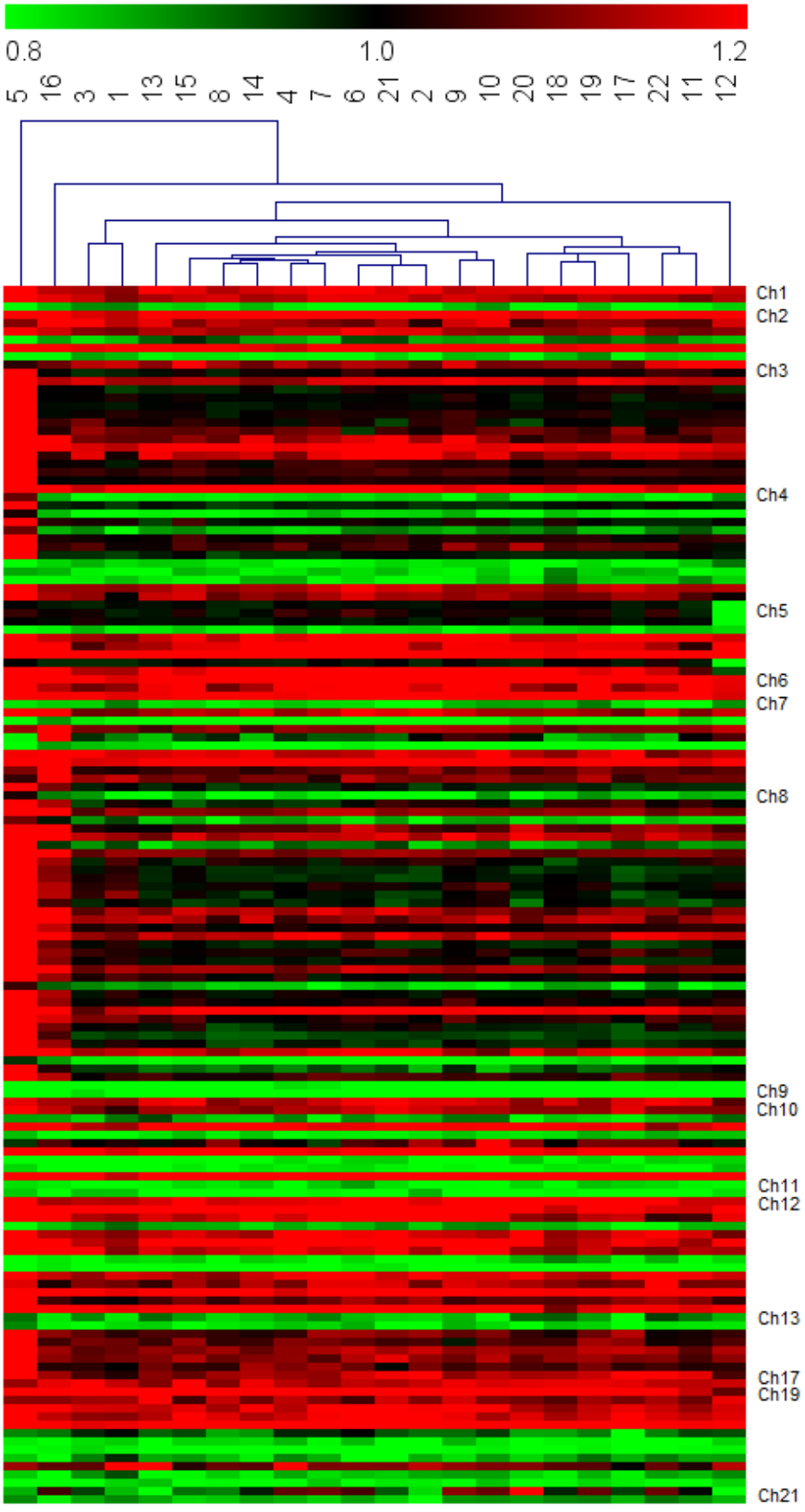


Figure 82**. Heatmap for copy number profiles in adenomas from patient1.**

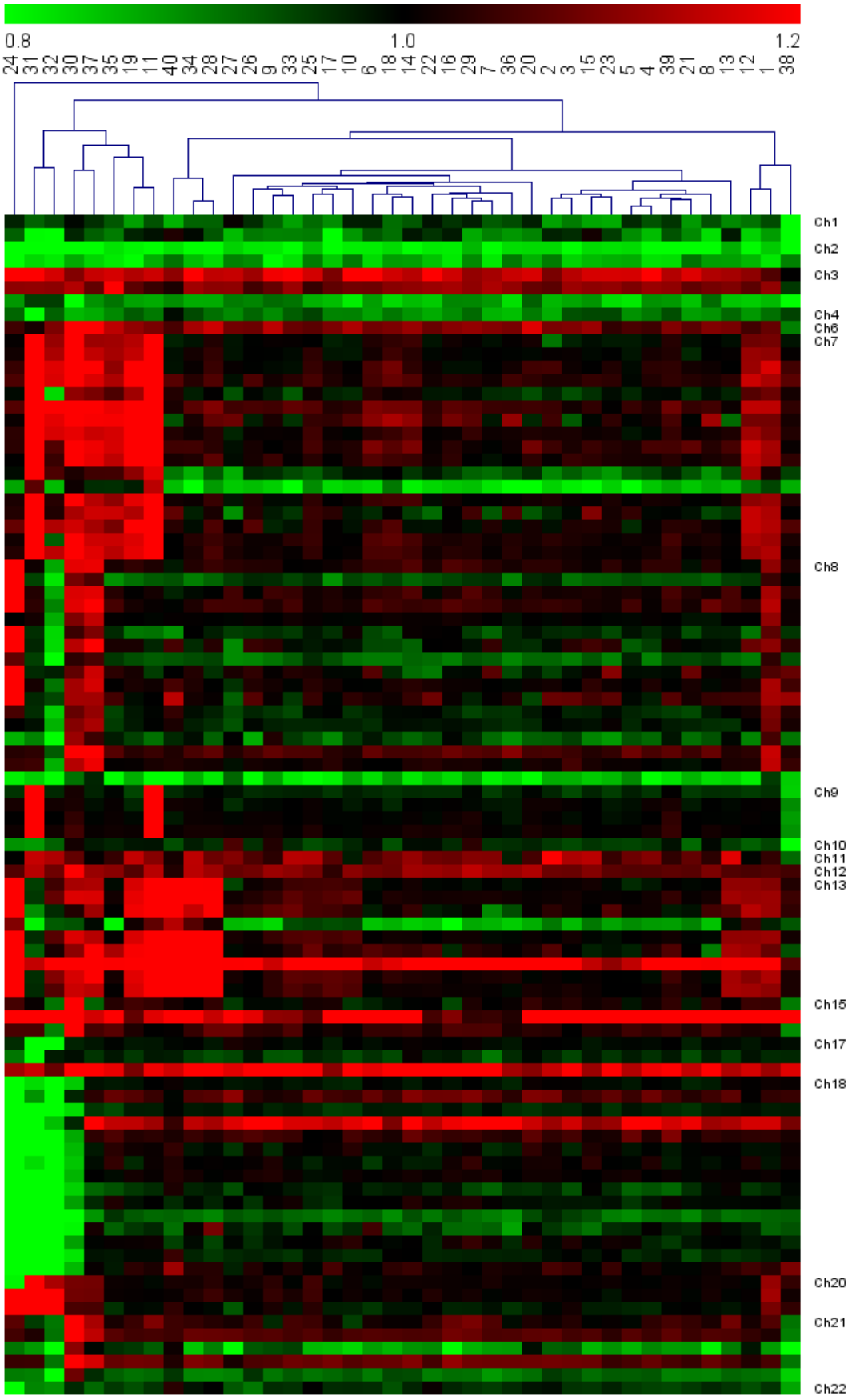


Figure 83. Heatmap for copy number profiles in adenomas from patient2.

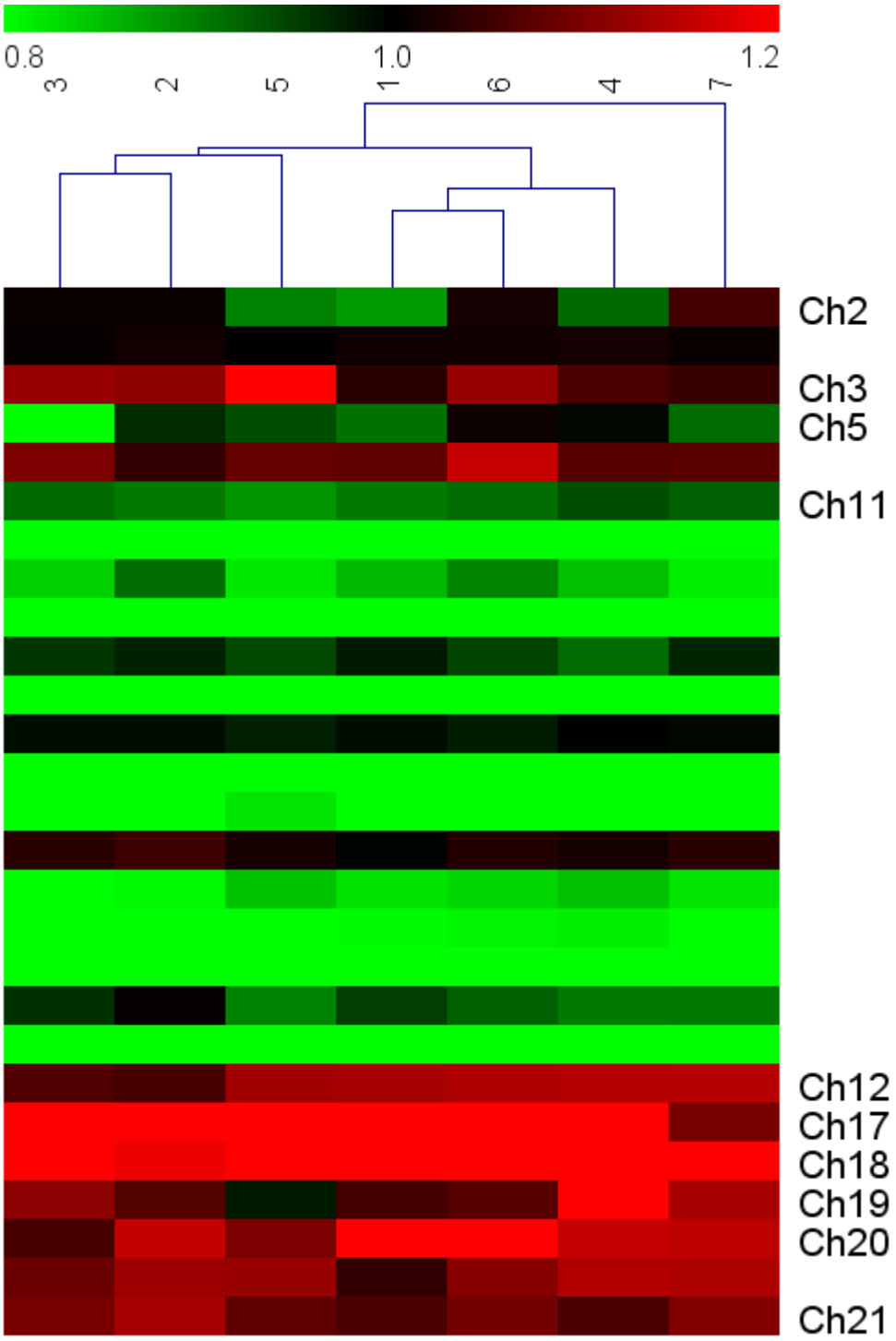


Figure 84. Heatmap for copy number profiles in adenomas from patient3.

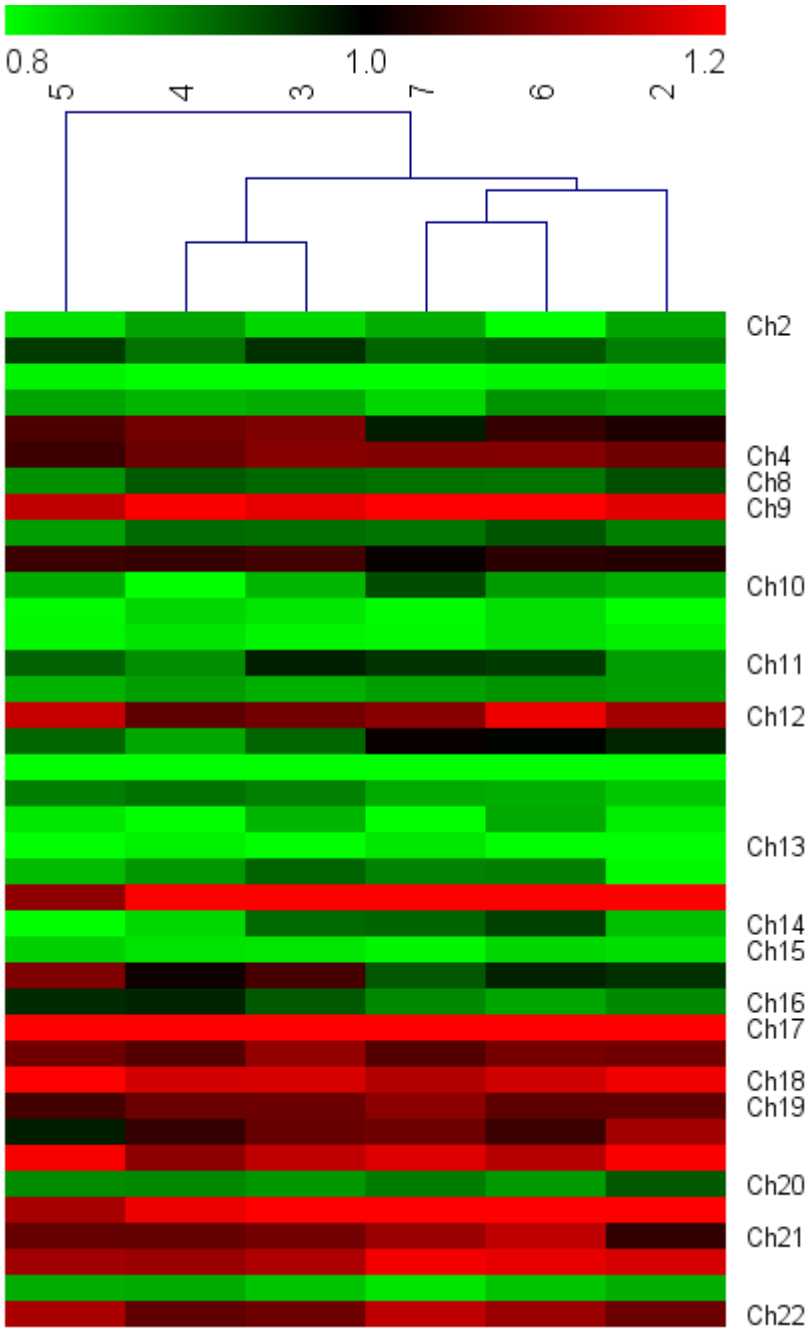


Figure 85. Heatmap for copy number profiles in adenomas from patient4.

Comparison between copy number and mutations

For each adenoma, the percentage of abnormal copy number across the genome versus percentage of mutations present in that adenoma of all shared mutations from that patient were plotted together (Figure 86). This was to determine whether copy number or mutations were having a dominant effect in any of the patients.

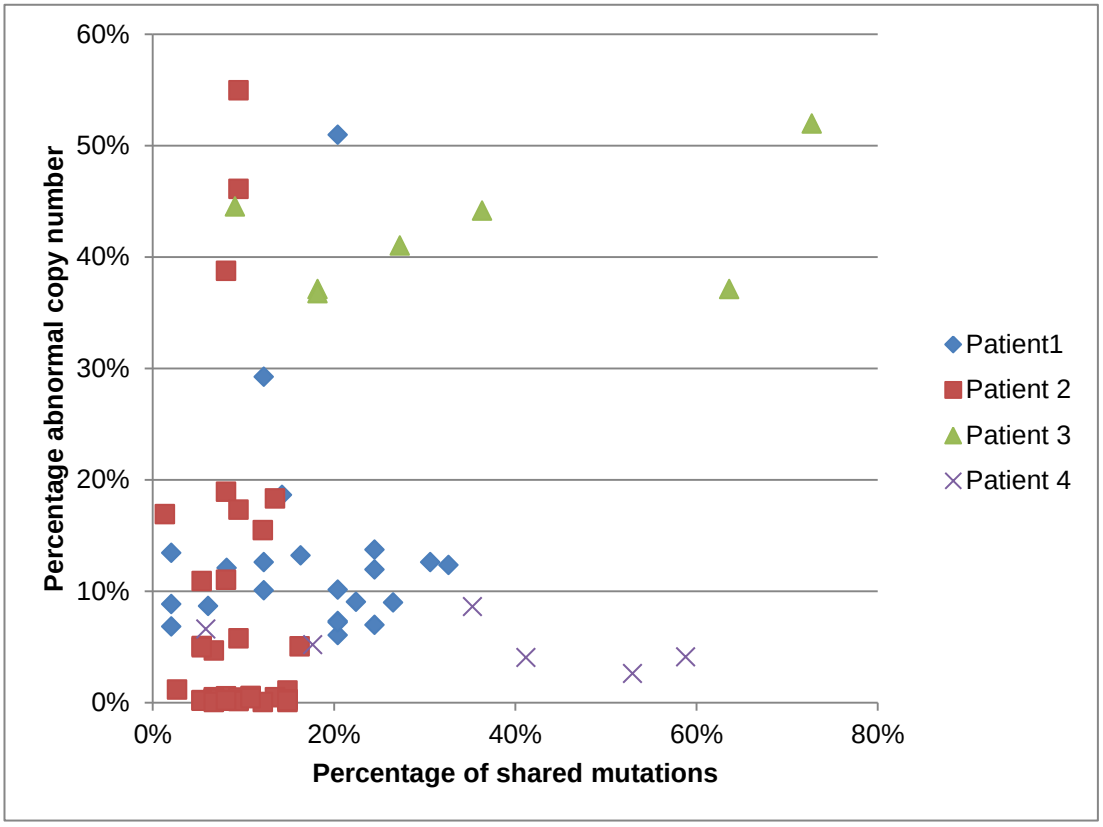


Figure 86. Plot of mutation rate against abnormal copy number rate for all adenomas from patients 1-4 showing clustering of patients according to the degree of mutations or abnormal copy number present.

4.6 Discussion

4.6.1 Use of Fluidigm for targeted sequencing

The ability of a technology to accurately detect mutations is highly dependent on the false positive rate and the ability to correctly identify and remove these false positives. By creating duplicate libraries using the Fluidigm Access Array (Fluidigm, San Francisco, USA) with a cut-off of 5%, it was revealed that 81% of mutations called in fresh tissue were present in both duplicates and only 11% in FFPE (Figure 64 and Figure 65). The mutations not detected in both duplicates are due to three main sources of error: formalin induced changes for FFPE samples, PCR error and sequencing error.

It has previously been described that formalin fixation results in deamination and depurination of cytosine and guanine residues (Williams et al., 1999, Kerick et al., 2011, Do et al., 2013) resulting in C>T and G>A artefacts in the PCR amplicons. The effects of formalin fixation have been previously explored in Chapter 2. In the cohort of FFPE samples that were sequenced alongside fresh samples, a significantly larger proportion of C>T and G>A changes were observed in mutations only detected in one of the repeats (Table 50). When the analysis was repeated for only mutations that were detected in both duplicates, this significant difference was removed, illustrating that by running samples in duplicate the bias in FFPE tissue for G>A and C>T changes can be removed. These artefacts have been reported to be significantly reduced by pre-treatment of gDNA with uracil-DNA glycosylase (UDG). Cytosine is converted to uracil through deamination and this is detected by the DNA polymerase as a thymine base which causes the resulting C>T change after PCR. UDG breaks uracil residues from their sugar phosphate backbone, causing the DNA polymerase to stop during amplification. As a result it has been reported that 60% - 81% of these artefacts can be removed from NGS targeted high coverage amplicon sequencing (Do et al., 2013). This information was available only after the experiment was run and this should be explored further.

For fresh tissue, as the minimum allele frequency cut-off for calling mutations was increased, the amount of agreement between each duplicate increased (Figure 66). This is likely due to the high number of starting template copies for fresh tissue. If a PCR error occurs in one template copy within an early cycle, the high number of template copies dilutes the effect of this error resulting in it having a lower allele frequency. For errors that occur in the later stages of the PCR, they are amplified less which results in them having a lower allele frequency. Therefore these errors can be filtered out by either having a higher minimum allele frequency cut-off or by running samples in duplicate. It would also be of value to run samples in triplicate and even quadruplicate to determine the point at which mutation calls do not change between multiple libraries from one sample.

For FFPE samples the amount of agreement between duplicates for mutation calling remained low, regardless of the minimum cut-off for mutant allele frequency. FFPE DNA is more fragmented than fresh (Hadd et al., 2013) and therefore contains fewer copies of starting template for PCR. Formalin can also cause cross-links to form between histones which may obstruct the polymerase, effectively reducing the amount of a starting template available (Teo and Shaunak, 1995). A low number of copies will mean that the error rate of the polymerase will have a more significant effect. Errors that occur within the early cycles of the PCR will be amplified and have a higher allele frequency. Therefore these cannot be filtered out by increasing the minimum allele frequency threshold as was seen for FFPE tissue.

Even with fresh tissue and a minimum allele frequency threshold of 10%, there were 11% of mutations that were not detected in both duplicates. This could partly be due to the polymerase used with the Fluidigm system. Although quoted to have a higher fidelity than standard taq polymerase, this enzyme is not the highest fidelity polymerase available and still induced errors within the PCR. These challenges can be overcome by running samples in duplicate and a higher fidelity enzyme should be optimised for the access array system. It may also be possible to determine the amount of starting template copies from FFPE tissue by use of qPCR, however, both of these methods have significant cost and time implications. However all of these options need to be tested to determine the optimal conditions for the Fluidigm access array in order to determine how reliable the assay can be.

4.6.2 Mutations in carcinoma and carcinoma associated normal tested in duplicate

For the cohort of fresh and FFPE carcinomas that were tested in duplicate, the distribution of mutations within the matched genes tested was similar to the distribution observed in the COSMIC database (Figure 60) (Bamford et al., 2004). Mutations were also seen in carcinoma-associated normal mucosa in APC, KRAS and SMAD4. There was a slightly higher number of SMAD4 mutations in the carcinomas (25%) than the COSMIC database (14%) but this was not significant (p=0.059). It has also been reported in other studies that SMAD4 mutation frequency in CRC is up to 35% (Miyaki and Kuroki, 2003). KRAS mutations have been previously observed in carcinoma associated normal mucosa (Minamoto et al., 1995, Yamada et al., 2005, Jonsson et al., 2010) as well as the cohort tested in Chapter 2 (ref). This experiment has shown KRAS mutations in normal mucosa in an independent cohort of samples with a different method of library preparation.

Mutations were also observed in APC in normal mucosa. APC loss is an early event within the development of colorectal cancer (Cunningham et al., 2010) and mutations in APC have been previously seen in very early lesions of colorectal cancer such as aberrant crypt foci (Orlando et al., 2008). When the minimum cut-off was set at 5%, the distribution of mutations in APC in carcinoma occurred across the gene and at particular sites that had a high frequency previously reported (Figure 62) suggesting that the data is reliable.

When the mutation minimum cut-off was lowered to 1% (Figure 63), 2 mutations in both carcinoma and carcinoma associated normal were clustered around the 3' end of exon 15. This has not been previously reported and the majority of mutations normally lie within the mutation cluster region. Only 2% of mutations reported in APC lie within this region of the gene (codon 1689 and above). It could be that these APC mutations are only weakly transforming and do not greatly affect the phenotype or as these mutations are low level (1% -5%) they could also likely represent PCR artefact in the amplicons that cover this region. Different polymerases can insert errors within the PCR at non-random points. Therefore it would be ideal to repeat the testing of this region of APC for these samples with a different polymerase to see if the mutations remained. Alternatively they could have

been falsely called by the VarScan variant caller software. It has previously been shown that there is poor concordance with different variant callers. A study that compared 4 of the most commonly used including VarScan, revealed that only 36 mutations were detected by all 4 variant callers out of a total of 2920 possible mutations (Figure 87) (Roberts et al., 2013). It would perhaps be beneficial to scan data with multiple variant callers however this does increase the time and computer power required to analyse data. Also without understanding the causes of the variation seen, it could be problematic and result in true findings being filtered out.

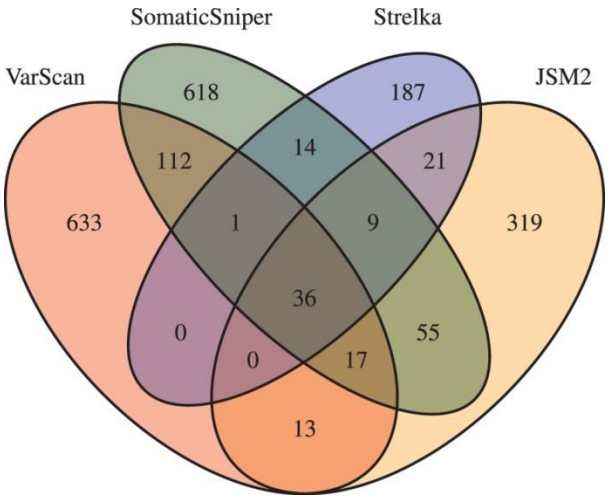


Figure 87. Overlaps between somatic single nucleotide variant sets from exome data. Reproduced with permission from (Roberts et al., 2013).

A mutation in CTNNB1 was seen in one of the samples of normal mucosa which also contained a mutation in the tumour. The mutation in the normal mucosa was different to that in the tumour. CTNNB1 mutations have been previously observed in histologically normal rat colon mucosa that has been treated with a carcinogen (Yamada et al., 2000).

The finding of SMAD4 mutations in carcinoma-associated normal mucosa is more unexpected. This could partly be explained by the presence of copy number aberrations. Loss of the 18q chromosome is reported in up to 60% of CRC (Woodford-Richens et al., 2001). Although the bioinformatic pipeline for calling mutations takes into account commonly reported SNPs there are always SNPs unique to that patient. These are normally filtered out by analysing the normal and

tumour samples in parallel. However, for cases where the tumour has 18q loss, a SNP in SMAD4 may appear to be a mutation in the normal mucosa. Obtaining copy number data for these samples alongside the mutational profiles would be useful to further investigate these SMAD4 mutations in normal tissue.

Figure 61 shows the mutational distribution when calling mutations at a minimum allele frequency of 1%. This resulted in a small increase in the number of tumours with mutations compared to 5%, which did not greatly affect the overall distribution for carcinomas. However, there was a much greater increase in the number of samples with mutations in the normal mucosa. This indicates that mutations in normal mucosa are either at a very low level in a small number of cells. This may represent the genetically damage background of normal mucosa in patients with carcinoma. Low level mutations in normal could either be carcinogen induced or due to faulty DNA repair or alternatively they could be artefact. To investigate this further, it could be possible to look for SNPs associated with cancer and see if this correlates with mutations in normal mucosa (Dunlop et al., 1997).

4.6.3 Carcinoma, adenoma and non-neoplastic normal

For the cohort of carcinoma, adenoma and their associated normals as well as non-neoplastic associated normal samples, libraries were not tested in duplicate since the error rate of the Fluidigm Access Array (Fluidigm, San Francisco, USA) was higher than anticipated. All samples were FFPE and therefore the mutational data obtained was not reliable enough to make mutation calls as the error rate was subsequently shown to be around 80%. However for the oncogenes tested, mutational calls could still be made with some confidence as only hotspots were interrogated. The probability of an error occurring within the one or two base positions of the oncogenes investigated compared to the entire 30kb of sequencing was low enough at 0.025% to call mutations at these positions with a degree of certainty. Furthermore a subset of the KRAS mutated carcinomas was validated by pyrosequencing. Where the mutant allele frequency was high enough, the mutations could be confirmed with pyrosequencing (Table 48).

For the 32 carcinomas tested, mutations were observed in hotspots of KRAS, NRAS and PIK3CA genes. Of the KRAS codons 12 and 13 mutated tumours, approximately a third of those also contained mutations within the normal mucosa. This is very similar to the other cohort of carcinomas tested for KRAS 12 and 13 mutations in chapter 3. For one sample, a mutation was observed in PIK3CA. It would be highly useful to repeat this sample to determine if this mutation is real as PIK3CA mutations have not previously been reported in normal mucosa. If this is a true mutation, there could be implications with regards to the use of aspirin as a chemopreventative agent. Aspirin has been shown to benefit those patients that have a PIK3CA-mutated tumour (Liao et al., 2012) and identifying PIK3CA mutations in normal mucosa could help identify those patients that might benefit from taking aspirin as a CRC preventative agent.

Of the 32 adenomas tested, mutations were seen in KRAS 12 and 13 and PIK3CA 545 and 546. There were no mutations seen in the adenoma-associated normal mucosa for these samples. Similarly, there were no mutations present within KRAS, NRAS or PIK3CA for samples of normal mucosa from patients with normal colonoscopies.

4.6.4 Mutational profiles of FAP adenomas

Mutations in FAP adenoma samples were determined by comparing the mutations present in each adenoma to a normal muscle control for each patient. This allowed for SNPs and PCR artefacts that occur at specific points within the amplicon to be filtered out. Thus identifying the germline APC mutation was not possible and due to ethical conditions the patients were anonymous therefore this information could not be obtained through medical records. For each patient, only sites where a mutation was present in one or more adenoma were included in the phylogenetic analysis. This allowed for random PCR errors to be removed from the analysis as the chance of a random error in 30kb of sequence to occur in same location for two or more samples was very low ($1x10^{-7}$%). Mutational profiles of adenomas from 4 FAP patients revealed different degrees of clustering based on their mutation similarities. Patients 1, 2 and 4 had adenomas sampled from the entire bowel, whereas patient 3 had adenomas located in the sigmoid and rectum.

For patient 1 the heatmap of mutations showed that the 11 adenomas on the right side of the colon contained more mutations (37) than the 11 adenomas on the left (20) (Figure 69). This difference was statistically significant (p=0.0105). This could be due to the microflora within this region of the bowel producing a higher number of carcinogens therefore causing more mutations to occur. The cluster analysis for patient 1 revealed that some lesions appeared to develop independently, whilst other lesions clustered together. The structure of the dendrogram revealed a cluster with adenomas 12, 16, 18 and 19 as these contained few shared mutations in the tested genes and were independent. Adenomas 8 and 17 contained the same APC mutation; however adenoma 17 also acquired a mutation in KRAS indicating that the APC mutation had occurred first. Adenomas 3 and 14 branched together and had the same mutations in APC and SMAD4. These lesions were located 10cm apart in the ascending colon which indicates that a mechanism allows for mutated cells to spread at least this distance.

Patient 2 had no significant difference in the number of mutations located in the left and right hand side of the bowel (Figure 70). The dendrogram from patient 2 showed a pattern of relatively independent adenomas throughout the bowel with a limited amount of clustering. A cluster of adenomas that did not branch further represented that these contained no shared mutations with any of the adenomas from this patient and were independent. On closer inspection of other parts of the dendrogram, there was branching of adenomas that were located in the bowel nearby. For adenomas 8 and 12, both from the ascending colon, they shared the same mutation in APC. It could also be seen where adenomas clustered due to them sharing a mutation but then one adenoma containing extra mutations such as adenomas 2 and 11. By chance, adenomas located closer together should share molecular lesions and adenomas located at a greater distance from each other should share less. However relationships are seen at a greater distance and therefore a mechanism is enabling clones to spread at greater distances throughout the bowel.

For patient 3 (Figure 71) adenomas 5 and 6 from the rectum shared multiple APC mutations. Adenomas 1 and 3 shared mutations in CTNNB1 and they were both located in the sigmoid. A similar pattern was also seen in patient 4 (Figure 72).

Adenomas 3 and 4 shared PIK3CA and multiple APC mutations. They were closely located in the colon but were embedded in two separate FFPE blocks. Adenoma 4 also contained extra low level APC mutations. The other adenomas from this patient branched separately, showing that they were relatively independent. Again, this shows that mutations spread locally however there may be independent events occurring in larger fields within the bowel.

Although some degree of clustering was revealed in adenomas through testing this gene panel, it is important to note that mutations in other genes could be occurring which may affect the relationships between adenomas. It would be very interesting to repeat this with whole exome sequencing however this would currently be very costly; both in sample preparation and the amount of sequencing required for sufficient coverage for sensitive detection. In order to check the reliability of the mutational phylogenetic trees it would be important to validate mutations. Therefore as this data stands, the copy number dendrograms are more reliable as the process is less prone to error.

4.6.5 Size of FAP adenomas and copy number aberrations

A trend could be seen for 3 out of the 4 patients between the size of the adenoma and the amount of abnormal copy number present across the genome. (Figure 73 - Figure 77). Aberrant copy number is associated with later stage tumour progression and therefore larger adenomas are more likely to have copy number aberrations. The most commonly seen chromosomal aberrations were gains in chromosomes 7, 8 and 13. These are also commonly reported in sporadic adenomas (Meijer et al., 1998, Jones et al., 2007).

4.6.6 Copy number profiles of FAP adenomas

Creating matrices from the copy number ratios was an effective way to observe clustering of adenomas as they enabled those adenomas that shared marker aberrations to cluster together on the phylogenetic tree. Copy number analysis of the same 80 adenomas from patients 1-4 revealed substantial clustering of adenomas. For patient 1 (Figure 82), many adenomas were fairly independent, similar to the mutational analysis dendrogram. However, adenomas 5, 12 and 16 contained much larger chromosome structural changes than the other adenomas tested. Interestingly, 5 and 16 both contained whole gains of chromosome 8, but adenoma 5 also had gains in 3, 13 and 14 whilst adenoma 16 had a chromosome 7 gain. This would indicate that these adenomas contain a common origin, but had then evolved independently. Adenomas 6 and 21 clustered according to their CNV profiles even though adenoma 6 is located within the ascending colon and adenoma 21 is from the rectum (Figure 88). Although they are located far apart within the colon, copy number analysis revealed these adenomas to share a molecular background even though they were not related by their mutational profiles. This suggests that multiple mechanisms drive adenoma formation and that molecular aberrations can spread over a distance within the bowel.

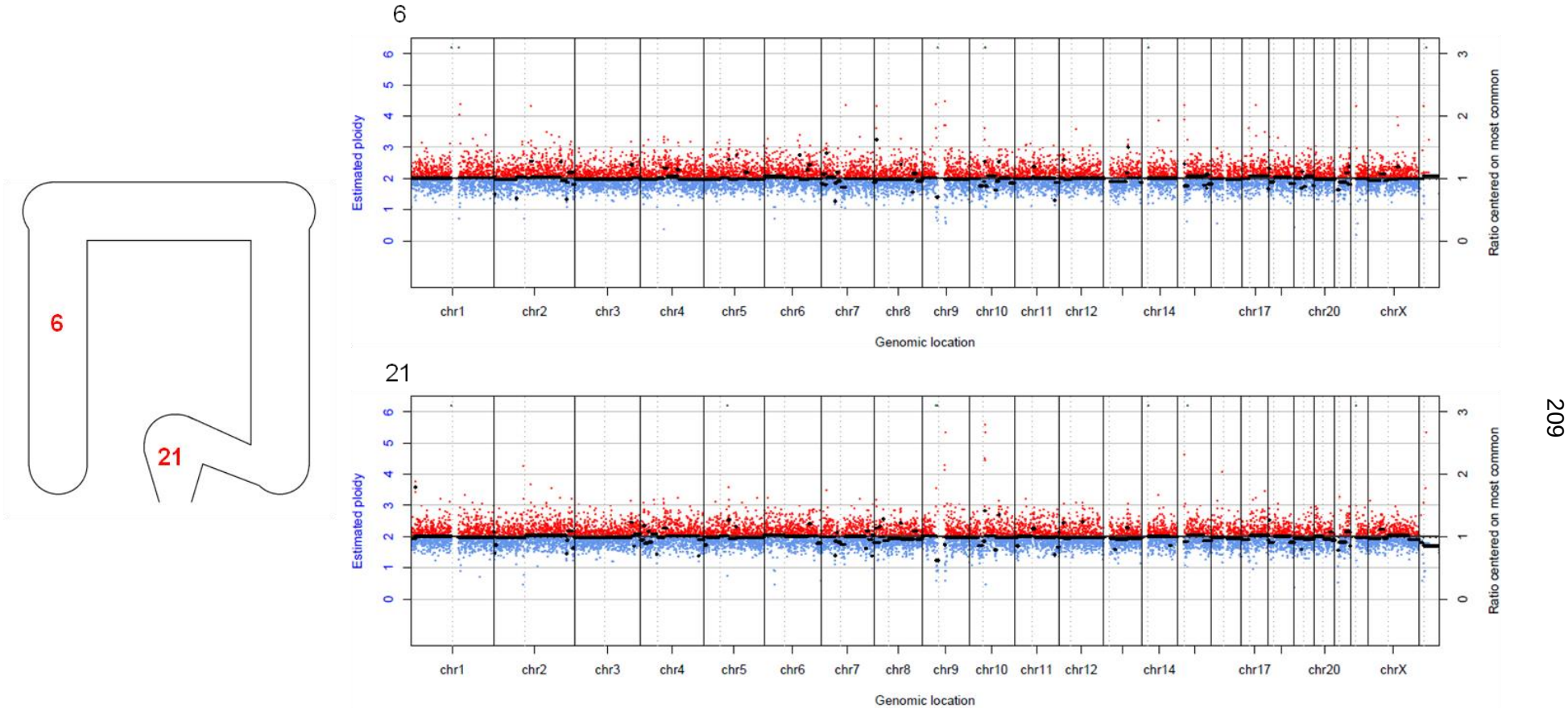


Figure 88. Copy number profiles of adenomas 6 and 21 from patient 1 located in the ascending colon and rectum respectively.

Adenomas from patient 2 clustered into two main subtrees with adenoma 24 clustering independently (Figure 83). This adenoma adenomas had gross chromosome structural changes and was independent of all other adenomas tested.

For the 8 adenomas that clustered together they all shared a chromosome 7 gain (Figure 89) and some also had a chromosome 13 gain (Figure 90). These lesions may have evolved from the same background, or chromosome 7 and 13 gains could have occurred independently at the proximal and distal ends of the bowel, creating two similar cancer fields. Out of all 40 adenomas tested in patient 2, 9 had a chromosome 7 gain (23%) and 11 had a chromosome 13 gain (28%). This is similar to the frequencies previously reported in adenoma (Meijer et al., 1998, Pino and Chung, 2010). Therefore the probability of an adenoma containing both by chance is 6% which equates to 2 out of the 40 adenomas. For patient 2, 6 out of the 40 adenomas (15%) contained both chromosome 7 and chromosome 13 gains which is not statistically significantly higher than the expected 5% of adenomas. Therefore it cannot be said that lesions from the caecum can spread through to the rectum and these adenomas may have both developed the same copy number aberrations by chance.

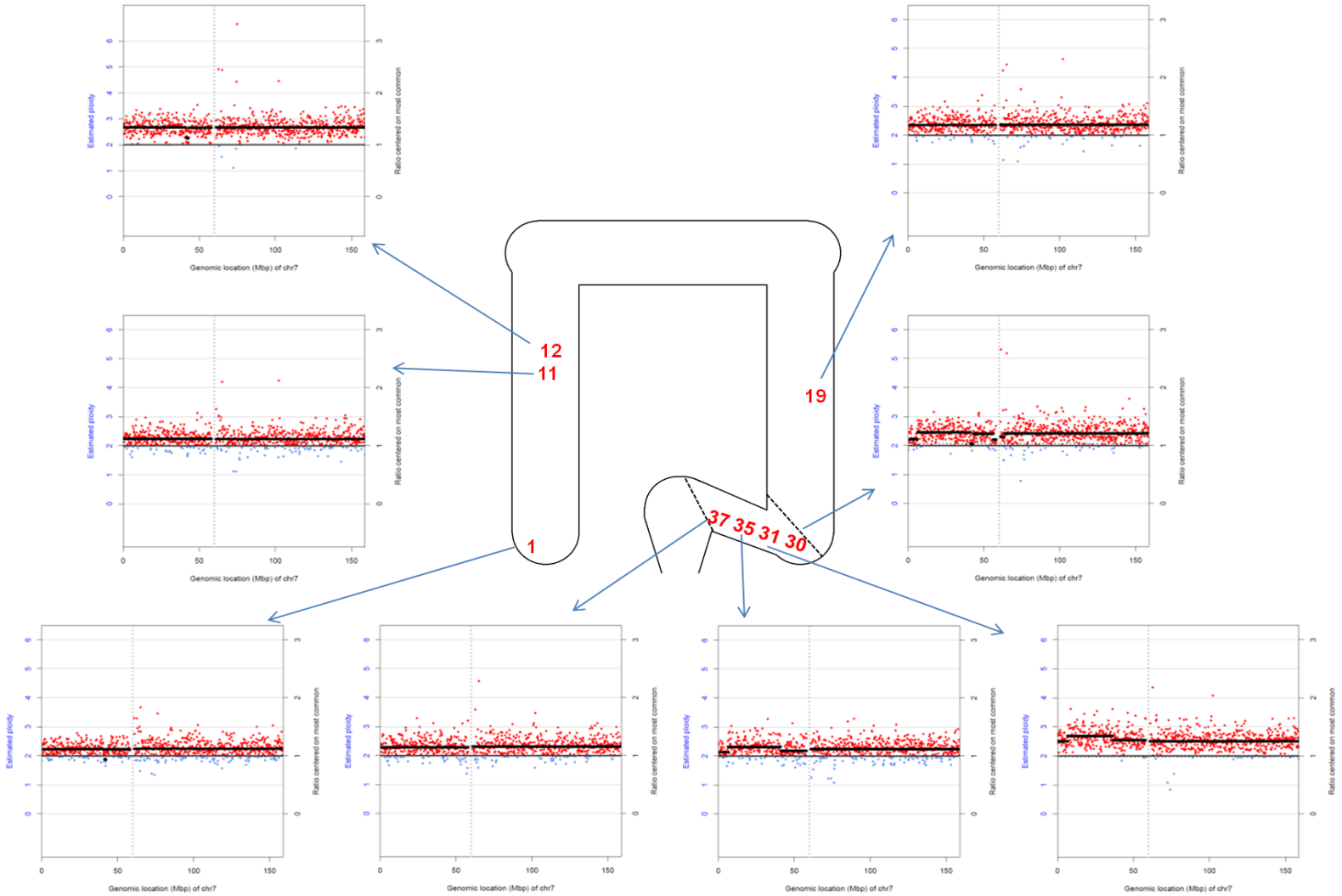


Figure 89. Adenomas with chromosome 7 gains for 8 adenomas clustered in cluster 1.

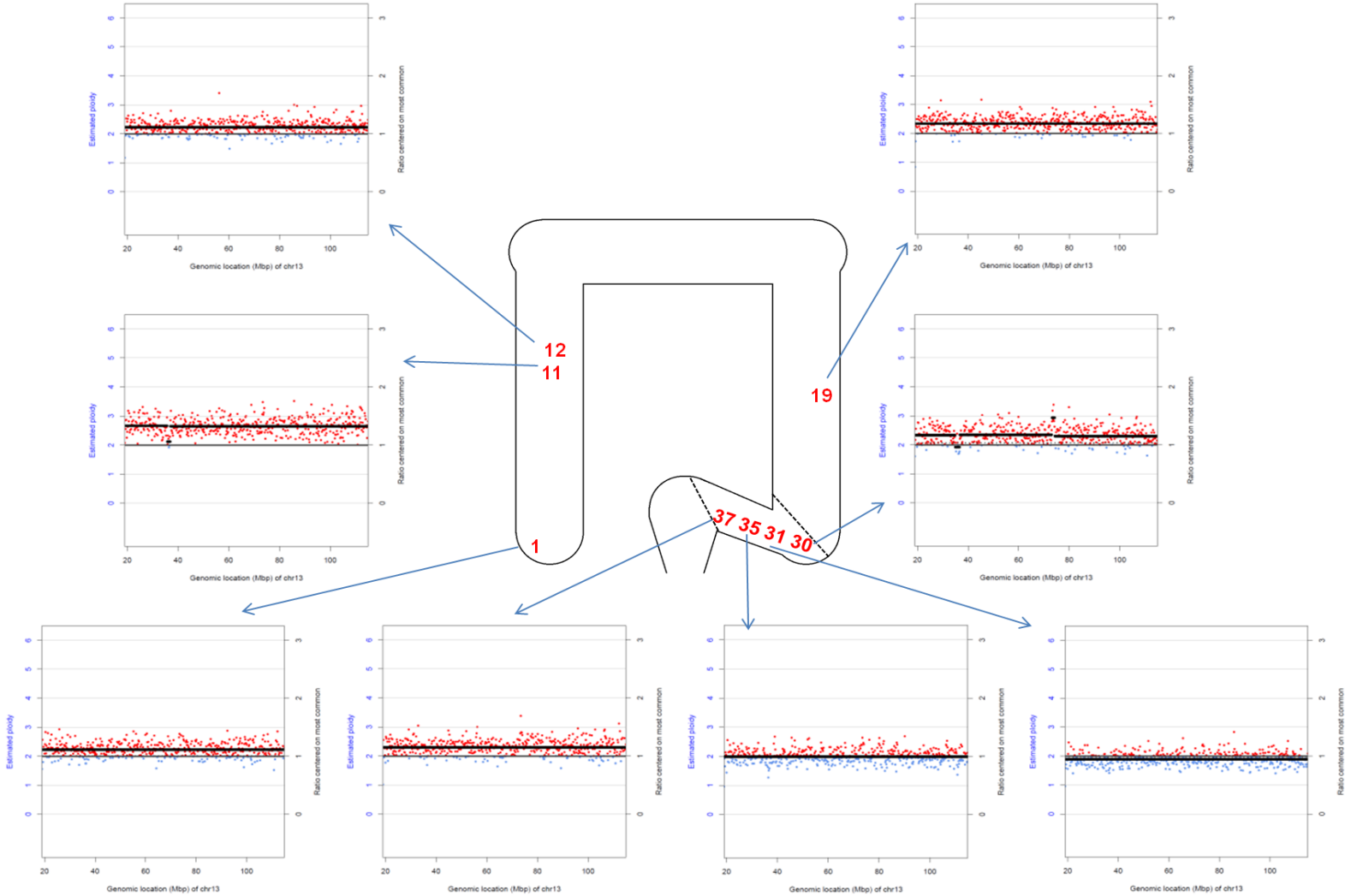


Figure 90. Adenomas with chromosome 13 gains for 6 out of 8 adenomas clustered in cluster 1.

For cluster 2 a cluster of 4 adenomas from the descending colon and rectum all contained a chromosome 13 gain. For the other adenomas there were very few copy number changes present. This cluster consisted of 25 adenomas located throughout the bowel. This pattern of grossly abnormal adenomas located nearby to adenomas with few copy number changes suggests that there are other factors involved in the evolving cancer field and that the dynamics are highly complex.

In contrast, adenomas from patients 3 and 4 contained relatively few large chromosome structural changes. For patient 3 (Figure 84), there was little clustering of the adenomas due to their copy number profiles being relatively normal. Adenomas 6 and 7 from the rectum both had a small gain at 12p, causing them to branch together on the nearest-neighbour tree. All of the adenomas shared a loss at chromosome 11, but the patterns of loss differed slightly between adenomas (Figure 81). This is evidence of the evolution of a shared change showing that the adenomas are related.

For patient 4 (Figure 85), lesion 7 had a 15q deletion but otherwise had very minimal changes. The other adenomas from this patient appeared to be relatively normal. The contrast in copy number changes not only throughout the bowel for each patient but also comparing different patients illustrates that chromosomal changes are only one of many contributing factors to adenoma development in FAP.

4.6.7 Comparison of copy number and mutations

For all four patients, there was very little similarity between the clustering of adenomas according to their mutational profiles and copy number aberrations. It is worth noting that for many of the adenomas, there is little copy number change and so these adenomas with cluster together with very short branching but may branch out and cluster differently as they contain more mutations. This difference in the dendrogram patterns may be due to mutational events and larger chromosomal changes occurring at different points within the development of the adenomas. However, it could also be due to the mutational data containing artefacts that could be interfering with the true clustering patterns.

By plotting all adenomas together according to how mutated they are and how abnormal their copy number is, it can be seen that adenomas cluster together and appear to have different trajectories (Figure 86). Patient 1 and patient 2 mainly cluster together having low mutation rate (30% and below) and low abnormal copy number rate (20% and below). There are a few adenomas from each patient that are outliers from this cluster that have a low mutation rate but a high abnormal copy number. For patient 1 the median adenoma size was 4.54mm (IQR = 3.61 – 8.55) and one of the outliers with an abnormal copy number rate of 30% was 10mm in diameter. Therefore the high copy number rate could be driving its growth. Similarly for patient 2, the median adenoma size was 4.70mm (IQR = 3.48 – 8.46) and the 2 adenomas with the largest abnormal copy number rate at 39%, 46% and 55% were 16mm, 7mm and 8mm in diameter respectively. Patient 4 has two lesions that cluster in the low mutation/low CNV region of the scatter plot, but also has 4 that have a higher mutation rate but still lower copy number. This could indicate that the adenomas from this patient are proliferating through a different mechanism that is more mutation-driven. Finally patient 3 has lesions that are all of a high abnormal copy number rate (35% and above) with a range of mutation rates. This patient is the oldest (27years) in comparison to patient 1 (16years), patient 2 (23 years) and patient 4 (18 years). This could be an indication why their adenomas have high mutation rates and high abnormal copy number as more aberrations have been acquired over a longer period of time. Alternatively for this patient the mutations could be having less impact than copy number during growth. Overall this plot suggests that there could be different mechanisms driving adenoma growth and one mechanism may be dominant or both may occur.

4.7 Chapter Summary

- The Fluidigm Access Array (Fluidigm, San Francisco, USA) as an NGS library preparation method allows for multiple mutations to be detected to a low allele frequency in highly multiplexed samples when tested in duplicate.
- FFPE material contains a significant proportion of G>A and C>T artefacts compared to fresh tissue.
- KRAS mutations can be detected in carcinoma-associated normal mucosa using the Fluidigm Access Array (Fluidigm, San Francisco, USA) in a separate cohort of patients, confirming this does occur.
- Mutations can also be detected in APC, SMAD4, CTNNB1 and PIK3CA in carcinoma-associated normal mucosa.
- No mutations have been detected in adenoma-associated normal or non-neoplastic normal mucosa in KRAS, NRAS BRAF or PIK3CA at 5% sensitivity.
- FAP adenomas show a positive correlation between size and degree of abnormal copy number across the genome.
- FAP adenomas cluster locally within the bowel both by their mutational and copy number profiles.
- Abnormalities, both mutational and copy number occur more frequently in the same areas of the bowel but can also occur at a great distance.
- Different mechanisms appear to drive different patient's adenoma growth processes.

5 Discussion, conclusions and future directions

5.1 Discussion

Colorectal cancer has a survival much poorer when detected at a late stage; Dukes' C and Dukes' D tumours have a 5-year survival rate of less than 50% (Cancer Research UK, 2012b). Therefore detecting changes early to prevent the development of cancer as well as finding it at an early more treatable stage is key to improving outcomes. This has already been seen with the implementation of the bowel cancer screening programme. Faecal occult blood testing has reduced the stage of presentation of carcinoma and the introduction of flexible sigmoidoscopy will identify and remove adenomas and as well can detect early cancer. Looking for genetic changes within the normal bowel mucosa before phenotypic changes occur might allow the identification of individuals who may progress to cancer at an even earlier stage.

5.1.1 Crypt isolation

In order to look for mutations in normal mucosa, it would be beneficial to isolate only epithelial cells. This is to allow for more sensitive detection as epithelial cells make up 51.7% of the mucosa. By extracting DNA from the whole mucosa, potential mutations are diluted and the detection technique has to be more sensitive to compensate.

Of the two isolation techniques investigated: mechanical dissociation of colonic crypts in the presence of a chelating agent (MD) and laser-capture microdissection (LCM), only MD was able to produce a consistently high enough yield of DNA for reliable downstream use. Despite a considerable time commitment for trouble-shooting LCM, attempts were unsuccessful and whole mucosa had to be used for FFPE DNA extraction. It is worth noting that previous studies using LCM for colonic crypt isolation have looked at a small number of gene targets with Sanger sequencing (Humphries et al., 2011) which is non-quantitative and insensitive and cannot detect low allele frequencies below 20%. For a targeted NGS approach it is important that a high quantity of gDNA is used to ensure a high starting number of template copies to reduce error. Therefore LCM may not be a suitable technique to use where NGS of whole mucosa is sufficiently sensitive.

5.1.2 Sensitivity of mutation detection

In order to be certain of the presence of mutations in normal mucosa it was necessary to obtain the highest sensitivity and specificity. Therefore 3 PCR based enrichment techniques were investigated (RFLP, COLD PCR, ICE COLD PCR) as well as pyrosequencing and ultimately NGS as it became available.

The COLD-PCR enrichment methods failed to produce any significant mutant allele enrichment. This is a relatively new approach that has not yet been replicated by other groups and has recently become available as a commercial PCR kit (Trangenomic Inc, 2013). It is highly temperature dependant and many environmental factors may interfere with the effectiveness of its enrichment. RFLP produced a moderate enrichment for pyrosequencing, but due to the increased chance of sample contamination and extra PCR cycles amplifying errors it was decided that the benefit of this technique could not justify its use when NGS was sufficiently sensitive.

The sensitivity of pyrosequencing was found to be relatively insensitive and lie between 5% to 25%, depending on the type of base substitution detected. This is comparable to other similar pyrosequencing studies which reported a 5% sensitivity, however they only tested a G>A base change which pyrosequencing is best at detecting (Tsiatis et al., 2010). NGS was found to have a sensitivity of 1%. Therefore in order to consistently be able to detect mutations at 5% an NGS approach was chosen.

5.1.3 Development of TALC

The original method of adaptor ligation was not an ideal method for preparing PCR amplicons for sequencing. This was due to the multiple steps the process involved which led to high risk of contamination as well as being time intensive and high in cost. Also this method required extra cycles of PCR, totalling the number of PCR cycles to 55 which is not ideal and introduces artefacts. Therefore a new method was developed to incorporate the NGS adaptors into the original PCR; targeted amplicon library creation (TALC). TALC was shown to produce equal results to the adaptor ligation method and reduced the preparation time and costs significantly.

5.1.4 Mutations in histologically normal mucosa

KRAS mutations in normal mucosa have been reported in both carcinoma-associated and also non-carcinoma associated normal mucosa in multiple studies (Ronai, 1994, Yamada et al., 2005, Parsons et al., 2010). Therefore the approach taken was firstly to investigate if this could be replicated. Through NGS sequencing of KRAS codon 12 & 13 amplicons run in duplicate, mutations could be detected in 11 out of 38 (29%) samples of carcinoma associated normal mucosa from patients with KRAS mutated tumours. The mutations detected in the normal mucosa had allele frequencies of between 1% and 9%, confirming that a highly sensitive technique is required to detect these low level mutations.

In a separate cohort, KRAS mutations were detected in 41% of carcinomas and 8% of the associated normal samples. There were no mutations found from normal mucosa where the tumour did not also contain a KRAS 12 & 13 mutation. No KRAS mutations were detected in the adenoma associated normal mucosa or the normal mucosa from patients with normal colonoscopies. The finding of RAS mutations in normal mucosa being restricted to cases with RAS mutations in the cancer might suggest that extreme carcinogens are inducing such mutations or that these patients have an inherent risk of suffering these mutations for an unknown reason.

The bioinformatic approach to analysing data for KRAS and other oncogene hotspots is relatively straightforward in comparison to interrogating large tumour suppressor genes where a mutation could occur at any point. For a cohort of carcinoma, adenoma and their associated normals as well as normal mucosa from patients with normal colonoscopies, mutations in the oncogenes were only detected in carcinoma associated normal mucosa and not mucosa from adenoma patients or normal mucosa unassociated with neoplasia. It may be that techniques are still not sufficiently sensitive or that they do not exist. At present searching for these mutations does not seem to be valuable.

There were KRAS mutations in 11% of the carcinoma associated normal mucosa and 31% contained a mutation in the tumour. This equates to 35% of mutated tumours also containing a KRAS mutation in their normal mucosa. These three separate experiments have confirmed the previously studies' findings that KRAS

codon 12 and 13 mutations are present in cancer associated normal mucosa and that these mutations occur at a low level. The finding of KRAS mutations in normal that in most cases differ to the tumour supports the cancer field theory. If the whole bowel is exposed to a carcinogen that induces a base change in KRAS, different bases may be substituted within different areas of the bowel. If an affected area then develops further mutations to give it a growth advantage, a tumour may develop. However the differing mutations in KRAS at other parts of the bowel unaffected by neoplasia show that the background mucosa is genetically abnormal. It would be valuable to widen the search for mutations in normal mucosa by looking at other genes.

It was seen that DNA that had been fixed in formalin had an increased number of G>A and C>T changes once amplified by PCR compared to fresh DNA. This is a known phenomenon and these artefacts could be removed by treating gDNA with UDG to remove the formalin effects of amination and depurination of guanine and cytosine residues.

The interpretation of tumour suppressor gene data from the samples of normal mucosa from patients with normal colonoscopies was not reliable due to the unavailability of blood samples to remove SNPs from the analysis. For the cohort of carcinomas and adenomas and their associated normals, SNPs could be filtered out, but the error rate from PCR amplification of formalin was too high to allow for reliable interpretation of tumour suppressor gene data as these samples were not tested in duplicate. It would therefore be useful to repeat this experiment with UDG pre-treatment of DNA to remove formalin induced errors and in duplicate or triplicate to remove PCR errors. This could be done by duplicate processing of samples to produce libraries and multiplexing them on a single NGS run.

From the cohort of carcinoma and carcinoma associated normal that were run in duplicate, as well as providing useful insights into the amount of PCR error generated it revealed mutations in normal mucosa in genes other than KRAS. Mutations were also seen in APC, SMAD4 and CTNNB1. This further shows that mutations occur in normal colonic mucosa as well as the cancer. This could be due to the bowel having a longer exposure to carcinogens and the patients with

carcinoma being on average 4.5 years older than the adenoma cohort (the median age of carcinoma cohort was 71years and 65.5 for the adenoma cohort).

This is not a longitudinal study and is only able to provide a snap shot of the mutational landscape of carcinoma associated normal mucosa. To address this question of how mutational damage occurs over time it would be useful to monitor the mutations present and see if they change over time as this would give a picture of population of mutant clones present at different times. This could possibly be enabled by taking samples of normal mucosa during flexisigmoidoscopy as part of the bowel cancer screening programme, as patients are monitored periodically and follow-up data would be available (Logan et al., 2012).

It is hypothesised that mutations within the normal mucosa are induced by the carcinogenic metabolic products of the bacteria that breakdown faecal material (Guarner and Malagelada, 2003). It has already been observed that fusobacterium is associated not only with carcinoma but also adenoma (McCoy et al., 2013). Therefore it would be interesting to classify the gut flora present within patients with mutations in normal mucosa compared to those without mutations to try and identify the most mutation-inducing bacteria. In this way patients could then be screened for their gut flora composition and those at higher risk of developing cancer could be identified and have their gut flora altered as a preventative measure for colorectal cancer.

5.1.5 FAP adenomas

In order to gain insight of how adenomas develop in the bowel, FAP patients were used as a model to study the molecular profiles of adenomas. The mutational and copy number profiles of the adenomas showed that clustering can occur locally within the bowel and sometimes at greater distances i.e. some adenomas within the ascending colon and transverse colon share molecular lesions. Conversely, in many cases, adenomas also appeared to develop independently. However, this is only according to their mutations in the genes tested and their copy number aberrations and it is highly likely that other mechanisms are driving adenoma development. It would therefore be useful to repeat this not only for more FAP patients but also looking at other molecular aberrations. This could include methylation, long non-coding RNA and micro RNA expression as well as using exome sequencing instead of targeted. By comparing all patients, it appeared that different mechanisms were driving the development of adenomas in different patients. By repeating with larger number of patients, this pattern may become clearer.

The copy number data from these adenomas is more reliable due to fewer chances of errors and artefacts occurring compared to PCR amplification of targets and mutation analysis. The copy number profiled revealed that adenomas within the bowel share both small and large chromosomal aberrations. The smaller aberrations in particular which are unique to that patient e.g. the pattern of loss of seen in chromosome 11 in patient 3 provide evidence that these adenomas have a clonal origin and then have developed further, independent aberrations.

5.2 Final conclusions

- Mechanical dissociation of colonic crypts in the presence of a chelating agent is an effective method of isolating colon epithelial cells from fresh tissue.
- The sensitivity of pyrosequencing varies depending on the type of base change being detected and lies between 5% and 25% mutant allele frequency.
- The sensitivity of NGS is 0.5% for fresh tissue and 1% for FFPE.
- PCR-based enrichment techniques such as RFLP are not reliable with NGS as they enrich a high proportion of error.
- Formalin induces G>A and C>T base changes in DNA once amplified and these can be removed bioinformatically by testing samples in duplicate.
- KRAS mutations are present at low level in carcinoma associated normal mucosa.
- Mutations were also found in APC, CTNNB1, PIK3CA and SMAD4 in carcinoma associated normal mucosa using the Fluidigm Access Array.
- Further validation is required of the Fluidigm Access Array to be certain of FAP adenoma mutation status by testing samples in duplicate.
- The most commonly seen chromosomal aberrations in FAP adenomas were gains of chromosomes 7,8 and 13 which is similar to reported aberrations in sporadic adenoma.
- FAP adenomas share marker aberrations indicating that some have a clonal origin.
- Wide variation in chromosomal aberrations was seen on top of clonal variations suggesting further independent evolution of adenomas.

References

AKHOONDI, S., SUN, D., VON DER LEHR, N., APOSTOLIDOU, S., KLOTZ, K., MALJUKOVA, A., CEPEDA, D., FIEGL, H., DAFOU, D., MARTH, C., MUELLER-HOLZNER, E., CORCORAN, M., DAGNELL, M., NEJAD, S. Z., NAYER, B. N., ZALI, M. R., HANSSON, J., EGYHAZI, S., PETERSSON, F., SANGFELT, P., NORDGREN, H., GRANDER, D., REED, S. I., WIDSCHWENDTER, M., SANGFELT, O. & SPRUCK, C. 2007. FBXW7/hCDC4 is a general tumor suppressor in human cancer. *Cancer research,* 67**,** 9006-12.

ALHOPURO, P., SAMMALKORPI, H., NIITTYMÄKI, I., BISTRÖM, M., RAITILA, A., SAHARINEN, J., NOUSIAINEN, K., LEHTONEN, H. J., HELIÖVAARA, E. & PUHAKKA, J. 2012. Candidate driver genes in microsatellite-unstable colorectal cancer. *International Journal of Cancer,* 130**,** 1558-1566.

ANASTAS, J. N. & MOON, R. T. 2012. WNT signalling pathways as therapeutic targets in cancer. *Nature reviews Cancer,* 13**,** 11-26.

ANDERSON, J., SWEDE, H., RUSTAGI, T., PROTIVA, P., PLEAU, D., BRENNER, B., RAJAN, T., HEINEN, C., LEVINE, J. & ROSENBERG, D. 2012. Aberrant crypt foci as predictors of colorectal neoplasia on repeat colonoscopy. *Cancer Causes & Control,* 23**,** 355-361.

ARADHYA, S., CHERRY, A. M. & GIRIRAJAN, S. 2013. Counting Chromosomes to Exons: Advances in Copy Number Detection. *Current Genetic Medicine Reports***,** 1-10.

ASHKTORAB, H., SCHAFFER, A. A., DAREMIPOURAN, M., SMOOT, D. T., LEE, E. & BRIM, H. 2010. Distinct genetic alterations in colorectal cancer. *PloS one,* 5**,** e8879.

ATKIN, W. S., EDWARDS, R., KRALJ-HANS, I., WOOLDRAGE, K., HART, A. R., NORTHOVER, J. M. A., PARKIN, D. M., WARDLE, J., DUFFY, S. W. & CUZICK, J. 2010. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *The Lancet,* 375**,** 1624-1633.

BAKER, A. M., GRAHAM, T. A. & WRIGHT, N. A. 2013. Pre-tumour clones, periodic selection and clonal interference in the origin and progression of gastrointestinal cancer: potential for biomarker development. *The Journal of pathology,* 229**,** 502-14.

BAMFORD, S., DAWSON, E., FORBES, S., CLEMENTS, J., PETTETT, R., DOGAN, A., FLANAGAN, A., TEAGUE, J., FUTREAL, P. A., STRATTON, M. R. & WOOSTER, R. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British journal of cancer,* 91**,** 355-8.

BASS, A. J., LAWRENCE, M. S., BRACE, L. E., RAMOS, A. H., DRIER, Y., CIBULSKIS, K., SOUGNEZ, C., VOET, D., SAKSENA, G., SIVACHENKO, A., JING, R., PARKIN, M., PUGH, T., VERHAAK, R. G., STRANSKY, N., BOUTIN, A. T., BARRETINA, J., SOLIT, D. B., VAKIANI, E., SHAO, W., MISHINA, Y., WARMUTH, M., JIMENEZ, J., CHIANG, D. Y., SIGNORETTI, S., KAELIN, W. G., SPARDY, N., HAHN, W. C., HOSHIDA, Y., OGINO, S., DEPINHO, R. A., CHIN, L., GARRAWAY, L. A., FUCHS, C. S., BASELGA, J., TABERNERO, J., GABRIEL, S., LANDER, E. S., GETZ, G. & MEYERSON, M. 2011. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nature genetics,* 43**,** 964-968.

BOLAND, C. R. & GOEL, A. 2010. Microsatellite Instability in Colorectal Cancer. *Gastroenterology,* 138**,** 2073-2087.e3.

BOSMAN, F. T., WORLD HEALTH ORGANIZATION. & INTERNATIONAL AGENCY FOR RESEARCH ON CANCER. 2010. *WHO classification of tumours of the digestive system,* Lyon, International Agency for Research on Cancer.

BOZIC, I., ANTAL, T., OHTSUKI, H., CARTER, H., KIM, D., CHEN, S., KARCHIN, R., KINZLER, K. W., VOGELSTEIN, B. & NOWAK, M. A. 2010. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences,* 107**,** 18545-18550.

BURRELL, R. A., MCCLELLAND, S. E., ENDESFELDER, D., GROTH, P., WELLER, M.-C., SHAIKH, N., DOMINGO, E., KANU, N., DEWHURST, S. M., GRONROOS, E., CHEW, S. K., ROWAN, A. J., SCHENK, A., SHEFFER, M., HOWELL, M., KSCHISCHO, M., BEHRENS, A., HELLEDAY, T., BARTEK, J., TOMLINSON, I. P. & SWANTON, C. 2013. Replication stress links structural and numerical cancer chromosomal instability. *Nature,* 494**,** 492-496.

CANCER RESEARCH UK. 2010. *Bowel cancer statistics - Key Facts* [Online]. London. Available: http://info.cancerresearchuk.org/cancerstats/types/bowel/?script=true [Accessed 25/10/2010 2010].

CANCER RESEARCH UK. 2011. *Bowel (colorectal) cancer - UK incidence statistics* [Online]. London. Available: http://info.cancerresearchuk.org/cancerstats/types/bowel/incidence/ [Accessed 01/06/2011 2011].

CANCER RESEARCH UK 2012a. Bowel cancer incidence statistics.

CANCER RESEARCH UK. 2012b. *Bowel cancer survival statistics* [Online]. Available: http://www.cancerresearchuk.org/cancer-info/cancerstats/types/bowel/survival/ [Accessed 05 August 2013].

CHAMBERS, P. A., STEAD, L. F., MORGAN, J. E., CARR, I. M., SUTTON, K. M., WATSON, C. M., CROWE, V., DICKINSON, H., ROBERTS, P., MULATERO, C., SEYMOUR, M., MARKHAM, A. F., WARING, P. M., QUIRKE, P. & TAYLOR, G. R. 2013. Mutation Detection by Clonal Sequencing of PCR Amplicons and Grouped Read Typing is Applicable to Clinical Diagnostics. *Human mutation,* 34**,** 248-254.

CHAN, T. L., ZHAO, W., LEUNG, S. Y. & YUEN, S. T. 2003. BRAF and KRAS mutations in colorectal hyperplastic polyps and serrated adenomas. *Cancer research,* 63**,** 4878-4881.

CHEN, X., LISTMAN, J. B., SLACK, F. J., GELERNTER, J. & ZHAO, H. 2012. Biases and errors on allele frequency estimation and disease association tests of next-generation sequencing of pooled samples. *Genetic epidemiology,* 36**,** 549-60.

CHENG, H., BJERKNES, M. & AMAR, J. 1984. Methods for the determination of epithelial cell kinetic parameters of human colonic epithelium isolated from surgical and biopsy specimens. *Gastroenterology,* 86**,** 78-85.

CRABTREE, M., SIEBER, O. M., LIPTON, L., HODGSON, S. V., LAMLUM, H., THOMAS, H. J., NEALE, K., PHILLIPS, R. K., HEINIMANN, K. & TOMLINSON, I. P. 2003. Refining the relation between 'first hits' and 'second hits' at the APC locus: the 'loose fit'model and evidence for

differences in somatic mutation spectra among patients. *Oncogene,* 22**,** 4257-4265.

CUNNINGHAM, D., ATKIN, W., LENZ, H. J., LYNCH, H. T., MINSKY, B., NORDLINGER, B. & STARLING, N. 2010. Colorectal cancer. *Lancet,* 375**,** 1030-47.

CUNNINGHAM, K. S. & RIDDELL, R. H. 2006. Serrated mucosal lesions of the colorectum. *Current opinion in gastroenterology,* 22**,** 48-53.

DAVIDSON, C. J., ZERINGER, E., CHAMPION, K. J., GAUTHIER, M. P., WANG, F., BOONYARATANAKORNKIT, J., JONES, J. R. & SCHREIBER, E. 2012. Improving the limit of detection for Sanger sequencing: A comparison of methodologies for KRAS variant detection. *Biotechniques,* 2012.

DE LA CHAPELLE, A. & HAMPEL, H. 2010. Clinical relevance of microsatellite instability in colorectal cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 28**,** 3380-7.

DIETERLE, C. P., CONZELMANN, M., LINNEMANN, U. & BERGER, M. R. 2004. Detection of isolated tumor cells by polymerase chain reaction-restriction fragment length polymorphism for K-ras mutations in tissue samples of 199 colorectal cancer patients. *Clinical cancer research : an official journal of the American Association for Cancer Research,* 10**,** 641-50.

DIXON, M. F. 2002. Gastrointestinal epithelial neoplasia: Vienna revisited. *Gut,* 51**,** 130-1.

DO, H., WONG, S. Q., LI, J. & DOBROVIC, A. 2013. Reducing Sequence Artifacts in Amplicon-Based Massively Parallel Sequencing of Formalin-Fixed Paraffin-Embedded DNA by Enzymatic Depletion of Uracil-Containing Templates. *Clinical chemistry*.

DUNLOP, M. G., FARRINGTON, S. M., CAROTHERS, A. D., WYLLIE, A. H., SHARP, L., BURN, J., LIU, B., KINZLER, K. W. & VOGELSTEIN, B. 1997. Cancer risk associated with germline DNA mismatch repair gene mutations. *Human molecular genetics,* 6**,** 105-10.

ENCYCLOPÆDIA BRITANNICA ONLINE. 2003. *large intestine: mucosa and musculature in humans* [Online]. Available: http://www.britannica.com/EBchecked/topic/330544/large-intestine [Accessed 01/06/2011 2011].

ESPINA, V., WULFKUHLE, J. D., CALVERT, V. S., VANMETER, A., ZHOU, W., COUKOS, G., GEHO, D. H., PETRICOIN, E. F. & LIOTTA, L. A. 2006. Laser-capture microdissection. *Nat. Protocols,* 1**,** 586-603.

EWING, B. & GREEN, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research,* 8**,** 186-94.

FLAHERTY, P., NATSOULIS, G., MURALIDHARAN, O., WINTERS, M., BUENROSTRO, J., BELL, J., BROWN, S., HOLODNIY, M., ZHANG, N. & JI, H. P. 2012. Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic acids research,* 40**,** e2.

FLORA, M., PIANA, S., BASSANO, C., BISAGNI, A., DE MARCO, L., CIARROCCHI, A., TAGLIAVINI, E., GARDINI, G., TAMAGNINI, I. & BANZI, C. 2012. Epidermal growth factor receptor (< i> EGFR</i>) gene copy number in colorectal adenoma-carcinoma progression. *Cancer genetics*.

GALANDIUK, S., RODRIGUEZ–JUSTO, M., JEFFERY, R., NICHOLSON, A. M., CHENG, Y., OUKRIF, D., ELIA, G., LEEDHAM, S. J., MCDONALD, S. A. &

WRIGHT, N. A. 2012. Field cancerization in the intestinal epithelium of patients with Crohn's ileocolitis. *Gastroenterology,* 142**,** 855-864. e8.

GALIATSATOS, P. & FOULKES, W. D. 2006. Familial Adenomatous Polyposis. *Am J Gastroenterol,* 101**,** 385-398.

GLOBOCAN. 2008. *Fast Stats* [Online]. Lyon: International Agency for Research on Cancer. Available: http://globocan.iarc.fr/factsheets/populations/factsheet.asp?uno=900 [Accessed 01/06/2011 2011].

GOODLAD, R. A., LEVI, S., LEE, C. Y., MANDIR, N., HODGSON, H. & WRIGHT, N. A. 1991. Morphometry and cell proliferation in endoscopic biopsies: evaluation of a technique. *Gastroenterology,* 101**,** 1235-41.

GORDON, D. J., RESIO, B. & PELLMAN, D. 2012. Causes and consequences of aneuploidy in cancer. *Nature reviews. Genetics,* 13**,** 189-203.

GRADY, W. M. & CARETHERS, J. M. 2008. Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology,* 135**,** 1079-99.

GRAHAM, T. A., HUMPHRIES, A., SANDERS, T., RODRIGUEZ-JUSTO, M., TADROUS, P. J., PRESTON, S. L., NOVELLI, M. R., LEEDHAM, S. J., MCDONALD, S. A. & WRIGHT, N. A. 2011a. Use of methylation patterns to determine expansion of stem cell clones in human colon tissue. *Gastroenterology,* 140**,** 1241-1250 e1-9.

GRAHAM, T. A., MCDONALD, S. A. & WRIGHT, N. A. 2011b. Field cancerization in the GI tract. *Future oncology,* 7**,** 981-93.

GREAVES, M. & MALEY, C. C. 2012. Clonal evolution in cancer. *Nature,* 481**,** 306-313.

GUARNER, F. & MALAGELADA, J.-R. 2003. Gut flora in health and disease. *The Lancet,* 361**,** 512-519.

GUSNANTO, A., WOOD, H. M., PAWITAN, Y., RABBITTS, P. & BERRI, S. 2012. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics,* 28**,** 40-47.

GUTIERREZ-GONZALEZ, L., DEHERAGODA, M., ELIA, G., LEEDHAM, S. J., SHANKAR, A., IMBER, C., JANKOWSKI, J. A., TURNBULL, D. M., NOVELLI, M., WRIGHT, N. A. & MCDONALD, S. A. 2009. Analysis of the clonal architecture of the human small intestinal epithelium establishes a common stem cell for all lineages and reveals a mechanism for the fixation and spread of mutations. *The Journal of pathology,* 217**,** 489-96.

HA, P. K. & CALIFANO, J. A. 2003. The molecular biology of mucosal field cancerization of the head and neck. *Crit Rev Oral Biol Med,* 14**,** 363-9.

HADD, A. G., HOUGHTON, J., CHOUDHARY, A., SAH, S., CHEN, L., MARKO, A. C., SANFORD, T., BUDDAVARAPU, K., KROSTING, J., GARMIRE, L., WYLIE, D., SHINDE, R., BEAUDENON, S., ALEXANDER, E. K., MAMBO, E., ADAI, A. T. & LATHAM, G. J. 2013. Targeted, High-Depth, Next-Generation Sequencing of Cancer Genes in Formalin-Fixed, Paraffin-Embedded and Fine-Needle Aspiration Tumor Specimens. *The Journal of Molecular Diagnostics,* 15**,** 234-247.

HALBRITTER, J., DIAZ, K., CHAKI, M., PORATH, J. D., TARRIER, B., FU, C., INNIS, J. L., ALLEN, S. J., LYONS, R. H., STEFANIDIS, C. J., OMRAN, H., SOLIMAN, N. A. & OTTO, E. A. 2012. High-throughput mutation analysis in

patients with a nephronophthisis-associated ciliopathy applying multiplexed barcoded array-based PCR amplification and next-generation sequencing. *Journal of medical genetics,* 49**,** 756-67.

HALF, E., BERCOVICH, D. & ROZEN, P. 2009. Familial adenomatous polyposis. *Orphanet Journal of Rare Diseases,* 4**,** 22.

HALIASSOS, A., CHOMEL, J. C., GRANDJOUAN, S., KRUH, J., KAPLAN, J. C. & KITZIS, A. 1989. Detection of minority point mutations by modified PCR technique: a new approach for a sensitive diagnosis of tumor-progression markers. *Nucleic acids research,* 17**,** 8093-9.

HAYES, J., TZIKA, A., THYGESEN, H., BERRI, S., WOOD, H., HEWITT, S., PENDLEBURY, M., COATES, A., WILLOUGHBY, L. & WATSON, C. 2013. Diagnosis of Copy Number Variation by Illumina Next Generation Sequencing is comparable in performance to Oligonucleotide Array Comparative Genomic Hybridisation. *Genomics*.

HEAPHY, C., GRIFFITH, J. & BISOFFI, M. 2009. Mammary field cancerization: molecular evidence and clinical importance. *Breast cancer research and treatment,* 118**,** 229-239.

HORNER, D. S., PAVESI, G., CASTRIGNANO, T., DE MEO, P. D., LIUNI, S., SAMMETH, M., PICARDI, E. & PESOLE, G. 2010. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in bioinformatics,* 11**,** 181-97.

HUMPHRIES, A., GRAHAM, T., MCDONALD, S. & WRIGHT, N. 2011. Altered stem cell dynamics in human colon adenoma crypts allow rapid expansion and fixation of mutations during clonal expansion. *Gut,* 60**,** A52-A53.

HUMPHRIES, A. & WRIGHT, N. A. 2008. Colonic crypt organization and tumorigenesis. *Nature reviews. Cancer,* 8**,** 415-24.

HUNYADY, B., MEZEY, E. & PALKOVITS, M. 2000. Gastrointestinal immunology: cell types in the lamina propria--a morphological review. *Acta physiologica Hungarica,* 87**,** 305-28.

ISSA, J. P. 2004. CpG island methylator phenotype in cancer. *Nature reviews Cancer,* 4**,** 988-93.

JANSSEN, A. & MEDEMA, R. H. 2013. Cancer: Stress mixes chromosomes. *Nature,* 494**,** 439-441.

JONES, A., THIRLWELL, C., HOWARTH, K., GRAHAM, T., CHAMBERS, W., SEGDITSAS, S., PAGE, K., PHILLIPS, R., THOMAS, H. & SIEBER, O. 2007. Analysis of copy number changes suggests chromosomal instability in a minority of large colorectal adenomas. *The Journal of pathology,* 213**,** 249-256.

JONSSON, C., STAL, P., SJOQVIST, U., AKERLUND, J. E., LOFBERG, R. & MOLLER, L. 2010. DNA adducts in normal colonic mucosa from healthy controls and patients with colon polyps and colorectal carcinomas. *Mutagenesis,* 25**,** 499-504.

KAWASAKI, T., OHNISHI, M., NOSHO, K., SUEMOTO, Y., KIRKNER, G. J., MEYERHARDT, J. A., FUCHS, C. S. & OGINO, S. 2008. CpG island methylator phenotype-low (CIMP-low) colorectal cancer shows not only few methylated CIMP-high-specific CpG islands, but also low-level methylation at individual loci. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc,* 21**,** 245-55.

KERICK, M., ISAU, M., TIMMERMANN, B., SULTMANN, H., HERWIG, R., KROBITSCH, S., SCHAEFER, G., VERDORFER, I., BARTSCH, G., KLOCKER, H., LEHRACH, H. & 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 medical genomics,* 4**,** 68.

KIM, M. S., LEE, J. & SIDRANSKY, D. 2010. DNA methylation markers in colorectal cancer. *Cancer metastasis reviews,* 29**,** 181-206.

KINDE, I., WU, J., PAPADOPOULOS, N., KINZLER, K. W. & VOGELSTEIN, B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America,* 108**,** 9530-5.

KING, B., TRIMARCHI, T., REAVIE, L., XU, L., MULLENDERS, J., NTZIACHRISTOS, P., ARANDA-ORGILLES, B., PEREZ-GARCIA, A., SHI, J., VAKOC, C., SANDY, P., SHEN, S. S., FERRANDO, A. & AIFANTIS, I. 2013. The Ubiquitin Ligase FBXW7 Modulates Leukemia-Initiating Cell Activity by Regulating MYC Stability. *Cell,* 153**,** 1552-66.

KOBOLDT, D. C., ZHANG, Q., LARSON, D. E., SHEN, D., MCLELLAN, M. D., LIN, L., MILLER, C. A., MARDIS, E. R., DING, L. & WILSON, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research,* 22**,** 568-76.

KRAUS, M. C., SEELIG, M. H., LINNEMANN, U. & BERGER, M. R. 2006. The balanced induction of K-ras codon 12 and 13 mutations in mucosa differs from their ratio in neoplastic tissues. *International journal of oncology,* 29**,** 957-64.

LABELLE, M., BEGUM, S. & HYNES, R. O. 2011. Direct signaling between platelets and cancer cells induces an epithelial-mesenchymal-like transition and promotes metastasis. *Cancer Cell,* 20**,** 576-590.

LAMLUM, H., ILYAS, M., ROWAN, A., CLARK, S., JOHNSON, V., BELL, J., FRAYLING, I., EFSTATHIOU, J., PACK, K. & PAYNE, S. 1999. The type of somatic mutation at APC in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's' two-hit'hypothesis. *Nature medicine,* 5**,** 1071-1075.

LANZA, G., MESSERINI, L., GAFÀ, R. & RISIO, M. 2011. Colorectal tumors: The histology report. *Digestive and Liver Disease,* 43, Supplement 4**,** S344-S355.

LASH, R. H., GENTA, R. M. & SCHULER, C. M. 2010. Sessile serrated adenomas: prevalence of dysplasia and carcinoma in 2139 patients. *J Clin Pathol,* 63**,** 681-6.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LI, H., RUAN, J. & DURBIN, R. 2008a. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res,* 18**,** 1851-8.

LI, J., WANG, L., MAMON, H., KULKE, M. H., BERBECO, R. & MAKRIGIORGOS, G. M. 2008b. Replacing PCR with COLD-PCR enriches variant DNA

sequences and redefines the sensitivity of genetic testing. *Nat Med,* 14**,** 579-584.

LIAO, X., LOCHHEAD, P., NISHIHARA, R., MORIKAWA, T., KUCHIBA, A., YAMAUCHI, M., IMAMURA, Y., QIAN, Z. R., BABA, Y. & SHIMA, K. 2012. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. *New England Journal of Medicine,* 367**,** 1596-1606.

LIU, J., WALKER, N. M., COOK, M. T., OOTANI, A. & CLARKE, L. L. 2012. Functional Cftr in crypt epithelium of organotypic enteroid cultures from murine small intestine. *American Journal of Physiology-Cell Physiology,* 302**,** C1492-C1503.

LOGAN, R. F., PATNICK, J., NICKERSON, C., COLEMAN, L., RUTTER, M. D. & VON WAGNER, C. 2012. Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. *Gut,* 61**,** 1439-1446.

MAMANOVA, L., COFFEY, A. J., SCOTT, C. E., KOZAREWA, I., TURNER, E. H., KUMAR, A., HOWARD, E., SHENDURE, J. & TURNER, D. J. 2010. Target-enrichment strategies for next-generation sequencing. *Nature methods,* 7**,** 111-8.

MANDEL, J. S., BOND, J. H., CHURCH, T. R., SNOVER, D. C., BRADLEY, G. M., SCHUMAN, L. M. & EDERER, F. 1993. Reducing Mortality from Colorectal Cancer by Screening for Fecal Occult Blood. *New England Journal of Medicine,* 328**,** 1365-1371.

MARDIS, E. R. 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet,* 9**,** 387-402.

MARKOWITZ, S. D. & BERTAGNOLLI, M. M. 2009. Molecular origins of cancer: Molecular basis of colorectal cancer. *N Engl J Med,* 361**,** 2449-60.

MCCOY, A. N., ARAUJO-PEREZ, F., AZCARATE-PERIL, A., YEH, J. J., SANDLER, R. S. & KEKU, T. O. 2013. Fusobacterium is associated with colorectal adenomas. *PloS one,* 8**,** e53653.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research,* 20**,** 1297-303.

MEIJER, G. A., HERMSEN, M., BAAK, J., VAN DIEST, P. J., MEUWISSEN, S., BELIEN, J., HOOVERS, J., JOENJE, H., SNIJDERS, P. & WALBOOMERS, J. 1998. Progression from colorectal adenoma to carcinoma is associated with non-random chromosomal gains as detected by comparative genomic hybridisation. *Journal of clinical pathology,* 51**,** 901-909.

MESTERI, I., BAYER, G., MEYER, J., CAPPER, D., SCHOPPMANN, S. F., VON DEIMLING, A. & BIRNER, P. 2013. Improved molecular classification of serrated lesions of the colon by immunohistochemical detection of BRAF V600E. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*.

METZKER, M. L. 2010. Sequencing technologies - the next generation. *Nature reviews. Genetics,* 11**,** 31-46.

MEYERSON, M., GABRIEL, S. & GETZ, G. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature reviews. Genetics,* 11**,** 685-96.

MILBURY, C. A., CORRELL, M., QUACKENBUSH, J., RUBIO, R. & MAKRIGIORGOS, G. M. 2012. COLD-PCR enrichment of rare cancer mutations prior to targeted amplicon resequencing. *Clinical chemistry,* 58**,** 580-9.

MILBURY, C. A., LI, J. & MAKRIGIORGOS, G. M. 2011. Ice-COLD-PCR enables rapid amplification and robust enrichment for low-abundance unknown DNA mutations. *Nucleic acids research,* 39**,** e2.

MINAMOTO, T., YAMASHITA, N., OCHIAI, A., MAI, M., SUGIMURA, T., RONAI, Z. & ESUMI, H. 1995. Mutant K-ras in apparently normal mucosa of colorectal cancer patients. Its potential as a biomarker of colorectal tumorigenesis. *Cancer,* 75**,** 1520-6.

MIYAKI, M. & KUROKI, T. 2003. Role of Smad4 (DPC4) inactivation in human cancer. *Biochemical and biophysical research communications,* 306**,** 799-804.

MOCH, H., BLANK, P. R., DIETEL, M., ELMBERGER, G., KERR, K. M., PALACIOS, J., PENAULT-LLORCA, F., ROSSI, G. & SZUCS, T. D. 2012. Personalized cancer medicine and the future of pathology. *Virchows Archiv : an international journal of pathology,* 460**,** 3-8.

MORAN, A. O., P. DE JUAN, C. FERNANDEZ-MARCELO, T. FRIAS, C. SANCHEZ-PERNAUTE, A. TORRES, A. J. DIAZ-RUBIO, E. INIESTA, P. BENITO, M. 2010. Differential colorectal carcinogenesis: Molecular basis and clinical relevance. *World J Gastroenterol,* 2**,** 151-158.

MORGAN, J. E., CARR, I. M., SHERIDAN, E., CHU, C. E., HAYWARD, B., CAMM, N., LINDSAY, H. A., MATTOCKS, C. J., MARKHAM, A. F., BONTHRON, D. T. & TAYLOR, G. R. 2010. Genetic diagnosis of familial breast cancer using clonal sequencing. *Human mutation,* 31**,** 484-91.

NEWTON, C. R., GRAHAM, A., HEPTINSTALL, L. E., POWELL, S. J., SUMMERS, C., KALSHEKER, N., SMITH, J. C. & MARKHAM, A. F. 1989. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic acids research,* 17**,** 2503-16.

NOFFSINGER, A. E. 2009. Serrated polyps and colorectal cancer: new pathway to malignancy. *Annual review of pathology,* 4**,** 343-64.

OBRADOR-HEVIA, A., CHIN, S.-F., GONZÁLEZ, S., REES, J., VILARDELL, F., GREENSON, J. K., CORDERO, D., MORENO, V., CALDAS, C. & CAPELLÁ, G. 2010. Oncogenic KRAS is not necessary for Wnt signalling activation in APC-associated FAP adenomas. *The Journal of pathology,* 221**,** 57-67.

OGINO, S. & GOEL, A. 2008. Molecular classification and correlates in colorectal cancer. *J Mol Diagn,* 10**,** 13-27.

OGINO, S., KAWASAKI, T., BRAHMANDAM, M., YAN, L., CANTOR, M., NAMGYAL, C., MINO-KENUDSON, M., LAUWERS, G. Y., LODA, M. & FUCHS, C. S. 2005. Sensitive sequencing method for KRAS mutation detection by Pyrosequencing. *J Mol Diagn,* 7**,** 413-21.

ORLANDO, F. A., TAN, D., BALTODANO, J. D., KHOURY, T., GIBBS, J. F., HASSID, V. J., AHMED, B. H. & ALRAWI, S. J. 2008. Aberrant crypt foci as

precursors in colorectal cancer progression. *Journal of surgical oncology,* 98**,** 207-13.

PARK, D. I., RYU, S., KIM, Y.-H., LEE, S.-H., LEE, C. K., EUN, C. S. & HAN, D. S. 2010. Comparison of guaiac-based and quantitative immunochemical fecal occult blood testing in a population at average risk undergoing colorectal cancer screening. *The American journal of gastroenterology,* 105**,** 2017-2025.

PARSONS, B. L., MARCHANT-MIROS, K. E., DELONGCHAMP, R. R., VERKLER, T. L., PATTERSON, T. A., MCKINZIE, P. B. & KIM, L. T. 2010. ACB-PCR quantification of K-RAS codon 12 GAT and GTT mutant fraction in colon tumor and non-tumor tissue. *Cancer investigation,* 28**,** 364-75.

PEETERS, M., OLINER, K. S., PARKER, A., SIENA, S., VAN CUTSEM, E., HUANG, J., HUMBLET, Y., VAN LAETHEM, J. L., ANDRE, T., WIEZOREK, J., REESE, D. & PATTERSON, S. D. 2013. Massively Parallel Tumor Multigene Sequencing to Evaluate Response to Panitumumab in a Randomized Phase III Study of Metastatic Colorectal Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*.

PINO, M. S. & CHUNG, D. C. 2010. The chromosomal instability pathway in colon cancer. *Gastroenterology,* 138**,** 2059-72.

POULOGIANNIS, G., ICHIMURA, K., HAMOUDI, R. A., LUO, F., LEUNG, S. Y., YUEN, S. T., HARRISON, D. J., WYLLIE, A. H. & ARENDS, M. J. 2010. Prognostic relevance of DNA copy number changes in colorectal cancer. *The Journal of pathology,* 220**,** 338-347.

QUAIL, M. A., SMITH, M., COUPLAND, P., OTTO, T. D., HARRIS, S. R., CONNOR, T. R., BERTONI, A., SWERDLOW, H. P. & GU, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13**,** 341.

QUINTERO, E., CASTELLS, A., BUJANDA, L., CUBIELLA, J., SALAS, D., LANAS, Á., ANDREU, M., CARBALLO, F., MORILLAS, J. D. & HERNÁNDEZ, C. 2012. Colonoscopy versus fecal immunochemical testing in colorectal-cancer screening. *New England Journal of Medicine,* 366**,** 697-706.

QUIRKE, P., RISIO, M., LAMBERT, R., VON KARSA, L. & VIETH, M. 2011. Quality assurance in pathology in colorectal cancer screening and diagnosis—European recommendations. *Virchows Archiv,* 458**,** 1-19.

RAJAGOPALAN, H., NOWAK, M. A., VOGELSTEIN, B. & LENGAUER, C. 2003. The significance of unstable chromosomes in colorectal cancer. *Nature reviews. Cancer,* 3**,** 695-701.

REDSTON, M. 2001. Carcinogenesis in the GI Tract: From Morphology to Genetics and Back Again. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc,* 14**,** 236-245.

REX, D. K., AHNEN, D. J., BARON, J. A., BATTS, K. P., BURKE, C. A., BURT, R. W., GOLDBLUM, J. R., GUILLEM, J. G., KAHI, C. J. & KALADY, M. F. 2012. Serrated lesions of the colorectum: review and recommendations from an expert panel. *The American journal of gastroenterology,* 107**,** 1315-1329.

ROBERTS, N. D., KORTSCHAK, R. D., PARKER, W. T., SCHREIBER, A. W., BRANFORD, S., SCOTT, H. S., GLONEK, G. & ADELSON, D. L. 2013. A

comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics*.

RONAGHI, M. 2001. Pyrosequencing sheds light on DNA sequencing. *Genome research,* 11**,** 3-11.

RONAI, Z. & MINAMOTO, T. 1997. Quantitative enriched PCR (QEPCR), a highly sensitive method for detection of K-ras oncogene mutation. *Human mutation,* 10**,** 322-5.

RONAI, Z. L., FENG C.; GRADIA, SCOTT; HART, WENDY J.; BUTLER, ROSS 1994. Detection of K-ras mutation in normal and malignant colonic tissues by an enriched PCR method *International Journal of Oncology* 4**,** 391-396.

SAEED, A., SHAROV, V., WHITE, J., LI, J., LIANG, W., BHAGABATI, N., BRAISTED, J., KLAPA, M., CURRIER, T. & THIAGARAJAN, M. 2003. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques,* 34**,** 374.

SANCHO, E., BATLLE, E. & CLEVERS, H. 2004. Signaling pathways in intestinal development and cancer. *Annu. Rev. Cell Dev. Biol.,* 20**,** 695-723.

SARTORE-BIANCHI, A., FIEUWS, S., VERONESE, S., MORONI, M., PERSONENI, N., FRATTINI, M., TORRI, V., CAPPUZZO, F., VANDER BORGHT, S. & MARTIN, V. 2012. Standardisation of EGFR FISH in colorectal cancer: results of an international interlaboratory reproducibility ring study. *Journal of clinical pathology,* 65**,** 218-223.

SCHMITT, M. W., KENNEDY, S. R., SALK, J. J., FOX, E. J., HIATT, J. B. & LOEB, L. A. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America,* 109**,** 14508-13.

SHENDURE, J. & JI, H. 2008. Next-generation DNA sequencing. *Nat Biotech,* 26**,** 1135-1145.

SLAUGHTER, D. P., SOUTHWICK, H. W. & SMEJKAL, W. 1953. "Field cancerization" in oral stratified squamous epithelium. Clinical implications of multicentric origin. *Cancer,* 6**,** 963-968.

SOBIN, L. H. & FLEMING, I. D. 1997. TNM classification of malignant tumors, (1997). *Cancer,* 80**,** 1803-1804.

SOREIDE, K., NEDREBO, B. S., KNAPP, J. C., GLOMSAKER, T. B., SOREIDE, J. A. & KORNER, H. 2009. Evolving molecular classification by genomic and proteomic biomarkers in colorectal cancer: potential implications for the surgical oncologist. *Surg Oncol,* 18**,** 31-50.

STEAD, L. F., SUTTON, K. M., TAYLOR, G. R., QUIRKE, P. & RABBITTS, P. 2013. Accurately Identifying Low-Allelic Fraction Variants in Single Samples with Next-Generation Sequencing: Applications in Tumor Subclone Resolution. *Human mutation*.

TAKAYAMA, T., KATSUKI, S., TAKAHASHI, Y., OHI, M., NOJIRI, S., SAKAMAKI, S., KATO, J., KOGAWA, K., MIYAKE, H. & NIITSU, Y. 1998. Aberrant Crypt Foci of the Colon as Precursors of Adenoma and Cancer. *New England Journal of Medicine,* 339**,** 1277-1284.

TAMURA, K., PETERSON, D., PETERSON, N., STECHER, G., NEI, M. & KUMAR, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution,* 28**,** 2731-2739.

TEO, I. A. & SHAUNAK, S. 1995. Polymerase chain reaction in situ: an appraisal of an emerging technique. *The Histochemical journal,* 27**,** 647-59.

THE CANCER GENOME ATLAS NETWORK 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature,* 487**,** 330-7.

THIRLWELL, C., WILL, O. C. C., DOMINGO, E., GRAHAM, T. A., MCDONALD, S. A. C., OUKRIF, D., JEFFREY, R., GORMAN, M., RODRIGUEZ–JUSTO, M., CHIN–ALEONG, J., CLARK, S. K., NOVELLI, M. R., JANKOWSKI, J. A., WRIGHT, N. A., TOMLINSON, I. P. M. & LEEDHAM, S. J. 2010. Clonality Assessment and Clonal Ordering of Individual Neoplastic Crypts Shows Polyclonality of Colorectal Adenomas. *Gastroenterology,* 138**,** 1441-1454.e7.

TIAN, S., ROEPMAN, P., POPOVICI, V., MICHAUT, M., MAJEWSKI, I., SALAZAR, R., SANTOS, C., ROSENBERG, R., NITSCHE, U. & MESKER, W. E. 2012. A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. *The Journal of pathology,* 228**,** 586-595.

TOMASETTI, C., VOGELSTEIN, B. & PARMIGIANI, G. 2013. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proceedings of the National Academy of Sciences,* 110**,** 1999-2004.

TRANGENOMIC INC. 2013. *REVEAL Kits* [Online]. Available: http://www.transgenomic.com/diagnostic-tools/genetic-analysis-kits/reveal-kits [Accessed 19th August 2013].

TSIATIS, A. C., NORRIS-KIRBY, A., RICH, R. G., HAFEZ, M. J., GOCKE, C. D., ESHLEMAN, J. R. & MURPHY, K. M. 2010. Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications. *The Journal of molecular diagnostics : JMD,* 12**,** 425-32.

TURNBULL JR, R. B., KYLE, K., WATSON, F. R. & SPRATT, J. 1967. Cancer of the colon: the influence of the no-touch isolation technic on survival rates. *Annals of surgery,* 166**,** 420.

VOELKERDING, K. V., DAMES, S. A. & DURTSCHI, J. D. 2009. Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry,* 55**,** 641-58.

VOGELSTEIN, B., FEARON, E. R., HAMILTON, S. R., KERN, S. E., PREISINGER, A. C., LEPPERT, M., NAKAMURA, Y., WHITE, R., SMITS, A. M. & BOS, J. L. 1988. Genetic alterations during colorectal-tumor development. *N Engl J Med,* 319**,** 525-32.

WATANABE, T., KOBUNAI, T., YAMAMOTO, Y., MATSUDA, K., ISHIHARA, S., NOZAWA, K., YAMADA, H., HAYAMA, T., INOUE, E., TAMURA, J., IINUMA, H., AKIYOSHI, T. & MUTO, T. 2012. Chromosomal instability (CIN) phenotype, CIN high or CIN low, predicts survival for colorectal cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 30**,** 2256-64.

WEST, N. P., DATTANI, M., MCSHANE, P., HUTCHINS, G., GRABSCH, J., MUELLER, W., TREANOR, D., QUIRKE, P. & GRABSCH, H. 2010. The proportion of tumour cells is an independent predictor for survival in colorectal cancer patients. *British journal of cancer,* 102**,** 1519-1523.

WHITEHEAD, R. H., BROWN, A. & BHATHAL, P. S. 1987. A method for the isolation and culture of human colonic crypts in collagen gels. *In vitro cellular*

& developmental biology : journal of the Tissue Culture Association, 23**,** 436-42.

WILL, O. C., LEEDHAM, S. J., ELIA, G., PHILLIPS, R. K., CLARK, S. K. & TOMLINSON, I. P. 2010. Location in the large bowel influences the APC mutations observed in FAP adenomas. *Familial cancer,* 9**,** 389-93.

WILLIAMS, C., PONTÉN, F., MOBERG, C., SÖDERKVIST, P., UHLÉN, M., PONTÉN, J., SITBON, G. & LUNDEBERG, J. 1999. A High Frequency of Sequence Alterations Is Due to Formalin Fixation of Archival Specimens. *The American journal of pathology,* 155**,** 1467-1471.

WOODFORD-RICHENS, K. L., ROWAN, A. J., GORMAN, P., HALFORD, S., BICKNELL, D. C., WASAN, H. S., ROYLANCE, R. R., BODMER, W. F. & TOMLINSON, I. P. M. 2001. SMAD4 mutations in colorectal cancer probably occur before chromosomal instability, but after divergence of the microsatellite instability pathway. *Proceedings of the National Academy of Sciences,* 98**,** 9719-9723.

XIE, T., D'ARIO, G., LAMB, J. R., MARTIN, E., WANG, K., TEJPAR, S., DELORENZI, M., BOSMAN, F. T., ROTH, A. D. & YAN, P. 2012. A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. *PloS one,* 7**,** e42001.

YAMADA, S., YASHIRO, M., MAEDA, K., NISHIGUCHI, Y. & HIRAKAWA, K. 2005. A novel high-specificity approach for colorectal neoplasia: Detection of K-ras2 oncogene mutation in normal mucosa. *International journal of cancer. Journal international du cancer,* 113**,** 1015-21.

YAMADA, Y., YOSHIMI, N., HIROSE, Y., KAWABATA, K., MATSUNAGA, K., SHIMIZU, M., HARA, A. & MORI, H. 2000. Frequent beta-catenin gene mutations and accumulations of the protein in the putative preneoplastic lesions lacking macroscopic aberrant crypt foci appearance, in rat colon carcinogenesis. *Cancer research,* 60**,** 3323-7.

YAMAMOTO, E., SUZUKI, H., YAMANO, H.-O., MARUYAMA, R., NOJIMA, M., KAMIMAE, S., SAWADA, T., ASHIDA, M., YOSHIKAWA, K., KIMURA, T., TAKAGI, R., HARADA, T., SUZUKI, R., SATO, A., KAI, M., SASAKI, Y., TOKINO, T., SUGAI, T., IMAI, K., SHINOMURA, Y. & TOYOTA, M. 2012. Molecular Dissection of Premalignant Colorectal Lesions Reveals Early Onset of the CpG Island Methylator Phenotype. *The American journal of pathology,* 181**,** 1847-1861.

YOST, S. E., SMITH, E. N., SCHWAB, R. B., BAO, L., JUNG, H., WANG, X., VOEST, E., PIERCE, J. P., MESSER, K., PARKER, B. A., HARISMENDY, O. & FRAZER, K. A. 2012. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic acids research,* 40**,** e107.

ZEKI, S. S., GRAHAM, T. A. & WRIGHT, N. A. 2011. Stem cells and their implications for colorectal cancer. *Nature reviews. Gastroenterology & hepatology,* 8**,** 90-100.

ZHANG, H., NORDENSKJÖLD, B., DUFMATS, M., SÖDERKVIST, P. & SUN, X. F. 1998. K-ras mutations in colorectal adenocarcinomas and neighbouring transitional mucosa. *European journal of cancer (Oxford, England : 1990),* 34**,** 2053-2057.

ZHU, D., KEOHAVONG, P., FINKELSTEIN, S. D., SWALSKY, P., BAKKER, A., WEISSFELD, J., SRIVASTAVA, S. & WHITESIDE, T. L. 1997. K-ras gene mutations in normal colorectal tissues from K-ras mutation-positive colorectal cancer patients. *Cancer Res,* 57**,** 2485-92.

6 Appendix

6.1 Solution recipes

6.1.1 Hanks Buffered Salt Solution (HBSS)

HBSS Stock Solutions:

Solution 1

Dissolve the following in 90ml of distilled H_2O:

- 8.0g NaCl
- 0.4g KCl

Make up final volume to 100ml with distilled H_2O

Solution 2

Dissolve the following in 90ml of distilled H_2O:

- 0.358g anhydrous Na_2HPO_4
- 0.60g KH_2PO_4

Make up final volume to 100ml with distilled H_2O

Solution 3

Dissolve 0.72g of $CaCl_2$ in 50ml of distilled H_2O

Solution 4

Dissolve 1.23g $MgSO_4.7H_2O$ in 50ml of distilled H_2O

Solution 5

Dissolve 0.35 g $NaHCO_3$ in 10ml of distilled H_2O

HBSS Premix

- 10.0 ml Solution 1
- 1ml Solution 2
- 1ml Solution 3
- 86.0 ml distilled H_2O
- 1ml Solution 4

HBSS

- 9.9 ml Hank's Premix
- 0.1ml Solution 5

6.1.2 Hanks Buffered Salt Solution Calcium and Magnesium Free (HBSS CMF)

Dissolve the following in 900ml of distilled H_2O:

- 8g NaCl
- 0.4g KCl
- 0.06g KH_2PO_4
- 0.35g $NaHCO_3$
- 0.012g $Na_2HPO_4.12H_2O$

Adjust the pH to 7.4 with 2M NaOH

Make up final volume to 1000ml with distilled H_2O

6.1.3 HBSS CMF + EDTA

Dissolve 2.23g EDTA in 150ml HBSS CMF

Make up final volume to 200ml with HBSS CMF

6.2 Spot counting analysis

	Section number	Number of spots counted	Number of spots on epithelium	Percentage epithelium (%)
Case 1	1	292	158	54.1
	2	294	139	47.3
	3	283	127	44.9
	4	275	137	49.8
	5	293	151	51.5
Case 2	1	276	137	49.6
	2	280	156	55.7
	3	285	150	52.6
	4	260	153	58.8
	5	276	149	54.0
Case 3	1	272	112	38.9
	2	285	126	40.5
	3	253	142	56.1
	4	276	143	51.8

Table 51. Spotcounting analysis for epithelial cell content for 14 frozen sections of normal colorectal mucosa.

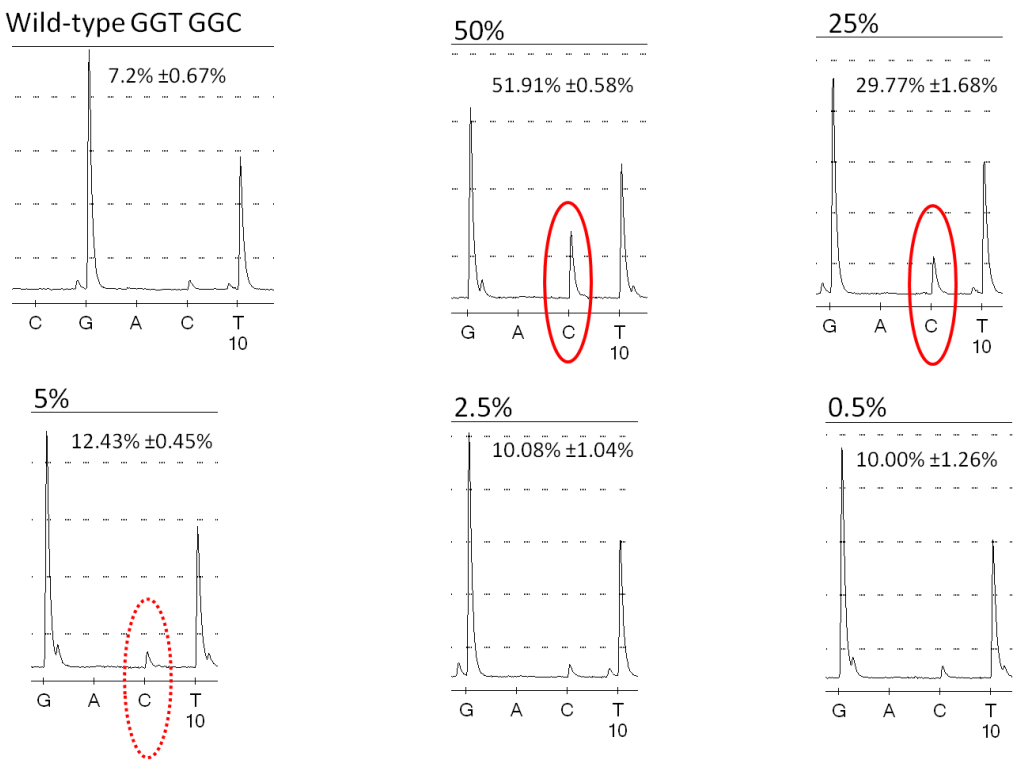
	Section number	Spots on lamina propria (%)	Spots on crypt Lumen (%)	Spots on muscle (%)	Spots on mucin (%)
Case 1	1	34.9	8.9	2.1	0.7
	2	37.7	11.6	3.7	-
	3	35.0	18.0	1.8	-
	4	33.5	14.9	1.8	-
	5	34.5	11.9	1.4	-
Case 2	1	37.7	12.0	0.7	-
	2	32.1	12.1	-	-
	3	33.7	11.6	1.1	-
	4	31.5	9.2	0.4	-
	5	33.3	11.6	1.1	-
Case 3	1	45.6	10.3	2.9	-
	2	34.7	18.2	2.5	0.4
	3	32.4	9.9	1.2	0.4
	4	36.6	10.1	1.1	0.4

Table 52. Spotcounting analysis for other cellular components, 14 frozen sections of normal colorectal mucosa.

6.3 Pyrosequencing of serial dilutions of mutant KRAS

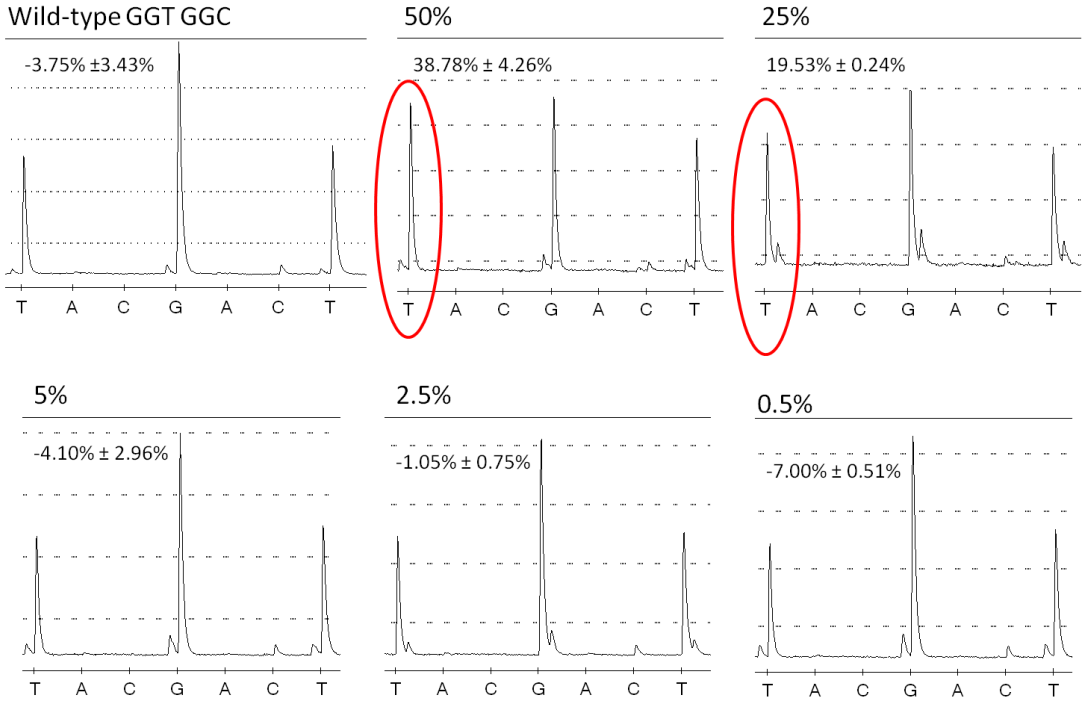
A

12A c.35G>C GCT GGC



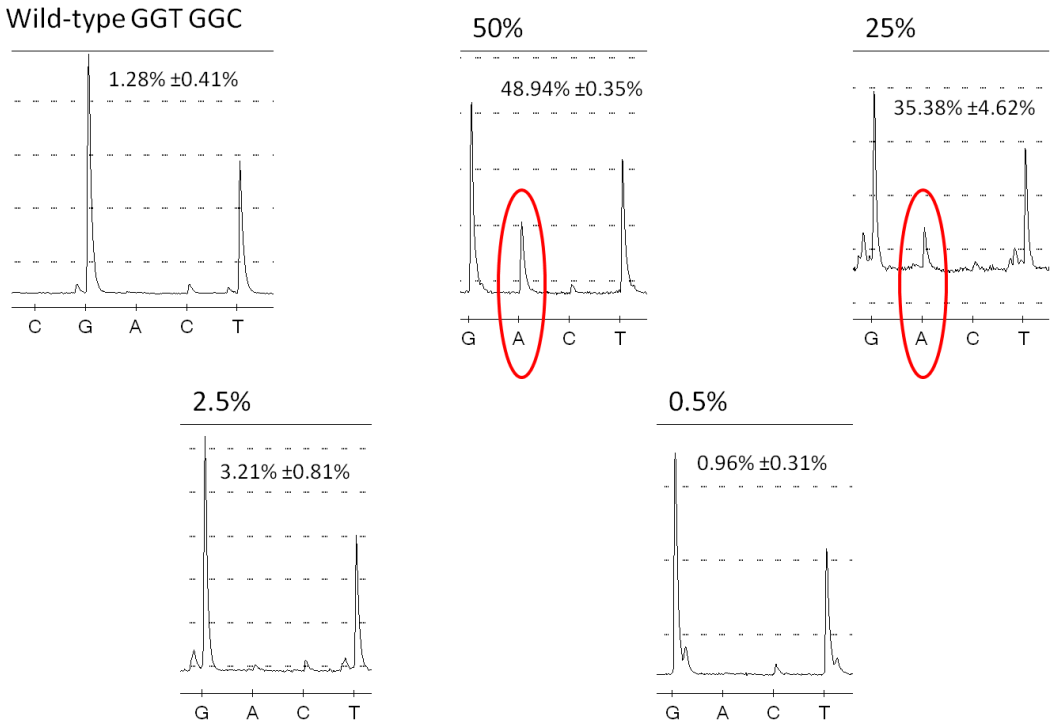
B

12C c.34G>T TGT GGC



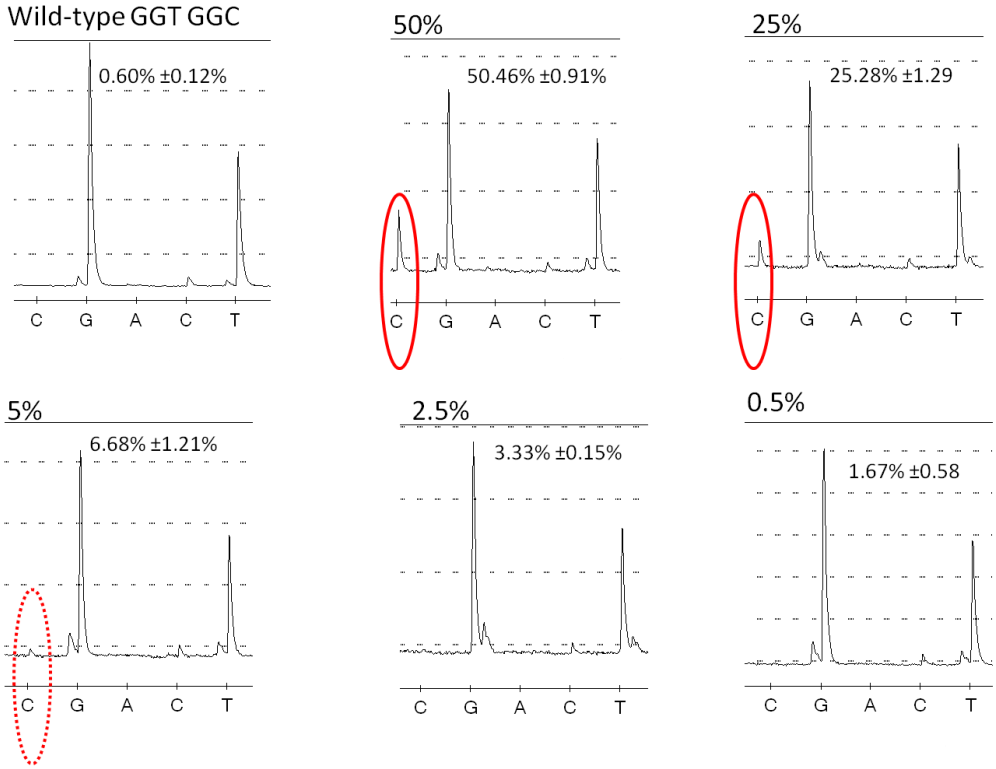
C

12D c.35G>A GAT GGC



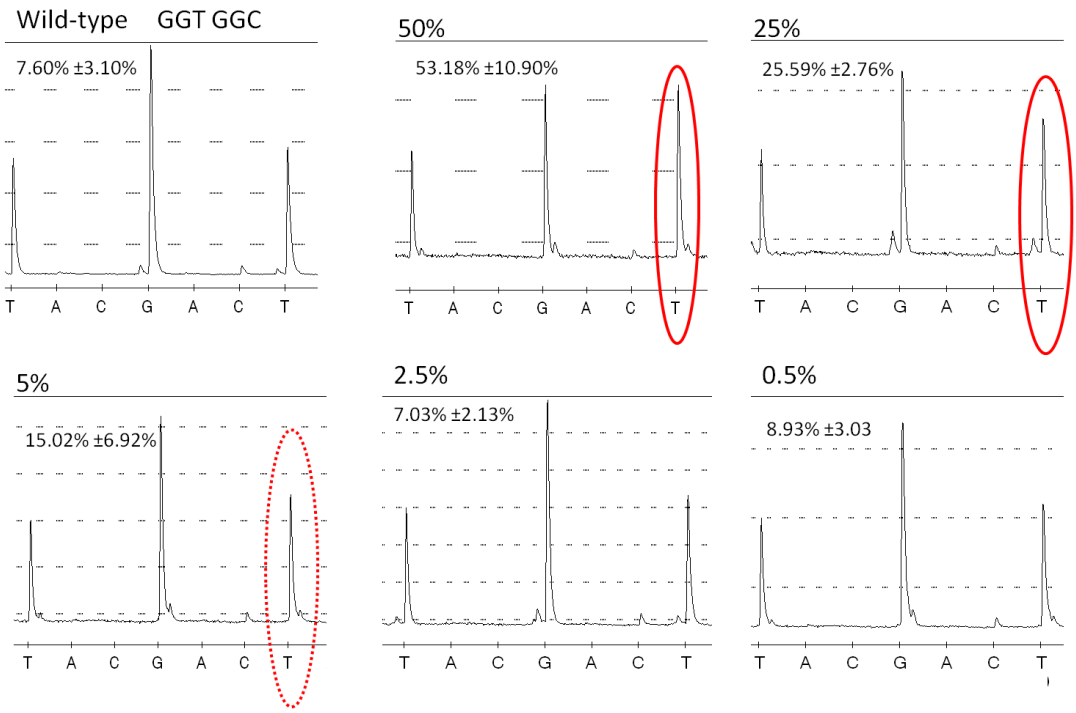
D

12R c.34 G>C CGT GGC



E

12V c.35G>T GTT GGC



F

13D c.38G>A GGT GAC

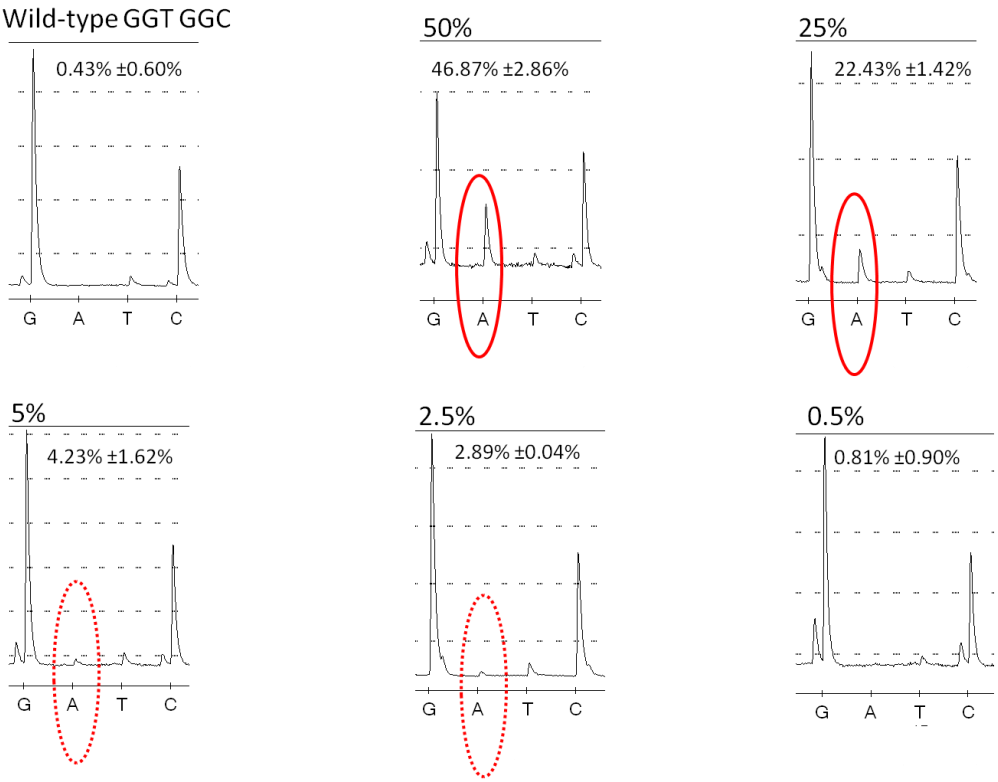
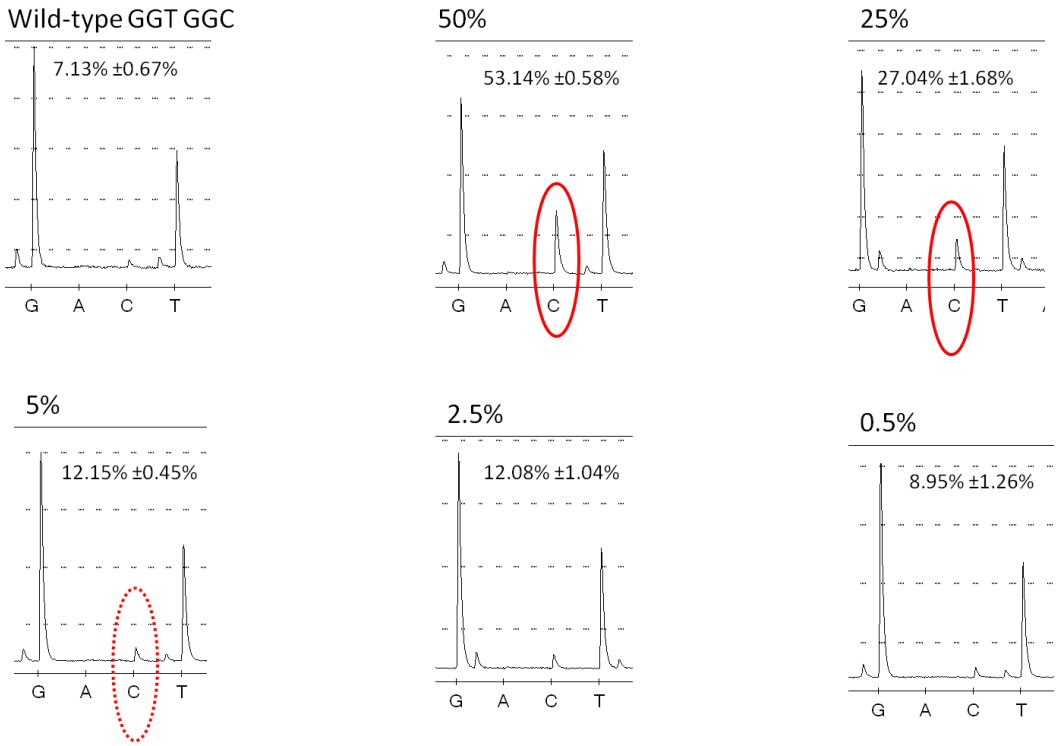


Figure 91. A-F Pyrosequencing of serial dilutions of 6 KRAS 12+13mutations, run2.

A

12A c.35G>C GCT GGC



B

12C c.34G>T TGT GGC

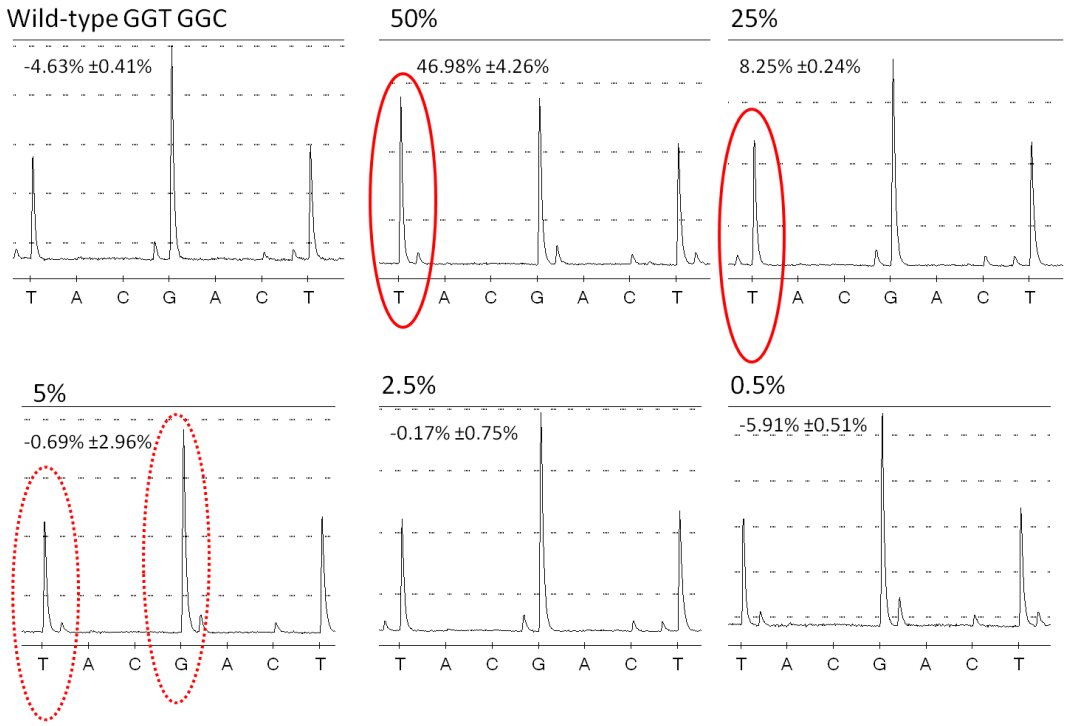


Figure 92. A-B Pyrosequencing of serial dilutions of 6 KRAS 12+13mutations, run3.

6.4 Scripts

6.4.1 Reference fasta file

This is the reference fasta file that was used for analysing targeted amplicons. Firstly reads were aligned to this reference using BWA alignment software. This file then acted as the reference input for the AllFreChecker.pl script. Pseudogene sequences were added to allow for these reads to be removed from the analysis.

```
>KRAS1213
GGCCTGCTGAAAATGACTGAATATAAACTTGTGGTAGTTGGAGCTGGTGGCGTAGGCAAGAGTGCCTTGACGA
TACAGCT
>KRAS1213_PSEUDO
GGCCTGCTGAAAATGACTGAATATAAACTTGCGGTAGTTGGAGCTGGTGGCGTAAGCAAAAGTGTCTTGACGA
TACAGCT
>KRAS61
AATTGATGGAGAAACCTGTCTCTTGGATATTCTCGACACAGCAGGTCAAGAGGAGTACAGTGCAATGAGGGAC
CAGTACA
>KRAS61_PSEUDO
AATTGATGGAGAAACCTGTCTCTTGGATATTCTTGACACAACAGGTCAAGAAGAGTACAATGCAATGAGGACCA
GTACA
>NRAS1213
CTTGCTGGTGTGAAATGACTGAGTACAAACTGGTGGTGGTTGGAGCAGGTGGTGTTGGGAAAAGCGCACTGA
CAATCCA
>NRAS61
GAAACCTGTTTGTTGGACATACTGGATACAGCTGGACAAGAAGAGTACAGTGCCATGAGAGACCAATACATGA
GGACAGG
>BRAF
TGAAGACCTCACAGTAAAAATAGGTGATTTTGGTCTAGCTACAGTGAAATCTCGATGGAGTGGGTCCCATCAGT
TTGAAC
>BRAF_PSEUDO
TGAAGACCTCACAGTGGAAATAGGTGATTTTGGTCTAGCCACAGTGAAATCTTGATGGAGTGGGTCCCATCAG
TTTGAAC
```

6.4.2 AllFreqChecker.pl

This script reports the number of A, G, T, C and N bases at each position within the reads of a SAM FILE for specified targets. Targets are inputted using the above reference fasta file containing the target sequences. The command line for the script is:

perl allele_freq.pl <reference file.fa> <input.sam>

The output is an excel file that shows the number of A, G, T, C and N bases at each position in the amplicon as well as the total coverage and finally the percentage of bases that do not align that that position. This output file can be easily manipulated in excel to produce a final report.

```
#!usr/bin/perl
use strict;
use warnings;

## Kate's super-duper script to give the allele frequencies from a fasta file
## usage = perl allele_freq.pl <ref file.fa> <input.sam>

print "usage = perl allele_freq.pl <ref file.fa> <input.sam>\n";

my $ref = $ARGV[0];
my $in = $ARGV[1];
my $out = "$in"."pfreqs.xls";
#my ($ref, $in) = shift(@ARGV);

open (REF, '<', $ref);
open (OUT, '>', $out);


my $line;
my $header;
my $amplicon;
my %seqs= ();

### READ FASTA FILE INTO HASH
while ($line = <REF>) {
        chomp $line;
        if ($line =~ /^>/) {
                $header = substr($line, 1);
    $header =~ s/\r$//;
        } else {
                $amplicon = $line;
        $seqs{$header} = $amplicon;
        };
};

### up to here all is ok!

foreach my $gene (keys %seqs) {
        next if ($gene =~m/PSEUDO/g);
   open (IN, '<', $in);
   my @countT = ("0") x (length $seqs{$gene});
   my @countG = ("0") x (length $seqs{$gene});
   my @countC = ("0") x (length $seqs{$gene});
   my @countA = ("0") x (length $seqs{$gene});
   my @countN = ("0") x (length $seqs{$gene});
   my @countTO = ("0") x (length $seqs{$gene});
        my $average;
        my $std;
        my @nonmatch;
        my $entry;
        my $result;
        print OUT "$gene\n";
         print  OUT "\tA\tG\tT\tC\tN\tTotal\t%Error\n";
        my $read;
        while ($read = <IN>) {
    chomp $read;
    next if ($read =~ m/^\@/);
    my @field = split(/\t/, $read);
                if (($field[5] eq "80M") and ($field[4] > 1) ) {
                        my $align = $field[2];
                        my $read = $field[9];
                        my $loc = $field[3];
                        my $n = $loc-1;
                        if ($align eq $gene) {
                                my $amp = $seqs{$gene};
                                my ($countT_ar, $countG_ar, $countC_ar, $countA_ar, $countN_ar,
$countTO_ar) = freq ($n, $amp, $read);
                                my @localCountT = @{$countT_ar};
                                my @localCountG = @{$countG_ar};
                                my @localCountC = @{$countC_ar};
                                my @localCountA = @{$countA_ar};
                                my @localCountN = @{$countN_ar};
                                my @localCountTO = @{$countTO_ar};
                                @countT = addCount(\@countT, \@localCountT);
                                @countG = addCount(\@countG, \@localCountG);
```

```
                                        @countC = addCount(\@countC, \@localCountC);
                                        @countA = addCount(\@countA, \@localCountA);
                                        @countN = addCount(\@countN, \@localCountN);
                                        @countTO = addCount(\@countTO, \@localCountTO);
                                };
                        };
                };
   close IN;
                for (my $a=0; $a<@countTO-1; $a++) {
                        my $d = $a+1;
                        if ($countTO[$a]>0) {
                                $nonmatch[$a] =
($countA[$a]+$countG[$a]+$countT[$a]+$countC[$a]+$countN[$a])/ $countTO[$a] * 100;
                                print OUT
"$d"."\t"."$countA[$a]"."\t"."$countG[$a]"."\t"."$countT[$a]"."\t"."$countC[$a]"."\t"."$countN[$a]"."\t"."$countTO[$a]"."\t
"."$nonmatch[$a]"."\n";
                        };
                };

                print OUT "\n";
};


sub freq {
                my ($n, $amp, $read) = @_;
                my @bases = split (//, $read);
                my @countT = ("0") x (length $amp);
                my @countG = ("0") x (length $amp);
                my @countC = ("0") x (length $amp);
                my @countA = ("0") x (length $amp);
                my @countN = ("0") x (length $amp);
                my @countTO = ("0") x (length $amp);
                for (my $x=0; $x<@bases; $x++) {
                        $countTO[$n+$x]++;
                        my @seq = split (//, $amp);
                        if (defined($seq[$n+$x])) {
                                if ($bases[$x] ne $seq[$n+$x]) {
                                        if ($bases[$x] eq "G") {
                                                $countG[$n+$x]++;
                };
                if ($bases[$x] eq "T") {
                                                $countT[$n+$x]++;
                                        };
                if ($bases[$x] eq "C") {
                                                $countC[$n+$x]++;
                                        };
                                        if ($bases[$x] eq "A") {
                                                $countA[$n+$x]++;
                                        };
                                        if ($bases[$x] eq "N") {
                                                $countN[$n+$x]++;
                                        };
                                };
                        };
                };
                return (\@countT,\@countG,\@countC,\@countA,\@countN,\@countTO);
};

sub addCount {
                my ($count_ar, $localcount_ar) = @_;
                my @count = @{$count_ar};
                my @localcount = @{$localcount_ar};
                for (my $n=0; $n<@count; $n++) {
                        $count[$n] = $localcount[$n] + $count[$n];
                };
                return @count;
};


sub average {
                my @nonmatch = @_;
                my $total = 0;
                ($total +=$_) for @nonmatch;
                my $average = ($total / @nonmatch);
                return $average;
```

```
};


sub stdev{
	my($average, @data,) = @_;
	my $sqtotal = 0;
	foreach(@data) {
		$sqtotal += ($average-$_) ** 2;
	}
	my $std = ($sqtotal / (@data-1)) ** 0.5;
	return $std;
};
```

6.4.3 Var_an.pl

This perl script allows for the output of VarScan 2 to be read more easily, opened in excel and acts as input for the matrix script mut_matrix.pl. It outputs an excel file that shows the Aligned Chromosome, position in that chromosome, the reference base at that position, the variant base at that position and finally the allele frequency of that variant.

```
#!usr/bin/perl
use strict;
use warnings;

my $in = shift(@ARGV);
my @name = split (/\./, $in);
my $out = "$name[0]".".xls";
open (IN, '<', $in);
open (OUT, '>', $out);

print OUT "Chromosome\tPosition\tRef\tVar\tAllFreq\n";
my $line;

while ($line = <IN>) {
	chomp $line;
	next if ($line =~ m/^C/);
	my @field = split (/\t/, $line);
	my @freq = split (/:/, $field[4]);
	print OUT "$field[0]\t$field[1]\t$field[2]\t$field[3]\t$freq[4]\n";
```

6.4.4 Mut_matrix_maker.pl

This perl script contains a mutational matrix that can be read in excel with shared mutational sites as the header and samples as the columns. Each cell within the matrix shows what base is present for each sample at that position. The samples are inputted as the output files from the var_an.pl script. The input also requires a reference file of shared mutations. Each line of the reference file has the mutational position followed by a tab followed by the WT base at that position. The command line for this script is as follows:

Perl mut_matrix_maker.pl <mutational reference file.txt> <sample file.xls>

```
#! usr/bin/perl

use strict;
use warnings;

my $in = $ARGV[1];
my $ref = $ARGV[0];
open (REF, '<', $ref);
open (IN, '<', $in);
my $out = "mutmatrix.txt";
open (OUT, '>>', $out);

my %muts;
my $entry;

while ($entry = <REF>) {
	chomp $entry;
	my @split = split (/\t/, $entry);
	$muts{$split[0]} = $split[1];
};

my $line;
my @list;
my %var;

while ($line = <IN>) {
	chomp $line;
	my @field = split (/\t/, $line);
	my $match = $field[6];
	push (@list, $match);
	$var{$match} = $field[4];
};

my %counter;
foreach my $rec (keys %muts) {
	$counter{$rec} = 0;
	for (my $n=0; $n<@list; $n++) {
		if($list[$n] =~ m/$entry/) {
			$counter{$entry}++;
		};
	};
};

print OUT "$in\t";
foreach my $record (keys %counter) {
	print OUT "$record\t";
};

print OUT "\n";

foreach my $record (keys %counter) {
	if ($counter{$record} == 0) {
		print OUT "\t$muts{$record}";
	} else {
		print OUT "\t$var{$record}";
	};
};

print OUT "\n";
```

6.4.5 CNVmatrix1.pl

This script is the first script to create a matrix for copy number variant. It checks the CNV ratio for each window in the sample file and records those that are above 1.2 or below 0.8. It also detects the centromeric and telomeric regions and removes these from the analysis. The input is a txt file with the window positions and the copy number ratio which is produced from the CNAnorm programme. The input also requires a reference text file that specifies the length of each chromosome and the position of the centromere. This can be created in excel from the sample output files. The output file from this script is a text file that can be inputted into the CNVmatrix2.pl script plus a target file that is used for the CNV2matrix2.pl scriptThe command line for this script is:

CNVmatrix1.pl <reference file.txt> <sample file.txt>

```
#! usr/bin/perl

use strict;
use warnings;

my $in = $ARGV[1];
my $ref = $ARGV[0];
my @name = split (/\./, $in);
open (REF, '<', $ref);
my %end1;
my %end2;
my %cen1;
my %cen2;
open (IN, '<', $in);
my $out = "$in.mi";
open (OUT, '>', $out);
my $newref = "targetfile.txt";
open (OOUT, '>>', $newref);

my $entry;
while ($entry = <REF>) {
	chomp $entry;
	my @col = split (/\t/, $entry);
	$end1{$col[0]} = $col[1];
	$end2{$col[0]} = $col[2];
	$cen1{$col[0]} = $col[3];
	$cen2{$col[0]} = $col[4];
};

my @oref;

#print OUT "Chr\tStart\tEnd\tScore\t\n";

my $line;
while ($line = <IN>) {
	chomp $line;
	next if 1..1;
	my @field = split (/\t/, $line);
	next if ($field[6] eq "NA");
	next if ($field[0] =~ m/Y/);
	next if ($field[0] =~ m/X/);
	if ( 0.9 > $field[6] or $field[6] > 1.1) {
		foreach my $rec (keys %end1) {
			if ($field[0] =~m/$rec/ and $field[1] > $end1{$rec}) {
				foreach my $rec1 (keys %cen1) {
					if ($field[0] =~m/$rec1/ and $field[1] < $cen1{$rec1}) {
```

```
                                print OUT "$name[0]\t$field[0]\t$field[1]\t$field[6]\n";
                                my $join = "$field[0]".":"."$field[1]";
                                push (@oref, $join);
                        };
                };
        };
        foreach my $rec3 (keys %cen2) {
                if ($field[0] =~m/$rec3/ and $field[1] > $cen2{$rec3}) {
                        foreach my $rec4 (keys %end2) {
                                if ($field[0] =~m/$rec4/ and $field[1] < $end2{$rec4}) {
                                        print OUT "$name[0]\t$field[0]\t$field[1]\t$field[6]\n";
                                        my $ojoin = "$field[0]".":"."$field[1]";
                                         push (@oref, $ojoin);
                                };
                        };
                };
        };
    };
};

my @final = uniq2(@oref);
for (my $n=0; $n<@final; $n++) {
       print OOUT "$final[$n]\n";
};


sub uniq2 {
   my %seen = ();
   my @r = ();
   foreach my $a (@_) {
      unless ($seen{$a}) {
         push @r, $a;
         $seen{$a} = 1;
      }
   }
   return @r;
}
```

6.4.6 CNVmatrix2.pl

This script uses the text file from the CNVmatrix1.pl script generated for each sample, plus the target file created from this first script. The output of CNVmatrix2.pl is an excel matrix with chromosomal location as the header (split according to windows from CNAnorm programme) and samples as the rows. Each cell indicates whether the copy number at that positon for that sample is normal (0), greater than 1.2 (1) or less than 0.8 (2). The command line for this script is as follows:

CNVmatrix2.pl <target file from CNVmatrix1> <sample file from CNV matrix1>

```
 #! usr/bin/perl

use strict;
use warnings;

my $in = $ARGV[1];
my $ref = $ARGV[0];
open (REF, '<', $ref);
open (IN, '<', $in);
my $out = "matrix.txt";
open (OUT, '>>', $out);

my @ref = <REF>;
chomp @ref;

my $line;
my @list;
my %scores;

while ($line = <IN>) {
	chomp $line;
	my @field = split (/\t/, $line);
	my $match = "$field[1]".":"."$field[2]";
	push (@list, $match);
	$scores{$match} = $field[3];
};

my %counter;

for my $entry (@ref) {
	$counter{$entry} =0;
	for (my $n=0; $n<@list; $n++) {
		if ($list[$n] =~ m/$entry/) {
			$counter{$entry}++;
		};
	};
};

print OUT "$in\t";
foreach my $record (keys %counter) {
	print OUT "$record\t";
};

print OUT "\n";

foreach my $record (keys %counter) {
	if ($counter{$record} == 0) {
		print OUT "\t0";
	} elsif ($scores{$record} > 1) {
		print OUT "\t1";
	} elsif ($scores{$record} < 1) {
		print OUT "\t2";
	};
};

print OUT "\n";
```